Estimating Risk-adjusted Process Performance with a Bias/Variance Trade-off

by
Patricia L. Cooper Barfoot

A thesis
presented to The University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2017

Author’s Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Decision makers responsible for managing the performance of a process commonly base their decisions on an estimate of present performance, a comparison of estimates across multiple streams, and the trend in performance estimates over time. Their decisions are well-informed when the risk-adjusted estimates of the performance measure (or parameter) are accurate and precise. The work is motivated by three applications to estimate a parameter at the present time from a stream of data where the parameter drifts slowly in an unpredictable way over time. It is common practice to estimate its value using either present time data only or using present and historical data. When sample sizes by time period are small, an estimate based on present time data is imprecise and can lead to uninformative or misleading conclusions. We can choose to estimate the parameter using an aggregate of historical and present time data but this choice trades more bias for less variability when the parameter is drifting over time. We propose to regulate the bias/variance trade-off using estimating equations that down-weight past data. We derive approximations for the variance of the estimator and the distribution of a hypothesis test statistic involving the estimator through known asymptotic properties of the estimating functions. We study the proposed approach relative to current practices with real or realistic data from each application. We offer simulations and analytic examples to generalize the comparisons and validate the approximations. We explore considerations related to implementing the proposed approach. We suggest future work to extend the applicability of this work.

Acknowledgements

I am extremely thankful to my village.

To Stefan and Jock, thanks for your dedication and patience and your valuable ideas. You have taught me and challenged me and I am wiser and stronger, thanks to you.

Thank you to my examining committee, Dr. S. Steiner, Dr. R.J. MacKay, Dr. M. Schonlau, Dr. Y. Zhu, Dr. P. Castagliola, and Dr. F. Gzara, for your attention to my work.

To Rick, I wouldn't have finished this without you. Thanks for being the wind beneath my wings.

To Alistair, Adrian, and Hillary, I've been working at this for as long as you can remember. Thanks for sharing your Mommy time selflessly and being my best cheering section. Please follow your passions and do hard things in your lives too.

To Mom and Dad, Christine and Candace and your families, and my friends… thanks for encouraging me and brightening my days at every turn.

Dedication

To my most precious people and the memory of those departed who give me confidence.

Table of Contents

List of Figures

Chapter 6: Hospital Performance Measure

List of Tables

List of Abbreviations

CMS	Centers for Medicare and Medicaid Services
DOH	Department of Health
EWMA	Exponentially Weighted Moving Average
FOBT	Fecal Occult Blood Test
GEE	Generalized Estimating Equations
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
IPWGEE	Inverse Probability Weighted Generalized Estimating Equations
LR	Likelihood Ratio
MLE	Maximum Likelihood Estimate
MSE	Mean Squared Error
MWLE	Maximum Weighted Likelihood Estimate
NPS	Net Promoter Score
NYS	New York State
PCI	Percutaneous Coronary Intervention
WEE	Weighted Estimating Equations
WI	Weighted Information

Chapter 1: Motivation and Introduction

Decision makers responsible for managing the performance of a process commonly base their decisions on an estimate of a process parameter such as the mean or a rate over time. Their decisions are well-informed when the current estimate of the parameter is accurate and precise. Specifically, a decision to focus re-engineering efforts to improve the performance of a product or service requires an efficient estimate of the present mean outcome. Validation of previous efforts requires that estimates be tracked and compared over time. Estimating the parameter can be particularly challenging when we observe the outcome from a small number of process subjects and performance changes over time in an unpredictable way.

Commonly there is one or more subject-level covariates that have an effect on the subject-level outcome that is observed. We may want to divide the subjects into multiple subgroups of interest which we refer to as streams. Considering that we observe a stream of outcome and possibly covariate data over time, we consider the objectives as follows,

- Monitor an estimate of the performance parameter that is risk-adjusted for a changing subject population over time. For example, Spiegelhalter et al. (2012) describe the importance of healthcare surveillance for rapid detection of emerging problems. Healthcare regulators need a measure of each healthcare provider's performance to assess against a relevant target or threshold. We may observe a small number of subjects in the current time period and the performance of a particular healthcare provider may drift over time in an unpredictable way.
- Monitor a comparison of parameter estimates across multiple streams. For example, Liu, MacKay, and Steiner (2008) describe the problem to monitor six wheel alignment characteristics in a truck assembly process where one of four possible gauges is used to measure the alignment characteristics on a particular truck. In order to maintain consistent testing, it is important that any differences among the mean outcome of the four gauges are detected. There may be a small number of trucks tested at one of the various gauges and the performance of a particular gauge may drift over time in an unpredictable way.

Such monitoring activities are used to identify problems, motivate the need for improvement, and quantify the extent to which improvement initiatives have been successful. These are different problems than usual statistical process monitoring applications where a mean measure or a

hypothesis test statistic is monitored relative to an in-control period determined by prior data (Montgomery, 2013). In the problem of study, decisions are based on an updated estimate of a parameter rather than a comparison to control limits.

There are two naïve approaches in common use to estimate the present value of a parameter based on a stream of data collected over time. The parameter estimate may be based on data observed at the present time only. This estimate is imprecise when the current sample size is small. The parameter estimate may be based on an aggregate of historical data without regard to the time period of the data. This estimate is biased when the true value of the parameter changes over time. We look for an approach to combine present and historical data that regulates the bias/variance trade-off for a more efficient estimate of the present parameter than either of the two naïve approaches. We consider the case where the parameter changes slowly over time.

This research is motivated by three applications having one or both of the stated objectives. Efficient estimates of a present time process parameter based on a stream of data are important for decision makers in all three applications. In each application, the common practice for analysis is one of the two naïve approaches previously discussed. In Section 1.1, we summarize the motivation, data, and objectives of these applications. More detail is provided in Chapters 4, 5, and 6.

1.1. Motivating applications

1.1.1. Customer loyalty measure

Compelling evidence (Reichheld and Markey, 2011) shows that a customer's response to the loyalty question coined the "ultimate question" is a good indicator of the likelihood of retaining that customer in the future. The ultimate question, 'How likely is it that you would recommend this company or product to a friend or colleague?', solicits a response on a scale from 0 to 10. The customer's response classifies them into one of the three categories

- detractors who respond six or below
- passives who respond seven or eight
- promoters who respond nine or ten.

Reichheld and Markey (2011) state that customers in these categories exhibit distinct loyalty behaviours. The actions that the company needs to take to encourage them to repurchase and promote the product are distinct. The difference between the proportions of customers who are promoters and detractors is a measure known as the Net Promoter Score (NPS). Increasing the

proportion of promoters, decreasing the proportion of detractors, or doing both simultaneously increases the value of NPS.

The NPS measure for customer loyalty is adopted in many industries selling consumer and enterprise products and services ("NPS Benchmarks", n.d.). Decision makers track NPS to plan efforts to improve customer loyalty and assess previous efforts. For these activities, representative population estimates of NPS are updated at frequent, regular intervals (Reichheld and Markey, 2011). We consider a real example for a smartphone vendor. In this application, a manager monitors the NPS estimates to plan loyalty-building efforts such as expanding the customer base in product lines or subpopulations where promoters are more numerous than detractors. For example, a decision could be made to adjust price or marketing campaigns to increase sales of a product line with a high NPS. The common practice is estimation using either data from the most recent time period or an aggregate of data across multiple time periods. Commonly, too little attention is paid to the uncertainty and bias in estimates by these practices and resulting estimates may be uninformative and misleading. We want to regulate a bias/variance trade-off in the estimate of NPS. Because there are many factors that impact customer loyalty and these change over time, we expect that true customer loyalty of the population drifts slowly over time in an unpredictable way.

Data

Consider responses to a survey asking the ultimate question from samples of customers over time.

- There are 19,981 customers who responded to the customer loyalty survey over a 42-week period.
- The customers providing the responses are not the same over time and are not identified.
- Each response on an 11-point scale is categorized into one of the three loyalty groups: detractor, passive, or promoter.
- The data are realistic.
- Responses are summarized in weekly subgroups.
- The number of responses by week varies and is small in some weeks.
- There are two covariates to describe the customer's smartphone:
    - product variant, an identifier having possible values $\{1,2,3,4\}$

    - tenure, lower bound of time in months that the smartphone has been in service having possible values $\{0,2,6,12,18,24\}$

- We assume there is no interaction among covariate effects and covariate effects do not change over time.

Objectives

Based on the stream of data, we want to:

1. estimate a mean NPS for the latest week
2. track trends in NPS over time
3. compare NPS for two product variants

Challenges

- There are varying numbers of customers observed weekly and in some weeks the sample size is small. As a result, the estimate of NPS may be imprecise for the latest week and differences in NPS over time may be difficult to detect.
- The true NPS changes slowly over time due to the effects of unobserved factors. As a result, estimates based on data from past time periods are biased estimates of NPS for the latest week.

1.1.2. Lab positive abnormal rate

In both Canada and the United States, a regulatory body oversees the proficiency of laboratories conducting medical diagnostic testing. Data from regular operation of the labs performing a particular test is monitored relative to international standards and compared among peers. Non-compliance and unfavourable performance measures have important implications for licensing continuance and for attracting patients. In the case of a test having two possible outcomes, the binomial distribution is a standard model to estimate positive abnormal rate from observations of either a positive or negative abnormal test outcome. Hypothesis tests based on the estimates of the positive abnormal rates and their uncertainties can compare labs to each other or to a standard. Test data are collected at regular intervals (for example, monthly) and so a periodic stream of test outcome data is available across multiple labs. We consider a real application where regulators monitor labs that perform a fecal occult blood test in Ontario. Here, the decision makers monitor the estimates of positive abnormal rate based on the data observed in the latest month by lab. We note that there are sizable differences in monthly sample size between the labs and some samples

sizes are small. Since sample size has an important impact on the probability that a lab is classified as non-conforming or different than the others, a better approach is sought. We want to reduce the uncertainty in the estimates and improve the power of the tests to detect differences across levels by leveraging historical data. We expect that true lab performance drifts slowly over time in an unpredictable way.

Data

Consider outcome data from the set of labs conducting the same fecal occult blood test (FOBT) in Ontario over time.

- There are 863,898 FOBT tests performed at one of seven labs in Ontario between January 2014 and June 2015.
- The data are observed by patient, by lab, and by month.
- The patients under test are not the same over time and are not identified.
- The data are real and were provided by Cancer Care Ontario.
- The number of patients by lab by month varies and is small for some labs in some months.
- There are two possible outcomes for each patient: positive or negative abnormal.
- There are no available covariate data describing labs or patients.

Objectives

Based on the stream of data, we want to:

1. estimate the mean positive abnormal rate ("positive rate") by lab for the latest month
2. compare the positive rates across all labs for the latest month
3. detect those labs which have a higher positive rate than their peers

Challenges

- There are varying numbers of patients tested by lab and by month and some are small. As a result, the estimates of positive rate by lab for the latest month may be imprecise and hypothesis tests may not detect important differences between labs.
- The true positive rate by lab changes slowly over time due to the effects of unobserved factors. As a result, estimates based on data from multiple time periods are biased estimates of positive rate for the latest month.

1.1.3. Hospital performance measure

Statistical models for predicting hospital performance are increasingly of interest to health planners, regulators, and patients (Clark, Hannan, and Wu, 2010, COPPS-CMS White Paper Committee, 2012). In the United States, the Centers for Medicare and Medicaid Services (CMS) has a congressional mandate to evaluate hospital performance using risk-adjusted mortality rates. Additionally, the New York State (NYS) Department of Health (DOH) publishes annual reports stating performance estimates for each of their hospitals performing percutaneous coronary intervention on patients with coronary artery disease (New York State Department of Health, 2015). For both applications, the performance estimates must reflect the quality of surgical care by adjusting for differences in patient health at admission across different hospitals but not adjust away differences related to the quality of the hospital. An additional requirement is that the reported performance measure should be affected as little as possible by the variability resulting from small numbers of patients treated at some hospitals.

The two applications use different approaches to estimate hospital performance. The NYSDOH estimates the hospital-specific performance measure through a risk-adjusted, naïve estimate of the observed mortality rate for a particular hospital. There is a high degree of instability and uncertainty in the NYSDOH estimate of performance for a low volume hospital in particular. The CMS uses an approach recommended by the COPPS-CMS White Paper Committee (2012) that estimates hospital-specific performance through risk-adjusted prediction of the mortality rate for a particular hospital. The predicted mortality rate is based on a hierarchical, random effects model that stabilizes the estimate of the hospital-specific performance measure. Criticism of the CMS approach points out that estimates for small, low volume hospitals have little value as they are close to the national mean (COPPS-CMS White Paper Committee 2012, pg. 24). Both approaches pool data over a three-year time period in order to improve estimates for low volume hospitals. We note that though pooling data reduces uncertainty, this approach increases bias in an estimate of the present time performance when performance changes over time. Considerable uncertainty may remain. We seek an alternative approach to improve estimates based on small samples utilizing the stream of test outcome data. We expect that hospital performance may drift slowly over time in an unpredictable way.

Data

Consider outcome data of coronary artery disease patients following percutaneous coronary intervention (PCI) over time.

- There are 467,401 patients who underwent PCI at one of 60 hospitals in NYS between 2004 and 2012.
- The PCI patients are not the same over time and are not identified.
- There are two possible patient outcomes: death or survival during the same hospital stay in which the patient underwent PCI or after hospital discharge but within 30 days of surgery.
- The observations of death or survival are available by patient, by hospital, by year.
- The data are realistic – they were derived to have the same characteristics as data in the annual NYS DOH reports over this period (New York State Department of Health, 2015).
- There are eight patient-level covariates to describe the risk of death for the patient at time of admission (not to include any attributes related to hospital performance):
    - patient age; an integer value in years greater than 55
    - hemodynamic state ∈ {'stable', 'unstable'}
    - ventricular ejection fraction $\in \{\geq 40\%, < 20\%, 20-29\%, 30-39\%\}$
    - pre-procedural myocardial infarction $\in \{$'none within 14 days', < 6 hrs, $6-11$ hrs, $12-23$ hrs, $1-14$ days$\}$
    - congestive heart failure ∈ {'no', 'current within 2 weeks'}
    - chronic lung disease ∈ {'no', 'yes'}
    - renal failure creatinine level $\in \{\leq 1.5, 1.6-2.0$ mg/dl, >2.0 mg/dl, 'requires dialysis'$\}$
    - malignant ventricular arrhythmia ∈ {'no', 'yes'}
- We assume that the patient-level covariate effects are the same for all hospitals.

Objective

Based on the stream of data, we want to:

- estimate a mean mortality rate by hospital for the latest year
- track trends in mortality rate by hospitals over time
- detect those hospitals which have a higher mortality rate than their peers

Challenges

- The number of patients who undergo PCI by year varies and may be small. There may be few or no observations for patients in a particular hospital in some years. As a result, the

estimates of mortality rate by hospital for the latest year may be imprecise and differences in mortality rate by hospital over time may be difficult to detect.

- The mortality rate changes slowly over time due to the effects of unobserved factors. As a result, estimates based on data from multiple time periods are biased estimates of mortality rate for the latest year.

1.2. General problem

We introduce the general problem and notation that applies to the three motivating examples which is used throughout this thesis.

Data and model

We consider the following data and model. At each time period t, we observe data d_t from a sample of n_t subjects. Note these are not panel data so the subject identifiers do not contain any information. Refer to subject j at time period t with $j = 1, \dots, n_t,\ t = 1, \dots, T$ where n_t is the number of subjects observed at time t and $t = T$ is the present time period. There may be a subject-specific characteristic of interest that divides the subjects into multiple streams. We identify the stream by subscript $m \in \{1, \dots, M\}$ and refer to subject j in stream m at time t with $j = 1, \dots, n_{mt}$, $m = 1, \dots, M,\ \sum_{m=1}^{M} n_{mt} = n_t$, and $t = 1, \dots, T$.

The data d_t includes an outcome response from each of the n_t subjects at time t which we refer to as $y_t = \{y_{jt}; j = 1, \dots, n_t\}$ or $y_t = \{y_{jmt}; j = 1, \dots, n_{mt}, m = 1, \dots, M\}$. The d_t may also include observed values of subject-specific covariates which we refer to as $x_t = \{x_{jt}; j = 1, \dots, n_t\}$ or $x_t = \{x_{jmt}; j = 1, \dots, n_{mt}, m = 1, \dots, M\}$. We refer to a $(s \times 1)$ vector of covariate values for subject j at time t as $x_{jt} = (x_{1,jt}, \dots, x_{s,jt})^T$ or for subject j in stream m at time t as $x_{jmt} = (x_{1,jmt}, \dots, x_{s,jmt})^T$. The covariate variables may be discrete or continuous.

There is a single random variable Y_{jmt} to describe the response y_{jmt} for subject j in stream m at time t (or Y_{jt} in the case of a single stream). The random variable may be continuous, $Y_{jmt} \in \mathcal{R}$, or categorical/ordinal, $Y_{jmt} \in \{k = 1, \dots, K\}$ with K possible levels. In the motivating applications to estimate rates in two or three groups, we consider the cases where $K = 2$ or $K = 3$. We assume that random variables $\{Y_{jmt}, j = 1, \dots, n_{mt}, m = 1, \dots, M, t = 1, \dots, T\}$ are independent for all j, m, and t, conditional on the values of covariates, $\{x_{jmt}, j = 1, \dots, n_{mt}, m = 1, \dots, M, t = 1, \dots, T\}$.

We assume that the response Y_{jmt} can be described by a generalized linear model (GLM) as a

function of the covariate vector x_{jmt} and a p-dimensional model parameter, θ_t. The elements of the parameter vector θ_t include α_t and may include either or both of δ_t, β_t, all of which may be vectors. In this thesis, α_t relates to the mean performance for a subject with a baseline level of the covariates at a baseline stream and δ_t relates to the effects of the various streams and β_t to the effects of the covariates on the performance mean.

According to usual practice (McCullagh and Nelder, 1989), denote the GLM by

- $f_Y(y)$, the distribution function for the response with mean $E(Y_{jmt}) = \pi_{jmt}$. Note that since the motivating applications involve rates, then we use the common notation π for a rate. Similarly, the common notation μ for the mean could be used.
- $\eta_{jmt} = h(x_{jmt}, m, \theta_t)$, a linear predictor, and θ_t and $g(\pi_{jmt}) = \eta_{jmt}$, a link function relating the parameter π_{jmt} to the linear predictor.

The selections of f, g, and h are based on the nature of the data and we assume that they do not change over time.

Objectives

Based on the data, we want an accurate and precise estimate of the current (at time T) parameter, $\theta = \theta_T$. Depending on the application of study, we might want to

- compare the estimate of the parameter to a target or benchmark value
- compare the risk-adjusted mean for stream m to a target or benchmark value
- compare risk-adjusted mean estimates across streams
- compare mean estimates for stream m across groups of subjects
- monitor estimates of the parameter over time
- monitor estimates of the risk-adjusted mean for stream m over time
- test hypotheses on elements of the parameter
- test hypotheses on the risk-adjusted mean for stream m

These objectives require the estimates $\hat{\theta}$ of $\theta = \theta_T$, the model parameter at the present time T, and $\hat{\pi}_m$ of π_m, the risk-adjusted mean performance for stream m at the present time T. As well, we require estimates of $var(\hat{\theta})$, $var(\hat{\pi}_m)$, and hypotheses test statistics involving $\hat{\theta}$. Estimates may be required over multiple time periods, over various groups of subjects, or subject to a null hypothesis. We specify the estimates required for each of these objectives in Chapter 3.

Approach

- Our objectives involve mean values that are adjusted for different distributions of the covariates among the samples in various streams or in various time periods. We estimate the risk-adjusted mean π_m for a single population known as the standard population which is a fixed set of values of the covariates representing subjects in a population of importance. Denote the subjects in the standard populations as $j^* = 1, \dots, J^*$ and select their covariate values $\left\{x_{j^*} = \left(x_{1,j^*}, \dots, x_{s,j^*}\right)^T \text{ for } j^* = 1, \dots, J^*\right\}$. Further discussion on the standard population is given in Section 1.3.
- Fit the GLM to the observed data $\{y_{jmt}\}$ and $\{x_{jmt}\}$ for $j = 1, \dots, n_{mt}, m = 1, \dots, M$, and $t = 1, \dots, T$ assuming that the associated random variables are independent, conditional on the values of the covariates. There are possibilities for which data to include, assumptions relating the various θ_t, $t = 1, \dots, T$ and methods to combine $\{\hat{\theta}_t\}$ as discussed in Chapters 2 and 3. The objective is an estimate for $\theta = \theta_T$ which we refer to as $\hat{\theta}$.
- Estimate π_m, the mean for subjects in the fixed standard population $\{x_{j^*}, j^* = 1, \dots, J^*\}$ in stream m. We refer to $\hat{\pi}_m$ as a risk-adjusted estimate.
- Calculate a hypothesis test statistic $\hat{S}$ involving $\hat{\theta}$ and estimates of some or all elements of $\hat{\theta}$ under a null hypothesis versus a specified alternative hypothesis.

Challenges

- There are varying sample sizes by time period and n_T or some $n_{mT}, m = 1, \dots, M$ may be small. As a result, the estimates of θ and π_m based on $\{y_{jmT}, j = 1, \dots, n_{mT}\}$ may be imprecise making inference difficult.
- Some elements of parameter θ_t may change slowly over time $t = 1, \dots, T$ in an unpredictable way due to the effects of unobserved factors, so the true value of π_m changes slowly over time. We need to be careful in the selection of what data to include and how to combine estimates across time periods to control bias in $\hat{\theta}$ and $\hat{\pi}_m$ based on $\{d_t, t = 1, \dots, T\}$. We do not want to assume a stochastic or deterministic model to describe the change in θ_t since the change may be hard to predict and we want our approach to be flexible.
- Some response or covariate data may be missing.

Out of scope

We recognize that the following conditions may impact the results but are not considered here. Consideration of these conditions is future work.

- changes in the true value of the parameter may not be small and may be predictable in some way
- the observations may be serially correlated
- there may be non-response bias and other types of sampling bias
- there may a time lag to gather or prepare the data for analysis
- measurements of the response may have error
- there may be outliers in the data
- important covariates describing subject-to-subject variation may be missing

1.3. Standard population

The first step in the approach to the general problem is to define a standard population which is a fixed set of values of the subject-level covariates representing subjects in a population of importance. The estimates of the mean response are made for subjects in the standard population. The use of a standard population is a risk-adjustment technique to adjust for differences among covariate levels observed in samples over time and to reliably compare estimates across time. It is important that the same standard population be used for each estimate over time.

The definition of the standard population is subjective but should reflect some population of importance. Some possible examples include the field population of subjects if this is known, an important segment of the population, or a typical subject. For appropriate interpretation, the definition of the standard population should be communicated with the estimates of the mean response for that population. It is possible to define more than one standard population and provide estimates of the mean response for each as long as the definitions of the standard populations are clearly communicated. The number of subjects in the standard population, J^*, depends on the definition of the standard population. In the previous examples, J^* may be the size of the field population or the segment of interest or $J^* = 1$ if there is a single subject of interest. Sample definitions of the standard population for the motivating applications follow.

Customer loyalty measure

- known field population of 10,000 customers at week 42 ($J^* = 10{,}000$)

- the most common customer segment: product variant 3 and tenure 6 months ($J^* = 1$)

Lab positive abnormal rate

- there is no standard population as there are no subject-specific covariates

Hospital performance measure

- population of patients who underwent PCI in 2012 across all NYS hospitals ($J^* = 47{,}045$)
- high risk patient segment ($J^* = 5$):
    - age $\in \{71 - 75\}$
    - hemodynamic state $= '\text{unstable}'$
    - ventricular ejection fraction $< 20\%$
    - pre-procedural myocardial infarction < 6 hrs
    - congestive heart failure = 'current within 2 weeks'
    - chronic lung disease $= '\text{yes}'$
    - renal failure creatinine level is $= '\text{requires dialysis}'$
    - malignant ventricular arrhythmia $= '\text{yes}'$

1.4. Models for motivating applications

In Table 1-1, we apply the general notation to models for the three motivating applications.

Table 1-1. Models for motivating applications

	Customer loyalty measure	**Lab positive abnormal rate**	**Hospital performance measure**
Data, d_t $t = 1, \dots, T$	y_{jt}: categorized response to ultimate question $x_{jt} = (x_{1,jt}, \dots, x_{4,jt})^T$: 3 indicator values representing product variant and interval value of tenure j: customer $\in \{1, \dots, n_t\}$ t: week $\in \{1, \dots, 42\}$	y_{jmt}: positive or negative abnormal test result j: subject $\in \{1, \dots, n_{mt}\}$ m: lab $\in \{1, \dots, 7\}$ t: month $\in \{1, \dots, 18\}$	y_{jmt}: death or survival at 30 days post-surgery $x_{jmt} = (x_{1,jmt}, \dots, x_{15,jmt})^T$: 1 integer value and 14 indicator values representing 8 patient risk factors at admission j: patient $\in \{1, \dots, n_{mt}\}$ m: hospital $\in \{1, \dots, 60\}$ t: year $\in \{1, \dots, 9\}$
Parameters of interest	π_1, π_3: proportions of customers who are detractors, promoters at time T	π_m: positive abnormal rate in lab m at time T	π_m: mortality rate following surgery in hospital m at time T
Distribution of $\mathcal{D}_t$	$Y_{jt} \sim multinomial(1, \pi_{1,jt}, 1 - \pi_{1,jt} - \pi_{3,jt}, \pi_{3,jt})$	$Y_{jmt} \sim binomial(1, \pi_{mt})$	$Y_{jmt} \sim binomial(1, \pi_{jmt})$
Linear predictor	$\eta_{1,jt} = \alpha_{1,t} + \beta_t^T x_{jt}$ $\eta_{2,jt} = \alpha_{2,t} + \beta_t^T x_{jt}$	$\eta_{mt} = \alpha_t + \delta_t I_m$ elements of $I_m \in \{0,1\}$ depending on m	$\eta_{jmt} = \alpha_t + \delta_t I_m + \beta_t^T x_{jmt}$ elements of $I_m \in \{0,1\}$ depending on m
Link function	$\eta_{1,jt} = \log\left\{\frac{\pi_{1,jt}}{1-\pi_{1,jt}}\right\}$ $\eta_{2,jt} = \log\left\{\frac{1-\pi_{3,jt}}{\pi_{3,jt}}\right\}$	$\eta_{mt} = \log\left\{\frac{\pi_{mt}}{1-\pi_{mt}}\right\}$	$\eta_{jmt} = \log\left\{\frac{\pi_{jmt}}{1-\pi_{jmt}}\right\}$
Model parameters	$\theta_t = (\alpha_t^T, \beta_t^T)^T$ $p = 6$	$\theta_t = (\alpha_t, \delta_t^T)^T$ $p = 7$	$\theta_t = (\alpha_t, \delta_t^T, \beta_t^T)^T$ $p = 75$
Objectives	1. Estimate $\theta = \theta_T$ 2. Estimate π_1, π_3, NPS for a standard population $\{x_{j^*}\}$ 3. Track estimates of NPS over time 4. Test H_0: $\beta_2 = \beta_3$	1. Estimate $\theta = \theta_T$ 2. Estimate π_m for $m \in \{1, \dots, 7\}$ 3. Test H_0: $\delta = \delta_0$ for some fixed value δ_0	1. Estimate $\theta = \theta_T$ 2. Estimate π_m, $var(\pi_m)$ for a standard population $\{x_{j^*}\}$ and $m \in \{1, \dots, 60\}$ 3. Track estimates of π_m over time
Assumptions	Over $t = 1, \dots, T$: α_t changing slowly β_t fixed	Over $t = 1, \dots, T$: α_t, δ_t changing slowly	Over $t = 1, \dots, T$: α_t, δ_t changing slowly β_t fixed

Notes on data and parameters

Customer loyalty measure

- There is a single stream.
- The customer-specific survey response for customer j in week t is y_{jt} where
    - $y_{jt} = 1$ if the customer response is $\in \{0,1,2,3,4,5,6\}$,
    - $y_{jt} = 2$ if the customer response is $\in \{7,8\}$,
    - $y_{jt} = 3$ if the customer response is $\in \{9,10\}$.
- The customer-specific covariate values for customer j in week t is $x_{jt} = \left(x_{1,jt}, \dots, x_{4,jt}\right)^T$ where
    - $x_{1,jt} = 1$ if product variant is 2, $x_{1,jt} = 0$ otherwise,
    - $x_{2,jt} = 1$ if product variant is 3, $x_{2,jt} = 0$ otherwise,
    - $x_{3,jt} = 1$ if product variant is 4, $x_{3,jt} = 0$ otherwise,
    - $x_{4,jt} \in \{0,2,6,12,18,24\}$, an interval variable designating the lower bound of tenure in months.
- The standard population specifies the levels of the x_{j^*} for each customer $j^* = 1, \dots, J^*$.
- The baseline level of the covariates ($x = (0,0,0,0)^T$) is $\text{product variant} = 1$ and $\text{tenure} = 0$.
- The elements of parameter $\alpha_t = \left(\alpha_{1,t}, \alpha_{2,t}\right)^T$ relate to the probabilities that a customer with baseline level of the covariates is a detractor or a promoter, respectively.
- The elements of parameter $\beta_t = \left(\beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \beta_{4,t}\right)^T$ relate to the effects of product variants 2, 3, and 4, and tenure, respectively, relative to the two baseline probabilities.

Lab positive abnormal rate

- The test result for subject j at lab m in month t is y_{jmt} where
    - $y_{jmt} = 1$ if the test result is 'positive abnormal' and $y_{jmt} = 0$ otherwise.
- There are no subject-level covariates.
- There are multiple streams relating to the seven labs where the FOBT test is conducted in Ontario.

- The parameter α_t relates to the positive abnormal rate ("positive rate") at the baseline lab 1.
- The elements of parameter $\delta_t = \left(\delta_{1,t}, \delta_{2,t}, \delta_{3,t}, \delta_{4,t}, \delta_{5,t}, \delta_{6,t}\right)^T$ relate to positive rates of labs 2 through 7, respectively, relative to the positive rate at the baseline lab.

