Some Statistical Models for Prediction

Jonathan Auerbach

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Some Statistical Models for Prediction

Jonathan Auerbach

This dissertation examines the use of statistical models for prediction. Examples are drawn from public policy and chosen because they represent pressing problems facing U.S. governments at the local, state, and federal level. The first five chapters provide examples where the perfunctory use of linear models, the prediction tool of choice in government, failed to produce reasonable predictions. Methodological flaws are identified, and more accurate models are proposed that draw on advances in statistics, data science, and machine learning. Chapter 1 examines skyscraper construction, where the normality assumption is violated and extreme value analysis is more appropriate. Chapters 2 and 3 examine presidential approval and voting (a leading measure of civic participation), where the non-collinearity assumption is violated and an index model is more appropriate. Chapter 4 examines changes in temperature sensitivity due to global warming, where the linearity assumption is violated and a first-hitting-time model is more appropriate. Chapter 5 examines the crime rate, where the independence assumption is violated and a block model is more appropriate. The last chapter provides an example where simple linear regression was overlooked as providing a sensible solution. Chapter 6 examines traffic fatalities, where the linear assumption provides a better predictor than the more popular non-linear probability model, logistic regression. A theoretical connection is established between the linear probability model, the influence score, and the predictivity.

Table of Contents

List of Tables vi

List of Figures vii

Chapter 1: Forecasting the Urban Skyline with Extreme Value Theory 1

1.1 Introduction 1

1.2 Modeling the Uncertainties of Skyscraper Development 3

1.3 Predicting the Quantity of Skyscrapers Completed by 2050 8

1.4 Predicting the Height of Skyscrapers Completed by 2050 10

1.5 Predicting the Number of Floors in Skyscrapers Completed by 2050 15

1.6 Discussion 25

1.7 Conclusion 27

1.8 Appendix 28

1.9 References 32

9. References 32

Chapter 2: Forecasting the 2020 American Presidential Election with a Generational Voting Model 37

2.1 Introduction 37

2.2 We use the Gallup Organization's presidential approval rating time series to compare voters across generations. 38

2.3 We model partisan preferences as a running tally of impressions left by the political events. . . . 40

2.4 Our model suggests the political events of the Trump administration will have smaller influence on average than the events experienced by previous generations of the same age. . . . 42

2.5 References . . . 49

Chapter 3: A Deeper Look at the Generational Voting Model . . . 50

3.1 Introduction . . . 50

3.2 Data and Preliminary Evidence . . . 53

3.3 Statistical Model . . . 59

3.4 Model Results . . . 63

3.4.1 Generational Effect . . . 63

3.4.2 Election Effect . . . 65

3.4.3 Explanatory Power . . . 66

3.5 Generations of Presidential Voting . . . 67

3.5.1 New Deal Democrats . . . 68

3.5.2 Eisenhower Republicans . . . 69

3.5.3 1960s Liberals . . . 71

3.5.4 Reagan Conservatives . . . 74

3.5.5 Millennials . . . 76

3.5.6 The Changing White Electorate . . . 78

3.6 Discussion . . . 80

3.7 References . . . 83

Chapter 4: Forecasting Declining Temperature Sensitivity with a First-hitting-time Model . 86

4.1 Introduction 86

4.2 A first-hitting-time model of leafout 87

4.2.1 Scenario 1: Using average daily temperature until the leafout date 89

4.2.2 Scenario 2: Using average daily temperature over a fixed window 91

4.3 Simulations 93

4.4 Long-term empirical data from PEP725 95

4.5 References 103

Chapter 5: Forecasting Migration Patterns with an Infinitesimal Block Model 105

5.1 Introduction 105

5.2 Model 107

5.2.1 Derivation 108

5.2.2 Identification 109

5.2.3 Specification 111

5.3 Inference 111

5.3.1 Parameterization 112

5.3.2 Update Schedule 113

5.4 Evaluation 113

5.4.1 Simulation 114

5.4.2 IRS Data 114

5.4.3 ACS Data 115

5.5 Extension 118

5.6 Conclusion . . . 120
5.7 Appendix . . . 123
5.7.1 Inhomogenous Extension . . . 123
5.7.2 Alternative Update Schedule . . . 124
5.7.3 Algorithm 1 Acceptance Probability . . . 124
5.7.4 High dimensional approximation . . . 127
5.7.5 Simulation . . . 130
5.7.6 Real Data . . . 130
5.8 References . . . 132

Chapter 6: Forecasting Pedestrian Fatalities with the Linear Probability Model . . . 136
6.1 Linear Probability Model for Partitions . . . 137
6.1.1 Interpreting the joint LPM . . . 138
6.1.2 Interpreting the conditional LPM . . . 139
6.2 The influence score assesses the fit of the LPM . . . 141
6.2.1 The general relationship between the influence score and predictivity . . . 142
6.2.2 Influence score as a measure of the conditional LPM's fit. . . . 143
6.3 Predicting NYC Traffic Fatalities . . . 145
6.3.1 Vision Zero dataset . . . 146
6.4 Appendix . . . 149
6.4.1 The 2x2 Contingency Table. . . . 149
6.4.2 Summarizing a 2x2 Contingency Table by its predictivity . . . 149
6.4.3 Alternative measures of uncertainty and association . . . 155

6.4.4 Connection between predictivity and co-predictivity. 159

6.4.5 Connection between predictivity and deviation. 162

6.4.6 Connection between the odds and deviation. 164

6.4.7 Connection between predictivity and odds. 166

6.4.8 Connection between predictivity and regression slope. 166

6.4.9 Derivations of the Influence Score 174

6.5 References 179

List of Tables

4.1 Climate and phenology statistics for two species (Betula pendula, Fagus sylvatica, across 45 and 47 sites respectively) from the PEP725 data across all sites with continuous data from 1950-1960 and 2000-2010. ST is spring temperature from 1 March to 30 April, ST.leafout is temperature 30 days before leafout, and GDD is growing degree days 30 days before leafout. Slope represents the estimated sensitivity using untransformed leafout and ST, while log-slope represents the estimated sensitivity using log(leafout) and log(ST). We calculated all metrics for for each species x site x 10 year period before taking mean or variance estimates. See also Fig. 4. 97

4.2 Climate and phenology statistics for two species (Betula pendula, Fagus sylvatica, across 17 and 24 sites respectively) from the PEP725 data across all sites with continuous data from 1950-2010. ST is spring temperature from 1 March to 30 April, ST.leafout is temperature 30 days before leafout, and GDD is growing degree days 30 days before leafout. Slope represents the estimated sensitivity using untransformed leafout and ST, while log-slope represents the estimated sensitivity using log(leafout) and log(ST). We calculated all metrics for for each species x site x 20 year period before taking mean or variance estimates. See also Fig. 5. 97

List of Figures

1.1 We predict the quantity of skyscrapers completed by 2050, thirty-three years after the data was collected at the end of 2017. A Poisson regression model suggests the number of skyscrapers will grow at a rate of eight percent a year. 11

1.2 We approximate the height distribution of tall skyscrapers using a GPD. The left panel shows maximum likelihood estimates and confidence intervals (50 and 95 percent) of the GPD shape parameter, excluding observations below a sequence of thresholds $\{u_i\}$. The right panel shows a Mean Excess plot (top) and a Hill plot (bottom) 16

1.3 We demonstrate the distribution of skyscraper heights has changed little over time by dividing extremely tall skyscrapers into sextiles based on the year completed and constructing a p-p plot for each sextile. The distribution of theoretical heights is the generalized Pareto distribution with parameters estimated by maximum likelihood. 17

1.4 The median height of extremely tall skyscrapers (blue points) has not increased significantly over the last forty years. A median regression indicates extremely tall skyscrapers grow .19 meters each year (p-value = .25). A 95 percent confidence interval for the median regression line is shaded blue. The median height of tall skyscrapers (red) is shown for comparison. 18

1.5 We predict the height of skyscrapers completed by 2050, thirty-three years after the data was collected at the end of 2017. The simulated density (blue) suggests the tallest building in the world is unlikely to exceed one mile (dashed line on right side). However, it will almost certainly be taller than the current tallest building, the Burj Khalifa (828 meters, dashed line on left side) and likely taller than the Jeddah Tower (one thousand meters), expected for completion in 2020. Had the same simulation been conducted in 1984, the density (red) would have found the tallest skyscraper in 2017 to be between six hundred and twelve hundred meters. . . 19

1.6 We estimate the number of floors in the tallest skyscrapers. The top-left panel shows the height and number of floors of tall skyscrapers—those exceeding 150 meters and 40 floors. The line represents the 3.8 meter rise per floor of the typical tall skyscraper. This ratio is not preserved as height increases. The other three panels compare the estimated conditional density of the number of floors in a skyscraper given its height to actual skyscraper designs. 23

1.7 We estimate the height and number of floors of the extremely tall skyscrapers that will be completed in major cities by 2050. Empirical medians in the top panel likely overestimate between-city variation. Model estimated medians in the bottom panel compromise between the noisy empirical medians in the top panel and the accurate, but global median estimated by the non-hierarchical model. The line represents the 3.8 meter rise per floor of the typical tall skyscraper. 24

1.8 We approximate the distribution of the number of floors of tall skyscrapers using a GPD. The left panel shows maximum likelihood estimates and confidence intervals (50 and 95 percent) of the GPD shape parameter, excluding observations below a sequence of thresholds $\{u_i\}$. The right panel shows a Mean Excess plot (top) and a Hill plot (bottom) 29

1.9 We demonstrate the distribution of the number of floors has changed little over time by dividing extremely tall skyscrapers into sextiles based on the year completed and constructing a p-p plot for each sextile. The distribution of theoretical floors is the generalized Pareto distribution with parameters estimated by maximum likelihood. 30

1.10 The median number of floors of extremely tall skyscrapers (blue points) has not increased significantly over the last forty years. A median regression indicates extremely tall skyscrapers grow by .06 floors each year (p-value = .11). A 95 percent confidence interval for the median regression line is shaded blue. The median number of floors of tall skyscrapers (red) is shown for comparison. 31

2.1 Comparing 18-year-olds (currently Generation Z, birth year 2002, magenta): Estimated Republican support from 18-year-olds since 1950 (top) and cumulative partisan impressions producing this support from birth to age 18 for select birth years (bottom). 44

2.2 Comparing 35-year-olds (currently Millennials, birth year 1985, blue): Estimated Republican support from 35-year-olds since 1950 (top) and cumulative partisan impressions producing this support from birth to age 35 for select birth years (bottom). 45

2.3 Comparing 52-year-olds (currently Generation X, birth year 1968, teal): Estimated Republican support from 52-year-olds since 1950 (top) and cumulative partisan impressions producing this support from birth to age 52 for select birth years (bottom). 46

2.4 Comparing 68-year-olds (currently Baby Boomers, birth year 1952, green): Estimated Republican support from 68-year-olds (Baby Boomers) since 1950 (top) and cumulative partisan impressions producing this support from birth to age 68 for select birth years (bottom). 47

2.5 Relative vote share: Estimated Republican support by age, race, ethnicity, and sex. These estimates can be post-stratified to predict the outcome of the 2020 election, which we plan on doing when data becomes available. 48

3.1 Raw data and loess curves, indicating the relationship between age and presidential voting preferences among non-Hispanic white voters for the 2000-2016 elections. From the left: (1) The relationship is non-monotonic and quite peculiar in 2012; instead of a linear or even quadratic relationship, the curve changes directions multiple times. (2) Non-monotonicity characterizes other elections as well. No clear pattern is apparent from this graph alone. (3) The true relationship emerges when the curves are lined up by birth year instead of age. The peaks and valleys occur at almost identical locations, indicating a generational trend. 53

3.2 The Gallup Organization's presidential approval rating time series, 1937-2016. The data reflects political events that influence voter's partisan preferences. 59

3.3 After removing survey respondents born before 1937, the analysis includes 215,693 survey respondents in total, here displayed by election year and year of birth. The data, and thus the analysis, have a strong emphasis towards the most recent four elections, and may be interpreted as weighted towards the contemporary political climate. The data encompass generational cohorts defined by their individual birth year from 1937-1998, with at least 1,000 responses for each birth year until 1986. . 60

3.4 Estimates of the generational effect. (L) We find the 14-24 age range is most import for the formation of long-term presidential voting preferences. Political events before 14 have little impact. After 24, the age weights decreases. (R) These weights, and the political socialization process implied by them, are substantially more important for non-Hispanic whites than for minorities as a whole. 63

3.5 (L) Estimates of the period effect. Minorities are consistently more likely to vote for Democratic presidents, and Southern whites have steadily trended pro-Republican over the past 50 years. (R) Election effects are similar between young and old minority voters and in the South. The evidence is inconclusive for non-Southern whites. 65

3.6 The model accounts for 91% of the macro-level variance in voting trends over the past half century, more than the simpler model incorporating only period/group effects. The model fits considerably better within race/region groups, particularly among non-Southern whites. 67

3.7 Presidential approval, and the cumulative generational effects, for Eisenhower Republicans born in 1941. The graph emphasizes peak years of socialization, according to age weights found by the model. Blue indicates pro-Democratic years, red for pro-Republican, grey in between. This generation missed most of the FDR years and was socialized through 10 straight pro-Republican years (Truman and Eisenhower). Their partisan voting tendencies were drawn back towards the neutral grey line by the pro-Democratic 1960s, and they reached a rough equilibrium by the end of the Nixon presidency. 70

3.8 The generation we refer to as 1960s Liberals are best epitomized by those born in 1952, whose presidential political events are emphasized here. Too young to be highly influenced by the Eisenhower years, they experienced an intense period of pro-Democratic sentiment during the 1960s. After 1968, however, roughly 25 years of near-consistent pro-Republican events neutralized their presidential voting preferences. 72

3.9 The Approval series as seen by the generation we call Reagan Conservatives, best epitomized by those born in 1968. This generation missed the Kennedy and Johnson years entirely, and their peak socialization fell under the popular Republican presidents Reagan and Bush I. By the time the Democratic president Clinton reached his peak popularity in the late 1990s, they were already roughly 30 years old. 75

3.10 The Approval series as seen by the last generation, the Millennials. Their experience had only lasted 31 years by the 2016 election, but the model indicates that these years should remain highly influential over the rest of their lives. Their formative years have been primarily characterized by the popular Democratic president Clinton and the unpopular Republican Bush II, resulting in their relatively strong pro-Democratic sentiment. 77

3.11 The cumulative preferences of each generation is shown, along with the weighted summation of the full white electorate. The generations are loosely defined so that the entire electorate can be plotted at once. The width of each curve indicates the proportion of the white electorate that each generation reflects at any given time. The model—in this graph reflecting only the approval time series and the age weights—explains much of the voting tendencies of the white electorate over time. 79

4.1 Shifts in temperature sensitivities with warming occur when using linear models for non-linear processes. Estimated sensitivities decline with warming in simulations (left) with no underlying change in the biological process when sensitivities were estimated with linear regression (estimated across 45 sites with a base temperature of normal(6,4)). This decline disappears when performing the regression on logged predictor and response variables (right). Such issues may underlie declining sensitivities calculated from observational data, including long-term observations of leafout across Europe ('observations,' using data for Betula pendula from PEP725 from for the 45 sites that had complete data for 1950-1960 and 2000-2010), which show a lower sensitivity with warming when calculated on raw data, but no change in sensitivity using logged data. Symbols and lines represent means ± standard deviations of regressions across sites. See SI for a discussion of why estimated sensitivities are -1 or lower in non-linear models. 94

4.2 Simulated leafout as a function of temperature across different temperatures highlights non-linearity of process. Here we simulated sets of data where leafout constantly occurs at 200 growing degree days (thermal sum of mean daily temperatures with 0°C as base temperature) across mean temperatures of 10, 15, 20 and 30°C (constant SD of 4), we calculated estimated mean temperature until leafout date (top row) or across a fixed window (bottom row, similar to estimates of 'spring temperature'). While within any small temperature range the relationship may appear linear, its non-linear relationship becomes clear across the greater range shown here (left). Taking the log of both leafout and temperature (right) linearizes the relationship. 98

4.3 A simple model generates declining sensitivities with warming. We show declines in estimated sensitivities with warming from simulations (top: using average temperature until leafout, bottom: using a fixed window) with no underlying change in the biological process when sensitivities were estimated with simple linear regression ("Simple linear regression"). This decline disappears using regression on logged predictor and response variables ("Using logged variables"). 99

4.4 Simulated leafout as a function of temperature across different levels of warming with shifts in underlying biology. Here we simulated sets of data where leafout occurs at a thermal sum of 200 (sum of mean daily temperatures with 0°C as base temperature) when chilling is met, and requires a higher thermal sum when chilling is not met. We show estimated sensitivities in the top panel, and the shifting cues in the bottom panel. 100

4.5 Sensitivities from PEP725 data using 10 year windows of data 101

4.6 Sensitivities from PEP725 data using 20 year windows of data 102

5.1 Two equivalent representations of a block migration pattern. Migrants move from one location in the red cluster to one location in the green cluster at rate B_{21} and from one location in the green cluster to one location in the blue cluster at rate B_{32}. 110

5.2 Two equivalent representations of a block migration pattern. Migrants move from two locations in the red cluster to one location in the green cluster at rate B_{21} and from one location in the green cluster to two locations in the blue cluster at rate B_{32}. 110

5.3 A comparison of the Infinitesimal Block Model (IBM) and a (hard) mixture of Poisson regressions (M-GLM) to a Poisson Stochastic Block Model (SBM). IBM and M-GLM are fit to longitudinal data, while SBM is fit to the underlying migration data. The first three panels show the cluster assignments from each algorithm. Assuming SBM found the "correct" assignments, the bottom-right panel shows the confusion matrix: IBM correctly identified 86 percent of counties; M-GLM 77 percent. 116

5.4 A comparison of the Infinitesimal Block Model (IBM) and a (hard) mixture of Poisson regressions (M-GLM) to the Lee-Carter model (SVD). Normalized mean squared error ("pseudo" R^2) is reported for in-sample predictions (top) and out-of-sample predictions (bottom). IBM and M-GLM are fit with both $K = 3$ and $K = 20$ clusters. IBM predictions are comparable with SVD. 118

5.5 The annual number of major felony offenses (crimes, top) and stops under the Stop, Question, and Frisk policy (stops, bottom) in New York City between 2002 and 2012. Crimes and stops appear negatively correlated, suggesting to policymakers that SQF prevents crime. The relationship at the precinct level is similar—albeit weaker. 121

5.6 The MAP cluster of New York City precincts identified by the extended Infinitesimal Block Model (IBM) with K = 3. Between 2002 and 2012 crime "migrated" from red precincts to blue precincts to green precincts. 122

5.7 Results from an IBM simulation study. The trajectory of 100 locations is colored by cluster assignment (top). Individuals move from green locations to red at instantaneous rate $b_{21} = 3e{-}04$ and from red locations to blue locations at rate $b_{32} = 3e{-}03$. Trace plots show convergence for b_{21} (left) and b_{32} (right) after around one thousand iterations. The red horizontal line marks the true data-generating value. 131

Chapter 1: Forecasting the Urban Skyline with Extreme Value Theory

with Phyllis Wan

The world's urban population is expected to grow fifty percent by the year 2050 and exceed six billion. The major challenges confronting cities, such as sustainability, safety, and equality, will depend on the infrastructure developed to accommodate the increase. Urban planners have long debated the consequences of vertical expansion—the concentration of residents by constructing tall buildings—over horizontal expansion—the dispersal of residents by extending urban boundaries. Yet relatively little work has predicted the vertical expansion of cities and quantified the likelihood and therefore urgency of these consequences.

We regard tall buildings as random exceedances over a threshold and use extreme value theory to forecast the skyscrapers that will dominate the urban skyline in 2050 if present trends continue. We predict forty-one thousand skyscrapers will surpass 150 meters and 40 floors, an increase of eight percent a year, far outpacing the expected urban population growth of two percent a year. The typical tall skyscraper will not be noticeably taller, and the tallest will likely exceed one thousand meters but not one mile. If a mile-high skyscraper is constructed, it will hold fewer occupants than many of the mile-highs currently designed. We predict roughly three-quarters the number of floors of the Mile-High Tower, two-thirds of Next Tokyo's Sky Mile Tower, and half the floors of Frank Lloyd Wright's The Illinois—three prominent plans for a mile-high skyscraper. However, the relationship between floor and height will vary considerably across cities.

1.1 Introduction

The world is urbanizing at an astonishing rate. Four billion people live in urban areas, up from two billion in 1985. By 2050, the United Nations predicts the urban population will ex-

ceed six billion. The increase is due to growth in both the world population and the proportion of the population that resides in urban areas. Roughly half the world's population is urban, up from forty-percent in 1985 and projected to rise above two-thirds in 2050 (UN, 2018). The future preponderance of cities suggests the major challenges confronting civilization will be urban challenges. Moreover, the particular nature of these challenges will depend on how cities choose to accommodate urbanization (Rose, 2016, p.15).

Cities change in response to population growth by either increasing density—the population per land area—or extending boundaries—the horizontal distance between city limits. The prevailing paradigm among urban planners is to preserve city boundaries and encourage density (Angel et al., 2011). It argues that density affords certain economies of scale, such as reducing the cost of infrastructure and social services like roads, water, safety, and health care. Density is also advocated to promote sustainability by preserving the city periphery for agriculture or wildlife (Swilling, 2016). Yet density, if not properly accommodated, can lead to overcrowding and impede quality of life. Nearly one third of urban residents in developing regions live in overcrowded slums that concentrate poverty (UN, 2015, p.2).

Some urban planners have argued that density requires vertical expansion, through the construction of skyscrapers, to prevent overcrowding and maintain quality of life (Gottmann (1966), Al-Kodmany (2012), Barr (2017)). This three-dimensional solution to a two-dimensional problem was stated as early as 1925 by the architect Le Corbusier: "We must decongest the centers of our cities by increasing their density" (Kashef, 2008). In this spirit, Glaeser (2011) recommends policies that ease height restrictions and increase financial incentives for skyscraper development.

Other urban planners warn that urbanization is too rapid to be adequately addressed by vertical expansion (James (2001, p.484), Cohen (2006, p.73), Canepari (2014)). Angel et al. (2011) argue cities must "make room" for urbanization by moving boundaries and recommend policies that extend the radius of public services like the transit system.

It stands to reason that cities will utilize multiple strategies to accommodate urbanization. For example, cities will incentivize some vertical expansion and extend the radius of some public

services. The challenges facing cities in 2050 will depend on which policies are implemented. Anticipating these challenges requires an answer to questions such as: if present trends continue, how much vertical growth will the typical city experience by 2050? How much farther will the typical city boundary extend?

This paper demonstrates that extreme value theory provides a principled basis for forecasting vertical growth. It regards tall buildings as random exceedances over a threshold and uses the probabilistic laws governing extreme values to extrapolate the characteristics of the skyscrapers that will dominate the urban skyline in 2050. Similar arguments have produced successful forecasts in a wide variety of fields, most notably those concerned with risk management, such as finance (Gencay and Selcuk (2004), Bao et al. (2006), Chan and Gray (2006), Herrera and González (2014)) and climate (Garreaud, RD. (2004), Ghil (2011), D'Amico et al. (2015), Thompson et al. (2017)). However, we know of no work that applies these arguments to forecast how cities will respond to rapid urbanization.

The findings are arranged into five sections. Section 2 motivates the use of statistical models to characterize the uncertainty of skyscraper development. The data are described, and a brief review of the skyscraper literature follows. Section 3 outlines the Poisson regression used to conclude that forty-one thousand skyscrapers will be completed by 2050. Section 4 outlines the generalized Pareto distribution (GPD) used to conclude that there is a seventy-three percent chance a skyscraper will exceed one thousand meters by 2050 and an eleven percent chance it will exceed a mile. Section 5 outlines the censored asymmetric bivariate logistic distribution used to conclude that a mile-high skyscraper, if built, will have around 250 floors. The paper concludes with two sections, highlighting the methodological and policy consequences of these predictions, respectively.

1.2 Modeling the Uncertainties of Skyscraper Development

This paper predicts the prevalence and nature of skyscrapers in the year 2050 if present trends continue. This section reviews the definition of a skyscraper, motivates the statistical modeling of tall buildings, and outlines our forecasting strategy.

Strictly speaking, the term skyscraper refers not to height but to the mode of construction. A skyscraper is defined as any multi-story building supported by a steel or concrete frame instead of traditional load-bearing walls (Curl and Wilson, 2015, p.710). The tall buildings capable of sustaining the dense cities of the future will almost certainly be skyscrapers. Over the past century, building beyond a few floors has required a supporting frame.

In contrast to the precise definition of a skyscraper, the definition of a tall building depends on context. From a public safety perspective, high-rises—multi-story buildings as short as 23 meters (75 feet)—are harder to evacuate than low-rises. From an environmental perspective, however, only a much taller structure (152 meters, 500 feet) would obstruct the migration pattern of birds (Brown and Caputo, 2007, p. 17). The Council on Tall Buildings and Urban Habitat (CTBUH) sets international standards for the purpose of research and arbitrating titles, such as the world's tallest building. They define buildings exceeding fifty meters as tall, three-hundred meters as super tall, and six-hundred meters as mega tall. Height is measured from "the level of the lowest, significant, open-air, pedestrian entrance to the architectural top of the building, including spires, but not including antennae, signage, flag poles or other functional-technical equipment" (Council on Tall Buildings and Urban Habitat, 2017).

This analysis retains the CTBUH measure of height but bases its definition of tall on statistical considerations. It relies on data from The Skyscraper Center, a CTBUH directory of every tall building worldwide. The directory is "the premier source for accurate, reliable information on tall buildings around the world" (Council on Tall Buildings and Urban Habitat, 2017). Nevertheless, it depends on CTBUH members and the public to add entries, resulting in the partial or complete omission of some smaller buildings. The data appear complete for buildings exceeding 150 meters and 40 floors, roughly the height of the Great Pyramid of Giza, the United Nations Headquarters, or the Seagram Building in New York City. We refer to these skyscrapers as tall. The Skyscraper Center catalogs 3,251 tall skyscrapers in 258 cities as of December 2017.

The height and year each tall skyscraper was completed is displayed in the top-left panel of Figure 1. It appears from the panel as though a simple statistical relationship might govern the

number of skyscrapers exceeding various heights, ignoring the hiatus between the Great Depression and the Second World War, which is assumed to be anomalous. If this is the case, forecasting the prevalence and nature of skyscrapers in the year 2050 is simply a matter of extrapolating that relationship. But it cannot be taken for granted that skyscrapers can be modeled statistically. These buildings are modern marvels, requiring enormous cooperation across teams of architects, engineers, financiers, and multiple levels of government. Extrapolation from the data is only meaningful if the determinants underlying skyscraper construction are varied and independent enough to be characterized statistically. The following is an argument supporting this view.

Skyscrapers appeared in the nineteenth century after technological innovations, such as the elevator brake and the mass production of steel, made building beyond a few floors economical. By the turn of the twentieth century, a handful of twenty floor buildings had been completed in major cities across the United States. Advances in the mid-twentieth century permitted mile-high buildings (1609.34 meters), and various architects have since proposed designs, such as the Houston Pinnacle (1 mile, 500 floors), the Ultima Tower (2 miles, 500 floors), and the Sky Mile Tower (1 mile, 400 floors). Perhaps most famously, the architect Frank Lolyd Wright proposed The Illinois in 1957, a mile-high building with 528 floors (Council on Tall Buildings and Urban Habitat, 2017).

But none of these designs have been realized. Despite the technical ability to build tall, mile-high skyscrapers are considered impractical because they are unlikely to turn a profit. In fact, ambitious skyscrapers frequently fail for financial reasons (Lepik, 2004, p.21-2). For example, the Jeddah Tower was originally planned for one mile (330 floors). After the Great Recession, the height was reduced by more than a third to one thousand meters (167 floors).

Skyscrapers are a commercial response to an economic phenomenon. They arise when land values produce rents that exceed the enormous cost of construction and maintenance. Aesthetics are a secondary concern (Clark and Kingston, 1930), (Willis, 1995), (Ascher and Vroman, 2011, p.12,22). Indeed, it was in the aftermath of the 1871 Chicago fire that the urgent need for new office space produced the First Chicago School of skyscrapers (Lepik, 2004, p.6). Construction

has since followed the rise of oil-rich Middle Eastern countries in the 1980s, the former Soviet-bloc countries in the 1990s, and the Pacific Rim countries in the 2000s (Sennott, 2004, p.1217). Barr and Luo (2017) find that half of the variation in China's skyscraper construction can be explained by population and gross city product alone. As the architect Cass Gilbert famously summarized: "A skyscraper is a machine that makes the land pay."

The enormous cost of a skyscraper is not only the result of additional construction materials. Taller buildings require concrete to be pumped at higher pressures (Ascher and Vroman, 2011, p.83). Nor is the cost entirely in raising the building itself. Taller buildings require more elevators, which reduces the floor area available for occupancy and thus the revenue potential of the building (Ascher and Vroman, 2011, p.33). Human comfort is also a factor. Excessive elevator speed (Ascher and Vroman, 2011, p.103) and building sway (Ascher and Vroman, 2011, p.1) can produce motion sickness even if safe. Additional considerations include government policy (permits and zoning, financial incentives, and public infrastructure), cultural values (equity, sustainability, and security), and environment (foundation quality, prevailing winds, and natural disaster frequency).

In short, a litany of factors must align favorably to produce a skyscraper. One could hypothetically observe every factor for every property and predict the future of skyscraper development. Yet it is convenient, and perhaps more accurate, to regard unobserved heterogeneity among these factors as a source of randomness that obeys the laws of probability. Statistical models quantify the probability a property possesses the factors that will produce a particular skyscraper, where the parameters of the models are determined from data.

In the following sections, we demonstrate that extreme value theory provides a principled strategy for choosing a statistical model. We adopt the Peaks Over Threshold approach: First, we select all observations exceeding a threshold. Then, we model the number of exceedances with Poisson distributions and the sizes of the exceedances with generalized Pareto distributions. The challenge with this approach is choosing the threshold. If the threshold is set too low, the assumptions underlying extreme value theory are not reasonably satisfied, and the model will not be flexible enough to approximate the data. If the threshold is set too high, there will not be enough data to estimate

the model parameters reliably.

Our forecasts use a different threshold for the Poisson and generalized Pareto distributions. We choose the lowest thresholds such that each model appears accurate, and we extrapolate the number or sizes of skyscrapers exceeding the respective threshold until the year 2050, thirty-three years after the data was collected in 2017. We repeat each forecast on a subset of the data to validate the results: We use data available only before 1984 to predict skyscraper development until 2017, thirty-three years later, and compare the predictions with the actual data. In Section 3, we find the threshold of 150 meters and 40 floors sufficiently high to forecast the number of buildings exceeding that threshold in 2050. However, it is too low to forecast the heights of those buildings. In Section 4, we find a threshold of 225 meters or 59 floors sufficient, roughly the size of One Penn Plaza in New York City. We refer to these buildings as extremely tall since the assumptions underlying extreme value theory appear to hold at this height. CTBUH catalogs 325 extremely tall skyscrapers in 81 cities as of December 2017, one tenth the number of tall skyscrapers.

Our approach assumes skyscraper development is independent given the parameters of these distributions. The assumption is certainly violated for contemporaneously completed skyscrapers within the same city—the aforementioned factors producing such skyscrapers are linked inextricably. But a series of recent investigations suggest that these factors are largely independent across cities and time periods. For example, Barr (2012) finds competition within cities is limited to periods close in time and space, Barr, Mizrach, and Mundra (2015) find skyscraper height is not a useful indicator of economic bubbles or turning points, and Barr and Luo (2017) find little evidence that cities in China compete for the tallest building.

The literature indicates that considerable economic pressure produces the factors that drive skyscraper development, and the catalyst for this pressure varies idiosyncratically by city and time period. It is perhaps because of this idiosyncratic variation—and the fact that no city possesses more than ten percent of all skyscrapers—that the prevalence and nature of skyscrapers so closely follows the distributions derived under extreme value theory, as demonstrated in the following sections, despite periods of economic and political turmoil within nearly every city since the Second

World War.

1.3 Predicting the Quantity of Skyscrapers Completed by 2050

Skyscraper construction has increased at a remarkably steady rate. The number of tall skyscrapers—skyscrapers exceeding 150 meters and 40 floors—has risen at the rate of eight percent a year since 1950. If this trend continues, the eight percent annual growth rate of skyscrapers will far outpace the two percent annual growth rate of urban populations anticipated by the UN (2018). Forty-one thousand tall skyscrapers will be completed by 2050, 6,800 per billion city residents compared to the roughly 800 per billion city residents today.

The eight percent rate was determined by fitting the following Poisson regression model. Let N_t be the number of tall skyscrapers completed in year t, $t = 1950, \ldots, 2017$. The N_t's are assumed to be independent and follow a Poisson distribution with mean

$$E[N_t] = \exp(\alpha + \beta t).$$

The Poisson distribution can be justified theoretically by regarding N_t as the sum of independent Bernoulli trials: Suppose $D(t)$ is the set containing every building in the world completed in year t. Define $D_u(t)$ to be the subset of all buildings in $D(t)$ exceeding the height threshold u. The Poisson Limit Theorem states that for sufficiently large u, the probability a building in $D(t)$ is also in $D_u(t)$ is small, and the number of buildings in $D_u(t)$, N_t, is well approximated by a Poisson distribution. See Coles (2001, p.124) for a more detailed discussion of the Poisson limit for threshold exceedances.

We use the `glm` function in the R Core Team (2018) package `stats` to obtain maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$ and to calculate their standard errors. The plug-in estimate, $\hat{E}[N_t] = \exp(\hat{\alpha} + \hat{\beta} t)$, is plotted against the data in the top-right panel of Figure 1.1. An inner 95 percent predictive interval for the years 2020 to 2050 is added in the bottom-left panel. A cumulative thirty-eight thousand skyscrapers is estimated for completion between 2018 and 2050 if present

trends continue (forty-one thousand total, with a standard error of seven thousand), about twelve times the current number.

To demonstrate the accuracy of the model for predicting 2050, thirty-three years after the data was collected in 2017, we predict the last thirty-three years using only data that would have been available before 1984. The bottom-right panel shows that if such predictions had been made in 1984 (dotted blue line), they would align closely with the actual number of skyscrapers built each year. The log-linear model anticipates 3,082 skyscrapers by the end of 2017 when in fact 2,988 were built between 1950 and 2017, a difference of three percent.

Numerous populations grow at a log-linear rate, $E[N_t] = \exp(\alpha + \beta t)$. The relationship arises when, at any instant, a population increases $100 \times \beta$ percent of its current size, $\frac{dN}{dt} = \beta N$—notwithstanding random fluctuation. For more complicated phenomena, such as skyscraper construction, the relationship provides an accurate yet parsimonious approximation. However, there is no guarantee the approximation will remain accurate in future years, and one might reasonably question the sensitivity of the predictions to the log-linear assumption. For example, the years 2015, 2016, and 2017 each saw fewer tall skyscrapers built than the year 2014. Perhaps it would be more accurate to conclude that skyscraper growth is slowing and will continue to slow until 2050.

The logistic-linear relationship is a more flexible alternative that can capture a slowing growth rate. Suppose for the moment that the N_t's are independent and follow a Poisson distribution, except now with mean

$$E[N_t] = \frac{\gamma}{1 + \exp(-\alpha' - \beta t)}.$$

This relationship arises when, at any instant, a population increases by $100 \times \beta(1 - \frac{N_t}{\gamma})$ percent of its current size, $\frac{dN}{dt} = \beta(1 - \frac{N}{\gamma})N$—notwithstanding random fluctuation. It is reasonable because, for $\frac{N_t}{\gamma} \approx 0$, a logistic-linear relationship is well approximated by a log-linear relationship, which was shown to fit the data well in Figure 1.1. Unlike a log-linear relationship, however, as the population increases, the growth rate of a logistic-linear relationship slows, and the population reaches its maximum capacity, γ. Indeed, the factor $(1 - \frac{N_t}{\gamma})$ encodes the remaining percent of

the capacity at time t. The number of tall skyscrapers might reach a maximum capacity if only a fixed number of skyscrapers is needed to accommodate demand or some other constraint prevents additional construction.

We use the `optim` function in the R Core Team (2018) package `stats` to obtain maximum likelihood estimates $\hat{\alpha}'$, $\hat{\beta}$, and $\hat{\gamma}$. However $\hat{\alpha}'$ and $\hat{\gamma}$ diverge, indicating that the number of tall skyscrapers increases with no discernible upper limit. The predictions from the limiting model are identical to the log-linear model.

Even if the data were known to follow a log-linear relationship, residual dependence could still produce predictions that underestimate uncertainty. We compare our results to a Box-Jenkins analysis (Box et al. 2015) of the log number of skyscrapers completed each year since 1950. We use the `auto.arima` function in the R Core Team (2018) package `forecast` to select the best ARIMA model according to the Akaike information criterion (Hyndman et al. (2019), Hyndman and Khandakar (2008)). Prediction paths are simulated from the selected ARIMA(0, 1, 2) with drift by nonparametric residual bootstrap using the `forecast.Arima` function. The forecast is essentially identical to the Poisson regression—only three percent higher. However, the standard error is 22,000—three times larger. We find this standard error conservative. For example, it suggests the number of skyscrapers might grow fourteen percent per year over the next 33 years, nearly double the 2007-2017 and 1950-2017 rates. At the other extreme, it suggests that no additional skyscrapers might built at all. Such scenarios would reflect a considerable departure from present trends. Nevertheless, we include these results as a reference for the reader.

1.4 Predicting the Height of Skyscrapers Completed by 2050

The tallest skyscraper has doubled in height since 1950. Yet the height increase of the typical extremely tall skyscraper—one exceeding 225 meters or 59 floors—is not statistically significant. In fact, the same distribution describes the heights of extremely tall skyscrapers since 1950. We conclude the tallest skyscraper is not increasing because skyscrapers are becoming taller. It is increasing because more buildings are being constructed and thus more buildings are eligible to

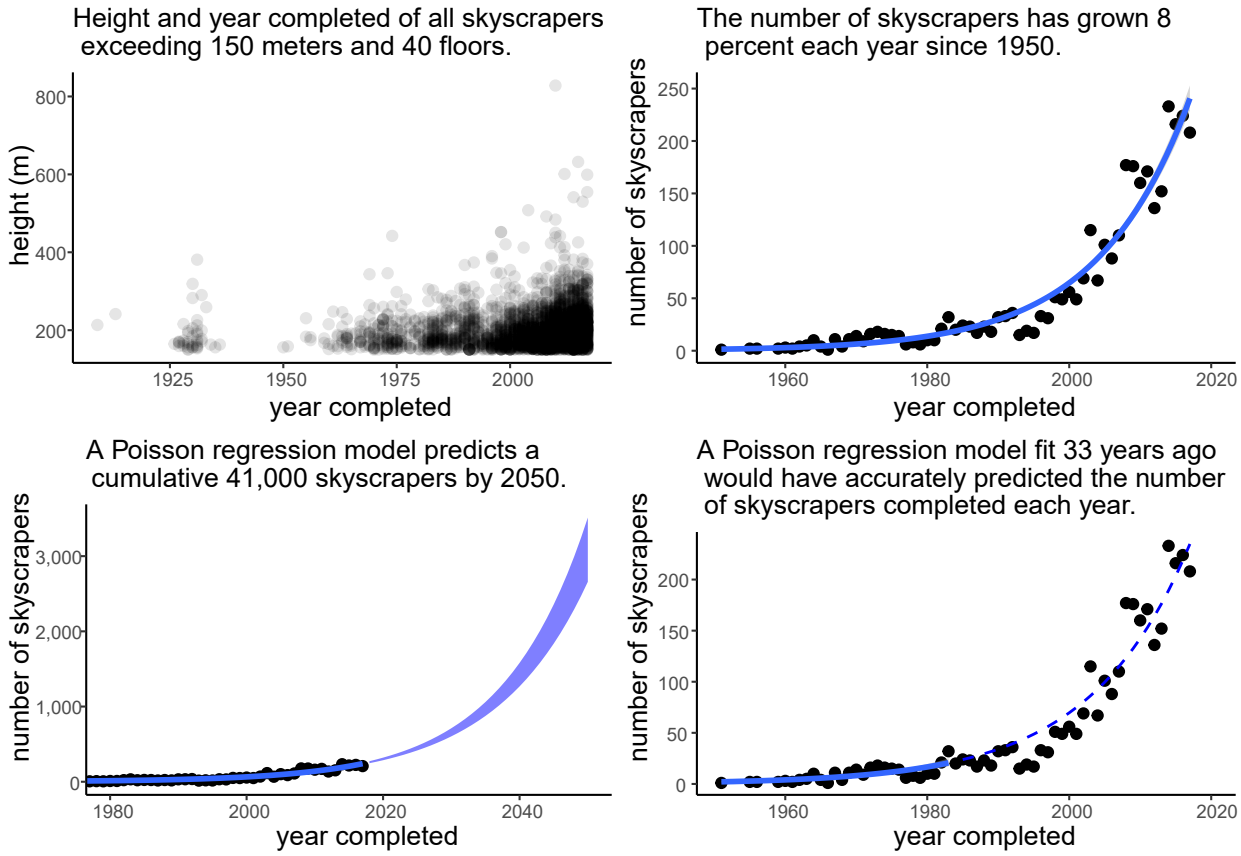


Figure 1.1: We predict the quantity of skyscrapers completed by 2050, thirty-three years after the data was collected at the end of 2017. A Poisson regression model suggests the number of skyscrapers will grow at a rate of eight percent a year.

be the tallest. Assuming the same distribution continues to describe the heights of skyscrapers completed by 2050, the probability a new building will exceed the current tallest—the Burj Khalifa (828 meters)—is estimated to be nearly 100 percent. The probability a new building will exceed the Jeddah Tower (1,000 meters)—scheduled to open in 2020—is 73 percent. The probability a new building will exceed one mile is 11 percent.

These probabilities were determined by approximating the heights of extremely tall skyscrapers with a generalized Pareto distribution (GPD). The GPD approximation can be justified theoretically by regarding tall buildings as random exceedances over a threshold: Suppose X is a random variable with an unknown distribution function F. Define F_u to be the exceedance conditional distribution—the distribution of $X - u$ given that X exceeds a threshold u, i.e.,

$$F_u(y) = \Pr(X - u \leq y | X > u).$$

The Pickands-Balkema-de Haan Theorem (Balkema and De Haan (1974), Pickands III (1975)) states that for a broad class of distributions, F,

$$F_u(y) \approx H(y) = 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)^{-1/\xi},$$

as u tends to $F^{\leftarrow}(1) = \sup\{y : F(y) < 1\}$, the right end point of F. When $\xi = 0$, $H(y)$ is defined to be its limit, $1 - \exp(-y/\sigma_u)$. The support of $H(y)$ is $y > 0$ when $\xi \geq 0$ and $0 \leq y \leq -\frac{\sigma_u}{\xi}$ when $\xi < 0$. Thus, for a large enough threshold u, the distribution $F_u(y)$ is governed by two parameters: a scale, σ_u, which depends on u, and a shape, ξ, which does not depend on u. The shape parameter is called the tail-index because it determines how fast the tail of the distribution, F, decays.

Now suppose D is the set containing every building in the world, and D_u is the subset of all buildings exceeding the height threshold u. Let X be the height of a building randomly selected from D_u. By setting $y = x - u$, the Pickands-Balkema-de Haan Theorem justifies, for sufficiently

large u, approximating the distribution of X by a GPD with parameters (u, σ_u, ξ),

$$\Pr(X \leq x | X > u) \approx H(x - u) = 1 - \left(1 + \frac{\xi(x-u)}{\sigma_u}\right)^{-1/\xi}.$$

The threshold, u, is a location parameter. See Coles (2001, p. 74) for a more detailed discussion of the GPD limit for threshold exceedances.

The threshold u must be set high enough for the GPD approximation to hold. Unlike Section 3, the default threshold of $u = 150$ meters may be insufficient. Were u sufficiently high such that $X - u | X > u$ followed a GPD with parameters (σ_u, ξ), $X - u' | X > u'$ would follow a GPD with $(\sigma_{u'}, \xi)$ for any $u' > u$. This suggests a strategy for choosing u: Produce a sequence of candidate thresholds $u_1 < \ldots < u_I$. At each threshold u_i, obtain estimates $(\hat{\sigma}_{u_i}, \hat{\xi}_{u_i})$ and their standard errors by maximizing the likelihood with dataset D_{u_i}. Select a threshold u for which the estimates $\hat{\xi}_{u'}$ appear consistent for $u' > u$ (Smith, 1985; Davison and Smith, 1990).

The left side of Figure 1.2 shows the maximum likelihood estimates for the shape parameter, $\hat{\xi}_{u_i}$, when fit to skyscraper heights exceeding a sequence of thresholds, u_i, from 150 to 350 meters. Point estimates are colored red, and thick (thin) lines represent 50 (95) percent confidence intervals. The point estimates increase with little uncertainty as the threshold increases from 0 at 150 meters to .18 at 225 meters, after which a ξ of .18 is consistent with the data. Computational details are discussed in the Appendix.

The choice of the 225 meter threshold and the .18 shape parameters is compared to a mean excess plot and a Hill plot on the right side of Figure 1.2. The mean excess plot (top of Figure 1.2) displays the empirical mean excess function, $\hat{E}[X - u | X > u]$, for different thresholds u (Bernktander and Segerdahl, 1960; Ghosh and Resnick, 2010). The Hill plot (bottom of Figure 1.2) displays the Hill estimator of ξ using the k largest observations for different choices of k (Hill, 1975). Dark blue regions represent inner 50 percent confidence intervals, and light blue regions represent inner 95 percent confidence intervals. When the data are consistent with a GPD, the mean excess plot is expected to be linear in u, and the Hill estimator is expected to stabilize

near ξ. We find the two plots are consistent with the 225 meter threshold and .18 shape parameters.

The GPD—and the choice of threshold and shape parameters—is validated by comparing simulations of skyscraper heights to the observed data. In fact, at the threshold of 225 meters, simulations from the same GPD describe the heights of skyscrapers completed at different time periods. We demonstrate this by first obtaining the maximum likelihood estimates for the GPD parameters, $\hat{\mu}_{225}$, $\hat{\sigma}_{225}$, and $\hat{\xi}_{225}$, where μ_{225} is the location of the GPD, considered unknown for the moment. We then partition the skyscrapers built after 1950 into sextiles according to the year in which they were completed. Since skyscraper growth is log-linear, the window lengths become shorter so that the same number of skyscrapers are in each bin. Figure 1.3 shows a p-p plot for each time period. That is, instead of a q-q plot, which plots the ordered skyscraper heights, $x_{(i)}$, against the ordered heights simulated from the GPD with parameters $\hat{\mu}_{225}$, $\hat{\sigma}_{225}$, and $\hat{\xi}_{225}$, we first apply the GPD probability integral transformation: $1 - [1 + \hat{\xi}_{225}(x_{(i)} - \hat{\mu}_{225})/\hat{\sigma}_{225})]^{(-1/\hat{\xi}_{225})}$. After transformation, the theoretical heights follow the standard uniform distribution.

The close fit in each time period suggests the height distribution of extremely tall skyscrapers does not change. This is investigated further in Figure 1.4, which shows the median height of extremely tall skyscrapers each year since 1950 (blue points). A median regression line is estimated using the R Core Team (2018) package `quantreg` (Koenker, 2018), and the blue region represents an inner 95 percent confidence interval. The height of the typical extremely tall skyscraper increases less than half a meter each year, a negligible 3.6 percent over sixty-eight years. The increase is not statistically significant (p-value = .25). For comparison, the same analysis is conducted on tall skyscrapers—skyscrapers exceeding 150 meters and 40 floors—and shows a parallel trend (red). We conclude that the urban skyline is driven primarily by the exponential increase in the number of buildings completed each year. Years with more construction are more likely to yield extremely tall skyscrapers, and the increase of the typical skyscraper is inconsequential by comparison.

For the purpose of predicting the height of the tallest skyscraper in 2050, we assume this trend will continue. We use the `gpdSim` function in the R Core Team (2018) package `fExtremes`

(Wuertz, Setz, and Chalabi, 2017) to draw from a GPD with parameters $(\hat{\mu}_{225}, \hat{\sigma}_{225}, \hat{\xi}_{225})$ roughly 8,400 times, where 8,400 is the number of new skyscrapers estimated to be completed by 2050 in the previous section multiplied by the empirical probability a tall skyscraper exceeds 225 meters. The estimated maximum height is retained. The distribution of this prediction is then approximated by parametric bootstrap. Figure 1.5 shows the result (blue). A right-sided 95 percent predictive interval ends at 1,900 meters.

To demonstrate the accuracy of this approach for predicting 2050, thirty-three years after the data was collected in 2017, we conduct a second simulation of 2017 using only data that would have been available before 1984. The GPD is fit using all skyscrapers above the 225 meter threshold, and we simulate the height of roughly five hundred skyscrapers, since, in 1984, three thousand tall skyscrapers would have been estimated for completion between 1984 and 2017, fifteen percent of which would have been expected to exceed 225 meters. The predictive distribution of the 2017 maximum height is approximated by parametric bootstrap and shown in Figure 1.5 (red). We find that the current tallest skyscraper, at 828 meters, would have been considered likely. The predicted height is 918 meters, eleven percent above this value.

We estimate an eleven percent chance the tallest skyscraper will exceed one mile in the year 2050. This assumes present trends continue: that the number of tall skyscrapers will continue to increase at its historic rate of eight percent a year. Deviations from this rate are possible, and they would influence the estimate. Using the 95 percent predictive interval from the Box-Jenkins analysis in the previous section, we find the probability of a mile high could be as low as 4 percent or as high as 30 percent. We reiterate that these probabilities would require a considerable departure from present trends. Nevertheless, we include them as a reference for the reader.

1.5 Predicting the Number of Floors in Skyscrapers Completed by 2050

The marginal number of floors in the typical skyscraper has decreased as height increases. Height alone overstates the ability of skyscrapers to accommodate a growing population. Assuming the marginal number of floors continues to decrease as height increases, the one thousand meter

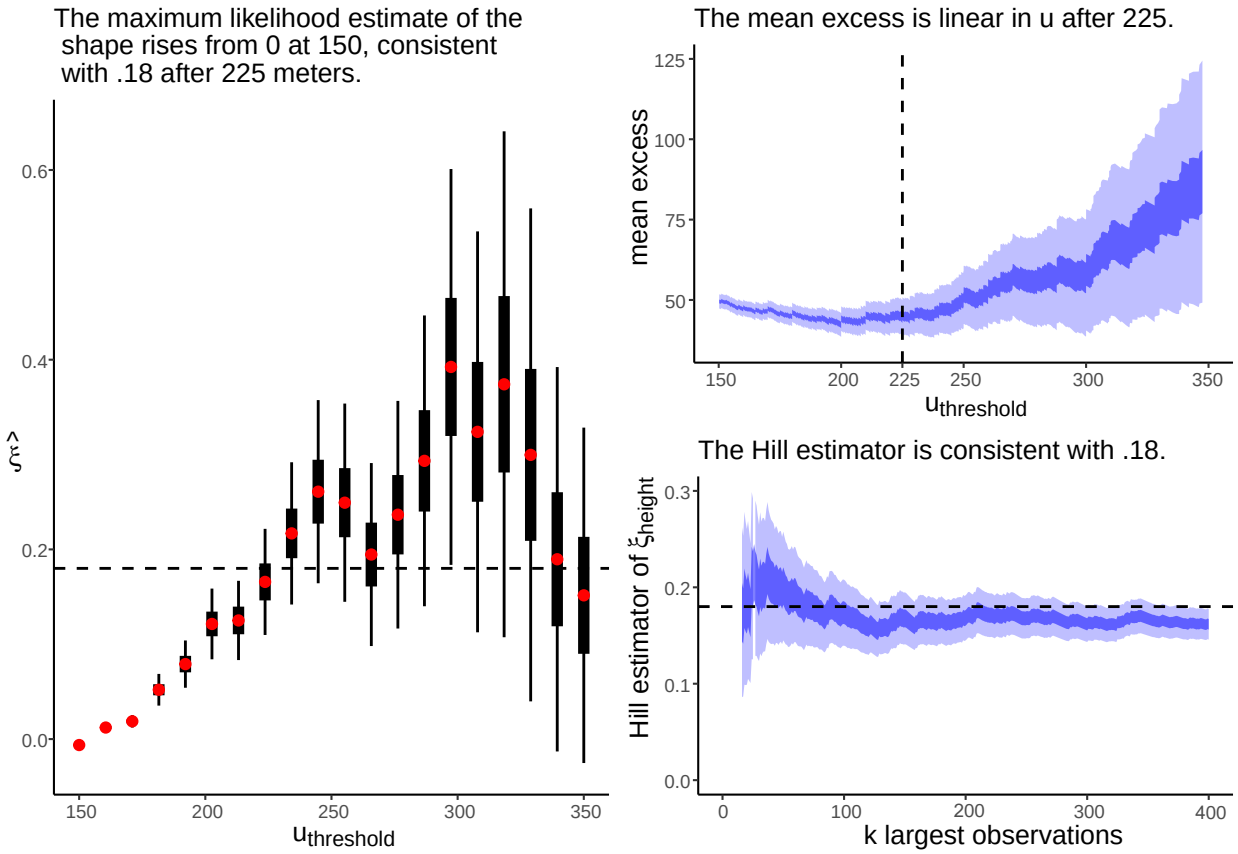


Figure 1.2: We approximate the height distribution of tall skyscrapers using a GPD. The left panel shows maximum likelihood estimates and confidence intervals (50 and 95 percent) of the GPD shape parameter, excluding observations below a sequence of thresholds $\{u_i\}$. The right panel shows a Mean Excess plot (top) and a Hill plot (bottom)

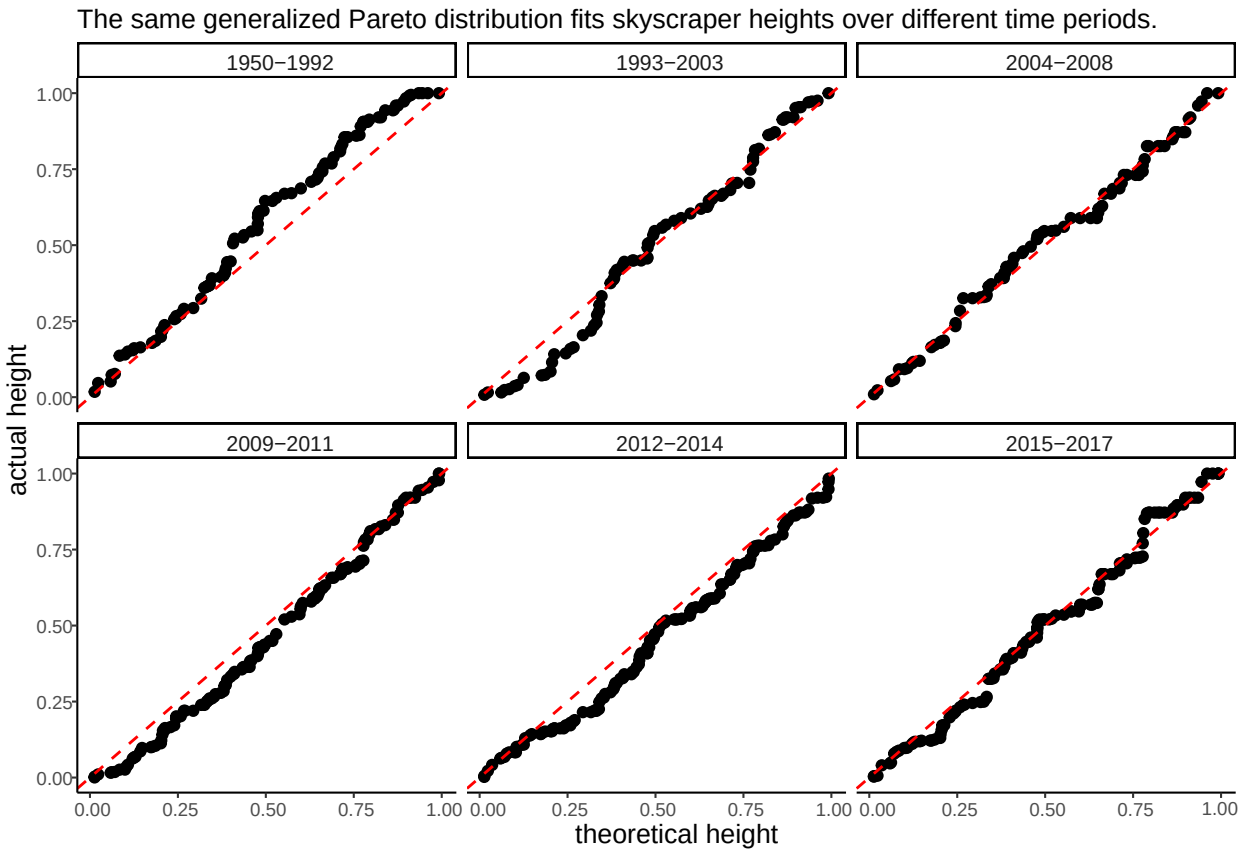


Figure 1.3: We demonstrate the distribution of skyscraper heights has changed little over time by dividing extremely tall skyscrapers into sextiles based on the year completed and constructing a p-p plot for each sextile. The distribution of theoretical heights is the generalized Pareto distribution with parameters estimated by maximum likelihood.