Hospital performance measure

- The result for patient j at hospital m in year t is y_{jmt} where
    - $y_{jmt} = 1$ if the patient is deceased during the same hospital stay in which he/she underwent PCI or after hospital discharge but within 30 days of surgery and $y_{jmt} = 0$ otherwise.
- There are multiple streams relating to the 60 hospitals performing PCI in New York State.
- The patient-level covariate values for patient j in hospital m in year t are $x_{jmt} = \left(x_{1,jmt}, \ldots, x_{15,jmt}\right)^T$ where
    - $x_{1,jmt}$ = patient age; an integer value in years greater than 55
    - $x_{2,jmt} = 1$ if hemodynamic state is 'unstable', $x_{2,jmt} = 0$ otherwise
    - $x_{3,jmt} = 1$ if ventricular ejection fraction is $< 20\%$, $x_{3,jmt} = 0$ otherwise
    - $x_{4,jmt} = 1$ if ventricular ejection fraction is $20 - 29\%$, $x_{4,jmt} = 0$ otherwise
    - $x_{5,jmt} = 1$ if ventricular ejection fraction is $30 - 39\%$, $x_{5,jmt} = 0$ otherwise
    - $x_{6,jmt} = 1$ if pre-procedural myocardial infarction is < 6 hrs, $x_{6,jmt} = 0$ otherwise
    - $x_{7,jmt} = 1$ if pre-procedural myocardial infarction is $6 - 11$ hrs, $x_{7,jmt} = 0$ otherwise
    - $x_{8,jmt} = 1$ if pre-procedural myocardial infarction is $12 - 23$ hrs, $x_{8,jmt} = 0$ otherwise
    - $x_{9,jmt} = 1$ if pre-procedural myocardial infarction is $1 - 14$ days, $x_{9,jmt} = 0$ otherwise
    - $x_{10,jmt} = 1$ if congestive heart failure is 'current within 2 wks', $x_{10,jmt} = 0$ otherwise
    - $x_{11,jmt} = 1$ if chronic lung disease is 'yes', $x_{11,jmt} = 0$ otherwise
    - $x_{12,jmt} = 1$ if renal failure creatinine level is 1.6 - 2.0 mg/dl, $x_{12,jmt} = 0$ otherwise
    - $x_{13,jmt} = 1$ if renal failure creatinine level is > 2.0 mg/dl, $x_{13,jmt} = 0$ otherwise

  - $x_{14,jmt} = 1$ if renal failure creatinine level is 'requires dialysis', $x_{14,jmt} = 0$ otherwise
  - $x_{15,jmt} = 1$ if malignant ventricular arrhythmia is 'yes', $x_{15,jmt} = 0$ otherwise
- The standard population specifies the levels of the x_{j^*} for each patient $j^* = 1, \dots, J^*$.
- The baseline level of the covariates ($x = (0,0, \dots, 0)^T$) is patient age $= 55$, hemodynamic state is 'stable', ventricular ejection fraction is $\geq 40\%$, pre-procedural myocardial infarction is 'none within 14 days', congestive heart failure is 'no', chronic lung disease is 'no', renal failure creatinine level is ≤ 1.5, malignant ventricular arrhythmia is 'no'. The baseline hospital is 1.
- The parameter α_t relates to the 30-day post-surgery mortality rate ("mortality rate") for a patient with baseline levels of the covariates at the baseline hospital.
- The parameter $\delta_t = \left(\delta_{1,t}, \delta_{2,t}, \dots, \delta_{60,t}\right)^T$ relates to the mortality rates of patients at hospitals 2 through 60, respectively, relative to the rate at the baseline hospital.
- The parameter $\beta_t = \left(\beta_{1,t}, \beta_{2,t}, \dots, \beta_{15,t}\right)^T$ relates to the mortality rates of patients at the various covariate levels relative to the baseline mortality rate.

1.5. Bias/variance trade-off

In the analysis of data collected over time, uncertainty in a parameter estimate based on data from the most recent time period is related to the number of observed responses. In the three motivating applications, the numbers of responses observed at the present time period relies on factors that cannot be controlled, such as response rates to a survey. Small sample sizes occur at some time periods and estimates based on these samples have large uncertainty and may negatively impact management decisions. We want to draw on data from multiple time periods to reduce uncertainty.

One alternative that improves precision in estimates from data collected at regular time intervals is to combine data across time periods. In the common situation where a parameter is drifting over time, a present time estimate that uses present and historical data is biased. Including historical data to reduce uncertainty due to small sample size trades bias for precision. To assess the trade-off, we consider the efficiency measure root mean squared error (MSE) $= \sqrt{\text{bias}^2 + \text{variance}^2}$ and prefer an estimator that has the smallest MSE among alternative estimators. Too much change in the parameter over time results in a large amount of bias and this trade-off is not viable. We

restrict our focus to problems where we expect that the true value of the parameter changes slowly over time.

1.6. Weighted estimating equations

We introduce the concept of weighted estimating equations to regulate the bias/variance trade-off that is the basis for this research.

Setup

- data $d_t, t = 1, \dots, T$ from a sample of $n_t = \sum_{m=1}^{M} n_{mt}$ subjects according to the general problem described in Section 1.2
- unknown parameter θ_t of dimension p at time $t = 1, \dots, T$
- likelihood function $\mathcal{L}_t(d_t; \theta_t)$ describing the probability of d_t given θ_t
- score function $\psi_t(\theta_t; d_t) = \frac{\partial l_t(\theta_t; d_t)}{\partial \theta_t}$ of dimension p where $l_t(\theta_t; d_t) = \log \mathcal{L}_t(\theta_t; d_t)$

The elements of the parameter vector θ_t include a parameter that relates to the mean performance for a subject having baseline levels of the covariates as well as the effects of covariates and/or multiple streams. We assume that the elements of the unknown parameter vector θ_t describe the same attributes of the process across time periods $t = 1, \dots, T$ and the unknown true value of one or more of the p elements may be drifting slowly in an unpredictable way.

Estimating functions

We may estimate θ using only the data d_T observed in the most recent time period.

$$Q_1(\theta; d_T) = \psi_T(\theta; d_T) \tag{1}$$

We may use all of the data $d = \{d_t; t = 1, \dots, T\}$ assuming $\theta = \theta_t$ for all t.

$$Q_2(\theta; d) = \sum_{t=1}^{T} \psi_t(\theta; d_t) \tag{2}$$

Steiner and MacKay (2014) propose weighted estimating functions as a means to use all historical data and down-weight the influence of historical data,

$$Q_3(\theta; d, w) = \sum_{t=1}^{T} w_t \psi_t(\theta; d_t) \tag{3}$$

for a selection of weights $w = \{w_t, t = 1, \dots, T\}$. Steiner and MacKay (2014) suggest selecting weights that decline exponentially from T to $T - 1$ and so on. The related weighted estimating equations (WEE) are $Q_3(\hat{\theta}; d, w) = [0]_{(p \times 1)}$ and solving these equations gives the estimate $\hat{\theta}$. The

motivation for using (3) over (1) or (2) is to regulate the bias/variance trade-off between estimates based on present time data only or based on aggregate of historical data.

1.7. Contribution and outline of this thesis

In this thesis, we extend the weighted estimating equations (WEE) approach originally proposed by Steiner and MacKay (2014) to three new application areas. These applications extend previous applications of this approach to deal with multiple covariates, multinomial outcomes, and tests of hypotheses. We show that this approach can have an important improvement in managing performance in these applications relative to current industry practices and other alternative approaches. We offer theoretical derivations of approximations for the measure of uncertainty of the WEE estimator and the distribution of the WEE likelihood ratio test statistic. We discuss various implementation considerations and improvements that are possible under previous knowledge or assumptions of the parameter of interest.

The thesis is organized as follows. Chapter 2 introduces existing approaches for the general problem of this research and highlights methodologies that are similar in some way to the weighted estimating equations formulation. Chapter 3 outlines the algorithm to estimate a risk-adjusted mean by the WEE approach and derives the approximations for an estimate of variance and the distribution of a test statistic based on WEE estimates. We give an analytic example to observe properties of these approximations in a simple case. Chapter 4 applies the WEE approach to estimate the customer loyalty measure based on a realistic dataset as well as on simulated data. This chapter assesses the approximation for the variance of the WEE estimate and compares the WEE approach to the exponentially weighted moving average approach. Chapter 5 applies the WEE approach to estimate the lab positive abnormal rate based on a real dataset as well as on simulated data. This chapter assesses the approximation for the distribution of the hypothesis test statistic and discusses implementation considerations including the selection of a historical time window and considerations for some large sample sizes. Chapter 6 applies the WEE approach to estimate the hospital performance measure based on realistic data. This chapter discusses implementation considerations including the selection of time subgroups and the weight parameter, alternatives for estimating covariate effects, and missing data and sampling zeros. Chapters 4, 5, and 6 discuss the current industry practice for each application and compare the WEE estimates to estimates through current practice as well as other naïve alternatives. Chapter 7 summarizes the results of this research, discusses limitations, and offers extensions as future work.

The theoretical derivations and analytic example in Chapter 3 and the WEE approach applied to the realistic customer loyalty dataset in Chapter 4 are the basis of a paper entitled "Bias/Variance Trade-off in Estimates of a Process Parameter based on Temporal Data" that has been submitted for publication (Cooper Barfoot, Steiner, and MacKay, 2016). Two rounds of reviewer feedback have suggested useful modifications that have been incorporated into this research. Additionally, we plan to reach the marketing and healthcare communities through applied papers which demonstrate the importance of considering the WEE approach for estimation in these applications.

Chapter 2: Literature Review and Related Methods

Under the framework of the general problem introduced in Section 1.2, we look at existing methods for estimating the risk-adjusted parameter of interest and its uncertainty based on a stream of data. Additionally, we review the similarities of weighted estimating equations to generalized estimating equations and relevance weighted likelihood. We also review methods to select weights pertinent to the problem.

2.1. Risk-adjustment

In Section 1.3, we introduce the first step in the approach to the general problem as the selection of a standard population. The use of a standard population is particularly important so that comparisons of estimates across time and across streams are reliable. The field of epidemiology uses the standard population concept in order to study patterns, causes, and effects of health and disease adjusted for risk factors of the people in the study population. The World Health Organization (WHO) states that most rates, such as incidence, prevalence, and mortality, are strongly age-dependent with risks rising or declining with age (Ahmad et al., 2001). The WHO publishes a current international population distribution by age group for practitioners to use as their standard population by age. They recommend direct standardization which is a weighted average of the age-specific rates for each of the populations to be compared. As in the motivating applications of this research, the choice of a standard population in epidemiology studies can affect the results and conclusions decisions based on the data and must be pertinent to the application.

Steiner (2014) gives a comprehensive discussion of the need for risk-adjusted monitoring of health care outcomes. The author highlights various risk-adjusted methods for monitoring and issues that need to be explored, one being the effect of estimation error and model specification error on the performance of a risk adjustment model used in conjunction with a monitoring chart. As in Steiner and MacKay (2014), the weighted estimating approach holds promise as an alternative to specifying and estimating parameters in a risk-adjustment model.

2.2. Non-parametric vs. parametric estimates

In Section 1.2, we introduce the problem to estimate π_m which is the mean of the random variable Y_{mt} at current time $t = T$ for a standard population of subjects in stream m. The estimate of π_m relies on an estimate of $\{\pi_{j^*m}, j^* = 1, \dots, J^*\}$ which is the present expected value of the

proportion or rate across subjects in the standard population. The standard population $\left\{x_{j^*} = \left(x_{1,j^*}, \ldots, x_{s,j^*}\right)^T \text{ for } j^* = 1, \ldots, J^*\right\}$ assigns values to the covariates as discussed in Section 1.3. The approach to estimate each π_{j^*m} introduced in Section 1.2 uses observed outcomes, associated values of the covariates, and parameter $\theta = \theta_T$ in a generalized linear model (GLM). Non-parametric estimates of the parameter of interest π_{j^*m} are also possible through the use of appropriate sample averages. We discuss advantages and limitations of non-parametric and parametric estimates.

Non-parametric estimates

A non-parametric estimate of π_{j^*m} can be made through sample averages of observations among subjects in stream m having the same covariate vector as subject j^* in the standard population. For a continuous covariate, a subject's observation is included in the average if its value of the covariate is within a specified close proximity to the covariate value for subject j^*. These estimates are generally simple to implement, well-understood, and estimates of their standard errors are straightforward to compute. One significant drawback to this non-parametric estimate is that the sample average is undefined when there are no observations among subjects having the same covariate vector (or a vector in close proximity) as subject j^*. Since the general problem under study exhibits small sample sizes at some time periods, then this is an important limitation to the non-parametric approach. Further, if there is more than one covariate, responses from subjects that have some, but not all, values of the covariates which are the same as j^* are not used to estimate π_{j^*m}. Notwithstanding these limitations, it is the author's experience that the simplicity of the non-parametric estimate and related standard error estimates make this a commonly used approach in practice.

Parametric estimates

Parametric methods of estimation require more assumptions than non-parametric methods. If those extra assumptions are valid, then the estimate is generally more accurate and precise. A parametric method requires a model to describe the mean as a function of a parameter $\theta = \theta_T$ and covariate values. A selection from the class of classical linear models or the wider class of GLMs is common and requires assumptions on the error structure of the fitted model. Agresti (2007) gives a thorough review of models for categorical data. McCullagh and Nelder (1989) is an important resource for GLMs.

Estimation of the parameter vector θ follows through maximum likelihood estimation (MLE) and the estimate $\hat{\theta}$ possesses desirable large sample properties:

- $\hat{\theta}$ is an unbiased minimum variance estimator as sample size increases
- $\hat{\theta}$ has an approximate normal distribution and its variance can be estimated
- likelihood ratios can be used to test hypotheses about models and parameters

With a MLE estimate of θ, an estimate of π_{j^*m} follows from the specified model and a specified standard population of covariate values. By the invariance property of MLEs (Casella and Berger 2002, p. 350), the estimate $\hat{\pi}_{j^*m}$ is itself a maximum likelihood estimate. By extension, $\hat{\pi}_{j^*m}$ possesses the same desirable large sample properties as $\hat{\theta}$.

There are additional advantages over estimation through a non-parametric approach. One advantage is that the estimate $\hat{\theta}$ contains important information about the covariate effects. The parameter θ is estimated with all of the data and so, unlike an estimate of π_{j^*m} by sample averages, an estimate by the parametric model uses more information than just observations from subjects with the same covariate vector (or a vector in close proximity) and the same stream. The parametric model estimates parameters to describe the covariate effects and through model restrictions, covariate effects can be allowed to vary or be held fixed over time. Another important advantage of the parametric model is that the estimate $\hat{\pi}_{j^*m}$ may be defined even if subjects having the same covariate vector as j^* in stream m are not present in the sample. Estimates of all elements of θ are possible as long as every level of each categorical covariate and each stream are present in the sample. For continuous covariates, then a minimum of two levels of each covariate must be present to estimate a linear effect and the covariates must not be collinear.

In general, the limitations of a parametric model include the assumptions on the model and the required solution of non-linear equations to estimate the parameter. With the vast availability of classical linear models, GLMs, and software for estimating MLEs, there is much flexibility in model selection and estimation.

2.3. Use of historical data for a present time estimate

Section 2.2 outlines non-parametric and GLM-based estimates as well as alternatives for estimation with small samples without considering any time ordering of the data. To address the general problem of Section 1.2, we can estimate π_{j^*m}, the mean response for subject j^* in stream m at current time $t = T$, using present time data or involving the historical data collected over time. Table 2-1 gives the mathematical formulations of the alternatives discussed in this section.

Use present time data only

We can estimate the present value of the parameter $\theta = \theta_T$ using data d_T observed from the present time only, either through non-parametric estimates or parametric estimates as in Section 2.2. The non-parametric and parametric estimates for π_{j^*m} are given in the first row of Table 2-1. In the general problem where the present sample size may be small and the parameter is changing slowly over time, the estimates have no bias and high variance. The limitation that the non-parametric estimate is undefined when there are no observations in the present time sample having the same covariate vector as subject j^* (or a vector in close proximity in the continuous covariate case) has an important detrimental effect when the present sample size is small.

A further limitation specific to the general problem outlined in Section 1.2 is the properties of an MLE estimate under small samples. The MLE approach relies on the assumption that the sample data are representative of the population and the relationship between the inputs and outputs is adequately represented. The amount of information in the sample data directly impacts the parameter estimates. It is well known that the MLE may be biased when the sample size or total Fisher information is small (Shenton and Bowman, 1977). Hence, an MLE estimate of parameter θ using data from the present time only may be biased when the sample size is small and increasing the sample size through the inclusion of historical data is desirable.

Use historical data weighted equally

Another option to estimate $\theta = \theta_T$ is to use data $d = \{d_t\}$ across all time periods $t = 1, \dots, T$ without regard for the time period of the data. We assume that the associated random variables $\mathcal{D} = \{\mathcal{D}_t\}$ over $t = 1, \dots, T$ are independent, conditional on the values of the covariates. We assume that the models of the various $\mathcal{D}_t, t = 1, \dots, T$ are described by the common parameter θ. Then, non-parametric and parametric estimates for π_{j^*m} based on all historical data are given in the second row of Table 2-1. All historical data are given equal weight in the estimate of the parameter and the time period of the data does not impact the estimate. Since more data are used for estimation, then the variance of the estimate is lower than when only present data are used. Further, since sample sizes are larger, then there are fewer cases where the non-parametric estimate is undefined. However, in our problem where the model parameter θ_t is changing slowly over time, the assumption of a single θ to describe all $\mathcal{D}_t$ over $t = 1, \dots, T$ is erroneous. The impact is a biased estimate of the parameter.

Exponentially weighted moving average

A compromise between using present time data only and using all historical data weighted equally is to use a weighted average of the estimates across time. We extend the concept of the exponentially weighted moving average (EWMA) from statistical process control literature since the EWMA chart is very effective in detecting small sustained shifts in a process (Montgomery, 2013). Weights are chosen to regulate the relative influence of present and historical data on the present time parameter estimate. We select weights $\{w_t, t = 1, \dots, T\}$ to have the largest value at the current time period T and decline exponentially across time periods in the further past. Selecting weights is discussed further in Section 2.8 and Section 3.1.

We can calculate an EWMA estimate of π_{j^*m} based on either non-parametric or GLM-based estimates of π_{j^*mt} at each time period $t = 1, \dots, T$. Table 2-1 shows the estimate of π_{j^*m} as weighted combinations of $\hat{\pi}_{j^*mt}, t = 1, \dots, T$, with weights $\{w_t\}$. The standard EWMA estimate for π_{j^*m} requires that estimates $\hat{\pi}_{j^*mt}$ be defined for all $t = 1, \dots, T$. As discussed in Section 2.2, the non-parametric estimate is undefined when there are no observations among subjects in d_t having the same covariate vector as subject j^* but the parametric estimate may still be possible depending on the observations in the entire dataset $d = \{d_1, d_2, \dots, d_T\}$. When one of the $\hat{\pi}_{j^*mt}$ over $t = 1, \dots, T$ is undefined, we have to make an adjustment to the standard EWMA approach such as adjusting the weight w_t to have value 0 and rescaling the remaining weights.

The weighted estimating equations approach introduced in Section 1.6 is also included in Table 2-1. Here, $I\left[x_{jmt} = x_{j^*}\right]$ is an indicator variable that takes the value of 1 when $x_{jmt} = x_{j^*}$ (or the two are in close proximity in the case of a continuous covariate) and 0 otherwise.

*Table 2-1. Approaches to estimate present time mean π_{j^*m} for subject j**

<table>
<tr><th colspan="2"></th><th>Non-parametric approaches</th><th>GLM-based approaches</th></tr>
<tr><td>Use present time data only</td><td>Naïve</td><td>$\hat{\pi}_{j^*m} = \frac{\sum_{j=1}^{n_{mT}} y_{jmT}\, I\left[x_{jmT} = x_{j^*}\right]}{\sum_{j=1}^{n_{mT}} I\left[x_{jmT} = x_{j^*}\right]}$</td><td>▪ Solve $Q_1(\hat{\theta}; d_T) = 0$ from (1) for $\hat{\theta}$
▪ $\hat{\pi}_{j^*m} = g^{-1}\left(h(x_{j^*}, m, \hat{\theta})\right)$</td></tr>
<tr><td rowspan="4">Use all historical data</td><td>Naïve</td><td>$\hat{\pi}_{j^*m} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{n_{mt}} y_{jmt}\, I\left[x_{jmt} = x_{j^*}\right]}{\sum_{t=1}^{T}\sum_{j=1}^{n_{mt}} I\left[x_{jmt} = x_{j^*}\right]}$</td><td>▪ Solve $Q_2(\hat{\theta}; d) = 0$ from (2) for $\hat{\theta}$
▪ $\hat{\pi}_{j^*m} = g^{-1}\left(h(x_{j^*}, m, \hat{\theta})\right)$</td></tr>
<tr><td>EWMA</td><td colspan="2">Select $w_t, t = 1, \ldots, T$ to be exponentially declining for $T, T-1, \ldots, 1$</td></tr>
<tr><td>EWMA</td><td>For each $t = 1, \ldots, T$,
▪ $\hat{\pi}_{j^*mt} = \frac{\sum_{j=1}^{n_{mt}} y_{jmt}\, I\left[x_{jmt}=x_{j^*}\right]}{\sum_{j=1}^{n_{mt}} I\left[x_{jmt}=x_{j^*}\right]}$
▪ $\hat{\pi}_{j^*m} = \frac{1}{\sum_{t=1}^{T} w_t}\sum_{t=1}^{T} w_t\, \hat{\pi}_{j^*mt}$</td><td>For each $t = 1, \ldots, T$,
▪ solve $Q_1(\hat{\theta}_t; d_t) = 0$ for $\hat{\theta}_t$
▪ $\hat{\pi}_{j^*mt} = g^{-1}\left(h(x_{j^*}, m, \hat{\theta}_t)\right)$
▪ $\hat{\pi}_{j^*m} = \frac{1}{\sum_{t=1}^{T} w_t}\sum_{t=1}^{T} w_t\, \hat{\pi}_{j^*mt}$</td></tr>
<tr><td>WEE</td><td>Not applicable</td><td>▪ Solve $Q_3(\hat{\theta}; d, w) = 0$ from (3) for $\hat{\theta}$
▪ $\hat{\pi}_{j^*m} = g^{-1}\left(h(x_{j^*}, m, \hat{\theta})\right)$</td></tr>
<tr><td colspan="4">▪ $\left\{x_{j^*} = \left(x_{1,j^*}, \ldots, x_{s,j^*}\right)^T, j^* = 1, \ldots, J^*\right\}$: covariate vectors for the standard population
▪ $\hat{\pi}_m = \frac{1}{J^*}\sum_{j^*=1}^{J^*} \hat{\pi}_{j^*m}$: estimate of mean for the standard population</td></tr>
</table>

Among the approaches discussed in Section 2.3, the EWMA GLM-based approach is the only one that describes the covariate effects using parameter estimates and controls the relative influence of past and present data for estimating the present value of a parameter. One important limitation of this approach is that the covariate effects are re-estimated at each time period using data observed at that time period only, even though we assume that covariate effects do not change over time. Uncertainties in the estimates at each time period add to the uncertainty of the present time estimate and so a small sample size at any time period has a detrimental effect on the precision of the present time estimate.

Like the EWMA approach, the weighted estimating equations approach regulates the trade-off between bias and variance with exponentially declining weights. However, the WEE approach addresses the EWMA shortcomings discussed. In Chapter 4, we pursue the favourability of the two approaches relative to the size of the change in the parameter over time, observed sample sizes, and the choice of weights.

2.4. Kalman Filter

The exponentially weighted moving average is a simplified state space approach since it combines an estimate from present data d_T with another estimate based on previous data $\{d_1, \dots, d_{T-1}\}$. State space models are based on the Markov property, which implies the independence of the present state of a process from its past, given the previous state. In such a system, the previous state of the process summarizes all the information from the past. A flexible state space model is the Kalman Filter (KF) which estimates the present state of a dynamic system (Grewal and Andrews, 2008). The KF comprises a system dynamic model which describes the evolution of the state vector and a measurement model which describes the generation of the observations from a given state vector. State estimate and covariance extrapolation and updating equations are solved recursively based on assumed initial conditions. Clearly, the KF is a flexible approach and it has been widely applied for the control of complex dynamic systems such as continuous manufacturing processes and spacecraft. In these applications, the system dynamic model and measurement model are built using subject matter expertise and observation of the process over long periods of time. At the other extreme where the system dynamic model involves no serial correlation, as in the random walk plus noise model, then Muth (1960) first pointed out that the steady-state solution of the Kalman Filter equations reduces to the EWMA estimator discussed in Section 2.3. In Section 7.1, we provide a qualitative comparison of the Kalman Filter and WEE approaches for the general problem of this work.

2.5. Estimates of uncertainty

In addition to the risk-adjusted estimate of the mean in stream m, an estimate of its uncertainty is important for statistical inference. We estimate the variance of the mean for the standard population, $\widehat{var}(\hat{\pi}_m)$, through

$$\widehat{var}(\hat{\pi}_m) = \frac{1}{J^{*2}} \sum_{j^*=1}^{J^*} \widehat{var}\left(\hat{\pi}_{j^*m}\right) \tag{4}$$

under the assumption that the random variables Y_{j^*mt} across $j^* = 1, \dots, J^*$ are independent, conditional on the value of the covariates. We require estimates of $var\left(\hat{\pi}_{j^*m}\right)$, the variance for the estimate of the mean for subject j^* in the standard population in stream m at the current time T. The estimates of uncertainty for the non-parametric estimates of a continuous response variable and a binary response variable for a single stream problem (π_{j^*m} becomes π_{j^*}) are given in Table 2-2. Extension to the multiple stream problem is straightforward. The extension to

categorical/ordinal response variables is also straightforward and is demonstrated through the customer loyalty measure in Chapter 4.

Table 2-2. Estimates of variance for non-parametric estimates of π_{j^*} (single stream problem)

		Continuous response	Binary response
Use present time data only	**Naïve**	$\widehat{var}(\hat{\pi}_{j^*}) = \frac{\underset{\{j\|x_{jT}=x_{j^*}\}}{\widehat{var}}(y_{jT})}{\sum_{j=1}^{n_T} I[x_{jT}=x_{j^*}]}$ [1]	$\widehat{var}(\hat{\pi}_{j^*}) = \frac{\hat{\pi}_{j^*}(1-\hat{\pi}_{j^*})}{\sum_{j=1}^{n_T} I[x_{jT}=x_{j^*}]}$
Use all historical data	**Naïve**	$\widehat{var}(\hat{\pi}_{j^*}) = \frac{\sum_{t=1}^{T} \underset{\{j\|x_{jt}=x_{j^*}\}}{\widehat{var}}(y_{jt})}{\sum_{t=1}^{T}\sum_{j=1}^{n_t} I[x_{jt}=x_{j^*}]}$ [1]	$\widehat{var}(\hat{\pi}_{j^*}) = \frac{\hat{\pi}_{j^*}(1-\hat{\pi}_{j^*})}{\sum_{t=1}^{T}\sum_{j=1}^{n_t} I[x_{jt}=x_{j^*}]}$
	EWMA	$\widehat{var}(\hat{\pi}_{j^*}) = \frac{1}{(\sum_{t=1}^{T} w_t)^2} \times \sum_{t=1}^{T} w_t^2 \frac{\underset{\{j\|x_{jt}=x_{j^*}\}}{\widehat{var}}(y_{jt})}{\sum_{j=1}^{n_t} I[x_{jt}=x_{j^*}]}$ [1]	$\widehat{var}(\hat{\pi}_{j^*}) = \frac{1}{(\sum_{t=1}^{T} w_t)^2} \times \sum_{t=1}^{T} w_t^2 \frac{\hat{\pi}_{j^*t}(1-\hat{\pi}_{j^*t})}{\sum_{j=1}^{n_t} I[x_{jt}=x_{j^*}]}$

[1]: $\widehat{var}(\cdot)$ refers to the sample variance of responses from subjects having specified covariate vector

When estimating the mean π_m with a GLM-based approach, the estimate $\widehat{var}(\hat{\pi}_m)$ combines the estimates $\widehat{var}(\hat{\pi}_{j^*m})$ across $j^* = 1, \dots, J^*$. The estimate $\widehat{var}(\hat{\pi}_{j^*m})$ follows from the multivariate delta method (Casella and Berger, 2002) as

$$\widehat{var}(\hat{\pi}_{j^*m}) = \sum_{p_1=1}^{p}\sum_{p_2=1}^{p} \widehat{var}(\hat{\theta})_{(p_1,p_2)} \left[\frac{\partial g^{-1}(h(x_{j^*},m,\theta_t))}{\partial \theta_{t,p_1}} \frac{\partial g^{-1}(h(x_{j^*},m,\theta_t))}{\partial \theta_{t,p_2}}\right]_{\theta_t=\hat{\theta}} \quad (5)$$

where $\hat{\theta}$ and $\widehat{var}(\hat{\theta})$ are the GLM estimates of the model parameter $\theta = \theta_T$ and its uncertainty and the functions $h(x_{jmt}, m, \theta_t)$ and $g(\pi_{jmt})$ and are the GLM linear predictor and link functions, respectively. Where parameter θ_t has dimension p, θ_{t,p_1} refers to the p_1 entry of θ_t, and $\widehat{var}(\hat{\theta})_{(p_1,p_2)}$ refers to the (p_1, p_2) entry of $\widehat{var}(\hat{\theta})$ for $p_1, p_2 \in \{1, \dots, p\}$. The matrix $var(\hat{\theta})$ is estimated by usual MLE methods for the GLM.

The calculation in (5) applies to the GLM-based estimates for π_{j^*m} based on either present time data only or the aggregate of all historical data. For the EWMA GLM-based estimate, the calculation in (5) must be made for each estimate by time period, $\widehat{var}(\hat{\pi}_{j^*mt})$, and combined through (6) for the estimate of $var(\hat{\pi}_{j^*m})$,

$$\widehat{var}(\hat{\pi}_{j^*m}) = \frac{1}{\left(\sum_{t=1}^{T} w_t\right)^2} \sum_{t=1}^{T} w_t^2 \widehat{var}\left(\hat{\pi}_{j^*mt}\right) \quad (6)$$

which is valid under the assumption that estimates by time period are independent, conditional on the values of the covariates.