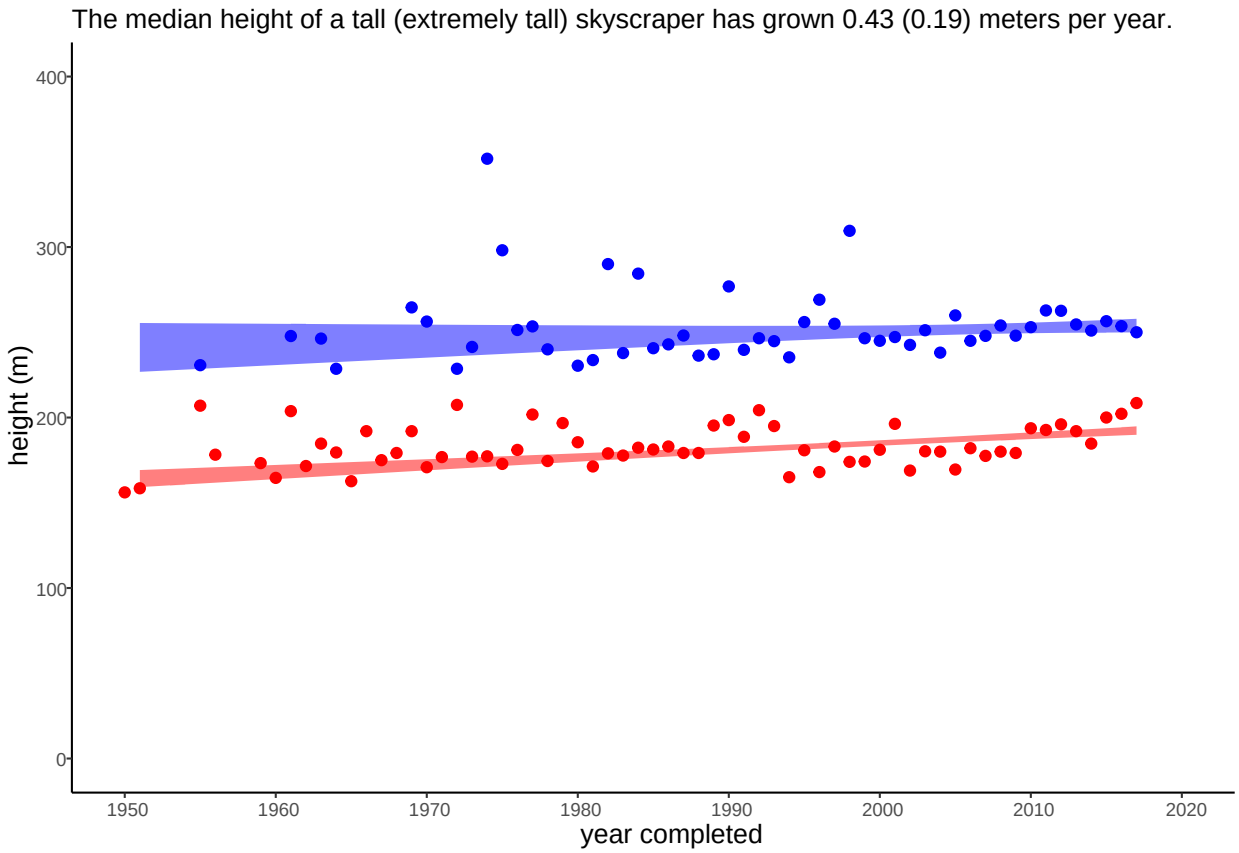


Figure 1.4: The median height of extremely tall skyscrapers (blue points) has not increased significantly over the last forty years. A median regression indicates extremely tall skyscrapers grow .19 meters each year (p-value = .25). A 95 percent confidence interval for the median regression line is shaded blue. The median height of tall skyscrapers (red) is shown for comparison.

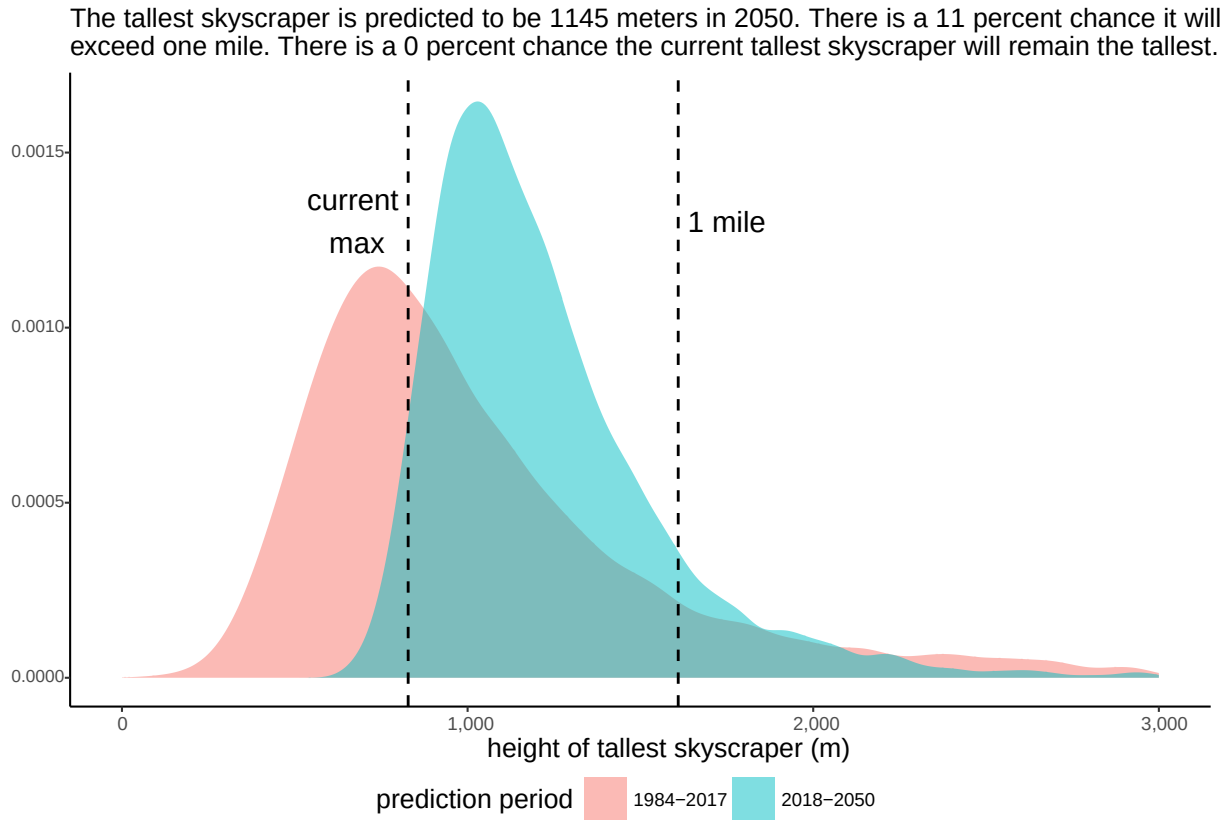


Figure 1.5: We predict the height of skyscrapers completed by 2050, thirty-three years after the data was collected at the end of 2017. The simulated density (blue) suggests the tallest building in the world is unlikely to exceed one mile (dashed line on right side). However, it will almost certainly be taller than the current tallest building, the Burj Khalifa (828 meters, dashed line on left side) and likely taller than the Jeddah Tower (one thousand meters), expected for completion in 2020. Had the same simulation been conducted in 1984, the density (red) would have found the tallest skyscraper in 2017 to be between six hundred and twelve hundred meters.

building is estimated to have seventy percent of the floors of the mile-high building—despite being sixty-two percent of the height. While diminishing marginal floors is reflected in most architectural designs, we find the number of floors will diminish faster with height than most designs anticipate. However, the exact relationship between height and number of floors will vary by city.

The joint distribution of skyscraper height and number of floors was extrapolated using the following bivariate extreme value model: Let (X, Y) be a random vector with GPD margins. Denote the respective parameter sets indexing the GPDs as (μ_x, σ_x, ξ_x) and (μ_y, σ_y, ξ_y), and consider the log probability integral transformations

$$\tilde{X} = -\log\left(\left[1 + \frac{\xi_x(X - \mu_x)}{\sigma_x}\right]^{-1/\xi_x}\right) \tag{1.1}$$

$$\tilde{Y} = -\log\left(\left[1 + \frac{\xi_y(Y - \mu_y)}{\sigma_y}\right]^{-1/\xi_y}\right). \tag{1.2}$$

Note that $(\tilde{X}, \tilde{Y})$ has standard exponential margins. The tail of the joint distribution $(\tilde{X}, \tilde{Y})$ is assumed to follow the asymmetric bivariate logistic distribution, i.e. given thresholds $u, v \gg 0$, for $X > u, Y > v$,

$$\Pr(\tilde{X} > x, \tilde{Y} > y) \propto \exp\left(-(1 - \theta_x)x - (1 - \theta_y)y - (x^r\theta_x^r + y^r\theta_y^r)^{1/r}\right), \tag{1.3}$$

where $\theta_x, \theta_y \in [0, 1]$ and $r \geq 1$. The asymmetric bivariate logistic distribution has the advantage of being simple and flexible, and it is suitable for larger sample problems. See Tawn (1988) and Coles (2001, p.142) for a detailed discussion of bivariate models for threshold exceedances.

Now suppose X is the height of a randomly selected skyscraper and Y the number of floors. We choose the thresholds $u = 225$ and $v = 59$ based on the fit of the marginal distribution as described in Section 4. Plots corresponding to the marginal analysis of skyscraper floors are displayed in Figures 1.8, 1.9, and 1.10 in the Appendix. The nine parameters μ_x, σ_x, ξ_x, μ_y, σ_y, ξ_y, θ_x, θ_y, and r are estimated using the heights and floors of all skyscraper exceeding 225 meters or 59 floors,

maximizing the censored likelihood:

$$L_c(\mu_x, \sigma_x, \xi_x, \mu_y, \sigma_y, \xi_y, \theta_x, \theta_y, r) = \prod_{x_i > u, y_i > v} f(x_i, y_i) \Pr(X > u, Y > v) \prod_{x_i \le u, y_i > v} f_Y(y_i) \Pr(X \le u) \prod_{x_i > u, y_i \le v} f_X(x_i) \Pr(Y \le v),$$

where f denotes the joint density function of (X, Y) implied by the transformations (1.1) and (1.2) and the distribution function (1.3), and f_X and f_Y denote the marginal density functions of X and Y, respectively.

The top left panel of Figure 1.6 displays the height and number of floors of every tall skyscraper, colored by its contribution to the censored likelihood, L_c. Skyscrapers below 225 meters and 59 floors (blue) do not contribute to the likelihood and are not used to estimate the parameters. Skyscrapers exceeding 225 meters and 59 floors (red) make up the first factor. The remaining skyscrapers (green) make up the second two factors. For example, a 250 meter skyscraper with 50 floors is treated like a 250 meter skyscraper whose floors are only known to be below 59. This approach is similar to the censored likelihood in Huser et al. (2016) except that skyscrapers at or below 225 meters and 59 floors are excluded from the analysis. Computation is discussed further in the Appendix.

The bivariate model fits the data well despite lacking the strong theoretical foundations that support the approximations in Sections 3 and 4. The first six parameter estimates from the bivariate model, $\hat{\mu}_x$, $\hat{\sigma}_x$, $\hat{\xi}_x$, $\hat{\mu}_y$, $\hat{\sigma}_y$, and $\hat{\xi}_y$, agree with the parameter estimates from the two marginal analyses on skyscraper heights and floors. In addition, simulations from the fitted model produce skyscrapers with heights and floors that are consistent with the data, as assessed visually and with predictive checks. For example, the average of the 325 extremely tall skyscrapers rises 4 meters per floor. The inner 95 percent of one thousand simulated averages rise between 3.7 and 4.5 meters per floor, with fifteen percent of simulated averages exceeding the observed average.

The maximum likelihood parameters are retained to estimate the conditional density of the number of floors for a one-thousand-meter skyscraper (Figure 1.6, top-right panel) and a one-mile

tall skyscraper (Figure 1.6, bottom-left panel). Dark blue regions represent a right-sided 50 percent interval, and light blue regions represent a right-sided 95 percent interval. These densities are compared with actual skyscraper plans (dotted line). The median one-thousand-meter skyscraper is estimated to have 107 percent the number of floors of the Jeddah Tower (to be completed in 2020). The median one-mile skyscraper is estimated to have roughly three-quarters the number of floors of the Mile-High Tower, two-thirds of Next Tokyo's Sky Mile Tower, and half the floors of Frank Lloyd Wright's The Illinois.

As in the previous two sections, the same analysis is performed with only data available before 1984. The bottom-right panel shows the conditional density of the 828 meter skyscraper as it would have been estimated using the threshold of 225 meters or 59 floors. In 1984, an 828 meter skyscraper would have been nearly twice the height of the current tallest building, the Willis Tower (then the Sears Tower, 442 meters, 108 floors). The conditional median predicts the typical 828 meter skyscraper would have 179 floors, ten percent more than the Burj Khalifa completed twenty-four years later. Simply scaling the Willis Tower to the height of the Burj Khalifa would yield 202 floors, overestimating the actual number by twenty-four percent. A linear regression model fit with extremely tall skyscrapers overestimates by fifteen percent.

The estimated relationship between floor and height varies considerably across cities. The top panel of Figure 1.7 shows the empirical median height and number of floors of extremely tall skyscrapers in select cities. The medians are each based on roughly ten observations, and thus sampling variation likely overstates the city-level differences between extremely tall skyscrapers in the year 2050.

We augment the bivariate model to estimate city-level medians. We allow the marginal GPD parameters to vary by city, according to a normal distribution with an unknown mean and variance. Such hierarchical models are often used to produce city-level estimates that have smaller errors on average than the corresponding stratified estimates. The use of a parameter hierarchy also has a Bayesian interpretation. See Coles (2001, p.169) and Vehtari (2017) for two discussions of Bayesian inference for threshold exceedances.

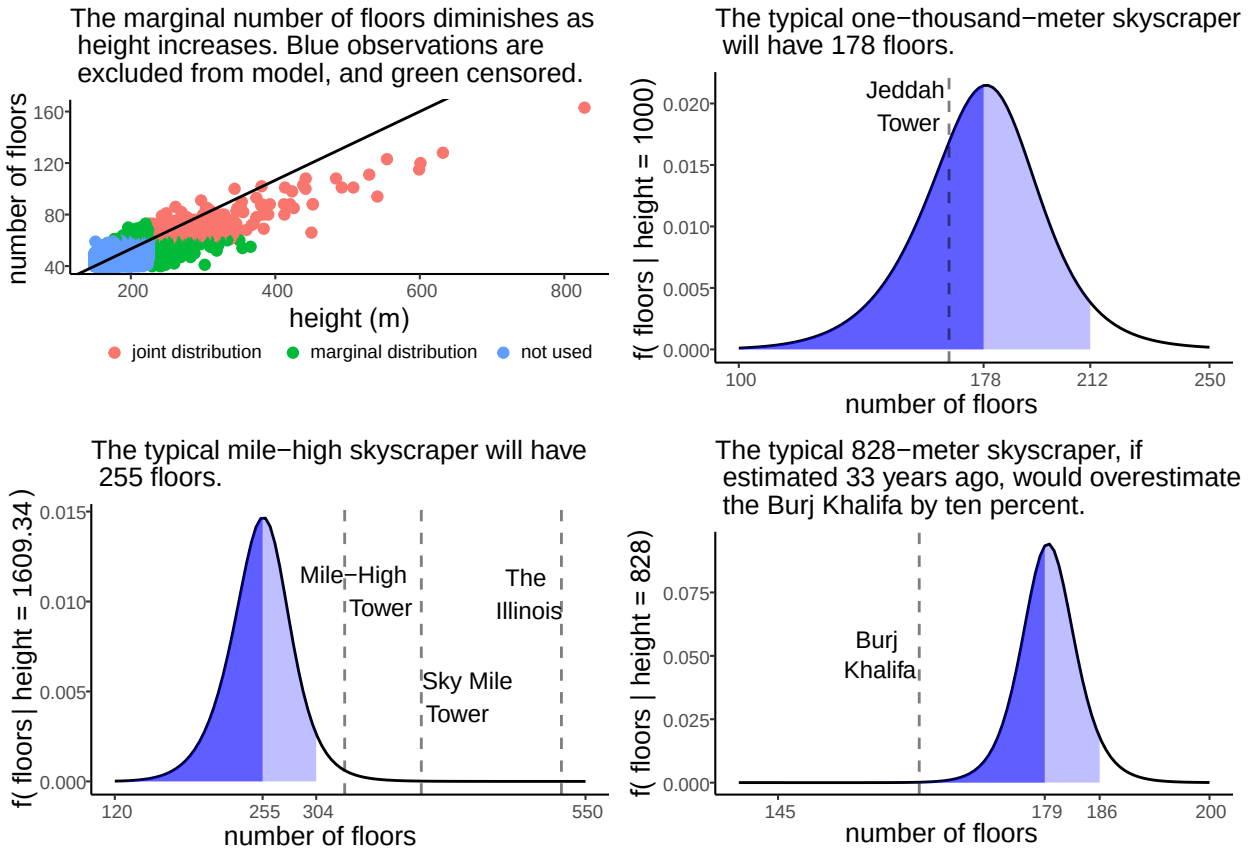


Figure 1.6: We estimate the number of floors in the tallest skyscrapers. The top-left panel shows the height and number of floors of tall skyscrapers—those exceeding 150 meters and 40 floors. The line represents the 3.8 meter rise per floor of the typical tall skyscraper. This ratio is not preserved as height increases. The other three panels compare the estimated conditional density of the number of floors in a skyscraper given its height to actual skyscraper designs.

The bottom panel of Figure 1.7 shows the estimated median for select cities from the hierarchical model with parameters selected by maximum likelihood. The median of the non-hierarchical model from the previous Figure is represented by a black dot. (Note that Hong Kong and New York City contain a disproportionately large number of extremely tall skyscrapers.) These city-level estimates can be seen as a compromise between the noisy empirical medians in the top panel and the more accurate but global median estimated by the non-hierarchical model. Yet despite partial pooling across cities, the typical height per floor ratio still spans a considerably large range: 3.6 (Hong Kong) to 4 (Moscow).

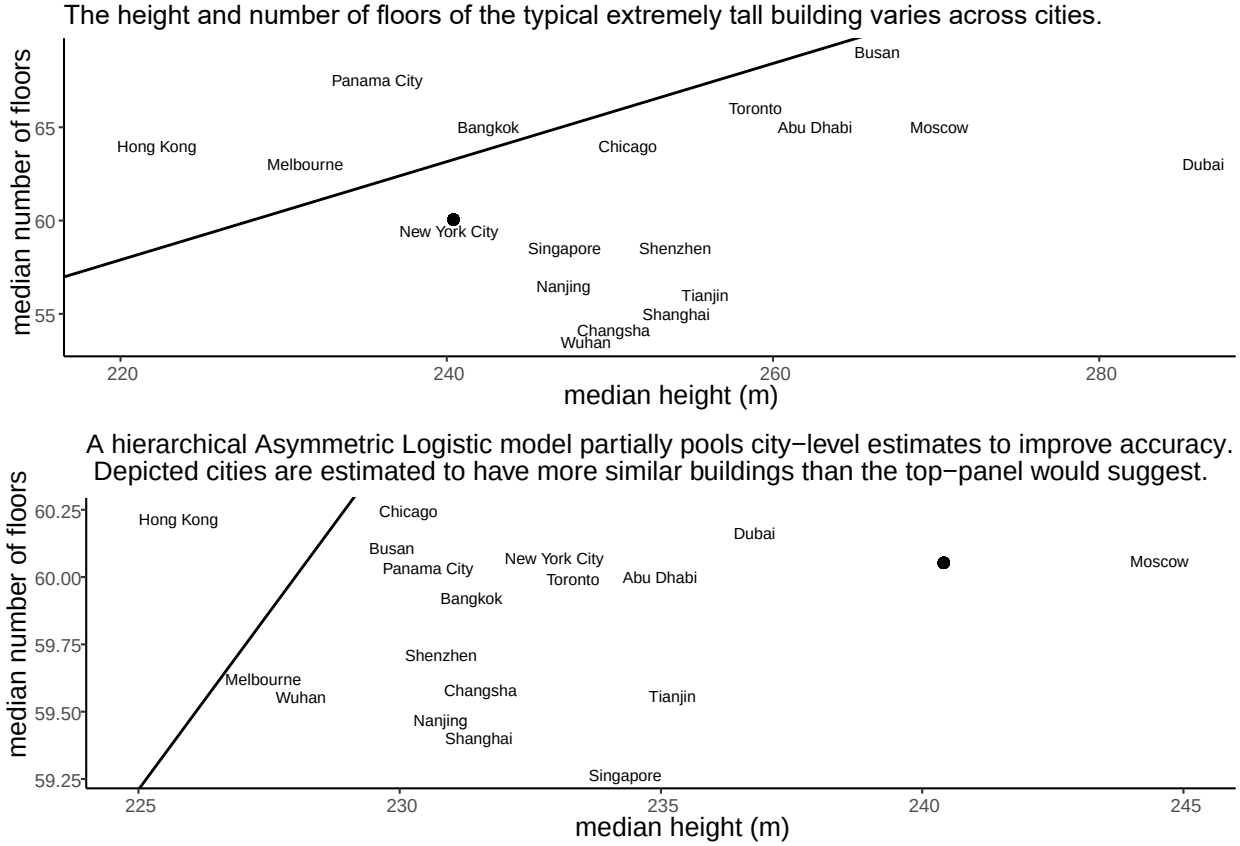


Figure 1.7: We estimate the height and number of floors of the extremely tall skyscrapers that will be completed in major cities by 2050. Empirical medians in the top panel likely overestimate between-city variation. Model estimated medians in the bottom panel compromise between the noisy empirical medians in the top panel and the accurate, but global median estimated by the non-hierarchical model. The line represents the 3.8 meter rise per floor of the typical tall skyscraper.

1.6 Discussion

This paper applied extreme value theory to predict the prevalence and nature of skyscrapers if present trends continue until the year 2050. The findings have methodological and policy consequences. This section discusses the methodological consequences, while the following section considers policy.

Section 3 found the number of skyscrapers completed each year followed a log-linear relationship. The relationship suggests a constant instantaneous growth rate—and extrapolation assumes construction will continue at this rate. When fit to the data, a logistic-linear relationship, which permits a declining growth rate, defaulted to a log-linear relationship and produced identical predictions. However, it would be incorrect to conclude the data is inconsistent with a declining growth rate in future years. It is possible growth is in fact logistic-linear, and the log-linear relationship found in Section 3 is only approximately true before the year 2018. In that case, the growth rate could decline substantially between 2018 and 2050. This scenario is speculative because it cannot be validated from the aggregated data. We therefore take the position that, though possible, a declining growth rate would reflect a departure from present trends.

Section 4 found that the heights of extremely tall skyscrapers are well described by a GPD with a positive shape, $\hat{\xi} \approx .2$. This means that the distribution of skyscraper heights has a heavy tail, and, theoretically speaking, the maximum does not exist. While this is obviously false—the laws of physics limit the height of any earthbound structure—it suggests sample averages and sums may be unreliable for inference and extrapolation. Researchers must be careful interpreting these quantities as evidence for their theories, especially with small sample sizes. For example, the fact that the average height does not increase with specific economic conditions may not indicate skyscraper construction is unrelated to those conditions. Conversely, increasing average height in recent years may provide a poor basis for predicting heights in future years. Quantiles, such as the median, can be more stable representatives of their theoretical analogs and may prove better alternatives for conducting inference and extrapolating as demonstrated in Figure 1.4. Cirillo and Taleb (2016)

make similar points for researchers using the total number of war casualties to determine whether humans are less violent than in the past and whether wide-scale war will return in the future.

Section 4 also found that the same GPD described extremely tall skyscraper heights over time. This suggests that the height of the tallest skyscrapers is driven by the exponential increase in the number of new buildings constructed each year and not a desire to build the typical building taller. The distinction is important for researchers investigating the factors that determine skyscraper height. For example, our findings are consistent with the theory that the demand for extremely tall buildings led to the use of innovative technologies, such as faster elevators. Had the reverse been true—had the development of innovative technologies prompted extremely tall skyscrapers—a systematic increase in skyscrapers would have been observed across the board as the technology became available, and the GPD parameters would have changed substantially over time. While it is not the intent of this paper to draw causal conclusions, we point out that forecasters need to be careful of "reverse-causality", attributing taller buildings as a consequence of a given factor instead of its cause. Bar (2016) documents a number of skyscraper myths that arise from the confusion of correlation with causation, for example the bedrock myth: that Manhattan's early skyscraper developers sought locations of shallow bedrock (p.210). These spurious correlations provide a poor basis for making predictions.

Section 5 demonstrated how city-level parameters might be estimated with a hierarchical model. The dataset contains 258 cities, although the typical city has only one skyscraper exceeding 150 meters and 40 floors. Predictions are still possible for these cities because the model borrows information across cities. Future research might augment the hierarchy to include country and region level parameters. Or alternatively, covariate information such as population and gross city product at the time each skyscraper was completed could be used instead. These covariates would make the modeling assumptions more plausible and may give insight into how cities might change policies to increase or decrease skyscraper activity—provided covariates are chosen judiciously and not based on spurious correlations. Spatiotemporal dependence could also be modeled directly as discussed by Bao et al. (2006), Chan and Gray (2006), and Ghil et al. (2011).

1.7 Conclusion

The major challenges confronting cities, such as sustainability, safety, and equality, will depend on the infrastructure developed to accommodate urbanization. Some urban planners suggest that vertical growth—the concentration of residents by constructing tall buildings—be used to accommodate density. Others argue that urbanization will be too rapid to be accommodated by vertical growth alone.

This paper finds that the construction of tall skyscrapers will outpace urbanization if present trends continue. Cities currently have around 800 tall skyscrapers per billion people. By 2050, cities are estimated to have 6,800 per billion people. The tallest among these will be fifty percent higher than those today and therefore able to accommodate more people. However, these skyscrapers will not have fifty percent more floors since the marginal capacity will diminish as heights increase. For example, the one thousand meter building will have seventy percent the floors of the mile-high building, despite being sixty-two percent of the height.

Future research might consider a different forecast horizon, although our choice of 2050 was not arbitrary. The UN World Population Prospects (UN, 2018) focuses on 2050, and many city planners consider 2050 to be the "not so distant future" (Lake (1996), Kennedy (2010), Wakefield (2013), Brown (2014)). Schuerman (2014) writes: "We chose the year 2050 for a reason: It is far enough away so that we can demonstrate dramatic changes in the climate, and yet near enough that many people alive today will still be living in the city."

A 2050 forecast was also practical. Skyscraper construction began in earnest after 1950, resulting in sixty-seven years of data. The year 2050 is thirty-three years away, which allows us to assess a thirty-three year forecast strategy by fitting the model on the first thirty-three years (1950-1983) and evaluating its performance on the following thirty-three years (1984-2017). The year 1984 is a meaningful break point. It marks the major shift in the second half of the twentieth century from the international period to contemporary skyscraper construction (Ascher and Vroman, 2011, p. 18).

This paper has not investigated whether skyscrapers will be constructed in the cities that need them most. Nor has it investigated whether skyscraper development should be used to accommodate density in the first place. Instead, extreme value theory provided a principled basis to forecast future trends and quantify uncertainty. It relies on the assumption that present trends continue, and there are a variety of reasons why future trends may deviate from the past sixty-eight years. For example, unprecedented technological change may result in new materials or methods that substantially reduce the cost of construction. Cultural changes could shift how residents live or work, perhaps freeing up commercial space for residential purposes. There is also the possibility of a hiatus due to global upheaval, not unlike the period spanning the Great Depression and Second World War.

We conclude by stressing that extreme value theory is one of many principled strategies that could be used to predict skyscraper development and the effects of urbanization more broadly. The previous sections could be augmented by integrating theories from architecture, engineering, policy, and social science. The incorporation of expert knowledge is always useful, but it is particularly desirable with extreme value analysis, as heavy tailed distributions are sensitive to outliers and benefit from the context afforded by the theory of other disciplines. More importantly, all disciplines will be necessary to anticipate how cities will respond to the greatest migration in human history and solve perhaps the principal challenge of our time.

1.8 Appendix

All models in Sections 4 and 5 are written in the Stan probabilistic programming language (Carpenter et al., 2017), and maximum likelihood estimates are computed using the LBFGS algorithm in the R Core Team (2018) package `rstan` (Stan Development Team, 2018), unless otherwise noted. Standard errors are computed from the observed Fisher information. Constrained parameters such as scales and shapes are transformed so that they are unconstrained, the Fisher information is calculated, and the constrained standard error is approximated using the delta method. Vehtari (2017) provides an introduction to univariate extreme value analysis with Stan. Carpen-

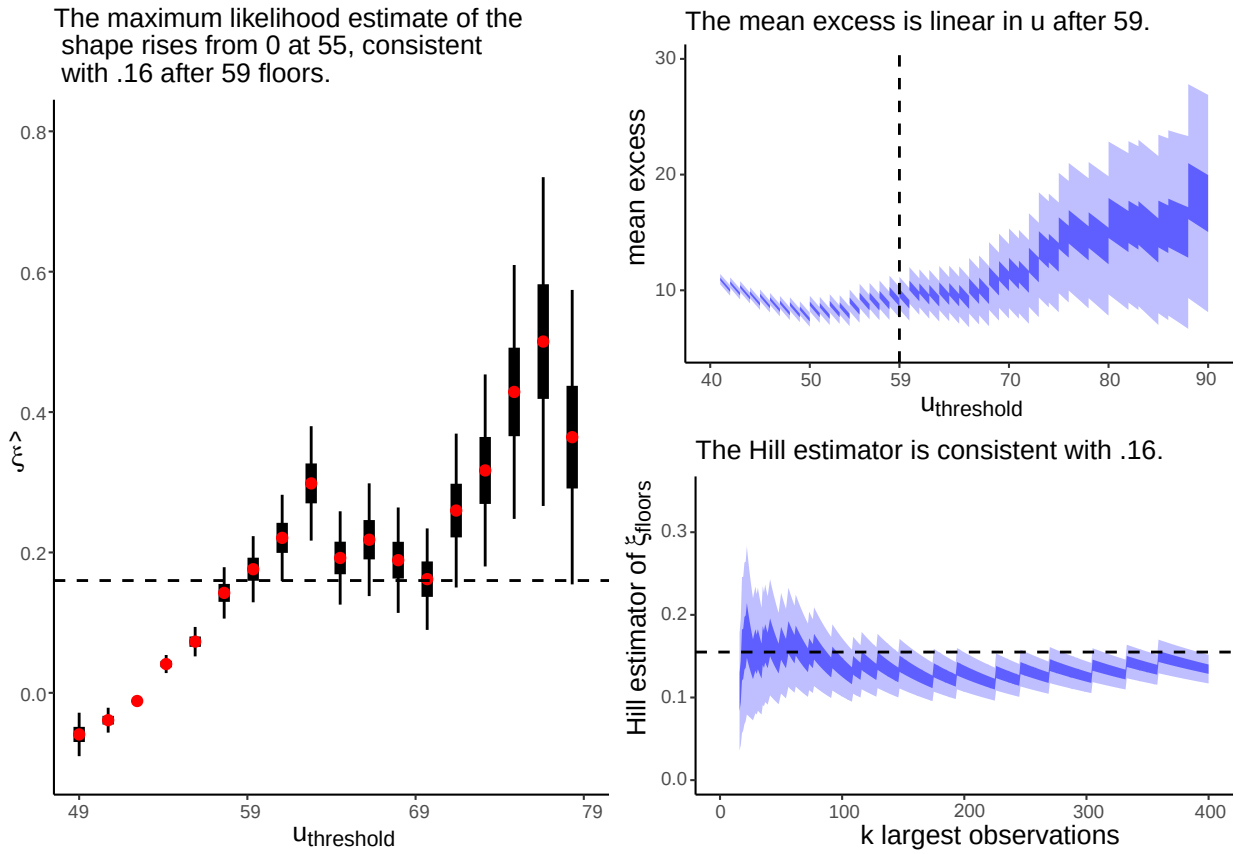


Figure 1.8: We approximate the distribution of the number of floors of tall skyscrapers using a GPD. The left panel shows maximum likelihood estimates and confidence intervals (50 and 95 percent) of the GPD shape parameter, excluding observations below a sequence of thresholds $\{u_i\}$. The right panel shows a Mean Excess plot (top) and a Hill plot (bottom)

ter et al. (2017, Section 2.6) discuss the LBFGS algorithm for maximum likelihood estimation. Predictive intervals are estimated by parametric bootstrap.

A 225 meter threshold for height (59 for floors) is used when fitting the GPD model. Note that in the limit, μ should equal the threshold u. We allow μ to vary ($0 < \mu < \min(x_i)$) in order to add flexibility to the model. This choice does not impact the conclusions in Section 4. Our point estimates match the output from the univariate `gpdFit` function in the package `fExtremes` (Wuertz, Setz, and Chalabi, 2017), which sets $\mu = u$.

Before fitting the bivariate models, the number of floors was scaled by 3.8. The median skyscraper rises 3.8 meters per floor, and scaling by this number aligns the marginal distributions.

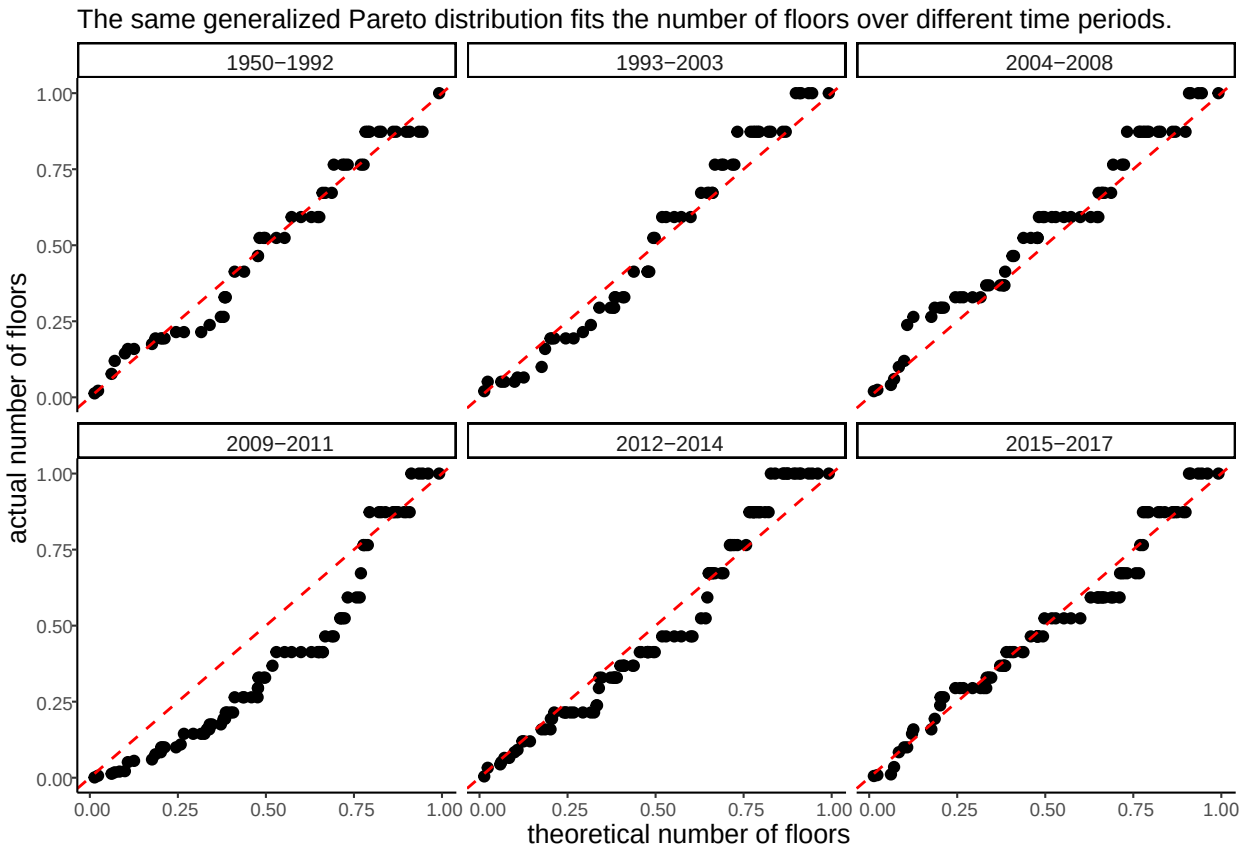


Figure 1.9: We demonstrate the distribution of the number of floors has changed little over time by dividing extremely tall skyscrapers into sextiles based on the year completed and constructing a p-p plot for each sextile. The distribution of theoretical floors is the generalized Pareto distribution with parameters estimated by maximum likelihood.

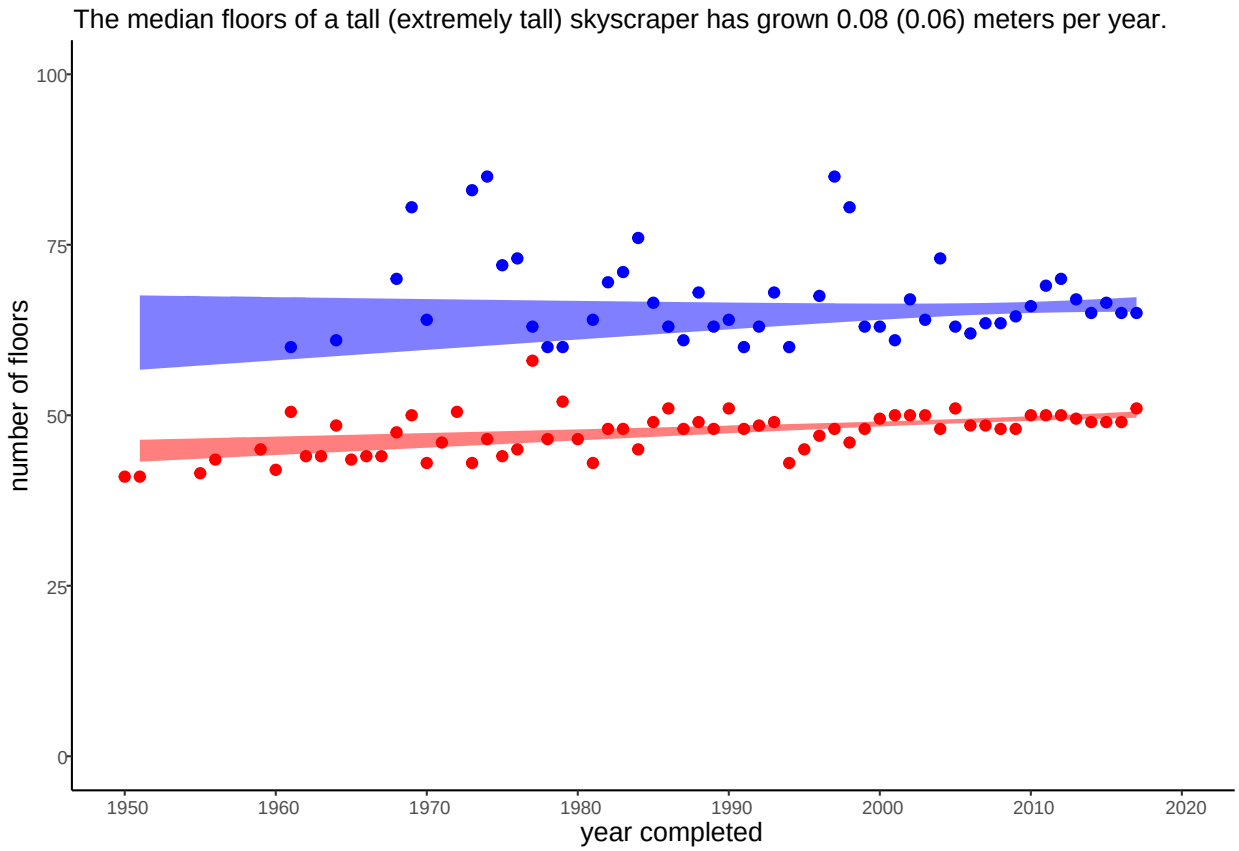


Figure 1.10: The median number of floors of extremely tall skyscrapers (blue points) has not increased significantly over the last forty years. A median regression indicates extremely tall skyscrapers grow by .06 floors each year (p-value = .11). A 95 percent confidence interval for the median regression line is shaded blue. The median number of floors of tall skyscrapers (red) is shown for comparison.

1.9 References

Al-Kodmany, Khei. 2012. “The Logic of Vertical Density: Tall Buildings in the 21st Century City.” International Journal of High Rise Buildings 1 (2).

Angel, Shlomo, Jason Parent, Daniel L Civco, and Alejandro M Blei. 2011. “Making Room for a Planet of Cities.” Lincoln Institute of Land Policy Cambridge, MA.

Ascher, Kate, and Rob Vroman. 2011. The Heights: Anatomy of a Skyscraper. Penguin Press London, England.

Balkema, August A, and Laurens De Haan. 1974. “Residual Life Time at Great Age.” The Annals of Probability. JSTOR, 792–804.

Barr, Jason. 2012. “Skyscraper Height.” The Journal of Real Estate Finance and Economics 45 (3). Springer: 723–53.

———. 2016. “Building the Skyline: The Birth and Growth of Manhattan’s Skyscrapers.” Oxford University Press.

———. 2017. “Asia Dreams in Skyscrapers.” The New York Times.

Barr, Jason, and Jingshu Luo. 2017. “Economic Drivers: Skyscrapers in Chinca.” CTBUH Research Report.

Barr, Jason, Bruce Mizrach, and Kusum Mundra. 2015. “Skyscraper Height and the Business Cycle: Separating Myth from Reality.” Applied Economics 47 (2). Taylor & Francis: 148–60.

Bao, Yong, Tae-Hwy Lee, and Burak Saltoglu. 2006. “Evaluating predictive performance of value-at-risk models in emerging markets: a reality check” Journal of forecasting 25 (2). Wiley Online Library: 101–128.

Benktander, Gunnar and C. Segerdahl. 1960. “On the analytical representation of claim distributions with special reference to excess of loss reinsurance”, in: XVIth International Congress of Actuaries, Brussels.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. 2015. “Time series analysis:

forecasting and control." John Wiley & Sons.

Brown, Hillary, and Steven Caputo. 2007. Bird-safe building guidelines. New York City Audubon Society. http://www.nycaudubon.org/pdf/BirdSafeBuildingGuidelines.pdf.

Brown, Lawrence A. 2014. "The City in 2050: A Kaleidoscopic Perspective." Applied Geography, 49: 4–11.

Canepari, Zack. 2014. "Essay: A Planet of Suburbs." Economist.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." Journal of Statistical Software 76 (1). Columbia Univ., New York, NY (United States); Harvard Univ., Cambridge, MA (United States).

Chan, Kam Fong and Philip Gray. 2006. "Using extreme value theory to measure value-at-risk for daily electricity spot prices' International Journal of Forecasting 22 (2). Elsevier: 283–300.

Cirillo, Pasquale and Nassim Nicholas Taleb. 2016. "On the statistical properties and tail risk of violent conflicts" Physica A: Statistical Mechanics and its Applications. 452: 29–45.

Clark, William Clifford, and John Lyndhurst Kingston. 1930. The Skyscraper: Study in the Economic Height of Modern Office Buildings. American Institute of Steel.

Cohen, Barney. 2006. "Urbanization in Developing Countries: Current Trends, Future Projections, and Key Challenges for Sustainability." Technology in Society 28 (1-2). Elsevier: 63–80.

Coles, Stuart, 2001. An Introduction to Statistical Modeling of Extreme Values. Vol. 208. Springer.

Council on Tall Buildings and Urban Habitat. 2017. "The Skyscraper Center."

Curl, James Stevens, and Susan Wilson. 2015. The Oxford Dictionary of Architecture. Oxford University Press, USA.

D'Amico, Guglielmo, Filippo Petroni, and Flavio Prattico. 2015. "Wind speed prediction for wind farm applications by extreme value theory and copulas" Journal of Wind Engineering and

Industrial Aerodynamics 145. Elsevier: 229–236.

Davison, Anthony C. and Richard L. Smith (1990). "Models for Exceedances over High Thresholds". Journal of the Royal Statistical Society: Series B 52, 393–442.

Dickey, David A. and Wayne A. Fuller. 1979. "Distribution of the estimators for autoregressive time series with a unit root." Journal of the American Statistical Association 74 (366a). 427–431.

Garreaud, RD. 2004. "Record-breaking climate anomalies lead to severe drought and environmental disruption in western Patagonia in 2016" Climate Research 74 (3): 217-229.

Gencay, Ramazan and Faruk Selcuk. 2004. "Extreme value theory and Value-at-Risk: Relative performance in emerging markets" International Journal of Forecasting 20 (2). Elsevier: 287–303.

Ghil M, P Yiou, S Hallegatte, BD Malamud, P Naveau, A Soloviev, P Friederichs, V Keilis-Borok, D Kondrashov, V Kossobokov, and O Mestre. 2011. "Extreme events: dynamics, statistics and prediction" Nonlinear Processes in Geophysics 18 (3). Copernicus: 295–350.

Ghosh, Souvik and Sidney I. Resnick. 2010. "Evaluating predictive performance of value-at-risk models in emerging markets: a reality check" Stochastic Processes and their Applications 120: 1492–1517.

Glaeser, Edward. 2011. "How Skyscrapers Can Save the City." The Atlantic.

Gottmann, Jean. 1966. "'Why the Skyscraper?" Geographical Review. JSTOR, 190–212.

Herrera, Rodrigo and Nicolás González. 2014. "The modeling and forecasting of extreme events in electricity spot markets" International Journal of Forecasting 30 (3). Elsevier: 477–490.

Hill, Bruce M. 1975. "A simple general approach to inference about the tail of a distribution" Annals of Statistics 3(5): 1163–1174.

Huser, Raphaël, Anthony C Davison, and Marc G Genton. 2016. "Likelihood Estimators for Multivariate Extremes." Extremes 19 (1). Springer: 79–103.

Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F. 2019. forecast: Forecasting functions for time series and linear models. 8.5. `http://pkg.robjhyndman.com/forecast`.

Hyndman RJ, Khandakar Y. 2008. "Automatic time series forecasting: the forecast package

for R." Journal of Statistical Software. 26 (3) 1-22.

James, Steele. 2001. "Architecture Today." Editura Phaidon, New York.

Kashef, Mohamad. 2008. "The Race for the Sky: Unbuilt Skyscrapers." CTBUH Journal, no. 1: 9–15.

Kennedy, Christine. 2010. "The City of 2050—An Age-Friendly, Vibrant, Intergenerational Community. " Generations, 34 (3): 70–75.

Koenker, Roger. 2018. Quantreg: Quantile Regression. https://CRAN.R-project.org/package=quantreg.

Lake, Andy. 1996. "The city in 2050: how sustainable?" World Transport Policy and Practice, 2 (1): 39–45.

Lepik, Andres. 2004. Skyscrapers. Prestel New York.

Pickands III, James. 1975. "Statistical Inference Using Extreme Order Statistics." The Annals of Statistics. JSTOR, 119–31.

R Core Team. 2018. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rose, Jonathan FP. 2016. "The Well-Tempered City: What Modern Science, Ancient Civilizations, and Human Nature Teach Us About the Future of Urban Life." Georgia Institute of Technology.

Schuerman, Matthew. 2014. NYC 2050: Climate Change and the Future of New York. WNYC. https://www.wnyc.org/story/nyc-2050-climate-change-and-future-new-york/.

Smith, Richard L. (1990). "Maximum likelihood estimation in a class of nonregular cases". Biometrika 72, 67–90.

Stan Development Team. 2018. "RStan: The R Interface to Stan." http://mc-stan.org/.

Sennott, R Stephen. 2004. Encyclopedia of 20th-Century Architecture. Routledge.

Swilling, Mark. 2016. "The Curse of Urban Sprawl: How Cities Grow, and Why This Has to

Change." The Guardian.

Tawn, Jonathan A. 1988. "Bivariate Extreme Value Theory: Models and Estimation." Biometrika 75 (3). Oxford University Press: 397–415.

Thompson, Vikki, Nick J Dunstone, Adam A Scaife,Doug M Smith, Julia M Slingo, Simon Brown, and Stephen E Belcher 2017. "High risk of unprecedented UK rainfall in the current climate" Nature communications 8 (1). Nature Publishing Group: 107.

UN, DESA. 2015. "World Urbanization Prospects: The 2014 Revision." United Nations Department of Economics and Social Affairs, Population Division: New York, NY, USA.

———. 2018. "World Urbanization Prospects: The 2018 Revision, Key Facts." United Nations Department of Economics and Social Affairs, Population Division: New York, NY, USA.

Vehtari, Aki. 2017. "Extreme Value Analysis and User Defined Probability Functions in Stan."

Wakefield, Jane. 2013. How will our future cities look?. BBC News. `https://www.bbc.com/news/technology-20770518`.

Willis, Carol. 1995. Form Follows Finance: Skyscrapers and Skylines in New York and Chicago. Princeton Architectural Press.

Wuertz, Diethelm, Tobias Setz, and Yohan Chalabi. 2017. FExtremes: Rmetrics - Modelling Extreme Events in Finance. `https://CRAN.R-project.org/package=fExtremes`.

Chapter 2: Forecasting the 2020 American Presidential Election with a Generational Voting Model

with Yair Ghitza and Andrew Gelman

We predict how the political events that occurred during the Trump administration will factor into the 2020 American Presidential Election. We begin by motivating the advantage of a generational voting model, in which voters' partisan preferences are modeled as a running tally of impressions left by the political events they live through. We then state the model explicitly. Finally, we predict how different generations will respond to the political events that occurred during the Trump presidency. Our results raise new questions about the comparability of generations in an era of extreme political polarization.

2.1 Introduction

The outcome of the 2020 American presidential election will depend on how voters assess the political events that occurred during the Trump administration—from the aftermath of the 2016 election, Mueller Report, and impeachment hearings to the coronavirus pandemic, recession, and widespread protests. Although unprecedented in several respects, it is because history repeats that past elections provide a scientific basis for predicting how voters will assess these events. The challenge is determining the manner in which history repeats, and thus the exact relevance of past elections, which may only be clear long after the 2020 election.

We predict how voters will assess the political events of the Trump administration using a generational voting approach. The model builds on political socialization theory, which describes how political events influence voting behavior (for example, see Hyman (1959), Campbell et al. (1960), and Jennings and Markus (1981)). Generally speaking, the value of this or any theory is that, if the

underlying assumptions are reasonable, the problem of predicting a complex phenomenon such as voting behavior is reduced to a manageable number of unknown factors. The advantage of political socialization theory specifically is that the assumptions are supported by decades of research, transparent, and easily discussed among experts or communicated to a lay audience. We believe this advantage is particularly desirable given the unusual circumstances in which the 2020 election will take place.

We proceed in three sections. We first motivate the main assumption underlying our generational voting model, which permits the comparison of voters across generations. We then describe the data and model (for related literature, see for example Burnham (1970), Beck and Jennings (1979), and Beck (1991)). Finally, we forecast how voters will assess the political events that occurred during the Trump presidency. Our main finding is that Generation Z has become more Democratic over the course of the Trump administration, but appear to lean only slightly Democratic overall when compared to previous generations—raising questions about the comparability of generations in our era of extreme political polarization.

2.2 We use the Gallup Organization's presidential approval rating time series to compare voters across generations.

Political socialization theory provides a broad framework for understanding how voters assess political events. To summarize, political events make impressions on voters, and the cumulative weight of those impressions determines which candidate a voter prefers. While every impression is unique, they typically coincide with a voter's demographic characteristics—such as age, race, ethnicity, and sex—reflecting the common conditions under which voters experience political events.

Forecasting an election with political socialization theory requires modeling how voters form their impressions. The model is then extrapolated to predict the impressions left by new political events. The problem is that both the political events and the manner in which voters assess them change over time, and the two must be disentangled before prediction is possible. For example, that older voters prefer conservative candidates could reflect the fact that voters become more

conservative with age or the fact that the current generation of older voters acquired conservative-leaning ideologies from political events unique to their adolescence. In the first case, we would expect future voters to become more conservative with age, while in the second case, we would expect future voters to retain their childhood ideologies, all else equal.

The natural solution is to compare voters from different generations when they were the same age and of arguably similar impressionability. Continuing the previous example, we could calculate the proportion of the older generation's conservative leanings that are due to adolescence by comparing the childhood of this generation to the childhoods of other generations. Such comparisons fill our public discourse, from the specific advice beginning "if I were your age ..." to the general nostalgia of a previous generation. However, there are any number of ways to compare two childhoods, and few articulate what exactly they mean if they were another's age. (Strictly speaking, if you were another's age, you would share their experiences and thus behave like them, although that is rarely what anyone means.) Even seasoned researchers speak casually about the effect of age without clarification.

The conceptual ambiguity inherent in intergenerational comparisons extends to all concomitant variables—variables that cannot be manipulated experimentally and represent abstract experiences whose relevance is subjective and case dependent, such as age, race, ethnicity, and sex. In order to fully realize the advantages of political socialization theory, namely transparent and easily communicated predictions, we must be explicit how different generations are compared. Our approach uses the Gallup Organization's presidential approval rating time series to represent political events and capture the changing zeitgeist; our main assumption is that voters who live through similar ratings adopt similar political ideologies. The approval ratings thus permit the comparison of voters across generations, not unlike how the consumer price index permits the comparison of prices across generations or the human development index permits the comparison of standard of living. The following section describes how exactly we make these comparisons, and the final section questions whether approval ratings sufficiently capture our new era of political polarization.

2.3 We model partisan preferences as a running tally of impressions left by the political events.

We assemble a dataset of responses to public opinion surveys about the American presidential election from five sources: (1) the ANES (elections 1952-2016), (2) the Gallup presidential polling dataset from the Roper Center's iPoll database (1952-2016), (3) the Annenberg National Election Studies (2000, 2004, and 2008), (4) the Greenberg Quinlan Rosner Research internal campaign polls (2012), and (5) the CNN/ORC and Pew polls (2016). The dataset has 215,693 complete responses after discarding undecided voters.

Let $a = \{1, 2, \ldots, 70\}$ denote the age of the respondent, $p = \{1956, 1960, \ldots, 2016\}$ the year of the response, $g = \{$non-Southern white male, Southern white male, minority male, non-Southern white female, Southern white female, and minority female$\}$ the demographic group of the respondent, and $h = \{$Annenberg, Gallup, NES, GQRR, CNN/ORC/Pew$\}$ the survey house that collected the response. We group all minority respondents together by sex since the data does not distinguish consistently between minority groups in early years; white refers to non-Hispanic white.

We also collect the Gallup Organization's long-running presidential approval rating time series from August 1937 to June 2020. Let x_t denote the Republican-directional presidential approval rating in year t, which is calculated by (1) subtracting 50% from the Gallup presidential approval rating in the year t, and (2) multiplying the difference by -1 if the sitting president was a Democrat.

The covariates, a, p, g and h, partition respondents into mutually exclusive cells. For each cell, j, we denote the age, period, group, and source of the responses by $a[j]$, $p[j]$, $g[j]$, and $h[j]$. Let y_j denote the number of respondents supporting the Republican candidate in cell j, and n_j the number preferring either the Republican or Democratic candidate. We model

$$y_j \sim \text{Binomial}(n_j, \theta_j),$$

where θ_j is the proportion supporting the Republican presidential candidate within cell j.

We decompose θ_j into four effects. We define a generational effect as the running tally of

impressions left by the political events experienced by the voters of cell j up until period $p[j]$,

$$\gamma_j = \Omega_{g[j]} \sum_{i=1}^{a[j]} w_i \; x_{p[j]-a[j]+i}$$

where w_i denotes the age-specific weight of the rating at age i, and Ω_g denotes the scale of the age-specific weights for group g.

We define a period effect for each group, β_{pg}, a period and age-weight interaction, $\lambda_g w_a \beta_{pg}$, and an election effect,

$$\begin{aligned} B_j &= \beta_{p[j]g[j]} + \lambda_{g[j]} w_{a[j]} \beta_{p[j]g[j]} \\ &= \left(1 + \lambda_{g[j]} w_{a[j]}\right) \beta_{p[j]g[j]} \end{aligned}$$

We also define an age effect α_a and a house effect η_h. Put together, these effects sum to θ_j on the logit scale,

$$\text{logit}(\theta_j) = \alpha_{a[j]} + B_j + \gamma_j + \eta_{h[j]}$$

We smooth the age weights,

$$w_i \sim \text{Normal}\left(w_{i-1}, 0.01\right),$$

and specify normal distributions for α, β, and η with mean zero and standard deviations $\sigma_\alpha, \sigma_\beta$, and σ_η. The scale parameters, λ and Ω, are constrained to be positive.

2.4 Our model suggests the political events of the Trump administration will have smaller influence on average than the events experienced by previous generations of the same age.

We fit our generational model with Stan (Carpenter et al. (2017)) using the R (R Core Team (2020)) package `cmdstanr` (Gabry and Češnovar (2020)). We sample 4 chains for 2000 iterations each. In each iteration, we generate $\hat{\theta}_j$, the proportion supporting the Republican presidential candidate, for all groups, ages 0 to 70, and years 1937 to 2020. NOTE: We set $\beta_{pg} = \beta_{2016g}$ and $\eta = 0$ so that the $\hat{\theta}_j$ are comparable across years—at this time there is insufficient data to estimate β_{2020g}. We ensure satisfactory diagnostic and posterior predictive performance (Gelman et al. (2013)) before using sample quantiles for estimates (median) and credible intervals in the following five figures. Darker intervals represent 50 percent posterior regions and lighter intervals represent 95 percent regions.

Figures 1-4 each examine four age groups of voters representing four different generations—18-year-olds (Generation Z), 35-year-olds (Millennials), 52-year-olds (Generation X), and 68-year-olds (Baby Boomers). The top half shows the estimated Republican support from that age group since 1950. Generation Z and Millennials lean Democratic compared to previous voters of that age, while Generation X and Baby Boomers lean Republican. The bottom half shows how the political events experienced by the different generations (labeled by birth year for consistency across Figures) have shaped their political ideology. Each generation is politically agnostic at age 0. They then develop unique trajectories of partisan preferences according to the political events they live through—displayed up until the age of the current generation. Under our model, younger generations have become more Democratic from the events experienced during the Trump administration. For non-minority Generation Z and Millennials, the shift is smaller than past generations, suggesting recent political events have had a smaller on average influence than the events experienced by older generations at the same age. For older generations, such as Generation X and Baby Boomers, and minorities of all generations, the change is consistent with previous generations.

We conclude by revisiting our main assumption: that approval ratings permit intergenerational comparisons. If this assumption is reasonable, one explanation for our findings is that the Internet and 24-hour news cycle have dampened the long-term influence of political events. If this assumption is not reasonable, as recent polls appear to suggest, approval ratings may understate the extent to which younger generations support the Democratic candidate. Future work should examine how these ratings might be adjusted to better facilitate intergenerational comparisons. However, any adjustment would likely forgo the eighty-three year length of the approval rating time series or overall interpretability of the model. Whether the benefit of adjustment exceeds the cost will be clearer after the 2020 American presidential election, which is sure to test the assumptions of every forecasting method.

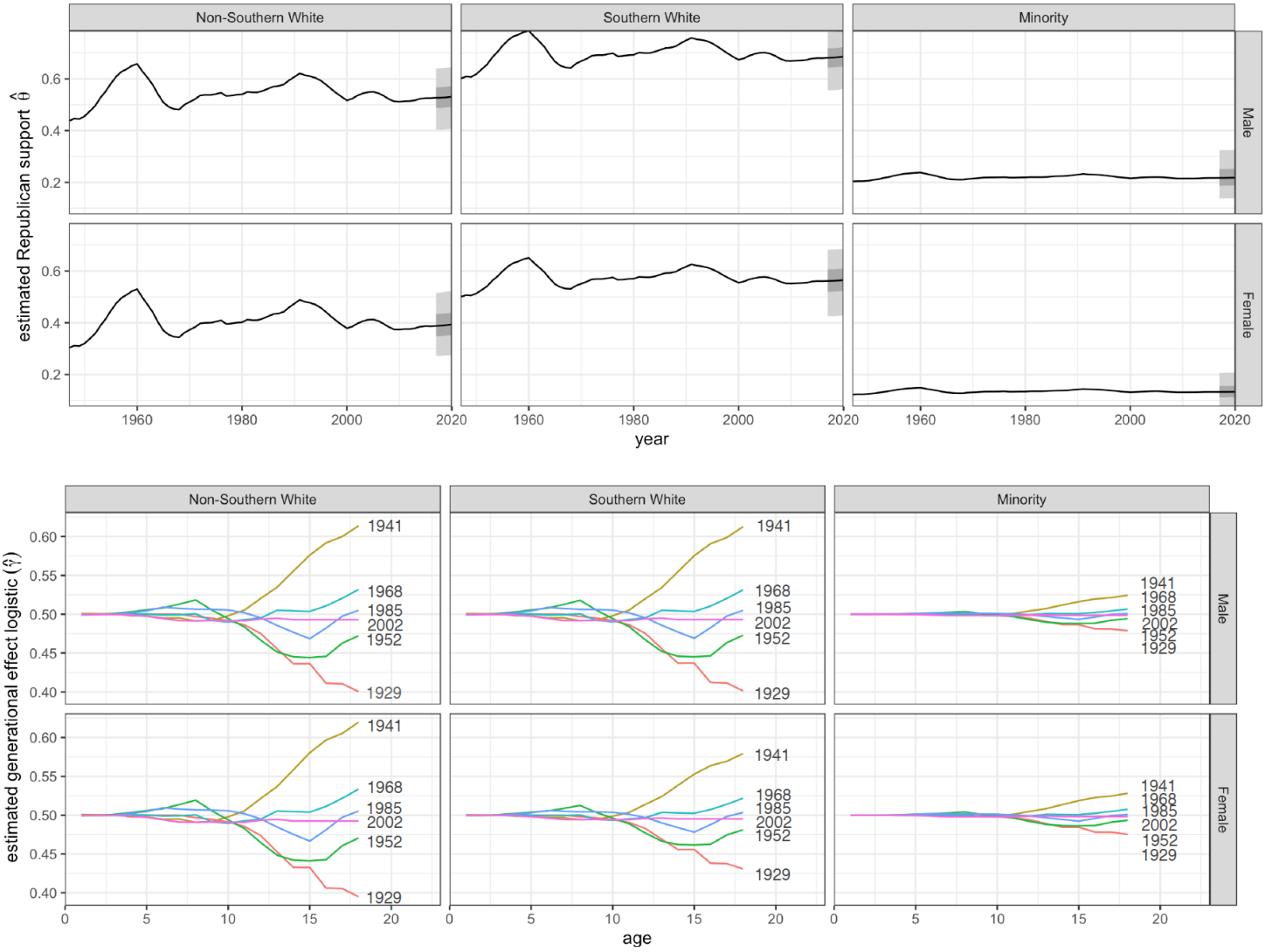


Figure 2.1: Comparing 18-year-olds (currently Generation Z, birth year 2002, magenta): Estimated Republican support from 18-year-olds since 1950 (top) and cumulative partisan impressions producing this support from birth to age 18 for select birth years (bottom).

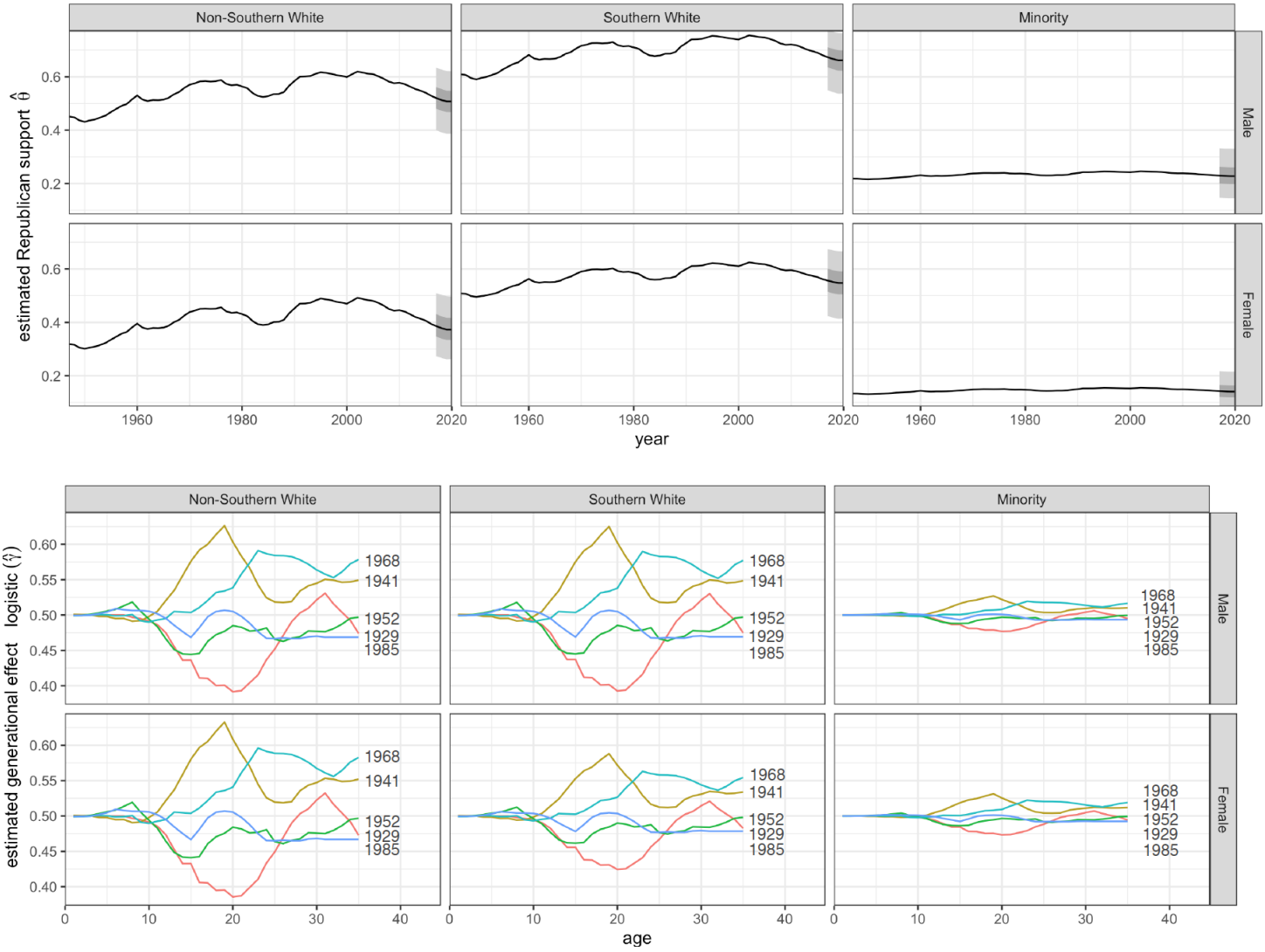


Figure 2.2: Comparing 35-year-olds (currently Millennials, birth year 1985, blue): Estimated Republican support from 35-year-olds since 1950 (top) and cumulative partisan impressions producing this support from birth to age 35 for select birth years (bottom).

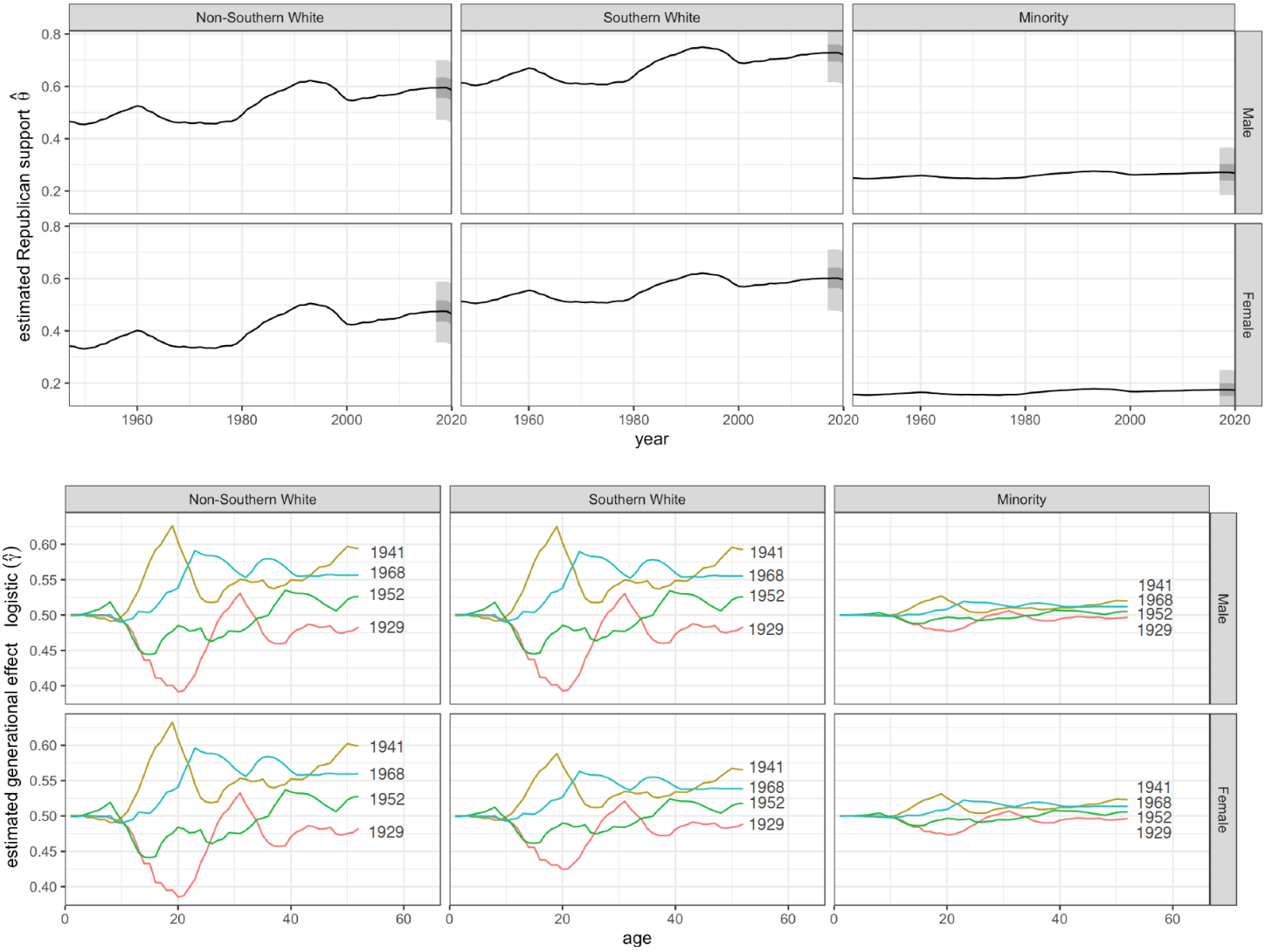


Figure 2.3: Comparing 52-year-olds (currently Generation X, birth year 1968, teal): Estimated Republican support from 52-year-olds since 1950 (top) and cumulative partisan impressions producing this support from birth to age 52 for select birth years (bottom).

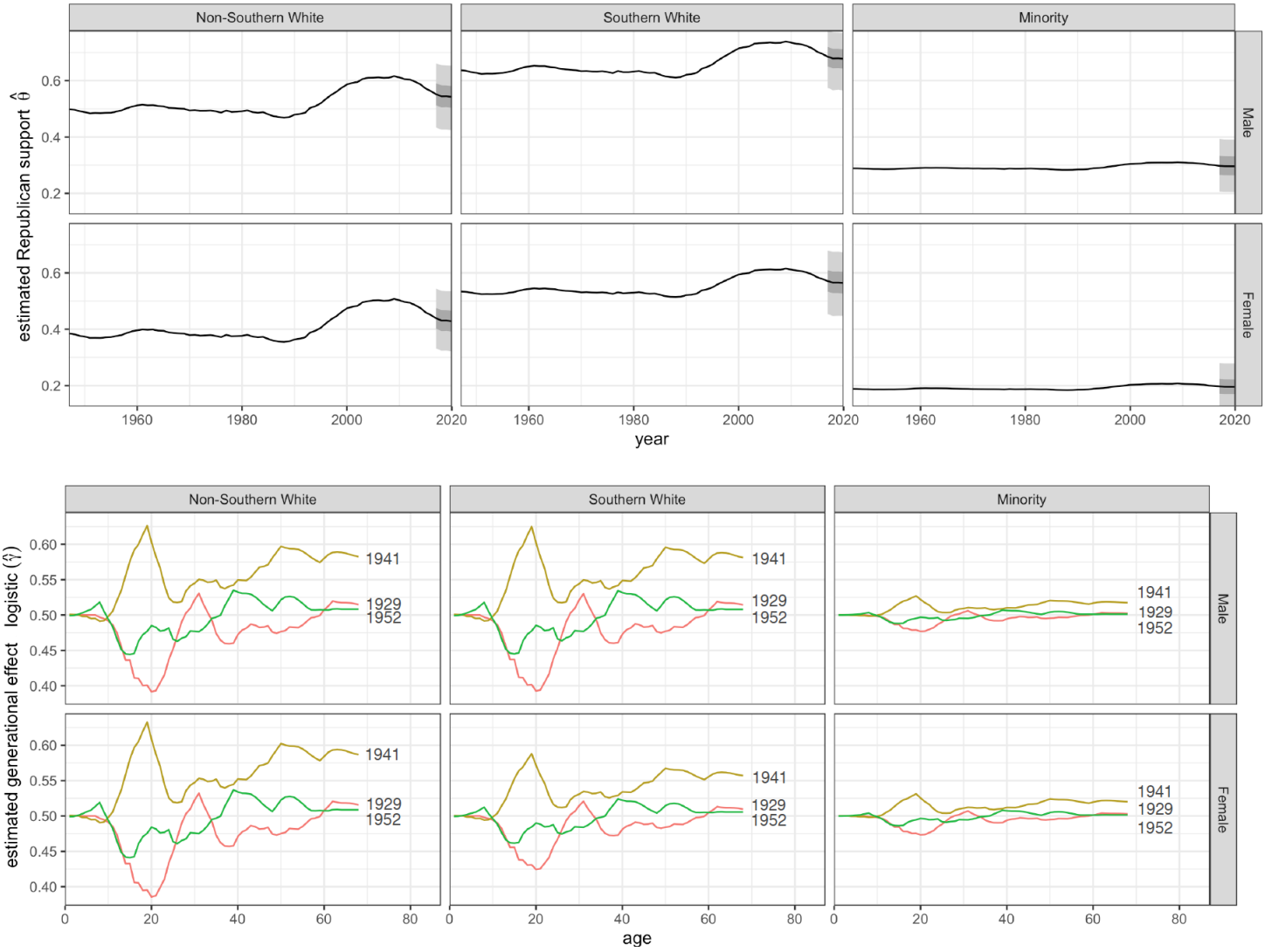


Figure 2.4: Comparing 68-year-olds (currently Baby Boomers, birth year 1952, green): Estimated Republican support from 68-year-olds (Baby Boomers) since 1950 (top) and cumulative partisan impressions producing this support from birth to age 68 for select birth years (bottom).

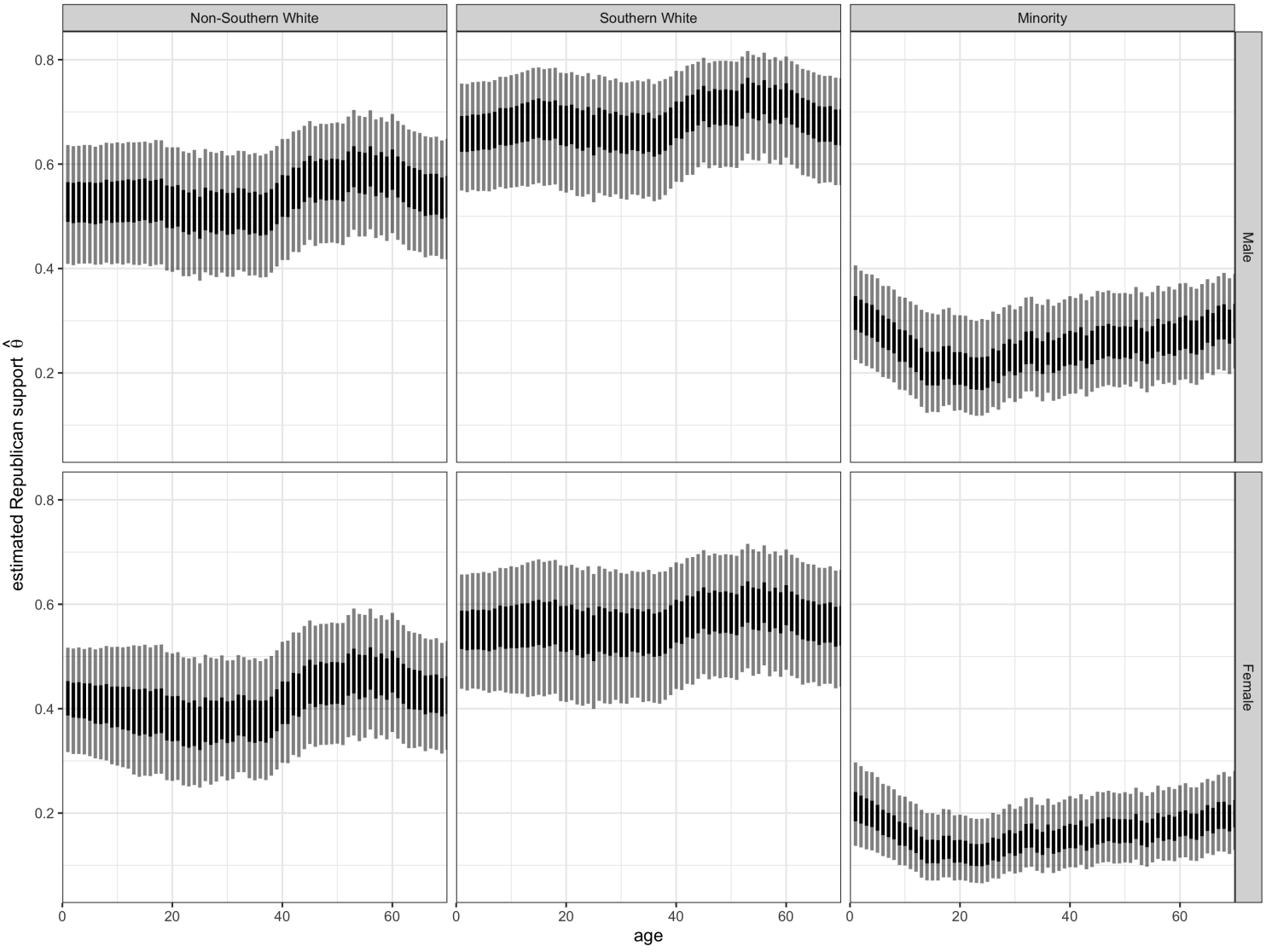


Figure 2.5: Relative vote share: Estimated Republican support by age, race, ethnicity, and sex. These estimates can be post-stratified to predict the outcome of the 2020 election, which we plan on doing when data becomes available.

2.5 References

Beck, Nathaniel. 1991. "Comparing Dynamic Specifications: The Case of Presidential Approval." Political Analysis, 51–87.

Beck, Paul Allen, and M Kent Jennings. 1979. "Political Periods and Political Participation." The American Political Science Review, 737–50.

Burnham, Walter Dean. 1970. Critical Elections and the Mainsprings of American Politics. New York: Norton.

Campbell, Angus, Philip E Converse, Warren E Miller, and Donald E Stokes. 1960. "The American Voter New York: Wiley."

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." Journal of Statistical Software 76 (1).

Gabry, Jonah, and Rok Češnovar. 2020. Cmdstanr: R Interface to 'Cmdstan'.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. Bayesian Data Analysis. CRC press.

Hyman, Herbert. 1959. "Political Socialization."

Jennings, M Kent, and Gregory B Markus. 1981. "Political Involvement in the Later Years: A Longitudinal Survey., in. Jennings, M. Kent & Niemi, Richard, 1981: Generations and Politics: A Panel Study of Young Adults and Their Parents." Princeton, NJ: Princeton University Press.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org/`.

Chapter 3: A Deeper Look at the Generational Voting Model

with Yair Ghitza and Andrew Gelman

We build a model of American presidential voting in which the cumulative impression left by political events determines the preferences of voters. The impression varies by voter, depending on their age at the time the events took place. We find the Gallup presidential approval rating time series reflects the major events that influence voter preferences, with the most influential occurring during a voter's teenage and early adult years. Our fitted model is predictive, explaining more than ninety percent of the variation in voting trends over the last half-century. It is also interpretable, dividing voters into five meaningful generations: New Deal Democrats, Eisenhower Republicans, 1960s Liberals, Reagan Conservatives, and Millennials. We present each generation in context of the political events that shaped its preferences, beginning in 1940 and ending with the 2016 election.

3.1 Introduction

We study generational voting in American presidential elections by modeling voters' partisan preferences as a running tally of impressions left by the political events they live through. When fit to data, the tally is weighted heavily by events that occur in a voter's teenage and early adult years. After early adulthood, voter preferences become consistent, and political events hold considerably less weight. The fitted model is predictive—explaining nearly all of the macro-level variation in voting trends over the past half-century—and interpretable—dividing voters into five meaningful generations.

Our model builds on a substantial literature in political science, sociology, and social psychology, beginning with the theory of "political socialization" (Hyman, 1959) and developed through

seminal works on American political behavior, such as The American Voter. These works used panels of high school students to establish the micro-level determinants of voting behavior. For example, Campbell et al. (1964) found party identification, the basis of political attitudes and voting behavior, is formed early in life and is influenced primarily by parents.[1]

However, these works were unable to agree on the macro-level determinants. For example, researchers observed that older voters were more likely to identify as Republican. Some argued this was the effect of aging: a social or psychological process pushed individuals towards a conservative viewpoint later in life. Others argued the effect was generational: the shared experiences of individuals from the same cohort happened to skew these voters Republican. Much ink was spilled attempting to disentangle the two. Crittendon (1962)emphasized age effects, while Cutler (1969) and Glenn and Hefner (1972) emphasized cohort effects.

Scholars soon discovered the problem with decomposing voter behavior into age, period, and cohort effects, the second of which refers to short-term influences of political attitudes that fail to leave a lasting impression. The effects are not identified because age, period, and cohort are collinear; a voter's age and cohort uniquely determine the period in which they vote (Converse, 1976; Glenn, 1976; Markus, 1983). Perfunctory attempts to estimate all three require model constraints that are difficult to interpret and cannot be validated from the data (Fienberg and Mason, 1979).

We resolve the age-period-cohort problem by directly modeling the impressions left by political

[1]Reviews of the early literature include Niemi and Sobieszek, 1977; Delli Carpini, 1989; Niemi and Hepburn, 1995), with Jennings and Niemi (1981) summarizing many of their substantial contributions. We briefly highlight how our approach compares to research in this area.

Burnham (1970) studies generational voting patterns over the long term, explaining system-wide shifts of roughly thirty-year increments. In contrast, our generations cover roughly fourteen-year increments, explaining the more rapid swing between liberal and conservative. The two definitions are best suited for studying their respective phenomenon; Burnham's theory does not explain rapid partisan swings between presidential administrations, while we do not model slower shifts, such as the gradual partisan shift of the South.

Beck and Jennings (1991) and Ostrom and Smith (1992) study the dynamics of presidential approval; a topic that remains relevant in our hyper-polarized era. Our work differs in that they model approval directly, while we use approval to model how voters choose candidates.

Beck and Jennings (1991) study the interaction between age and cohort effects in American politics, focusing on the period during the late 1960s and early 1970s when young voters were a major force in American politics. Our work follows their insight that "opportunities for political action ... vary with changes in the political stimuli across different periods." Beck and Jennings (1982), Beck (1991) use panel data to study political socialization of young Americans. We again follow their idea that adult political attitudes are a product of individual and social inputs.

events that researchers typically interpret as cohort effects. We use the Gallup presidential approval rating time series to instantiate these events for three reasons. First, the president is the most public and notable in American politics. The position is prominently associated with major political events, even when those events are unrelated to the presidency. Second, presidential elections are among the most salient events in American politics. By a wide margin, presidential turnout is higher than any other form of political participation. Lastly, the series continuously measures the public's evaluation of the president since the 1930s.

Because presidential approval ratings reflect the political events that determine presidential voting, we need only estimate the influence of those events at each age—along with a relatively small number of additional parameters discussed in the following sections. As a result, our model not only resolves the age-period-cohort problem, but, when fit to our massive dataset, quantifies generational trends with a precision unprecedented in the literature. Our three main findings are:

First, the political events that influence partisan preferences occur largely between the ages of 14-24, and a generation's preferred party is essentially locked-in by 40. This influence varies by race and region. It is strongest among non-Southern whites and relatively weak among minorities, suggesting considerable differences in the political socialization process.

Second, the impressions left by these events delineate five distinct generations. For example, consider white voters born in 1952 and socialized during the Kennedy and Johnson administrations. These voters are consistently 5-10 percentage points more likely to support Democratic presidential candidates than those born in 1968, who came of age during the presidencies of Carter, Reagan, and Bush I. We name these generations New Deal Democrats, Eisenhower Republicans, 1960s Liberals, Reagan Conservatives, and Millennials.

Third, period effects are important despite our focus on generations. But a simple model of period effects is insufficient for explaining voter preferences, even when voters are further divided by race and region. Our model explains significantly more macro-level variation, especially among non-Southern white voters. This suggests a single defining political event is less important in the formation of voter preferences than the cumulative impression left by a lifetime of events.

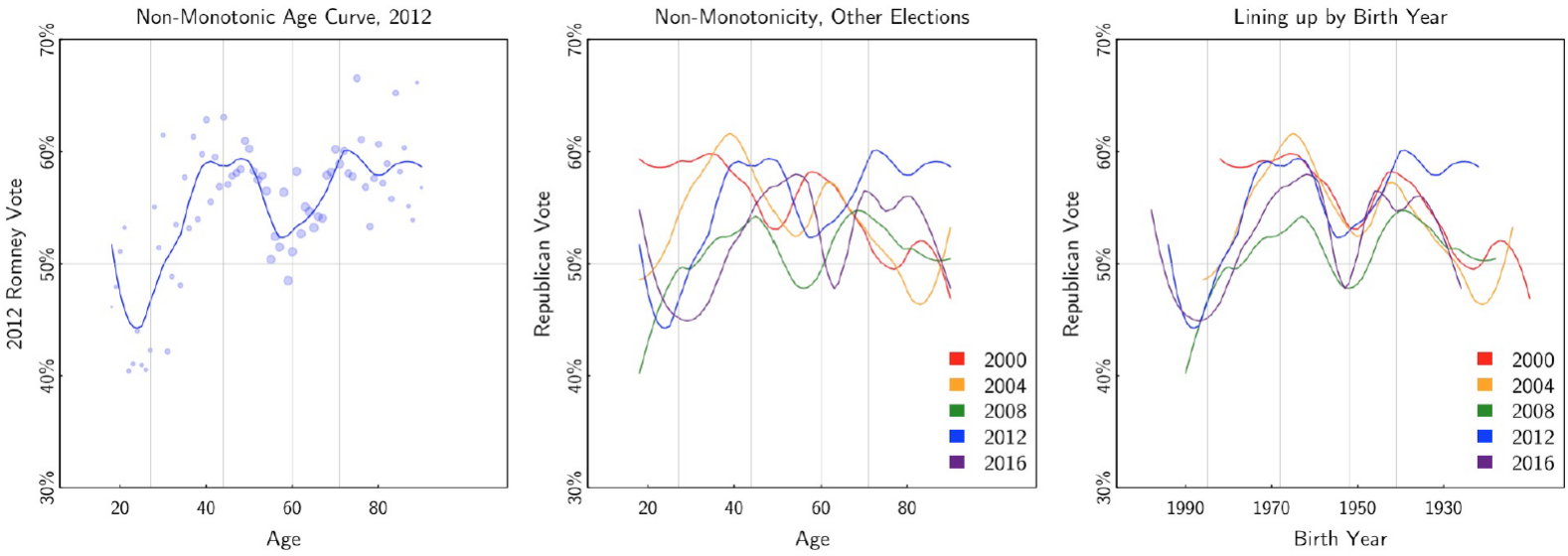


Figure 3.1: Raw data and loess curves, indicating the relationship between age and presidential voting preferences among non-Hispanic white voters for the 2000-2016 elections. From the left: (1) The relationship is non-monotonic and quite peculiar in 2012; instead of a linear or even quadratic relationship, the curve changes directions multiple times. (2) Non-monotonicity characterizes other elections as well. No clear pattern is apparent from this graph alone. (3) The true relationship emerges when the curves are lined up by birth year instead of age. The peaks and valleys occur at almost identical locations, indicating a generational trend.

We present the details and additional findings in five sections: (1) We describe the data and motivate our model in the context of the age-period-cohort problem. (2) We present the technical details of our model. (3) We fit the model to the data and interpret the results. (4) We narrate presidential preference over the past half-century, using the fitted model to quantify how political events left differential impressions on five generations of American voters. (5) We conclude with a brief discussion.

3.2 Data and Preliminary Evidence

We assemble a massive dataset from five sources: (1) the ANES cumulative dataset covering elections (1952-2016), (2) the Gallup presidential polling dataset from the Roper Center's iPoll database (1952-2016), (3) the Annenberg National Election Studies (2000, 2004, and 2008), (4) the Greenberg Quinlan Rosner Research internal campaign polls (2012 election cycle), and (5) the

CNN/ORC and Pew polls (2016 election cycle). We only use responses collected during presidential election years. There are 319,678 observations after removing incomplete records.[2]

Graphing the combined data provides strong—albeit preliminary—evidence of generational voting. To illustrate the reasoning behind our assessment and motivate the importance of our model, consider Figure 3.1, which displays the relationship between age and presidential vote choice for white respondents across all data sources.

The left panel combines the preference of voters in 2012 by age. That is, for each age group on the x-axis, the y-axis indicates the percent supporting the Republican candidate. The size of each bubble represents the number of voters surveyed. A best-fit curve is estimated using locally weighted regression (LOESS).

From the panel, it is clear that support for the 2012 Republican candidate, Romney, varied by age: The youngest white voters slightly supported Romney; voters around the age of 24 preferred Obama, the Democratic incumbent; Romney's vote grew steadily with age until 45, only to reverse direction until 60; it then climbed one last time to 70, before finally flattening. But the reason for this pattern is not obvious.

The center panel overlays curves for all presidential elections from 2000 to 2016. We remove the bubbles for clarity. As with the left panel, the patterns in each election are difficult to interpret. Moreover, there is no common trend across elections.

It is only in the right panel—when the data are combined by birth year (birth cohort) instead of age—that a common pattern emerges. The five curves align. Their peaks and valleys coincide, and, with the exception of the 2008 election, all curves are essentially on top of each other. This is especially true for voters born between 1940 and 1970, where the bulk of the data lie.

The common pattern establishes that voters share preferences with their birth cohort and maintain these preferences across elections. We reason this is evidence of generational voting: Voters from the same cohort live through the same interval of history; the cumulative impression left by

[2]Variables of interest are presidential vote choice, race/ethnicity, sex, state of residence, and age (or equivalently birth year (birth cohort), defined as the year of the survey response minus age). Responses are not weighted. Throughout this paper, white refers to non-Hispanic white.

the events of that interval produce a stable political identity, which explains the stable partisan preferences we observe in the right panel.

The weight of the evidence appears strong because of the simplicity of the explanation—the curve has two peaks around the birth years of 1941 and 1968 and one pro-Democratic valley around 1952, delineating five meaningful generations—and its consistency—the curve repeats over five elections, measured across multiple surveys conducted by different organizations, and unaltered by any complicated adjustment or statistical model. In comparison, the center panel suggests neither a simple nor consistent explanation.

Yet the evidence is preliminary because our reasoning does not explicitly account for non-generational explanations of vote choice. For example, political events during election years have been known to influence preferences monotonically across cohorts and create "uniform swings" (Ghitza and Gelman, 2013). We noticed a uniform swing earlier in the right panel, where the 2008 curve is lower for nearly every birth cohort.

Uniform swings suggest events, such as recessions, natural disasters, or war, can influence voters during the period in which they occur but not subsequent periods. It also stands to reason that life-cycle events, such as education, marriage, and retirement, can influence voters at the age in which they occur but not subsequent ages. We do not consider such period-specific and age-specific influences generational voting because they do not leave the lasting impression that defines a generation.

We build a model that includes both generational and non-generational determinants of voter preference in order to identify the events that define a generation and quantify their import. Our model avoids the well-known limitation of the traditional age-period-cohort model, which we briefly review in order to motivate our approach. An authoritative discussion of age-period-cohort models is given by Fienberg and Mason (1979).

The traditional age-period-cohort model for categorical data decomposes the log-odds additively into static age components (α), period components (β), and cohort components (γ). Let θ_{ap} denote the expected proportion of voters supporting the Republican candidate at age a during

election period p with birth year (birth cohort) $c = p - a$. We assume without loss of generality that the age and period indices have been centered: $\sum_a a = \sum_p p = 0$. Then the traditional model is written

$$\text{logit}(\theta_{ap}) = \alpha_a + \beta_p + \gamma_{p-a} \tag{3.1}$$

If we knew the summands, α_a, β_p, and γ_{p-a}, we could indirectly determine the generational import of political events. We might identify two cohorts, c_1 and c_2, differentially affected by those events and compute the difference, $\gamma_{c_1} - \gamma_{c_2}$. (Indeed, when we interpreted the curves in Figure 3.1, we compared the difference between cohorts.) However, the summands are not known, only the sum, and as it stands the system of equations given by (3.1) is indeterminate; the parameters α_a, β_p, and γ_{p-a} cannot be determined uniquely from $\text{logit}(\theta_{ap})$ and are said to be unidentified.

We divide the identification problem into two cases for ease of explanation. Contrasting the two sheds light on what exactly can be learned from age-period-cohort data. The first case is routine in categorical data analysis. If some solution, α'_a, β'_p, γ'_{p-a}, did exist, we could obtain an observationally equivalent second solution, α''_a, β''_p, γ''_{p-a}, by adding and subtracting a constant of magnitude δ to the right side of (3.1):

$$\begin{aligned}
\text{logit}(\theta_{ap}) &= \alpha'_a + \beta'_p + \gamma'_{p-a} \\
&= \alpha'_a + \beta'_p + \gamma'_{p-a} \pm \delta \\
&= \alpha'_a + (\beta'_p + \delta) + (\gamma'_{p-a} - \delta) \\
&= \alpha''_a + \beta''_p + \gamma''_{p-a}
\end{aligned}$$

This case poses no difficulty because δ is the same across cohorts, and therefore the difference between cohorts remains the same: $\gamma'_{c_1} - \gamma'_{c_2} = \gamma''_{c_1} - \gamma''_{c_2}$ regardless of δ. We impose the restriction $\sum_a \alpha_a = \sum_p \beta_p = 0$ to avoid this problem, justifying our choice on the basis that any other restriction, for example $\alpha_0 = \beta_0 = 0$, results in the same intercohort comparisons.

But even with this restriction, equation (3.1) is unidentified. A consequence of the linear relationship between age, period, and cohort, $c = p - a$, is that we can add and subtract $(p - a)\delta$ to the right side of (3.1):

$$\begin{aligned}
\text{logit}(\theta_{ap}) &= \alpha'_a + \beta'_p + \gamma'_{p-a} \\
&= \alpha'_a + \beta'_p + \gamma'_{p-a} \pm (p - a)\delta \\
&= (\alpha'_a - a\delta) + (\beta'_p + p\delta) + (\gamma'_{p-a} - (p - a)\delta) \\
&= \alpha''_a + \beta''_p + \gamma''_{p-a}
\end{aligned}$$

This second case poses significant difficulty because $(p - a)\delta = c\delta$ is not the same across cohorts. For any solution, α'_a, β'_p, γ'_{p-a}, we can generate an observationally equivalent second solution, α''_a, β''_p, γ''_{p-a}, whose difference,

$$\begin{aligned}
\gamma''_{c_1} - \gamma''_{c_2} &= (\gamma'_{c_1} - c_1\delta) - (\gamma'_{c_2} - c_2\delta) \\
&= (\gamma'_{c_1} - \gamma'_{c_2}) + (c_2 - c_1)\delta
\end{aligned} \tag{3.2}$$

can be made arbitrarily small or large by choosing δ accordingly. Put simply, the data cannot distinguish between the determinants that produce two different cohorts and the chance timing of age and period determinants.[3]

We could impose additional restrictions to force a unique solution, but, unlike the first case, we cannot justify our choice of restriction on the basis that all restrictions are equivalent. The fact

[3]It is important to note that not all cohort comparisons are unidentified. For example, relative differences can be estimated. By equation (3.2),

$$\begin{aligned}
&(\gamma''_{c_1} - \gamma''_{c_2}) - (\gamma''_{c_2} - \gamma''_{c_3}) \\
&\quad = (\gamma'_{c_1} - \gamma'_{c_2}) - (\gamma'_{c_2} - \gamma'_{c_3}) - (c_3 + c_1 - 2c_2)\delta \\
&\quad = (\gamma'_{c_1} - \gamma'_{c_2}) - (\gamma'_{c_2} - \gamma'_{c_3})
\end{aligned}$$

for equally spaced cohorts, $c_3 - c_2 = c_2 - c_1$. But this difference in differences does not serve our purpose because it only establishes the relative influence of events.

that no perfunctory solution identifies the system of equations (3.1) is called the age-period-cohort problem.[4]

We resolve the age-period-cohort problem, not by restricting the cohort parameters that indirectly determine the generational import of political events, but by directly modeling the generational voting process that explains the pattern in Figure 1. We build a dynamic "running tally" model in which the cumulative impression left by political events—in addition to age and period determinants—influence the preferences of voters.[5]

Our running tally consists of two parts. We use the Gallup Organization's long-running presidential approval rating time series, displayed in Figure 3.2, to measure the political events voters experience. We then weight this measure according to the age of the voter at the time the events took place. That is, we replace γ_{p-a} in equation (3.1) with $\sum_{i=0}^{a} w_i \; x_{p-a+i}$ where x_{p-a+i} is the (observed) measurement of the political events when a voter from cohort $c = p - a$ was age i, and w_i is the (unobserved) influence of the events at age i.

Unlike the static age-period-cohort model, our cohort parameter is dynamic and changes with a, and we therefore call it the generational parameter. Our model also accounts for the sex, race, and region of voters, the survey house that collected the data, and choice interactions. We present the technical details of our model in the following section.

One limitation of Gallup's approval ratings is that, despite being one of the longest-running time series available for the study of American political behavior, it is "only" available from 1937 onward. Because this analysis examines the formation of preferences over a voter's entire life

[4]Although called the age-period-cohort problem, the linear relationship that produces this identification problem arises whenever exposure to a phenomenon of interest is not measured directly but approximated from the timing of a life event such as birth year, graduation, employment, or retirement. For example, consider the linear model with explanatory variables age, number of years married, and number of years not married. The problem also extends to interactions, which are not identified since the relationship $c = p - a$ implies $c^2 = p^2 - 2ap - a^2$, $cp = p^2 - ac$, and $ca = pc - a^2$.

[5]In the typical "running tally" model, voters choose their partisan identification by evaluating each party's performance over their lifetime (Fiorina, 1981; Achen, 1992). The simplest versions give each evaluation equal weight regardless of age or recency. Several papers have generalized the model, for example see Gerber and Green (1998). In another example, independent of our work, Bartels and Jackman (2014) combine age-specific weights with period-specific shocks. Both parameters are estimated from the American National Election Study (ANES) cumulative dataset. While these parameters are not underidentified, see footnote 17, (Bartels and Jackman, 2014: pg 14), the model is statistically underpowered; the age-specific weights oscillate between negative and positive, and the uncertainty bounds are large with almost none statistically distinguishable from zero.

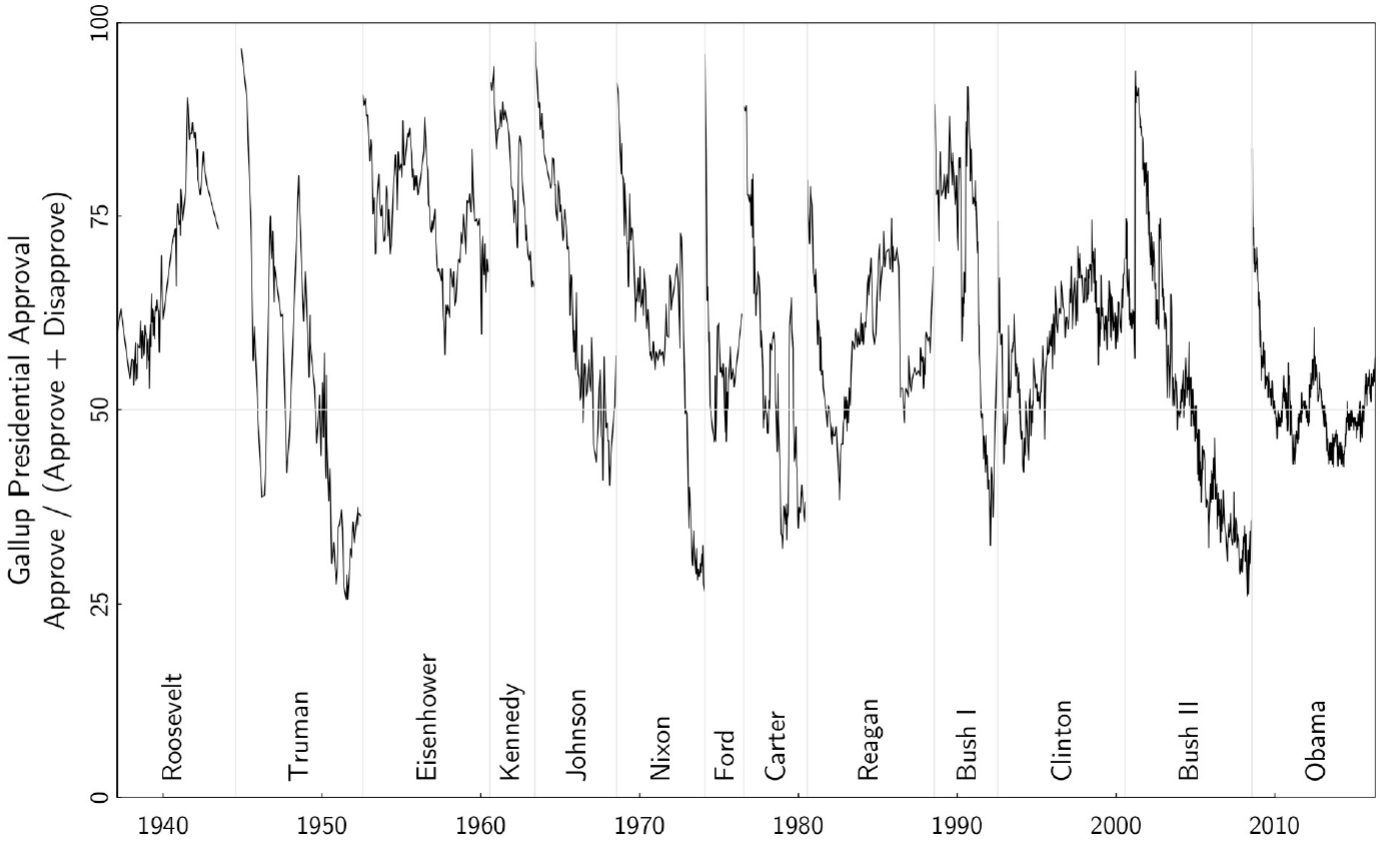


Figure 3.2: The Gallup Organization's presidential approval rating time series, 1937-2016. The data reflects political events that influence voter's partisan preferences.

cycle, and due to the importance of early life political socialization indicated in the literature, we discard observations for which we do not have presidential approval data over the respondents' entire life span. That is, we drop respondents born before 1937, leaving 215,693 responses. The data are plotted by election year and year of birth in Figure 3.3. They cover the 1960-2016 elections and sixty-one birth-year cohorts (1937-1998), with at least 1,000 responses for any individual year.

3.3 Statistical Model

We model the partisan preference of the survey respondents described in the previous section. We index each response by five identifiers: (1) the age of the respondent $a = \{1, 2, \ldots, 70\}$, (2) the year of the response $p = \{1960, 1961, \ldots, 2016\}$, (3) the race/region group of the respondent

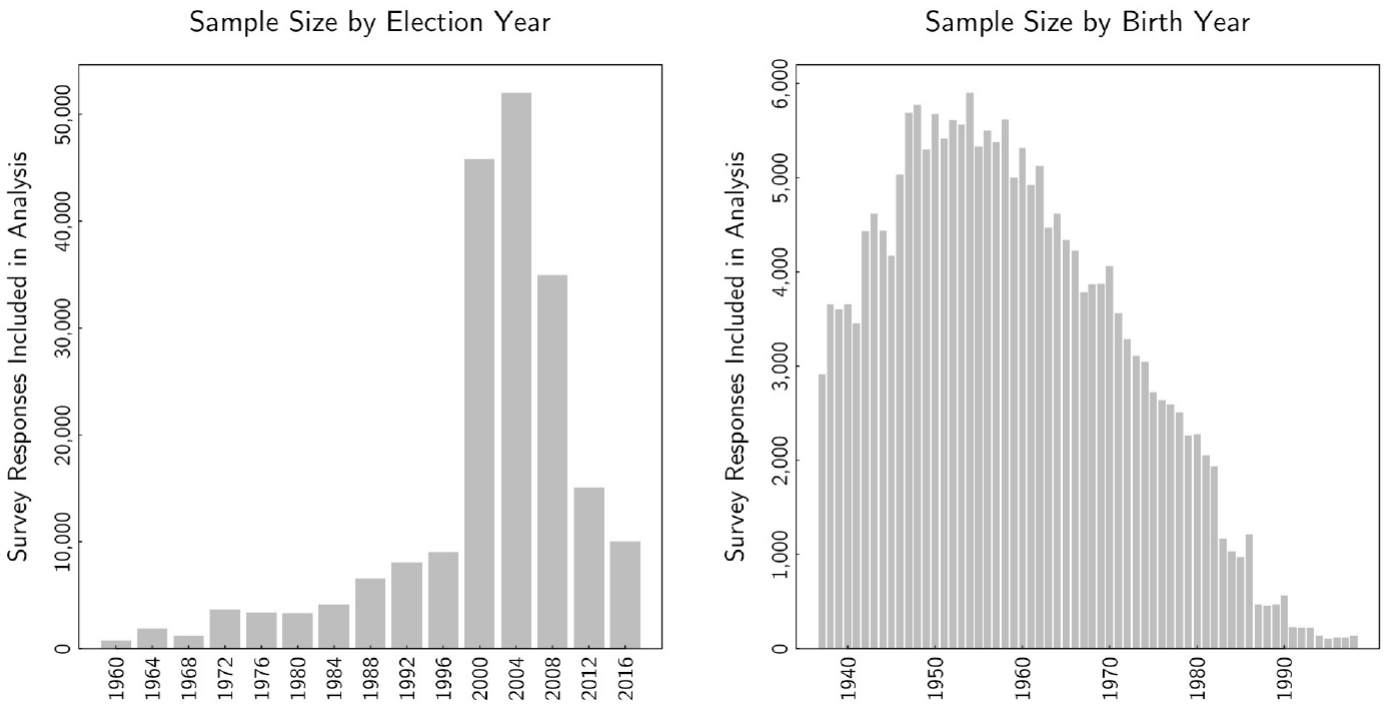


Figure 3.3: After removing survey respondents born before 1937, the analysis includes 215,693 survey respondents in total, here displayed by election year and year of birth. The data, and thus the analysis, have a strong emphasis towards the most recent four elections, and may be interpreted as weighted towards the contemporary political climate. The data encompass generational cohorts defined by their individual birth year from 1937-1998, with at least 1,000 responses for each birth year until 1986.

g = {non-Southern white, Southern white, and minority}, (4) the sex of the respondent s = {female, male}, and (5) the survey house that collected the response h = {Annenberg, Gallup, NES, GQRR, CNN/ORC/Pew}.[6]

These identifiers partition the responses into mutually exclusive cells. For each cell j, we denote the age, period, group, sex, and source of the responses by $a[j]$, $p[j]$, $g[j]$, $s[j]$, and $h[j]$.