Sandwich estimate of variance

When estimating parameter $\theta = \theta_T$ through the weighted estimating equations approach, Steiner and MacKay (2014) use the estimate of $var(\hat{\theta})$ for the asymptotic variance of the MLE for a misspecified model introduced by White (1982). The so-called sandwich estimate of variance for $\hat{\theta}$ is

$$\widehat{var}_S(\hat{\theta}) = \hat{V}^{-1}\left(\hat{\theta}\right)\hat{B}(\hat{\theta})\hat{V}^{-1}(\hat{\theta}) \quad (7)$$

where $\hat{\theta}$ is the solution of a p-dimensional estimating function vector $\psi(\theta; d)$ based on data d. The derivation based on a misspecified model is found in Geyer (2013). Here $\hat{V}(\hat{\theta}) = -E\left[\frac{\partial \psi(\theta;d)}{\partial \theta}\right]_{\theta=\hat{\theta}}$ is the expected Hessian matrix and $\hat{B}(\hat{\theta}) = var[\psi(\theta; d)]_{\theta=\hat{\theta}}$ is the variance of the estimating function vector. Both are evaluated at the estimate of the parameter. Note that when we estimate the parameter based on either present time data only or all historical data weighted equally, then $-E\left[\frac{\partial \psi(\theta;d)}{\partial \theta}\right] = var[\psi(\theta; d)]$ and (8) simplifies to the usual estimate of variance. When we estimate the parameter by the WEE approach with non-trivial weights, then $\hat{V}(\hat{\theta}) \neq \hat{B}(\hat{\theta})$. Since the sandwich estimate of variance combines the variance estimate for the specified data distribution with a variance matrix constructed from the data, then the variance estimate is sometimes called the empirical variance estimate. With estimate $\widehat{var}_S(\hat{\theta})$ for WEE estimate $\hat{\theta}$, then estimates of uncertainty for $\hat{\pi}_{j^*m}$ and $\hat{\pi}_m$ follow as in (5) and (4).

The sandwich estimate of variance in (7) applies generally to a vector of unbiased estimation equations (Hardin and Hilbe, 2013). The estimating equations may be generalized estimating equations (GEE) which extend a GLM for longitudinal or batch correlated data. These estimating equations are derived without specifying the joint distribution of a subject's observations but with model components for the mean and covariance of the marginal distributions of the subject's observations. The parameter estimates and sandwich estimates of parameter variances are consistent and robust to misspecification of the covariance structure under regularity conditions (Liang and Zeger, 1986). Since the general problem outlined in Section 1.2 is to estimate the present value of a parameter that may change over time, the sandwich estimate of variance may

be a biased estimate of its variance. We study this estimate in relation to the WEE approach in Section 3.3.

2.6. Generalized estimating equations

We look to theory on weighted generalized estimating equations (GEE) pertinent to the weighted estimating equation approach. We begin with the general GEE formulation (Godambe and Kale, 1991).

Setup

- Y_{it}: a random variable describing the response for subject i at time t; unlike the setup in Section 1.2, reference i refers to the same subject with repeated measures over time periods $t = 1, \dots, T$
- $y_i = (y_{i1}, \dots, y_{it}, \dots, y_{iT})^T$: a $(T \times 1)$ column vector of data from random variables Y_{it}
- $x_{it} = \left(x_{1,it}, \dots, x_{s,it}\right)^T$: a $(s \times 1)$ column vector of the s covariates associated with y_{it} and $x_i = (x_{i1}\ x_{i2} \dots x_{iT})_{(s \times T)}$
- $\mu_{it} = E[Y_{it}|x_i]$, $\mu_i = (\mu_{i1}\ \mu_{i2} \dots\ \mu_{iT})^T$: a $(T \times 1)$ column vector of mean values for subject i by time period t
- $v_{it} = var(Y_{it}|x_i)$ and $V_i = cov(Y_i|x_i)$: a $(T \times T)$ matrix having v_{it} as diagonal elements, known as the working covariance matrix for the response Y_i
- user-specified regression model of μ and v based on x and s-dimensional parameter θ: $\mu_{it} = h^{-1}(x_{it}^T\theta)$ and $v_{it} = k^{-1}(\mu_{it}) = k^{-1}\left(h^{-1}(x_{it}^T\theta)\right)$

Formulation

Subject to regularity conditions and the definition of optimality (Godambe and Kale, 1991, p. 12), then an optimal estimating function for θ is $U(\theta) = \sum_{i=1}^{n}\left(\frac{\partial \mu_i^T}{\partial \theta}\right)V_i^{-1}(Y_i - \mu_i) = 0$. The equation $U(\theta) = 0$ is called the generalized estimating equation (GEE). Solve $U(\hat{\theta}) = 0$ for $\hat{\theta}$ which is called the GEE estimate of θ. Note that $E[U(\theta)] = 0$ when $E(Y_i|x_i) = \mu_i$ under the implicit assumptions that $E(Y_{it}|x_i) = E(Y_{it}|x_{it})$ and $E(Y_{it}|x_i)$ does not depend on V_i. In practice, V_i is often replaced by a working covariance matrix. Lipsitz, Kim, and Zhao (1994) extend the GEE formulation for multinomial data.

Weighted generalized estimating equations

Chen, Yi, and Cook (2010) and Robins, Rotnitzky, and Zhao (1995) propose a weighted GEE approach for handling incomplete response and covariate data. For this problem, the authors recommend inverse probability weighted generalized estimating equations (IPWGEE) that adjust the usual GEE formulation with weights inversely related to the conditional probability of complete data given the response vector and covariates. They refer to the matrix containing subject-specific weights by time period as $\Delta_i(\alpha)$ where α is a set of regression parameters modeling the missing-data process. In practice, the parameters α of the missing-data model are unknown and must be estimated. Then, the IPWGEE formulation is $U^*(\theta, \alpha) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i^T}{\partial \theta}\right) V_i^{-1} \Delta_i(\alpha)(Y_i - \mu_i) = 0$. The authors show that IPWGEE estimators are consistent subject to correct specification of the missing-data process and simulations show negligible bias in finite samples. Further, they give asymptotic calculations for expected bias with different types of missing data model misspecification.

The estimating equations and weighted estimation equations given by (1), (2), and (3) in Section 1.6 to address the general problem of this research can be derived from the formulations of GEE and IPWGEE stated here. The data of the general problem is not longitudinal or batch correlated data, so the V_i matrix is a diagonal matrix. The general problem assumes a distribution for the response which then defines the diagonal entries of V_i by time period. The marginal mean of the subject-specific response is defined by the GLM in terms of the parameters of the model. We replace subject subscript i by subscript j to remind the reader of the difference that subject i has repeated measures over time but subject j measures across time are independent, conditional on the values of the covariates. In (1), the vectors and matrices defined by the GEE formulation have single entries since there are data from one time period only. Formulation (2) follows from the GEE formulation with sizes of vectors and matrices related to the number of time periods in the data. (Let n in the GEE formulation be $\max_{t=1,\ldots,T} n_t$ from Section 1.2 and set GEE components to 0 whenever $j > n_t$). The WEE formulation (3) follows from the IPWGEE formulation with the weight matrix defined as $\Delta_j(\alpha) = diag(w_t, 1 \leq t \leq T)$ for all $j = 1, \ldots, n$. Since some parameters may be changing slowly over time in the general problem, then the assumption $E(Y_{it}|x_i) = E(Y_{it}|x_{it})$ may not hold. As expected, we may have a biased estimating equation in this case.

Properties of generalized estimating equations

Liang and Zeger (1986) look at asymptotic properties of the GEE estimate under the assumption that the number of independent subjects goes to infinity and the cluster sizes are finite with an upper bound. Xie and Yang (2003) present asymptotic results when either the number of independent subjects or the cluster sizes or both go to infinity. Qu, Yi, Song, and Wang (2011) propose a test to examine the unbiasedness of the weighted estimating functions assuming that the mean structure is correctly specified. Unbiasedness of the weighted estimating functions indicates that the conditional probabilities of complete data are consistently estimated. These properties of GEEs may become useful as we study the properties of the WEE formulation in further research.

2.7. Relevance weighted likelihood

We consider the similarity of the weighted estimating equations approach to relevance weighted likelihood methods (Hu and Zidek 2002, Hu and Rosenberger 2000) where contributions to the likelihood from similar populations are weighted by a relevance measure. Consider T independent populations labelled $t = 1, \dots, T$. Suppose that for each t, Y_t represents a measurable attribute or vector of attributes and Y_t are assumed to be independently distributed. The unknown population distribution of Y_t has probability density function f_t. The probability distributions for Y_t are not necessarily identical across $t = 1, \dots, T$, but we assume that they each resemble the others to some extent. Let $Y = (Y_1, \dots, Y_T)$ be the vector or matrix of measurable attributes across T populations. From each population t, $n_t \geq 0$ items are randomly and independently sampled so we have $Y_t = \{Y_{jt}, j = 1, \dots, n_t\}$ and each of the Y_{jt} are independent and identically distributed with f_t. The development of the relevance weighted likelihood formulation follows.

Akaike (1978) formulates statistical inference as the problem of determining the probability distribution $f(y)$ of an observation y. He suggests the entropy measure $B(f, g) = -\int \frac{f(y)}{g(y)} \log\left(\frac{f(y)}{g(y)}\right) g(y) dy$ for density function $g(y)$ as an estimate of the true distribution $f(y)$. The inference problem is finding $g(y)$ to maximize entropy. Equivalently, the entropy measure can be written as $B(f, g) = \int f(y) \log\big(g(y)\big)\, dy - \int f(y) \log\big(f(y)\big)\, dy$. Since the second term is a constant, then maximizing the first term maximizes $B(f, g)$. That is, we want to maximize $E_f[\log g(y)] = \int f(y) \log\big(g(y)\big)\, dy$.

Hu and Zidek (2002) specify $g(y)$ as $\hat{f}_Y(y)$, the predictive density for Y based on observations $y_1, \dots, y_T$. Since the T populations are sampled independently, then we require $\hat{f}_Y$ to be a product

of predictive densities for the individual populations, $\hat{f}_Y = \hat{f}_1 \times \ldots \times \hat{f}_T$. It then follows that we may find the optimum $\hat{f}_Y$ by finding the optimum $\hat{f}_t$ for each $t = 1, \ldots, T$. We restrict possibilities for $\hat{f}_t$ to the class of densities $f_t(y|\theta_t)$ where f_t are specified functions so that only θ_t need to be estimated. Under this formulation, the degree to which the distribution of a population t resembles that of t' can be represented by

$$\int f_{t'}(y) \log\big(f_t(y|\theta_t)\big)\, dy \geq c_{tt'},\ t \neq t' \tag{8}$$

where $t' = \{1, \ldots, t-1, t+1, \ldots, T\}$ for some constants $c_{tt'}$ representing similarity. Considering that each θ_t must be estimated to maximize the similarity measure to all other populations, the problem becomes finding $\theta_t = \hat{\theta}_t$ that maximizes $\sum_{t'=1}^{T} w_{tt'} \int f_{t'}(y) \log\big(f_t(y|\theta_t)\big)\, dy$ among all possible θ_t where $\{w_{tt'}, t' = 1, \ldots, T\}$ are constants to ensure that (8) holds. Since $f_{t'}(y)$ are unknown, then we must estimate them. Under the conditions that Y_t are discrete and $\int f_{t'}(y) \log\big(f_t(y|\theta_t)\big)\, dy, t \neq t'$, is continuously differentiable on θ_t and using Lagrange's method, the objective function to maximize is

$$\begin{aligned}
\sum_{t'=1}^{T} w_{tt'} \int f_{t'}(y) \log\big(f_t(y|\theta_t)\big)\, dy &= \sum_{t'=1}^{T} w_{tt'} \int \log\big(f_t(y|\theta_t)\big)\, dF_{t'}^{emp}(y) \\
&= \sum_{t'=1}^{T} w_{tt'} \frac{\sum_{j=1}^{n_{t'}} \log f_t\big(y_{jt'}|\theta_t\big)}{n_{t'}} \qquad \text{for} \quad t = 1, \ldots, T \\
&= \sum_{t'=1}^{T} \sum_{j=1}^{n_{t'}} \log f_t^{\,w_{tt'}/n_{t'}}\big(y_{jt'}|\theta_t\big)
\end{aligned}$$

subject to $\int f_t\big(y|\hat{\theta}_t\big)dy = 1$. Applying the exponential function and combining over $t = 1, \ldots, T$ gives the joint all-population relevance weighted likelihood objective function to maximize as $\prod_{t=1}^{T} \prod_{t'=1}^{T} \prod_{j=1}^{n_{t'}} f_t^{\,w_{tt'}/n_{t'}}\big(y_{jt'}|\theta_t\big)$. When $n_{t'} = 0$, let $\frac{w_{tt'}}{n_{t'}} = 0$. The relevance weighted likelihood function relative to one population only is $\prod_{t=1}^{T} \prod_{j=1}^{n_t} f_T^{w_t}\big(y_{jt}|\theta_T\big)$. Note t' is replaced by t and w_{Tt}/n_t is replaced by w_t. The maximum weighted likelihood estimate (MWLE) is the vector of parameters $\big(\hat{\theta}_1, \ldots, \hat{\theta}_T\big)^T$ or $\hat{\theta}_T$ that maximizes the appropriate objective function.

The MWLE $\hat{\theta}_T$ that maximizes the relevance weighted likelihood function relative to a single population only is equivalent to the solution of $Q_3(\hat{\theta}; d, w) = 0$ in (3) for $\hat{\theta}$ as an estimate for $= \theta_T$. The WEE formulation is equivalent to relevance weighted likelihood where the single population of interest is the current time period and the related populations are previous time periods. Samples within and across time periods are assumed to be drawn independently from a distribution f_t by time period. As in relevance weighted likelihood, $Q_3(\theta; d, w)$ requires constants $\{w_t\}$ that represent the relevance of distribution f_t to f_T which are discussed in Section 2.8.

Properties of maximum weighted likelihood estimators

The works Hu (1997) and Wang, van Eeden, and Zidek (2002) extend the classical large sample theory for the MLE to the MWLE under two asymptotic paradigms. Hu (1997) increases the number of populations in close proximity to θ and Wang et al. (2002) increases the number of observations from each population with the number of populations remaining fixed. The Hu (1997) work is less relevant to the problem of this research since weights assigned to estimating functions involving data in the further distant past decline exponentially, and so the present time estimate effectively depends on data from a finite number of populations. Alternatively, the asymptotic paradigm of Wang et al. (2002) is relevant to the problem of interest in the situation where sample sizes from the present time period as well as historical time periods are increasingly large. We study the properties of the WEE estimate under this asymptotic paradigm in Sections 3.3 and 3.4.

Wang et al. (2002) show that the sequence of maximum weighted likelihood estimators are consistent and asymptotically normal as the number of observations from the populations increase under appropriate conditions. The authors give the conditions for consistency and asymptotic normality including the following assumption governing the selection of weights (adapted from Wang et al., 2002, p. 14).

- $w^{(n_T)} = \left(w_1^{(n_T)}, \dots, w_T^{(n_T)}\right)^T$ satisfies $w^{(n_T)} \to (v_1, \dots, v_T)^T \triangleq (0, \dots .,0,1)^T$ while $\max_{1\le t\prime\le T} n_{t\prime}^2 \max_{1\le t\le T} \left|v_t - w_t^{(n_T)}\right|^2 \le \mathcal{O}\left(n_T^{1-\delta}\right)$ as $n_T \to \infty$ for some $\delta > 0$

Here, the superscript (n_T) indicates that the select weights $\{w_t, t = 1, \dots, T\}$ are fixed for a particular value of n_T but different weights are selected as $n_T \to \infty$. The assumption requires that $w_t^{(n_T)} \le K \frac{n_T^{1/2-\delta}}{n_t}$ for some constants K and $\delta > 0$ and $t = 1, \dots, T-1$. Consider the case where relative sample size defined by $c_t = \frac{n_t}{N}$ remains constant for each t so that $n_T \to \infty$ and $n_t \to \infty$ at the same rate and $n_t = \mathcal{O}(n_T)$. Then, the assumption requires that $w_t^{(n_T)} \le K n_T^{-1/2-\delta}$. Under this requirement, the upper bound on $w_t^{(n_T)}$ gets smaller as $n_T \to \infty$. As sample sizes from each population increase, weights given to the terms for the related populations must decrease in order that consistency and asymptotical normality holds. This requirement supports the bias/variance trade-off provided by the WEE approach whereby in the case of a larger sample size at the present time period, we prefer to reduce the weight given to historical data in the estimating function. We consider the selection of weights in Section 3.1.

Wang et al. (2002) compares the performance of the MLE and MWLE for several examples. Since the true values of the parameters are unknown, then the authors replace the unknown quantity by the MLE. In the given disease mapping problem with data from each of seven years, the MWLE reduces the average mean squared error (MSE) by about 25%. In general, the impact of the MWLE depends on the values of the weights and the differences across populations and the authors state that the MWLE does not always reduce MSE when weights are selected independently from the data.

Since the WEE formulation in Section 1.6 is equivalent to the relevance weighted likelihood formulation of Hu and Zidek (2002), then the previous results can be applied to the WEE estimator. The WEE estimate is consistent and asymptotically normal under the conditions in Wang et al. (2002) including the condition stated above. An additional condition is that the random variables $\{Y_{jt}, j = 1, \dots, n_t\}$ are independent and identically distributed for each $= 1, \dots, T$. Since the problem at hand expects that the parameter may change slowly over time, then this condition is not satisfied. The effect on the asymptotic results can be studied through the authors' proof.

There is an important conceptual distinction between the WEE formulation and relevance weighted likelihood. The motivation for the WEE formulation stems from the problem to estimate a parameter that changes slowly over time. At each new time period, there is a new contribution to the estimating function. The estimating function contributions from the past time periods have increasing bias to the present time parameter and so it makes sense to further down-weight their contributions to the present time estimating function. Conceptually, under the WEE approach, the time order defines the relevance of past data.

2.8. Selecting weights

Further to the discussion in Section 2.7, we consider methods for selecting weights pertinent to the problem at hand. The moving average estimate is common in analysis of data collected at regular time intervals, where points within a defined window of the current observation are given equal weight and points outside that window are given zero weight. A moving average serves to smooth out short-term variability and dampen out unwanted periodic fluctuations. Additionally, weights related to an uncertainty measure such as known or observed variance or sample size are used in some applications to improve precision of the estimate or to correct for under-dispersion or over-dispersion in observed data. Survey data may be weighted relative to non-response by subgroups. Two other common approaches to weighting, exponentially declining weights and adaptive weights, are discussed here in more detail.

Exponentially declining weights

Exponentially weighted moving average (EWMA) control charts are an effective alternative to Shewhart control charts in detecting various types of process changes, including small sustained shifts in the process (Montgomery, 2013). Study has shown that the EWMA and suitable modifications have optimal properties for monitoring or estimating a process mean for a wide class of applications (Box, Jenkins, and MacGregor, 1974, Lucas and Saccucci, 1990). Let $X_1, X_2, \dots X_T$ be a sequence of observations collected at fixed intervals of time. The EWMA statistics are $Q_t = \lambda X_t + (1-\lambda)Q_{t-1}, t = 1,2, \dots$ where λ is a smoothing constant, $\lambda \in (0,1]$ and Q_0 is the initial value. More generally,

$$\begin{aligned} Q_T &= \sum_{t=1}^{T} \lambda(1-\lambda)^{T-t} X_t \\ &= \sum_{t=1}^{T} w_t X_t \end{aligned} \tag{9}$$

where $w_t = \lambda(1-\lambda)^{T-t}$. In all applications, a value for the constant λ must be selected. In quality monitoring applications, typical values for the parameter λ are between 0.05 and 0.25, and larger values may be used in forecasting and control applications (Steiner, 1999). Lucas and Saccucci (1990) give suggestions for λ that result in a desired minimum average run length under a specified shift in the process. The authors show that the optimal value of λ increases as the shift in the process increases.

Adaptive weights

Weights that adapt to properties of the data offer a mechanism for the data to self-weight according to some relevance measure and criteria. An application of MWLE studied in Wang et al. (2002) estimates a Poisson rate parameter for incidences of a disease in one location, based on a time series of event data from that location and other locations close in geographical proximity. The authors derive a model describing mean squared error of the parameter estimates as a function of the values of the weights, quantities describing the relevance of the locations to each other, assumed variances of data by location, and observed correlations. Then, the optimal values of the weights are estimated to minimize the mean squared error of the parameter estimates. The authors shows that these estimated weights are optimal in minimizing mean squared error of the parameter estimates.

An adaptive exponentially weighted estimation scheme was suggested by Yashchin (1995) to improve the estimation of a current process mean subject to abrupt changes. The estimate relies on the sequence of observations $X_1, \dots, X_T$ and $\{w_t\}$ as in (9) with $\{w_t\}$ dependent on the data. Yaschin recommends a scheme for setting $\{w_t\}$ that involves an estimate of r_T, the number of

observations preceding the observation at T, that are stationary relative to some threshold criteria. With $\hat{r}_T$ and smoothing parameter $\lambda \in (0,1]$, the current estimate of the process mean only involves those observations since the last point of change.

For the general problem of this research, observations are collected at regular time intervals and since we assume that parameters describing the outcome change slowly with time, then exponentially declining weights seem to be the most appropriate among the alternatives. We may be able to improve the mean squared error of estimates through a method of adaptive weights, but the need to estimate the weights is undesirable and the impact on the estimate of variance is not clear. Instead, the research focuses on the impact of the smoothing parameter and guidelines for its selection.

This literature review outlines some methodologies that have pertinence to the weighted estimating equations formulation for the general problem of this research. These methodologies support or provide alternatives or generalizations of the WEE approach. The rich foundation of literature provides opportunities to explore and expand the theoretical and applied aspects of the WEE approach. We discuss the fundamental aspects of the WEE approach for the motivating applications in Chapter 3.

Chapter 3: Weighted Estimating Equations Approach

The motivating problem of this research is to regulate a bias/variance trade-off in the estimate of the present value of a performance measure based on a stream of data collected over time. We expect that there may be a small number of subjects observed in the present time period and the model parameter may change slowly over time. We describe the data, objectives, and general approach to the problem in Section 1.2.

In the motivating applications of this research, changes to the model parameter over time may occur due to many complex factors. For example, customer loyalty may change due to continuous improvement in the product or process, new competitive products in the market, and changing media views of the product. We do not want to assume a stochastic or deterministic model to describe the change in the parameter since the change may be hard to predict and we want our approach to be flexible. Instead, we prefer to estimate the present value of the parameter assuming that the changes in the parameter over time are slow and so past data have relevance related to the proximity of the time period when they were observed to the current time period. For data d_t from each time period t up to the current time period T, we combine score contributions by time into an estimating equation that down-weights the contributions of historical data and estimates a single parameter $\hat{\theta}$. We call $\hat{\theta}$ the weighted estimating equation (WEE) estimate. We know that the WEE estimate $\hat{\theta}$ is a biased estimate of θ_T assuming that $\theta_T \neq \theta_t$ for $t = 1, \dots, T-1$, but $\hat{\theta}_T$ has less uncertainty than if we estimate it based on d_T alone. Since the sample size in the current time period is small, reducing uncertainty by incorporating historical data becomes important. This is the bias/variance trade-off which is the motivation for using the weighted estimating equations approach.

We remind the reader of the weighted estimating function (3) as well as the two naïve alternatives (1) and (2).

- The weighted estimating function down-weights the influence of historical data in the estimate of the present time parameter $\theta = \theta_T$ (Steiner and MacKay, 2014):

$$Q_3(\theta; d, w) = \sum_{t=1}^{T} w_t \psi_t(\theta; d_t) \tag{3}$$

  for a selection of weights $w = \{w_t, t = 1, \dots, T\}$ that decline over time periods $T, T-1, \dots, 1$. The weighted estimating equation is $\sum_{t=1}^{T} w_t \psi_t(\hat{\theta}; d_t) = 0$ and we denote the solution as the WEE estimate, $\hat{\theta}$.

- A naïve alternative is an estimating function based only on the data d_T observed in the most recent time period:

$$Q_1(\theta; d_T) = \psi_T(\theta; d_T) \quad (1)$$

- A naïve alternative is an estimating function based on all of the data $d = \{d_t; t = 1, \dots, T\}$ weighted equally:

$$Q_2(\theta; d) = \sum_{t=1}^{T} \psi_t(\theta; d_t) \quad (2)$$

The WEE approach to obtain a risk-adjusted estimate of the parameter and related inference follows.

Select standard population, model, and weights

- Select a standard population of J^* subjects that is important for inference in the application at hand. The objective is to estimate a risk-adjusted mean value of performance for this standard population to reliably compare estimates across time or across streams. Assign the value of the covariates for subjects in the standard population, $\left\{x_{j^*} = \left(x_{1,j^*}, \dots, x_{s,j^*}\right)^T, j^* = 1, \dots, J^*\right\}$.
- Select a model for the random variable Y_{jmt} in terms of a s-dimensional covariate vector x_{jmt}, stream m, and a p-dimensional model parameter, θ_t.
- Select weights $w = \{w_t, t = 1, \dots, T\}$ where $w_t \geq w_{t-1}$ for all t.

Define weighted estimating functions

- Define $Q(\theta; d, w)$ as in (3) which is a p-dimensional weighted estimating function for $\theta = \theta_T$ involving weighted score terms based on weights w and data $d = \{d_{jmt}\}$ observed on subjects $j = 1, \dots, n_{mt}$, in streams $m = 1, \dots, M$, at time periods $t = 1, \dots, T$.

Solve and calculate estimates

- Solve $Q(\hat{\theta}; d, w) = 0$ for the WEE estimate $\hat{\theta}$ of $\theta = \theta_T$.
- Estimate the expected value of the response for each of the standard population subjects in stream m, $\hat{\pi}_{j^*m}, j^* = 1, \dots, J^*$, using $\hat{\theta}$ and x_{j^*} (see Table 2-1).
- Estimate the mean for the standard population in stream m, $\hat{\pi}_m$, using $\{\hat{\pi}_{j^*m}, j^* = 1, \dots, J^*\}$ (see Table 2-1).
- Estimate the variance of $\hat{\theta}$. The sandwich estimate in (7) is one possibility.
- Estimate the variances of $\hat{\pi}_{j^*m}$ from (5) and $\hat{\pi}_m$ from (4).

Inference

According to the general objectives as stated in Section 1.2, we might want to

- Compare the estimate of the parameter to a target or benchmark value: use $\hat{\theta}$, $\widehat{var}(\hat{\theta})$
- Compare the risk-adjusted mean in stream m to a target or benchmark value: use $\hat{\pi}_m, \widehat{var}(\hat{\pi}_m)$
- Compare risk-adjusted mean estimates across streams: use $\hat{\pi}_{m_i}$, $\widehat{var}(\hat{\pi}_{m_i})$ vs. $\hat{\pi}_{m_j}$, $\widehat{var}\left(\hat{\pi}_{m_j}\right)$ for $m_i \neq m_j$.
- Compare mean estimates in stream m across groups of subjects: use $\hat{\pi}_{m_1}$, $\widehat{var}(\hat{\pi}_{m_1})$ vs. $\hat{\pi}_{m_2}$, $\widehat{var}(\hat{\pi}_{m_2})$ based on two standard populations $\{x_{j_1^*} \text{ for } j^* = 1, \dots, J_1^*\}$ and $\{x_{j_2^*} \text{ for } j^* = 1, \dots, J_2^*\}$.
- Monitor estimates of the parameter over time: use $\hat{\theta}$, $\widehat{var}(\hat{\theta})$ and historical $\hat{\theta}$, $\widehat{var}(\hat{\theta})$.
- Monitor estimates of the risk-adjusted mean in stream m over time: use $\hat{\pi}_m, \widehat{var}(\hat{\pi}_m)$ and historical $\hat{\pi}_m, \widehat{var}(\hat{\pi}_m)$. Note that $\hat{\pi}_m$ at each time period must be estimated for the same standard population. Recalculating $\hat{\pi}_m$ with the estimates $\hat{\theta}$ in previous time periods is possible if the standard population of interest changes.
- Test hypotheses on elements of the parameter: use $\hat{\theta}$, $\widehat{var}(\hat{\theta})$ and $\hat{\theta}_0$, $\widehat{var}(\hat{\theta}_0)$ which are estimates under the null hypothesis.
- Test hypotheses on the risk-adjusted mean in stream m: use $\hat{\pi}_m, \widehat{var}(\hat{\pi}_m)$ and $\hat{\pi}_{m_0}, \widehat{var}(\hat{\pi}_{m_0})$ which are based on estimates $\hat{\theta}_0$, $\widehat{var}(\hat{\theta}_0)$.

In Chapter 3, we discuss various aspects of this approach including weights selection, effective sample size, approximations for the variance of $\hat{\theta}$ and the distribution of a hypothesis test statistic involving $\hat{\theta}$, and implementing the WEE approach through SAS software. Some aspects of these results are explored through an analytic example in Section 3.6. The approach is demonstrated through the customer loyalty measure in Chapter 4, the lab positive abnormal rate in Chapter 5, and the hospital performance measure in Chapter 6. Considerations to implement this approach are also discussed in Chapter 6.

3.1. Selecting weights

Relative values of the weights $w = \{w_t, t = 1, \dots, T\}$ control the trade-off between bias and variance so w needs to be selected appropriately. In the general problem of this research, the score

functions by time period have a natural ordering and we expect that one or more of the p elements of parameter θ_t may drift slowly with time. Accordingly, we use weights that decrease (exponentially) for time periods further in the past. In particular, we propose to use a weight parameter, λ, having possible values $0 < \lambda < 1$ and to define the weights as in

$$w_t = \frac{\lambda(1-\lambda)^{T-t}}{\sum_{t=1}^{T} \lambda(1-\lambda)^{T-t}} \tag{10}$$

for each $t = 1, \dots, T$. These are exponentially declining weights as we discuss in Section 2.8. We select these since the exponentially weighted moving average control chart has desirable properties when there are small shifts in a process (see discussion in Section 2.3). Other definitions of decreasing weights are possible. With (10), the weight for the most recent time period is proportional to λ, the time period before that has weight proportional to $\lambda(1-\lambda)$, the time period before that $\lambda(1-\lambda)^2$, and so on. For convenience, we divide each weight by the same constant $\sum_{t=1}^{T} \lambda(1-\lambda)^{T-t}$ so that $\sum_{t=1}^{T} w_t = 1$. Note that this rescaling does not change the estimate of θ or its properties. Under (10), increasing the value of λ increases the relative weight of present data which reduces bias and increases variance of the estimator (assuming the parameter is changing over time). There is subjectivity in the selection of λ. Based on the guidelines for λ for exponentially declining weights as discussed in Section 2.8, the value $\lambda = 0.1$ is reasonable when the parameter drifts slowly over time.