Let y_j denote the number of respondents preferring the Republican candidate in cell j, and n_j the number preferring either the Republican or Democratic candidate. Undecided voters are

[6]Our index includes non-election years because voters continuously form preferences even though they only express those preferences in election years. We also group all minority respondents together. Although we prefer to separate African Americans, Hispanic Americans, Asian Americans, etc., the data does not distinguish consistently between minority groups in early years.

discarded. We model

$$y_j \sim \text{Binomial}(n_j, \theta_j),$$

where θ_j is the proportion of Republican presidential support within cell j.

Our primary goal is to quantify the generational effect: the extent to which $\text{logit}(\theta_j)$ is explained by the cumulative impression left by political events. We use the Gallup presidential approval rating time series to instantiate these events as follows.

Let x_t denote the Republican-directional presidential approval rating in year t. The rating is calculated by (1) subtracting 50% from the Gallup presidential approval rating in the year t, and (2) multiplying the difference by -1 if the sitting president was a Democrat. It is positive under two conditions: a Republican president had ratings above 50% or a Democratic president had ratings below 50%. Conversely, it is negative under a popular Democratic or an unpopular Republican president.

Respondents of cell j experience the rating $x_{p[j]-a[j]+i}$ at age i. For example, respondents surveyed at $a = 53$ in $p = 2012$ were born in $c = 2012 - 53 = 1959$ at $i = 0$. In 1960 ($i = 1$), the average approval rating for Republican president Eisenhower was 71%, so $x_{2012-53+1} = (71-50) = +21\%$. In 1961 (i = 2), the presidency flipped to Democratic president Kennedy, who had an average rating of 88%, yielding $x_{2012-53+2} = -1 \times (88 - 50) = -38\%$.[7]

The generational effect is defined as

$$\gamma_j = \Omega_{g[j]} \sum_{i=1}^{a[j]} w_i \; x_{p[j]-a[j]+i}$$

where w_i denotes the age-specific weight of the rating at age i, and Ω_g denotes the scale of the age-specific weights for group g.

In addition to the generational effect, we define a period effect for each group, β_{pg}, and a

[7] We set $x_{p[j]-a[j]+i} = 0$ if $i > a$. We top-censor x at age 70 because few approval ratings are observed above that age.

period and age-weight interaction, $\lambda_g w_a \beta_{pg}$. The interaction accounts for the impressionability of respondents to political events at election time, and λ_g denotes the scale of the interaction for group g, similar to Ω_g in the generational effect. Put together, these define the election effect,

$$\begin{aligned} B_j &= \beta_{p[j]g[j]} + \lambda_{g[j]} w_{a[j]} \beta_{p[j]g[j]} \\ &= \left(1 + \lambda_{g[j]} w_{a[j]}\right) \beta_{p[j]g[j]}. \end{aligned}$$

We also define an age effect α_a, a house effect η_h, and the following linear-in-period sex effect

$$\delta_{sp} = \begin{cases} -\frac{1}{2}(\delta_0 + \delta_1\ p) \text{ if female} \\ \\ \ \ \frac{1}{2}(\delta_0 + \delta_1\ p) \text{ if male} \end{cases}$$

The log-odds is the sum of the effects as in equation (3.1)

$$\text{logit}(\theta_j) = \alpha_{a[j]} + B_j + \gamma_j + \eta_{h[j]} + \delta_{s[j]p[j]}$$

We complete the model by smoothing the age weights,

$$w_i \sim \text{Normal}\left(w_{i-1}, 0.005\right),$$

and specifying normal distributions for α, β, and η with mean zero and standard deviations σ_α, σ_β, and σ_η. The scale parameters, λ, and Ω are constrained to be positive.

We fit the model using Stan (Stan Development Team, 2013) and R (R Core Team, 2012). Stan runs a No U-Turn (NUTS) sampler (Hoffman and Gelman, 2014), an extension to Hamiltonian Monte Carlo (HMC) sampling, which is itself a form of Markov Chain Monte Carlo. We generate 4 chains for 5000 iterations. The final 2500 iterations of each chain converge as indicated by post-modeling diagnostics such as Gelman-Rubin $\hat{R}$ (Gelman et al., 2004). We ensure satisfactory posterior predictive model performance (Gelman et al., 2004) before using sample means (for

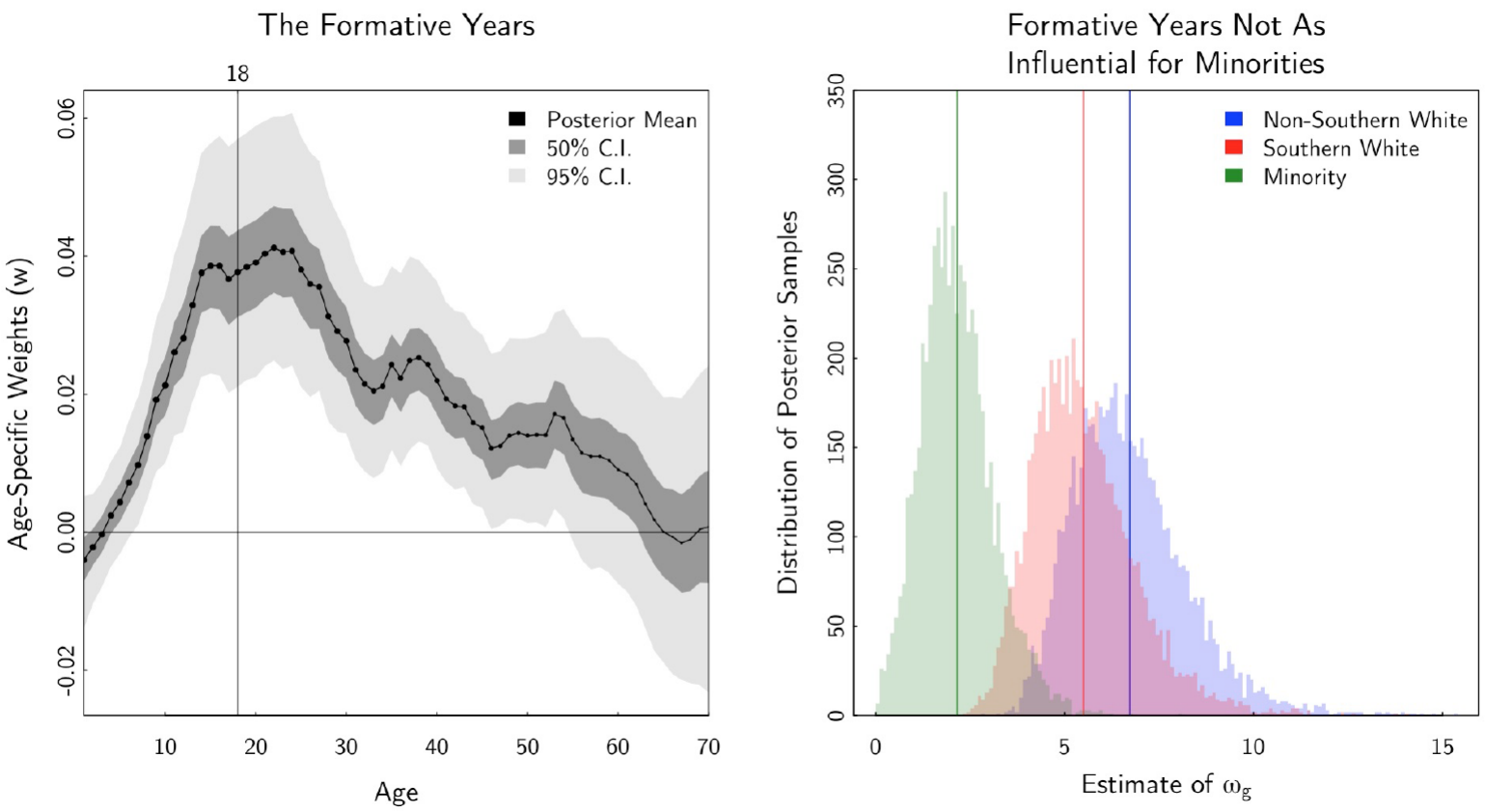


Figure 3.4: Estimates of the generational effect. (L) We find the 14-24 age range is most import for the formation of long-term presidential voting preferences. Political events before 14 have little impact. After 24, the age weights decreases. (R) These weights, and the political socialization process implied by them, are substantially more important for non-Hispanic whites than for minorities as a whole.

estimates) and sample quantiles (for credible intervals) in the following section.

3.4 Model Results

We interpret the fitted model with a series of graphs. In Figure 3.4 we examine the generational effect, Figure 3.5 the election effect, and Figure 3.6 the amount of variation explained by the model.

3.4.1 Generational Effect

The left side of Figure 3.4 shows the estimated age-specific weights, w_i, along with 50% and 95% credible intervals. The weights quantify the formative years of political socialization with

precision unprecedented in the literature: At a very young age, political events leave virtually no impression—the weight at age 1, w_1, is essentially zero. The weights then increase steadily, peaking around 14-24 and gradually decreasing thereafter. At the height of their influence, around the age of 18, events are nearly three times as meaningful as those later in life.

The importance of adolescence and early adulthood in the socialization process is supported by an enormous literature. For example, Erikson, MacKuen and Stimson (2002) also find political events have the largest impact at age 18-19 before declining. Yet despite the decline—and the fact that a generation's preferred party is all but locked-in by 40—we find political events continue to influence voter preferences. The age-weights only return to zero around age 60.

No children were interviewed, leaving one to perhaps wonder how the model can determine the impressions left by childhood events. To understand how, consider a year in a respondent's childhood, say the year the respondent was 14 years old. We know the age the respondent was interviewed and therefore the year in which the respondent was 14. We also know the political events of that year, as reflected in the presidential approval rating. Our model uses this data to "back out" the size of the impression left by the events the respondent experienced at age 14.

For example, a 45-year old who was interviewed in 2012 would have been 14 in 1981. President Reagan had an average approval of 66% in 1981. The fitted model estimates the weight 14-year-olds must give events instantiated by a 66% approval rating in order to explain the preference of voters 31 years later, in 2012.

The right side of Figure 3.4 shows Ω_g, the amount the age-weights are scaled to produce the generational effect for each group. The estimated generational effect is found to be over twice as large for non-Hispanic whites as for minorities as a whole, suggesting considerable differences in the political socialization process.

The difference could reflect the fact that African Americans are consistent Democratic voters, and Hispanic or Asian American immigrants may not have been in the United States during peak socialization to experience the political events captured by the Gallup series. In addition, the political participation of naturalized citizens has been shown to vary depending on their community

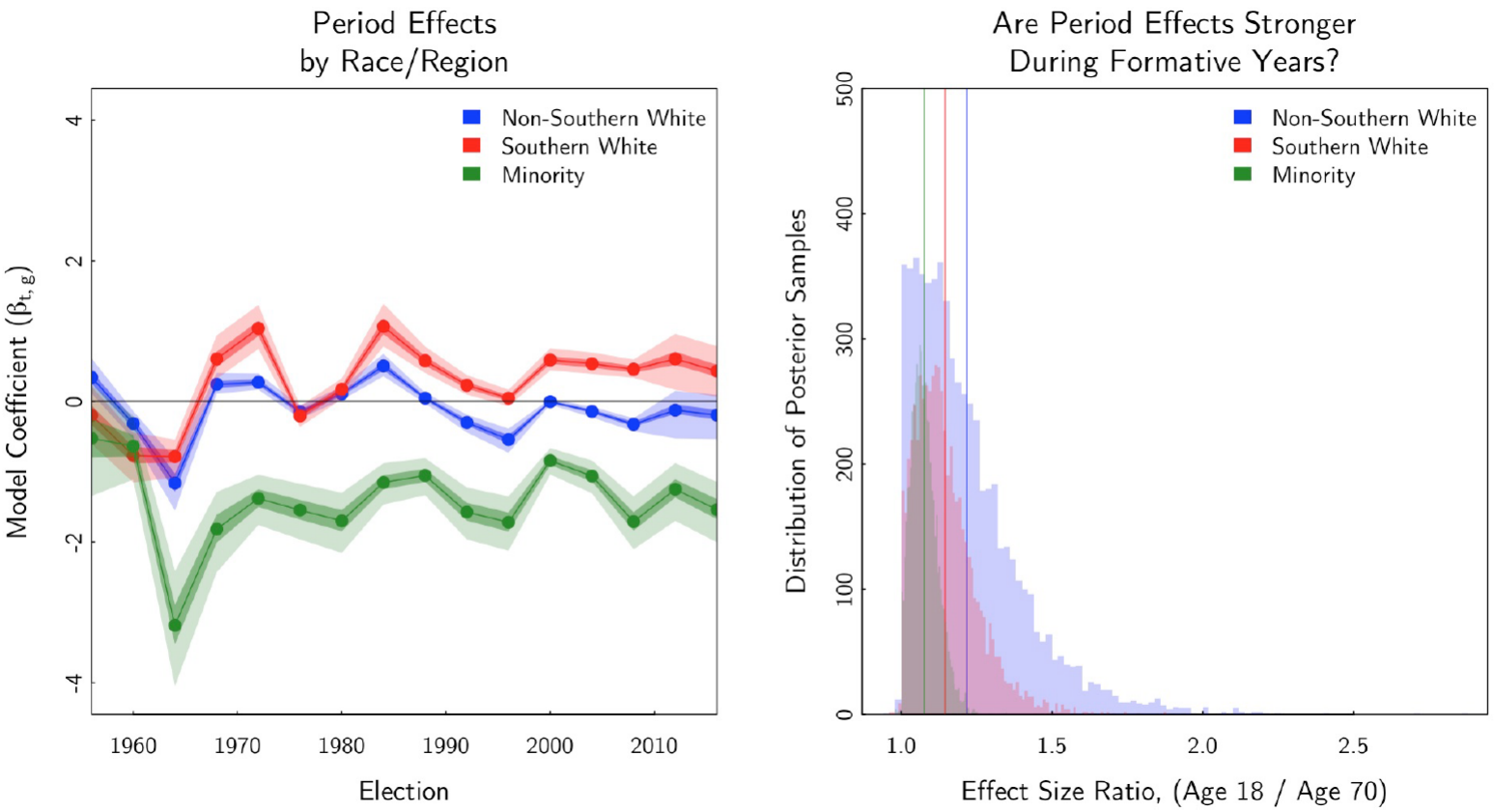


Figure 3.5: (L) Estimates of the period effect. Minorities are consistently more likely to vote for Democratic presidents, and Southern whites have steadily trended pro-Republican over the past 50 years. (R) Election effects are similar between young and old minority voters and in the South. The evidence is inconclusive for non-Southern whites.

(Pantoja, Ramirez and Segura, 2001).

Whatever the reason, the political socialization process observed with white voters is less evident with minority voters. A more rigorous investigation would separate minority subgroups, which, unfortunately, we are unable to do from the data.

3.4.2 Election Effect

The left side of Figure 3.5 shows a time series plot of the estimated period effects, β_{pg}, along with 50% and 95% credible intervals. The effects vary by race/region group, reflecting 50 years of political polarization. Minorities are consistently more likely to vote Democratic, and Southern whites, Republican.

The right side shows the interaction between period and age-weight, $\lambda_g w_a \beta_{pg}$. The interaction allows us to determine whether the election effects are more pronounced during the formative years shown in Figure 3.4. However, interactions are difficult to interpret directly (Gelman and Hill, 2007). Instead we examine the following ratio, where the numerator is the first factor of B_j, $\left(1 + \lambda_g w_{18}\right)$, for an 18-year old voter (one of the most impressionable ages as determined by the peak of the age-weight curve), and the denominator is the corresponding factor, $\left(1 + \lambda_g w_{70}\right)$, for a 70-year old voter (one of the least impressionable ages as determined by the nadir).

We do not find clear evidence that the election effect varies according to the age-weights. For Southern whites and minorities, the mode of the ratio gathers at the boundary 1.0, implying no difference. For non-Southern whites, the effect has substantial mass between 1.0 to 1.4. That is, the model indicates that election effect for non-Southern whites are between 0% and 40% greater for young voters than old voters.

3.4.3 Explanatory Power

Figure **??** compares the sample R^2 of the fitted model against a simpler model with only period and race/region effects. The comparison is made overall and within each group. We use R^2 because of its simplicity and near-universal recognition among researchers. However, we note that R^2 is one of many possible measures of explanatory power, with other choices typically trading between reliability and interpretability. We weight R^2 by the size of the j cells.

Overall, the model explains 91% of the variance in the data. Much of this variation, 89%, is also explained by the simpler model. However, that merely reflects the enormous difference in voting preferences between groups and across elections.

Within race/region groups, our model explains considerably more variation—although the improvement is not equal across all groups. For non-Southern whites, the fit increases nearly twenty percentage points, from 51 to 69%. For Southern whites, it improves a modest seven, from 47 to 54%. For minorities, there is little difference.

We conclude that our model accounts for a substantial portion of the variation in presidential

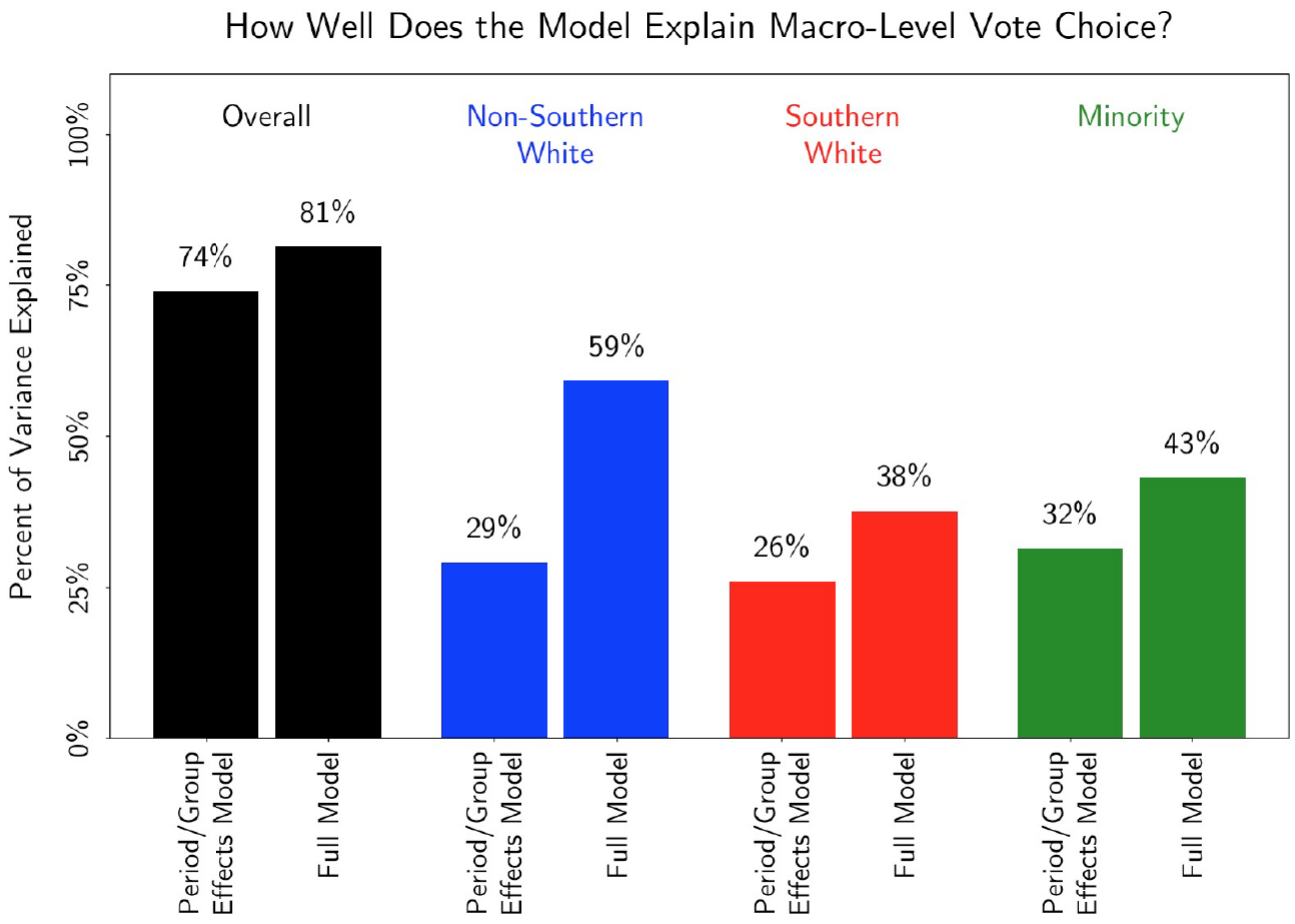


Figure 3.6: The model accounts for 91% of the macro-level variance in voting trends over the past half century, more than the simpler model incorporating only period/group effects. The model fits considerably better within race/region groups, particularly among non-Southern whites.

voting over the last half century. It is a demonstrable improvement over a model with only period and race/region, suggesting a single defining political event is less important in the formation of voter preferences than the prolonged impression left by a lifetime of events.

3.5 Generations of Presidential Voting

We demonstrate how the fitted model aids the study of elections. We provide a narrative of the presidential approval time series, emphasizing how political events, generally associated with the presiding administration, formed the preferences of five distinct generations: New Deal Democrats,

Eisenhower Republicans, 1960s Liberals, Reagan Conservatives, and Millennials. Each generation is epitomized by the birth years in which partisan preferences were the strongest: 1929 (pro-Democrat), 1941 (Republican), 1952 (Democrat), 1968 (Republican), and 1985 (Democrat).

These labels are for convenience and should not be taken too literally; generations are not the byproduct of a single year, but rather the result of sustained periods of partisan influence. The following figures plot model results for white voters. Under our model, the generational effects for minority groups are proportional. However, as discussed in the previous section, the evidence of generational voting is weaker for minority voters, and thus any generalizations made from the model are weaker.

3.5.1 New Deal Democrats

Since the approval ratings begin in 1937, the model is limited in its description of the New Deal Democrats, who are epitomized by the 1929 birth year. Nevertheless, we use the general principles learned from the model in other generations to piece together how political events influenced this group.

New Deal Democrats include a large and diverse group, dominated by a single towering figure: Franklin Delano Roosevelt. As president, FDR guided the country through the Great Depression and World War II, and, with the New Deal, laid the foundation for the modern American welfare state. He was enormously popular, winning four elections and serving for twelve years, more than any president in American history.

For the first half of this group, voters born between 1910 and 1920, their peak formative years were spent during the Great Depression and World War II. They experienced Republican president Hoover's inability to help a struggling United States, followed by economic recovery and the greatest war in world history—both under Democrat FDR.

To these voters, the United States became a leader of the free world under Roosevelt's watch. This left a strong impression that remains to the present day. Recall Figure 3.1, where these now elderly voters continue to have comparatively pro-Democratic preferences in the 2000-2016

elections.

For the second half of this group, voters born after 1930, their exposure to FDR was limited. Their formative years occurred after the country recovered from the Depression, and, for many, after World War II as well. Though they lived through the tail end of his presidency, during which FDR remained enormously popular, their peak years were spent with Truman at the helm. Truman had mixed and limited popularity over his two terms, ending his presidency at 36% approval. As a result, these voters' long-term voting preferences are mixed.

3.5.2 Eisenhower Republicans

The approval ratings are available for the entire life span of the remaining generations. As a result, we can directly interpret the fitted model, which we do with the aid of the two panels in Figure 3.7.

The top panel shows the approval ratings, highlighted to emphasize the generational import of each time period: The ratings are colored red to blue, with red reflecting pro-Republican approval ratings, blue pro-Democratic, and shades of grey in between. The width and darkness of the line correspond to the estimated, age-specific weights w. Thus, the darkest and widest lines emphasize the peak formative years, when the events represented by the approval ratings were most influential.

The bottom panel shows the cumulative weighted approval ratings, which define the generational effect of the political events experienced up until the age indicated on the x-axis. The series starts at the grey line (age 0).

With this Figure, we examine the presidential preferences of the Eisenhower Republican generation, epitomized by voters born in 1941. These voters were too young to remember FDR's many accomplishments, instead entering their years of peak socialization in anti-Democratic and pro-Republican times; their earliest impressions were formed in 1951 when Truman, who had barely won reelection three years earlier, sent American troops into Korea. After the unconditional victory of World War II, Americans were unaccustomed to the apparent stalemate in Korea, and Truman's popularity plummeted.

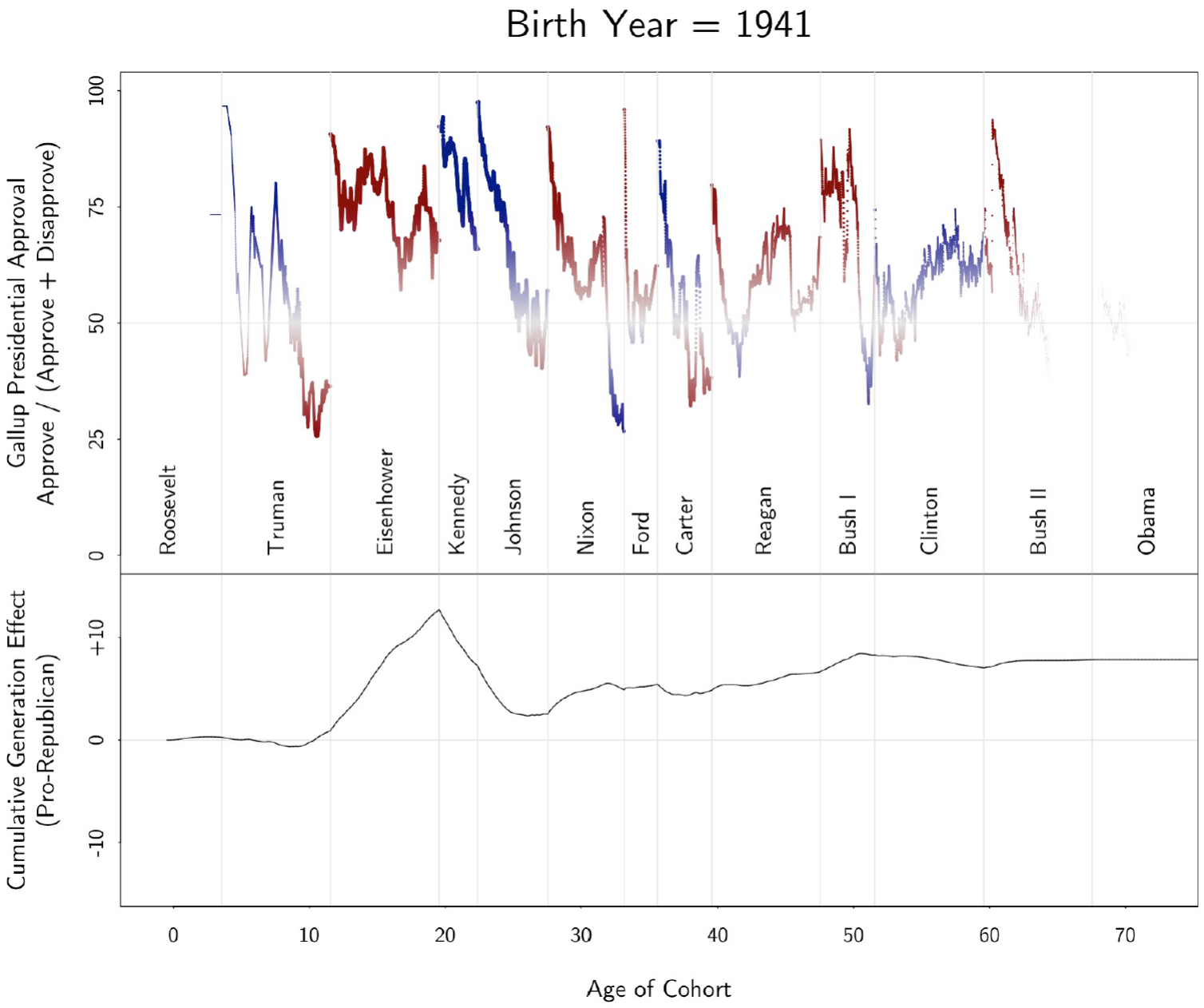


Figure 3.7: Presidential approval, and the cumulative generational effects, for Eisenhower Republicans born in 1941. The graph emphasizes peak years of socialization, according to age weights found by the model. Blue indicates pro-Democratic years, red for pro-Republican, grey in between. This generation missed most of the FDR years and was socialized through 10 straight pro-Republican years (Truman and Eisenhower). Their partisan voting tendencies were drawn back towards the neutral grey line by the pro-Democratic 1960s, and they reached a rough equilibrium by the end of the Nixon presidency.

When Eisenhower assumed office in 1953, his approval was a near unanimous 91%. While most presidential terms begin with high ratings (Erikson, MacKuen and Stimson, 2002), Eisenhower remained popular over his entire presidency. The heroic World War II general promised to end the Korean War during his campaign and quickly did so. Although he did not end the larger Cold War, as he desired, international conflicts were relatively minor over his tenure. The 1950s were a time of relative peace, prosperity, and progress.

The most prominent dip in Eisenhower's popularity came around 1957-1958. The country was in a recession, the Soviet Union had launched Sputnik and appeared to be winning the space race, and Eisenhower was forced to send federal troops to Little Rock to enforce a federal desegregation policy, indicative of national tensions over civil rights. But the dip was short lived, reaching a bottom point of 57% in March of 1958 and rebounding quickly back to the 70-80% range. Eisenhower left office with a 69% approval rating.

The Eisenhower Republican generation experienced 10 straight years of pro-Republican presidential evaluations, much within the peak years of socialization. The impact of this period on their long-term presidential voting preferences is apparent in the bottom panel of Figure 3.7. The curve ascends steeply, peaking at the end of the Eisenhower administration. Their preferences were then moderated by the Kennedy and Johnson years, reaching equilibrium by the end of the Nixon presidency.

3.5.3 1960s Liberals

The generation of the 1960s Liberals is epitomized by voters born in 1952. As can be seen in Figure 3.8, these voters came of age during the Kennedy, Johnson, and Nixon years.

Kennedy, like Eisenhower, began his presidency with immense popularity, at a 92% approval rating. The political mood of the country was at a liberal high-point (Stimson, 1991), and there was widespread optimism about the role of government. That optimism was reflected in Kennedy's bold "New Frontier" agenda, in which he committed to sending a man to the moon by the end of the decade, and in his sweeping initiatives to combat poverty, expand medical care, increase

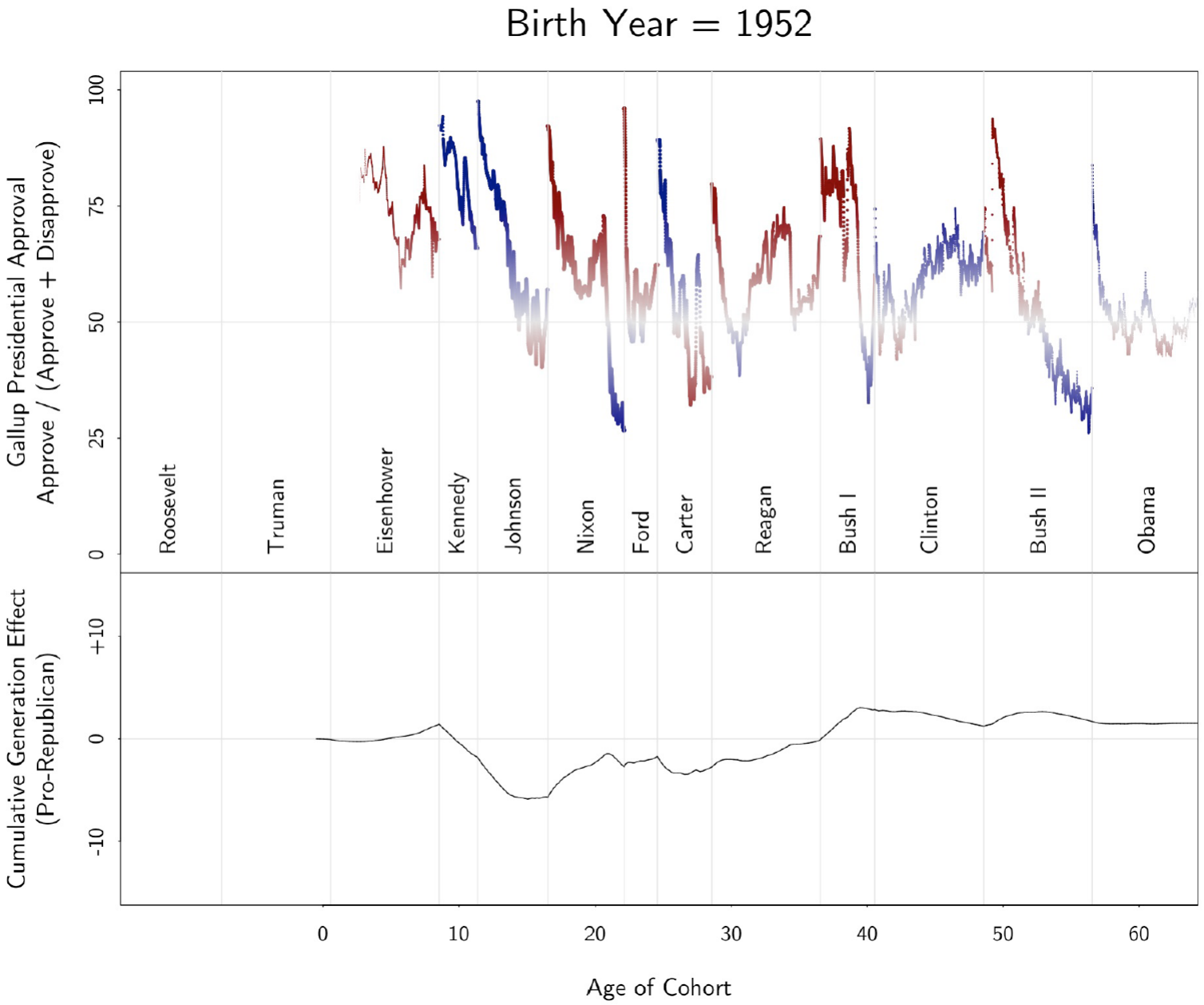


Figure 3.8: The generation we refer to as 1960s Liberals are best epitomized by those born in 1952, whose presidential political events are emphasized here. Too young to be highly influenced by the Eisenhower years, they experienced an intense period of pro-Democratic sentiment during the 1960s. After 1968, however, roughly 25 years of near-consistent pro-Republican events neutralized their presidential voting preferences.

educational aid, and progress the cause of civil rights.

Kennedy succeeded in passing a number of initiatives, but his presidency was overshadowed by a series of tumultuous, foreign policy events. He was at the helm during the failed Bay of

Pigs invasion and the Cuban Missile Crisis. Though Kennedy averted war, many questioned his strength as a leader in the face of the Soviet Union. His approval ratings declined steadily over his three-year presidency with only a short positive burst following the Cuban Missile Crisis.

Right before Kennedy's assassination in November 1963, his approval ratings bottomed out at 66%. The assassination, however, resulted in an enormous popularity spike—the second pro-Democratic spike in less than a three-year time span. Perhaps for this reason, the 1960s Liberals remember Kennedy more for his charisma, his beautiful and sophisticated family, and his optimistic vision of the future.

When Johnson assumed the presidency, he was the second Democratic president to start in the 90% range, this time at 97% approval, the highest in the series. In the name of the fallen president, and as the quintessential Washington insider, Johnson promoted his vision of a "Great Society". He signed foundational legislation, such as the Civil Rights Act of 1964 and the Voting Rights Act of 1965. He established landmark programs, such as Medicare, Medicaid, food stamps, and Project Head Start. He expanded student loans and increased federal funding, including universities and the nongovernmental Corporation for Public Broadcasting. He protected the environment, regulating pollution through the Water Quality Act and Air Quality Act and establishing the national wilderness, rivers, and trails systems.

Johnson enjoyed immense popularity for an extended period of time, as reflected in his high approval ratings and landslide election victory over Barry Goldwater in 1964. These dramatic events socialized the 1960s Liberals generation unusually early, with a steep pro-Democratic shift from 1961-1966 (corresponding roughly to 9-14 years old)—well before the peak years of socialization (14-24). This shift was strong enough to influence their preferences for decades to come.

The Vietnam War and increasing racial and social tension in the late 1960s, however, marred Johnson's presidency and legacy. By 1967, his approval ratings had fallen, and in 1968 the once powerful president decided against running for reelection.

In that election, Nixon played on the generation gap, speaking to the "silent majority" and explicitly denouncing the political concerns of the 1960s Liberals. Four years later, after the national

voting age was lowered to 18, the generation gap reached its largest. White voters under the age of 25 (first-time voters in 1972) supported Nixon at 53%, compared to 70% for white voters 25 or older. This 17-point gap is the largest in the dataset, not exceeding 9 points in any other election.

The continuation of the Vietnam war under Nixon, when this generation reached draft age, and ultimately the Watergate scandal, helped keep the 1960s Liberals pro-Democratic until their 40s. Yet, the cumulative curve of Figure 3.8 shows that the 1960s Liberals never returned to their 1968 pro-Democratic highpoint.

3.5.4 Reagan Conservatives

The generation of the Reagan Conservatives is epitomized, ironically, by the 1968 birth cohort—the year the 1960s Liberals hit their pro-Democratic highpoint. These voters were not alive for the popular, pro-Democratic Kennedy and Johnson years, and the Nixon and Ford presidencies had little impact, as shown in Figure 3.9. Their political socialization started with president Carter.

Carter was initially popular, but his ratings plummeted as adverse political events overtook his presidency. By the time he left office, an energy crisis, stagflation, and the Iran hostage crisis, among other events, left him with approval ratings in the 30-40% range.

Reagan captivated this generation with his optimistic vision of America as a shining city on a hill. Though his early years were defined by a lack of economic recovery and Republican defeats in the 1982 midterm elections, Reagan's popularity dipped below 50% for only a short period. The recovery hit full swing shortly thereafter, and Reagan, whose campaign famously declared that it was "Morning in America" again, was reelected in a landslide.

The Reagan "Revolution" had a powerful impact on this generation, who, at 16 years old, were reaching their peak years of socialization. Despite the Iran-Contra scandal and ballooning deficits near the end of his second term, Reagan ended his presidency with a 68% approval rating.

Figure 3.9 suggests President Bush I's presidency extended pro-Republican sentiment in ways that are perhaps underestimated in the collective public memory. From a foreign policy perspective, Bush was enormously successful. The fall of the Berlin Wall and the end of the Cold War came

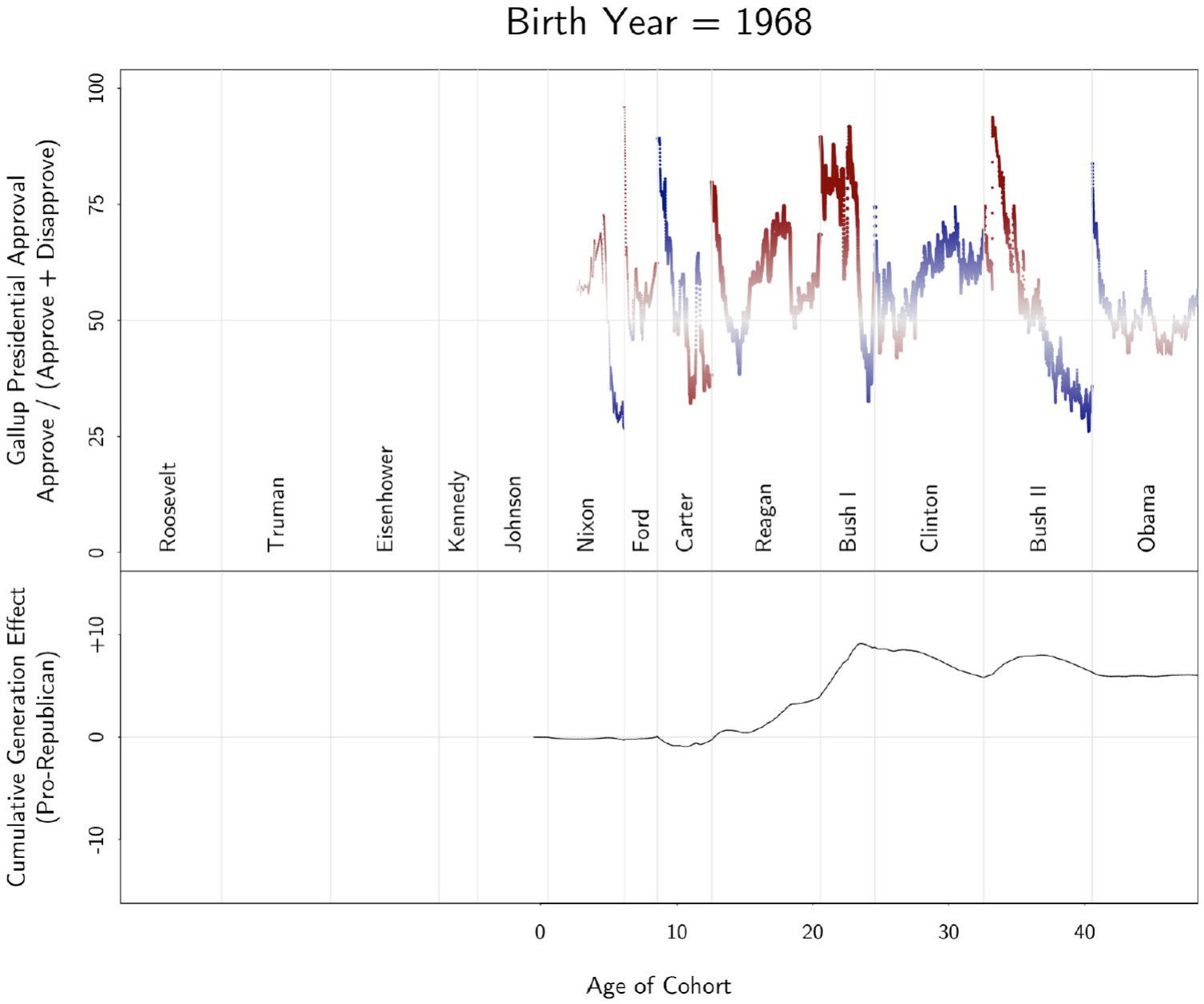


Figure 3.9: The Approval series as seen by the generation we call Reagan Conservatives, best epitomized by those born in 1968. This generation missed the Kennedy and Johnson years entirely, and their peak socialization fell under the popular Republican presidents Reagan and Bush I. By the time the Democratic president Clinton reached his peak popularity in the late 1990s, they were already roughly 30 years old.

under his watch, and Operation Desert Storm demonstrated the power of American leadership in the post-Cold War era. As a result, Bush's ratings rarely fell below 80% for over 2 years, only dipping below 50% right near the end of his term.

Ultimately, economic problems at home doomed his presidency. The Clinton campaign mocked, "It's the Economy, Stupid," winning the presidency in 1992 and ending over a decade of nearly continuous pro-Republican sentiment. The pro-Democratic Clinton years curbed this generation's long term preferences, but the Reagan Conservatives were roughly 30 years old by the time Clinton reached the height of his popularity in the late 1990s, past the peak age of socialization.

3.5.5 Millennials

For the last generation—the Millennials, epitomized by the 1985 birth year—we only observe 31 years of preferences. Nevertheless, the political events that have shaped their voting preferences are clear in Figure 3.10. If past generations are any guide, these impressions will continue to influence their preferences for the rest of their lives.

The Millennials did not experience the uncertainty of the Cold War or the foreign policy successes of the Reagan and Bush I administrations. The Clinton years were the first to influence their voting patterns.

Clinton's biggest political defeat, the Republicans' Contract with America, took place in 1994 when Millennials were around 9 years old. But as Millenials entered their peak socialization years, America had become the globe's lone superpower, and the country was experiencing tremendous economic growth. Clinton enjoyed positive approval ratings for roughly four straight years, and despite his impeachment, ended his presidency at 67% approval.

Republican Bush II took office in 2001, beginning one of the most turbulent presidencies in American history. After an initial popularity of 94% following the 9/11 terrorist attacks, his approval declined precipitously. His administration undertook costly and unpopular wars in two countries. Some supported the president's vision of America as a crusader for democracy, but many grew to oppose the war in Iraq, in particular. On the domestic front, Bush II's most notable accomplishment was his 2001 tax cuts, which created massive federal deficits. He ended his presidency amid the largest financial crisis since the Great Depression. Eleventh hour legislation, the Troubled Asset Relief Program (TARP), was unable to avert the crisis.

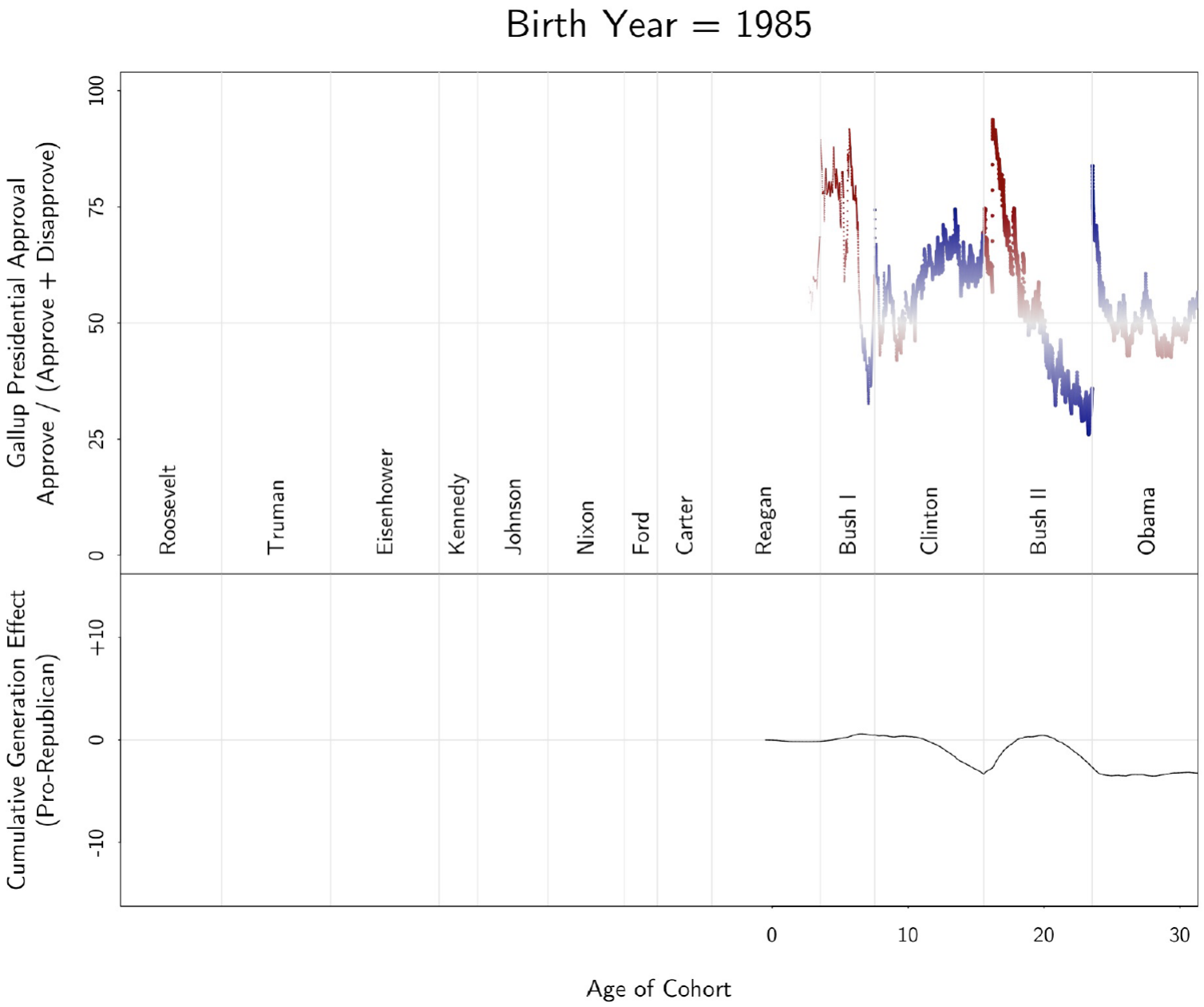


Figure 3.10: The Approval series as seen by the last generation, the Millennials. Their experience had only lasted 31 years by the 2016 election, but the model indicates that these years should remain highly influential over the rest of their lives. Their formative years have been primarily characterized by the popular Democratic president Clinton and the unpopular Republican Bush II, resulting in their relatively strong pro-Democratic sentiment.

These events are reflected in his approval ratings. Bush II fell below 50% in 2004, barely winning reelection that year. He fell below 50% again in 2005, and his ratings stayed in negative territory for the remainder of his presidency—almost an entire four years, by far the longest stretch

in the series. His approval hit its low point of 26% in October of 2008, in the midst of the financial crisis.

We conclude with Democratic president Barack Obama, who ends the series. Obama, like the other presidents, began with a high 76% rating—less than the 90% levels of earlier presidents, but in line with the start of the Clinton and Bush II presidencies. His popularity quickly declined, dropping to 50% in February 2010 and hovering around 50% for the remainder of his presidency.

The Millennials' preferences thus far reflect the popular Democrat Clinton and the deeply unpopular Republican Bush II. But consider the youngest voters, born in 1998 and 18 years old during the 2016 election. They were barely alive during Clinton's presidency and were only ten years at Obama's election, essentially missing both of these consequential time periods. Instead, they were socialized during the relatively even Obama years. Referring back to Figure 3.1, we can see that they trended Republican compared to their slightly older counterparts. However, their ultimate life-long voting patterns remain to be seen.

3.5.6 The Changing White Electorate

We examine the white electorate as the changing composition of five generations. Figure 3.11 combines each of the generational curves from the previous sections on a single graph. However, instead of plotting each generation by its representative birth cohort, we broaden each generation to the scale of decades. The narrative remains the same though; narrow definitions of generations are not indicated by the data.

The changing width of each curve reflects the proportion of the electorate that each generation contributes at any given time: At the start of the series, the oldest generation comprises the entire electorate. As time marches on, they become a smaller and smaller portion, and by 2016 all five generations are represented.[8] The overall electorate is shown in black.

Before the 1960s, partisan preferences moved back and forth between Republican and Demo-

[8]Instead of plotting each generation's full curve from age zero onward, we only plot the curves from their first election onward. We have also included New Deal Democrats and older voters in this graph, even though they were not included in the statistical model.

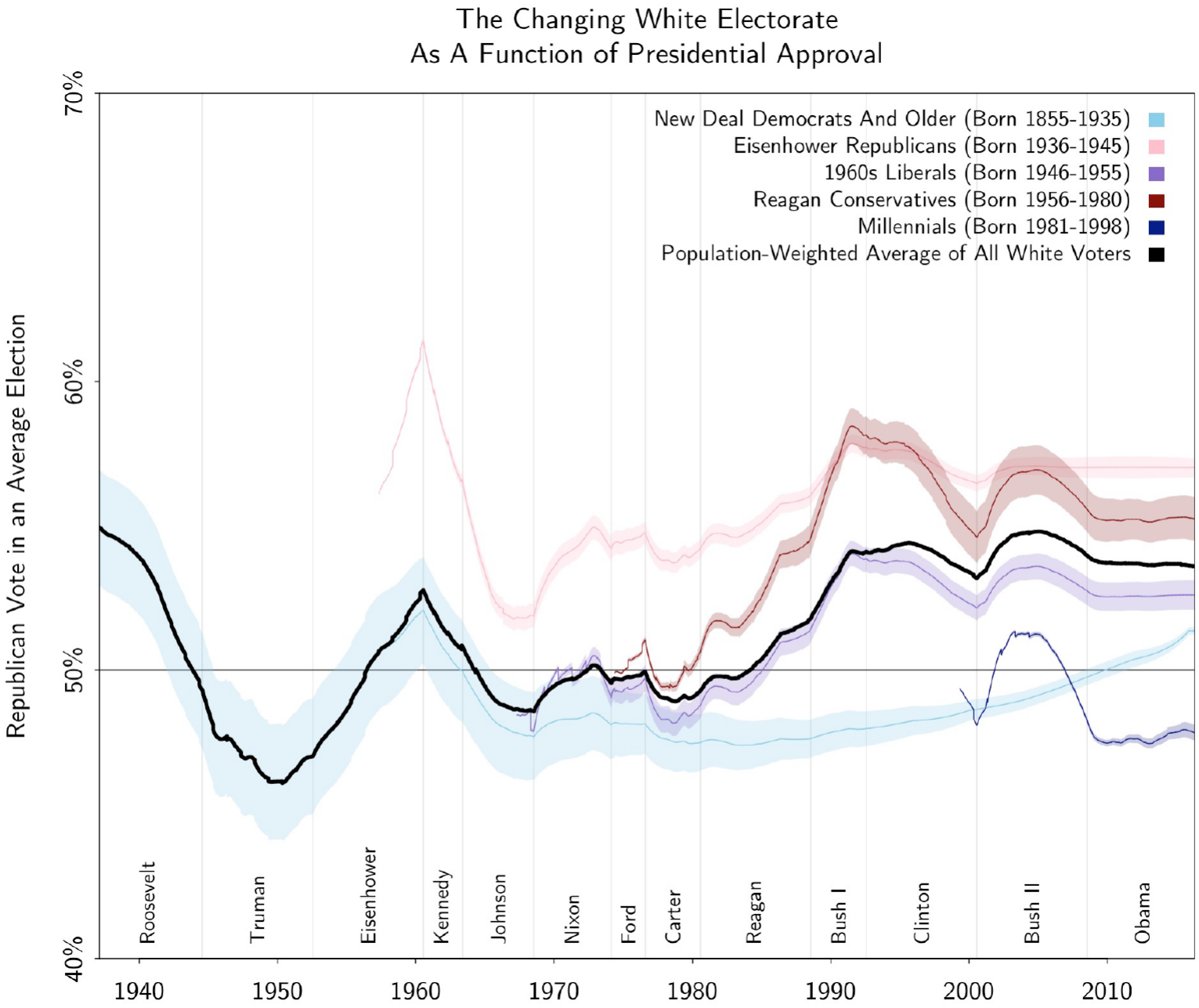


Figure 3.11: The cumulative preferences of each generation is shown, along with the weighted summation of the full white electorate. The generations are loosely defined so that the entire electorate can be plotted at once. The width of each curve indicates the proportion of the white electorate that each generation reflects at any given time. The model—in this graph reflecting only the approval time series and the age weights—explains much of the voting tendencies of the white electorate over time.

crat, in response to the popularity of Roosevelt, Truman, and Eisenhower. The Kennedy/Johnson years pulled the electorate Democratic over the course of the 1960s. Then followed a long period

of Republican ascendancy—slightly trending upward through the Nixon and Ford years, slowed in part by the entry of the 1960s Liberals.

The Reagan administration moved all generations upward. The New Deal Democrats were too old for a large change, but the remaining generations, especially the Reagan Conservatives, moved dramatically, with the black curve crossing the 50% boundary line in late 1984. The Reagan Conservatives were then moderated by the Clinton presidency, but not enough to shift the electorate as a whole. A Bush II spike followed 9/11, only to drop under his slow and steady decline.

Our model is relatively simple, but it explains a substantial amount of the voting character of the electorate. Figure 3.11 is driven entirely by approval ratings and age weights. It recreates the familiar "parallel lines" of public opinion, in which different groups exhibit respond proportionally to political events (Page and Shapiro, 1992). Though the fluctuations in partisan preference may seem small—the black curve spans only 10 percentage points altogether—they are large enough to determine an election and thus the political direction of the United States.

3.6 Discussion

We build a generational model of American presidential voting. In our model, the political events voters experience leave lasting impressions that inform their partisan preferences. The size of the impression depends on the age of the voter at the time the event took place.

We demonstrate the fitted model is both predictive—explaining a substantial portion of voting trends over the last half-century—and interpretable—dividing voters into five meaningful generations. The predictivity and interpretability remain even after controlling for changes in cohort composition, such as age, race, region, and sex. We conclude the data strongly support generational voting.

Our analysis is at the macro-level in that we do not study specific voters or political events in detail, but rather the broad strokes of events across the electorate. We believe the quantification of macro-level trends is an important contribution in its own right; we illustrate in detail how the fitted model aids the study of elections. Nevertheless, we believe our work has two important

implications for micro-level analysis.

First, while many of the events we identify with our model have been suggested in the literature, the length and size of our dataset allow us to assess their importance with unprecedented precision. We find some events were so impactful, they left an impression on individuals we would not typically consider impressionable. For example, the events during the Kennedy and Johnson administrations defined the 1960s Liberals while they were still children. Researchers can further study these events using detailed surveys or quasi-experiments.

Second, the age-period-cohort problem arises in micro-level as well as macro-level analyses. It occurs whenever exposure to the phenomenon of interest is not measured directly but backed out from the timing of a life event such as birth year, graduation, employment, or retirement. Moreover, it occurs regardless of whether additional individual-level covariates are included or age, period, and cohort are treated as continuous measurements.

We believe our approach is an effective solution to the age-period-cohort problem, which continues to challenge researchers despite its discovery nearly a century ago. For example, consider a variant of the problem, which puzzled pollsters after the 2012 presidential election: In 2008, 55% of white voters aged 18-29 voted for then-Democratic candidate Obama. In 2012, that advantage flipped to 54% in favor of Republican candidate Romney. Why did this happen? Was this a temporary shift in the preferences of young voters? Or would young white voters support the Republican candidate in 2016?

Our model provides a clear answer. Heading into 2008, young, impressionable voters had only experienced the popular Clinton and unpopular Bush II years. The winds were in Obama's favor. By 2012, however, the years of poor Bush II performance that had swayed the young voters of 2008 were replaced by the more recent, mediocre ratings of Obama himself. The shift of young, white voters to the Republican Party was not temporary. In fact, our model predicted it in 2012, and the 2016 election confirmed this trend.

We could paint these events in a positive light for the Democrats. The year 2008 was special, similar to 1972, in that a strongly pro-Democratic cohort entered the electorate following a

deeply unpopular Republican president. The impression left by the Clinton and Bush II years may be strong enough to keep an entire generation of voters pro-Democratic throughout their entire lifetime.

We conclude on this note. When we think about generations of presidential voting, it is important not to think about a single defining political event. Rather, generations are formed through prolonged periods of presidential excellence: FDR and the New Deal, Eisenhower, Kennedy and Johnson's Great Society, the Reagan/Bush conservative revolution, and the Clinton years. Each is characterized by long periods of high approval ratings. Each defined a generation by slowly and steadily winning over the electorate's most impressionable voters.

3.7 References

Achen, Christopher H. 1992. "Social Psychology, Demographic Variables, and Linear Regression: Breaking the Iron Triangle in Voting Research." Political Behavior, 14(3):195–211.

Bartels, Larry M and Simon Jackman. 2014. "A Generational Model of Political Learning." Electoral Studies, 33:7-18.

Beck, Nathaniel. 1991. "Comparing Dynamic Specifications: The Case of Presidential Approval." Political Analysis, 3:51–87.

Beck, Paul Allen and M Kent Jennings. 1979. "Political Periods and Political Participation." American Political Science Review, 73(3):737–750.

Beck, Paul Allen and M Kent Jennings. 1982. "Pathways to Participation." American Political Science Review, 76(1):94–108.

Beck, Paul Allen and M. Kent Jennings. 1991. "Family Traditions, Political Periods, and the Development of Partisan Orientations." The Journal of Politics, 53(3):742–763.

Burnham, Walter Dean. 1970. "Critical Elections and the Mainsprings of American Politics." Norton.

Campbell, Angus, Philip E. Converse, Warren E. Miller and Donald E. Stokes. 1964. "The American Voter." New York: Wiley.

Converse, Phillip E. 1976. "The Dynamics of Party Support: Cohort-Analyzing Party Identification." California: SAGE Publications.

Crittendon, John. 1962. "Aging and Party Affiliation." Public Opinion Quarterly, 26(4):648–657.

Cutler, Neal E. 1969. "Generation, Maturation, and Party Affiliation: A Cohort Analysis." Public Opinion Quarterly, 33(4):583–588.

Delli Carpini, Michael X. 1989. "Age and History: Generations and Sociopolitical Change" In Political Learning in Adulthood: A Sourcebook of Theory and Research, ed. Roberta S. Sigel. Illinois: University of Chicago Press.

Erikson, Robert S., Michael B. MacKuen and James A. Stimson. 2002. "The Macro Polity."

New York: Cambridge University Press.

Fienberg, Stephen E and William M Mason. 1979. "Identification and Estimation of Age-Period-Cohort Models in the Analysis of Discrete Archival Data." Sociological Methodology 10:1–67.

Fiorina, Morris P. 1981. "Retrospective Voting in American National Elections." Connecticut: Yale University Press.

Gelman, Andrew and Jennifer Hill. 2007. "Data Analysis Using Regression and Multilevel/Hierarchical Models." New York: Cambridge University Press.

Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin. 2004. "Bayesian Data Analysis." Florida: Chapman and Hall/CRC.

Gerber, Alan and Donald P. Green. 1998. "Rational Learning and Partisan Attitudes." American Journal of Political Science 42(3):794–818.

Ghitza, Yair and Andrew Gelman. 2013. "Deep Interactions With MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups." American Journal of Political Science 57(3):762–776.

Glenn, Norval D. 1976. "Cohort Analysts' Futile Quest: Statistical Attempts to Separate Age, Period and Cohort Effects." American Sociological Review 41(5):900–904.

Glenn, Norval D and Ted Hefner. 1972. "Further Evidence on Aging and Party Identification." Public Opinion Quarterly 36(1):31–47.

Hoffman, Matthew D and Andrew Gelman. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." Journal of Machine Learning Research15:1351–1381.

Hyman, Herbert. 1959. "Political Socialization: A Study in the Psychology of Political Behavior." New York: Free Press.

Jennings, M. Kent and Richard Niemi. 1981. "Generations and Politics: A Panel Study of Young Adults and Their Parents." New Jersey: Princeton University Press.

Markus, Gregory P. 1983. 'Dynamic Modeling of Cohort Change: The Case of Political Parti-

sanship." American Journal of Political Science 27(4):717–739.

Niemi, Richard G and Barbara I Sobieszek. 1977. "Political Socialization." Annual Review of Sociology 3:209–233.

Niemi, Richard G. and Mary A. Hepburn. 1995. "The Rebirth of Political Socialization." Perspectives on Political Science 24(1):7–16.

Ostrom, Charles W and Renee M Smith. 1992. "Error Correction, Attitude Persistence, and Executive Rewards and Punishments: A Behavioral Theory of Presidential Approval." Political Analysis 4:127–183.

Page, Benjamin I. and Robert Y. Shapiro. 1992. "The Rational Public: Fifty Years of Trends in Americans' Policy Preferences." Illinois: University of Chicago Press.

Pantoja, Adrian D., Ricardo Ramirez and Gary M. Segura. 2001. "Citizens by Choice, Voters by Necessity: Patterns in Political Mobilization by Naturalized Latinos." Political Research Quarterly 54(4):729–750.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. URL: http://www.R-project.org/

Stan Development Team. 2013. "Stan: A C++ Library for Probability and Sampling, Version 1.3.". URL: http://mc-stan.org/

Stimson, James A. 1991. "Public Opinion in America: Moods, Cycles, and Swings." Colorado: Westview Press.

Chapter 4: Forecasting Declining Temperature Sensitivity with a First-hitting-time Model

with E. M. Wolkovich, C. J. Chamberlain, D. M. Buonaiuto, A. K. Ettinger, I. Morales-Castilla, and A. Gelman

Temperature sensitivity—the magnitude of a biological response per °C—is a fundamental concept across scientific disciplines, especially biology, where temperature determines the rate of many plant, animal and ecosystem processes. Recently, a growing body of literature in global change biology has found temperature sensitivities decline as temperatures rise (Fu et al., 2015; Güsewell et al., 2017; Piao et al., 2017; Dai et al., 2019).. Such observations have been used to suggest climate change is reshaping biological processes, with major implications for forecasts of future change. Here we present a simple alternative explanation for observed declining sensitivities: the use of linear models to estimate non-linear temperature responses. Corrections for the non-linearity of temperature response in simulated data and long-term phenological data from Europe remove the apparent decline. Our results show that rising temperatures combined with linear estimates based on calendar time are expected to produce observations of declining sensitivity—without any shift in the underlying biology.

4.1 Introduction

Climate change has reshaped biological processes around the globe, with shifts in the timing of major life history events (phenology), carbon dynamics and other ecosystem processes (IPCC, 2014). With rising temperatures, a growing body of literature has documented changes in temperature sensitivity—the magnitude of a biological response scaled per °C. Many studies have found declining responses to temperature in recent decades (Fu et al., 2015; Güsewell et al., 2017; Piao

et al., 2017; Dai et al., 2019), and some have reported more uniform sensitivities across elevation (Vitasse et al., 2018), or lower sensitivities in warmer, urban areas areas (Meng et al., 2020).

Most studies attribute changes in temperature sensitivity to shifts in underlying biological processes. For example, researchers have suggested weaker temperature sensitivities are evidence of increased light limitation in the tundra (Piao et al., 2017), or a decline in the relative importance of warm spring temperatures for spring phenological events (e.g., leafout, insect emergence) in the temperate zone (Fu et al., 2015; Meng et al., 2020), as other environmental triggers (e.g., winter temperatures that determine 'chilling') play a larger role. Yet, despite an increase in studies reporting declining or shifting temperature sensitivities, none have provided strong evidence of the biological mechanisms underlying these changes (e.g., Fu et al., 2015; Meng et al., 2020). The missing mechanisms may be hidden in the data: environmental factors moderate biological processes in complex ways (Chuine et al., 2016; Güsewell et al., 2017), are strongly correlated in nature (e.g., Fu et al., 2015), and temperature variance shifts over time and space (Keenan et al., 2020).

Here we propose a simpler alternative explanation: the use of linear models for non-linear responses to temperature. Researchers generally use methods with assumptions of linearity to calculate temperature sensitivities, often relying on some form of linear regression to compute a change in a quantity—days to leafout or carbon sequestered over a fixed time, for example—per °C, thus ignoring that many biological responses to temperature are non-linear. We show, theoretically then with simulated and empirical data, how the use of linear methods for non-linear processes can produce an illusion that the mechanisms underlying biological processes are changing.

4.2 A first-hitting-time model of leafout

Many observed biological events are the result of continuous non-linear processes that depend on temperature, which are discretized into temporal units for measurement. For example, a biological response, such as leafout, occurs when a certain thermal sum is reached, and plants will reach this threshold more quickly—in calendar time—when average daily temperatures are warmer

(Chuine, 2000). Biologically, however, the plants may require the same temperature sum. Indeed any process observed or measured as the time until reaching a threshold is inversely proportional to the speed at which that threshold is approached. Temperature determines the speed of many biological processes. Thus, at very low temperatures plants would never leaf out and at higher temperatures they could leaf out in only a matter of days—yet sensitivities estimated from linear regression at higher (warmer) temperatures would appear much lower than those observed at lower temperatures. Warming acts to step on the biological accelerator, and thus may produce declining sensitivities without any change in the underlying process.

We show this by deriving the relationship between a biological response and temperature using simple stochastic model, which describes the first time a random process hits a threshold (called a 'first-hitting-time model). Our model holds the temperature threshold for leafout constant. Even though the mechanism by which temperature leads to leafout does not change, the model produces declining sensitivity—as measured in days per °C—with warming. Indeed, under this model constant temperature sensitivity would be evidence that the temperature threshold is not constant and the mechanisms underlying the leafout process have changed.

We derive the relationship between daily temperature and leafout in two common scenarios. In the first, we take the average daily temperature up until the leafout date. In the second, we take the average daily temperature over a fixed window, such as March 1st to April 30th. In both cases, we discretize time since, although many biological processes depend continuously on time, research typically measures time in discretized units, such as days, weeks, or months. We also assume normality—appealing implicitly to the Markov chain central limit theorem.

4.2.1 Scenario 1: Using average daily temperature until the leafout date

We use the following notation:

n = day since temperatures start to accumulate, $n = 0, 1, ..., N$

X_n = observed temperature on day n

$S_0^n = \sum_{i=0}^{n} X_i$, the cumulative daily temperature from day 0 to day n

$M_0^n = \frac{S_0^n}{n}$, the average daily temperature from day 0 to day n

β = the threshold of interest, $\beta > 0$, (thermal sum required for leafout)

$n_\beta = \underset{n}{\text{argmin}}\ S_n > \beta$, the first day the cumulative daily temperature passes the threshold

(for example, day of year (doy) of leafout).

We model X_n as a Gaussian random walk, $X_n \overset{\text{i.i.d}}{\sim} \text{normal}\,(\alpha_0 + \alpha_1 n, \sigma)$, where $\alpha_0 > 0$ is the average temperature on day $n = 0$, $\alpha_1 > 0$ is the day-over-day increase in average temperatures, and σ is the standard deviation. This model differs from the traditional Gaussian random walk because of the factor n.

This model has two important consequences:

(1) Leafout time is inversely related to average temperature at leafout time.

Under this model, $M_0^{n_\beta}$ and n_β are inversely proportional. To see why, assume for the moment that the cumulative daily temperature hits the threshold exactly on leafout day. That is, $S_0^{n_\beta} = \beta$. Then

$$M_0^{n_\beta} = \frac{S_0^{n_\beta}}{n_\beta} = \frac{\beta}{n_\beta}$$

rearranging yields

$$n_\beta = \frac{\beta}{M_0^{n_\beta}}$$

Many global change biology studies use linear regression to quantify the relationship between n_β and $M_0^{n_\beta}$ (or similar metrics, see Wolkovich et al., 2012; Piao et al., 2017; Keenan et al., 2020, for examples). Regressing n_β on $M_0^{n_\beta}$ finds a best fit line to the inverse curve, $n_\beta = \frac{\beta}{M_0^{n_\beta}}$. The relationship is linearized with the logarithm transformation: $\log(n_\beta) = \log(\beta) - \log(M_0^{n_\beta})$. That is, $\log(n_\beta)$ is linear in log-average daily temperature with slope -1 and intercept $\log(\beta)$.

(2) The variance of the average temperature may decreases as temperatures rise.

Under the model, the mean and variance of M_0^n are $\mathrm{E}(M_0^n|\alpha_0, \alpha_1) = \frac{1}{n}\sum_{i=0}^{n}(\alpha_0 + \alpha_1 i) = \alpha_0 + \alpha_1\frac{(n+1)}{2}$ and $\mathrm{Var}(M_0^n|\alpha_0, \alpha_1) = \frac{\sigma^2}{n}$.

By the law of total variance,

$$\begin{aligned}
\mathrm{Var}(M_0^n) &= \mathrm{E}(\mathrm{Var}(M_0^n|\alpha_0, \alpha_1)) + \mathrm{Var}(\mathrm{E}(M_0^n|\alpha_0, \alpha_1)) \\
&= \frac{\sigma^2}{n} + \mathrm{Var}(\alpha_0 + \alpha_1\frac{n+1}{2}) \\
&= \frac{\sigma^2}{n} + \mathrm{Var}(\alpha_0) + \frac{(n+1)^2}{4}\mathrm{Var}(\alpha_1) + (n+1)\mathrm{Cov}(\alpha_0, \alpha_1)
\end{aligned}$$

As temperatures rise and leafout date becomes earlier, the variance of the average temperature will decline—provided the variation in temperatures, σ, is sufficiently small.

4.2.2 Scenario 2: Using average daily temperature over a fixed window

We slightly modify the notation:

$$
\begin{aligned}
n &= \text{day since temperatures start to accumulate, } n = 0, \ldots, a, \ldots, b \\
X_n &= \text{observed temperature on day } n \\
S_a^n &= \sum_{i=a}^{n} X_i, \text{ the cumulative daily temperature from day } a \text{ to day } n \\
M_a^n &= \frac{S_a^n}{n-a}, \text{ the average daily temperature from day } a \text{ to day } n \\
\beta &= \text{the threshold of interest, } \beta > 0, \text{ (thermal sum required for leafout)} \\
n_\beta &= \operatorname*{argmin}_n S_0^n > \beta, \text{ the first day the cumulative daily temperature passes the threshold} \\
&\quad \text{(for example, day of year (doy) of leafout).}
\end{aligned}
$$

As before, we model X_n as a Gaussian random walk, $X_n \overset{\text{i.i.d}}{\sim} \text{normal}\,(\alpha_0 + \alpha_1 n, \sigma)$, where $\alpha_0 > 0$ is the average temperature on day $n = 0$, $\alpha_1 > 0$ is the day-over-day increase in average temperatures, and σ is the standard deviation. We make the additional assumption that $X_n \geq 0$ for all n and $a < n_\beta < b$. That is, the cumulative temperature acquired by the plant always increases.

Note that

$$S_a^b \sim \text{normal}\left(\alpha_0(b-a) + \tfrac{\alpha_1}{2}(b-a)(b+a+1), \sigma\sqrt{b-a}\right)$$

$$M_a^b \sim \text{normal}\left(\alpha_0 + \tfrac{\alpha_1}{2}(b+a+1), \tfrac{\sigma}{\sqrt{b-a}}\right)$$

$$S_n^b - S_0^a \sim \text{normal}\left(\alpha_0(b-a-n) + \tfrac{\alpha_1}{2}((b-n)(b+n+1) - a(a+1)), \sigma\sqrt{b+a-n}\right)$$

so that

$$\begin{aligned}
Pr\left(n_\beta \le n \mid M_a^b = m\right) &= Pr\left(n_\beta \le n \mid S_a^b = (b-a)m\right) \\
&= Pr\left(S_0^n \ge \beta \mid S_a^b = (b-a)m\right) \\
&= Pr\left(S_n^b \le (b-a)m + S_0^a - \beta\right) \\
&= Pr\left(S_n^b - S_0^a \le (b-a)m - \beta\right) \\
&= \Phi\left(\frac{(b-a)m - \beta - [\alpha_0(b-a-n) + \frac{\alpha_1}{2}((b-n)(b+n+1) - a(a+1))]}{\sigma\sqrt{b+a-n}}\right)
\end{aligned}$$

The distribution of M_a^b shows that consequence (2) above still holds with this model. Consequence (1) no longer holds directly, but will in many situations where average daily temperature until an event correlates strongly with average daily temperature because the window is chosen based, in part, on the expected hitting time (Figs. 4.2-4.3). We note two additional consequences:

(3) The conditional median is quadratic in n:

$$\begin{aligned}
&\frac{1}{2} \overset{\text{set}}{=} Pr\left(n_\beta \le n \mid M_a^b = m\right) \\
&\quad\Rightarrow 0 = (b-a)m - \beta - [\alpha_0(b-a-n) + \frac{\alpha_1}{2}((b-n)(b+n+1) - a(a+1))] \\
&\quad\Rightarrow m = \frac{1}{(a-b)}[-\beta - \alpha_0(b-a-n) - \frac{\alpha_1}{2}((b-n)(b+n+1) - a(a+1))] \\
&\qquad = \frac{1}{(a-b)}[-\beta - \alpha_0(b-a) - \frac{\alpha_1}{2}(b-a)(b+a+1)] + \frac{\alpha_0 + \frac{\alpha_1}{2}}{(a-b)}n + \frac{\frac{\alpha_1}{2}}{(a-b)}n^2 \\
&\qquad := \gamma_0 + \gamma_1 n + \gamma_2 n^2
\end{aligned}$$

(4) The conditional mean and variance are sums of negative sigmoids, according to the follow-

ing identities

$$E\left(n_\beta \mid M_a^b = m\right) = \sum_{n=0}^{\infty} Pr\left(n_\beta \geq n \mid M_a^b = m\right)$$

$$E\left(n_\beta^2 \mid M_a^b = m\right) = \sum_{n=0}^{\infty} n\ Pr\left(n_\beta \geq n \mid M_a^b = m\right)$$

4.3 Simulations

Simulations show that correcting for non-linearity removes apparent declines in temperature sensitivity (Fig. 4.1, 4.3, code link). Assuming a model where warming increases the required thermal sum for a biological event—a common hypothesis for declining sensitivities in spring phenological events—yields declining sensitivities that remain after correcting for non-linearity (Fig. 4.4).

Further, after correcting for non-linearity in long-term leafout data from Europe, we find little evidence for declining sensitivities with warming (Figs. 4.1, 4.5, 4.6). An apparent decline in sensitivity for silver birch (Betula pendula) from -4.3 days/°C to -3.6 days/°C from 1950-1960 compared to 2000-2010 disappears using a log-log regression (-0.17 versus -0.22). We see similar corrections using 20-year windows, and a potential increase in sensitivity for European beech (Fagus sylvatica, see Tables 4.1-4.2). Moreover, the variance of the leafout dates of both species declines as temperatures rise—(declines of roughly 50%, see Tables 4.1-4.2), which is expected under our model as warming accelerates towards the thermal threshold that triggers leafout (and in contrast to predictions from changing mechanisms, see Ford et al., 2016).

Our theoretical model and empirical results show that rising temperatures are sufficient to explain declining temperature sensitivity. It is not necessary to invoke changes to the mechanisms that underlie the biological processes themselves.

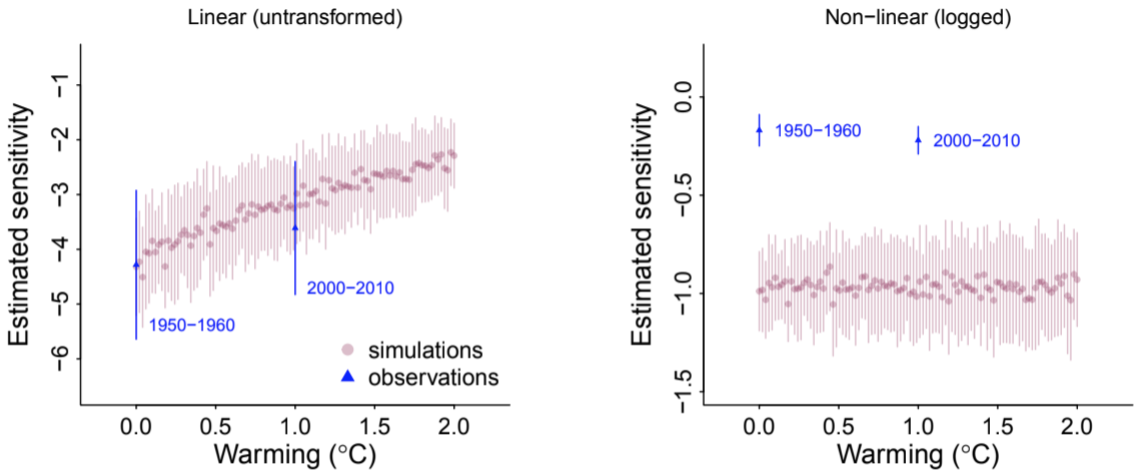


Figure 4.1: Shifts in temperature sensitivities with warming occur when using linear models for non-linear processes. Estimated sensitivities decline with warming in simulations (left) with no underlying change in the biological process when sensitivities were estimated with linear regression (estimated across 45 sites with a base temperature of normal(6,4)). This decline disappears when performing the regression on logged predictor and response variables (right). Such issues may underlie declining sensitivities calculated from observational data, including long-term observations of leafout across Europe ('observations,' using data for Betula pendula from PEP725 from for the 45 sites that had complete data for 1950-1960 and 2000-2010), which show a lower sensitivity with warming when calculated on raw data, but no change in sensitivity using logged data. Symbols and lines represent means ± standard deviations of regressions across sites. See SI for a discussion of why estimated sensitivities are -1 or lower in non-linear models.

4.4 Long-term empirical data from PEP725

To examine how estimated sensitivities shift over time, we selected sites of two common European tree species (silver birch, Betula pendula, and European beech, Fagus sylvatica) that have long-term observational data of leafout, through the Pan European Phenology Project (PEP725, Templ et al., 2018). We used a European-wide gridded climate dataset (E-OBS, Cornes et al., 2018) to extract daily minimum and maximum temperature for the grid cells where observations of leafout for these two species were available. We used sites with complete leafout data across both our 10-year (and 20-year) windows to avoid possible confounding effects of shifting sites over time (see Tables 4.1-4.2 for numbers of sites per species x window).

Our estimates of temperature sensitivity from a linear model using untransformed variables show a decline in sensitivity with recent warming for Betula pendula over 10 and 20-year windows, but no decline for Fagus sylvatica; using logged variables estimates appeared more similar over time or sometimes suggested an increase in sensitivity (see Figs. 4.5-4.6, Tables 4.1-4.2). This shift in estimated sensitivity when regressing with untransformed versus logged variables suggests the declining estimates with untransformed variables may not be caused by changes in the underlying mechanisms of leafout (i.e., reduced winter chilling) and driven instead by using linear regression for a non-linear process. This hypothesis is supported further by large declines in variance of leafout in recent decades.

Shifts in variance provide another hurdle to robust estimates of temperature sensitivity. Previous work has highlighted how shifting temperature variance (over space and/or time) could lead to shifting estimates of temperature sensitivities (Keenan et al., 2020), but our results stress that variance in both leafout and temperature are shifting. If both shift in step, estimates would not be impacted by changes in temperature variance, but our results suggest variance in temperature—for these data—has declined more than variance in leafout, though both have declined substantially in

recent decades (Tables 4.1-4.2).

Estimated sensitivities for the empirical data (PEP725) using logged variables are far lower than the value obtained in our simulations (-1). This likely results from a contrast between our simulations—where we can accurately define the temperature plants experience and the temporal window that drives leafout—and our empirical data, where we do not know how measured temperatures translate into the temperatures that plants accumulate and where we have no clear method to define the relevant temporal window (Güsewell et al., 2017).

These results highlight how the acceleration of biological time due to climate change requires researchers to clarify their assumptions. Expecting temperature sensitivity to remain constant as temperatures rise assumes the relationship between response and temperature is proportional. But the underlying biological processes suggest this relationship is seldom proportional, or even linear. In fact, when our model holds, declining sensitivity with rising temperatures should be the null hypothesis of any analysis of temperature sensitivity based on linear regression or similar methods.

Inferring biological processes from statistical artifacts is not a new problem (e.g., Nee et al., 2005), but climate change provides a new challenge in discerning mechanism from measurements because it affects biological time, while researchers continue to use calendar time. Other fields focused on temperature sensitivity often use approaches that acknowledge the non-linearity of responses (e.g., Yuste et al., 2004). Researchers have called for greater use of process-based models (Keenan et al., 2020), which often include non-linear responses to temperature, but rely themselves on exploratory methods and descriptive analyses for progress (Chuine et al., 2016). The challenge, then, is to interrogate the implicit and explicit models we use to interpret data summaries, and to develop null expectations that apply across biological and calendar time.

years	species	mean (ST)	mean (ST.leafout)	var (ST)	var (leafout)	mean (GDD)	slope	log-slope
1950-1960	*Betula pendula*	5.6	7.0	3.4	110.5	71.7	-4.3	-0.17
2000-2010	*Betula pendula*	6.6	6.8	1.2	47.0	64.6	-3.6	-0.22
1950-1960	*Fagus sylvatica*	5.6	7.5	3.3	71.9	83.8	-2.8	-0.11
2000-2010	*Fagus sylvatica*	6.7	7.7	1.2	38.3	86.7	-3.4	-0.20

Table 4.1: Climate and phenology statistics for two species (Betula pendula, Fagus sylvatica, across 45 and 47 sites respectively) from the PEP725 data across all sites with continuous data from 1950-1960 and 2000-2010. ST is spring temperature from 1 March to 30 April, ST.leafout is temperature 30 days before leafout, and GDD is growing degree days 30 days before leafout. Slope represents the estimated sensitivity using untransformed leafout and ST, while log-slope represents the estimated sensitivity using log(leafout) and log(ST). We calculated all metrics for for each species x site x 10 year period before taking mean or variance estimates. See also Fig. 4.

years	species	mean (ST)	mean (ST.leafout)	var (ST)	var (leafout)	mean (GDD)	slope	log-slope
1950-1970	*Betula pendula*	5.8	7.1	2.6	79.9	72.5	-4.3	-0.19
1970-1990	*Betula pendula*	5.9	7.2	1.3	104.8	72.2	-6.1	-0.33
1990-2010	*Betula pendula*	6.8	6.7	0.9	36.2	60.0	-3.3	-0.21
1950-1970	*Fagus sylvatica*	5.6	7.6	2.7	63.4	86.0	-3.1	-0.12
1970-1990	*Fagus sylvatica*	5.6	7.5	1.3	56.2	81.3	-2.5	-0.16
1990-2010	*Fagus sylvatica*	6.7	7.3	1.2	31.4	76.0	-3.4	-0.19

Table 4.2: Climate and phenology statistics for two species (Betula pendula, Fagus sylvatica, across 17 and 24 sites respectively) from the PEP725 data across all sites with continuous data from 1950-2010. ST is spring temperature from 1 March to 30 April, ST.leafout is temperature 30 days before leafout, and GDD is growing degree days 30 days before leafout. Slope represents the estimated sensitivity using untransformed leafout and ST, while log-slope represents the estimated sensitivity using log(leafout) and log(ST). We calculated all metrics for for each species x site x 20 year period before taking mean or variance estimates. See also Fig. 5.

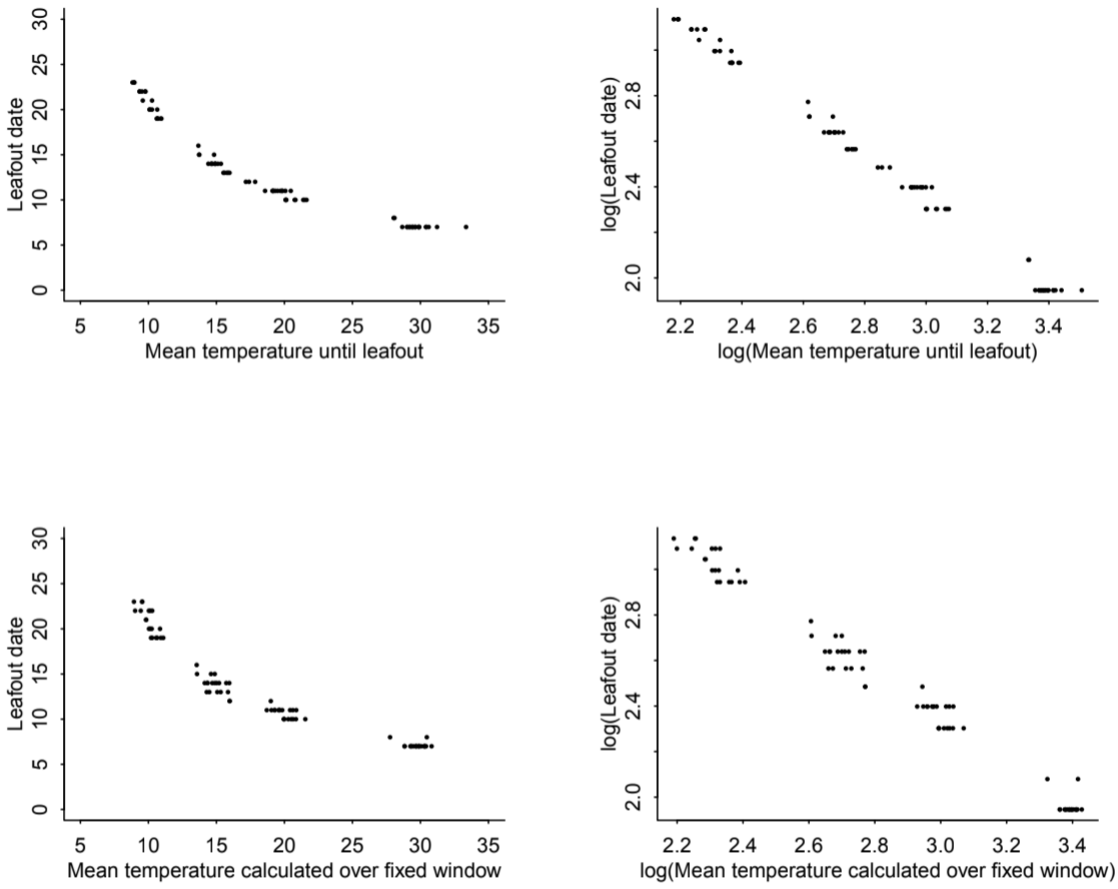


Figure 4.2: Simulated leafout as a function of temperature across different temperatures highlights non-linearity of process. Here we simulated sets of data where leafout constantly occurs at 200 growing degree days (thermal sum of mean daily temperatures with 0°C as base temperature) across mean temperatures of 10, 15, 20 and 30°C (constant SD of 4), we calculated estimated mean temperature until leafout date (top row) or across a fixed window (bottom row, similar to estimates of 'spring temperature'). While within any small temperature range the relationship may appear linear, its non-linear relationship becomes clear across the greater range shown here (left). Taking the log of both leafout and temperature (right) linearizes the relationship.

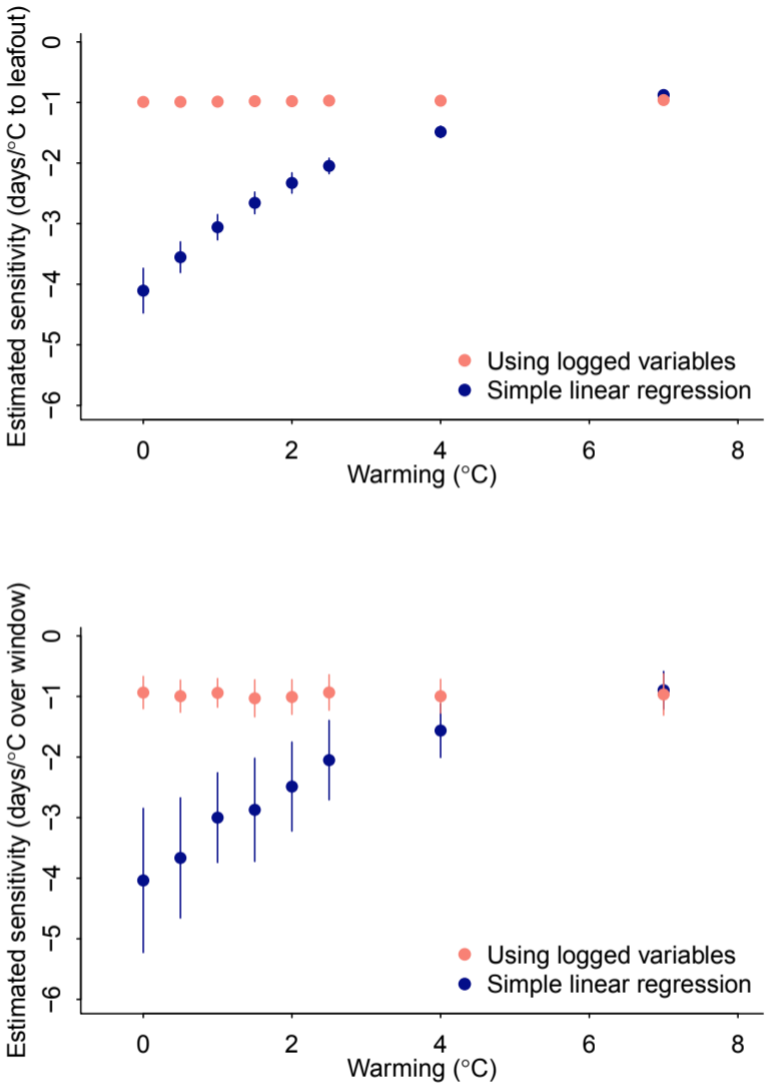


Figure 4.3: A simple model generates declining sensitivities with warming. We show declines in estimated sensitivities with warming from simulations (top: using average temperature until leafout, bottom: using a fixed window) with no underlying change in the biological process when sensitivities were estimated with simple linear regression ("Simple linear regression"). This decline disappears using regression on logged predictor and response variables ("Using logged variables").

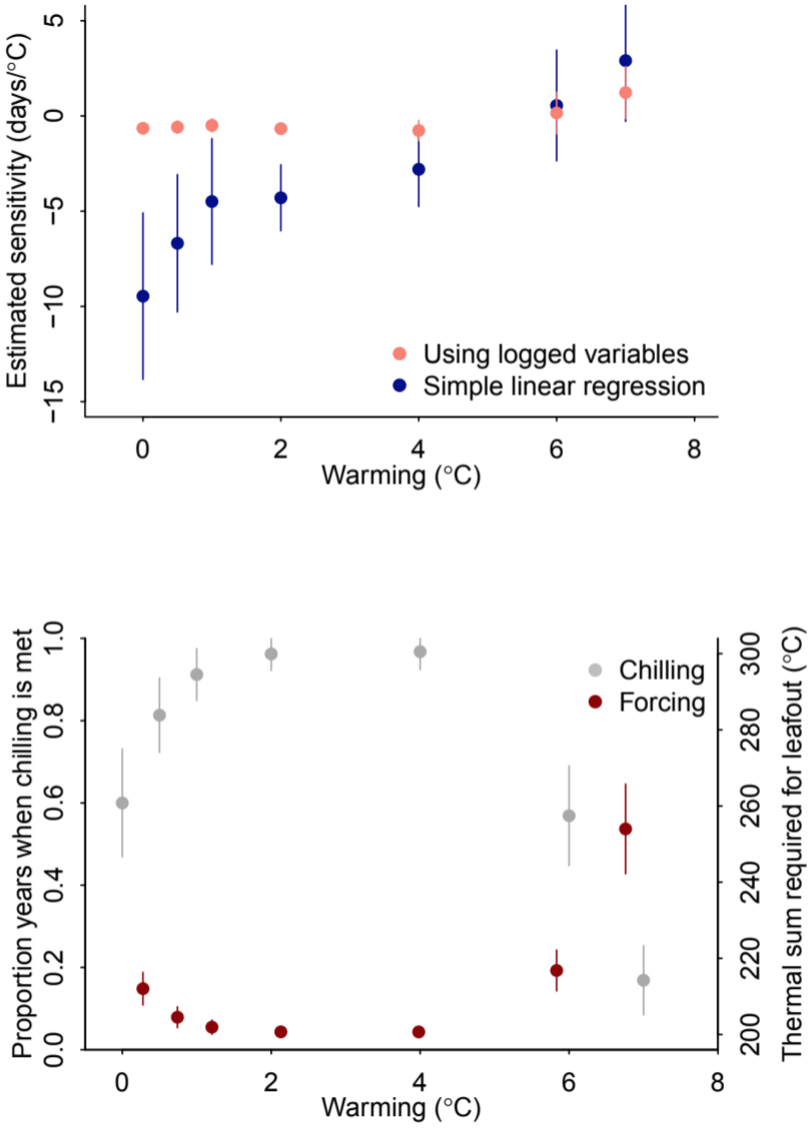


Figure 4.4: Simulated leafout as a function of temperature across different levels of warming with shifts in underlying biology. Here we simulated sets of data where leafout occurs at a thermal sum of 200 (sum of mean daily temperatures with 0°C as base temperature) when chilling is met, and requires a higher thermal sum when chilling is not met. We show estimated sensitivities in the top panel, and the shifting cues in the bottom panel.

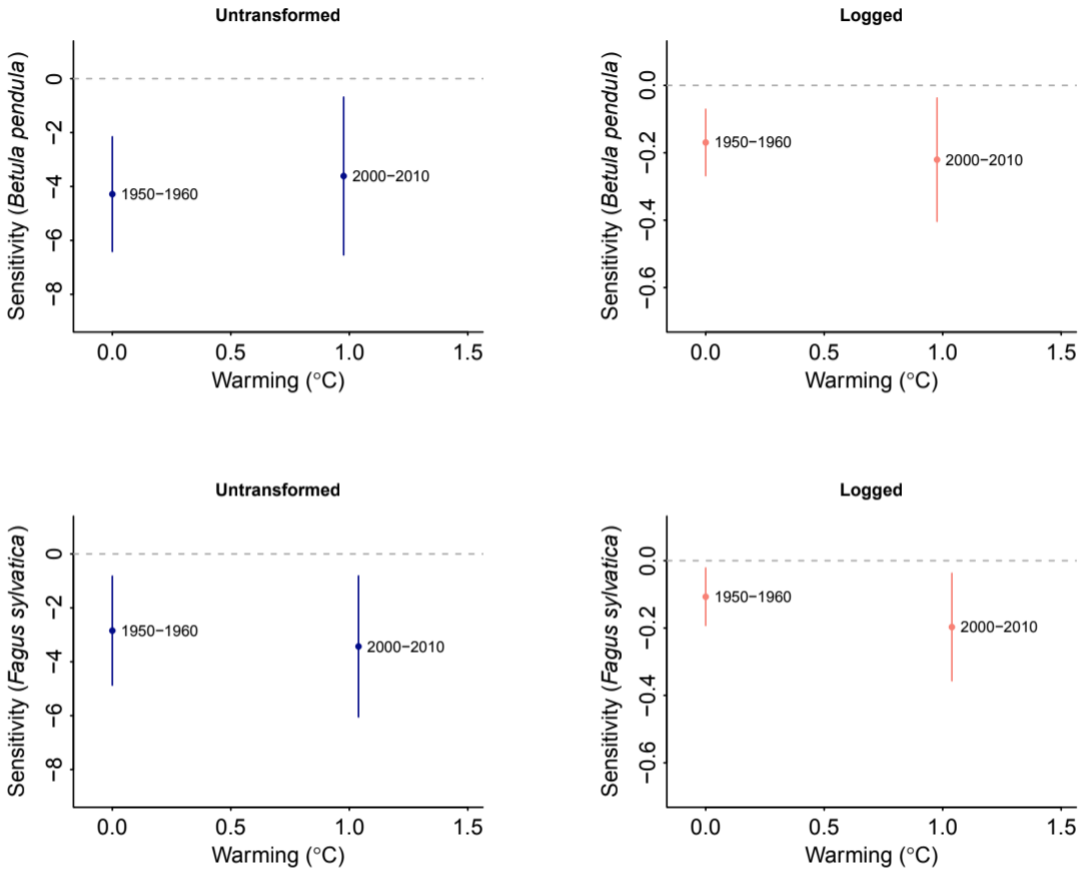


Figure 4.5: Sensitivities from PEP725 data using 10 year windows of data for two species (top – Betula pendula, bottom – Fagus sylvatica; all lines show 78% confidence intervals from linear regressions). Amounts of warming are calculated relative to 1950-1960 and we used only sites with leafout data in all years shown here. See Table 4.1 for further details.

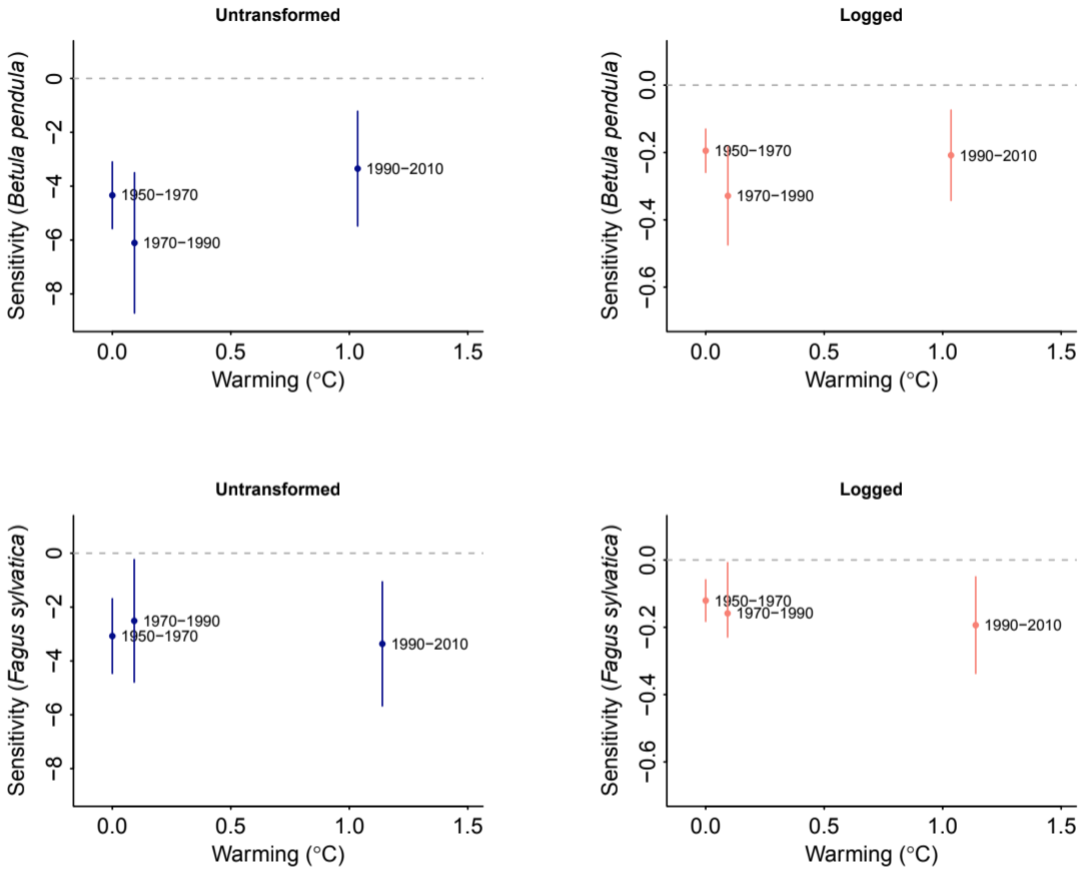


Figure 4.6: Sensitivities from PEP725 data using 20 year windows of data for two species (top – Betula pendula, bottom – Fagus sylvatica; all lines show 78% confidence intervals from linear regressions). Amounts of warming are calculated relative to 1950-1970 and we used only sites with leafout data in all years shown here. See Table 4.2 for further details.

4.5 References

Chuine, I. 2000. "A unified model for budburst of trees." Journal of Theoretical Biology 207:337–347.

Chuine, I., M. Bonhomme, J.-M. Legave, I. Garcıa de Cortazar-Atauri, G. Charrier, A. Lacointe, and T. Ameglio. 2016. "Can phenological models predict tree phenology accurately in the future? The unrevealed hurdle of endodormancy break." Global Change Biology 22:3444–3460.

Cornes, R. C., G. van der Schrier, E. J. van den Besselaar, and P. D. Jones. 2018. "An ensemble version of the E-OBS temperature and precipitation data sets." Journal of Geophysical Research: Atmospheres 123:9391–9409.

Dai, W. J., H. Y. Jin, Y. H. Zhang, T. Liu, and Z. Q. Zhou. 2019. "Detecting temporal changes in the temperature sensitivity of spring phenology with global warming: Application of machine learning in phenological model." Agricultural and Forest Meteorology 279.

Ford, K. R., C. A. Harrington, S. Bansal, J. Gould, Peter, and J. B. St. Clair. 2016. "Will changes in phenology track climate change? A study of growth initiation timing in coast Douglas–fir." Global Change Biology 22:3712–3723.

Fu, Y. S. H., H. F. Zhao, S. L. Piao, M. Peaucelle, S. S. Peng, G. Y. Zhou, P. Ciais, M. T. Huang, A. Menzel, J. P. Uelas, Y. Song, Y. Vitasse, Z. Z. Zeng, and I. A. Janssens. 2015. "Declining global warming effects on the phenology of spring leaf unfolding." Nature 526:104–107.

Ǵusewell, S., R. Furrer, R. Gehrig, and B. Pietragalla. 2017. "Changes in temperature sensitivity of spring phenology with recent climate warming in Switzerland are related to shifts of the preseason." Global Change Biology 23:5189–5202.

IPCC. 2014. "Climate Change 2014: Impacts, Adaptation, and Vulnerability." Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Keenan, T. F., A. D. Richardson, and K. Hufkens. 2020. "On quantifying the apparent temperature sensitivity of plant phenology." New Phytologist 225:1033–1040.

Meng, L., J. Mao, Y. Zhou, A. D. Richardson, X. Lee, P. E. Thornton, D. M. Ricciuto, X.

Li, Y. Dai, X. Shi, and G. Jia. 2020. "Urban warming advances spring phenology but reduces the response of phenology to temperature in the conterminous united states." Proceedings of the National Academy of Sciences 117:4228.

Nee, S., N. Colegrave, S. A. West, and A. Grafen. 2005. "The illusion of invariant quantities in life histories." Science 309:1236–1239.

Piao, S., Z. Liu, T. Wang, S. Peng, P. Ciais, M. Huang, A. Ahlstrom, J. F. Burkhart, F. Chevallier, I. A. Janssens, et al. 2017. "Weakening temperature control on the interannual variations of spring carbon uptake across northern lands." Nature climate change 7:359.

Templ, B., E. Koch, K. Bolmgren, M. Ungersböck, A. Paul, H. Scheifinger, T. Rutishauser, M. Busto, F.-M. Chmielewski, L. H'ajkov'a, S. Hodzi'c, F. Kaspar, B. Pietragalla, R. Romero Fresneda, A. Tolvanen, V. Vu'ceti'c, K. Zimmermann, and A. Zust. 2018. "Pan European Phenological database (PEP725): a single point of access for European data." International Journal of Biometeorology 62:1109–1113.

Vitasse, Y., C. Signarbieux, and Y. H. Fu. 2018. "Global warming leads to more uniform spring phenology across elevations." Proceedings of the National Academy of Sciences 115:1004–1008.

Wolkovich, E. M., B. I. Cook, J. M. Allen, T. M. Crimmins, J. L. Betancourt, S. E. Travers, S. Pau, J. Regetz, T. J. Davies, N. J. B. Kraft, T. R. Ault, K. Bolmgren, S. J. Mazer, G. J. McCabe, B. J. McGill, C. Parmesan, N. Salamin, M. D. Schwartz, and E. E. Cleland. 2012. "Warming experiments underpredict plant phenological responses to climate change." Nature 485:494–497.

Yuste, J., I. A. Janssens, A. Carrara, and R. Ceulemans. 2004. "Annual Q10 of soil respiration reflects plant phenological patterns as well as temperature sensitivity." Global Change Biology 10:161–169.

Chapter 5: Forecasting Migration Patterns with an Infinitesimal Block Model

Learning migration patterns is of central importance in business, government, and the social and biological sciences. But migration data—the flow of individuals between locations or "units"—are not easily observed. Only longitudinal data—population snapshots at each unit—are readily available. We propose a model that identifies latent migration patterns from longitudinal data by combining a continuous-time Markov chain (CTMC) with a stochastic block model (SBM). In our model, migrations are represented by a system of differential equations, and the infinitesimal generator by a low-rank, block matrix. We outline a Markov Chain Monte Carlo (MCMC) algorithm for drawing samples from the posterior distribution of this matrix, and a variational approximation for high dimensions. We then use these algorithm to discover migration patterns from two datasets: the U.S. Internal Revenue Service Population Migration data and the U.S. Census American Community Survey. We find our model retains the interpretability of an SBM and captures migration patterns better than a group-based trajectory model. In addition, our model explains nearly all the variation in the data, and its predictions are competitive with the predictive but less interpretable Lee-Carter matrix factorization model. Finally, we extend our model to solve a problem uniquely suited to our approach: estimating the number of crimes prevented by the NYPD practice Stop, Question, and Frisk. We find SQF prevented relatively few crimes, with 99 percent of the observed crime reduction between 2002 and 2012 explained by migrations.

5.1 Introduction

We propose a model that learns latent migration patterns from longitudinal data. Learning these migrations is of considerable interest in business, government, and the social and biological

sciences. But migration data are difficult to obtain because cost, privacy restrictions, and other extenuating circumstances typically preclude the following of individuals across time, which is necessary to collect migration data. The only data available are longitudinal—the number of individuals in each location or "unit" at various time periods.

For example, demographers want to learn how individuals move, but all they observe are the counts of neighborhood populations each year; epidemiologists want to learn how contaminants spread, but all they observe are the number of disease cases admitted to hospitals each week; firms want to learn how buyer preferences change, but all they observe are the quantity of competing products sold each quarter. In these and many other cases, the units of interest are measured at multiple time points, but the interaction between units is not. Migrations of individuals between units—typically represented as the edges of a directed graph or "flow network"—cannot be observed directly.