Note that the two naïve approaches involving either present time data as in (1) or the aggregate of historical data weighted equally as in (2) are particular cases of (3) at the two limiting values of the weight parameter λ.

- As λ approaches 1, w_T approaches 1 and $w_t, t < T$ approaches 0. The estimating function involves the present time data d_T only.
- As λ approaches 0, w_t approaches $\frac{1}{T}$ for all $t = 1, \dots, T$. The estimating function involves the aggregate of data $\{d_1, d_2, \dots, d_T\}$ weighted equally.

3.2. Effective sample size

We consider the notion of effective sample size to compare the various approaches for the general problem of interest. We refer to effective sample size of an estimator as N_{eff}. Consider the value N_{eff} to be the number of independent observations that gives an estimate by the unweighted estimating function in (2) that has the same precision as the estimator involving N samples. For the WEE estimator, as $\lambda \to 1$ we give more weight to fewer observations and the

estimate has less precision. At the extreme ($\lambda \sim 1$), $N_{eff} = n_T$ for the naïve estimator involving d_T only. Conversely, as $\lambda \to 0$, we increase the weight across more of the observations and the estimate has more precision. At the extreme ($\lambda \sim 0$), $N_{eff} = N = \sum_{t=1}^{T} n_t$ for the naïve estimator involving $\{d_{jt}, j = 1, \dots, n_t, t = 1, \dots, T\}$ weighted equally. The WEE estimate based on a selection of λ between the two extremes has $n_T < N_{eff} < N$.

We compare the effective sample size of the WEE estimator to that of the EWMA approach. Consider a simple example where random variables $Y_{jt} \sim binomial(1, \theta_t)$ are independent over $j = 1, \dots, n_t$ and $t = 1, \dots, T$. We observe data $\{y_{jt}\}$ on subjects $j = 1, \dots, n_t$ over $t = 1, \dots, T$. There are no covariates and so the estimate of $\theta = \theta_T$ by the EWMA approach is

$$\hat{\theta}_{EWMA} = w_1 \bar{y}_1 + w_2 \bar{y}_2 + \cdots + w_T \bar{y}_T$$

where $\hat{\theta}_t = \bar{y}_t = \frac{\sum_{j=1}^{n_t} y_{jt}}{n_t}$ and $\sum_{t=1}^{T} w_t = 1$. The variance of the estimator is $ar(\theta_{EWMA}) = \sum_{t=1}^{T} \frac{w_t^2 \theta_t (1-\theta_t)}{n_t}$. The estimate of $\theta = \theta_T$ based on (2) where the *N* observations are weighted equally is

$$\hat{\theta}_{(2)} = \frac{\sum_{t=1}^{T} \sum_{j=1}^{n_t} y_{jt}}{N}$$

and the variance of the estimator is $var(\theta_{(2)}) = \frac{\sum_{t=1}^{T} n_t \theta_t (1-\theta_t)}{N^2}$ and so $N_{eff} = \left(\sum_{t=1}^{T} \frac{w_t^2}{n_t} \right)^{-1}$ for the EWMA estimator under the assumption that $\theta_t = \theta_T$ for all $t = 1, \dots, T$. In comparison, we can show that the parameter estimate by the WEE approach is

$$\hat{\theta}_{WEE} = \frac{w_1 n_1}{\sum_{t=1}^{T} w_t n_t} \bar{y}_1 + \frac{w_2 n_2}{\sum_{t=1}^{T} w_t n_t} \bar{y}_2 + \cdots + \frac{w_T n_T}{\sum_{t=1}^{T} w_t n_t} \bar{y}_T$$

and $var(\theta) = \frac{\sum_{t=1}^{T} w_t^2 n_t \theta_t (1-\theta_t)}{\left(\sum_{t=1}^{T} w_t n_t \right)^2}$. For the WEE estimator, $N_{eff} = \frac{\left(\sum_{t=1}^{T} w_t n_t \right)^2}{\sum_{t=1}^{T} w_t^2 n_t}$ under the assumption that $\theta_t = \theta_T$ for all $t = 1, \dots, T$. By this simple example, we see that the WEE estimate can be rewritten as an EWMA estimate where the total weight given to the statistic based on d_t is proportional to $w_t n_t$. The WEE estimate weights a time period statistic proportional to the sample size in that time period and declining for the further past. Under the EWMA approach, the sample size by time period does not affect the weight given to the statistic at that time period. Here, we see by the expression for N_{eff} that a small sample size at any time period greatly decreases the N_{eff} of the EWMA estimator, illustrating the shortcoming of the EWMA approach discussed in

Section 2.3. In the limit, as the sample size in one time period approaches zero, N_{eff} approaches zero. We see by this simple example that this is not a problem for the WEE estimator. In general, the two estimators give the same estimates when sample sizes are the same across all time periods; however, when there are small sample sizes in some time periods, as in the general problem of this research, the comparison of effective sample sizes favours the WEE estimator over the EWMA estimator.

3.3. Estimate of variance

Inference based on the WEE estimate $\hat{\theta}$ requires an estimate of its uncertainty. For example, a manager may want to assess whether the mean performance estimate based on $\hat{\theta}$ is significantly different than the competitive benchmark. We derive approximations for the variance of $\hat{\theta}$ using the usual asymptotic properties of the information and score functions in the model based on data by time period. We assume that the model parameter θ_t does not change over time $t = 1, \dots, T$. So there are two sources of error in the approximations; first, the usual error due to the asymptotics and a second error due to the fact that the parameter may have drifted.

We assume that the model $\mathcal{L}_t(\tilde{\theta}; \mathcal{D}_t)$ holds for each $t = 1, \dots, T$ and that the random variables $\mathcal{D}_t$ are independent over t. In the case where the model depends on covariates, then we assume that $\mathcal{D}_t$ are independent over t, conditional on the values of the covariates. Note that we do not model changes in the covariates. For θ, the unknown model parameter, $I_t(\theta) = -E\left(\frac{\partial^2 \log \mathcal{L}_t(\theta; \mathcal{D}_t)}{\partial \theta^2}\right)$ is the matrix of expected information about θ at time t, $i_t(\theta) = -l_t''(\theta; d_t)$ is the observed information matrix, and the two are related by $E\big(i_t(\theta)\big) = I_t(\theta)$ (Small, 2010). Since the weighted estimating functions combine the usual score functions by time period, then we consider an estimate of $var\big(\tilde{\theta}\big)$ through the known asymptotic properties of the corresponding information and score functions.

We consider the asymptotic properties of the information and score functions in the case where the total sample size $N = \sum_{t=1}^{T} n_t$ approaches infinity and the number of time periods T remains fixed. In order to preserve the usual asymptotic properties of these functions by time period as $\to \infty$, we need to preserve some uniformity in the relative distributions of $I_t(\theta)$ by time period $t = 1, \dots, T$. We require that the relative sample size defined by $c_t = \frac{n_t}{N}$ remains constant for each t so that $n_t \to \infty$ as $\to \infty$. In the case where the model does not depend on covariates, then each individual unit has the same expected information and so, for fixed c_t, the relative distributions of

$I_t(\theta), t = 1, \dots, T$, stay the same as $N \to \infty$. In the more general problem where the model depends on covariates, then some uniformity in the distribution of samples across the covariate space must be maintained as each $n_t \to \infty$ so that $\frac{I_t(\theta)}{n_t} \to g_t(\theta)$ for some constant matrix $g_t(\theta)$. We derive an approximation for $var(\tilde{\theta})$ under this asymptotic paradigm.

First, we sketch a proof to show that $\tilde{\theta}$ is a consistent estimator of the true value $\theta = \theta_T$ under usual regularity conditions and under the condition that θ_t does not change over time $= 1, \dots, T$. A rigorous proof of consistency of the WEE estimator would follow the method in Wald (1949) for a MLE estimator. We denote θ_0 as the true value of $\theta = \theta_t$ for $t = 1, \dots, T$.

Lemma: For any $\theta \neq \theta_0$ we have $E(l_t(X, \theta)) < E(l_t(X, \theta_0))$ where X is a random variable having distribution $f(x, \theta_0)$ and $l_t(x_1, \dots, x_{n_t}, \theta) = \sum_{j=1}^{n_t} log\, f(x_j, \theta)$. See Wald (1949) for proof.

Theorem: Under usual regularity conditions on the family of distributions, the WEE estimate $\hat{\theta}$ is consistent; that is, $\hat{\theta} \xrightarrow{p} \theta_0$ as $N \to \infty$.

Sketch of proof: We have the following facts:

- $\hat{\theta}$ is a maximizer of $\sum_{t=1}^{T} w_t l_t(x, \theta)$ by definition
- θ_0 is the maximizer of $E(l_t(X, \theta))$ by the Lemma. It follows that θ_0 is also the maximizer of $E(\sum_{t=1}^{T} w_t l_t(X, \theta))$.

By the Law of Large Numbers, $\sum_{t=1}^{T} w_t l_t(x, \theta) \xrightarrow{p} E(\sum_{t=1}^{T} w_t l_t(X, \theta))$ for all θ as $N \to \infty$. Since two functions are getting closer, then the points of maximum should also get closer which means that $\hat{\theta} \xrightarrow{p} \theta_0$ as $N \to \infty$.

Next, we derive the estimate of the variance of $\tilde{\theta}$ for a model that does not depend on covariates. For $I(\theta)$, the expected information from a single sample, and $I_t(\theta) = n_t I(\theta)$, the expected information from all samples at t,

$$var(\psi_t(\theta; \mathcal{D}_t)) = I_t(\theta) = n_t I(\theta) = N c_t I(\theta)$$

since $c_t = \frac{n_t}{N}$ for all t. Then, by the Central Limit Theorem,

$$\frac{\psi_t(\theta; \mathcal{D}_t)}{\sqrt{n_t}} \xrightarrow{D} N_p(0, I(\theta))$$

since ψ_t is the sum of n_t terms each with mean vector 0_p and covariance matrix $I(\theta)$ for each t as $n_t \to \infty$. Since $\mathcal{D}_t$ are assumed to be independent across time $t = 1, \dots, T$ and w_t and c_t are constants, then

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{T} w_t\psi_t(\theta;\mathcal{D}_t) \xrightarrow{D} N_p(0,\sum_{t=1}^{T} w_t^2 c_t I(\theta))$$

We consider the first order Taylor Series approximation of $\psi(\hat{\theta})$ for $\hat{\theta}$ near θ,

$$\left(\hat{\theta}-\theta\right) \approx [-\psi'(\theta)]^{-1}\psi(\theta)$$

since $\psi\left(\hat{\theta}\right)=0$. We extend this to an approximation for the corresponding random variable $\left(\tilde{\theta}-\theta\right)$ with observed information at time t, $i_t(\theta)=-\psi_t'(\theta)$, so

$$\sqrt{N}\left(\tilde{\theta}-\theta\right) \approx \left(\frac{1}{N}\sum_{t=1}^{T} w_t i_t(\theta)\right)^{-1} \frac{1}{\sqrt{N}}\sum_{t=1}^{T} w_t\psi_t(\theta)$$

since $\tilde{\theta}$ is consistent. Then, by Slutsky's Theorem,

$$\sqrt{N}\left(\tilde{\theta}-\theta\right) \xrightarrow{D} (\sum_{t=1}^{T} w_t c_t I(\theta))^{-1} Z$$

as $N\to\infty$, since $E[i_t(\theta)]=I_t(\theta)=Nc_tI(\theta)$ and Z is the asymptotic distribution of $\frac{1}{\sqrt{N}}\sum_{t=1}^{T} w_t\psi_t(\theta;\mathcal{D}_t)$. Then, with the previous result for Z,

$$\sqrt{N}\left(\tilde{\theta}-\theta\right) \xrightarrow{D} N_p(0,(\sum_{t=1}^{T} w_t c_t I(\theta))^{-1}\sum_{t=1}^{T} w_t^2 c_t I(\theta)\,(\sum_{t=1}^{T} w_t c_t I(\theta))^{-1})$$

Then, an estimate for the asymptotic variance of $\tilde{\theta}$ is

$$\widehat{var}_{WI}\left(\tilde{\theta};\hat{\theta}\right)=\left(N\sum_{t=1}^{T} w_t c_t I\left(\hat{\theta}\right)\right)^{-1} N\sum_{t=1}^{T} w_t^2 c_t I\left(\hat{\theta}\right)\left(N\sum_{t=1}^{T} w_t c_t I\left(\hat{\theta}\right)\right)^{-1}$$

More generally in the case where the model depends on the covariates and $\frac{I_t(\theta)}{n_t}\to g_t(\theta)$ as $n_t\to\infty$ for $g_t(\theta)$ a matrix of constants, we extend this estimate as

$$\widehat{var}_{WI}\left(\tilde{\theta};\hat{\theta}\right)=\left(\sum_{t=1}^{T} w_t I_t\left(\hat{\theta}\right)\right)^{-1}\sum_{t=1}^{T} w_t^2 I_t\left(\hat{\theta}\right)\left(\sum_{t=1}^{T} w_t I_t\left(\hat{\theta}\right)\right)^{-1} \tag{11}$$

given weights $\{w_t\}$ and expected information matrices evaluated at the WEE estimate, $\left\{I_t\left(\hat{\theta}\right), t=1,\dots,T\right\}$. We refer to (11) as the weighted information (WI) estimate of variance. We use this approximation for the variance of the random variable $\tilde{\theta}$ to estimate the variance of the WEE estimate $\hat{\theta}$. Note that the approximation for the variance of $\tilde{\theta}$ in (11) does not change if we scale each w_t by the same constant.

Note that result (11) at the two special cases of weight values described previously gives the usual estimates of variance. In the case where $w_T=1$ and $w_t=0$ for all $t<T$, then

$$\begin{aligned}\widehat{var}_{WI}\left(\hat{\theta}\right) &= I_T^{-1}\left(\hat{\theta}\right) I_T\left(\hat{\theta}\right) I_T^{-1}\left(\hat{\theta}\right)\\ &= I_T^{-1}\left(\hat{\theta}\right)\end{aligned}$$

In the case where $w_t = \frac{1}{T}$ for all t, then

$$\widehat{var}_{WI}(\hat{\theta}) = \left(\sum_{t=1}^{T} \frac{I_t(\hat{\theta})}{T}\right)^{-1} \sum_{t=1}^{T} \frac{I_t(\hat{\theta})}{T^2} \left(\sum_{t=1}^{T} \frac{I_t(\hat{\theta})}{T}\right)^{-1}$$
$$= \left(\sum_{t=1}^{T} I_t(\hat{\theta})\right)^{-1}$$

We compare the weighted information estimate of variance to the sandwich estimate of variance in (7). Based on the definitions of the components of (7), we can show that $\hat{V}(\hat{\theta}) = \sum_{t=1}^{T} w_t I_t(\hat{\theta})$ and $\hat{B}(\hat{\theta}) = \sum_{t=1}^{T} w_t^2 I_t(\hat{\theta})$ and so $\widehat{var}_S(\hat{\theta}) = \widehat{var}_{WI}(\hat{\theta})$. The sandwich estimate of variance was proposed for maximum likelihood estimates of a misspecified model or under missing covariate data (White, 1982). Through this work we justify its use as an estimate of the variance of the WEE estimator relative to the specified asymptotic paradigm.

3.4. Distribution of hypothesis test statistic

A specific application of interest may require a test of hypothesis involving the WEE estimate at the current time T. For example, a process quality manager responsible for checking the consistency of multiple parallel gauges may want to monitor a test statistic for the hypothesis that the parameters of the model describing the gauge effects are the same. This activity requires an approximation for the distribution of a test statistic involving the WEE estimate under a null hypothesis versus a specified alternative hypothesis. We assume that the model parameter θ_t does not change over time $t = 1, \ldots, T$.

Here, we consider a test statistic based on a likelihood ratio (LR), though a Wald or score test statistic could also be constructed (Lehmann and Romano, 2005). Consider a partition of the parameter vector $\theta = \theta_T$ into $\theta = (\delta^T, \alpha^T)^T$ where δ is the vector of parameters of interest for testing and α is the vector of unrestricted parameters. Let the number of independent restrictions on parameters in δ be r. For example, when monitoring the consistency of M binary gauges, suppose the parameter α represents the pass rate for a baseline gauge and parameter $\delta = (\delta_1, \ldots, \delta_{M-1})^T$ represents the pass rate of the other gauges relative to the baseline. We test for consistency across the M gauges through a test of the null hypothesis $H_0: \delta_1 = \ldots = \delta_{M-1} = 0$ versus the alternative H_A: at least one $\delta_1, \ldots, \delta_{M-1} \neq 0$.

The general null hypothesis of interest is $H_0: \delta = \delta_0$. To construct a LR test statistic, we estimate $\theta = (\delta^T, \alpha^T)^T$ under the unrestricted model and α_0 when δ is restricted to δ_0. The weighted estimating function (3) gives WEE estimates $\hat{\theta}$ and $\hat{\alpha}_0$. The WEE approach extends the

usual LR test statistic by comparing the weighted log-likelihood contributions by time under the unrestricted and restricted models. The WEE LR test statistic is

$$\hat{S} = 2\left(\sum_{t=1}^{T} w_t l_t\left(\hat{\theta}; d_t\right) - \sum_{t=1}^{T} w_t l_t(\delta_0, \hat{\alpha}_0; d_t)\right) \quad (12)$$

at WEE estimates $\hat{\theta}$ and $\hat{\alpha}_0$. We consider the distribution of the corresponding random variable $\tilde{S}$ under the null hypothesis and the asymptotic paradigm discussed in Section 3.3. First, we derive an approximate distribution for $\tilde{S}$ when $\dim(\theta) = 1$, that is in a model with no covariates and a single parameter. Later, we show that the result holds when the model has covariates and $\frac{I_t(\theta)}{n_t} \to g(\theta)$; that is, the average expected information in the limit is the same for all t.

The likelihood ratio test of the simple null hypothesis $H_0: \theta = \theta_0$ against the alternative hypothesis $H_A: \theta \neq \theta_0$ is based on the likelihood ratio random variable

$$\tilde{S} = 2\left(\sum_{t=1}^{T} w_t l_t\left(\tilde{\theta}\right) - \sum_{t=1}^{T} w_t l_t(\theta_0)\right)$$

Consider the second degree Taylor Series approximation of $\sum_{t=1}^{T} w_t l_t(\theta_0)$ for θ_0 near $\hat{\theta}$,

$$\sum_{t=1}^{T} w_t l_t(\theta_0) \approx \sum_{t=1}^{T} w_t l_t\left(\hat{\theta}\right) + \left(\theta_0 - \hat{\theta}\right)^T \sum_{t=1}^{T} w_t l_t'\left(\hat{\theta}\right) + \frac{1}{2}\left(\theta_0 - \hat{\theta}\right)^T \sum_{t=1}^{T} w_t l_t''\left(\hat{\theta}\right)\left(\theta_0 - \hat{\theta}\right)$$

Since $\sum_{t=1}^{T} w_t l_t'\left(\hat{\theta}\right) = 0$ and observed information matrix $i_t(\theta) = -l_t''(\theta)$, then

$$\hat{S} = 2\left(\sum_{t=1}^{T} w_t l_t\left(\hat{\theta}\right) - \sum_{t=1}^{T} w_t l_t(\theta_0)\right) \approx \sqrt{N}\left(\hat{\theta} - \theta_0\right)^T \frac{1}{N} \sum_{t=1}^{T} w_t i_t\left(\hat{\theta}\right) \sqrt{N}\left(\hat{\theta} - \theta_0\right)$$

We extend this result for $\hat{S}$ to the random variable $\tilde{S}$. We consider the case where the model does not depend on covariates. Then, $\tilde{S}$ has the same asymptotic distribution as

$$\sqrt{N}\left(\tilde{\theta} - \theta_0\right)^T \sum_{t=1}^{T} w_t c_t I\left(\tilde{\theta}\right) \sqrt{N}\left(\tilde{\theta} - \theta_0\right)$$

since $E[i_t(\theta)] = N c_t I(\theta)$. In Section 3.3, we show that under regularity conditions and consistency,

$$\sqrt{N}\left(\tilde{\theta} - \theta_0\right) \xrightarrow{D} N_p\left(0, \left(\sum_{t=1}^{T} w_t c_t I(\tilde{\theta})\right)^{-1} \sum_{t=1}^{T} w_t^2 c_t I(\tilde{\theta}) \left(\sum_{t=1}^{T} w_t c_t I(\tilde{\theta})\right)^{-1}\right) \text{ as } N \to \infty.$$

It follows that

$$\sqrt{N}\left(\tilde{\theta} - \theta_0\right)^T \left(\sum_{t=1}^{T} w_t c_t I(\tilde{\theta})\right) \left(\sum_{t=1}^{T} w_t^2 c_t I\left(\tilde{\theta}\right)\right)^{-1} \left(\sum_{t=1}^{T} w_t c_t I(\tilde{\theta})\right) \sqrt{N}\left(\tilde{\theta} - \theta_0\right) \xrightarrow{D} \chi_p^2$$

as $N \to \infty$. With this asymptotic result, we state an approximation for the distribution of $\tilde{S} \sim \sqrt{N}\left(\tilde{\theta} - \theta_0\right)^T \sum_{t=1}^{T} w_t c_t I\left(\tilde{\theta}\right) \sqrt{N}\left(\tilde{\theta} - \theta_0\right)$ in the case that $\dim(\theta) = 1$. Since $I(\theta)$ is a scalar, then

$$(\sum_{t=1}^{T} w_t^2 c_t)^{-1}(\sum_{t=1}^{T} w_t c_t)\, \tilde{S} \xrightarrow{D} \chi_1^2 \quad \text{as } N \to \infty$$

under the null hypothesis.

More generally where the model depends on the covariates, we consider the case where the average expected information in the limit is the same for all t so that $\frac{I_t(\theta)}{n_t} \to g(\theta)$. In the limit as n_t and N get large, then

$$I_t(\theta) \approx n_t g(\theta) \approx N c_t g(\theta)$$

for each $t = 1, \dots, T$. The previous results in Section 3.3 involving $I(\theta)$ extend to results involving $g(\theta)$. Then, in the case where $g(\theta)$ is a scalar, it follows that

$$(\sum_{t=1}^{T} w_t^2 c_t)^{-1}(\sum_{t=1}^{T} w_t c_t)\, \tilde{S} \xrightarrow{D} \chi_1^2 \quad \text{as } N \to \infty$$

under the null hypothesis. These results extend to the more general case where $\dim(\theta) = p \geq 1$,

$$(\sum_{t=1}^{T} w_t^2 c_t)^{-1}(\sum_{t=1}^{T} w_t c_t)\, \tilde{S} \xrightarrow{D} \chi_p^2 \quad \text{as } N \to \infty$$

under the simple null hypothesis. For testing $r < p$ restrictions on θ, then we can show by a similar argument that

$$(\sum_{t=1}^{T} w_t^2\, c_t)^{-1}(\sum_{t=1}^{T} w_t c_t)\, \tilde{S} \xrightarrow{D} \chi_r^2 \quad \text{as } N \to \infty$$

under the null hypothesis. In practice, we replace c_t by $\frac{n_t}{N}$ and use these results to approximate the distribution of the weight-adjusted test statistic $\frac{\sum_{t=1}^{T} w_t n_t}{\sum_{t=1}^{T} w_t^2 n_t}\hat{S}$. With this asymptotic result, we approximate the distribution of the weighted random variable $\tilde{S}$,

$$\frac{\sum_{t=1}^{T} w_t n_t}{\sum_{t=1}^{T} w_t^2 n_t}\tilde{S} \overset{\text{approx}}{\sim} \chi_p^2$$

under a simple null hypothesis. For testing $r < p$ restrictions on θ, then

$$\frac{\sum_{t=1}^{T} w_t n_t}{\sum_{t=1}^{T} w_t^2 n_t}\tilde{S} \overset{\text{approx}}{\sim} \chi_r^2 \tag{13}$$

under the null hypothesis. Note that at the two special cases of weight values described previously, (13) gives the usual results using present time data only or all data weighted equally. The extension of (13) to the most general case where $\dim(\theta) \geq 1$ and the average expected information in the limit is not the same for all time periods is not straightforward. This remains as future work. Note

that the distribution of the weight-adjusted test statistic in (13) does not change if we scale each w_t by the same constant.

The argument for consistency and the derivations of the approximate results (11) and (13) assume that the true value of parameter θ_t is the same across the $t = 1, \ldots, T$ time periods. The general problem of this research expects that the parameter may drift over time and so these results are approximations. Since we restrict our focus to slow changes in the parameter over time, then we expect that these results are reasonable approximations. In Section 3.6, we show an example where the parameter changes slowly over time. Here, the WEE approach with an appropriate weight parameter gives an estimate with lower mean squared error than a naïve approach where no weights are used. This property holds for a wide variety of problems.

3.5 Criteria for comparing WEE to alternative approaches

We consider a criterion for comparing the WEE estimator to the naïve and EWMA estimators in this research. Since the objective is to set up a bias/variance trade-off in the estimate of a parameter, then we consider the efficiency measure

$$\text{root mean squared error}(\tilde{\theta}) = \sqrt{\text{bias}(\tilde{\theta})^2 + \text{variance}(\tilde{\theta})}$$

We refer to root mean squared error as MSE and we prefer the estimator having the minimum value of $\text{MSE}(\tilde{\theta})$ over the alternatives. This efficiency measure is widely used in statistical learning problems and is closely related to expected prediction error (Hastie, Tibshirani, and Friedman, 2009). In Section 4.2 and Section 5.2, the comparison of the efficiency measure across estimates from various alternatives is based on simulated data where the true value is known and so bias can be computed. We compare WEE to alternative methods based on the sensitivity of MSE to the selection of weight values, sample sizes, and the speed of change in the true value of the parameter. For the lab positive abnormal rate where we want a test of hypothesis, we compare power of test and Type II error among test statistics from various alternatives.

3.6 Analytic example

We look at an example of a simple process with multiple streams to look at properties of the WEE parameter estimate, the WI estimate of variance, and the WEE LR test statistic. The simple process generates binary observations from units in two streams over time. The observations are the quantities of passed units y_{1t}, y_{2t} among n_{1t}, n_{2t} units tested at time t arising from two gauges

performing the same test. The objective is to monitor the difference in the pass rates from the gauges over time. The simplicity of the example is convenient for demonstration purposes. Similar demonstrations can be made over a wide class of models.

We consider random variables $Y_{mt} \sim binomial(n_{mt}, \pi_{mt})$ for $m = 1,2$ that we assume are each independent over $t = 1, \dots, T$. For $\pi = \pi_2 - \pi_1$, the difference of the mean pass rates at the two streams at the present time, we want to test the null hypothesis $H_0: \pi = 0$ versus the alternative $H_A: \pi \neq 0$. We expect that one or both of the true values of the elements of $\theta_t = (\pi_{1t}, \pi_{2t})$ may change slowly over time.

Assuming that $\pi_{mt} = \pi_m, m = 1,2$ for each t, a closed-form solution for the WEE estimate $\hat{\theta}$ is possible for this simple example,

$$\hat{\pi}_1 = \frac{\sum_{t=1}^{T} w_t y_{1t}}{\sum_{t=1}^{T} w_t n_{1t}}, \hat{\pi}_2 = \frac{\sum_{t=1}^{T} w_t y_{2t}}{\sum_{t=1}^{T} w_t n_{2t}}$$

The expected information matrix is $I_t(\theta) = \begin{bmatrix} \frac{n_{1,t}}{\pi_1(1-\pi_1)} & 0 \\ 0 & \frac{n_{2,t}}{\pi_2(1-\pi_2)} \end{bmatrix}$ and so the estimate of variance of $\tilde{\theta}$ by (11) is

$$\widehat{var}(\hat{\pi}_m) = \frac{\sum_{t=1}^{T} w_t^2 n_{mt}\, \hat{\pi}_m(1-\hat{\pi}_m)}{\left(\sum_{t=1}^{T} w_t n_{mt}\right)^2}, m = 1,2$$

The parameter of interest to compare the pass rates between the two streams is $\pi = \pi_2 - \pi_1$. Based on the preceding estimates,

$$\hat{\pi} = \frac{\sum_{t=1}^{T} w_t y_{2t}}{\sum_{t=1}^{T} w_t n_{2t}} - \frac{\sum_{t=1}^{T} w_t y_{1t}}{\sum_{t=1}^{T} w_t n_{1t}}, \quad \widehat{var}(\hat{\pi}) = \sum_{m=1}^{2} \frac{\hat{\pi}_m(1-\hat{\pi}_m) \sum_{t=1}^{T} w_t^2 n_{mt}}{\left(\sum_{t=1}^{T} w_t n_{mt}\right)^2}$$

To test the hypothesis $H_0: \pi = 0$, the WEE LR test statistic (12) is

$$\begin{aligned} \hat{S} = {} & 2 \sum_{m=1}^{2} (log\, \hat{\pi}_m \sum_{t=1}^{T} w_t y_{mt} + log(1-\hat{\pi}_m) \sum_{t=1}^{T} w_t (n_{mt} - y_{mt})) \\ & -2\, log\, \hat{\pi}_0 \sum_{m=1}^{2} \sum_{t=1}^{T} w_t y_{mt} - 2\, log(1-\hat{\pi}_0) \sum_{m=1}^{2} \sum_{t=1}^{T} w_t (n_{mt} - y_{mt}) \end{aligned}$$

for $\hat{\pi}_1$ and $\hat{\pi}_2$ as previously stated and $\hat{\pi}_0 = \frac{\sum_{m=1}^{2} \sum_{t=1}^{T} w_t y_{mt}}{\sum_{m=1}^{2} \sum_{t=1}^{T} w_t n_{mt}}$ under the null hypothesis. The usual MLE estimates involving the historical observations are special cases of these estimates with $w_t = 1$ (or $w_t = \frac{1}{T}$) for all t. Based on the parameter and test statistic estimates for this simple problem, we consider four properties as follows.

i. the estimate of $var(\tilde{\theta})$ in (11) is appropriate

Given the simple model, we estimate $var(\tilde{\pi})$ directly by the distributions of the random variables $\{Y_{1,t}, Y_{2,t}, t = 1, \dots, T\}$. The WI estimate of variance by (11) is the same as the closed-form expression of variance derived directly from the distributions of the random variables. Since no asymptotic assumptions are required for the latter formulation, then the weighted information estimate of variance is a suitable estimate even when there are small samples for this simple example.

ii. small sample sizes have less impact on the estimate of $var(\tilde{\theta})$ based on the WEE approach than based EWMA approach

The estimates of π and $var(\tilde{\pi})$ by the EWMA and WEE approaches are given in Table 3-1.