Our goal is to learn network flows from longitudinal stocks. We assume the data can be represented by a matrix whose rows are time series, columns are cross sections, and entries are counts. Our model builds primarily on two bodies of work that are commonly applied to such data. In the matrix factorization literature (Lee and Carter, 1992; Kanjilalet al., 1997; Lee, 2000; Pedroza, 2006), researchers decompose the data matrix according to its spectrum. In the mixed membership literature (Lin et al., 2002; Hannah et al., 2011; Lecci, 2014; Manrique-Vallier, 2014), researchers decompose each time series into a few, simple trajectories.

These and similar algorithms work by taking advantage of the correlation between units induced by latent migrations. Yet neither literature explicitly identifies the latent migration pattern. Our contribution is the direct estimation of migration flows from longitudinal data. We represent these flows with a system of differential equations and approximate the infinitesimal generator by a low-rank matrix. Since the infinitesimal generator is an adjacency matrix, the block assumption made in the Stochastic Block Model (SBM) is natural here. For reference purposes, we call the proposed model the Infinitesimal Block Model (IBM).

Our work is conceptually similar to—but meaningfully distinct from—the literature on longi-

tudinal network data. With longitudinal network data, a network is observed over multiple time periods. In our framework, network data is not observed at all. Complicated models that have proven successful with longitudinal network data, for example the logistic map, the Hawkes process, and VARIMA, are generally unstable when applied to longitudinal data; A single migration can drastically alter subsequent migrations, making algorithms extremely sensitive to initial conditions and unreliable for inference.

5.2 Model

In our model, each unit is one state of a continuous-time Markov chain (CTMC). We refer to units as "locations" among which "individuals" migrate. However, the proposed model is applicable to a wide range of phenomenon. For ease of presentation, we begin with a system that is homogeneous and closed, and at all times individuals reside in exactly one of n locations. The model is easily generalized to accommodate inhomogeneous, open systems, which we demonstrate in Section 5 and discuss further in the Appendix.

The instantaneous migration rate between locations is represented by the $n \times n$ matrix $Q = [q_{ij}]$. Off-diagonal element q_{ij} is the rate at which individuals move to location i from location j. Diagonal element q_{ii} is the rate at which individuals leave location i. It follows that $q_{ij} > 0$ and $q_{ii} = -\sum_{j \neq i} q_{ij}$ (Lawler, 2018).

Were migration data collected—perhaps a sample of individuals observed across time—a weighted adjacency matrix $A = [a_{ij}]$ could be constructed and Q inferred by a Stochastic Block Model (SBM) (Holland et al., 1983; Aicher et al.,2014). The SBM assumes each location i belongs to one of $K \leq n$ clusters or "communities", denoted $k_i \in \{1, \ldots, K\}$, and models a_{ij} as independently distributed $F(k_i, k_j)$. The clusters are frequently interpreted as a discretization of a (asymmetric) graphon (Bickel and Chen, 2009), the functional which parameterizes the set of exchangeable (directed) graphs (Diaconis and Janson, 2007).

We take this interpretation here. The graphon encodes the migration rate from one location to any other and discretization delineates the major flow or "blocks" of migration. But the migration

data necessary to fit this model are not available—we can only observe the number of individuals at each location each time period. In this section, we derive the data as a function of Q and then apply a block approximation.

5.2.1 Derivation

Let $y_i(t)$ denote the population of location i at time t. We assume the data are balanced: every location $i \in \{1, \ldots, n\}$ is measured at every time $t \in \{1, \ldots, T\}$. The result is an $n \times T$ matrix of counts $y = [y_i(t)]$, whose ith row is denoted y_i and tth column $y(t)$.

In an arbitrarily small amount of time, Δt, the net migration to location i, Δy_i, is the difference between the number of immigrants—all individuals leaving locations $j \neq i$ for location i—and the number of emigrants—all individuals leaving location i. We assume migration between times t and $t + \Delta t$ is linear in the population so that migrants arrive at a constant rate q_{ij}. Since $q_{ii} = -\sum_{j \neq i} q_{ij}$, net migration to location i is:

$$\Delta y_i = (q_{i1} y_1 + q_{i2} y_2 + \cdots + q_{in} y_n)\, \Delta t.$$

Sending $\Delta t \to 0$ yields a system of ordinary differential equations:

$$\begin{aligned}
\frac{dy_1}{dt} &= q_{11} y_1 + q_{12} y_2 + \cdots + q_{1n} y_n \\
\frac{dy_2}{dt} &= q_{21} y_1 + q_{22} y_2 + \cdots + q_{2n} y_n \\
&\vdots \\
\frac{dy_n}{dt} &= q_{n1} y_1 + q_{n2} y_2 + \cdots + q_{nn} y_n,
\end{aligned}$$

whose solution is

$$y(t) = e^{Qt} y(0)$$

where e^{Qt} is the matrix exponential of Qt (Leonard, 1996).

The matrix Q, called the infinitesimal generator, is given low-rank block structure

$$Q = ZBZ^T - \text{diag}(ZBZ^T\mathbf{1}).$$

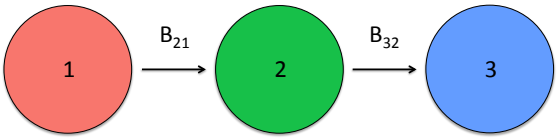
Z is an $n \times K$ binary matrix of cluster assignments, i.e. $Z_{ik_i} = 1$ and 0 otherwise. B is a $K \times K$ positive matrix of instantaneous migration rates, and $\mathbf{1}$ is a $K \times 1$ vector of ones. ZBZ^T is the SBM block decomposition (Holland et al. 1983). Since the system is closed, we account for emigrations by subtracting column sums from the diagonal of ZBZ^T, i.e. $\text{diag}(ZBZ^T\mathbf{1})$.

5.2.2 Identification

Not all migration patterns can be identified from longitudinal data. Cyclical patterns cannot be represented uniquely; It is impossible to learn from the data if five individuals moved from i to j and five individuals moved from j to i, or there were no migrations at all. Therefore, we restrict our model to monotonic, acyclic migration patterns by constraining the lower subdiagonal of B to be monotonic and zero elsewhere—except for the diagonal, which is always chosen so that the columns sum to zero.

Figures 5.1 and 5.2 show two such block migration patterns. On the left side, migrations are shown as a directed acyclic graph (DAG). On the right side, they are shown as an infinitesimal generator Q. In both Figures, migrants move from all red locations to all green locations at rate B_{21} and from all green locations to all blue locations at rate B_{32}.

We note that this restriction eliminates the label-switching problem and is sufficiently flexible to learn the mass migrations generally of interest to researchers—e.g. displacement by conflict or gentrification—as we demonstrate in Section 5. The triangular parameterization of Q also has computational benefits. However, the matrix can be reparametrized for other patterns.

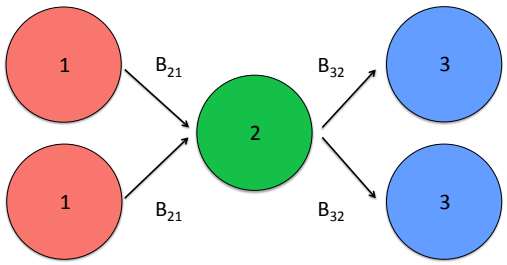


(a) Directed Acyclic Graph (DAG)

$$\begin{bmatrix} -B_{21} & 0 & 0 \\ B_{21} & -B_{32} & 0 \\ 0 & B_{32} & 0 \end{bmatrix}$$

(b) Block Infinitesimal Generator Q

Figure 5.1: Two equivalent representations of a block migration pattern. Migrants move from one location in the red cluster to one location in the green cluster at rate B_{21} and from one location in the green cluster to one location in the blue cluster at rate B_{32}.



(a) Directed Acyclic Graph (DAG)

$$\begin{bmatrix} -B_{21} & 0 & 0 & 0 & 0 \\ 0 & -B_{21} & 0 & 0 & 0 \\ B_{21} & B_{21} & -2B_{32} & 0 & 0 \\ 0 & 0 & B_{32} & 0 & 0 \\ 0 & 0 & B_{32} & 0 & 0 \end{bmatrix}$$

(b) Infinitesimal Generator Q

Figure 5.2: Two equivalent representations of a block migration pattern. Migrants move from two locations in the red cluster to one location in the green cluster at rate B_{21} and from one location in the green cluster to two locations in the blue cluster at rate B_{32}.

5.2.3 Specification

We assume individuals who do not follow the block migration pattern do so randomly. We model the number of individuals at location i as distributed i.i.d. Poisson with mean equal to the ith element of $e^{Qt}y(0)$. This allows locations to experience population increases despite long-term out-migrations. The final result is the Infinitesimal Block Model (IBM):

$$y(t) \sim \text{Poisson}(e^{Qt}y(0)),$$

where the $K-1$ subdiagonal elements of B and n rows of $Z = (Z_1, Z_2, \ldots, Z_n)^T$ follow the distributions typical of mixture models:

$$\begin{aligned} B_{k(k-1)} &\sim \text{Gamma}(\theta, \phi) \\ Z_i &\sim \text{Categorical}(\pi) \\ \pi &\sim \text{Dirichlet}(\alpha). \end{aligned}$$

5.3 Inference

The posterior distribution of Q cannot be expressed in closed form because the model is not conjugate. Markov Chain Monte Carlo (MCMC) algorithms are commonly used to draw (approximate) samples from a posterior. We implement the Metropolis-within-Gibbs sampler shown in Algorithm 1 (Gelfand and Smith, 1990; Metropolis et al. 1953; Geman and Geman, 1987)—we detail a large-sample approximation in the Appendix. The state of the sampler at iteration t is the set of all $n + K - 1$ parameters: the n rows of the cluster assignment matrix, $z^{(t)} = \left(z_1^{(t)}, z_2^{(t)}, \ldots, z_n^{(t)}\right)^T$, and the $K-1$ elements of the migration matrix subdiagonal, $b^{(t)} = \left(b_{21}^{(t)}, b_{32}^{(t)}, \ldots, b_{K(K-1)}^{(t)}\right)$. Metropolis proposals are made sequentially across $z^{(t)}$ ($b^{(t)}$) by simple (Gaussian) random walk: A simple random step is proposed for each $z_i^{(t)}$, followed by one multivariate Gaussian random step for all $b^{(t)}$. The Appendix details the acceptance probability, and an R language (2020) code implementation is available upon request. In the remainder of this section, we highlight two important

considerations when implementing Algorithm 1.

Algorithm 1: IBM Metropolis-within-Gibbs Sampler

Input: $z^0 = \left(z_1^0, \ldots, z_n^0\right)$, $b^0 = \left(b_{21}^0, b_{32}^0, \ldots b_{K(K-1)}^0\right)$, y, convergence criteria.
repeat
  for $i = 1$ to n **do**
    propose z_i^*
    $U \sim \text{uniform}(0, 1)$
    if $U < \dfrac{p\left(z_i^* \middle| b^{(t-1)}, z_{1:(i-1)}^{(t)}, z_{(i+1):n}^{(t-1)}, y\right)}{p\left(z_i^{(t-1)} \middle| b^{(t-1)}, z_{1:(i-1)}^{(t)}, z_{(i+1):n}^{(t-1)}, y\right)}$ **then**
      $z_i^{(t)} \leftarrow z_i^*$
    else
      $z_i^{(t)} \leftarrow z_i^{(t-1)}$
    end if
  end for
  propose b^*
  $U \sim \text{uniform}(0, 1)$
  if $U < \dfrac{p\left(b^* \middle| z^{(t)}, y\right)}{p\left(b^{(t-1)} | z^{(t)}, y\right)}$ **then**
    $b^{(t)} \leftarrow b^*$
  else
    $b^{(t)} \leftarrow b^{(t-1)}$
  end if
  $t \leftarrow t + 1$
until convergence criteria met

5.3.1 Parameterization

The subdiagonal of B is constrained to be monotonic, which may make Algorithm 1 inefficient. We recommend two different parameterizations of the IBM, depending on whether K is small or large. For small K, we suggest a Gaussian random walk in η-space

$$\eta_1 = b_{21}, \quad \eta_2 = b_{32} - b_{21}, \quad \ldots \quad , \quad \eta_{K-1} = b_{K(K-1)} - b_{(K-1)(K-2)},$$

rejecting samples that fail to satisfy the constraint $\eta_j > 0$. For large K, most samples will be rejected, and we recommend the unconstrained transformation

$$\eta_1 = \log b_{21}, \;\; \eta_2 = \log\left(b_{32} - b_{21}\right), \;\; \ldots \;\; , \;\; \eta_{K-1} = \log\left(b_{K(K-1)} - b_{(K-1)(K-2)}\right).$$

The Appendix describes the Jacobian adjustment necessary for the second transformation.

5.3.2 Update Schedule

Algorithm 1 does not force each cluster to have a location, which may result in a posterior sample where one or more clusters are assigned no locations. This problem of vanishing clusters is common in clustering algorithms. We find it typically arises during the first few iterations of Algorithm 1 when the state of migration matrix b is in a low probability region, and the cluster assignments, z_i, are forced to compensate for the implausible b.

We recommend the alternative update schedule shown in the Appendix (Algorithm 2) to mitigate the problem of vanishing clusters. With probability $\beta \in [0, 1]$, a new z_i is proposed by swapping the current cluster assignment of location i with a randomly chosen location in a neighboring cluster j. With probability $1 - \beta$, a new z_i is proposed by simple random walk—switching the label of location i to a randomly chosen, neighboring cluster. Swapping does not change the number of locations in each cluster so large β protects against vanishing clusters.

Swapping has additional benefits. For example, the matrix exponential is computationally expensive. When a swap occurs, the proposal acceptance probability can be computed using the matrix exponential from the previous iteration of the sampler and permuting the data.

5.4 Evaluation

We evaluate the IBM with both simulated and real-world data. We first test whether Algorithm 1 can recover the infinitesimal generator Q using computer-generated data from the IBM likelihood. We then test whether Algorithm 1 can discover real migration patterns using two pub-

licly available datasets: the U.S. Internal Revenue Service (IRS) Population Migration data and the U.S. Census Bureau (Census) American Community Survey (ACS).

All evaluations use the R language (2020). In the first real-world dataset, the true migration rate is known, and the IBM is compared to a SBM using the package `Blockmodels` (Leger, 2016) and a group-based trajectory model (Manrique-Vallier, 2014) (referred to as M-GLM) using the package `FlexMix` (Leisch, 2004). In the second real-world dataset, the true migration rate is not known, and predictions from the IBM are compared to predictions from the trajectory model and the Lee and Carter (1992) (referred to as SVD) matrix factorization model.

These comparison algorithms are popular tools for clustering and prediction, and we demonstrate that the IBM is a competitive alternative. However, we stress that neither competitor explicitly learn migration patterns from longitudinal data, which is the express purpose of the IBM.

5.4.1 Simulation

Algorithm 1 can recover the infinitesimal generator Q in a wide variety of simulations. Generally, the step size of the Gaussian proposal must be tuned by grid search, which is easily parallelized. See the Appendix for simulation details.

5.4.2 IRS Data

IBM correctly identifies major migration patterns from real data. The IBM approximation is more accurate than a group-based trajectory model.

We use the U.S. Internal Revenue Service's Population Migration data, a rare dataset that follows the county-to-county migrations of U.S. taxpayers each year. We restrict our attention to the last $T = 10$ available years of the IRS data $(2006, \ldots, 2015)$ and the $n = 35$ counties that make up the New York–Newark, NY–NJ–CT–PA Combined Statistical Area (CSA). This area is the largest urban economic agglomeration in the Americas and experienced rapid urbanization over the last decade. The question is how well IBM captures that migration. Both the data and the CSA are described in greater detail in the Appendix.

We create two matrices from this data. The first is the $n \times T$ matrix of county populations, y, where $y_i(t)$ is the number of taxpayers who file in county i in year t. The second is an $n \times n$ matrix of net-migrations, $A = [a_{ij}]$, where a_{ij} is the number of taxpayers that immigrated to county i from county j between 2006 and 2015 minus the number that emigrated. Negative numbers are set to zero. The former is used to fit IBM and the latter is used to validate the results.

We run Algorithm 1 using $y_i(t)$ and $K = 3$. We sample one thousand chains for one thousand iterations each and retain the maximum a posteriori (MAP) cluster assignment, $\hat{Z}_{\text{MAP}}$. We normally recommend clustering on the trajectories, y_i, to initialize IBM. However, since these clusters are the benchmark algorithm, we initialize cluster assignments randomly.

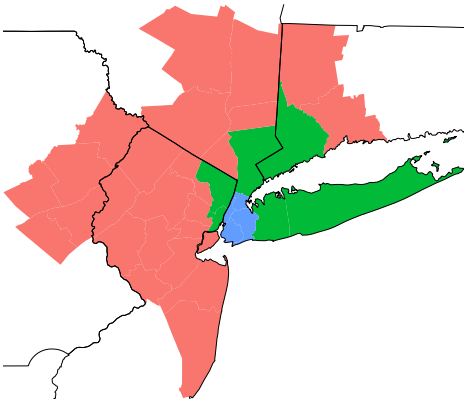
The top left panel of Figure 5.3 plots the MAP assignment for each county in the CSA. As in Section 2, individuals migrate from the red counties to the green counties and from the green counties to the blue counties. The IBM well-captures urbanization: the blue counties are the four major boroughs of New York City, the urban center of the CSA. The green reflect the suburban areas that commute to New York City, and the remaining are satellite cities and rural areas.

While we cannot observe the infinitesimal generator, the net-migration matrix, A, is a reasonable check. We fit a Poisson SBM and plot the clusters assignments in the top right panel of Figure 5.3. The top two plots agree on 30/35 assignments.

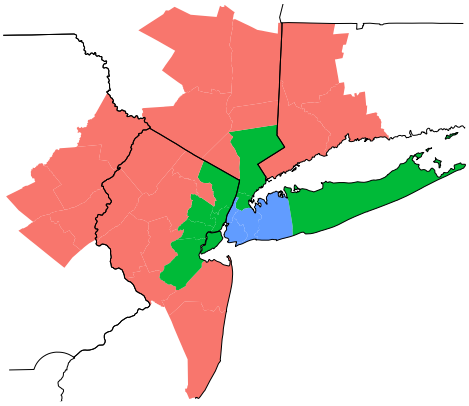
We then compare the IBM to an algorithm that clusters on each trajectory, y_i, directly. We fit a $K = 3$ hard-mixture of Poisson regressions with a log link (M-GLM) using the EM algorithm. We run the algorithm one thousand times and choose the cluster assignments that maximize the likelihood. The bottom left panel of Figure 5.3 plots the results. M-GLM agrees with SBM on 27/35 assignments, picking up some of the structure caused by the migration patterns without explicitly modeling migration. The confusion matrix is shown in the bottom right panel.

5.4.3 ACS Data

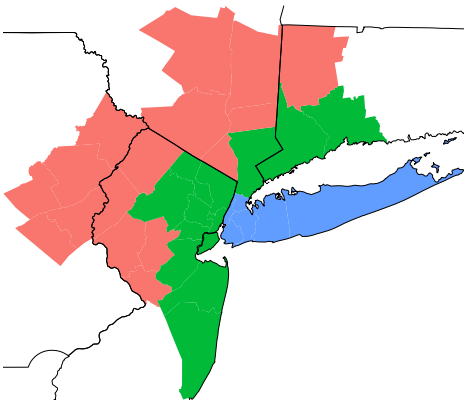
The IBM explains nearly all the variation in a real dataset, and its predictions are competitive with the predictive but less interpretable Lee-Carter matrix factorization model.



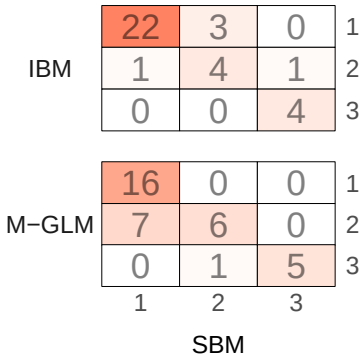
(a) Infinitesimal Block Model (IBM)



(b) Stochastic Block Model (SBM)



(c) Mixed Generalized Linear Model (M-GLM)



(d) Confusion Matrix $n = 35$

Figure 5.3: A comparison of the Infinitesimal Block Model (IBM) and a (hard) mixture of Poisson regressions (M-GLM) to a Poisson Stochastic Block Model (SBM). IBM and M-GLM are fit to longitudinal data, while SBM is fit to the underlying migration data. The first three panels show the cluster assignments from each algorithm. Assuming SBM found the "correct" assignments, the bottom-right panel shows the confusion matrix: IBM correctly identified 86 percent of counties; M-GLM 77 percent.

We use data from the U.S. Census Bureau's American Community Survey, perhaps the most common source of U.S. population statistics. Instead of 35 counties observed over ten years, the ACS reports the CSA population for the more granular $n = 196$ Public Use Microdata Areas (PUMAs) over $T = 6$ years (2012-2017). We use the first four years as training and evaluate predictive performance on the last two years. Thus, $[y_i(t)]$ is the 196×4 matrix of PUMA populations, and $[y_i(t^{\text{new}})]$ the 196×2 matrix of PUMA populations to be predicted.

We run Algorithm 1 with $y_i(t)$ for one thousand chains of one thousand iterations each, first with $K = 3$ and then with $K = 20$. We retain the MAP infinitesimal generator:

$$\hat{Q}_{\text{MAP}} = \hat{Z}_{\text{MAP}} \hat{B}_{\text{MAP}} \hat{Z}_{\text{MAP}}^T - \text{diag}(\hat{Z}_{\text{MAP}} \hat{B}_{\text{MAP}} \hat{Z}_{\text{MAP}}^T \mathbf{1}),$$

and make predictions

$$\hat{y}(t^{\text{new}}) = e^{\hat{Q}_{\text{MAP}} t^{\text{new}}} y(0)$$

Figure 4 summarizes the quality of the predictions using a pseudo R^2 that normalizes mean squared error. For a set of locations I and time periods T, we define

$$R^2(T) = 1 - \frac{\sum_T \sum_I (y_i(t) - \hat{y}_i(t))^2}{\sum_T \sum_I (y_i(t) - \bar{y}_i(t))^2}.$$

The pseudo R^2 is loosely interpreted as the amount of variation explained. Larger values are better. In sample—i.e. for $T = \{2012, 2013, 2014, 2015\}$ and $I = \{1, \ldots, 196\}$—IBM with $K = 3$ (20) explains 98 (99) percent of the variation. Out of sample—i.e. for $T = \{2016, 2017\}$ and $I = \{1, \ldots, 196\}$—IBM explains roughly 94 (93) percent of the variation. IBM beats a (hard) mixture of Poisson regressions (M-GLM) by a considerable margin, and its performance is comparable to the Lee-Carter model (SVD).

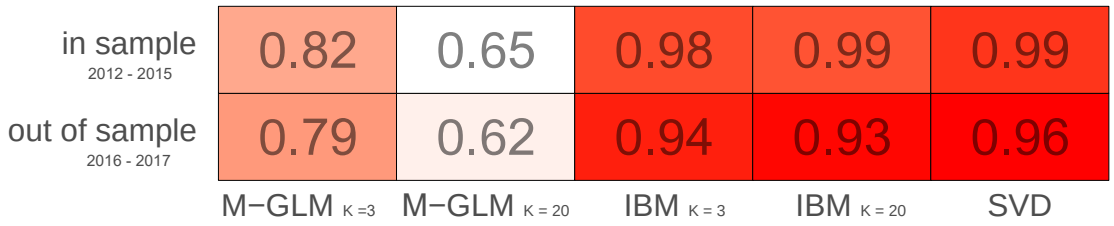


Figure 5.4: A comparison of the Infinitesimal Block Model (IBM) and a (hard) mixture of Poisson regressions (M-GLM) to the Lee-Carter model (SVD). Normalized mean squared error ("pseudo" R^2) is reported for in-sample predictions (top) and out-of-sample predictions (bottom). IBM and M-GLM are fit with both $K = 3$ and $K = 20$ clusters. IBM predictions are comparable with SVD.

5.5 Extension

Our main example is a problem well suited to the IBM: estimating the number of crimes prevented by the New York Police Department (NYPD) practice of Stop, Question, and Frisk (SQF). SQF refers to the temporary detention, interrogation, and search of pedestrians for illegal contraband like guns and drugs. Proponents argue enforcing lower-level crimes such as these prevents violent offenses—citing New York City's reduction in crime following the increasing use of SQF. Critics argue the reduction in crime is better explained by gentrification—victims and/or perpetrators were gradually displaced from New York City. See Zimring (2011) for details.

We estimate the relationship between violent crime and SQF stops using two publicly available datasets: the the NYPD Stop, Question, and Frisk database and the number of crimes reported to the NYPD. Crime is limited to the number of seven major felony offenses: murder, rape, robbery, assault, burglary, larceny, larceny (motor vehicle); Misdemeanors and violations are excluded.

To motivate our application of IBM, consider Figure 5.5. The top half of Figure 5.5 plots the total number of crimes in New York City for the years 2002 to 2012. The bottom half plots the number of stops under SQF. Proponents of SQF might quantify the number of crimes prevented by SQF by specifying the simple relationship:

$$y'(t) = -\alpha x(t),$$

where $y(t) = [y_i(t)]$ is a 76×1 vector containing the number of crimes in each New York City precinct at time t, $t \in \{1, \ldots, 10\}$ and $x(t) = [x_i(t)]$ is a 76×1 containing the number of stops. α is a positive scalar, representing the reduction in the instantaneous crime rate from an additional stop. The relationship captures the belief that one stop can prevent multiple future crimes.

The equation can be solved

$$y(t) = y(0) - \alpha \int_0^t x(s)ds,$$

where the integral is simply the cumulative number of stops. We assume crime is measured with error, yielding the model

$$y(t) \sim \text{Poisson}(y(0) - \alpha \int_0^t x(s)ds).$$

Since α is one-dimensional, the posterior is calculated sufficiently accurately by grid approximation, and the number of crimes prevented by stops is estimated to be 120,000. A 95 percent uncertainty interval is 119,000 to 122,000.

However, this simple relationship does not capture the rapid gentrification that occurred in New York City between 2002 and 2012. We now expanded to allow crime to "migrate" between precincts. Crime might migrate between precincts because of gentrification; The composition of precincts changed drastically over the 2002-2012 decade, and the declining number of crimes in Figure 5.5 could be explained by victims and/or perpetrators moving through and out of New York City.

We apply IBM to the previous model by defining the inhomogenous linear differential equations,

$$y(t) = Qy(t) - \alpha x(t),$$

where Q is as defined in Section 3, except the system is open: crime migrates into cluster one locations and out of cluster K locations at unknown rates B_{10} and $B_{(K+1)K}$. The solution to this equation is

$$y(t) = e^{Qt}y(0) - \alpha \int_0^t e^{Q(t-s)}x(s)ds$$

and implies the likelihood

$$y(t) \sim \text{Poisson}(e^{Qt}y(0) - \alpha \int_0^t e^{Q(t-s)}x(s)ds).$$

We run a modified version of Algorithm 1, alternating between sampling α and Q using $y_i(t)$, $x_i(t)$, and K= 3. We sample one thousand chains for one thousand iterations each and retain the maximum a posteriori (MAP) cluster assignment, $\hat{Z}_{\text{MAP}}$, which is plotted in Figure 5.6. We cluster on the trajectories, y_i, to initialize the Algorithm.

The number of crimes prevented by stops is estimated to be 460. A 95 percent uncertainty interval is 440 to 480. Migration therefore explains $1 - 460/120000 = 99$ percent of the crime reduction between 2002 and 2012.

5.6 Conclusion

We proposed a model that identifies latent migration patterns from longitudinal data. These patterns are of considerable interest in business, government, and the social and biological sciences, where migration data is difficult to obtain. We outlined an algorithm for fitting the model and demonstrated the result is flexible, interpretable, and predictive. We then applied the algorithm to estimate the number of crimes prevented by the NYPD practice Stop, Question, and Frisk, decomposing the instantaneous crime rate into a migration effect and a stop effect.

The proposed model has other straightforward extensions not considered here. For example, the initial state, $y(0)$, could be treated as unobserved and estimated as a parameter. For long time series, the homogeneity assumption may only be appropriate over short time intervals. Multiple

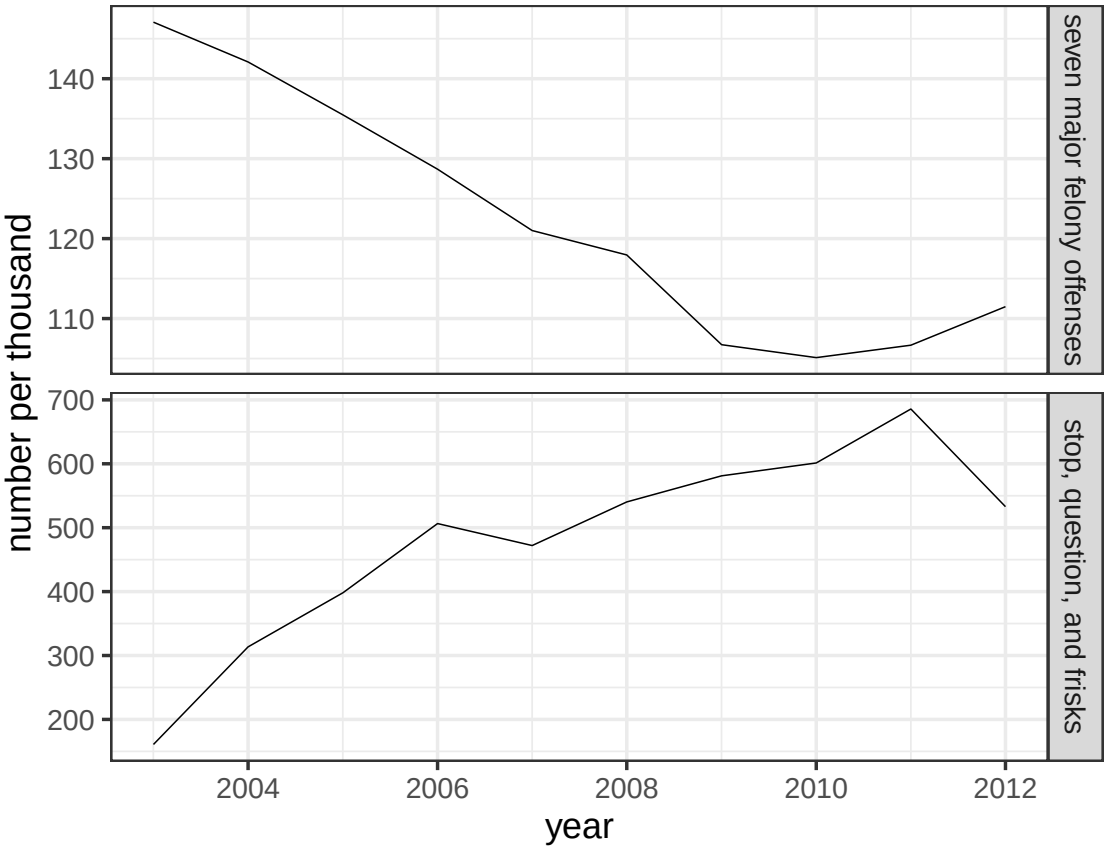


Figure 5.5: The annual number of major felony offenses (crimes, top) and stops under the Stop, Question, and Frisk policy (stops, bottom) in New York City between 2002 and 2012. Crimes and stops appear negatively correlated, suggesting to policymakers that SQF prevents crime. The relationship at the precinct level is similar—albeit weaker.

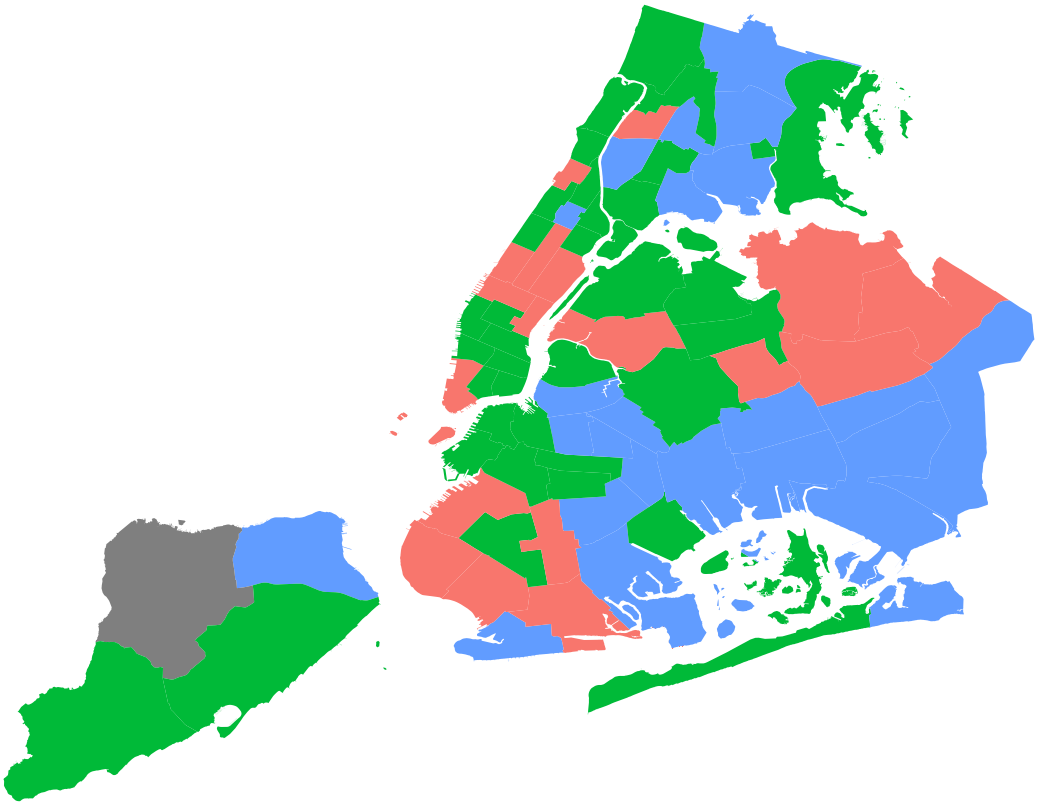


Figure 5.6: The MAP cluster of New York City precincts identified by the extended Infinitesimal Block Model (IBM) with K = 3. Between 2002 and 2012 crime "migrated" from red precincts to blue precincts to green precincts.

IBMs can be combined piecewise: The series could be divided into four or five year increments, and a different IBM fit to the data in each interval. Each IBM could assign the same unit to the same cluster, or the same unit could be assigned to different but correlated clusters.

The fit of the IBM at the end of the interval could form the initial state, $y(0)$, of the following IBM. Sometimes a longitudinal survey coincides with an infrequent, but more accurate census. For example, the annual American Community Survey coincides with the U.S. decennial Census. In such cases, the Census could serve as the initial state of each interval.

5.7 Appendix

5.7.1 Inhomogenous Extension

Consider the inhomogeneous linear differential equations,

$$\begin{aligned}
\frac{dy_1}{dt} &= q_{11}y_1 + q_{12}y_2 + \cdots + q_{1n}y_n + f_1 \\
\frac{dy_2}{dt} &= q_{21}y_1 + q_{22}y_2 + \cdots + q_{2n}y_n + f_2 \\
&\vdots \\
\frac{dy_n}{dt} &= q_{n1}y_1 + q_{n2}y_2 + \cdots + q_{nn}y_n + f_n,
\end{aligned}$$

whose solution is

$$y(t) = e^{Qt}y(0) + \int_0^t e^{Q(t-s)}f(s)ds$$

where $f(t) = [f_i(t)]$ is an $n \times 1$ vector of counts, whose ith element is denoted f_i at time t. e^{Qt} is the matrix exponential of Qt.

As before, we interpret the coefficients q_{ij} as the instantaneous rate at which individuals migrate from location i to location j. However in addition, we have a time-varying filter, $f(t)$: $f_i(t)$ denotes the number of individuals removed from location i at time t. The integral can be evaluated empirically or $f(t)$ can be approximated by a polynomial in t,

$$f(t) = a_0 = a_1 t + a_2 t^2 + \ldots + a_p s^p.$$

For example, for $p = 2$, integration by parts yields

$$y(t) = e^{Qt}y(0) + e^{Qt}y(0) + Q^{-1}(e^{Qt} - I)a_0 + Q^{-2}(e^{Qt} - Qt - I)a_1 + Q^{-3}(2e^{Qt} - Q^2t^2 - 2Qt - 2I)a_2$$

Function of polynomials, such as splines, can also be easily accommodated in this way.

5.7.2 Alternative Update Schedule

Algorithm 2: Z Update

for $i = 1$ to n **do**
  $V \sim \text{uniform}(0, 1)$.
  if $V < \beta$ **then**
    propose cluster to swap, k^*
    propose z_j from cluster k^*
    Swap cluster assignments of z_i and z_j
  else
    propose z_i^*
    $U \sim \text{uniform}(0, 1)$
    if $U < \frac{p\left(z_i^* \middle| b^{(t)}, z_{1:(i-1)}^{(t)}, z_{(i+1):n}^{(t-1)}, y\right)}{p\left(z_i^{(t-1)} \middle| b^{(t)}, z_{1:(i-1)}^{(t)}, z_{(i+1):n}^{(t-1)}, y\right)}$ **then**
      $z_i^{(t)} \leftarrow z_i^*$
    end if
  end if
end for

5.7.3 Algorithm 1 Acceptance Probability

For ease of presentation, we suppress the iteration number of the sampler, t, in this section. Let Z be the current matrix of cluster assignments and Z^* be the proposed matrix of cluster assignments, where the i-th row z_i is replaced with proposal z_i^*. Similarly, let B^* denote the proposed

block matrix, where subdiagonal element $b_{j,j-1}$ is replaced with $b^*_{j,j-1}$. Define

$$D = \text{diag}(ZBZ^T\mathbf{1}) \quad D^* = \text{diag}(Z^*BZ^{*T}\mathbf{1}) \quad D^*_B = \text{diag}(ZB^*Z^T\mathbf{1}).$$

Then the posterior conditional is proportional to

$$\begin{aligned}
p(z_i|b, z_{-i}, y) &\propto p(y|z, b)p(z, b) \\
&= \prod_{i,t} \text{Poisson}(y_i(t)|[e^{(ZBZ^T-D)t}y(0)]_i) \prod_i \text{Multinomial}(z_i|1, \pi) \prod_k \text{Gamma}(b_{k,k-1}|\phi, \psi) \\
&\propto \prod_{i,t} \exp\left\{-[e^{(ZBZ^T-D)t}y(0)]_i\right\} [e^{(ZBZ^T-D)t}y(0)]_i^{y_i(t)} \prod_{i,k} \pi_{ik}^{z_{ik}} \prod_k b_{k,k-1}^{\theta-1} e^{-\phi b_{k,k-1}} \\
&\propto \prod_{i,t} \exp\left\{-\sum_j [e^{(ZBZ^T-D)t}]_{ij} y(0)^j\right\} \left[\sum_j [e^{(ZBZ^T-D)t}]_{ij} y(0)_j\right]^{y_i(t)} \prod_{i,k} \pi_{ik}^{z_{ik}} \\
&\quad \times \prod_k b_{k,k-1}^{\theta-1} e^{-\phi b_{k,k-1}} \\
&\propto \prod_{i,t,j} \exp\left\{-[e^{(ZBZ^T-D)t}]_{ij} y(0)_j\right\} \left[\sum_j [e^{(ZBZ^T-D)t}]_{ij} y(0)_j\right]^{y_i(t)} \prod_{i,k} \pi_{ik}^{z_{ik}} \\
&\quad \times \prod_k b_{k,k-1}^{\theta-1} e^{-\phi b_{k,k-1}},
\end{aligned}$$

and the rejection probability is

$$\begin{aligned}
\frac{p(z_i^*|b, z_{-i}, y)}{p(z_i|b, z_{-i}, y)} &= \frac{\prod_{i,t,j} \exp\left\{-[e^{(Z^*BZ^{*T}-D^*)t}]_{ij} y(0)_j\right\} \left[\sum_j [e^{(Z^*BZ^{*T}-D^*)t}]_{ij} y(0)_j\right]^{y_i(t)}}{\prod_{i,t,j} \exp\left\{-[e^{(ZBZ^T-D)t}]_{ij} y(0)_j\right\} \left[\sum_j [e^{(ZBZ^T-D)t}]_{ij} y(0)_j\right]^{y_i(t)}} \\
&\quad \times \frac{\prod_k \pi_{ik}^{z_{ik}^*}}{\prod_k \pi_{ik}^{z_{ik}}}.
\end{aligned}$$

Similarly,

$$\frac{p(b^*_{j,j-1}|b_{-(j,j-1)},z,y)}{p(b_{j,j-1}|b_{-(j,j-1)},z,y)} = \frac{\prod_{i,t,j}\exp\left\{-[e^{(ZB^*Z^T-D^*_B)t}]_{ij}y(0)_j\right\}\left[\sum_j[e^{(ZB^*Z^T-D^*_B)t}]_{ij}y(0)_j\right]^{y_i(t)}}{\prod_{i,t,j}\exp\left\{-[e^{(ZBZ^T-D)t}]_{ij}y(0)_j\right\}\left[\sum_j[e^{(ZBZ^T-D)t}]_{ij}y(0)_j\right]^{y_i(t)}} \times \frac{\prod_k b^{*\theta-1}_{j,j-1}e^{-\phi b^*_{j,j-1}}}{\prod_k b^{\theta-1}_{j,j-1}e^{-\phi b_{j,j-1}}}.$$

Note we take $\frac{p(b^*_{j,j-1}|b_{-(j,j-1)},z,y)}{p(b_{j,j-1}|b_{-(j,j-1)})} = 0$ whenever $b^*_{j,j-1}$ does not satisfy the constraint: $0 < \cdots < b_{j-2,j-1} < b^*_{j-1,j} < \cdots < b_{K-1,K}$.

For the sampler to remain efficient when the number of clusters is high, we reparameterize to the unconstrained space:

$$\eta_1 = \log b_{21},\ \ \eta_2 = \log\left(b_{32} - b_{21}\right),\ \ \ldots\ ,\ \ \eta_{K-1} = \log\left(b_{K(K-1)} - b_{(K-1)(K-2)}\right).$$

We require the following adjustment to the rejection probability. Let f be the function such that $f(b) = \eta$.

$$\begin{aligned} p(\eta_j|\eta_{-j},z,y) &\propto p(y|\eta,z)p(\eta)p(z) \\ &= p(y|\eta,z)p(z)p(f^{-1}(\eta))|Jf^{-1}(\eta)|. \end{aligned}$$

Let H be the matrix B with each $b_{j+1,j}$ replaced with $\eta_j = f(b_{j+1,j})$ and D_H be diagonal matrix populated with the negative column sums of H.

$$\begin{aligned} \frac{p(\eta^*_j|\eta_{-j},z,y)}{p(\eta_j|\eta_{-j},z,y)} = &\frac{\prod_{i,t,j}\exp\left\{-[e^{(ZH^*Z^T-D_{H^*})t}]_{ij}y(0)_j\right\}\left[\sum_j[e^{(ZH^*Z^T-D_{H^*})t}]_{ij}y(0)_j\right]^{y_i(t)}}{\prod_{i,t,j}\exp\left\{-[e^{(ZHZ^T-D_H)t}]_{ij}y(0)_j\right\}\left[\sum_j[e^{(ZHZ^T-D_H)t}]_{ij}y(0)_j\right]^{y_i(t)}} \\ &\times \frac{((e^{\eta^*_j}+e^{\eta_{j-1}})(e^{\eta^*_j}+e^{\eta_{j+1}}))^{\theta-1}\exp(-\phi(e^{\eta^*_j}+e^{\eta_{j-1}})-\phi(e^{\eta^*_j}+e^{\eta_{j+1}}))}{((e^{\eta_j}+e^{\eta_{j-1}})(e^{\eta_j}+e^{\eta_{j+1}}))^{\theta-1}\exp(-\phi(e^{\eta_j}+e^{\eta_{j-1}})-\phi(e^{\eta_j}+e^{\eta_{j+1}}))} \\ &\times \frac{e^{\eta^*_j}}{e^{\eta_j}}. \end{aligned}$$

5.7.4 High dimensional approximation

When the dimension is to large for rejection sampling, we recommend the following approximation, which requires the following notation.

1. $y_i(t)$ is the population of location $i \in \{1 \dots n\}$ at time $t \in \{1 \dots T\}$, and $y = [y_i(t)]$ is an $n \times T$ matrix of counts, whose ith row is denoted y_i and tth column $y(t)$. $\mu_i(t) := E(y_i(t))$ and $x := y(0)$.

2. $k_i \in \{1, \dots, K\}$ is the cluster assignment of location i, Z is an $n \times K$ binary matrix of cluster assignments, i.e. $Z_{ij} = \delta_{j=k_i}$, and $n_{k_i} = \sum_{j=1}^{n} Z_{jk_i}$ is the number of locations in cluster k_i.

3. $B_{ij} \geq 0$ is the instantaneous migration rate from a location in cluster $j \in \{1, \dots, K-1\}$ to a location in cluster $i \in \{2, \dots, K\}$, and $B = [B_{ij}]$ is a $K \times K$ migration matrix. Individuals migrate across clusters sequentially: $B_{ij} > 0$ only if $i = j + 1$. The system is open: $B_{01}, B_{(K+1)K} > 0$ where 0 and $K + 1$ denote locations outside of the system.

4. $m_{k_i} = n_{(k_i+1)} B_{(k_i+1)k_i}$ is the emigration rate $B_{(k_i+1)k_i}$ from location i in cluster k_i to all $n_{(k_i+1)}$ locations in cluster $k_i + 1$, and $m = [m_{k_i}]$ is the $n \times 1$ vector of emigration rates. The emigration rate from a location in cluster 0 is $m_0 = nB_{01}$, and from cluster K, $m_K = nB_{(K+1)K}$. In addition, define

$$r_{k_i k_j} = \begin{cases} 1 & \text{if } k_i = k_j \\ \prod_{l=1}^{k_i - k_j} \dfrac{m_{(k_j+l-1)}}{m_{(k_j+l)} - m_{k_j}} & \text{if } k_i > k_j \end{cases}$$

$$r^{-1}_{k_j k_p} = \begin{cases} 1 & \text{if } k_j = k_p \\ \prod_{l=1}^{k_j - k_p} \dfrac{m_{(k_j-l)}}{m_{(k_j-l)} - m_{k_j}} & \text{if } k_p < k_j \end{cases}$$

5. $Q = ZBZ^T - \text{diag}(m)$ is an $n \times n$ instantaneous migration matrix. The diagonal of Q is the total emigration rates. The rows of the off-diagonal are the immigration rates to location i from each location in cluster $k_i - 1$, and the columns of Q are the emigration rates from location i to each location in cluster $k_i + 1$.

   Assume the entries of m, $m_1, \ldots, m_k$, are nonzero and distinct. Define $n \times n$ matrices Λ and V:

$$\Lambda_{ij} = \begin{cases} Q_{jj} = -m_{k_j} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$V_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \text{ and } k_i \leq k_j \\ \frac{r_{k_i k_j}}{n_{k_i}} & \text{if } k_i > k_j \end{cases}$$

   Then $V\Lambda V^{-1}$ is a orthogonal (though not orthonormal) eigendecomposition of Q where V^{-1} has entries

$$V^{-1}_{jp} = \begin{cases} 1 & \text{if } p = j \\ 0 & \text{if } p \neq j \text{ and } k_p \geq k_j \\ \frac{r^{-1}_{k_j k_p}}{n_{k_j}} & \text{if } k_p < k_j \end{cases}$$

   If k_i is monotonically increasing, Q is lower triangular. Otherwise, Q can be premultiplied by a set of row-switching transformations so that it is lower triangular. In either case, the diagonal of Λ is the diagonal of Q.

   V can then be found by forward substitution: Fix the jth column of V, v_j. The identity $(Q - \lambda_j I)v_j = (Q + m_{k_j} I)v_j = 0$ is the system of linear equations

$$V_{ij}(m_{k_j} - m_{k_i}) + \sum_{i':k_{i'}=k_i-1} \frac{m_{k_{i-1}}}{n_{k_i}} V_{i'j} = 0$$

Since the entries of m are distinct, V_{ij} must be 0 if $k_i < k_j$. Setting $V_{jj} = 1$ and $V_{ij} = 0$ for $k_i = k_j$, the remaining elements of v_j are obtained recursively from $i : k_i = k_j + 1, \ldots, K$.

$$V_{ij} = \frac{1}{n_{k_i}} \frac{m_{k_{i-1}}}{(m_{k_i} - m_{k_j})} \sum_{i':k_{i'}=k_i-1} V_{i'j}$$

V^{-1} can be similarly found by backward substitution: Fix the jth row of V^{-1}, v_j^{-1}. The identity $v_j^{-1}(Q + m_{k_j} I) = 0$ is the system of linear equations

$$V_{jp}^{-1}(m_{k_j} - m_{k_p}) + \sum_{p':k_{p'}=k_p+1} \frac{m_{k_p}}{n_{k_p-1}} V_{jp}^{-1} = 0$$

This time V_{jp}^{-1} is 0 if $k_p > k_j$. Setting $V_{jj} = 1$ and $V_{jp} = 0$ for $k_p = k_j$, the remaining elements of v_j are obtained recursively from $p : k_p = k_j - 1, \ldots, 0$.

$$V_{jp}^{-1} = \frac{1}{n_{k_p-1}} \frac{m_{k_p}}{(m_{k_p} - m_{k_j})} \sum_{p':k_{p'}=k_p+1} V_{jp'}$$

Assume $[n_1, \ldots, n_K] \sim \text{Multinomial}(n; [\alpha_1, \ldots, \alpha_K])$ and $\sum_{i:k_i=k} \frac{x_i}{n_{k_i}} \underset{n\to\infty}{\longrightarrow} \bar{x}_{k_i}$ so that

1. $\frac{m_{k_i}}{n} \underset{n\to\infty}{\longrightarrow} \bar{m}_{k_i} := \alpha_{(k_i+1)} B_{(k_i+1)k_i}$
2. $r_{k_i k_j} \underset{n\to\infty}{\longrightarrow} \bar{r}_{k_i k_j} := \prod_{l=1}^{k_i-k_j} \frac{\bar{m}_{(k_j+l-1)}}{\bar{m}_{(k_j+l)} - \bar{m}_{k_j}}$
3. $r_{k_j k_p}^{-1} \underset{n\to\infty}{\longrightarrow} \bar{r}_{k_j k_p}^{-1} := \prod_{l=1}^{k_j-k_p} \frac{\bar{m}_{(k_j-l)}}{\bar{m}_{(k_j-l)} - \bar{m}_{k_j}}$
4. $$-\sum_{j:k_j=k} \sum_{p=1}^{n} V_{ij} V_{jp}^{-1} \frac{m_0}{m_{k_j}} \delta_{k_p=1} \underset{n\to\infty}{\longrightarrow} \begin{cases} 0 & \text{if } k_i < k_j \\ \bar{d}_{k_i k_j} := \frac{\alpha_1}{\alpha_i} \bar{r}_{k_i k_j} \frac{\bar{m}_0}{\bar{m}_{k_j}} \bar{r}_{k_j 1}^{-1} & \text{if } k_i \geq k_j \end{cases}$$

5.

$$\sum_{j:k_j=k} V_{ij} \sum_{p=1}^{n} V_{jp}^{-1} x_p \underset{n\to\infty}{\longrightarrow} \begin{cases} 0 & \text{if } k_i < k_j \\ \bar{c}_i := x_i + \sum_{k_p=1}^{k_i-1} \frac{\alpha_p}{\alpha_i} \bar{r}_{k_i k_p}^{-1} \bar{x}_{k_p} & \text{if } k_i = k_j \\ \bar{c}_{k_i k_j} := \frac{\alpha_j}{\alpha_i} \bar{r}_{k_i k_j} (\bar{x}_{k_j} + \sum_{k_p=1}^{k_j-1} \frac{\alpha_p}{\alpha_j} \bar{r}_{k_j k_p}^{-1} \bar{x}_{k_p}) & \text{if } k_i > k_j \end{cases}$$

Combining 1-6 yields the following large sample approximation,

$$\mu_i(t) \approx e^{-m_{k_i} t}(\bar{c}_i - \bar{d}_{k_i k_i}) + \bar{d}_{k_i k_i} + \sum_{k_j=1}^{k_i-1} [e^{-m_{k_j} t}(\bar{c}_{k_i k_j} - \bar{d}_{k_i k_i}) + \bar{d}_{k_i k_i}]$$

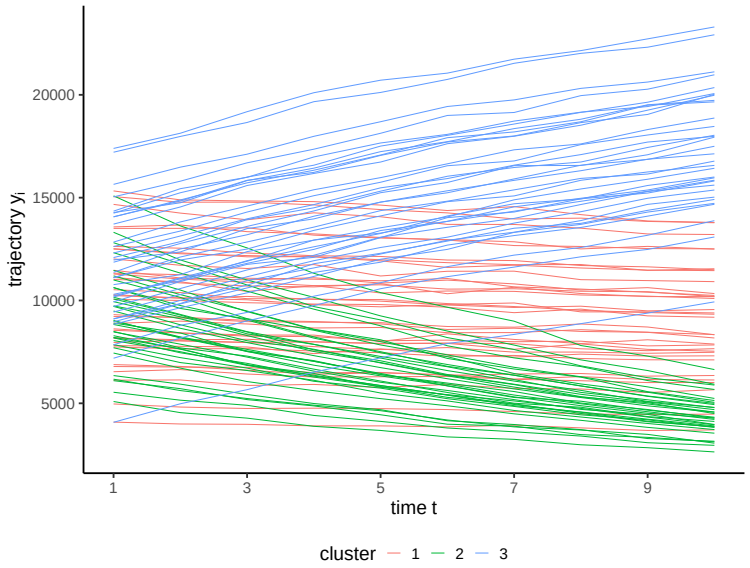
5.7.5 Simulation

We demonstrate Algorithm 1 can recover the infinitesimal generator Q by simulation. We generate longitudinal data from the IBM with $K = 3$ clusters and $n = 100$ locations. One such simulation is shown in Figure 5.7. The top panel shows the trajectory of all 100 locations, colored according to their cluster assignment. Individuals move from green locations to red at instantaneous rate $b_{21} = 3e{-}04$ and from red locations to blue locations at rate $b_{32} = 3e{-}03$.