Table 3-1. Estimates of π and $var(\tilde{\pi})$ by EWMA and WEE Approaches

	Estimate of π	**Estimate of variance of $\tilde{\pi}$**
EWMA	$\hat{\pi}_{EWMA} = \sum_{t=1}^{T} w_t \left(\frac{y_{2t}}{n_{2t}} - \frac{y_{1t}}{n_{1t}}\right)$	$\widehat{var}_{EWMA} = \sum_{m=1}^{2} \hat{\pi}_m(1-\hat{\pi}_m) \sum_{t=1}^{T} \frac{w_t^2}{n_{mt}}$
WEE	$\hat{\pi}_{WEE} = \frac{\sum_{t=1}^{T} w_t y_{2t}}{\sum_{t=1}^{T} w_t n_{2t}} - \frac{\sum_{t=1}^{T} w_t y_{1t}}{\sum_{t=1}^{T} w_t n_{1t}}$	$\widehat{var}_{WI} = \sum_{m=1}^{2} \frac{\hat{\pi}_m(1-\hat{\pi}_m) \sum_{t=1}^{T} w_t^2 n_{mt}}{\left(\sum_{t=1}^{T} w_t n_{mt}\right)^2}$

We compare the impact of small samples on these estimates. If there are constant sample sizes across time periods, $n_{1t} = n_1$ and $n_{2t} = n_2$ for all t, then $\hat{\pi}_{EWMA} = \hat{\pi}_{WEE}$ and $\widehat{var}_{EWMA} = \widehat{var}_{WI}$. We discussed this property of equal variance estimates in the comparison of effective sample sizes in Section 3.2. The two approaches are equivalent when the sample sizes are the same across all time periods.

Consider the case where one sample size by time period is different. Specifically, let the sample size at the current time period T be one tenth as large as the rest so $n_{1t} = n_1$ and $n_{2t} = n_2$ for $t = 1, \dots, T-1$ and $n_{1T} = 0.1n_1$ and $n_{2T} = 0.1n_2$. The estimates for π are different in this case,

$$\hat{\pi}_{EWMA} = \sum_{t=1}^{T-1} w_t \left(\frac{y_{2t}}{n_2} - \frac{y_{1t}}{n_1}\right) + 10w_T \left(\frac{y_{2T}}{n_2} - \frac{y_{1T}}{n_1}\right)$$

$$\hat{\pi}_{WEE} = \frac{\sum_{t=1}^{T} w_t y_{2t}}{n_2\left(\sum_{t=1}^{T-1} w_t + 0.1w_T\right)} - \frac{\sum_{t=1}^{T} w_t y_{1t}}{n_1\left(\sum_{t=1}^{T-1} w_t + 0.1w_T\right)}$$

The estimates of variance of $\tilde{\pi}$ by the two approaches are

$$\widehat{var}_{EWMA} = \sum_{m=1}^{2} \frac{\hat{\pi}_m(1-\hat{\pi}_m)}{n_m} \left(\sum_{t=1}^{T-1} w_t^2 + 10w_T^2\right)$$

$$\widehat{var}_{WI} = \sum_{m=1}^{2} \frac{\hat{\pi}_m(1-\hat{\pi}_m)}{n_m} \frac{\sum_{t=1}^{T-1} w_t^2 + 0.1w_T^2}{\left(\sum_{t=1}^{T-1} w_t + 0.1w_T\right)^2}$$

Note that for $w_T = 1$ and $w_t = 0$ for $t < T$, then $\widehat{var}_{EWMA} = \widehat{var}_{WI}$. Otherwise, for $0 < w_t < 1$ for all t, then $\widehat{var}_{EWMA} > \widehat{var}_{WI}$. Except in the naïve approach using present time data only, the estimate of variance of the EWMA estimator is larger than the estimate of variance of the WEE estimator. By extension, with a small sample size at any time period, there is less precision in the EWMA estimate of θ than in the estimate by the WEE approach.

iii. the WEE estimate and WI estimate of $var(\tilde{\theta})$ is suitable with small changes in θ_t over time periods $t = 1, \dots, T$

We study the effect of a change in true value π on the $bias(\tilde{\pi}) = E(\tilde{\pi}) - \pi$ and $\widehat{var}(\tilde{\pi})$. For this study, we choose arbitrary values:

- $T = 10$ time periods of data observed
- sample sizes $n_{1t} = n_1 = 100$ and $n_{2t} = n_2 = 60$ for all $t = 1, \dots, T$
- stream 2 experiences a positive step change in rate π_{2t} of size Δ at time $t = 6$

Streams 1 and 2 have the same initial pass rates, so $\pi_{1,1} = \pi_{2,1}$. We vary initial values $\pi_{1,1} = \pi_{2,1}$ and size of change $\Delta = \pi_{2,6} - \pi_{2,5}$ to compare the properties of the WEE estimator over various profiles. Note that under a change in pass rate at stream 2, the true value of $\pi_{2,t}$ is $\pi_{2,1}$ for $t < 6$ and $\pi_{2,1} + \Delta$ for $t \geq 6$. Then, the quantity $E(Y_{2t})$ in $E(\tilde{\pi})$ and $I_t(\tilde{\pi})$ depends on t and the size of the change Δ for $t \geq 6$. At the present time $T = 10$, the expected value of estimator $\tilde{\pi}$ is $E(\tilde{\pi}; \Delta) = E\left(\frac{\sum_{t=1}^{T} w_t Y_{2t}}{\sum_{t=1}^{T} w_t n_{2t}} - \frac{\sum_{t=1}^{T} w_t Y_{1t}}{\sum_{t=1}^{T} w_t n_{1t}}\right) = \pi_{2,1} + \Delta \sum_{t=6}^{10} w_t - \pi_{1,1}$ and its true value is $\pi = \pi_{2,1} + \Delta - \pi_{1,1}$. The bias in estimator $\tilde{\pi}$ at time T is

$$bias(\tilde{\pi}; \Delta) = \Delta\left(\sum_{t=6}^{10} w_t - 1\right)$$

The weighted information estimate of variance of $\tilde{\pi}$ based on (11) is

$$\widehat{var}_{WI}(\tilde{\pi}; \Delta) = \frac{\hat{\pi}_{1,1}\left(1-\hat{\pi}_{1,1}\right)\sum_{t=1}^{10} w_t^2}{n_1} + \frac{\frac{1}{\hat{\pi}_{2,1}(1-\hat{\pi}_{2,1})}\sum_{t=1}^{5} w_t^2 + \frac{1}{(\hat{\pi}_{2,1}+\Delta)(1-\hat{\pi}_{2,1}-\Delta)}\sum_{t=6}^{10} w_t^2}{n_2\left(\frac{1}{\hat{\pi}_{2,1}(1-\hat{\pi}_{2,1})}\sum_{t=1}^{5} w_t + \frac{1}{(\hat{\pi}_{2,1}+\Delta)(1-\hat{\pi}_{2,1}-\Delta)}\sum_{t=6}^{10} w_t\right)^2}$$

We study the bias and variance of the WEE estimator through root mean squared error for various sizes of change, Δ. We calculate $\text{MSE}(\tilde{\pi}, \Delta)$ for values of $\pi_{1,1} = \pi_{2,1}$ in the range of 0.02 to 0.20

and values of Δ in the range of 25% to 100% of each of the starting values $\pi_{1,1} = \pi_{2,1}$. Figure 3-1 gives contour plots of the relatives values of MSE for the values of $\pi_{1,1} = \pi_{2,1}$ and Δ. The relative values compare MSE for the WEE estimator with weight parameter $\lambda = 0.1$ to that of each of the two naïve estimators having limiting values of the weight parameters.

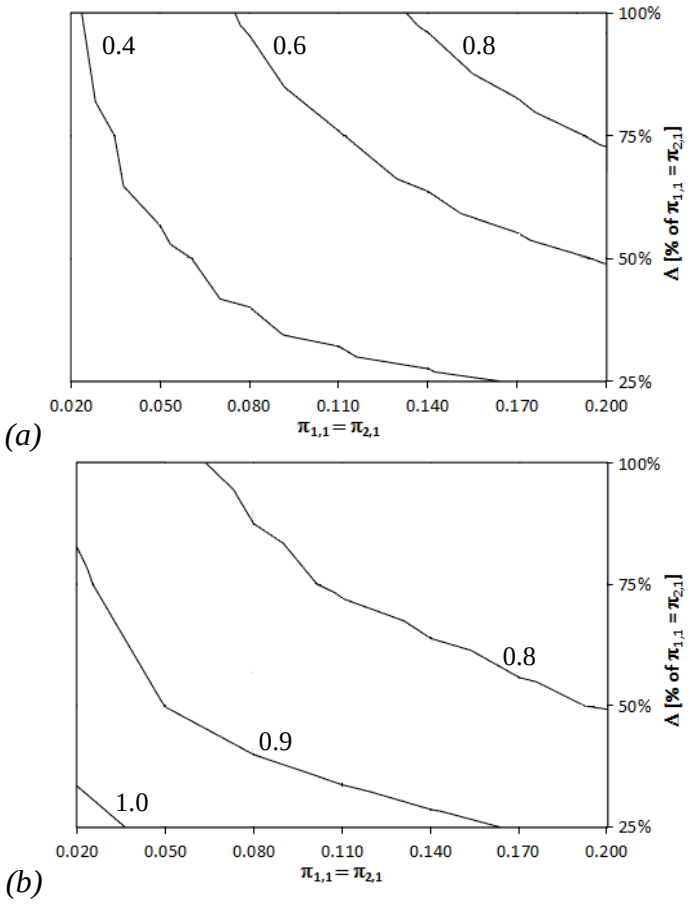


Figure 3-1. Contour plots of relative MSE vs. pass rates $\pi_{1,1} = \pi_{2,1}$ and size of step change: (a) relative MSE = MSE_{WEE} / $MSE_{naïve,\ \lambda \to 1}$, (b) relative MSE = MSE_{WEE} / $MSE_{naïve,\ \lambda \to 0}$

Figure 3-1 shows that the WEE estimator has lower MSE than either of the naïve estimators for most of the values of $\pi_{1,1} = \pi_{2,1}$ and Δ. The advantage of the WEE estimator over the estimator based on present time data only is more important when the change in the parameter is small and present time sample size is small. The advantage of the WEE estimator over the estimator based on all historical data weighted equally is more pronounced for larger changes in the parameter. We see that the WEE estimator provides a trade-off between bias and variance relative to the two naïve approaches for this simple example.

For this simple example, we also calculate variance by the distributions of the random variables $\{Y_{1t}, Y_{2t}, t = 1, \dots, T\}$. At the present time $T = 10$, the variance of estimator $\tilde{\pi}$ is

$$var_{dist}(\tilde{\pi}; \Delta) = var\left(\frac{\sum_{t=1}^{T} w_t Y_{2t}}{\sum_{t=1}^{T} w_t n_{2t}} - \frac{\sum_{t=1}^{T} w_t Y_{1t}}{\sum_{t=1}^{T} w_t n_{1t}}\right)$$
$$= \frac{\sum_{t=1}^{T} w_t^2 \pi_{1,1}(1-\pi_{1,1})}{n_1} + \frac{\sum_{t=1}^{5} w_t^2 \pi_{2,1}(1-\pi_{2,1}) + \sum_{t=6}^{10} w_t^2 (\pi_{2,1}+\Delta)(1-\pi_{2,1}-\Delta)}{n_2}$$

We compare the two variances by the ratio of standard deviations which we denote as $e(\tilde{\pi}, \Delta) = \frac{\sqrt{\widehat{var}_{WI}(\tilde{\pi};\Delta)}}{\sqrt{var_{dist}(\tilde{\pi};\Delta)}}$. Figure 3-2 gives a contour plot of the values of $e(\tilde{\pi}, \Delta)$ for the WEE estimator with weight parameter $\lambda = 0.1$.

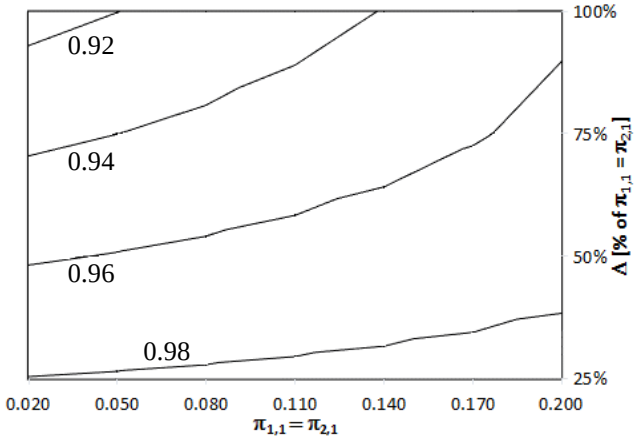


Figure 3-2. Contour plot of $e(\tilde{\pi}, \Delta)$ by $\pi_{1,1} = \pi_{2,1}$ and Δ

Figure 3-2 shows that $\widehat{var}_{WI}(\tilde{\pi}; \Delta)$ and the variance based on the distributions of $\{Y_{1t}, Y_{2t}, t = 1, \dots, T\}$ are close for these values of $\pi_{1,1} = \pi_{2,1}$ and Δ. We see that the weighted information variance using WEE estimates is a good estimate of variance for this simple example, especially when there is a small change in the parameter.

iv. the distribution of the weight-adjusted random variable, $\frac{\sum_{t=1}^{T} w_t n_t}{\sum_{t=1}^{T} w_t^2 n_t} \tilde{S}$, is approximately χ_r^2

At time t, consider a test of null hypothesis $H_0: \pi = 0$ versus the alternative hypothesis $H_A: \pi \neq 0$. We show by properties of the random variables that $E\left(\frac{\sum_{t=1}^{T} w_t n_t}{\sum_{t=1}^{T} w_t^2 n_t} \tilde{S}\right) = 1$ under the null hypothesis which agrees with the first moment of the distribution in (13). We validate the second and third moments and 95th percentile of the distribution in (13) through comparison to approximate distributions based on simulated data. Table 3-2 confirms that the approximate

distribution $\frac{\sum_{t=1}^{T} w_t n_t}{\sum_{t=1}^{T} w_t^2 n_t} \tilde{S} \overset{approx}{\sim} \chi_1^2$ is suitable when N is very large ($N = 1 \times 10^7$) and a useful approximation when N is small ($N = 100$). The details that follow in this section may be skipped by the reader.

We approximate a distribution for $\tilde{S}$ in order to test a hypothesis based on test statistic $\hat{S}$. The random variable $\tilde{S}$ in terms of random variables Y_{mt}, sample sizes n_{mt}, and weights w_t, $t = 1, \dots, T, m = 1,2$ is

$$\tilde{S} = 2 \sum_{m=1}^{2} \sum_{t=1}^{T} w_t \left(Y_{mt} \, log \frac{\tilde{\pi}_m}{\tilde{\pi}_{null}} + (n_{mt} - Y_{mt}) \, log \left(\frac{1-\tilde{\pi}_m}{1-\tilde{\pi}_{null}} \right) \right)$$

We approximate $\tilde{S}$ through second order Taylor Series approximations for those terms involving logarithms of the random variables:

- $\log(x)$ for $\sum_{t=1}^{T} w_t Y_{mt}$ around $\sum_{t=1}^{T} w_t \pi_m n_{mt}$ for $m = 1,2$
- $\log(x)$ for $\sum_{m=1}^{2} \sum_{t=1}^{T} w_t Y_{mt}$ around $\left(\frac{\pi_1 + \pi_2}{2} \right) \sum_{m=1}^{2} \sum_{t=1}^{T} w_t n_{mt}$

$$\begin{aligned}
\tilde{S} \approx\ & 2 \sum_{m=1}^{2} \left(\frac{\left(\sum_{t=1}^{T} w_t Y_{mt} \right)^2}{\pi_m \sum_{t=1}^{T} w_t n_{mt}} - (1 - log\, \pi_m) \sum_{t=1}^{T} w_t Y_{mt} - \frac{\sum_{t=1}^{T} w_t Y_{mt} \left(\sum_{t=1}^{T} w_t (\pi_m n_{mt} - Y_{mt}) \right)^2}{2\pi_m^2 \left(\sum_{t=1}^{T} w_t n_{mt} \right)^2} \right) \\
& -2 \sum_{m=1}^{2} \left(\frac{\sum_{t=1}^{T} w_t Y_{mt} \sum_{t=1}^{T} w_t (n_{mt} - Y_{mt})}{(1-\pi_m) \sum_{t=1}^{T} w_t n_{mt}} + \left(\frac{\pi_m}{1-\pi_m} + log(1-\pi_m) \right) \sum_{t=1}^{T} w_t (n_{mt} - Y_{mt}) \right) \\
& -2 \sum_{m=1}^{2} \frac{\sum_{t=1}^{T} w_t (n_{mt} - Y_{mt}) \left(\sum_{t=1}^{T} w_t (\pi_m n_{mt} - Y_{mt}) \right)^2}{2(1-\pi_m)^2 \left(\sum_{t=1}^{T} w_t n_{mt} \right)^2} + 2 \frac{\left(\sum_{m=1}^{2} \sum_{t=1}^{T} w_t Y_{mt} \right)^2}{\sum_{m=1}^{2} \frac{\pi_m}{2} \sum_{m=1}^{2} \sum_{t=1}^{T} w_t n_{mt}} \\
& -2 \left(1 - log \left(\sum_{m=1}^{2} \frac{\pi_m}{2} \right) \right) \sum_{m=1}^{2} \sum_{t=1}^{T} w_t Y_{mt} - 2 \frac{\sum_{m=1}^{2} \sum_{t=1}^{T} w_t Y_{mt} \sum_{m=1}^{2} \sum_{t=1}^{T} w_t (n_{mt} - Y_{mt})}{\left(\sum_{m=1}^{2} \frac{1-\pi_m}{2} \right) \sum_{m=1}^{2} \sum_{t=1}^{T} w_t n_{mt}} \\
& +2 \left(\frac{\sum_{m=1}^{2} \pi_m}{\sum_{m=1}^{2} 1-\pi_m} + log \left(\sum_{m=1}^{2} \frac{1-\pi_m}{2} \right) \right) \sum_{m=1}^{2} \sum_{t=1}^{T} w_t (n_{mt} - Y_{mt}) \\
& -2 \left(\frac{\sum_{m=1}^{2} \sum_{t=1}^{T} w_t (n_{mt} - Y_{mt})}{2 \left(\sum_{m=1}^{2} \frac{1-\pi_m}{2} \right)^2} + \frac{\sum_{m=1}^{2} \sum_{t=1}^{T} w_t Y_{mt}}{2 \left(\sum_{m=1}^{2} \frac{\pi_m}{2} \right)^2} \right) \frac{\left(\sum_{m=1}^{2} \sum_{t=1}^{T} w_t \left(\left(\sum_{m=1}^{2} \frac{\pi_m}{2} \right) n_{mt} - Y_{mt} \right) \right)^2}{\left(\sum_{m=1}^{2} \sum_{t=1}^{T} w_t n_{mt} \right)^2}
\end{aligned}$$

We find the expected value of the approximation for $\tilde{S}$ based on the assumptions $Y_{1t} \sim binomial(n_{1t}, \pi_1)$, $Y_{2t} \sim binomial(n_{2t}, \pi_2)$, and Y_{1t}, Y_{2t} independent for each t. Under the null hypothesis with $\pi = \pi_1 - \pi_2 = 0$, then $E[\tilde{S}] \approx \frac{\sum_{t=1}^{T} w_t^2 n_t}{\sum_{t=1}^{T} w_t n_t}$ and $E\left[\frac{\sum_{t=1}^{T} w_t n_t}{\sum_{t=1}^{T} w_t^2 n_t} \tilde{S} \right] \approx 1$.

We validate higher moments of the distribution of $\tilde{S}$ through simulation. We consider the empirical distribution of $\hat{S}$ for 100,000 datasets that are generated with $T = 10, \pi_1 = \pi_2 = 0.04$, $\lambda = 0.1$, and $n_{1t} = n_1, n_{2t} = n_2$ for all t. We repeat the simulation study for large $N = 1 \times 10^7$ and small $N = 100$. Table 3-2 gives the empirical moments of the distributions of $\hat{S}$.

Table 3-2. Moments of approximate distributions of weight-adjusted hypothesis test statistic

Approximate Distribution	Mean	Variance	Skew	95th percentile
$\frac{\sum_{t=1}^{T} w_t n_t}{\sum_{t=1}^{T} w_t^2 n_t} \tilde{S} \sim \chi_1^2$	1	2	2.828	3.841
simulated distribution with $N = 1 \times 10^7$	1.000	2.008	2.852	3.860
simulated distribution with $N = 100$	1.019	2.086	2.869	3.914

Table 3-2 shows that the empirical distributions based on simulations are close to the approximation distribution given in (13) for this simple example under the select simulation conditions. This simple example is convenient for demonstrating the properties of the WEE estimator and the approximations for the variance and hypothesis test statistic estimates. Similar demonstrations can be made over a wide class of models.

3.7 SAS routines

The weighted estimating equations corresponding to the estimating functions in (3) can be solved in most regression programs that allow for weights. In SAS, the weighted estimating equations can be solved using PROC GENMOD. Details on this procedure and other resources to use SAS are available at "Resources to help you learn and use SAS" (n.d.). Consider an example dataset called SAMPLE_DATA with one row for each subject that is observed. The dataset contains fields for an index 'case', covariate values 'x_1, x_2', $\{w_t\}$ 'weights', outcome 'y'. The parameter to estimate includes elements for the mean outcome for a baseline subject and two covariates effects, $\theta_T = (\alpha_T, \beta_{1,T}, \beta_{2,T})$. The SAS statements to estimate $\theta = \theta_T$ by the WEE approach assuming a binomial GLM with a logit link function for SAMPLE_DATA are given in Figure 3-3. This SAS PROC GENMOD routine also provides the weighted information estimate of the variance of $\hat{\theta}$ given in (7).

```
PROC GENMOD data=SAMPLE_DATA order=internal descending;
      class case;
      weight weights;
      MODEL y = x1 x2 / expected dist=binomial;
      repeated subject=case / type=ind ecovb;
      ods output GEEEmpPEst=theta_est GEERCov=covmatrix_est;
RUN;
```

Figure 3-3. SAS code for WEE analysis of an example dataset

The SAS GENMOD procedure is used in other applications to solve modifications of a GLM, such as a weighted response reflecting prior knowledge of varying dispersion among the data.

The SAS PROC GENMOD procedure also computes the WEE likelihood ratio test statistic for the null hypothesis H_0: $\delta_1 = \delta_2 = \cdots = \delta_6 = 0$ versus the alternative H_A: at least one element of $\delta_T \neq 0$. The dataset contains fields for indicator variables $m1, \ldots, m6$ to indicate the stream m where an observation is made. The SAS code to estimate the WEE LR test statistic in (12) assuming a binomial model of the response for dataset SAMPLE_DATA is given in Figure 3-4.

```
PROC GENMOD data=SAMPLE_DATA descending;
      weight weights;
      freq freq;
      MODEL y = Lab1 Lab2 Lab3 Lab4 Lab5 Lab6 / dist=binomial;
      contrast 'LR' Lab1 1, Lab2 1, Lab3 1, Lab4 1, Lab5 1, Lab6 1;
RUN;
```

Figure 3-4. SAS code for WEE estimate of test statistic for H_0 vs. H_A

The convenience of the existing software functionality for solving the weighted estimating equations and calculating the hypothesis test statistic makes it convenient to implement the WEE approach and update the estimates over time.

The discussion of the WEE approach in Chapter 3 give all of the aspects that are necessary for applying the approach to the motivating applications of this research. In Chapters 4, 5, and 6 we apply this approach to real, realistic, and simulated datasets in order to explore the impact that the WEE approach can have over current industry practices.

Chapter 4: Customer Loyalty Measure

One popular business management philosophy prioritizes actions for driving growth around improving customer loyalty (Reichheld and Markey, 2011). The measure known as Net Promoter Score (NPS) is commonly used to focus process and product improvements to drive customer loyalty and achieve business success. The estimate of this measure is based on customer responses to a survey asking the ultimate question, "On a scale of 0-10, how likely is it that you would recommend this company or product to a friend or colleague?" The customer's response classifies them into one of the three categories

i. detractors who respond six or below
ii. passives who respond seven or eight
iii. promoters who respond nine or ten

The quantity NPS is defined as the difference between the proportions of customers who are promoters and detractors. Increasing the proportion of customers who are promoters, decreasing the proportion of detractors, or doing both simultaneously increases the value of NPS. Publicly available information such as NPS Benchmarks (2014) shows that many diverse companies report NPS quantities as a measure of business performance. Efficient estimation of NPS is thus a topic of importance.

The management consulting firms, Bain & Company and Satmetrix, provide insights into best practises for shaping a business through driving observed NPS values to targets; however, little is written on analysis considerations. Markey, Reichheld, and Dullweber (2013) recommend, "You can analyse NPS by business, region, or any other subcategory, and you can track it from week to week to see how your loyalty-building efforts are working." Done this way, NPS estimates are based on observed proportions of promoters and detractors among respondents in various streams (e.g. region) by week. The current industry practice is a naïve estimate for NPS based on sample proportions among the most recent sample (Markey et al., 2013) Estimates by time period are then compared to benchmarks and targets and tracked in a trend chart over time as seen in the example for a telecommunications application in Figure 4-1 (Nowinski, 2009).

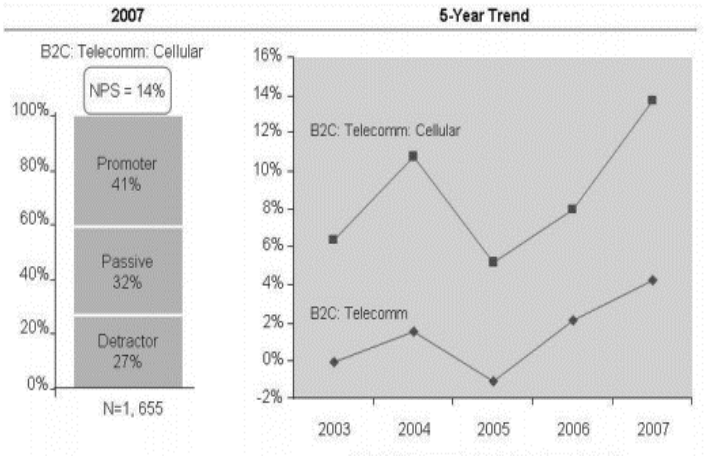


Figure 4-1. Net Promoter Score analysis for a telecommunications application

In a typical presentation such as Figure 4-1, little or no attention is paid to the impact of sample size, covariate effects, and changing populations over time. Depending on the survey design and fluctuations in response rates, small samples are likely in some time periods. We draw on data from multiple time periods to reduce uncertainty. In the common situation where performance drifts over time, a present time estimate that uses present and historical data is biased. We illustrate the application of the WEE approach to set up a bias/variance trade-off in the present time estimate of NPS with a smartphone customer loyalty example drawn from the author's experience. Further, we study a simulated dataset in order to compare the WEE approach to competing alternatives. We follow the notation and model for this application given in Table 1-1.

4.1. Smartphone Net Promoter Score

We study the estimates from the weighted estimating equations approach with a realistic customer loyalty dataset from weekly surveys by a smartphone vendor. We expect that overall customer loyalty drifts slowly from week to week in an unpredictable way due to the effect of improvement efforts and other factors not included in our dataset. As well, data are observed from different individuals among a changing customer population. In order to reliably compare estimates across time, we adjust the present time estimate of the parameter for the different covariate distributions among the samples. Further, we illustrate a test to compare estimates across levels of a covariate of interest.

Data

The data arising from customers' responses to the survey asking the ultimate question are described in Section 1.1. The dataset contains sample responses from 19,981 customers over 42 weeks. The number of customers responding by week over this period, $n_t, t = 1, \ldots, 42$, is given in Figure 4-2.

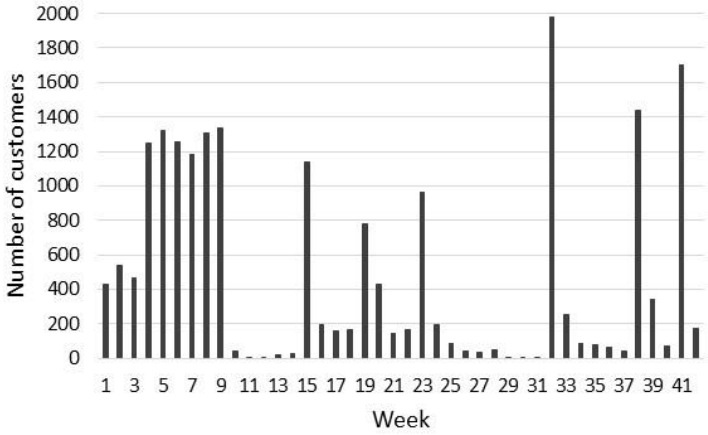


Figure 4-2. Smartphone customer loyalty dataset: sample size by week

Figure 4-2 shows that the number of customer responses by week varies considerably. There are as few as 4 customer responses and as many as 2000 responses in one week. There are 175 customer responses in the current week. For each sample, we observe the categorized customer response to the ultimate question taking a value from $y = \{1\ (\text{detractor}), 2\ (\text{passive}), 3\ (\text{promoter})\}$. In addition to the response, we also observe two covariate values for each customer: their product variant and the amount of time since their purchase of the product (tenure). The indicator variables x_1, x_2, x_3 describe the product variant and the interval variable x_4 describes the tenure. The levels of these variables and the baseline level of the covariates are described in Section 1.4.

Estimation by weighted estimating equations

We use the WEE approach since some sample sizes are small and we expect that the mean proportions of detractors and promoters among customers having some fixed values of the covariates may drift over time in an unpredictable way. In Section 1.3, we list two possible standard populations of interest for this application. The field population refers to the actual values of the covariates among all current customers. These are known for a smartphone vendor with

access to sales and service records. For the example under study having $10{,}000$ present customers, the distribution of their covariate values is given in Table 4-1.

Table 4-1. Field population distribution for 10,000 customers

		Tenure [months]					
		[0-2]	**[2-6]**	**[6-12]**	**[12-18]**	**[18-24]**	**[24+]**
Product variant	**1**	5	21	95	92	289	1071
	2	20	67	353	490	557	743
	3	64	228	1188	931	522	227
	4	524	1379	1133	1	0	0

With the quantities in Table 4-1 we define the standard population $\left\{x_{j^*} = \left(x_{1,j^*}, \dots, x_{4,j^*}\right)^T \text{ for } j^* = 1, \dots, 10{,}000\right\}$.

Table 1-1 introduces the GLM that is selected for this problem based on π_1 and π_3 as the multinomial proportions of responses 1 (detractor) and 3 (promoter), respectively. We assume that the random variables are independent across $t = 1, \dots, T$, conditional on the values of the covariates. The mean level of the two proportions at the baseline level of the covariates are modelled by α_1 and α_2, respectively. The effects of the three product variants relative to the baseline value are modelled by $\beta_1, \beta_2, \beta_3$ and tenure is modelled by β_4. For customer j observed at time t having covariate value x_{jt}, the multinomial proportions relate to the model parameter $\theta_t = \left(\alpha_{1,t}, \alpha_{2,t}, \beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \beta_{4,t}\right)^T$ through the inverse link functions

$$\begin{aligned} \pi_{1,jt}(\theta_t; d_t) &= \frac{\exp\left(\alpha_{1,t}+\beta_{1,t}x_{1,jt}+\beta_{2,t}x_{2,jt}+\beta_{3,t}x_{3,jt}+\beta_{4,t}x_{4,jt}\right)}{1+\exp\left(\alpha_{1,t}+\beta_{1,t}x_{1,jt}+\beta_{2,t}x_{2,jt}+\beta_{3,t}x_{3,jt}+\beta_{4,t}x_{4,jt}\right)} \\ \pi_{3,jt}(\theta_t; d_t) &= \frac{1}{1+\exp\left(\alpha_{2,t}+\beta_{1,t}x_{1,jt}+\beta_{2,t}x_{2,jt}+\beta_{3,t}x_{3,jt}+\beta_{4,t}x_{4,jt}\right)} \end{aligned} \tag{14}$$

The proportional odds property assumes that the effect of the covariates is identical for the two logits. In this example, we assume that the covariate effects $\beta_t = \left(\beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \beta_{4,t}\right)^T$ are fixed over time, but one or both elements of $\alpha_t = \left(\alpha_{1,t}, \alpha_{2,t}\right)^T$ may change slowly due to the influence of other factors affecting customer loyalty over time. The log-likelihood function describing the probability of data $d_t = \left\{\left(x_{jt}, y_{jt}\right), j = 1, \dots, n_t\right\}$ including all observations at time period t is

$$l_t(\theta_t; d_t) = \sum_{j=1}^{n_t} I\left[y_{jt} = 1\right] \log \pi_{1,jt} + I\left[y_{jt} = 2\right] \log\left(1 - \pi_{1,jt} - \pi_{3,jt}\right) + I\left[y_{jt} = 3\right] \log \pi_{3,jt} \tag{15}$$

for indicator variables $I[y_{jt}=1], I[y_{jt}=2]$, and $I[y_{jt}=3]$.

We select weights $\{w_t, t=1,\dots,T\}$ by (10) with weight parameter value $\lambda = 0.1$. The WEE estimates are compared to the estimates of the naïve approach using the two special cases of the weights described in Section 3.1.

Under (3), the weighted estimating function vector of length 6 is

$$Q(\theta; d, w) = \sum_{t=1}^{T} w_t \psi_t(\theta; d_t)$$

$$= \begin{bmatrix} \sum_{t=1}^{T} w_t \sum_{j=1}^{n_t} I\left[y_{jt}=1\right]\left(1-\pi_{3,jt}\right) + \frac{\pi_{3,jt}\left(1-\pi_{3,jt}\right)-\pi_{1,jt}\left(1-\pi_{1,jt}\right)}{1-\pi_{1,jt}-\pi_{3,jt}} - I\left[y_{jt}=3\right]\left(1-\pi_{1,jt}\right) \\ \sum_{t=1}^{T} w_t \sum_{j=1}^{n_t} x_{1,jt}\left(I\left[y_{jt}=1\right]\left(1-\pi_{3,jt}\right) + \frac{\pi_{3,jt}\left(1-\pi_{3,jt}\right)-\pi_{1,jt}\left(1-\pi_{1,jt}\right)}{1-\pi_{1,jt}-\pi_{3,jt}} - I\left[y_{jt}=3\right]\left(1-\pi_{1,jt}\right)\right) \\ \sum_{t=1}^{T} w_t \sum_{j=1}^{n_t} x_{2,jt}\left(I\left[y_{jt}=1\right]\left(1-\pi_{3,jt}\right) + \frac{\pi_{3,jt}\left(1-\pi_{3,jt}\right)-\pi_{1,jt}\left(1-\pi_{1,jt}\right)}{1-\pi_{1,jt}-\pi_{3,jt}} - I\left[y_{jt}=3\right]\left(1-\pi_{1,jt}\right)\right) \\ \sum_{t=1}^{T} w_t \sum_{j=1}^{n_t} x_{3,jt}\left(I\left[y_{jt}=1\right]\left(1-\pi_{3,jt}\right) + \frac{\pi_{3,jt}\left(1-\pi_{3,jt}\right)-\pi_{1,jt}\left(1-\pi_{1,jt}\right)}{1-\pi_{1,jt}-\pi_{3,jt}} - I\left[y_{jt}=3\right]\left(1-\pi_{1,jt}\right)\right) \\ \sum_{t=1}^{T} w_t \sum_{j=1}^{n_t} x_{4,jt}\left(I\left[y_{jt}=1\right]\left(1-\pi_{3,jt}\right) + \frac{\pi_{3,jt}\left(1-\pi_{3,jt}\right)-\pi_{1,jt}\left(1-\pi_{1,jt}\right)}{1-\pi_{1,jt}-\pi_{3,jt}} - I\left[y_{jt}=3\right]\left(1-\pi_{1,jt}\right)\right) \\ \sum_{t=1}^{T} w_t \sum_{j=1}^{n_t}\left(1-I\left[y_{jt}=1\right]\right)\frac{\pi_{3,jt}\left(1-\pi_{3,jt}\right)}{1-\pi_{1,jt}-\pi_{3,jt}} - I\left[y_{jt}=3\right]\left(1+\frac{\pi_{1,jt}\pi_{3,jt}}{1-\pi_{1,jt}-\pi_{3,jt}}\right) \end{bmatrix} \quad (16)$$

given the present time value of the parameter, $\theta = \theta_T = (\alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3, \beta_4)^T$, data $d = \{d_t, t=1,\dots,T\}$, weights $w = \{w_t, t=1,\dots,T\}$, and inverse link functions $\pi_1(\theta_t; d_t)$ and $\pi_3(\theta_t; d_t)$.

The WEE estimate $\hat{\theta}$ is the solution of $Q(\hat{\theta}; d, w) = 0$. Through (11), we estimate the weighted information estimate of variance, $\widehat{var}(\tilde{\theta}; \hat{\theta})$, involving $I_t(\hat{\theta})$ which is the expected information function at each time period evaluated at the WEE estimate. With estimate $\hat{\theta}$, we use Table 2-1 to compute the proportions estimates $\hat{\pi}_{1,j^*}$ and $\hat{\pi}_{3,j^*}$ for each of the standard population customers $j^* = \{1,\dots,10{,}000\}$ given $\{x_{j^*}\}$ and to compute the estimates $\hat{\pi}_1$ and $\hat{\pi}_3$ for the entire standard population. Then, the estimate of NPS for the standard population at time T is $\widehat{NPS} = \hat{\pi}_3 - \hat{\pi}_1$. Similarly, with estimate $\widehat{var}(\hat{\theta})$, we compute estimates $\widehat{var}(\hat{\pi}_{1,j^*})$ and $\widehat{var}(\hat{\pi}_{3,j^*})$ through (5) and estimates $\widehat{var}(\hat{\pi}_1)$ and $\widehat{var}(\hat{\pi}_3)$ through (4). Additionally, we require $\widehat{covar}(\hat{\pi}_{1,j^*}, \hat{\pi}_{3,j^*})$ which we compute through the multivariate delta method (Casella and Berger, 2002) in order to get estimates $\widehat{covar}(\hat{\pi}_1, \hat{\pi}_3)$ and $\widehat{var}(\widehat{NPS})$.

Results

We compare the WEE estimate for NPS to those by the naïve and the EWMA approaches. For the two naïve approaches, estimates $\hat{\pi}_{1,j^*}$ and $\hat{\pi}_{3,j^*}$ and estimates of their variances are calculated through the WEE approach with one of the limiting values of the weight parameter. For the EWMA approach, estimates $\hat{\pi}_{1,j^*t}$ and $\hat{\pi}_{3,j^*t}$ are MLE estimates based on data at each time period and combined across time with weights. Estimates of the variances of $\hat{\pi}_{1,j^*}$ and $\hat{\pi}_{3,j^*}$ are calculated from the estimates of the variances of each $\hat{\pi}_{1,j^*t}$ and $\hat{\pi}_{3,j^*t}, t = 1, \dots, T$.

The naïve approach commonly used in practice (Markey et al., 2013) estimates NPS based on present time data only without attention to the values of the covariates among customers in the sample. Here, estimates $\hat{\pi}_1$ and $\hat{\pi}_3$ are sample proportions. Clearly, these estimates do not account for a changing customer population. Since covariate levels are discrete in this application, then a better non-parametric approach involving the standard population is to estimate π_{1,j^*} and π_{3,j^*} based on observations among customers having the same covariate values as standard population subject j^*. Variances of the sample proportion estimates are estimated by usual methods. The estimate of variance is large when there are few observations having covariate value x_{j^*} and the approach is infeasible when no similar customers are observed. The other naïve approach involves sample proportions estimates $\hat{\pi}_{1,j^*}$ and $\hat{\pi}_{3.j^*}$ based on the aggregate of historical data weighted equally. Here, the number of observations having covariate value x_{j^*} is larger. The non-parametric EWMA approach involves sample proportion estimates π_{1,j^*t} and π_{3,j^*t} across time periods. The various approaches are summarized in Table 2-1.

With estimates $\hat{\pi}_{1,j^*}$ and $\hat{\pi}_{3.j^*}$ by a non-parametric or a GLM-based approach, an estimate of NPS is possible for any standard population. We study estimates for the field population of 10,000 customers at week 42 defined in Table 4-1. Figure 4-3 shows estimates $\widehat{NPS}$ and the corresponding 95% confidence interval based on $\widehat{var}(\widehat{NPS})$ assuming normality.

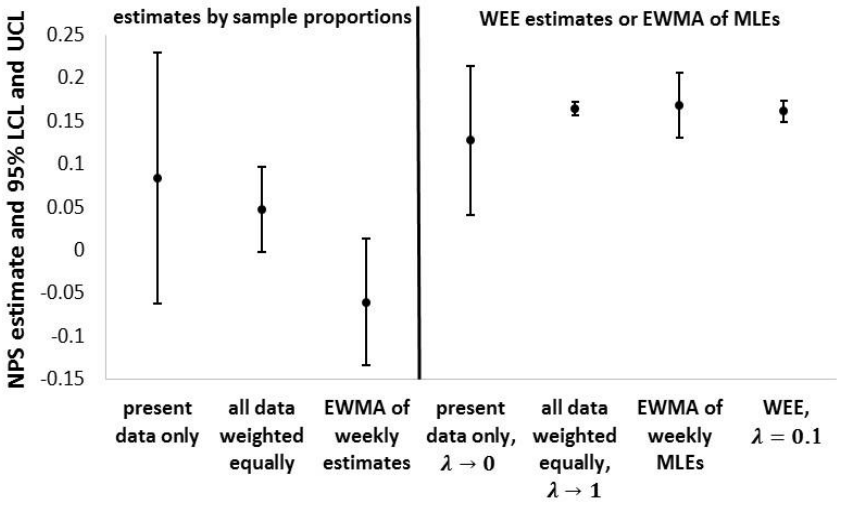


Figure 4-3. Estimates of field population NPS by various approaches

Figure 4-3 shows that the estimate by the recommended WEE approach ($\lambda = 0.1$) has less uncertainty than those estimates using present time data only. Its uncertainty is comparable to those of the other two estimates by the WEE formulation that use all historical data. There are some differences between the estimates by the various approaches, but we are unable to assess bias since the true value is unknown. The advantage of the recommended WEE approach over the other approaches depends on the sample sizes and the change in the parameter over time. In Section 4.2, we study the bias and variance of the GLM-based estimates through simulation.

Decision makers track the NPS esimates over time to regularly assess and plan improvement activities. The author has seen an NPS business intelligence dashboard designed with filters to allow decision makers to subdivide and summarize the data by their selection of time period. Changing populations, changing overall customer loyalty over time, and small samples sizes have a detrimental effect on data viewed in this way. To demonstrate, in Figure 4-4 we compare the trends in the field population estimates between the common naïve approach involving sample proportions based on present time data only and the WEE approach. Note that there is a difference in the scales of the two vertical axes.

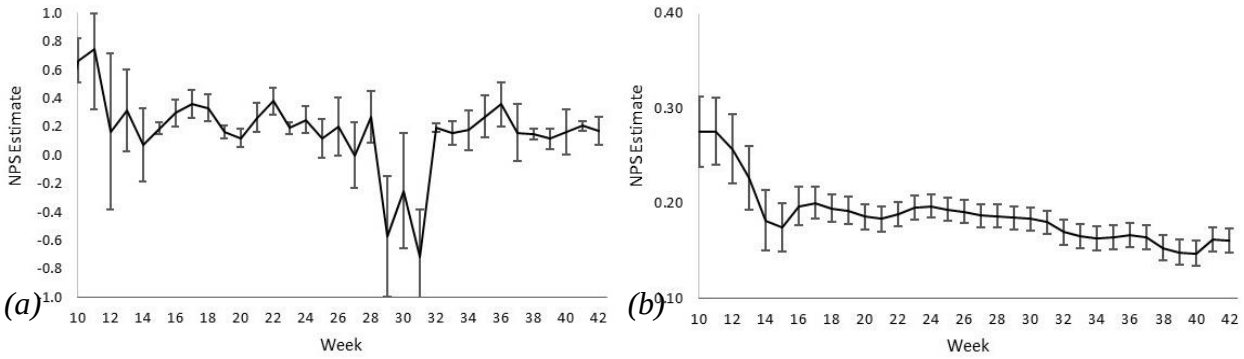


Figure 4-4. Trends in field population NPS estimates: (a) sample proportion estimates using present data only, (b) WEE approach, $\lambda = 0.1$

Figure 4-4 shows a vast difference in the trend of NPS estimates by the two approaches over time. The population-adjusted estimates by the WEE approach are much more precise and show a trend that is not apparent on the other graph. The WEE approach can have an important impact on the decisions taken by decision makers to drive loyalty and growth though a trade-off in bias and variability in population-adjusted NPS estimates and reliable comparisons across time.

In this application, a decision maker may want to compare NPS estimates across subgroups of the customer population. For example, superior results for a particular product variant may encourage decision makers to target sales of this variant or focus efforts to bring the NPS of other product variants to comparable levels. We consider the test of the hypothesis that NPS for customers with product variant 4 is the same as NPS for customers with product variant 3. In terms of the parameters, we test the null hypothesis $H_0: \beta_3 - \beta_2 = 0$ versus the alternative $H_A: \beta_3 - \beta_2 \neq 0$. The WEE estimate and relevant quantities to test H_0 versus H_A are given in Table 4-2.

Table 4-2. WEE hypothesis test quantities for $H_0: \beta_3 - \beta_2 = 0$ vs. $H_A: \beta_3 - \beta_2 \neq 0$

Unconstrained model	**WEE estimate of $\boldsymbol{\theta}$**	$\hat{\theta} = (\text{-}0.695, 0.380, \text{-}0.0928, \text{-}0.147, \text{-}0.414, \text{-}5.84\ \text{E-}6)^T$
	Weighted log likelihood	$\sum_{t=1}^{T} w_t l_t(\hat{\theta}; d_t) = \text{-}486.218$
Constrained model	**WEE estimate of $\boldsymbol{\theta}$**	$\hat{\theta}_0 = (\text{-}0.899, 0.173, \text{-}0.0514, \text{-}0.144, \text{-}0.144, 0.0100)^T$
	Weighted log likelihood	$\sum_{t=1}^{T} w_t l_t(\hat{\theta}_0; d_t) = \text{-}486.754$
WEE LR test statistic (12)		$\hat{S} = 1.072$
Weight-adjusted test statistic		$\frac{\sum_{t=1}^{T} w_t n_t}{\sum_{t=1}^{T} w_t^2 n_t} \hat{S} = 17.2$
p-value for H_0 under (13)		$\Pr\left(\chi_1^2 > \frac{\sum_{t=1}^{T} w_t n_t}{\sum_{t=1}^{T} w_t^2 n_t}\ \hat{S}\right) < 0.01$

Table 4-2 gives evidence to reject the null hypothesis $H_0: \beta_3 - \beta_2 = 0$ in favour of $H_A: \beta_3 - \beta_2 \neq 0$ for a size 0.05 test. The estimates of the proportions are $\hat{\pi}_{1,\text{variant=3}} = 0.301$, $\hat{\pi}_{3,\text{variant=3}} = 0.442$, $\hat{\pi}_{1,\text{variant=4}} = 0.248$, and $\hat{\pi}_{3,\text{variant=4}} = 0.509$. Then, the estimates of NPS for customers at the baseline level of tenure ($x_4 = 0$) with the two model variants are $\widehat{NPS}_{\text{variant=3}} = 0.14$ and $\widehat{NPS}_{\text{variant=4}} = 0.26$. Decision makers have evidence that NPS of product variant 4 is superior to that of product variant 3.

A decision maker may track the estimate of the difference between NPS values of the two product variants over time in order to monitor the similarity of the two streams. The graph of $\widehat{NPS}_{\text{variant=4}} - \widehat{NPS}_{\text{variant=3}}$ based on data over the range $T = 10, \ldots, 42$ is shown in Figure 4-5. The 95% confidence interval of each estimate $\widehat{NPS}_{\text{variant=4}} - \widehat{NPS}_{\text{variant=3}}$ is based on the WI estimate of variance for $\hat{\theta}$ at that point in time assuming normality. The dotted line shows the p-value for H_0 under (13) at each point in time.

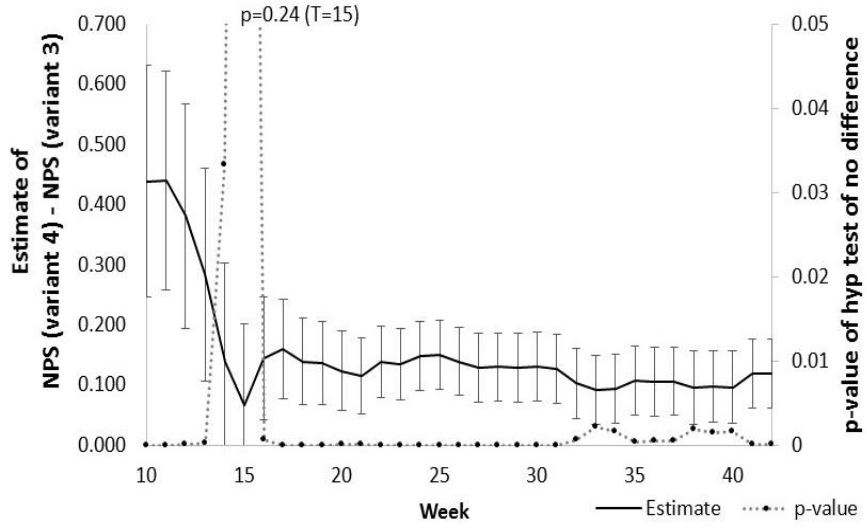


Figure 4-5. WEE estimates of difference in NPS for product variants 3 and 4

Figure 4-5 shows that the estimate of NPS for product variant 4 is consistently larger than that of product variant 3 and there is evidence to reject the size 0.05 test of no difference between the two at all points in time except $T = 15$. The earliest customers using product variant 4 are observed in week 10 and so the uncertainty of $\widehat{NPS}_{\text{variant=4}} - \widehat{NPS}_{\text{variant=3}}$ decreases after week 10 as more data on this variant are observed. There is a decrease in the estimate of the difference in NPS of the two product variants from week 10 to week 14. The difference between the two product variants is stable from week 15 to week 42. Alternatively, we could monitor the similarity between the mean performance at the two covariate levels through a graph of the weighted WEE LR test statistic over time.

4.2. Simulation study

We simulate data that is similar to the customer loyalty dataset to study the bias and variance of NPS estimates by the various approaches. We simulate data from four profiles of change in field population NPS over the 42 time periods. For each profile, the design value of the population NPS at time t, referred to as NPS_t, starts at $NPS_1 = -0.05882$ and either stays constant or changes linearly over the time span. The rate of change and design values of field population NPS for the four profiles are given in Table 4-3.

Table 4-3. Field population NPS design values, 4 profiles

Design profile	$NPS_{t+1} - NPS_t$	NPS_{42}	$NPS_{42} - NPS_1$
I	-2.439E-3	-0.1588	-0.1
II	0	-0.05882	0
III	2.439E-3	0.04118	0.1
IV	4.478E-3	0.1412	0.2

We fix the values of the five parameters,
$(\alpha_{2,t}, \beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \beta_{4,t}) = (0.6, 0.0195,\ -0.0355, -0.269, 0.0149)$ over $t = 1, \dots, 42$. The design values for NPS_1 and $(\alpha_{2,t}, \beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \beta_{4,t})$ are quite precise but have no particular significance. We calculate the design value for the $\{\alpha_{1,t,}\ t = 1, \dots, 42\}$ using these values, calculations of $\pi_{1,t}$ and $\pi_{3,t}$ in (14), and the field population distribution in Table 4-1.

For each simulated dataset, sample sizes by time period $\{n_t\}$ remain fixed at the values in Figure 4-2. The covariate vector x_{jt} for each sample is allocated by a multinomial distribution having the same proportions in each covariate group as the field population distribution at t. As in the example dataset, product variant 4 is introduced into the field population at time period $t = 10$. The response to the ultimate question, y_{jt}, is simulated for each sample by the multinomial distribution with the design probabilities $(\pi_{1,jt}, 1 - \pi_{1,jt} - \pi_{3,jt}, \pi_{3,jt})$. For each of the four NPS design profiles, we simulate 5000 datasets of d_t across 42 time periods.

We estimate NPS from $\{d_t, t = 1, \dots, 42\}$ by the GLM-based approaches in Table 2-1 since we want to model covariate effects and expect sparse data. We add analysis by the sample proportions approach using only present time data to highlight the advantages of a GLM-based approach. We refer to the five approaches as follows:

- estimates by sample proportions using d_T only, "Prop t=T"
- estimates by GLM using d_T only, "GLM t=T"
- estimates by GLM using all historical data with no weighting, "GLM t≤T"
- estimates by EWMA of weekly GLM estimates, "GLM EWMA"
- estimates by weighted estimating equations using all historical data, "WEE"

The weights required for the GLM EWMA and WEE methods are selected as outlined in Section 3.1 with weight parameter $\lambda = 0.1$. Following estimation of $\theta = \theta_{42}$ for each dataset by each approach, we compute the estimate of NPS_{42} for the present field population. In Figure 4-6, we give boxplots for the 5000 estimates of NPS_{42} and confidence interval estimates assuming

normality for each approach and design profile. The horizontal dotted lines show the design value of NPS at $T = 42$. The difference between the estimate and design values of NPS is a measure of bias.

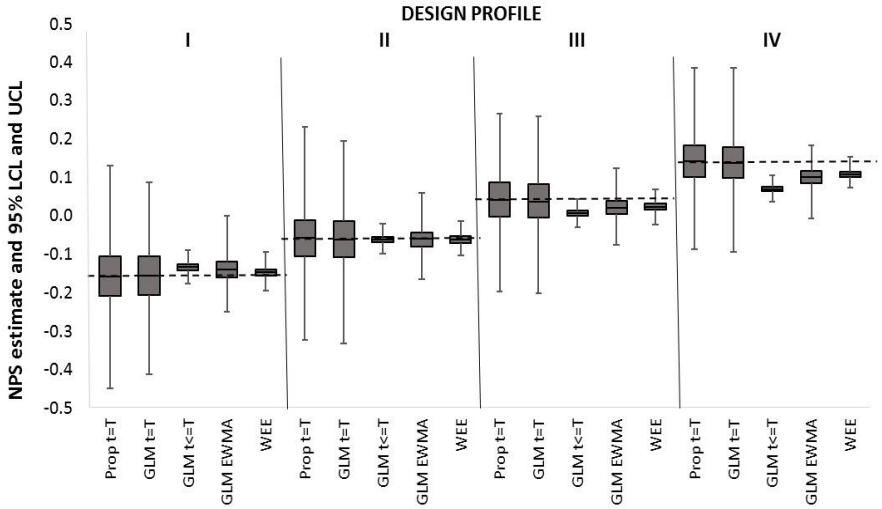


Figure 4-6. Estimates of NPS at T=42 by design profile and various approaches

Figure 4-6 shows that the approaches that depend only on the data at the present time, Prop t=T and GLM t=T, give estimates with the least amount of bias and these biases fluctuate very little with the size of the NPS change. The approaches that use all historical data, GLM t≤T, GLM EWMA, and WEE, add bias when the performance is changing over time relative to the size of the change. The difference between bias from the GLM t≤T and WEE approaches shows that down-weighting the estimating equation contributions of data from the further past reduces bias over using the historical data without weights. Further, the uncertainty in the estimates differs between those based on present time data and those based on all historical data. The variations in the WEE and GLM t≤T estimates are similar for each design profile and noticeably smaller than variations of the estimates using present time data only. Uncertainties of the GLM-based EWMA estimates are more than twice as large as those based on the other two approaches that use all historical data.

Figure 4-7 combines the bias and standard deviation of the 5000 estimates of NPS into the root mean squared error (MSE) for each design profile and analysis approach.

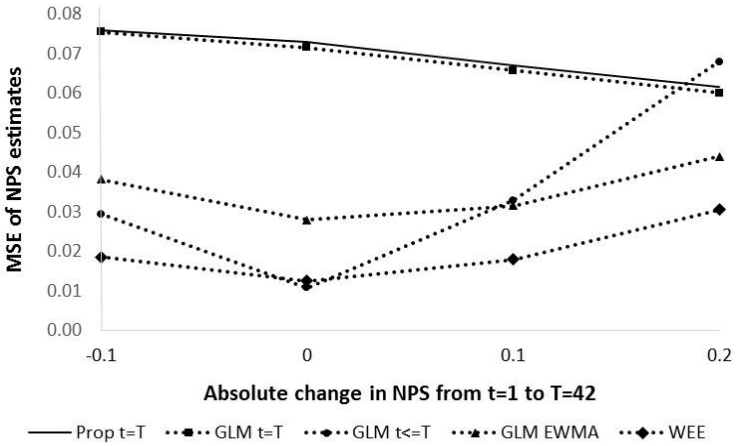


Figure 4-7. Root mean squared error of estimates over design profiles by approach

Figure 4-7 shows that the estimates from the WEE approach have the smallest MSE when θ_t (and hence NPS) changes over time. If θ_t is not changing over time, then the best approach is to fit a GLM to all the data without weights. The approaches that use data from all time periods are more efficient than approaches that use present time data only with the exception of the GLM t≤T approach for the largest design change. Because it regulates bias, the WEE approach has added efficiency over the GLM t≤T approach as the size of the performance change increases. Approaches based on present time data only suffer due to lack of precision in the estimates and their MSE values show that the reduced bias does not make up for the lack of precision relative to the other approaches. Significantly more uncertainty results from using present time data only or from using EWMA with all historical data, but there is only a modest increase in uncertainty when using the WEE approach. In summary, the WEE is the most efficient approach for estimates of the parameter when the parameter changes slowly over time.

The MSE values in Figure 4-7 use the standard deviation of each group of 5000 estimates as the estimate of standard error for an NPS estimate for each simulated dataset by approach. The weighted information (WI) estimate of variance in (11) can be used to estimate the standard error of the NPS estimate. We study the suitability of the WI estimate of variance by comparing this estimate to the standard deviation of each simulated group of 5000 estimates. The plot in Figure 4-8 gives the distribution of WI estimates of the standard deviation of NPS by the WEE approach relative to the observed standard deviation of the group of 5000 NPS estimates for each design profile. The dotted lines represent the observed standard deviation of the 5000 NPS estimates based on the WEE approach and the box and whiskers show the distribution of the 5000 estimates of the standard deviation based on the WI estimate of variance of the parameters.

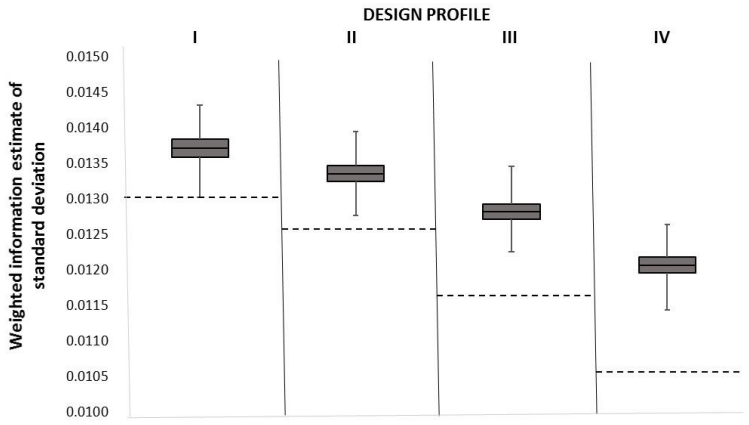


Figure 4-8. Distribution of weighted information estimates of standard deviation relative to observed values

Figure 4-8 shows that the weighted information estimate is a biased estimate of the standard deviation of the NPS estimate. The size of the bias is less than 2% of the true NPS value at week 42. In this simulation study, the weighted information estimate of variance is a reasonable approximation to the true variance of the estimate of NPS.

4.3. Summary and discussion

In a Harvard Business Review article, Frank Reichheld claims that NPS is the "one number you need to grow" (Riechheld, 2003 p.54) for business success. Since that time, surveys asking the ultimate question are commonplace and "as academics debate the details, managers are putting the [Net Promoter] scores into practice" (McGregor, 2006 p. 94). The only documented approach to estimate the NPS measure from a periodic stream of customer data collected over time is a naïve estimate based on sample proportions. Little or no attention is paid to sample size, covariate effects, and changing populations over time.