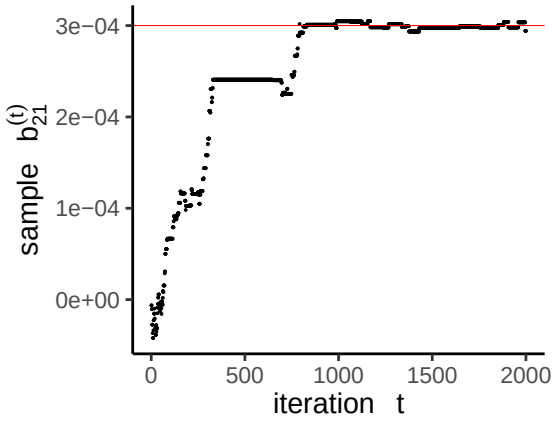
We run one chain of Algorithm 1 for two thousand iterations. Cluster assignments are initalized randomly and $b_{21}^{(0)} = b_{32}^{(0)} = 0$. The algorithm correctly identifies all 100 assignments and the true B. Trace plots show convergence for b_{21} (left) and b_{32} (right) after around one thousand iterations. The second thousand iterations are thinned to produce one hundred samples from the posterior. The red horizontal line marks the true data-generating value.

5.7.6 Real Data

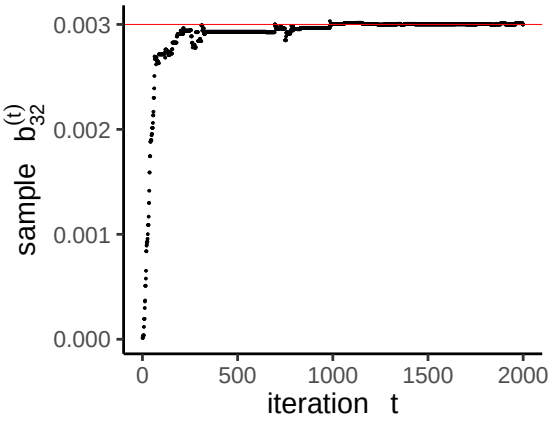
This paper identifies migration patterns in the New York–Newark, NY–NJ–CT–PA Combined Statistical Area (CSA) between 2006 and 2017. The CSA includes the five counties of New York City, and its surrounding suburban and rural communities. The counties are highly connected,



(a) Data



(b) Trajectory 1



(c) Trajectory 2

Figure 5.7: Results from an IBM simulation study. The trajectory of 100 locations is colored by cluster assignment (top). Individuals move from green locations to red at instantaneous rate $b_{21} = 3e{-}04$ and from red locations to blue locations at rate $b_{32} = 3e{-}03$. Trace plots show convergence for b_{21} (left) and b_{32} (right) after around one thousand iterations. The red horizontal line marks the true data-generating value.

forming the largest urban agglomeration of economies in the Americas. Two datasets are used, one from the U.S. Internal Revenue Service (IRS) and one from the U.S. Census Bureau (Census).

The Statistics of Income Division (SOI) of the U.S. Internal Revenue Service (IRS) maintains a database of all income tax forms filed. They tabulate the number of county-to-county migrations of tax filers and their dependents each year, and release the data to the public.

The data have two notable limitations. First, non-filing and late-filing households are not represented in all years so the data under-represent the very poor, the very wealthy, and the elderly (Gross, 2003). Second, the methodology changed starting with the 2011-2012 data (Pierce, 2015). For the purpose of estimating major migration flows, however, these limitations are not particularly consequential.

The American Community Survey has been conducted monthly by the U.S. Census Bureau since 2005. Roughly 1 percent of the U.S. population is sampled each year, and participation is required by law. We use the "ACS-1-year" population estimates from this survey. Public Use Micro Areas (PUMAs) is the smallest area for which year-over-year data is publicly available for all locations. PUMAs contain around 100,000 residents, conceptually the size of a neighborhood. The PUMA boundaries were redrawn in 2012 according to the 2010 decennial Census. We confine our study to 2012-2017, however, the model can be adjusted for a longer time frame.

5.8 References

Aicher, Christopher, Abigail Z Jacobs, and Aaron Clauset. 2014. " Learning latent block structure in weighted networks." Journalof Complex Networks 3(2):221–248.

Airoldi, Edoardo, David Blei, Elena A Erosheva, and Stephen E Fienberg. 2014. Handbook of mixed membership modelsand their applications. CRC press.

Bickel, Peter J, and Aiyou Chen. 2009. "A nonparametric view of network models and newman–girvan and other modularities." Proceedings of the National Academy of Sciences 106(50):21068–21073.

Buchholz, Peter, Jan Kriege, and Iryna Felko. 2014. "Input modeling with phase-type distri-

butions and Markov models: theory and applications." Springer, 2014.

Carmichael, Gordon A. 2016. "Fundamentals of demographic analysis: Concepts, measures and methods." Springer, 2016.

Diaconis, Persi and Svante Janson. 2016. "Graph limits and exchangeable random graphs." arXiv preprint arXiv:0712.2749.

Gelfand, Alan E and Adrian FM Smith. 1990. "Sampling-based approaches to calculating marginal densities." Journal of the American statistical association 85(410):398–409.

Geman, Stuart and Donald Geman. 1987. "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images." In Readings in computer vision, pages 564–584. Elsevier.

Gross, Emily. 2003. "US population migration data: Strengths and limitations." Internal Revenue Service Statistics of IncomeDivision, Washington, DC. http://www. irs. gov/pub/irs-soi/99gross$_update$.

Hannah, Lauren A, David M Blei, and Warren B Powell. 2011. "Dirichlet process mixtures of generalized linear models." Journal of Machine Learning Research 12(Jun):1923–1953.

Holland, Paul W, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. "Stochastic blockmodels: first steps." Social networks 5(2):109–137.

Kanjilal, Partha Pratim, Sarbani Palit, and Goutam Saha. 1997. "Fetal ecg xtraction from single-channel maternal ecg using singular value decomposition." IEEE Transactions on Biomedical Engineering 44(1):51–59.

Lawler, Gregory F. 2018. "Introduction to stochastic processes." Chapman and Hall/CRC.

Lecci, Fabrizio. 2014. "An analysis of development of dementia through the extended trajectory grade of membership model." Handbook of mixed membership models. Chapman Hall, pages 189–200.

Lee, Ronald. 2000. "The lee-carter method for forecasting mortality, with various extensions and applications." North American actuarial journal 4(1):80–91.

Lee, Ronald D and Lawrence R Carter. 1992. "Modeling and forecasting us mortality." Journal

of the American statistical association 87(419):659–671.

Leger, Jean-Benoist 2016. "Blockmodels: A r-package for estimating in latent block model and stochastic block model, withvarious probability functions, with or without covariates." arXiv preprint arXiv:1602.07587.

Leisch, Friedrich. 2004. "FlexMix: A general framework for finite mixture models and latent class regression in R." Journal of Statistical Software 11(8):1–18.

Leonard, IE. 1996. "The matrix exponential." SIAM review 38(3):507–512.

Lin, Haiqun, Bruce W Turnbull, Charles E McCulloch, and Elizabeth H Slate. 2002. "Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer." Journal of the American statistical association 97(457):53–65.

Manrique-Vallier, Daniel. 2014. "Longitudinal mixed membership trajectory models for disability survey data." The annals of applied statistics 8(4):2268.

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. 1953. " Equation of state calculations by fast computing machines." The journal of chemical physics 21(6):1087–1092.

Pedroza, Claudia. 2006. "A bayesian forecasting model: predicting us male mortality." Biostatistics 7(4):530–550.

Pierce, Kevin. 2015. "Soi migration data: a new approach: Methodological improvements for soi's united states population migration data, calendar years 2011-2012." Statistics of Income. SOI Bulletin 35(1).

R Core Team.R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.16

Raftery, Adrian E, Leontine Alkema, and Patrick Gerland. 2014. "Bayesian population projections for the united nations." Statistical science: a review. 29(1):58.

Villegas, Andrés, Vladimir K Kaishev, and Pietro Millossovich. 2015. "Stmomo: An r package

for stochastic mortality modelling."

Liang Xiong, Xi Chen, Tzu-Kuo Huang, Jeff Schneider, and Jaime G Carbonell. 2010. "Temporal collaborative filtering withbayesian probabilistic tensor factorization." In Proceedings of the 2010 SIAM International Conference on Data Mining. 211–222.

Zimring, Franklin E. 2011. "The city that became safe: New York's lessons for urban crime and its control." Oxford University Press.

Chapter 6: Forecasting Pedestrian Fatalities with the Linear Probability Model

with Shaw-Hwa Lo

The linear probability model (LPM) approximates the probability of a dichotomous variable with a linear combination of covariates. The model is as old as least squares and remains popular in applied statistics and data science: It is used to approximate nonlinear curves and marginal effects (for example, the test for trend (Armitage 1958), (Cochran 1955), and (Cox 1958, p. 222), and the derivative rule (Berkson 1951), (King 1998, p. 105), and (Gelman 2006, p. 81); it provides an initial solution to iterated algorithms like feasible generalized least squares and iteratively weighted least squares. (Freedman 2009, p. 123); and it is a powerful predictor, serving as a point of comparison to more complex classifiers. Friedman (2001, p. 104) note that "on many problems it [the LPM] gives similar results to more standard linear methods for classification".

The popularity of the LPM is due to its analytic and computational simplicity: It is easy to prove theoretical properties (such as the order of the error when approximating nonlinear curves) and compute the fit (such as via least squares or weighted least squares). Yet despite its popularity, the LPM has been largely dismissed as a stand-alone tool for statistical inference. It seldom appears in inferential tasks such as collapsing contingency tables (or variable selection in general), analysis of variance, or causal inference. This is primarily because setting the coefficients of the LPM to zero does not correspond to an independence assumption and has been heretofore considered uninterpretable (see Cox (1989, p. 118) and Bishop (2007, p. 23-24)). It is because of this limitation that secondary considerations, such as heteroskedasticity and boundary violations (when the model fit falls outside the unit interval), has led researchers to prefer alternative models—even though these limitations are not specific to the LPM and are therefore not themselves

a sufficient basis for its rejection (Cochran 1940).

We reevaluate the role of the LPM for describing contingency tables in three sections. We first demonstrate that the LPM parameters are related to the Bayes error rate, serving the same role for predictivity as log-linear models serve for independence. This allows researchers to assess the predictivity of a covariate as a parametric inference problem. We then demonstrate the relationship between ordinary least squares and the influence score, which have both been demonstrated effective in problems of predictivity. For example, see Battey (2019) on the robustness of least squares for LPMs and Chernoff (2009) and Lo (2016) on prediction with the influence score. The influence score is key to any analysis with the LPM because, unlike traditional measures of model fit such as χ^2 or G^2, it does not fail when the estimated probabilities are close to 0 and 1 or outside the unit interval. Our final section compares the LPM and the log-linear model (LLM) using a data set describing the New York City traffic safety policy, Vision Zero.

6.1 Linear Probability Model for Partitions

Consider a class of partitions, Π, where each partition $\pi \in \Pi$ divides n objects into J groups, $1 \leq J \leq n$. Our goal is to choose the partition that best summarizes a gold standard variable, Y_i. Unlike the traditional classification problem, we assume Y_i is measured without error. Instead, any proposed partition is subject to classification error according to a linear probability model—reflecting the fact that the partitions under consideration were chosen with the gold standard in mind. Though the goal is to predict Y_i on the same or similar units in the future, we find the best partition by calibration, modeling $\Pi|Y$ and solving for Y.

To fix ideas, suppose Y_i measures the risk of a fatality on a randomly chosen traffic intersection between 2011 and 2013, and $X_i = [X_1, \ldots, X_J]$ is a multivariate random vector indicating whether the road segment belongs to group $j = 1, \ldots, J$. We assume that for each intersection, classification is made with error according to a linear probability model; the partition imperfectly separates roads into groups of different risk levels.

In the fist subsection, we consider the joint LPM, and in the second subsection, the conditional

LPM. In both, we suppress the index i for clarity. The results extend to higher dimensional tables (both the $I \times J$ table and K 2×2 tables), as we partially outline in the Appendix, but the single 2×2 table is sufficient for the task of choosing the best, summarizing partition $\pi \in \Pi$ for a single, gold-standard measure Y.

6.1.1 Interpreting the joint LPM

In the remainder of this section, we assume Y is dichotomous so that Y and X_j partition the n observations into four cells, which form the following 2x2 contingency table; see Appendix for details. To avoid overloading the notation in this section, we denote the arbitrary X_j by Z. Also, let p_Y denote $\Pr(Y = 1)$, q_Y denote $\Pr(Y = 0)$, p_Z denote $\Pr(Z = 1)$, q_Z denote $\Pr(Z = 0)$, p_{YZ} denote $\Pr(Y = Z)$, and q_{YZ} denote $\Pr(Y \neq Z)$. It will be convenient to represent the entries of the contingency table with letters:

	Z = 0	Z = 1
Y = 0	a	b
Y = 1	c	d

where

$$a = \Pr(Y = 0, Z = 0)$$

$$b = \Pr(Y = 0, Z = 1)$$

$$c = \Pr(Y = 1, Z = 0)$$

$$d = \Pr(Y = 1, Z = 1)$$

The (saturated) joint LPM is defined as

$$\Pr(Y = i, Z = j) = \alpha + \alpha_Y(i) + \alpha_Z(j) + \alpha_{YZ}(ij)$$

The model is unidentified, and it is common to use the centered parameterization: $\alpha_Y := \alpha_Y(1) = -\alpha_Y(0)$, $\alpha_Z := \alpha_Z(1) = -\alpha_Z(0)$, and $\alpha_{YZ} := \alpha_{YZ}(00) = \alpha_{YZ}(11) = -\alpha_{YZ}(10) = -\alpha_{YZ}(01)$

For a 2x2 table, $i = j = 0, 1$ so there are 9 parameters and 6 constraints for a total of 9 - 6 = 3 "free parameters" to be estimated from 4 observations. It follows from the constraints and the total law of probability that:

$$\alpha = \frac{a+b+c+d}{4} = \frac{1}{4}$$
$$\alpha_Y(1) = -\alpha_Y(0) = \frac{c+d-a-b}{4} = \frac{1}{4}(p_Y - q_Y)$$
$$\alpha_Z(1) = -\alpha_Z(0) = \frac{b+d-a-c}{4} = \frac{1}{4}(p_Z - q_Z)$$
$$\alpha_{YZ}(0,0) = \alpha_{YZ}(1,1) = -\alpha_{YZ}(0,1) = -\alpha_{YZ}(1,0) = \frac{a+d-b-c}{4} = \frac{1}{4}(p_{YZ} - q_{YZ})$$

The constraints only identify the magnitude of the coefficients; an equivalent decomposition is given by $\mu_{ij} = \alpha - \alpha_Y(i) - \alpha_Z(j) - \alpha_{YZ}(ij)$. Let θ_Y be the marginal predictivity of Y, defined as $\theta_Y = p_Y \vee q_Y$, θ_Z the marginal predictivity of Z, defined as $\theta_Z = p_Z \vee q_Z$, and θ_{YZ} the co-predictivity of Y and Z, defined as $\theta_{YZ} = p_{YZ} \vee q_{YZ}$ (see Appendix for details). From Identity I.1: $|p - q| = 2\theta - 1$,

$$|\alpha_Y| = \frac{1}{4}|p_Y - q_Y| = \frac{1}{2}\theta_Y - \frac{1}{4} = \frac{1}{2}\theta_Y - \alpha$$
$$|\alpha_Z| = \frac{1}{4}|p_Z - q_Z| = \frac{1}{2}\theta_Z - \frac{1}{4} = \frac{1}{2}\theta_Z - \alpha$$
$$|\alpha_{YZ}| = \frac{1}{4}|p_{YZ} - q_{YZ}| = \frac{1}{2}\theta_{YZ} - \frac{1}{4} = \frac{1}{2}\theta_{YZ} - \alpha$$

When the LPM coefficient, α_Y, is large, Y is much more predictable than a random guess ($2\alpha = \frac{1}{2}$). When the LPM coefficient is zero, Y is no more predictable than a random guess. In comparison, the parameters of the log-linear model correspond with the odds ratio, which is a relative measure of risk and whose size is uninformative of the marginal or co-predictivity (unless it is zero, see Appendix).

6.1.2 Interpreting the conditional LPM

We retain the setup from the previous subsection, where dichotomous variables Y and X_j partition the n observations and form a 2x2 contingency table, and Z denotes an arbitrary X_j. Let $p_{Z|y}$

denote $\Pr(X_j = 1|Y = y)$, $q_{Z|y}$ denote $\Pr(X_j = 0|Y = y)$, p_Y denote $\Pr(Y = 1)$, and q_Y denote $\Pr(Y = 0)$. Consider again the joint LPM from the previous section,

$$\Pr(Y = i, Z = j) = \alpha + \alpha_Y(i) + \alpha_Z(j) + \alpha_{YZ}(ij)$$

As before, the parameters are not identified, and we use the centered parameterization, $\alpha_Y := \alpha_Y(1) = -\alpha_Y(0)$, $\alpha_Z := \alpha_Z(1) = -\alpha_Z(0)$, and $\alpha_{YZ} := \alpha_{YZ}(00) = \alpha_{YZ}(11) = -\alpha_{YZ}(10) = -\alpha_{YZ}(01)$ so that

$$\begin{aligned}
\Pr(Y = i) &= (\Pr(Y = i, Z = 0)) + (\Pr(Y = i, Z = 1)) \\
&= (\alpha + \alpha_Y(i) + \alpha_Z(0) + \alpha_{YZ}(i0)) + (\alpha + \alpha_Y(i) + \alpha_Z(1) + \alpha_{YZ}(i1)) \\
&= (\alpha + \alpha_Y(i) + \alpha_Z(0) + \alpha_{YZ}(i0)) + (\alpha + \alpha_Y(i) - \alpha_Z(0) - \alpha_{YZ}(i0)) \\
&= 2(\alpha + \alpha_Y(i))
\end{aligned}$$

and the conditional mass function is

$$\Pr(Z = 1|Y = i) = \frac{\alpha + \alpha_Y(i) + \alpha_Z(1) + \alpha_{YZ}(i1)}{2(\alpha + \alpha_Y(i))} = \frac{1}{2} + \frac{\alpha_Z(1) + \alpha_{YZ}(i1)}{2(\alpha + \alpha_Y(i))} = \begin{cases} \frac{1}{2} + \frac{\alpha_Z - \alpha_{YZ}}{q_Y} & \text{if } i = 0 \\ \frac{1}{2} + \frac{\alpha_Z + \alpha_{YZ}}{p_Y} & \text{if } i = 1 \end{cases}$$

Let $\theta_{Z|y}$ be the conditional predictivity, defined as $\theta_{Z|y} = p_{Z|y} \vee q_{Z|y}$. From Identity I.1: $|p - q| = 2\theta - 1$,

$$
\begin{aligned}
2\theta_{Z|y=i} - 1 &= |\Pr(Z = 1|Y = i) - \Pr(Z = 0|Y = i)| \\
&= |2\Pr(Z = 1|Y = i) - 1| \\
&= \frac{|\alpha_Z(1) + \alpha_{YZ}(i1)|}{|\alpha + \alpha_Y(i)|} \\
&= \begin{cases} \frac{2|\alpha_Z - \alpha_{YZ}|}{q_Y} & \text{if } i = 0 \\ \frac{2|\alpha_Z + \alpha_{YZ}|}{p_Y} & \text{if } i = 1 \end{cases}
\end{aligned}
$$

Thus, $\theta_{Z|1} = \frac{1}{2} + \frac{|\alpha_Z + \alpha_{YZ}|}{p_Y}$, $\theta_{Y|0} = \frac{1}{2} + \frac{|\alpha_Z - \alpha_{YZ}|}{q_Y}$, and $\mathrm{E}_Z(\theta_{Y|Z}) = \frac{1}{2} + |\alpha_Z + \alpha_{YZ}| + |\alpha_Z - \alpha_{YZ}|$. Similar to the predictivity measures of the previous section, the conditional predictivity is a simple function of α_Y and α_{YZ}, except now weighted to the population of Y.

If we were to consider the "reduced form" model, $\Pr(Z = 1|Y = i) = m\ i + b$, we would identify the intercept b with $\frac{1}{2} + \frac{\alpha_Z - \alpha_{YZ}}{q_Y}$ and the slope m with $\Pr(Z = 1|Y = 1) - \Pr(Z = 1|Y = 0) = \frac{\alpha_Z + \alpha_{YZ}}{p_Y} - \frac{\alpha_Z - \alpha_{YZ}}{q_Y}$. It follows from the triangle inequality that

$$
|m| = \left| \frac{\alpha_Z + \alpha_{YZ}}{p_Y} - \frac{\alpha_Z - \alpha_{YZ}}{q_Y} \right| \geq \left| \frac{\alpha_Z + \alpha_{YZ}}{p_Y} \right| - \left| \frac{\alpha_Z - \alpha_{YZ}}{q_Y} \right| = \theta_{Z|1} - \theta_{Z|0}
$$

Thus, for small m, Y can be ruled out as a predictor of Z. This is not true for the conditional log-linear model (or logistic regression, whose parameters are relative measures of conditional risk, see Appendix for details). If J models are fit for each of the J X_j resulting in slopes $m = [m_1, \ldots, m_K]$, the total predictivity can be assessed: $||m||_2^2$, as we justify in the following section.

6.2 The influence score assesses the fit of the LPM

We demonstrate the role of the influence score in fitting the LPM. The first subsection introduces the influence score and its relationship to the covariance. The covariance is itself related to the predictivity (see Appendix, which generalizes the relationship between the influence score and marginal predictivity found in Lo (2016)). The second subsection demonstrates how comparing

a partition to a gold standard reduces to evaluating the influence score. The main advantage of the influence score for summarizing contingency tables over traditional measures of association is that the influence score does not require the estimated probabilities to lie in the interior of the unit interval. For example, chi-square and model based measures (like AIC or G^2) do not work if $\hat{p} = 0$ or $\hat{p} = 1$.

6.2.1 The general relationship between the influence score and predictivity

Consider a partition of Y, π, that associates with each observation one of J cells. Let n_j denote the number of observations associated with the jth cell. Let $X = [X_1, \ldots, X_J]$ be a vector of dichotomous variables denoting which cell contains the jth observation. Without loss of generality, we suppose Y_{ij} is standardized, $Y_{ij} = \frac{Y'_{ij} - \bar{Y}'}{\sigma_{Y'}}$; the right side can be substituted for the left to examine the non-standardized case. We begin by defining the (average) influence score to be the weighted sum of squares: $\sum_j (\frac{n_j}{n} \bar{Y}_j)^2$. Since $\frac{n_j}{n} \underset{n\to\infty}{\to} \Pr(X_j = 1)$ and $\bar{Y}_j \underset{n\to\infty}{\to} \mathrm{E}(Y|X_j = 1)$,

$$
\begin{aligned}
\sum_j [(\frac{n_j}{n})(\bar{Y}_j)]^2 &\underset{n\to\infty}{\to} \sum_j [\Pr(X_j = 1)\mathrm{E}(Y|X_j = 1)]^2 \\
&= \sum_j [1 \Pr(X_j = 1)\mathrm{E}(Y|X_j = 1) + 0 \Pr(X_j = 0)\mathrm{E}(Y|X_j = 0)]^2 \\
&= \sum_j \mathrm{E}(X_j \mathrm{E}(Y|X_j))^2 \\
&= \sum_j \mathrm{E}(\mathrm{E}(Y X_j|X_j))^2 \\
&= \sum_j \mathrm{E}(Y X_j)^2 \\
&= \sum_j \mathrm{Cov}(Y, X_j)^2
\end{aligned}
$$

where the last equality follows from the fact that Y has zero mean.

Let $\beta = [\beta_1, \ldots, \beta_J]$ be the slopes from J regressions of X_j on Y. Using the least squares estimator, $\hat{\beta}_j = \widehat{\mathrm{Cov}}(Y, X_j)$ so that the (average) influence score can be seen as the norm of the

estimated regression slopes: $||\beta||_2^2 = \sum_j \widehat{\text{Cov}}(Y, X_j)^2$. This measure arises naturally in the analysis of variance (see next section). The covariance has several ideal properties, such as symmetry and invariance. It also bounds the predictivity tighter than other summary measures. (We compare other common measures of contingency tables in the Appendix.) We point out here that the relationship to covariance is similar but meaningfully distinct from χ^2, which, for 2x2 tables, is equivalent to $\rho_j^2 = \frac{\text{Cov}(Y,X_j)}{\text{Var}(Y)\text{Var}(X_j)}$.

6.2.2 Influence score as a measure of the conditional LPM's fit.

Recall that the goal is to assess how well a partition of n objects, $i = \{1, \ldots, n\}$, represents a gold standard variable Y_i. For example, Y_i could be the number of fatalities on the ith road. As in the previous sections, we suppress the index i, writing Y for the measurement of an randomly chosen unit, and we assume Y has been standardized so that $\text{E}(Y) = 0$ and $\text{Var}(Y) = 1$.

Let the partition be represented by a J-vector of dichotomous variables $X = [X_1, \ldots, X_J]$. We assume $X \sim \text{Multinomial}(p = [\alpha_1, \ldots, \alpha_J])$ and use the explained variation, $\text{E}(\text{Var}(X|Y))$, as the measure of assessment. This measure arises in the multivariate analysis of variance (MANOVA, although usually normal errors are assumed).

We assume a conditional linear probability model: $\Pr(X_j = 1|Y = y) = \beta_{0j} + \beta_{1j}y$. However, since $\beta_{0j} = \alpha_j$ by the iterated law of expectation,

$$\alpha_j := \Pr(X_j = 1) = \text{E}(\Pr(X_j = 1|Y)) = \text{E}(\beta_{0j} + \beta_j Y) = \beta_{0j}$$

we simply write $\Pr(X_j = 1|Y = y) = \alpha_j + \beta_j y$.

Since X is multinomial, the total variation is

$$\mathrm{E}(X_j) = \alpha_j$$

$$\mathrm{Var}(X_j) = \alpha_j(1-\alpha_j)$$

$$\mathrm{Cov}(X_{j_1}, X_{j_2}) = -\alpha_{j_1}\alpha_{j_2}$$

and, by the iterated laws of variance and covariance, the between variation is

$$\mathrm{Var}(\mathrm{E}(X_j|Y)) = \mathrm{Var}(\alpha_j + \beta_j Y) = \beta_j^2$$

$$\mathrm{Cov}(\mathrm{E}(X_{j_1}|Y), \mathrm{E}(X_{j_2}|Y)) = \mathrm{Cov}(\alpha_{j_1} + \beta_{j_1}Y, \alpha_{j_2} + \beta_{j_2}Y) = \beta_{1j_1}\beta_{1j_2}$$

and the within variation is

$$\mathrm{E}(\mathrm{Var}(X_j|Y)) = \mathrm{E}[(\alpha_j + \beta_j Y)(1 - \alpha_j - \beta_j Y)] = \alpha_j(1-\alpha_j) - \beta_j^2$$

$$\mathrm{E}(\mathrm{Cov}(X_{j_1}, X_{j_2}|Y)) = \mathrm{E}[-(\alpha_{j_1} + \beta_{j_1}Y)(\alpha_{j_2} + \beta_{j_2}Y)] = -\alpha_{j_1}\alpha_{j_2} - \beta_{j_1}\beta_{j_2}$$

In matrix form, we write $\Sigma_T = \mathrm{diag}(\alpha) - \alpha\alpha^T, \Sigma_B = \beta\beta^T$, and $\Sigma_W = \Sigma_T - \Sigma_B = \mathrm{diag}(\alpha) - \alpha\alpha^T - \beta\beta^T$, where α and β are column vectors $[\alpha_1, \ldots, \alpha_J]$ and $[\beta_1, \ldots, \beta_J]$. T, B, and W stand for the total, between and within.

Due to the decomposition $\Sigma_T = \Sigma_B + \Sigma_W$, a one-dimensional summary of Σ_B (or Σ_B and Σ_T) is used to assess fit. For example, Wilk's lambda $\Lambda = \Sigma_B\Sigma_W^{-1} = \Sigma_B(\Sigma_T - \Sigma_B)^{-1}$. Since our goal is prediction, we are interested in the magnitude of Σ_B (see (Miller 1997) and (Breiman 1984) for details). We use $\mathrm{tr}(\Sigma_B) = \sum_{j=1}^{J} \beta_j^2$ as our summary.

We estimate β_j using the least squares estimator: $\hat{\beta}_j = \frac{\sum_i (y_i - \bar{y})(x_{ij} - \bar{x}_j)}{\sum_i (y_i - \bar{y})^2} = \sum_i y_i x_{ij} = n_j \bar{y}_j$, where n_j is the number of observations in partition cell j and $\bar{y}_j$ is the average of y for the observations in partition j. The plug-in estimator of $\text{tr}(\Sigma_B)$ is thus $\text{tr}(\hat{\Sigma}_B) = \sum_{j=1}^{J} (\sum_i y_i x_{ij})^2 = \sum_{j=1}^{J} n_j^2 \bar{y}_j^2$. Consistency of the estimator follows from the consistency of least squares and the continuous mapping theorem.

Other measures are possible, although not explored in detail here. For the percent of the variation explained, we can estimate $\text{tr}(\hat{\Sigma}_B \hat{\Sigma}_T^{-1})$. However, for $\hat{\Sigma}_T$ to be invertible, the last row/column of Σ_B and Σ_T must be removed and the constraints $\sum_{j=1}^{J} \hat{\alpha}_j = 1$ and $\sum_{j=1}^{J} \hat{\beta}_j = 0$ enforced. The first constraint follows from the fact that the $\hat{\alpha}_j$ are probabilities. The second from the fact that $\sum_j \hat{\beta}_j = \sum_j n \, \text{Cov}(Y, X_j) = n \, \text{Cov}(Y, \sum_j X_j) = n \, \text{Cov}(Y, 1) = 0$.

By the Sherman-Morrison Identity, $\hat{\Sigma}_T^{-1} = \text{diag}([\frac{1}{\hat{\alpha}_1}, \ldots, \frac{1}{\hat{\alpha}_{J-1}}]) + \frac{1}{\hat{\alpha}_J} 1_{J-1 \times J-1}$, where $1_{J-1 \times J-1}$ is a $J-1 \times J-1$ matrix of 1s. Therefore,

$$\text{tr}(\hat{\Sigma}_B \hat{\Sigma}_T^{-1}) = \hat{\beta} \hat{\beta}^T (\text{diag}([\frac{1}{\hat{\alpha}_1}, \ldots, \frac{1}{\hat{\alpha}_{J-1}}]) + \frac{1}{\hat{\alpha}_J} 1_{J-1 \times J-1}) = \sum_{j=1}^{J} \frac{\hat{\beta}_j^2}{\hat{\alpha}_j} = n \sum_{j=1}^{J} n_j \bar{y}_j^2$$

6.3 Predicting NYC Traffic Fatalities

We examine a dataset of traffic fatalities occurring in New York City intersections between the years 2011-2013 and 2016-2018. The years were chosen to capture the conditions before and after the implementation of the traffic-safety policy, Vision Zero. (Though Vision Zero began in 2014 and is ongoing, we compare the fatalities between 2016-2018 on the roads impacted by Vision Zero between 2014-2015 to the fatalities between 2011-2013 on similar roads as of 2011.)

Vision Zero aims to eliminate traffic fatalities by reengineering roads so that drivers are forced to reduce their speeds. Typically, traffic safety countermeasures are added to roads that selectively combine capital investments, like speed humps and signal retimings, and community investments, like education and outreach. The ultimate policy goal is to evaluate which countermeasures reduced fatalities. Our narrower goal here, however, is to determine whether the countermeasures

installed up until 2013, and the fatalities between 2011-2013, are predictive of the fatalities between 2016-2018 on roads that received the new countermeasures between 2014-2015.

Vision Zero researchers view traffic fatalities as deterministic (as the predictable consequence of road features), and New York City selected traffic safety countermeasures for Vision Zero based on the fatalities between 2011-2013. For this reason, we model traffic fatalities as non-random, and the partition formed by traffic safety countermeasures as random (given traffic fatalities in the before period). An alternative interpretation is that we model the treatment assignment process, which is a common strategy in causal inference.

We compare our approach of using the LPM with linear regression (assuming normal errors) and the conditional log-linear model (logistic regression). We choose logistic regression as a comparison for several reasons. First, because it is perhaps the most popular prediction algorithm for binary variables. Second, because the linear and log-linear models are related: the log-linear model is a linear hazard model; a linear model of the instantaneous death rate corresponds to a log-linear model of the prevalence (number of deaths per unit period). See Appendix for details.

6.3.1 Vision Zero dataset

We begin by comparing the three approaches on a subset of the data. The number of New York intersections is broken down in the following 8 2x2 contingency tables (Table 1a and 1b), the first four by whether there was a fatality between 2011 and 2013 (Death) and three of the twelve road characteristics (Leading Pedestrian Interval (Interval), Speed Hump (Hump), and Slow Zone (Slow)), and the second four by whether there was a fatality between 2016 and 2018. A 0 indicates the intersection lacked that characteristic, and a 1 indicates the intersection has that characteristic.

We fit the conditional linear and log-linear models (logistic regression) for all 11 possible models predicting the number of 2011-2013 fatalities given Interval (I), Hump (H), and Slow (S). The results are shown in Table 2. The first column records the model fit using hierarchical notation (see Agresti (2003)). We summarize the log-linear model with AIC (assuming Y|X binomial with

Table 1a: Vision Zero Data (2011-2013)

Slow = 0, Interval = 0	Hump = 0	Hump = 1
Death = 0	4277	1316
Death = 1	9	3

Slow = 1, Interval = 0	Hump = 0	Hump = 1
Death = 0	635	127
Death = 1	5	2

Slow = 0, Interval = 1	Hump = 0	Hump = 1
Death = 0	1019	355
Death = 1	3	5

Slow = 1, Interval = 1	Hump = 0	Hump = 1
Death = 0	488	135
Death = 1	7	1

Table 1b: Vision Zero Data (2016-2018)

Slow = 0, Interval = 0	Hump = 0	Hump = 1
Death = 0	4156	1286
Death = 1	130	33

Slow = 1, Interval = 0	Hump = 0	Hump = 1
Death = 0	601	121
Death = 1	39	8

Slow = 0, Interval = 1	Hump = 0	Hump = 1
Death = 0	966	346
Death = 1	56	14

Slow = 1, Interval = 1	Hump = 0	Hump = 1
Death = 0	467	116
Death = 1	28	20

logistic link, AIC1) and the linear model with the influence score (IS) and AIC (assuming Y|X normal, AIC2). We then calculate the mean absolute error on the 2016-2018 data using the log-linear fit (MAE1, per 100) and linear fit (MAE2, per 100). The superscript ^{a} denotes the model selected from comparing all subsets. The superscript ^{b} denotes the model selected from iteratively collapsing the saturated model, denoted IHS.

Table 2: All subsets for I, H, and S

model	AIC1	AIC2	I	MAE1	MAE2
	455.4	-22183	0.0	4.91	4.91
I	449.5	-22190.0	3.3	4.96	4.96
H	456.1	-22182.3	0.5	4.88	4.88
S	444.1	-22198.3	4.8	4.90	4.90
IH	452.1	-22188.9	5.1^{ab}	4.59	4.59
IS	443.4^{a}	-22198.9	5.1	4.85	4.85
HS	445.6	-22196.2	4.6	4.88	4.88
IH IS	445.6	-22198.2	3.8	4.86	4.85
IH HS	445.1^{b}	-22198.7	3.8	4.86	4.85
IS HS	445.1	-22196.7	3.8	4.86	4.88
IHS	445.6	-22201.3^{ab}	0.0	4.88	4.88

MAE1 and MAE2 differ little for each model, likely because the probability of a fatality is so low that it is closely resembles the incidence. What is different is the selection criteria. AIC1 and AIC2 favor larger models than I and miss a more predictive subset.

The prediction gain from using I is modest in Table 2, but it becomes magnified when we search for the best model over the entire Vision Zero dataset. Table 3 shows the result from iteratively collapsing the saturated model with all 12 covariates (best subsets is not possible): Leading Pedestrian Interval, Speed Hump, and Slow Zone, Enhanced Crossings, Street Improvement, Traffic Calming, Signal Retiming, Street Teams, Senior Centers, Outreach, facility type (direction of road: one way, two way, etc.), system type (purpose of road: major collector, minor, arterial, etc.), and through type (number of through lanes: 1, 2, 3, etc.) I leads to a nearly 50 percent reduction in overall error.

Comparing selection criteria from iteratively collapsing saturated model (backwards step-wise selection)

selection criteria	best value	MAE1	MAE2
AIC1	443.4	4.85	4.85
AIC2	-22406.9	4.85	4.85
I	7.2	2.62	2.62

6.4 Appendix

6.4.1 The 2x2 Contingency Table.

Let $Y \sim \mathrm{B}(p_Y)$ and $Z \sim \mathrm{B}(p_Z)$ be Bernoulli random variables. The goal is to best summarize the relationship between Y and Z.

Denote $p_Y = \Pr(Y = 1)$, $q_Y = \Pr(Y = 0)$, $p_Z = \Pr(Z = 1)$, and $q_Z = \Pr(Z = 0)$. Also denote the entries of the 2x2 contingency table formed by Y and Z by

	Z = 0	Z = 1
Y = 0	a	b
Y = 1	c	d

with

$$a = \Pr(Y = 0, Z = 0)$$

$$b = \Pr(Y = 0, Z = 1)$$

$$c = \Pr(Y = 1, Z = 0)$$

$$d = \Pr(Y = 1, Z = 1)$$

Note that $p_Y = c + d$, $q_Y = a + b$, $p_Z = b + d$, and $q_Z = a + c$. For completeness, denote $p_{YZ} = \Pr(Y = Z) = a + d$ and $q_{YZ} = \Pr(Y \neq Z) = b + c$.

6.4.2 Summarizing a 2x2 Contingency Table by its predictivity

Define the predictivity (Lo 2016) of Y and Z to be

$$\theta_Y = \max(p_Y, q_Y) = p_Y \vee \ q_Y = (c + d) \vee (a + b)$$

$$\theta_Z = \max(p_Z, q_Z) = p_Z \vee \ q_Z = (b + d) \vee (a + c)$$

and the co-predictivity of Y and Z to be

$$\theta_{YZ} = \max(p_{YZ}, q_{YZ}) = p_{YZ} \vee \ q_{YZ} = (a + d) \vee (b + c)$$

Three identities relate predictivity to more common parameters. Suppressing subscripts,

I.1 Net Risk := $|p - q| = 2\theta - 1$

I.2 Var := $pq = \theta(1 - \theta)$

I.3 Bayes rate := $p \wedge \ q = 1 - \theta$

The net risk is also caled the excess probability (Breslow 1980, p. 51-55), Var denotes the variance, and Bayes rate the Bayes optimal error rate. The first identity (Lo 2016, eq. 2) follows from the fact that $p \vee q + p \wedge q = 1$ and $p \vee q - p \wedge q = |p - q|$. Adding the right and left sides of the equations yields $2(p \vee q) = 1 + |p - q|$.

The second follows directly from the definition of predictivity: $\theta(1-\theta) = (p \vee q)(p \wedge q) = pq$. The third can be defined as the frequentist error rate if the estimator $1_{p>q}$ were used for a large number of binary decisions. c.f. (Gelman 2006, p. 99), (Friedman 2001, p. 20), (Bickel 2015, ch. 12.2.2, p. 320).

The predictivity corresponds with the L_∞ norm of the centered random variable $Y - p_Y$.

The Bayes rate corresponds with L_0. These follow from the definition of the L_r norm,

$$||Y - p_Y||_r = \sqrt[r]{\mathrm{E}|Y - p_Y|^r} = \sqrt[r]{p_Y^r q_Y + q_Y^r p_Y}$$

where $||Y - p_Y||_r$ measures the amount of "uncertainty".

It turns out that different values of r produce different averages of p_Y and q_Y:

Setting $r = 2$ produces the geometric mean (or sd), $||Y - p_Y||_2 = \sqrt{p_Y^2 q_Y + q_Y^2 p_Y} = \sqrt{p_Y q_Y}$.

Setting $r = 1$ produces the harmonic mean (or mad), $||Y - p_Y||_1 = |p_Y q_Y + q_Y p_Y| = 2p_Y q_Y = \frac{2}{p_Y^{-1} + q_Y^{-1}}$.

Letting $r \to \infty$, produces the max-norm (or predictivity). Multiplying and dividing by $(p_Y \vee q_Y)$ yields,

$$\lim_{r\to\infty} ||Y - p_Y||_r = \lim_{r\to\infty} (p_Y \vee q_Y) \sqrt[r]{\left(\frac{p_Y}{p_Y \vee q_Y}\right)^r q_Y + \left(\frac{q_Y}{p_Y \vee q_Y}\right)^r p_Y} = (p_Y \vee q_Y)$$

since $\left(\frac{p_Y}{p_Y \vee q_Y}\right)^r \underset{r\to\infty}{\to} 1_{p_Y \geq q_Y}$.

Letting $r \to 0$, produces the min-norm (or Bayes rate). Following the same argument as $r \to \infty$, except multiplying and dividing by $(p_Y \wedge q_Y)$ and noting that $\left(\frac{p_Y}{p_Y \vee q_Y}\right)^r \underset{r\to 0}{\to} 1$.

n.b. if $Y_i \perp\!\!\!\perp Y_j \mid p_Y$, then for any r, $||Y_i - Y_j||_r = E(1^r_{Y_i \neq Y_j}) = p_Y q_Y + q_Y p_Y = 2p_Y q_Y = ||Y_i - p_Y||_1$.

An immediate consequence is the following inequalities:

$$\theta_Y = p_Y \vee q_Y \geq \frac{p_Y + q_Y}{2} = \frac{1}{2} \geq \sqrt{p_Y q_Y} \geq \frac{2}{p_Y^{-1} + q_Y^{-1}} \geq 1 - \theta_Y = p_Y \wedge q_Y$$

Predictivity corresponds with a linear model of a 2x2 table:

Consider the following ANOVA model with centered parameterization

$$\mu_{ij} = \alpha + \alpha_Y(i) + \alpha_Z(j) + \alpha_{YZ}(ij)$$

where $\sum_i \alpha_Y(i) = \sum_j \alpha_Z(j) = \sum_i \alpha_{YZ}(ij) = \sum_j \alpha_{YZ}(ij) = 0$.

For a 2x2 table, $i = j = 0, 1$ so there are 9 parameters and 6 constraints for a total of 3 “free

parameters" to be estimated from 4 observations. It follows from $\mu_{ij} = \Pr(Y = i, Z = j)$ and the constraints that:

$$\begin{aligned}
\alpha &= \frac{a+b+c+d}{4} = \frac{1}{4} \\
\alpha_Y(1) = -\alpha_Y(0) &= \frac{c+d-a-b}{4} = \frac{1}{4}(p_Y - q_Y) \\
\alpha_Z(1) = -\alpha_Z(0) &= \frac{b+d-a-c}{4} = \frac{1}{4}(p_Z - q_Z) \\
\alpha_{YZ}(0,0) = \alpha_{YZ}(1,1) = -\alpha_{YZ}(0,1) = -\alpha_{YZ}(1,0) &= \frac{a+d-b-c}{4} = \frac{1}{4}(p_{YZ} - q_{YZ})
\end{aligned}$$

The constraints only identify the magnitude of the coefficients; an equivalent decomposition is given by $\mu_{ij} = \alpha - \alpha_Y(i) - \alpha_Z(j) - \alpha_{YZ}(ij)$. From identity $|p - q| = 2\theta - 1$,

$$\begin{aligned}
|\alpha_Y| &= \frac{1}{4}|p_Y - q_Y| = \frac{1}{2}\theta_Y - \frac{1}{4} = \frac{1}{2}\theta_Y - \alpha \\
|\alpha_Z| &= \frac{1}{4}|p_Z - q_Z| = \frac{1}{2}\theta_Z - \frac{1}{4} = \frac{1}{2}\theta_Z - \alpha \\
|\alpha_{YZ}| &= \frac{1}{4}|p_{YZ} - q_{YZ}| = \frac{1}{2}\theta_{YZ} - \frac{1}{4} = \frac{1}{2}\theta_{YZ} - \alpha
\end{aligned}$$

The intercept $\alpha = \frac{1}{4}$ is the maximum uncertainty if no data are collected and thus each cell is assumed to have $\mu_{ij} = \frac{1}{4}$. The strength of marginal and joint association is measured in L_∞ distance–e.g. one of $\alpha_Y(0)$ and $\alpha_Y(1)$ is $\frac{1}{2}\theta_Y$ above α, the other is $\frac{1}{2}\theta_Y$ below alpha, and the two are θ_Y apart. If $\theta_{YZ} = 2\alpha = \frac{1}{2}$, the marginal measures θ_Y and θ_Z completely characterize the table and the table is said to be "collapseable". c.f. Section 6.4.3, this does not imply Y and Z are independent, the collapseability condition for the log-linear model.

Put another way, suppose the interaction term α_{YZ} were not included and $\hat{\mu}_{ij} = \alpha + \alpha_Y(i) + \alpha_Z(j)$. Then the mean absolute error $\text{MAE}_{-\alpha_{YZ}} := \sum_{ij} \mu_{ij}|\hat{\mu}_{ij} - \mu_{ij}| = |\alpha_{YZ}| \sum_{ij} \mu_{ij} = \frac{|p_{YZ} - q_{YZ}|}{4} = \frac{1}{2}\theta_{YZ} - \frac{1}{4}$. If the co-predictivity $\theta_{YZ} = \frac{1}{2}$, then no error is made by excluding the term α_{YZ}.

That the marginal terms of the linear model corresponds to the marginal probabilities of the

contingency table was recognized by (Bishop 2007, p. 23-24). However, they did not seem to recognize that a unit change in the margins (from row 1 to row 2, or column 1 to column 2) correspond to the predictivity. In fact, their recommendation to use the log-linear model presented in Section 6.4.3 over the linear model rests on the assumption that the interaction of the linear model has no interpretation.

Generalization for an IxJ table The predictivity interpretation of the ANOVA model with centered parameterization generalizes to the IxJ case. Letting $i \in \{1, \ldots, I\}$ and $j \in \{1, \ldots, J\}$, assume without loss of generality that $I \leq J$. There are $1 + I + J + IJ$ parameters and $I + J + 2$ constraints, resulting in $IJ - 1$ "free parameters" to be estimated from IJ observations. It follows from $\mu_{ij} = \Pr(Y = i, Z = j)$ and the constraints that:

$$1 = \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij} = IJ\alpha$$

$$\Pr(Y = i) - \Pr(Y \neq i) = 2\Pr(Y = i) - 1 = 2 \sum_{j=1}^{J} \mu_{ij} - 1 = 2J(\alpha + \alpha_Y(i)) - 1$$

$$\Pr(Z = j) - \Pr(Z \neq j) = 2\Pr(Z = j) - 1 = 2 \sum_{i=1}^{I} \mu_{ij} - 1 = 2I(\alpha + \alpha_Z(j)) - 1$$

$$\Pr(Y = Z) - \Pr(Y \neq Z) = 2\Pr(Y = Z) - 1 = 2 \sum_{i=1}^{I} \mu_{ii} - 1 = 2(I\alpha + \sum_{i=1}^{I} \alpha_Z(i) + \sum_{i=1}^{I} \alpha_{YZ}(i, i)) - 1$$

If $I = J$, the last equation reduces to $\Pr(Y = Z) - \Pr(Y \neq Z) = 2(I\alpha + \sum_{i=1}^{I} \alpha_{YZ}(i, i)) - 1$

Conditional predictivity corresponds with a weighted linear model of a 2x2 table:

Conditional predictivity, the focus of Section **??**, is defined to be $\theta_{Z|y} = p_{Z|y} \vee q_{Z|y}$, where $p_{Z|y}$ denotes $\Pr(Z = 1|Y = y)$ and $q_{Z|y}$ denotes $\Pr(Z = 0|Y = y)$. It satisfies all the identities of Section 6.4.1. For this Subsection, Y is assumed dichotomous so that the conditional predictivity can be

derived directly from the linear model of Subsection 6.4.2:

$$\mu_{ij} = \alpha + \alpha_Y(i) + \alpha_Z(j) + \alpha_{YZ}(ij)$$

Due to the centered parameterization, $\alpha_Z(1)) = -\alpha_Z(0))$ and $\alpha_{YZ}(i1)) = -\alpha_{YZ}(i0))$ so that

$$\begin{aligned} p_Y = \Pr(Y = i) &= (\Pr(Y = i, Z = 0)) + (\Pr(Y = i, Z = 1)) \\ &= (\alpha + \alpha_Y(i) + \alpha_Z(0) + \alpha_{YZ}(i0)) + (\alpha + \alpha_Y(i) + \alpha_Z(1) + \alpha_{YZ}(i1)) \\ &= (\alpha + \alpha_Y(i) + \alpha_Z(0) + \alpha_{YZ}(i0)) + (\alpha + \alpha_Y(i) - \alpha_Z(0) - \alpha_{YZ}(i0)) \\ &= 2(\alpha + \alpha_Y(i)) \end{aligned}$$

The conditional mass function is

$$\Pr(Z = 1|Y = i) = \frac{\alpha + \alpha_Y(i) + \alpha_Z(1) + \alpha_{YZ}(i1)}{2(\alpha + \alpha_Y(i))} = \frac{1}{2} + \frac{\alpha_Z(1) + \alpha_{YZ}(i1)}{2(\alpha + \alpha_Y(i))} = \begin{cases} \frac{1}{2} + \frac{\alpha_Z - \alpha_{YZ}}{q_Y} & \text{if } i = 0 \\ \frac{1}{2} + \frac{\alpha_Z + \alpha_{YZ}}{p_Y} & \text{if } i = 1 \end{cases}$$

where $\alpha_Y = \alpha_Y(1)$ and $\alpha_{YZ} = \alpha_{YZ}(11)$. By Identity I.1, the conditional predictivity satisfies

$$\begin{aligned} 2\theta_{Z|i} - 1 &= |\Pr(Z = 1|Y = i) - \Pr(Z = 0|Y = i)| \\ &= |2\Pr(Z = 1|Y = i) - 1| \\ &= \frac{|\alpha_Z(1) + \alpha_{YZ}(i1)|}{|\alpha + \alpha_Y(i)|} \\ &= \begin{cases} \frac{|\alpha_Z - \alpha_{YZ}|}{2q_Y} & \text{if } i = 0 \\ \frac{|\alpha_Z + \alpha_{YZ}|}{2p_Y} & \text{if } i = 1 \end{cases} \end{aligned}$$

Thus, $\theta_{Z|1} = \frac{1}{2} + \frac{|\alpha_Z + \alpha_{YZ}|}{4p_Y}$ and $\mathrm{E}_Y(\theta_{Z|Y}) = \frac{1}{2} + \frac{|\alpha_Z + \alpha_{YZ}| + |\alpha_Z - \alpha_{YZ}|}{4}$.

As in the previous section, conditional predictivity is directly interpreted from the coefficients of the linear. However, where in the previous model assessing predictivity involved observing the magnitude of a coefficient, conditional predictivity requires assessing the sum of two coefficients.

Also as in the previous section, the mean absolute error of the conditional model is a function of the co-predictivity. Suppose α_{YZ} was thought to be zero so that the probability $\Pr(Z = 1|Y = 1)$ were estimated simply by $\hat{\Pr}(Z = 1|Y = 1) = \frac{1}{2} + \frac{\alpha_z}{p_Y}$. Then

$$\begin{aligned}\mathrm{MAE}_{-\alpha_{YZ}} &= |\hat{\Pr}(Z = 0|Y = 1) - \Pr(Z = 0|Y = 1)|\Pr(Z = 0|Y = 1) \\ &\quad + |\hat{\Pr}(Z = 1|Y = 1) - \Pr(Z = 1|Y = 1)|\Pr(Z = 1|Y = 1) \\ &= \frac{|\alpha_{YZ}|}{p_Y}\end{aligned}$$

and as before $\mathrm{E}_Y(\mathrm{MAE}_{-\alpha_{YZ}}) = |\alpha_{YZ}| = \frac{1}{2}\theta_{YZ} - \frac{1}{4}$, and no error is made if and only if the co-predictivity is zero.

6.4.3 Alternative measures of uncertainty and association

For comparison, consider two traditional measures of uncertainty—the standard deviation, $\mathrm{SD}(Y) = \sqrt{\mathrm{Var}(Y)} = \sqrt{p_Y q_Y}$ and the (marginal) odds (Yule 1912), $\mathrm{O}(Y) = \frac{p_Y}{q_Y}$, and two traditional measures of association—the covariance $\mathrm{Cov}(Y, Z) = d - p_Y p_Z$, and the (co) odds ratio $\mathrm{OR}(Y, Z) = \frac{ad}{bc}$. The odds is typically normalized: $\gamma_Y = \frac{\mathrm{O}(Y)-1}{\mathrm{O}(Y)+1}$ and $\gamma_{YZ} = \frac{\mathrm{OR}(Y,Z)-1}{\mathrm{OR}(Y,Z)+1}$.

The covariance is a common measure of association as the angle or "inner product", and the variance is the corresponding "norm". It arises frequently in linear models.

The normalized odds can be rewritten

$$\gamma_Y = \frac{\mathrm{O}(Y) - 1}{\mathrm{O}(Y) + 1} = \frac{\frac{p_Y}{q_Y} - 1}{\frac{p_Y}{q_Y} + 1} = \frac{\frac{1}{q_Y}(p_Y - q_Y)}{\frac{1}{q_Y}(p_Y + q_Y)} = p_Y - q_Y$$

so that $|\alpha_Y| = \frac{1}{4}|\gamma_Y|$, where α_Y was defined in Section 1.1b. An identical argument shows that $|\alpha_Z| = \frac{1}{4}|\gamma_Z|$ and $|\alpha_{YZ}| = \frac{1}{4}|\gamma_{YZ}|$. n.b. $\text{OR}(Y, Z) \neq \text{O}(YZ)$.

This measure arises in calculating the unconditional net risk $\Pr(Y = 1, Z = 0) - \Pr(Y = 0, Z = 1) = p_Y(1 - p_Z) - (1 - p_Y)p_Z = p_Y - p_Z$. The conditional net risk (for matched pairs) produces the marginal odds ratio $\text{MOR}(Y, Z) = \frac{p_Y(1-p_Z)}{(1-p_Y)p_Z}$ (not to be confused with the (co) odds ratio defined above, which arises in log-linear models, see 6.4.3):

$$\Pr(Y = 0, Z = 1|Y + Z = 1) = \frac{(1 - p_Y)p_Z}{(1 - p_Y)p_Z + p_Y(1 - p_Z)} = \frac{1}{1 + \text{MOR}(Y, Z)}$$

so that $\Pr(Y = 1, Z = 0|Y + Z = 1) - \Pr(Y = 0, Z = 1|Y + Z = 1) = \frac{\text{MOR}(Y,Z)-1}{\text{MOR}(Y,Z)+1}$.

The odds ratio corresponds with a log-linear model of a 2x2 table.