We analyse a realistic set of customer responses to the ultimate question from a smartphone vendor. We use the weighted estimating equations approach since we expect that mean NPS for a set of customers with fixed values of the covariates may drift slowly over time in an unpredictable way and some sample sizes may be small. We compare the WEE estimate of NPS to estimates by naïve approaches based on present time data only and all historical data weighted equally and the EWMA approach. The various approaches produce NPS estimates that vary considerably from one another. The WEE estimate has similar precision to the estimate by the GLM based on all historical data weighted equally and to the estimate by the EWMA of the weekly MLE estimates.

There is a vast difference in the trend of population-adjusted NPS estimates over time from the WEE approach compared to current industry practice.

We explore the bias and variance of estimates by the various approaches through simulated data designed with varying changes in the model parameter over time. The approaches that use all historical data add bias when θ_t (and NPS) changes over time relative to the size of the change. We see that down-weighting the estimating equation contributions of data from the further past through the WEE approach reduces bias over using the historical data without weights. The estimates by the WEE approach have the lowest mean squared error among the approaches under consideration when NPS changes slowly over time. We validate the usefulness of the weighted information estimate of variance through the simulated data. The work of this chapter indicates that the WEE approach could have an important effect on a manager's ability to drive business growth based on estimates of the NPS customer loyalty measure.

Selecting covariates

As in regression analysis, the selection of covariates to include in the analysis is important for the estimates and interpretation of results. Best practices on variable selection for model parsimony, numerical stability, and generalizability of the results from regression analysis (Bursac et al. 2008) should be followed. The analysis of the customer loyalty example data indicates that business decisions to increase the relative size of the customer base towards product variant 4 are expected to improve future performance. Not all covariates (e.g., tenure) can be influenced among the customer population but should be included if they explain significant variation in sample responses.

Comparison of the WEE and EWMA approaches

In general, the argument in favour of the WEE approach over a naïve or EWMA approach is not uniformly conclusive. Comparing the performance of the various approaches through a simulation study is not conclusive since there are many parameter values, covariate values, sample sizes, and ways that the parameter might change over time. We suspect that the mean squared error of the estimate is not uniformly lower for one approach relative to the other.

We consider a qualitative comparison of the WEE and EWMA approaches. The key difference between the two approaches is the order of the weighting and estimating operations. Estimates by an EWMA approach are based on data at each time period separately and combined with weights, whereas the WEE approach weights contributions to the estimating functions and estimates parameters involving data across all time periods. Under an EWMA approach, covariate effects

are re-estimated at each time period even though these effects may not change or change slowly over time. Uncertainties in the estimates by time period add to the uncertainty of the present time estimate and so, as we have seen in Sections 3.2 and 3.6, a small sample size at any time period has a negative effect on the precision of the present time estimate. Parameter estimates by time period are not sufficient statistics and so information may be lost in a present time estimate that combines estimates by time period as under the EWMA approach. Under the WEE approach, covariate effects estimates are based on all observations and so uncertainties in these estimates are smaller. The score functions that contribute to the weighted estimating function are each sufficient statistics summarizing data by time period and so the present time estimate uses all the information in the data. Both EWMA and WEE approaches require the solution of estimating equations with p unknowns in the present time period. For the customer loyalty application involving discrete-valued covariates, there needs to be an instance of each level of the covariates in the sample in order to estimate the related covariate effect. In the realistic dataset under consideration, there are insufficient data to estimate all of the covariate effects in roughly half of the time periods if we base estimates on data from each time period separately. Here, a standard implementation of the EWMA approach is not possible.

Based on the quantitative comparison under the particular conditions of this application and the preceding qualitative comparison, we prefer the WEE approach over the EWMA approach for down-weighting the influence of historical data in an estimate of present performance. The remainder of the thesis focuses on the important comparison of the WEE approach to the naïve approaches that are common in industry which combine data across time periods without weights or use present time data only.

Chapter 5: Lab Positive Abnormal Rate

In the United States, the Centers for Medicare and Medicaid Services regulate all laboratory testing performed on humans through the Clinical Laboratory Improvement Amendments (CLIA). In Ontario, the Institute for Quality Management in Healthcare (IQMH) is an independent agency with a provincial mandate to assess the ability of laboratories to perform medical testing. To equip medical professionals with quality data for decisions impacting patient health, the mission of the regulatory agencies is to provide rigorous, objective, third-party evaluation of the medical diagnostic testing systems according to international standards. Various laboratories may be performing the same test; however, differences between test methodologies, instrumentation, and operations can contribute to measurable differences between observed responses across the various labs. Proficiency testing is the term used by the CLIA relating to regular assessment of a laboratory's ability to provide an acceptable standard of service by comparison with peers. For a non-destructive test, one approach to proficiency testing may compare test results conducted on a single reference population at various laboratories. Here, a single set of subjects is selected and tested at each of the laboratories and the test measurements or discrete test outcomes are compared directly across peers. Challenges with this approach to proficiency testing include how to select the single set of subjects and the cost and logistics to circulate the samples across labs without degrading or destroying the samples. Additionally, for a test that has a binary outcome, a large sample is required to detect small but important differences between populations, adding cost and logistical difficulty. An alternative approach is to base proficiency tests on data observed from regular operation of the labs. Using data from regular operation of the labs avoids the challenge of selecting a single sample as well as the cost and logistics to transport samples among labs; however, having adequate sample size for the analysis is still a concern. Further, the number of patients tested at various labs may vary widely and so varying precision in the results by lab needs to be considered.

5.1. Fecal occult blood test positive abnormal rate

The application under study relates to proficiency testing of laboratories testing for indications of colorectal cancer. In Ontario, the Colon Cancer Check program was initiated in 2008 as the first population-based, province-wide, organized screening program designed to raise screening rates and reduce deaths from colorectal cancer. Those individuals who are deemed to be at risk for

developing colorectal cancer are encouraged to have a fetal occult blood test (FOBT) every two years. A kit is provided to the patient who draws the FOBT sample at home and sends their sample to a lab for testing. At the lab, a technician tests the sample and assigns a positive or negative abnormal result which informs the medical professional whether or not to conduct further testing. Studies show that screening with a FOBT every two years reduced death from colorectal cancer by 16 per cent over a decade (Cancer Care Ontario, 2008). There are seven licensed community medical laboratories providing FOBT testing services in Ontario. Unlike most other diagnostic tests, oversight of the proficiency testing of the seven labs testing FOBT samples is assigned to a committee comprised exclusively of laboratory representatives.

This research highlights shortcomings with the approach to proficiency testing of the labs carried out by the committee responsible for overseeing the laboratories testing FOBT samples in Ontario and suggests a more effective approach. The approach used by the committee as of May 2014 which we refer to as the "Ontario FOBT proficiency test" is as follows. Monthly, each of the seven labs report their positive abnormal rate which is calculated as the number of samples tested with a positive abnormal result relative to the total number of samples tested. The monthly positive abnormal rate for each lab is compared to an acceptance interval and a rate outside this interval indicates that the lab is in potential non-compliance. Three consecutive months of this status prompts a letter of concern from the committee and can escalate to requests for re-training, peer visits, or a recommendation to the Ministry of Health that the non-compliant lab cease performing tests. The acceptance interval is determined by three standard deviations above and below the 12-month moving average of results across all seven labs. As the positive abnormal rate for each lab is compared to the acceptance interval, no consideration is given to the uncertainty of the rate resulting from sample size. This research shows that under this approach, differences in monthly sample sizes by lab have an important impact on the probabilities of classifying a lab in error and need to be considered. We suggest the weighted estimating equations (WEE) approach to regulate the bias/variance trade-off in estimates of the positive abnormal test rate since we expect that the true rate drifts slowly in an unpredictable way over time and sample sizes at any single time period may be small. We want to increase the power of a hypothesis test comparing positive abnormal rates by lab by increasing sample size and down-weighting the influence of historical data.

Data

The data arising from the seven labs conducting the FOBT in Ontario are described in Section 1.1. The dataset contains observed test outcomes from 863,898 patients who were tested at FOBT labs in Ontario over the 18-month period from January 2014 to June 2015. Figure 5-1

gives the number of patients by month, the observed positive abnormal rate ("positive rate") over time across all labs, and the linear trend line in positive rate.

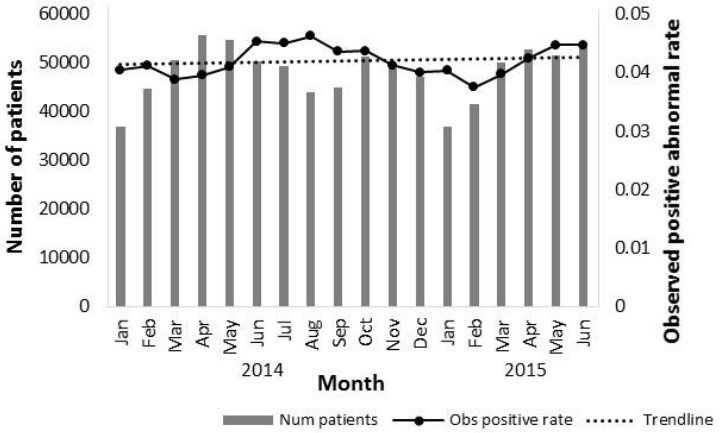


Figure 5-1. Observed positive rate and sample size of FOBT labs in Ontario

Figure 5-1 shows that the positive rate across all labs drifts slowly over time in an unpredictable way. A physician recommending a FOBT usually refers their patient to a particular lab for testing. In Ontario, a lab may service patients from as few as 100 or as many as several thousand referring physicians. As such, the number of samples tested by month varies considerably from lab to lab. The sample size and observed positive rate of each FOBT lab in the latest month (June 2015) are given in Figure 5-2.

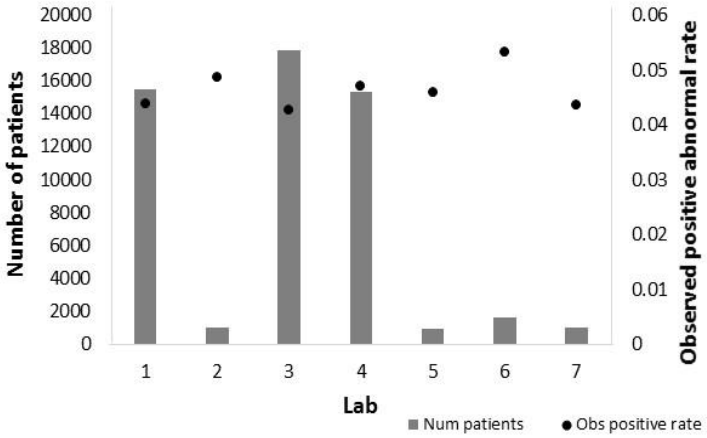


Figure 5-2. Observed positive rate and sample size of FOBT labs in Ontario in June 2015

Figure 5-2 shows that there are large differences in the numbers of patients who are tested across the various labs. In general, the number of FOBT samples tested varies from approximately 600

per month to approximately 20,000 per month. The varying number of monthly samples tested by lab impacts the power of the Ontario FOBT proficiency test to correctly classify a lab as acceptable or non-compliant based on the acceptance interval approach. We demonstrate this impact through an example of three groups having the same true rate but varying sample sizes. In May 2014, the estimate of the overall positive rate based on the moving average was 0.042 and the acceptance interval reported by Cancer Care Ontario was (0.037, 0.047). If we consider the true rate to be 0.045, the binomial distributions of the observed rates among groups having sizes 600, 5000, and 20,000 are given in Figure 5-3.

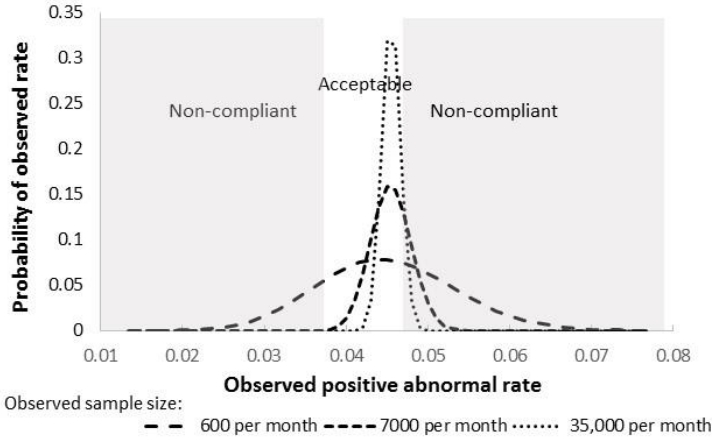


Figure 5-3. Distribution of observed positive rates by sample size for true rate = 0.045

Figure 5-3 shows that the true rate of each sample, 0.045, lies within the Cancer Care Ontario acceptance interval; however, we notice that we may observe a rate outside of this interval due to variation related to sample size. There is a 0.55 probability that a lab that tests 600 samples per month observes a rate outside the acceptance interval. The probability that the observed positive rate is outside this acceptance interval for a lab that tests 7000 samples per month is 0.11 and for a lab that tests 20,000 per month is 0.019. Clearly, the probability that the Ontario FOBT proficiency test incorrectly classifies a lab as non-compliant depends on the sample size. There is relatively high probability that a small lab will be classified as non-compliant in error.

The previous example points out that when one or more of the labs test few samples relative to other labs, the probability that the Ontario FOBT proficiency test incorrectly classifies the lab as acceptable or non-compliant may be large. Further, the acceptance interval is known to be calculated based on an average of results across all labs. Depending on how it is calculated, the data observed at a larger lab could have larger influence on the acceptance interval than a small lab. Changes in performance at a larger lab could move the acceptance interval over time and a

smaller lab that experiences no change may become non-compliant relative to the latest acceptance interval. Due to the wide disparity in sample sizes between the seven labs testing FOBT in Ontario, the power of the Ontario FOBT proficiency test to correctly classify small labs as acceptable or non-compliant is a concern.

We consider a more rigorous alternative to the Ontario FOBT proficiency test to compare positive rates at seven labs performing the FOBT through tests of hypotheses. We consider the test with null hypothesis H_0: all labs have same positive rate for the latest month versus the alternative H_A: at least one of the labs has a different positive rate than the others. If H_0 is rejected, then there is statistically significant evidence that there are differences between test results across the labs. The committee can review estimates from each of the labs and carry out follow-up analysis to identify the nature of the differences across labs. Three characteristics of the hypothesis test are considered:

- size, α: upper bound on the probability that the test is rejected for values of the parameter in the region where the null hypothesis is true
- power, $\beta(\theta)$: the probability that the test is rejected at a particular value of the parameter, θ
- unbiasedness: the condition that the power for values of the parameter in the region where the null hypothesis is false is at least as large as the size of the test

For tests with a select value of size α, we want the power of the test to be as large as possible on alternative values of the parameter θ among all unbiased tests. The power of a test is limited by the number of observations and so we look for an approach with the highest power of the test for H_0 versus H_A given some relatively small sample sizes by lab. Since increasing sample size increases the power of a hypothesis test (Lehmann and Romano, 2005), then a possibility to improve the power of a test is to combine data across time periods. However, including data observed in time periods before a change occurs reduces the power of the test aimed at detecting the change. The decision whether to use present time data only or to include some or all observed historical data depends on the sample size of the lab experiencing the change and the size of the change which are both unknown. We consider the WEE approach to combine present and historical data that increases power of the test over either naïve approach. In the application under consideration, there are no patient-level covariate data and so risk-adjustment is not needed.

Estimation by weighted estimating equations

Table 1-1 introduces the GLM that is selected for this problem based on π_m, the binomial positive rate at lab m for $m = 1, \dots, 7$. We assume that the random variables are independent across $t = 1, \dots, T$. The mean positive rate at the baseline lab (lab 1) is modelled by α. The positive rates of the other labs relative to the baseline lab are modelled by $\delta_1, \delta_2, \dots, \delta_6$. For patient j tested at hospital m at time t, the binomial positive rate π_{mt} relates to the model parameter $\theta_t = \left(\alpha_t, \delta_{1,t}, \delta_{2,t}, \dots, \delta_{6,t}\right)^T$ through the inverse link function

$$\pi_{mt}(\theta_t; d_t) = \frac{\exp\left(\alpha_t + \delta_{1,t} I_m[1] + \delta_{2,t} I_m[2] + \cdots, \delta_{6,t} I_m[6]\right)}{1 + \exp\left(\alpha_t + \delta_{1,t} I_m[1] + \delta_{2,t} I_m[2] + \cdots, \delta_{6,t} I_m[6]\right)} \tag{17}$$

where I_m is a size 6 vector with elements that are either 0 or 1 depending on the lab that the patient attended and $I_m[i]$ is the i^{th} element of I_m. We expect that levels α_t and $\delta_t = \left(\delta_{1,t}, \delta_{2,t}, \dots, \delta_{6,t}\right)^T$ may change slowly over $t = 1, \dots, T$ due to the influence of factors that are not included in the analysis. The observed test result for subject j at lab m at time t is recorded as $y_{jmt} = 1$ if the result is positive and $y_{jmt} = 0$ otherwise. The log-likelihood function describing the probability of data $d_t = \{y_{jmt}, j = 1, \dots, n_{mt}, m = 1, \dots, 7\}$ is

$$l_t(\theta_t; d_t) = \sum_{m=1}^{7} \sum_{j=1}^{n_{mt}} I\left[y_{jmt} = 1\right] \log \pi_{mt} + I\left[y_{jmt} = 0\right] \log(1 - \pi_{mt}) \tag{18}$$

for indicator variables $I\left[y_{jmt} = 0\right]$ and $I\left[y_{jmt} = 1\right]$.

We select weights $\{w_t, t = 1, \dots, T\}$ by (10) with weight parameter value $\lambda = 0.1$. The WEE estimates will be compared to the estimates by the two naïve approaches using the two special cases of the weights described in Section 3.1.

Under (3), the weighted estimating function vector of length 7 is

$$Q(\theta; d, w) = \sum_{t=1}^{T} w_t \psi_t(\theta; d_t) = \begin{bmatrix} \sum_{t=1}^{T} w_t \sum_{m=1}^{7} \sum_{j=1}^{n_{mt}} I\left[y_{jmt} = 1\right] - \pi_{mt} \\ \sum_{t=1}^{T} w_t \sum_{j=1}^{n_{2t}} I\left[y_{j2t} = 1\right] - \pi_{2\mathrm{t}} \\ \sum_{t=1}^{T} w_t \sum_{j=1}^{n_{3t}} I\left[y_{j3t} = 1\right] - \pi_{3\mathrm{t}} \\ \sum_{t=1}^{T} w_t \sum_{j=1}^{n_{4t}} I\left[y_{j4t} = 1\right] - \pi_{4\mathrm{t}} \\ \sum_{t=1}^{T} w_t \sum_{j=1}^{n_{5t}} I\left[y_{j5t} = 1\right] - \pi_{5\mathrm{t}} \\ \sum_{t=1}^{T} w_t \sum_{j=1}^{n_{6t}} I\left[y_{j6t} = 1\right] - \pi_{6\mathrm{t}} \\ \sum_{t=1}^{T} w_t \sum_{j=1}^{n_{7t}} I\left[y_{j7t} = 1\right] - \pi_{7\mathrm{t}} \end{bmatrix} \tag{19}$$

given the present time value of the parameter, $\theta = \theta_T = (\alpha, \delta_1, \delta_2, \dots, \delta_6)^T$, data $d = \{d_t, t = 1, \dots, T\}$, weights $w = \{w_t, t = 1, \dots, T\}$, and inverse link function $\pi_{mt}(\theta_t; d_t)$.

The WEE estimate $\hat{\theta}$ is the solution of $Q(\hat{\theta}; d, w) = 0$. Through (11), we estimate the weighted information estimate of variance, $\widehat{var}(\hat{\theta})$ involving $I_t(\hat{\theta})$, the expected information function at each time period evaluated at the WEE estimate. With estimate $\hat{\theta}$, we use Table 2-1 to compute an estimate $\hat{\pi}_m$ for each lab $m = 1, \dots, 7$. Similarly, with estimate $\widehat{var}(\hat{\theta})$, we compute estimate $\widehat{var}(\hat{\pi}_m)$ for $m = 1, \dots, 7$ through (5).

The null hypothesis H_0: all labs have the same positive rate for the current month in terms of parameter θ is $H_0: \delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = \delta_6 = 0$. We write this as $H_0: \delta = \delta_0$ with $\delta_0 = (0,0,0,0,0,0)^T$. The alternative hypothesis that allows for a different positive rate at one or more of the labs is H_A: at least one element of $\delta \neq 0$. There are $r = 6$ parameters of interest for testing and one remaining parameter, α.

Through (19) we calculate the unrestricted WEE estimate $\hat{\theta}$ as the solution of $Q(\hat{\theta}; d, w) = 0$. To calculate a likelihood ratio (LR) test statistic as specified in Section 3.4, we additionally estimate $\alpha = \alpha_0$ when δ is restricted to δ_0. The weighted estimating function gives the restricted WEE estimate $\hat{\alpha}_0$. The WEE LR test statistic is given by (12) and involves the log-likelihood function in (18) and WEE estimates $\hat{\theta}$ and $\hat{\alpha}_0$. An approximation for the distribution of the WEE LR test statistic under the null hypothesis that restricts six parameters is given by (13).

Results

We compare the WEE estimates for $\pi_m, m = 1, \dots, 7$ to those by the two naïve approaches discussed in Section 2.2. For the naïve approaches, estimates $\hat{\pi}_m$ and estimates of their variances are calculated through the WEE approach with one of the limiting values of the weight parameter. Figure 5-4 gives the estimates $\hat{\pi}_m, m = 1, \dots, 7$ based on the WEE approach and the two naïve approaches and the corresponding 95% confidence intervals assuming normality.

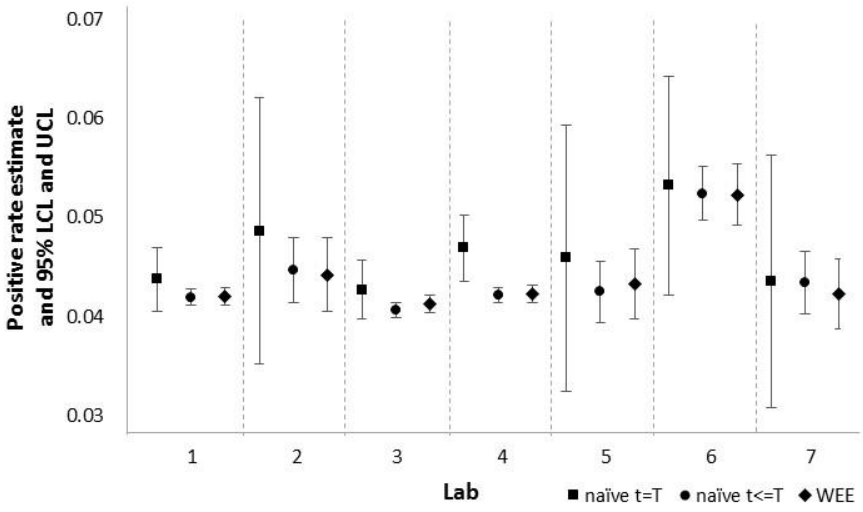


Figure 5-4. Estimates of positive rate for FOBT labs in June 2015 by various approaches

Figure 5-4 shows that estimates by the recommended WEE approach ($\lambda = 0.1$) have less uncertainty than estimates using present data only across all labs. The uncertainties of the WEE estimates are comparable to those of the naïve estimates that use all historical data. The WEE estimates of positive rates agree closely to those of the naïve approach with all historical data and these are significantly different than those of the naïve approach with present data only for labs 1, 2, 3, and 4. This is an indication that there has been some significant change in actual positive rates at these labs over the 18-month period. A WEE analysis with a larger selection of λ may be considered in this example in order to better balance the trade-off between bias and variance. Guidelines to select λ relative to the expected change in the true value of the parameter over time are discussed further in Section 6.2.

The WEE estimates and relevant quantities to test $H_0: \delta = \delta_0$ versus $H_A: \delta \neq \delta_0$ for the FOBT dataset are given in Table 5-1.

Table 5-1. WEE hypothesis test quantities for H_0: $\delta = \delta_0$ vs. H_A: $\delta \neq \delta_0$

Unconstrained model	**WEE estimate of $\boldsymbol{\theta}$**	$\hat{\theta} =$ (-3.126, 0.0528, -0.0187, 7.85E-3, 0.0315, 0.230, 6.41E-3)^T
	Weighted log likelihood	$\sum_{t=1}^{T} w_t l_t(\hat{\theta}; d_t) =$ -151,234.5420
Constrained model	**WEE estimate of $\boldsymbol{\theta}$**	$\hat{\theta}_0 =$ (-3.120)^T
	Weighted log likelihood	$\sum_{t=1}^{T} w_t l_t(\hat{\theta}_0; d_t) =$ -151,269.1618
WEE LR test statistic (12)		$\hat{S} = 69.24$
Weight-adjusted test statistic		$\frac{\sum_{t=1}^{T} w_t n_t}{\sum_{t=1}^{T} w_t^2 n_t} \hat{S} = 53.81$
p-value for H_0 under (13)		$\Pr\left(\chi_6^2 > \frac{\sum_{t=1}^{T} w_t n_t}{\sum_{t=1}^{T} w_t^2 n_t}\ \hat{S}\right) < 0.01$

Table 5-1 shows evidence to reject the null hypothesis H_0: $\delta = \delta_0$ in favour of the alternative H_A: $\delta \neq \delta_0$ for a size 0.05 test. The p-values for the same hypothesis test by the naïve approach with all historical data weighted equally is $p < 0.01$ and by the naïve approach based on present time data only is $p = 0.35$. An approach based on the latest monthly data is less sensitive at detecting differences among labs for this dataset. This is the current industry practice among the committee that oversees FOBT labs in Ontario.

A follow-up test of hypothesis is directed at detecting differences at a specific lab. Such a test is relevant to the management of a particular lab or the committee responsible for overseeing all laboratories. The null hypothesis remains as H_0: $\delta_1 = \delta_2 = \ldots = \delta_6 = 0$ and the alternative hypothesis becomes H_k: $\begin{cases} \delta_k \neq 0 \\ \delta_i = 0 \text{ for } i = 1, \ldots, 6, i \neq k \end{cases}$ for $k = 1, \ldots, 6$. The WEE likelihood ratio statistic corresponding to lab k is

$$\hat{S}_k = 2\left(\sum_{t=1}^{T} w_t l_t\left(\delta_i, \hat{\alpha}_k, \hat{\delta}_k; d_t\right) - \sum_{t=1}^{T} w_t l_t(\delta_0, \hat{\alpha}_0; d_t)\right) \quad (20)$$

where $\hat{\alpha}_0$ is the WEE estimate under the null hypothesis and $\hat{\alpha}_k$ and $\hat{\delta}_k$ are the WEE estimates under the specified alternative H_k. The approximate distribution for $\hat{S}_k$ under the null hypothesis follows from (13) with $r = 1$. The test statistics, $\hat{S}_k$, for the test of the null hypothesis against the lab-specific alternatives, H_k, for this dataset are given in Table 5-2.

Table 5-2. WEE hypothesis test quantities for $H_0: \delta = \delta_0$ vs. $H_k: \delta_k \neq 0$

k	related lab	WEE LR test statistic (20) $\widehat{S}_k$	Weight-adjusted test statistic	p-value for H_0 under (13)
1	**2**	1.42	1.11	0.29
2	**3**	9.61	7.46	<0.01
3	**4**	0.07	0.0514	0.82
4	**5**	0.45	0.353	0.55
5	**6**	62.5	48.5	<0.01
6	**7**	0.00	3.11E-4	0.99

Table 5-2 shows evidence to reject the null hypothesis in favour of alternative H_2 or H_5 for a size 0.05 test. There is evidence that positive rates are significantly different at lab 3 and lab 6 relative to lab 1. There is evidence to reject the null hypothesis in favour of the same two alternative hypotheses by the naïve approach with all historical data weighted equally. By the naïve approach based on present time data only, there is no evidence to reject the null hypothesis in favour of any of the alternative hypotheses. Once again, we see for this dataset that an approach based only on the latest monthly data is less sensitive at detecting differences among labs.

In addition to comparing the WEE LRT statistic to a critical value, it is useful to track the trend of the weighted WEE LR test statistic for the test of H_0 versus H_A over time. The trend in the weighted test statistic $\frac{\sum_{t=1}^{T} w_t n_t}{\sum_{t=1}^{T} w_t^2 n_t} \hat{S}$ at successive months from for the Ontario FOBT dataset is given in Figure 5-5.

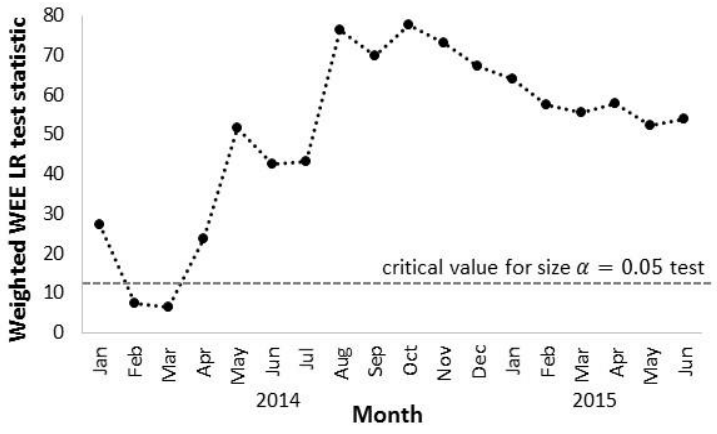


Figure 5-5. Weighted WEE LR test statistic H_0 vs. H_A by month for Ontario FOBT dataset

Figure 5-5 shows evidence to reject null hypothesis H_0 in favour of alternative H_A from April 2014 to June 2015 for a size 0.05 test. There are significant differences in the positive rate at one or more of the labs over this period. Note that this graph does not point to a particular lab and so there may be different outlier lab(s) from period to period. The graph points to a change at one of the labs that began around March 2014. Further, there is a downward trend that starts around September 2014. The downward trend from September 2014 to June 2015 may indicate that positive rates across the labs are becoming more consistent with one another. A distinctive trend in the weighted WEE LR test statistic should be investigated with the follow-up analysis discussed previously.

Formal process monitoring could be used to provide quicker detection of small sustained shifts and control the misclassification rate at a desired level. Liu et al. (2008) propose a control chart statistic based on the likelihood ratio test for monitoring multiple stream processes to detect a change in both the overall process mean and changes in the individual stream means. The authors show that this test does not require a phase 1 sample which saves cost. This work could be extended to develop a control chart for the WEE LR test statistic to improve time to detection and misclassification rate.