The odds ratio can be derived using the ANOVA model from 6.4.2, except with a log-link: Consider the following ANOVA model with centered parameterization

$$\log \mu_{ij} = \beta + \beta_Y(i) + \beta_Z(j) + \beta_{YZ}(ij)$$

where $\sum_i \beta_Y(i) = \sum_j \beta_Z(j) = \sum_i \beta_{YZ}(ij) = \sum_j \beta_{YZ}(ij) = 0$. The relationship in 1.1b holds with differences being replaced by ratios:

$$\begin{aligned}
\beta &= \frac{1}{4} \log abcd \\
\beta_Y &= \frac{1}{4} \log \frac{cd}{ab} = \frac{1}{4} \log \text{OR}(Z, YZ) \\
\beta_Z &= \frac{1}{4} \log \frac{bd}{ac} = \frac{1}{4} \log \text{OR}(Y, YZ) \\
\beta_{YZ} &= \frac{1}{4} \log \frac{ad}{bc} = \frac{1}{4} \log \text{OR}(Y, Z)
\end{aligned}$$

Thus, testing whether the reduced marginal model or "independent" model holds is equivalent to testing whether the odds ratio is one or γ_{YZ} is zero.

The log-linear model is multiplicative on the original probability scale, where the coefficients measure the relative difference (relative risk) between any two rows/columns. Relative risk may be preferred to absolute risk and the linear model because it is invariant to conditioning: For any events A, B, and C, $\frac{\Pr(A,C)}{\Pr(B,C)} = \frac{\Pr(A|C)}{\Pr(B|C)}\frac{\Pr(C)}{\Pr(C)} = \frac{\Pr(A|C)}{\Pr(B|C)}$. In other words, the relative risk does not depend on the prevalence of C in the population. In contrast, the risk difference scale with the prevalence of B: $\Pr(A, C) - \Pr(B, C) = \Pr(A|C)\Pr(C) - \Pr(B|C)\Pr(C) \leq \Pr(A|C) - \Pr(B|C)$. The fact that predictivity and the Bayes rate depend on the prevalence suggests the log-linear model may not be ideal tools for prediction.

The linear and log-linear parameterizations are not equivalent.

The linear and log-linear probability models correspond with uniform and exponential thresholding function.

Zero covariance or zero log-odds are necessary and sufficient conditions for zero independence. Predictivity is neither a necessary or sufficient condition for independence. However, "large" statistical dependence, as measured by either covariance or the odds implies high predictivity.

The following tables (Fienberg 2007) show an undesirable property of the odds ratio when cells are sparse/zero. Here, $\text{OR}(Y, Z) = \infty$, but Table 2 is clearly stronger evidence of a relationship than Table 1. The problem is fixed by adding ϵ to all cells, with large samples, (Cox 1989, p. 33) and (Agresti 2003, p. 617) for $\epsilon = \frac{1}{2}$ (it eliminates the bias by an order of n as demonstrated by expansion), however $\text{OR}(Y, Z)$ is then sensitive to ϵ.

Alternatively, the difference in the tables is captured by predictivity: $\theta_Y = .8$ in both tables, but θ_Z and θ_{YZ} are larger in the second table (.6 and .8) compared to (.8 and 1).

Table 1	Z = 0	Z = 1
Y = 0	60	20
Y = 1	0	20

Table 2	Z = 0	Z = 1
Y = 0	80	0
Y = 1	0	20

Independent tables range from $\theta_{YZ} = \frac{1}{2}$ to 1. For example, in Table 2, OR(Y, Z) is 1 and Y and Z are independent for any $d = \{0, \ldots, \infty\}$. As d increases, the predictivity $\frac{(d-1)^2}{(d+1)^2}$ rises from $\frac{1}{2}$ to 1.

Table 2	Z = 0	Z = 1
Y = 0	1	d
Y = 1	d	d^2

Log-linear probability model implies a linear model of average risk.

Reparameterization clarifies the relationship between predictivity and the odds ratio.

Let probability $p_Y(t)$ denote the net risk of Y between time 0 and $t < 1$. Let λ_Y denote the incidence or instantaneous percent of new cases at time t: $\lambda_Y(t) = \lim_{h \to 0} \frac{1}{h} \frac{p_Y(t) - p_Y(t+h)}{p_Y(t)} = -\frac{\frac{d}{dt} p_Y(t)}{p_Y(t)}$ (Breslow 1980, p. 51). The average incidence $\Lambda_Y(t) = \int_0^1 \lambda_Y(s) ds = -\log(p_Y(t))$. Taking $t = 1$ and suppressing subscripts yields, $\Lambda_Y = -\log(p_Y)$. Thus, a log-linear model that summarizes a table based on odds ratios is equivalent to a linear model of average (negative) risk.

The empirical relationship is asymptotic: Let $Y \sim$ Binomial(n, p_Y). Define empirical estimates $\hat{p}_Y = \frac{Y}{n}$ and $\hat{\Lambda} = \sum_{i=n-Y}^{n} \frac{1}{i}$. Each term in the sum $\hat{\Lambda}_Y$ is the empirical incidence at each positive event—the number of positive units at the event divided by the number at risk—and the sum converges to the integral in Λ_Y.

That the sum convergences at rate n^{-1} follows from the bound $\sum_{i=1}^{c} \frac{1}{i} \leq \frac{1}{c+1} + \log(c) + \gamma$, where γ is the Euler-Mascheroni constant and

$$\begin{aligned}\Pr(|\sum_{i=1}^{Y}\frac{1}{i}-\log(Y)-\gamma|\geq\frac{\epsilon}{n}) &\leq \frac{n}{\epsilon}\,\mathrm{E}|\sum_{i=1}^{Y}\frac{1}{i}-\log(Y)-\gamma| \\ &\leq \frac{n}{\epsilon}\,\mathrm{E}(\frac{1}{c+1}) \\ &= \frac{n}{\epsilon}\frac{1-(1-p)^{n+1}}{(n+1)p} \\ &\underset{n\to\infty}{\to} \frac{1}{p\epsilon}\end{aligned}$$

$$\hat{\Lambda}_Y=\sum_{n-Y}^{n}\frac{1}{i}=\sum_{1}^{n}\frac{1}{i}-\sum_{1}^{Y}\frac{1}{i}=\log(n)-\log(Y)+\mathrm{O}_p(n^{-1})=-\log(\hat{p}_Y)+\mathrm{O}_p(n^{-1})$$

Since $\beta_{YZ}=\frac{1}{4}\log\mathrm{OR}(Y,Z)=\frac{1}{4}\log\frac{ad}{bc}=\frac{1}{4}(\Lambda_a+\Lambda_d-\Lambda_b-\Lambda_c)$, coefficient β_{YZ} measures whether the average incidence is different if $Y=Z$ or $Y\neq Z$. Thus, if subjects are divided into two groups, assessing the predictivity of the incidence is identical to assessing the odds of a case. However, it is easy to imagine events with large incidences and low prevalence—suggesting the log-linear model is not a replacement for the linear model for prediction.

Conversely, the $p(t)$ could be interpreted as the hazard rate of some underlying life time: $p(t)=\frac{f(t)}{S(t)}$. The same argument shows the linear model of the average hazard is itself implied by a log-linear survival model: $p_Y(t)=-\frac{\frac{d}{dt}S_Y(t)}{S_Y(t)}\Rightarrow\log(S_Y(t))=-\int_0^t p_Y(s)ds:=-P_Y(t)$. Suppressing subscripts yields $P_Y=-\log(S_Y)$. If the Kaplan-Meier estimate is $\hat{S}_Y=\prod_{i=1}^{Y}\frac{n-i}{n-i+1}$, then $\hat{P}_Y=-\log\left(\prod_{i=1}^{Y}\frac{n-i}{n-i+1}\right)=\sum_{i=1}^{Y}\left(\sum_{j=1}^{n-i+1}\frac{1}{j}-\sum_{j=1}^{n-i}\frac{1}{j}\right)+\mathrm{O}_p(n^{-1})=\sum_{i=n-Y}^{n}\frac{1}{i}+\mathrm{O}_p(n^{-1})$

6.4.4 Connection between predictivity and co-predictivity.

This and the following subsections investigate the relationship between predictivity and other measures of uncertainty/association. In general, marginal measures of association are 1-1 functions of each other, whereas joint measures (interactions) are not. Thus, choosing a marginal measure and dropping the interaction term forgoes specific information—unless that term was zero. How

much information is forgone depends on how tight the bound is between measures. Similar bounds exist in information theory (e.g. Fano's Inequality).

Predictivity is a sufficient, but not necessary condition of dependence. In comparison, (non-zero) correlation and (non-unitary) log-odds are necessary and sufficient conditions for dependence.

Invariance to permutation:

Referring to θ_Y and θ_Z as the predictivity and θ_{YZ} as the co-predictivity is somewhat arbitrary. Permuting elements below produces a table (left) with margins Z and YZ and diagonal Y, and a table (right) with margins Y and YZ and diagonal Z. Therefore, any formula relating θ_{YZ} with θ_Y and θ_Z could be rewritten to relate θ_Y with θ_Z and θ_{YZ}.

	Z = 0	Z = 1
YZ = 0	c	b
YZ = 1	a	d

	YZ = 0	YZ = 1
Y = 0	b	a
Y = 1	c	d

Any two measures of predictivity, θ_Y, θ_Z, and θ_{YZ} are unrelated:

Individually, $\frac{1}{2} \leq \theta_Y, \theta_Z, \theta_{YZ} \leq 1$, and knowing any one measure does not narrow down the range of the other two. For example, if $a = b = \frac{1}{2}, c = d = 0$, then $\theta_Y = 1$ and $\theta_Z = \theta_{YZ} = \frac{1}{2}$. If $a = 1, b = c = d = 0$, then $\theta_Y = \theta_Z = \theta_{YZ} = 1$. However, not all combinations of θ_Y, θ_Z, θ_{YZ} are possible; no contingency table can yield, for example, $\theta_Y = \frac{1}{2}$, $\theta_Z = \theta_{YZ} = 1$. (See next point.)

All three measures of predictivity are related:

For any 1-1 assignment of $\theta_Y, \theta_Z, \theta_{YZ}$ to X_1, X_2, X_3,

$$X_2 + X_3 \leq 1 + X_1$$

For example, $\theta_{YZ} \leq \theta_Y + \theta_Z \leq 1 + \theta_{YZ}$. This follows from the definitions of θ_Y, θ_Z, θ_{YZ}:

$$\begin{aligned}
\theta_Y + \theta_Z &= (a+b) \vee (c+d) + (a+c) \vee (b+d) \\
&= (2a+b+c) \vee (a+2c+d) \vee (a+2b+d) \vee (b+c+2d) \\
&= (a+1-d) \vee (c+1-b) \vee (b+1-c) \vee (d+1-a) \\
&\leq (a+1) \vee (c+1) \vee (b+1) \vee (d+1) \\
&= 1 + a \vee b \vee c \vee d \\
&\leq 1 + (a+d) \vee (b+c) \\
&= 1 + \theta_{YZ}
\end{aligned}$$

As mentioned in Section 6.4.1, an identical argument establishes the other relationships, such as $\theta_Y \leq \theta_{YZ} + \theta_Z \leq 1 + \theta_Y$.

Two consequences follow immediately. The first is that, since the arithmetic mean is an upper bound to the geometric and harmonic means, we also have

$$\frac{1}{2} + \frac{1}{2}X_1 \geq \frac{X_2 + X_3}{2} \geq \sqrt{X_2 X_3} \geq \frac{2}{X_2^{-1} + X_3^{-1}} \geq \frac{1}{2} \geq \frac{1}{2}X_1$$

The averages reach the upper bound when $X_2 = X_3$. For example, if $\theta_{YZ} = \frac{1}{2}$, then the average marginal predictivity of Y and Z cannot exceed $\frac{3}{4}$, with equality if $\theta_Y = \theta_Z = \frac{3}{4}$. Conversely, if $\theta_Y = 1$ and $\theta_Z = \frac{1}{2}$, then $\theta_{YZ} = \frac{1}{2}$, which was observed at the end of Section 6.4.2.

The second consequence is that the estimands, $2\times\text{average}(\theta_Y, \theta_Z) - \theta_{YZ}$ and $1 - (2\times\text{average}(\theta_Y, \theta_Z) - \theta_{YZ}) = 1 + \theta_{YZ} - 2\times\text{average}(\theta_Y, \theta_Z)$, are measures of association. They are between 0 and 1 for any average, and indicate the percent of the marginal predictivity of Y and Z not captured (captured) by the co-predictivity.

6.4.5 Connection between predictivity and deviation.

The standard deviation is a deterministic function of the predictivity.

$\mathrm{SD}(Y) = \sqrt{\theta_Y(1-\theta_Y)}$

From Section 6.4.1, Identity 1.2, $\theta_Y(1-\theta_Y) = (p_Y \vee q_Y)(p_Y \wedge q_Y) = p_Y q_Y = \mathrm{Var}(Y)$.

An immediate consequence is that

$$\theta_Y = 1 - \frac{\mathrm{Var}(Y)}{\theta_Y} \geq 1 - 2\mathrm{Var}(Y)$$

Quadratic formula yields

$$\theta_Y = \frac{1}{2} + \sqrt{\frac{1}{4} - \mathrm{SD}(Y)^2}$$

This also follows from Section 6.4.1, Identity 1.1,

$$(p_Y + q_Y)^2 = 1 \iff p_Y^2 + q_Y^2 - 2p_Y q_Y = 1 - 4p_Y q_Y \iff |p_Y - q_Y| = \sqrt{1 - 4\mathrm{Var}(Y)}$$

A second immediate consequence is that, dropping α_Y from the linear probability model and calculating $\hat{\mu}_{ij} = \alpha + \alpha_Z(j) + +\alpha_{YZ}(i)$ produces residual $\pm\alpha_Y = \pm\frac{1}{4}(p_Y - q_Y)$. The resulting mean squared error is $\mathrm{MSE}_{-\alpha_Y} := \sum_{ij} \mu_{ij}(\hat{\mu}_{ij} - \mu_{ij})^2 = \frac{1}{16}(p_Y - q_Y)^2 \sum_{ij} \mu_{ij} = \frac{1}{4}(\theta_Y - \frac{1}{2})^2 = \frac{1}{16} - \frac{1}{4}\mathrm{Var}(Y)$. As with the mean absolute error from Section 1.1b, if the predictivity of Y is $\theta_Y = \frac{1}{2}$, then there is no error by excluding the term α_{YZ}. While MAE is linear in θ_Y, MSE is linear in $\mathrm{Var}(Y)$.

Finally, by the Cauchy-Schwarz inequality, $\mathrm{Cov}(Y,Z)^2 \leq \mathrm{Var}(Z)\mathrm{Var}(Y) \leq \mathrm{Var}(Y)$ implies that $\theta_Y \leq \frac{1}{2} + \sqrt{\frac{1}{4} - \mathrm{Cov}(Y,Z)^2}$. For example, if $Y \approx Z$ and $\mathrm{Cov}(Y,Z) \approx \mathrm{Var}(Y) = .25$, then the predictivity of either θ_Y or θ_Z is $\frac{1}{2} - \sqrt{\frac{1}{4} - \frac{1}{16}} = \frac{2+\sqrt{3}}{4} \approx .93$.

The co-predictivity is greater than twice the standard codeviation:

$\theta_{YZ} \geq 2\sqrt{|\mathrm{Cov}(Y,Z)|}$

First note that $\text{Cov}(Y, Z) = ad - bc$ since

$$\begin{aligned}
\text{Cov}(Y, Z) &= \text{E}(Y, Z) - \text{E}(Y)\text{E}(Z) \\
&= d - (c + d)(b + d) \\
&= d - (d^2 + bd + cd + bc) \\
&= (1 - b - c - d)d - bc \\
&= ad - bc
\end{aligned}$$

The inequality is established by considering the four cases where either of $a + d \geq b + c$ or $ad \geq bc$ is true or false. If $a + d \geq b + c$ and $ad \geq bc$, then

$$\theta_{YZ}^2 = (a + d)^2 = 4ad + (a - d)^2 \geq 4|ad - bc| = 4|\text{Cov}(Y, Z)|$$

If $a + d \geq b + c$ and $ad < bc$, then $\theta_{YZ} \geq \frac{1}{2}$ and $b + c \leq \frac{1}{2}$ together imply $bc \leq \frac{1}{16}$ and

$$\theta_{YZ}^2 \geq \frac{1}{4} \geq 4bc \geq 4bc - 4ad = 4|ad - bc| = 4|\text{Cov}(Y, Z)|$$

The argument establishing the other two cases is identical, except with the role of $a + d$ and $b + c$ reversed.

The relationship between θ_{YZ} and $\sqrt{|\text{Cov}(Y, Z)|}$ is not deterministic. Equality holds iff $a = d = \frac{1}{2}, b = c = 0$ or $a = d = 0, b = c = \frac{1}{2}$. Since, for a 2x2 table, $\text{Cov}(Y, Z) = 0 \iff Y \perp\!\!\!\perp Z$, an immediate consequence is that dependence is a sufficient (but not necessary) condition for predictivity.

6.4.6 Connection between the odds and deviation.

The standard deviation is a deterministic function of the odds:

$\mathrm{SD}(Y) = \frac{\sqrt{\mathrm{O}(Y)}}{\mathrm{O}(Y)+1}$

This follows directly from the definition of the odds $\frac{\sqrt{\mathrm{O}(Y)}}{\mathrm{O}(Y)+1} = \frac{\sqrt{\frac{p_Y}{q_Y}}}{\frac{p_Y}{q_Y}+1} = \sqrt{\frac{p_Y}{q_Y}}q_Y = \sqrt{p_Y q_Y} = \mathrm{SD}(Y)$.

The odds can be "normalized" to produce a measure of "certainty", $\gamma_Y = \frac{\mathrm{O}(Y)-1}{\mathrm{O}(Y)+1}$, where γ_Y takes the value -1 iff $p_Y = 0$, the value 0 iff $p_Y = .5$, and the value 1 iff $p_Y = 1$. The same normalization of the odds ratio produces a measure of association. Recall that

$$\gamma_Y = \frac{\mathrm{O}(Y) - 1}{\mathrm{O}(Y) + 1} = \frac{\frac{p_Y}{q_Y} - 1}{\frac{p_Y}{q_Y} + 1} = \frac{\frac{1}{q_Y}(p_Y - q_Y)}{\frac{1}{q_Y}(p_Y + q_Y)} = p_Y - q_Y$$

It follows that

$$\begin{aligned}
1 - \gamma_Y^2 &= p_Y + q_Y - \gamma_Y^2 \\
&= p_Y + q_y - (p_Y - q_Y)^2 \\
&= p_y + q_y + p_y^2 + q_Y^2 + 2p_y q_y \\
&= p_Y(1 - p_Y) + q_Y(1 - q_Y) + 2p_Y q_Y \\
&= 4p_Y q_Y \\
&= 4\mathrm{SD}(Y)^2
\end{aligned}$$

and therefore $|\gamma_Y| = \sqrt{1 - 4\mathrm{SD}(Y)^2}$

Combining this result with Section 6.4.6

$$|\gamma_Y| = \sqrt{1 - 4\mathrm{SD}(Y)^2} \leq 1 - 2\mathrm{Var}(Y) \leq \theta_Y$$

with equality iff the variance is zero.

The squared normalized odds ratio is greater than twice the standard codeviation:

$\gamma_{YZ}^2 \geq 2\sqrt{|\text{Cov}(Y,Z)|}$

Similar to Section , and assuming $\gamma_{YZ} \geq 0$,

$$\gamma_{YZ} = \frac{\text{OR}(Y,Z)-1}{\text{OR}(Y,Z)+1} = \frac{\frac{ad}{bc}-1}{\frac{ad}{bc}+1} = \frac{ad-bc}{ad+bc} \geq 4(ad-bc) = 4\text{Cov}(Y,Z)$$

The inequality is reversed if $\gamma_{YZ} < 0$. A tighter bound is found by using the following monotonic transformation of γ_{YZ}:

$$\frac{1-\sqrt{1-\gamma_{YZ}^2}}{\gamma_{YZ}} = \frac{\sqrt{\text{OR}(Y,Z)}-1}{\sqrt{\text{OR}(Y,Z)}+1} = \frac{\sqrt{\frac{ad}{bc}}-1}{\sqrt{\frac{ad}{bc}}+1} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \geq 4\text{Cov}(Y,Z)$$

where the inequality follows from

$$\begin{aligned} &\frac{1}{2} \geq \sqrt{ad}+\sqrt{bc} \\ \Rightarrow\ &1 \geq 4(\sqrt{ad}+\sqrt{bc})^2 \\ \Rightarrow\ &\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \geq 4(\sqrt{ad}+\sqrt{bc})(\sqrt{ad}-\sqrt{bc}) = 4(ad-bc) \end{aligned}$$

Solving for γ_{YZ} yields $|\gamma_{YZ}| \geq \frac{2(4\text{Cov}(Y,Z))}{1+(4\text{Cov}(Y,Z))^2} = \frac{\text{Cov}(Y,Z)}{\frac{1}{8}+2\text{Cov}(Y,Z)^2}$

Covariance is zero iff the odds ratio is one. For a 2x2 table, both conditions are equivalent to independence:

This follows directly from Section 6.4.6, since $\text{OR}(Y,Z) = \frac{ad}{bc} = 1$ iff $\text{Cov}(Y,Z) = ad-bc = 0$. Independence is defined as $d = (b+d)(c+d)$, which is true iff $\text{Cov}(Y,Z) = d-(b+d)(c+d) = 0$.

In contrast, predictivity is unrelated to independence.

6.4.7 Connection between predictivity and odds.

The predictivity is a deterministic function of the odds:

Combining the results from 2.1 and 3.1 yields

$$\theta_Y = \frac{1}{2} + \sqrt{\frac{1}{4} - \frac{\mathrm{O}(Y)}{(\mathrm{O}(Y)+1)^2}}$$

Furthermore, the same sections imply $\theta_Y = \frac{1}{2} + \sqrt{\frac{1}{4} - \mathrm{SD}(Y)^2} = \frac{1}{2} + \sqrt{\frac{1}{4} - \frac{1}{4}(1-\gamma_Y^2)} = \frac{1}{2} + \sqrt{\frac{\gamma_Y^2}{4}} = \frac{1}{2} + \frac{|\gamma_Y|}{2}$.

This also follows immediately from identity I.1 $\theta_Y = \frac{1}{2} + \frac{1}{2}|p_Y - q_Y|$.

6.4.8 Connection between predictivity and regression slope.

The slope of the linear probability model is the difference in conditional probabilities:

$$\beta := \frac{\mathrm{Cov}(Y,Z)}{\mathrm{Var}(Y)} = \Pr(Z=1|Y=1) - \Pr(Z=1|Y=0)$$

Proof:

$$
\begin{aligned}
\beta &= \frac{\Pr(Z=1, Y=1) - \Pr(Z=1)\Pr(Y=1)}{\mathrm{Var}(Y)} \\
&= \frac{\Pr(Z=1|Y=1)\Pr(Y=1) - \Pr(Z=1)\Pr(Y=1)}{\Pr(Y=1)\Pr(Y=0)} \\
&= \frac{\Pr(Z=1|Y=1) - [\Pr(Z=1)]}{\Pr(Y=0)} \\
&= \frac{\Pr(Z=1|Y=1) - [\Pr(Z=1|Y=1)\Pr(Y=1) + \Pr(Z=1|Y=0)\Pr(Y=0)]}{\Pr(Y=0)} \\
&= \frac{\Pr(Z=1|Y=1)[1 - \Pr(Y=1)] + \Pr(Z=1|Y=0)\Pr(Y=0)]}{\Pr(Y=0)} \\
&= \frac{\Pr(Z=1|Y=1)[\Pr(Y=0)] + \Pr(Z=1|Y=0)\Pr(Y=0)]}{\Pr(Y=0)} \\
&= \Pr(Z=1|Y=1) - \Pr(Z=1|Y=0)
\end{aligned}
$$

The slope of the linear probability model bounds the difference in conditional predictivity:

$$|\beta| := \left|\frac{\mathrm{Cov}(Y,Z)}{\mathrm{Var}(Y)}\right| \geq |\theta_{Z|Y=1} - \theta_{Z|Y=0}|$$

where

$$\theta_{Z|Y=1} := \mathrm{Pr(Z{=}1|Y{=}1)} \vee \mathrm{Pr(Z{=}0|Y{=}1)} = \frac{1}{2} + \frac{1}{2}|\mathrm{Pr(Z{=}1|Y{=}1)} - \mathrm{Pr(Z{=}0|Y{=}1)}|$$

and

$$\theta_{Z|Y=0} := \mathrm{Pr(Z{=}1|Y{=}0)} \vee \mathrm{Pr(Z{=}0|Y{=}0)} = \frac{1}{2} + \frac{1}{2}|\mathrm{Pr(Z{=}1|Y{=}0)} - \mathrm{Pr(Z{=}0|Y{=}0)}|$$

Proof (reverse triangle inequality):

$$
\begin{aligned}
||\theta_{Z|Y=1}| - |\theta_{Z|Y=0}|| &\leq |\theta_{Z|Y=1} - \theta_{Z|Y=0}| \\
&= \frac{1}{2}||\text{Pr(Z=1|Y=1)} - \text{Pr(Z=0|Y=1)}| - |\text{Pr(Z=1|Y=0)} - \text{Pr(Z=0|Y=0)}|| \\
&\leq \frac{1}{2}|\text{Pr(Z=1|Y=1)} - \text{Pr(Z=0|Y=1)} - \text{Pr(Z=1|Y=0)} + \text{Pr(Z=0|Y=0)}| \\
&= \frac{1}{2}|\text{Pr(Z=1|Y=1)} - (1 - \text{Pr(Z=1|Y=1)}) - \text{Pr(Z=1|Y=0)} + (1 - \text{Pr(Z=1|Y=0)})| \\
&= \frac{1}{2}|2[\text{Pr(Z=1|Y=1)} - \text{Pr(Z=1|Y=0)}]| \\
&= |\text{Pr(Z=1|Y=1)} - \text{Pr(Z=1|Y=0)}| \\
&= |\beta|
\end{aligned}
$$

The slope of the linear probability model bounds the difference in conditional and unconditional predictivity:

$$|\beta| := \left|\frac{\text{Cov}(Y, Z)}{\text{Var}(Y)}\right| \geq |\theta_{Z|Y=1} - \theta_Z|$$

where

$$\theta_{Z|Y=1} := \text{Pr(Z=1|Y=1)} \vee \text{Pr(Z=0|Y=1)} = \frac{1}{2} + \frac{1}{2}|\text{Pr(Z=1|Y=1)} - \text{Pr(Z=0|Y=1)}|$$

and

$$\theta_Z := \text{Pr(Z=1)} \vee \text{Pr(Z=0)} = \frac{1}{2} + \frac{1}{2}|\text{Pr(Z=1)} - \text{Pr(Z=0)}|$$

Cor: The coding $Y = 1$ is arbitrary; an identical argument establishes that $\beta \geq |\theta_{\theta_{Z|Y=0} - \theta_Z}$.

Proof of the upper bound $\theta_Z \leq |\beta| + \theta_{Z|Y=1}$ (triangle inequality):

$$
\begin{aligned}
2\theta_Z - 1 &= |\Pr(Z=1) - \Pr(Z=0)| \\
&= |\Pr(Z=1, Y=1) + \Pr(Z=1, Y=0) - \Pr(Z=0, Y=1) - \Pr(Z=0, Y=0)| \\
&\leq |\Pr(Z=1, Y=1) - \Pr(Z=0, Y=1)| + |\Pr(Z=1, Y=0) - \Pr(Z=0, Y=0)| \\
&= |\Pr(Z=1|Y=1) - \Pr(Z=0|Y=1)|\Pr(Y=1) + |\Pr(Z=1|Y=0) - \Pr(Z=0|Y=0)|\Pr(Y=0) \\
&= (2\theta_{Z|Y=1} - 1)\Pr(Y=1) + (2\theta_{Z|Y=0} - 1)\Pr(Y=0) \\
&= 2\theta_{Z|Y=1}\Pr(Y=1) + 2\theta_{Z|Y=0}\Pr(Y=0) - 1 \\
\Rightarrow \theta_Z &\leq \theta_{Z|Y=1}\Pr(Y=1) + \theta_{Z|Y=0}\Pr(Y=0) \\
&= \theta_{Z|Y=1}(1 - \Pr(Y=0)) + \theta_{Z|Y=0}\Pr(Y=0) \\
&= -(\theta_{Z|Y=1} - \theta_{Z|Y=0})\Pr(Y=0) + \theta_{Z|Y=1} \\
&\leq |\beta|\Pr(Y=0) + \theta_{Z|Y=1} \\
&\leq |\beta| + \theta_{Z|Y=1}
\end{aligned}
$$

Proof of the lower bound $\theta_Z \geq \theta_{Z|Y=1} - |\beta|$ (reverse triangle inequality):

First note that

$$
\begin{aligned}
\Pr(Z=1) &= \Pr(Z=1|Y=1)\Pr(Y=1) + [(\Pr(Z=1|Y=0)]\Pr(Y=0) \\
&= \Pr(Z=1|Y=1)\Pr(Y=1) + [\Pr(Z=1|Y=1) - \beta]\Pr(Y=0) \\
&= \Pr(Z=1|Y=1) - \beta\Pr(Y=0)
\end{aligned}
$$

and

$$\begin{aligned}\Pr(Z=0) &= \Pr(Z=0|Y=1)\Pr(Y=1) + [(\Pr(Z=0|Y=0)]\Pr(Y=0) \\ &= \Pr(Z=0|Y=1)\Pr(Y=1) + [\Pr(Z=0|Y=1) + \beta]\Pr(Y=0) \\ &= \Pr(Z=0|Y=1) + \beta\Pr(Y=0)\end{aligned}$$

imply $\Pr(Z=1) - \Pr(Z=0) = \Pr(Z=1|Y=1) - \Pr(Z=0|Y=1) - 2\beta\Pr(Y=0)$ so that

$$\begin{aligned}2\theta_Z - 1 &= |\Pr(Z=1) - \Pr(Z=0)| \\ &= |\Pr(Z=1|Y=1) - \Pr(Z=0|Y=1) - 2\beta\Pr(Y=0)| \\ &\geq |\Pr(Z=1|Y=1) - \Pr(Z=0|Y=1)| - |2\beta\Pr(Y=0)| \\ &= 2\theta_{Z|Y=1} - 1 - 2|\beta|\Pr(Y=0) \\ \Rightarrow \theta_Z &\geq \theta_{Z|Y=1} - |\beta|\Pr(Y=0) \\ &\geq \theta_{Z|Y=1} - |\beta|\end{aligned}$$

The average conditional predictivity exceeds the unconditional predictivity:

$$\mathrm{E}\theta_{Z|Y} = \theta_{Z|Y=1}\Pr(Y=1) + \theta_{Z|Y=0}\Pr(Y=0) \geq \theta_Z$$

Proof (triangle inequality)

$$
\begin{aligned}
2\theta_Z - 1 &= |\Pr(Z=1) - \Pr(Z=0)| \\
&= |\Pr(Z=1, Y=1) + \Pr(Z=1, Y=0) - \Pr(Z=0, Y=1) - \Pr(Z=0, Y=0)| \\
&\geq |\Pr(Z=1, Y=1) - \Pr(Z=0, Y=1)| + |\Pr(Z=1, Y=0) - \Pr(Z=0, Y=0)| \\
&= |\Pr(Z=1|Y=1) - \Pr(Z=0|Y=1)|\Pr(Y=1) + |\Pr(Z=1, Y=0) - \Pr(Z=0, Y=0)|\Pr(Y=0) \\
&= (2\theta_{Z|Y=1} - 1)\Pr(Y=1) + (2\theta_{Z|Y=0} - 1)\Pr(Y=0) \\
&= 2\theta_{Z|Y=1}\Pr(Y=1) + 2\theta_{Z|Y=0}\Pr(Y=0) - 1 \\
\Rightarrow \theta_Z &\leq \theta_{Z|Y=1}\Pr(Y=1) + \theta_{Z|Y=0}\Pr(Y=0) \\
&= \mathrm{E}\theta_{Z|Y}
\end{aligned}
$$

Example 1:

	Z = 0	Z = 1	Pr(Y)
Y = 0	**10d**	**d**	11d
Y = 1	**d**	**1**	d+1
Pr(Z)	11d	d + 1	12d + 1

$$
\Pr(Z=1|Y=1) = \frac{1}{d+1} \qquad \Pr(Z=0|Y=1) = \frac{d}{d+1}
$$

$$
\Pr(Z=1|Y=0) = \frac{1}{11} \qquad \Pr(Z=0|Y=0) = \frac{10}{11}
$$

$$
\Pr(Z=1) = \frac{d+1}{12d+1} \qquad \Pr(Z=0) = \frac{11d}{12d+1}
$$

$$\alpha = \Pr(Z = 1|Y = 1) - \Pr(Z = 1|Y = 0) = \frac{1}{d+1} - \frac{1}{11} \underset{d\to\infty}{\to} -\frac{1}{11} \approx -.09$$

$$\beta = \log\left(\frac{\Pr(Z = 1|Y = 1)}{\Pr(Z = 1|Y = 0)}\right) = \log\left(\frac{11}{d+1}\right) \underset{d\to\infty}{\to} -\infty$$

$$\gamma = \log\left(\frac{\Pr(Z = 1|Y = 1)\Pr(Z = 0|Y = 0)}{\Pr(Z = 1|Y = 0)\Pr(Z = 0|Y = 1)}\right) = \log\left(\frac{10d}{d^2}\right) \underset{d\to\infty}{\to} -\infty$$

$$\theta_Z = \frac{1}{2} + \frac{1}{2}|\Pr(Z = 1) - \Pr(Z = 0)| = \frac{1}{2} + \frac{1}{2}\left|\frac{d+1-11d}{12d+1}\right| \underset{d\to\infty}{\to} \frac{11}{12} \approx .92$$

$$\theta_{Z|Y=1} = \frac{1}{2} + \frac{1}{2}|\Pr(Z = 1|Y = 1) - \Pr(Z = 0|Y = 1)| = \frac{1}{2} + \frac{1}{2}\left|\frac{1-d}{d+1}\right| \underset{d\to\infty}{\to} 1$$

$$\theta_{Z|Y=0} = \frac{1}{2} + \frac{1}{2}\,|\Pr(Z = 1|Y = 0) - \Pr(Z = 0|Y = 0)| = \frac{1}{2} + \frac{1}{2}\left|\frac{1-10}{11}\right| = \frac{10}{11} \approx .91$$

$$\mathrm{E}(\theta_{Z|Y}) = \theta_{Z|Y=1}\Pr(Y = 1) + \theta_{Z|Y=0}\Pr(Y = 0) \underset{d\to\infty}{\to} 1 \times \frac{1}{12} + \frac{10}{11} \times \frac{11}{12} = \frac{11}{12}$$

The influence score (divided by n) is the sum of squared covariances:

$$\frac{1}{n}I := \frac{1}{n}\frac{\sum_j n_j^2(\bar{Y}_j - \bar{Y})^2}{n\sigma_Y^2} = \sum_j \frac{\mathrm{Cov}(X_j, Y)^2}{\sigma_Y^2}$$

where n_j is the number of the n observations falling in cell j and $\Pr(X_j = 1)$ is the probability an observation falls in cell j.

Proof:

$$
\begin{aligned}
\frac{1}{n}I &= \frac{\sum_j n_j^2(\bar{Y}_j - \bar{Y})^2}{n^2\sigma_Y^2} \\
&= \sum_j \left(\frac{n_j}{n}\right)^2 \left(\frac{\bar{Y}_j - \bar{Y}}{\sigma_Y}\right)^2 \\
&= \sum_j \left(\frac{n_j}{n} Y_j'\right)^2 \\
&= \sum_j \Big(\Pr(X_j = 1)\mathrm{E}(Y'|X_j = 1)\Big)^2 \\
&= \sum_j \Big(1\Pr(X_j = 1)\mathrm{E}(Y'|X_j = 1) + 0\Pr(X_j = 0)\mathrm{E}(Y'|X_j = 0)\Big)^2 \\
&= \sum_j \mathrm{E}(X_j Y')^2 \\
&= \sum_j \mathrm{Cov}(X_j, Y')^2 \\
&= \sum_j \frac{\mathrm{Cov}(X_j, Y)^2}{\sigma_Y^2}
\end{aligned}
$$

where $\mathrm{E}(Y'|X_j = 1) := \frac{1}{n_j}\sum_{i=1}^{n_j} \frac{Y_{ij} - \bar{Y}}{\sigma_Y} = \frac{\frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij} - \frac{1}{n_j}\sum_{i=1}^{n_j} \bar{Y}}{\sigma_Y} = \frac{\bar{Y}_j - \bar{Y}}{\sigma_Y} = Y_j'$ and $\Pr(X_j = 1) := \frac{n_j}{n}$

The influence score (divided by n) summarizes the between variation of the multivariate LPM:

Let $\Pr(X_j|Y = y) = \mathrm{E}(X_j|Y = y) = \alpha_j + \beta_j y$. Then

$$\frac{1}{n}I := \frac{1}{n}\frac{\sum_j n_j^2(\bar{Y}_j - \bar{Y})^2}{n\sigma_Y^2} = \sum_j \hat{\beta}_j^2$$

where $\hat{\beta}_j = \frac{\mathrm{Cov}(X_j, Y)}{\sigma_Y^2}$

Proof: By the iterated laws of variance and covariance,

$$\begin{aligned}
\mathrm{Var}(\mathrm{E}(X_j|Y)) &= \mathrm{Var}(\alpha_j + \beta_j Y) \\
&= \beta_j^2 \sigma_Y^2 \\
&\approx \hat{\beta}_j^2 \sigma_Y^2 \\
&= \frac{\mathrm{Cov}(X_j, Y)^2}{\sigma_Y^2}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Cov}(\mathrm{E}(X_{j_1}|Y), \mathrm{E}(X_{j_2}|Y)) &= \mathrm{Cov}(\alpha_{j_1} + \beta_{j_1} Y, \alpha_{j_2} + \beta_{j_2} Y) \\
&= \beta_{1j_1} \beta_{1j_2} \mathrm{Cov}(Y, Y) \\
&= \beta_{1j_1} \beta_{1j_2} \sigma_Y^2 \\
&\approx \hat{\beta}_{1j_1} \hat{\beta}_{1j_2} \sigma_Y^2 \\
&= \frac{\mathrm{Cov}(X_{j_1}, Y)\mathrm{Cov}(X_{j_2}, Y)}{\sigma_Y^2}
\end{aligned}$$

6.4.9 Derivations of the Influence Score

The influence score summarizes the explained variation.

The goal is to assess how well a partition of n objects, $i = \{1, \ldots, n\}$, represents a gold standard variable Y_i. For example, Y_i could be the height of the ith subject. In this section, we suppress the index i, writing Y for the measurement of an randomly chosen subject. We assume Y has been standardized so that $\mathrm{E}(Y) = 0$ and $\mathrm{Var}(Y) = 1$.

Let the partition be represented by a J-vector of dichotomous variables $X = [X_1, \ldots, X_J]$. For example, these variables could represent demographic characteristics like sex, race, and ethnicity.

However, we assume X is random, $X \sim \text{Multinomial}(p = [\alpha_1, \ldots, \alpha_J])$, that is, the partition has been randomly selected from a set of possible summarizing partitions, while Y, being the gold standard, is measured without error. We use the explained variation, $\text{E}(\text{Var}(X|Y))$, as the measure of assessment. Since the dependent variable is a vector, the use of this measure in this way is often called an Multivariate Analysis of Variance (MANOVA).

We assume a conditional linear probability model: $\Pr(X_j = 1|Y = y) = \beta_{0j} + \beta_{1j}y$. However, $\beta_{0j} = \alpha_j$ by the iterated law of expectation,

$$\alpha_j := \Pr(X_j = 1) = \text{E}(\Pr(X_j = 1|Y)) = \text{E}(\beta_{0j} + \beta_j Y) = \beta_{0j}$$

so we simply write $\Pr(X_j = 1|Y = y) = \alpha_j + \beta_j y$.

Since X is multinomial, we have the total variation

$$\begin{aligned}
\text{E}(X_j) &= \alpha_j \\
\text{Var}(X_j) &= \alpha_j(1 - \alpha_j) \\
\text{Cov}(X_{j_1}, X_{j_2}) &= -\alpha_{j_1}\alpha_{j_2}
\end{aligned}$$

and by the iterated laws of variance and covariance, the between variation,

$$\begin{aligned}
\text{Var}(\text{E}(X_j|Y)) &= \text{Var}(\alpha_j + \beta_j Y) = \beta_j^2 \\
\text{Cov}(\text{E}(X_{j_1}|Y), \text{E}(X_{j_2}|Y)) &= \text{Cov}(\alpha_{j_1} + \beta_{j_1}Y, \alpha_{j_2} + \beta_{j_2}Y) = \beta_{1j_1}\beta_{1j_2}
\end{aligned}$$

and within variation

$$\mathrm{E}(\mathrm{Var}(X_j|Y)) = \mathrm{E}[(\alpha_j + \beta_j Y)(1 - \alpha_j - \beta_j Y)] = \alpha_j(1 - \alpha_j) - \beta_j^2$$

$$\mathrm{E}(\mathrm{Cov}(X_{j_1}, X_{j_2}|Y)) = \mathrm{E}[-(\alpha_{j_1} + \beta_{j_1} Y)(\alpha_{j_2} + \beta_{j_2} Y)] = -\alpha_{j_1}\alpha_{j_2} - \beta_{j_1}\beta_{j_2}$$

In matrix form, we write $\Sigma_T = \mathrm{diag}(\alpha) - \alpha\alpha^T, \Sigma_B = \beta\beta^T$, and $\Sigma_W = \Sigma_T - \Sigma_B = \mathrm{diag}(\alpha) - \alpha\alpha^T - \beta\beta^T$, where α and β are column vectors $[\alpha_1, \ldots, \alpha_J]$ and $[\beta_1, \ldots, \beta_J]$. T, B, and W stand for the total, between and within.

Typically, a one-dimensional summary of Σ_B (or Σ_B relative to Σ_T) is used to assess fit. For example, Wilk's lambda $\Lambda = \Sigma_B\Sigma_W^{-1} = \Sigma_B(\Sigma_T - \Sigma_B)^{-1}$. Since our goal is prediction, we want to see whether Σ_B is large (see Miller (1997) and Breiman (1984). We use $\mathrm{tr}(\Sigma_B) = \sum_{j=1}^J \beta_j^2$ as our summary—this ignores the correlation structure in the data.

We estimate β_j using the least squares estimator: $\hat{\beta}_j = \frac{\sum_i (y_i - \bar{y})(x_{ij} - \bar{x}_j)}{\sum_i (y_i - \bar{y})^2} = \sum_i y_i x_{ij} = n_j \bar{y}_j$, where n_j is the number of observations in partition cell j and $\bar{y}_j$ is the average of y for the observations in partition j. The plug-in estimator of $\mathrm{tr}(\Sigma_B)$ is thus $\mathrm{tr}(\hat{\Sigma}_B) = \sum_{j=1}^J (\sum_i y_i x_{ij})^2 = \sum_{j=1}^J n_j^2 \bar{y}_j^2$. Consistency of the estimator follows from the consistency of least squares and the continuous mapping theorem.

For the percent of the variation explained, we can estimate $\mathrm{tr}(\hat{\Sigma}_B\hat{\Sigma}_T^{-1})$. However, for $\hat{\Sigma}_T$ to be invertible, the last row/column of Σ_B and Σ_T must be removed and the constraints $\sum_{j=1}^J \hat{\alpha}_j = 1$ and $\sum_{j=1}^J \hat{\beta}_j = 0$ enforced. The first constraint follows from the fact that the $\hat{\alpha}_j$ are probabilities. The second from the fact that $\sum_j \hat{\beta}_j = \sum_j n\,\mathrm{Cov}(Y, X_j) = n\,\mathrm{Cov}(Y, \sum_j X_j) = n\,\mathrm{Cov}(Y, 1) = 0$.

By the Sherman-Morrison Identity, $\hat{\Sigma}_T^{-1} = \mathrm{diag}([\frac{1}{\hat{\alpha}_1}, \ldots, \frac{1}{\hat{\alpha}_{J-1}}]) + \frac{1}{\hat{\alpha}_J} 1_{J-1\times J-1}$, where $1_{J-1\times J-1}$ is a $J-1 \times J-1$ matrix of 1s. Therefore,

$$\mathrm{tr}(\hat{\Sigma}_B\hat{\Sigma}_T^{-1}) = \hat{\beta}\hat{\beta}^T(\mathrm{diag}([\frac{1}{\hat{\alpha}_1}, \ldots, \frac{1}{\hat{\alpha}_{J-1}}]) + \frac{1}{\hat{\alpha}_J} 1_{J-1\times J-1}) = \sum_{j=1}^J \frac{\hat{\beta}_j^2}{\hat{\alpha}_j} = n \sum_{j=1}^J n_j \bar{y}_j^2$$

Adjustment for other LPM.

The least squares algorithm can also be applied in a similar fashion to the joint LPM or conditional LPM $Y|X$, although the partition and resulting influence score would need to be adjusted as we briefly outline with the joint LPM here.

We would first augment the partition $\Pi' \supset \Pi$ to include Y and regress the proportion of observations in each cell, $W = \frac{n_j}{n}$, on some or all of the partition indicators (no intercept), $x = [x_1, \ldots, x_{J'}]$ where $J' > J$. Under the linear probability model,

$$\begin{aligned} \text{Var}(\text{E}(W|X)) &= \text{Var}(X\beta) \\ &= \sum_{ij} \beta_i \beta_j \text{Cov}(X_i, X_j) \end{aligned}$$

and β can be estimated via least squares,

$$\hat{\beta} = (X^T X)^{-1}(X^T W) = \text{diag}([\frac{1}{m_1}, \ldots, \frac{1}{m_{J'}}])[m_1 \bar{W}_1, \ldots, m_{J'} \bar{W}_{J'}] = [\bar{W}_1, \ldots, \bar{W}_{J'}]$$

If the model is saturated, the $m_j = 1$. Otherwise, the least squares solution is the average of all partitions with the same coefficients.

The influence score as the unweighted residual sum of squares.

Suppose $Y_{ij} \sim \text{Bernoulli}(\theta_j), i \in \{1, \ldots, n_j\}, j \in \{1, \ldots, J\}$, with θ_j unknown. The question is whether θ_j can be approximated by $\theta_j^\star$ without appreciable loss. However, unless $\theta_j^\star$ is close to zero or one, observations must be combined to assess the goodness of fit.

Typically, the goodness of fit of $\theta_j^\star$ is assessed by combining the standardized residuals: $R_{ij} =$

$\frac{Y_{ij}-\theta_j^\star}{\sqrt{\theta_j^\star(1-\theta_j^\star)}}$. If $\theta_j^\star = \theta$, $E(Y_{ij}) = 0$ and $\text{Var}(Y_{ij}) = 1$, and the combined can be evaluated using the central limit theorem.

If residuals with the same fit are grouped together, inference can be done on the individual fit θ_j; this is possible when the observations have similar predictors. The grouped standardized residuals are, $R_{+j} = \sum_{i=1}^{n_j} R_{ij} = \frac{n_j(\bar{Y}_j-\hat{\theta}_j)}{\sqrt{\hat{\theta}_j(1-\hat{\theta}_j)}}$. When $\hat{\theta} = \theta$, $R_{ij} \sim \text{Normal}(0,1)$ and $R_{+j} \sim \text{Normal}(0, n_j)$. The standardized residuals can then be combined in any number of ways:

$$\begin{aligned}
\text{WLS}_{R_{ij}} = \text{OLS}_{R_{ij}} &= \sum_{ij} R_{ij}^2 = \sum_{ij} \frac{(Y_{ij}-\hat{\theta}_j)^2}{\hat{\theta}_j(1-\hat{\theta}_j)} \\
\text{WLS}_{R_{+j}} &= \sum_{j} \frac{R_{\cdot j}^2}{n_j} = \sum_{i} \frac{n_j(\bar{X}_j-\hat{\theta}_j)^2}{\hat{\theta}_j(1-\hat{\theta}_j)} = \sum_{ij} \frac{(\bar{Y}_j-\hat{\theta}_j)^2}{\hat{\theta}_j(1-\hat{\theta}_j)} \\
\text{OLS}_{R_{+j}} &= \sum_{j} R_{\cdot j}^2 = \sum_{i} \frac{n_j^2(\bar{Y}_j-\hat{\theta}_j)^2}{\hat{\theta}_j(1-\hat{\theta}_j)} = \sum_{ij} \frac{n_j(\bar{Y}_j-\hat{\theta}_j)^2}{\hat{\theta}_j(1-\hat{\theta}_j)}
\end{aligned}$$

$\text{WLS}_{R_{+j}}$ is equivalent to the Wald statistic and Neyman chi square approaches for categorical data (Bishop 2007, p. 352-354)

$\text{OLS}_{R_{+j}}$ provides a fair comparison of nested partitions:

Let Y_{ij} be cross classified data: $Y_{ij} \sim F_{ij}$ where $i \in \{1,\ldots,n_j\}, j \in \{1,\ldots,J\}, E(Y_{ij}) = \mu_{ij}$ and $\text{Var}(Y_{ij}) = \tau_{ij}$. Suppose Y_{ij} is to be approximated by θ_j where $\text{Var}(Y_{ij}|\theta_j) = \sigma_{\theta_j}^2$. The challenge is to assess the goodness of fit of $\theta = \{\theta_1,\ldots,\theta_J\}$.

Without knowledge of F, it is common to standardize the observations and appeal to the central limit theorem: Define standardized residual $R_{ij} = \frac{Y_{ij}-\theta_j}{\sigma_{\theta_j}}$ and grouped residual $R_{+j} = \sum_{i=1}^{n_j} R_{ij} = \frac{n_j(\bar{Y}_j-\theta_j)}{\sigma_{\theta_j}}$. If $\mu_{ij} = \theta_j$, $E(R_{ij}) = 0$, $\text{Var}(R_{ij}) = 1$, and R_{+j} is approximately normal with mean zero and standard deviation $\sqrt{n_j}$. A small sum of squared residuals, $\sum_{j=1}^{J} R_{+j}^2$, is consistent with fit $\mu_{ij} = \theta_j$, and a large sum of squared residuals is inconsistent.

If the residuals are uncorrelated, the expected sum of squared residuals is not reduced with a finer partition. Suppose the jth group were subdivided into K_j subgroups of sizes $\{n_1, \ldots, n_{k_j}, \ldots, n_{K_j}\}$ so that $\sum_{k_j=1}^{K}{}_j n_{kj} = n_j$. Let R_{ik_j} denoting the ith observation in the k_jth group.

The sum of residuals is preserved: $R_{+j} = \sum_{i=1}^{n_j} R_{ij} = \sum_{k_j=1}^{K}{}_j \sum_{i=1}^{n_{kj}} R_{ik_j} = \sum_{k_j=1}^{K_j} R_{+k_j}$. It follows that

$$R_{+j}^2 = (\sum_{k_j=1}^{K_j} R_{+k_j})^2 = \sum_{k_j=1}^{K_j} R_{+k_j}^2 + \sum_{k_j}^{K_j} \sum_{k_j' \neq k_j}^{K_j} R_{+k_j} R_{+k_j'} \Rightarrow R_{+j}^2 - \sum_{k_j=1}^{K_j} R_{+k_j}^2 = \sum_{k_j}^{K_j} \sum_{k_j' \neq k_j}^{K_j} R_{+k_j} R_{+k_j'}$$

where $\sum_{k_j}^{K_j} \sum_{k_j' \neq k_j}^{K_j} R_{+k_j} R_{+k_j'} = \sum_{k_j}^{K_j} R_{+k_j} \sum_{k_j' \neq k_j}^{K}{}_j R_{+k_j'} = \sum_{k_j}^{K_j} R_{+k_j} R_{+(-k_j)}$

If the residuals are uncorrelated,

$$E(R_{+j}^2 - \sum_{k_j=1}^{K_j} R_{+k_j}^2) = E(\sum_{k_j}^{K_j} \sum_{k_j' \neq k_j}^{K_j} R_{+k_j} R_{+k_j'}) = \sum_{k_j}^{K_j} \sum_{k_j' \neq k_j}^{K_j} E(R_{+k_j}) E(R_{+k_j'}) = 0$$

6.5 References

Agresti, Alan. 2003. Categorical data analysis. Vol. 482. John Wiley Sons.

Armitage, Peter. 1958. "Numerical studies in the sequential estimation of a binomial parameter." Biometrika 45.1/2, 1–15.

Battey, HS, DR Cox, and MV Jackson. 2019. "On the linear in probability model for binary data." RoyalSociety open science 6.5, 190067.

Berkson, Joseph. 1951. "Why I prefer logits to probits." Biometrics 7.4, 327–339.

Bickel, Peter J and Kjell A Doksum. 2015. Mathematical Statistics: Basic Ideas and Selected Topics, Volumes II CRC Press.

Bishop, Yvonne M, Stephen E Fienberg, and Paul W Holland. 2007. Discrete multivariate

analysis: theory and practice. Springer Science Business Media.

Breiman, Leo, Jerome H Friedman, Richard A Olshen, and Charles J Stone. 1984 Classification and regression trees. Wadsworth Brooks/Cole Advanced Books. Monterey, CA.

Breslow, Norman E, Nicholas E Day, and Elisabeth Heseltine. 1980. Statistical methods in cancer research. Vol. 1. International Agency for Research on Cancer.

Chernoff, Herman, Shaw-Hwa Lo, and Tian Zheng. 2007. Discovering influential variables: a method of partitions. The Annals of applied statistics. 1335–1369.

Cochran, William G. 1940. The analysis of variance when experimental errors follow the Poisson or binomial laws. The Annals of Mathematical Statistics 11.3 p. 335–347.

Cochran, William G. 1955. "A test of a linear function of the deviations between observed and expected numbers." Journal of the American Statistical Association 50.270, 377–397.

Cox, David R. 1958. "The regression analysis of binary sequences." Journal of the Royal Statistical Society: Series B (Methodological). 20.2, 215–232.

Cox, David R and Joyce Snell 1989. Analysis of binary data. Vol. 32. CRC press.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. The elements of statistical learning. Springer series in statistics New York.

Gelman, Andrew and Jennifer Hill. 2006. Data analysis using regression and multilevel/hierarchical models. Cambridge university press.

Lo, Adeline, Herman Chernoff, Tian Zheng, and Shaw-Hwa Lo. 2016. "Framework for making better predictions by directly estimating variables' predictivity". Proceedings of the National Academy of Sciences. 113.50 14277–14282.

Miller, Rupert G Jr. 1997. Beyond ANOVA: basics of applied statistics. CRC press

Nelder, John Ashworth and Robert WM Wedderburn. 1972. "Generalized linear models." Journal of the Royal Statistical Society: Series A (General). 135.3, 370–384.

King, Gary. 1998. Unifying political methodology: The likelihood theory of statistical inference. University of Michigan Press.

Yule, Udny 1912. "On the methods of measuring association between two attributes." Journal

of the Royal Statistical Society 75.6, 579–652.