5.2. Simulation study

We simulate data that resembles the fecal occult blood test in Ontario dataset to study the power and unbiasedness of the size α tests of hypotheses by the various approaches. We compare the WEE approach to the two naïve approaches. We discuss the limitations of the results and the impact of certain characteristics of the data.

We simulate datasets with sample sizes similar to the Ontario FOBT lab problem where the number of samples per month ranges from 600 to 35,000 across the seven labs and the total sample size is 60,000 observations per month. Figure 5-6 gives the sample sizes n_{mt} for labs $m = 1, \dots, 7$ which are the same for each month $t = 1, \dots, T$.

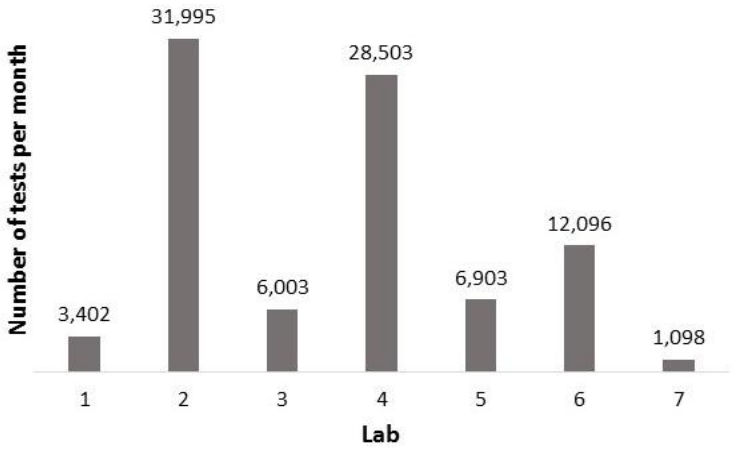


Figure 5-6. Sample size by lab per month

Each simulated dataset contains observations by lab per month over a period of 18 months. As stated, the objective of the analysis is to regularly assess a laboratory's ability to provide an acceptable standard of service by comparison with peers and so parallel changes at all labs simultaneously are not of interest in this problem. Each dataset is designed with positive rate at the first time period equal to $\pi_{m,1} = 0.042$ for each lab $m = 1, \ldots, 7$. Following the first time period, a change is introduced into a single lab and positive rates at the remaining labs are unchanged. We simulate a change at either the largest lab or the smallest lab in order to study the power and unbiasedness of the hypothesis test at the extremities of lab sample sizes. Many changes are possible; we simulate a step or linear change that increases or decreases the positive rate over an 18-month period. We add a profile for the base case where the positive rate stays constant at all labs over time. Under these conditions, there are nine profiles of change in positive rates over time as summarized in Table 5-3. The profile lettering refers to the design values for the positive rates given in Figure 5-7.

Table 5-3. Lab positive rate design profiles, 9 profiles

Design profile	Size of the lab undergoing the change	Type of change	Direction of change	Positive rate for lab m at month t, π_{mt}
I	none	none	none	profile a for all m
II	small lab	step change	increase	profile a for $m = \{1, \dots, 6\}$ profile b for $m = 7$
III	small lab	step change	decrease	profile a for $m = \{1, \dots, 6\}$ profile c for $m = 7$
IV	small lab	linear change	increase	profile a for $m = \{1, \dots, 6\}$ profile d for $m = 7$
V	small lab	linear change	decrease	profile a for $m = \{1, \dots, 6\}$ profile e for $m = 7$
VI	large lab	step change	increase	profile a for $m = \{1,2,3,5,6,7\}$ profile b for $m = 4$
VII	large lab	step change	decrease	profile a for $m = \{1,2,3,5,6,7\}$ profile c for $m = 4$
VIII	large lab	linear change	increase	profile a for $m = \{1,2,3,5,6,7\}$ profile d for $m = 4$
IX	large lab	linear change	decrease	profile a for $m = \{1,2,3,5,6,7\}$ profile e for $m = 4$

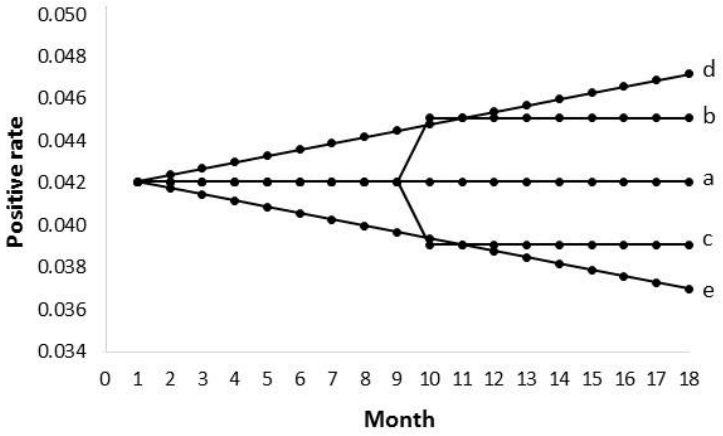


Figure 5-7. Positive rate design profiles a-e

A positive or negative test response y_{jmt} is simulated for each sample j at lab m at time t by the binomial distribution with the appropriate positive rate design value. For each of the nine design profiles, we simulate 5000 datasets of $d_S = \{y_{jmt}, j = 1, \dots, n_{mt}, m = 1, \dots, 7, t = 1, \dots, 18\}$

based on the sample size by lab, n_{mt}, given in Figure 5-6 and the positive rate design profile over time by lab, π_{mt}, given in Table 5-3 and Figure 5-7.

For each of the 5000 simulations of d_S for each of the nine design profiles, we calculate the WEE LR test statistic and reject or do not reject H_0 versus H_A based on the asymptotic approximation for its distribution under the null hypothesis in (13). We do this for every value of $T = 1, \ldots, 18$ and for each of the naïve and WEE approaches to study the power and unbiasedness of the tests statistics by the various approaches over successive time periods. With these simulation results, we evaluate the size for the design profile where the null hypothesis is known to be true and the power and unbiasedness for design profiles where the null hypothesis is known to be false. Figure 5-8 gives the percentage of LR test statistics by the WEE and naïve approaches where the null hypothesis is rejected at $\alpha = 0.05$ for data simulated by Profile I where the null hypothesis is known to be true.

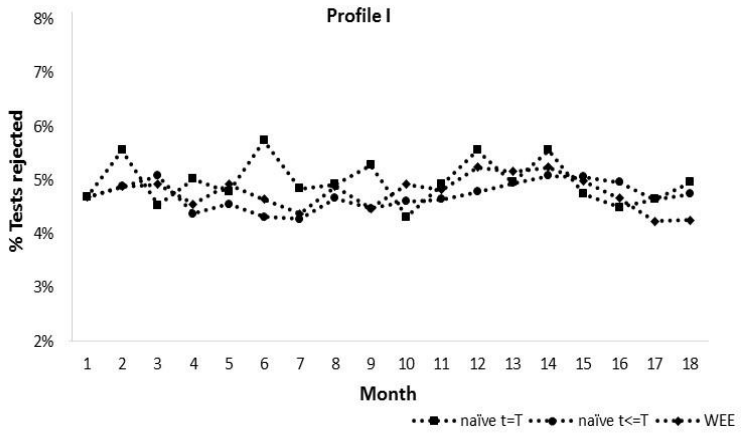
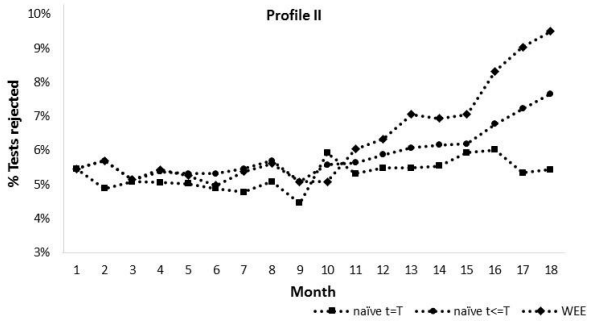


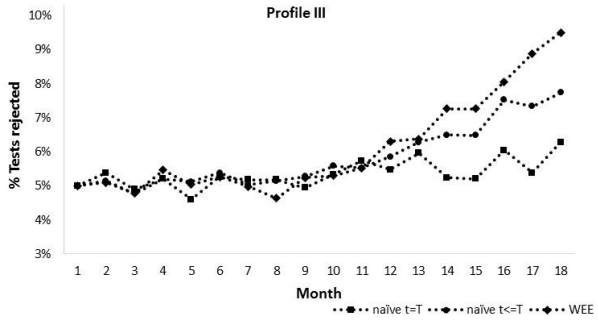
Figure 5-8. Percentage of tests of H_0 rejected for profile I (no change)

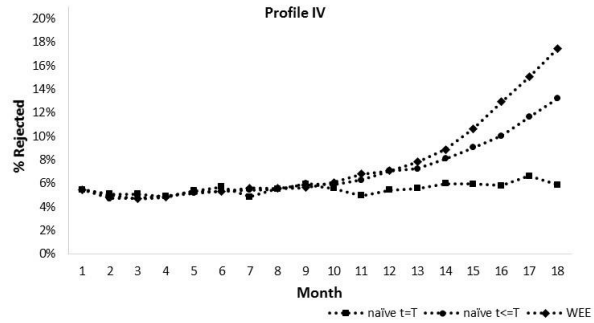
Figure 5-8 shows that the percentage of tests rejected over time is similar for the WEE and naïve approaches. The LR test statistic by the WEE approach rejects the null hypothesis for 4.8% of datasets, and those by the naïve approach with all historical data and only present time data reject for 4.7% and 5.0% of datasets, respectively. The closeness of the observed sizes of the tests compared to the design value for the size of test (5%) is expected and indicates that the approximations for the critical values of the test statistics are reasonable. The observed differences in actual sizes of the tests among the three approaches do not have an important impact on the interpretation of the power of the tests to follow.

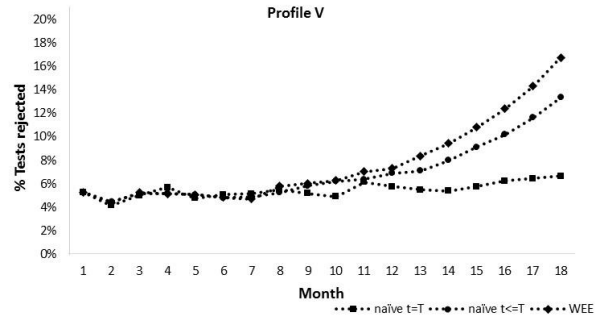
Figure 5-9 gives the percentage of LR test statistics by the WEE and naïve approaches where the null hypothesis H_0 is rejected in favour of the alternative H_A based on data simulated with each

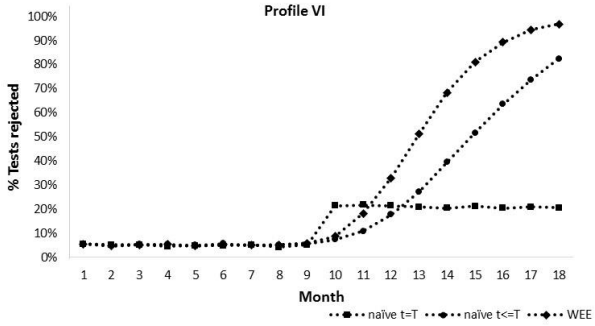
of the eight design profiles where the null hypothesis is known to be false. The graphs are interpreted as the observed power of the various test statistics to reject the null hypothesis with sizes of the test close to 0.05. We expect the observed power to increase according to the known change in positive rate. Note that there are differences in the scales of the vertical axes.

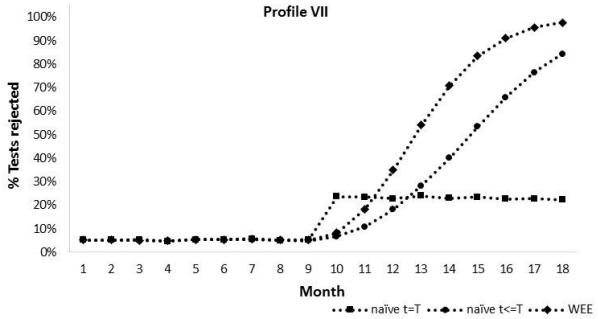












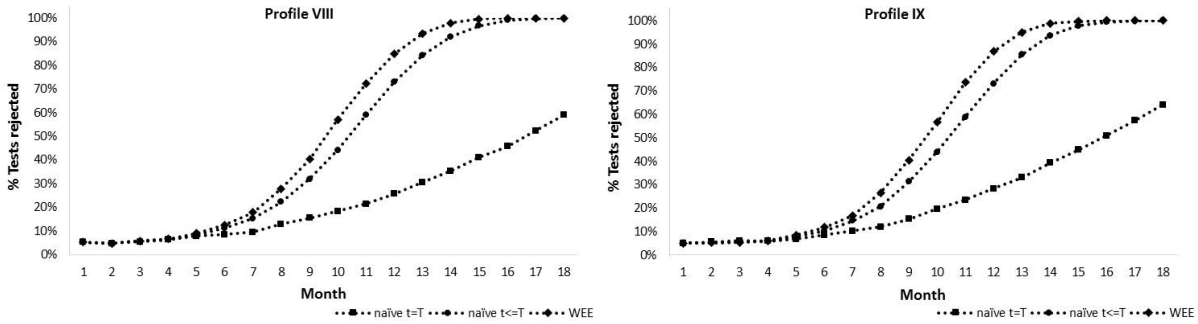


Figure 5-9. Percentage of tests of H_0 rejected for profiles II-IX

Figure 5-9 shows that the power usually increases with more time periods since the step change and as the linear change gets larger. The exception is under the naïve approach based only on current time data where power does not increase with more time periods following a step change (naïve t=T for Profiles II, III, VI, VII). In general, the WEE approach has higher power to detect a change after a given number of time periods and requires fewer time periods to achieve a particular level of power.

We investigate how the power to detect a change increases as the size of the change increases in a follow-up simulation study. In this study, we simulate data where the true value of the positive rate does not change for time periods $t = 1, \ldots, 9$ and then either a linear or step change of various sizes occurs at the small lab 7. Figure 5-10 gives the observed power to detect the linear or step change of various sizes at three time periods following the change ($T = 12$) by test statistics from the various approaches.

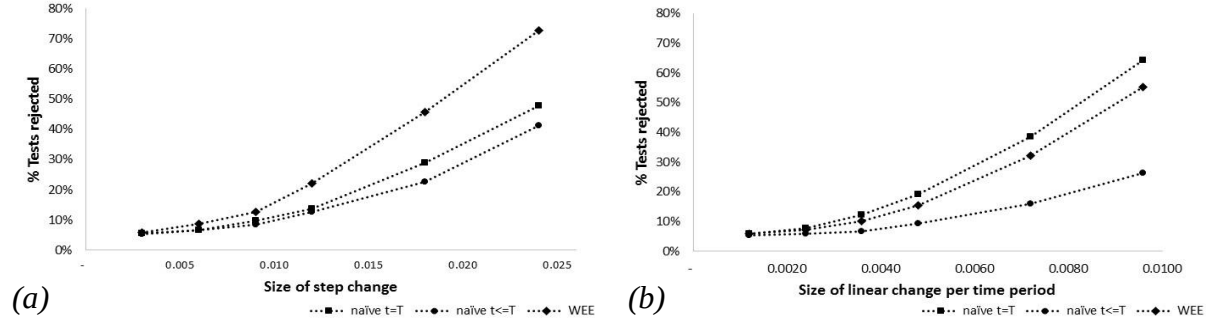


Figure 5-10. Power of test to detect change at a small lab after three months: (a) following a step change, (b) following a linear change

Figure 5-10(a) shows that the power of the WEE approach to detect a step change of 0.024 in the positive rate (from 0.042 to 0.066) at lab 7 at three months since the change is favourable at 72.8%.

As the size of the step change increases over the range from 0.003 to 0.024, the WEE approach has increasingly more power to detect the change than either naïve approach. Figure 5-10(b) shows that the power of the WEE approach to detect a linear change of 0.01 per month (from 0.042 to 0.072 after three months) at lab 7 at three months is 55.4%. The naïve approach using current time data only has slightly more power than the WEE approach for this example since the change is relatively large and the total amount of data since the change is relatively small compared to that from before the change. As more time passes since the start of the change, we expect that the power of the WEE approach to detect a change will surpass the power of the naïve t=T approach. The nine time periods of data observed before the change considerably reduce the power of the naïve approach using all historical data weighted equally compared to the other approaches. This study shows that there is favourable power of the WEE approach for detecting a change at a small lab within a short time frame depending on the size of the change.

5.3. Summary and discussion

The proficiency test to assess the ability of laboratories to perform fecal occult blood tests (FOBT) in Ontario compares the observed positive abnormal rate from various labs to an acceptance interval based on data across all labs. When one or more of the labs test few samples relative to the other labs, the probability that the Ontario FOBT proficient test incorrectly classifies the lab acceptable or non-compliant may be large. There is wide disparity in sample sizes between the seven labs testing FOBT in Ontario. The power of the Ontario FOBT proficient test to correctly classify small labs is a concern.

We analyse real Ontario FOBT outcome data from seven labs over a period of 18 months. We use the weighted estimating equations approach since we expect that test performance may drift slowly over time in an unpredictable way and some sample sizes may be small. We compare the WEE estimate of the positive rate to estimates by naïve approaches based on present time data only and all historical data weighted equally. The various approaches produce positive rate estimates that vary considerably from one another. The WEE estimate has similar precision to the estimate based on all historical data weighted equally. Based on the WEE approach and the naïve approach based on all historical data weighted equally, we reject a test of the null hypothesis that all labs have the same positive rate in favour of an alternative hypothesis that not all labs have the same positive rate. We do not reject this null hypothesis based on the analysis of present data only. Similarly, two of the tests against lab-specific alternative hypotheses are rejected based on the WEE approach and the naïve approach using all historical data, but not rejected based on the

analysis of present data only. There are important differences in the results based on the WEE approach compared with those in line with current industry practice.

We explore the power of the hypothesis test to detect difference between labs by the various approaches through simulated data designed with varying changes in positive rate over time. The conditions for the simulation study reflect those of the Ontario FOBT proficiency test at May 2014 as a prototype example, including the number of labs, sample sizes by lab, and initial positive abnormal rates. In general, the WEE approach has higher power to detect a change after a given number of time periods and requires fewer time periods to achieve a particular level of power. As the size of a step change increases, the WEE approach has increasingly more power to detect the change than either naïve approach under the particular simulation conditions. Under a linear change, initially the approach based on present time data only has higher power, but the power of the WEE approach surpasses the power of the naïve approach within a short time frame depending on the size of the change. The work of this chapter indicates that a more reliable Ontario FOBT proficiency test can be constructed based on the WEE approach that has suitable power to detect changes at a lab of any size and reduces the risk of classifying a lab as non-compliant in error.

Multiple testing

In the lab positive rate application, the tests of H_0 versus H_A and H_0 versus H_k are multiple testing problems since we test the significance of multiple stream effects simultaneously. An alternative is to test separate hypotheses for each stream effect; for example, $H_0: \delta_1 = 0$ versus $H_A: \delta_1 \neq 0$. Lehmann and Romano (2008, p.349) point out that the probability of a false rejection rises rapidly with the number of tests, here -1. When the number of true hypotheses is large, we are nearly certain to reject some of them. Lehmann and Romano (2008) discuss strategies such as the Bonferroni procedure and the Holm procedure for controlling the probability of one or more false rejections for multiple testing problems. In some applications, there may be thousands of treatments under test in which case an adjustment for multiplicity is important. In the lab positive rate problem, the number of hypotheses is fairly small at $M = 7$. We proceed without an adjustment for multiplicity.

Selecting historical time window

In an analysis involving historical data, we must select a time window for the data to include in the analysis. In the case where we expect that the true value of the parameter has had a significant, sustained change, we want to restrict our analysis to data following the change. In the case where we expect that the parameter changes slowly, the effect of the time window is related to sample

sizes and the values of the weights over time. In Section 3.2, we discuss that precision of an estimate is related to effective sample size and show for a binomial model without covariates that $N_{eff} = \frac{\left(\sum_{t=1}^{T} w_t n_t\right)^2}{\sum_{t=1}^{T} w_t^2 n_t}$ for weights $\{w_t, t = 1, \ldots, T\}$ and sample sizes $\{n_t, t = 1, \ldots, T\}$. Under the WEE approach with exponentially declining weights, the increase in N_{eff} declines as we widen the size of the time window and effectively approaches an upper bound. In Figure 5-11, we explore the values of N_{eff} for various sizes of the historical time window based on the number of FOBT patients observed over time as given in Figure 5-1 and $\{w_t, t = 1, \ldots, T\}$ with weight parameter $\lambda = 0.1$.

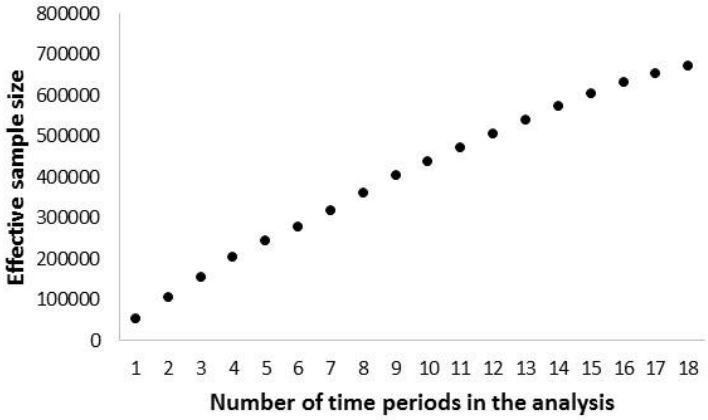


Figure 5-11. Effective sample size vs. size of historical time window for FOBT dataset

Figure 5-11 shows that the effective sample size increases as we expand the size of the historical time window over this range. The curve will level off as data from more historical time periods become available, at which time widening the historical time window will have little effect on the estimates. In contrast, under the naïve approach involving all historical data weighted equally, there is no effective upper bound on the effective sample size. Here it is more important to select the time window with care in order to reduce the possibility for added bias since the historical data have equal weight. We can make an arbitrary selection of the desired effective sample size, $N_{eff} = \sum_{t=1}^{T} n_t$. Since the WEE approach down-weights historical data with exponentially declining weights, then the selection of a time window has less impact on the estimates compared to the naïve approach with all historical data weighted equally.

Considerations for some large sample sizes

The WEE approach is motivated by the need for a bias/variance trade-off when the parameter changes slowly over time and sample size in the present time period is small. In the data on FOBT labs in Ontario, Figure 5-2 shows that three of the seven labs have few samples (< 1000) relative to three large labs ($> 15{,}000$ samples) in the present time period. The simulation results show that the WEE approach has higher power to detect a change at the small lab relative to either of the naïve approaches. The same observation holds for a change at the large lab, though the power values are higher for all approaches. For testing a hypothesis related to a large lab, suitable power can be achieved when using present data only. We prefer to use present data in the case where the present sample size is sufficient so that we minimize the potential for adding bias when the parameter may change over time. We consider two alternatives to the standard WEE approach where there are sufficient data in the present time period for at least one lab and small sample sizes at other labs.

One alternative to the standard WEE approach is to exclude the historical data $\{d_{m',1}, d_{m',2}, \dots, d_{m',T-1}\}$ observed at a large lab m' from the analysis. The mean of the particular lab is estimated through the present data $d_{m',T}$ only. The formulations of the WEE approach in Chapter 3 apply directly in this case. This alternative has the effect of reducing the relative weight given to data from that particular lab in the estimation of covariate effects (within a model that includes covariates). The covariate effect estimates are less precise since less data are used for estimation. In the case where there are no covariates in the model, then this is the best alternative. Further consideration of the approximations of the estimate of the variance of $\hat{\theta}$ and the distribution of the hypothesis test statistic involving $\hat{\theta}$ under this alternative is required.

A second alternative that maintains the precision of the covariate effect estimates is to separate the estimation of covariate effects from the estimation of the lab effects in a two-stage approach. In stage 1, the estimation of covariate effects is based on all present and historical data from all labs. In stage 2, the covariate effects estimates are used as fixed values in the estimation of lab effects. Under this alternative, stage 2 estimation of the effect for a large lab m' involves its present data $d_{m',T}$ only. Since the covariate effects are fixed in stage 2, then the estimation of the various lab effects can be separated. Usual MLE results for the estimate of uncertainty and distribution of the hypothesis test statistic apply to the estimates based on present time data only and WEE results apply to the estimates based on weighted estimating equations as before. The two-stage approach is discussed further as future work in Section 7.2.

Chapter 6: Hospital Performance Measure

Complications of surgical care are a major cause of death and disability worldwide (World Health Organization, 2009). Confronted with this problem, the World Health Assembly adopted a resolution urging countries to strengthen the safety of health care and monitoring systems in 2002. In the United States, the Centers for Medicare and Medicaid Services (CMS) have a congressional mandate to evaluate hospital performance using risk-adjusted mortality rates. The CMS began publicly reporting hospital 30-day mortality rates for patients with acute myocardial infarction and heart failure in June 2007 and for pneumonia in 2008. In Canada, the Canadian Institute for Health Information (CIHI) provides information on Canada's health system under the mandate to accelerate improvements in health system performance. One of their goals is to expand their analytical tools to support measurement of health systems (Canadian Institute for Health Information, 2016). Clearly, statistical methods for assessing patient outcomes following surgery is an issue of substantial public importance.

In the context of surgical performance, the patient outcome following surgery is an important indicator of quality at the hospital where the patient is treated. Patient outcomes vary across hospitals due to individual patient health at admission (patient risk factors) as well as the quality of the surgical process and post-surgical care. A performance measure for surgical performance at a particular hospital must adjust for the risk factors of the patients it has treated but not adjust for differences related to its surgical process and post-surgical care quality. With an appropriate performance measure, Spiegelhalter et al. (2012) discuss three primary functions of this measure. Specifically, a regulator or stakeholder may want to

- compare performance to target
- screen performance to decide which hospitals to inspect
- monitor performance for arising problems

Of particular importance is the ability of these functions to inform stakeholders who can accelerate improvements in patient outcomes. Uncertainty in the performance measure is also important for stakeholders to consider and will be affected by the number of cases seen at the various hospitals and other factors.

The New York State (NYS) Department of Health (DOH) has studied the effects of patient and treatment characteristics on outcomes for patients with heart disease for over 20 years. A common procedure performed on patients with coronary artery disease is percutaneous coronary

intervention (PCI). Annually, the NYS Department of Health publishes a report based on information collected on patients over a three-year period who underwent PCI in NYS hospitals (New York State Department of Health, 2015). Their hospital-specific performance measure adjusts for its patients' health at admission through an estimate of risk-adjusted mortality rate for a mix of patients identical to the statewide mix. The current practice to estimate the measure for a particular hospital involves estimates of its observed mortality rate, expected mortality rate for its observed patient mix, and the observed statewide mortality rate. The mathematical formulation is given in more detail in Section 6.1. The estimate of its observed morality rate is a naïve estimate based on the observed patient outcomes for a particular hospital. As such, there is a high degree of instability and uncertainty in the NYSDOH estimates of performance for a low volume hospital in particular. This instability and uncertainty limit the usefulness of the performance measure. For example, when a hospital treats two patients in a time period, then its estimate of observed mortality rate can take one of the three possible values 0, 50%, or 100% and binomial uncertainty interval estimates for these quantities are extremely large. A risk-adjusted measure involving this estimate of the observed mortality rate is not able to serve the necessary functions. The NYSDOH pools data over a three-year time window to increase the number of cases observed by hospital. We point out that though pooling data reduces uncertainty, this approach increases bias in an estimate of the present time performance when performance changes over time. Further, this approach reduces the sensitivity to identify changes over time which is one important function of this measure. Considerable uncertainty may remain.

The Centers for Medicare and Medicaid Services (CMS) uses an approach recommended by the COPPS-CMS White Paper Committee (2012) that similarly pools data over a three-year time period. The estimate of the performance measure by the CMS approach involves a risk-adjusted prediction of the mortality rate for a particular hospital. The key difference to the NYSDOH approach is that the CMS approach stabilizes the estimate of the hospital-specific performance measure through a hierarchical, random effects model. The random effects model estimates fixed covariate effects and predicts hospital-specific effects. The model is fit with data observed at all hospitals in the country that perform the particular surgery and so the fixed covariate effect estimates borrow strength across hospitals. The predicted mortality rate estimates through the random effects model are closer to the overall mortality rate across all hospitals and have lower standard error than the naïve observed mortality rate estimates (COPPS-CMS White Paper Committee, 2012). Thus, this model is referred to as a shrinkage model. The mathematical formulation is given in more detail in Section 6.1.

Another alternative is to estimate the observed mortality rate with a fixed effects model. On average, the stabilized predicted mortality rate estimates by the shrinkage model are closer to the overall mortality rate and have lower mean squared error (MSE) than estimates based on fixed effects. Kalbfleisch and Wolfe (2013) point out that the MSE advantage of the shrinkage model is achieved by smaller error among the large number of hospitals near the centre of the distribution at the expense of larger error among hospitals with exceptional outcomes. They state that if the goal is to have high power for identifying hospitals with exceptional outcomes and to estimate the difference from the expected outcome for such exceptional hospitals, then fixed effects methods are better than random effects methods. Another criticism of this approach is that shrinkage has the effect of producing estimates for low volume hospitals that are close to the overall mean. Some stakeholders argue in favour of different shrinkage models depending on the volume of the hospital and others argue for no shrinkage at all (COPPS-CMS White Paper Committee, 2012). Further, the shrinkage model may be hard for hospital performance stakeholders to understand.

There is an opportunity to improve the bias and uncertainty in estimates of present performance beyond the approaches of pooling data and borrowing strength across hospitals and as an alternative to the shrinkage model. The weighted estimating equations approach borrows information from the past in order to manage a bias/variance trade-off in an estimate of present performance when surgical performance may drift slowly over time in an unpredictable way and some sample sizes may be small. Where stakeholders pay regular attention to hospital performance issues, we expect that surgical quality changes slowly over time. The WEE approach increases the statistical information for estimation by involving past data through the weighted estimating functions. Then, estimates of present performance have less bias than pooling data without weights and less uncertainty than using present data only. Similar to the CMS approach, the WEE approach also borrows strength across hospitals for estimates of covariate effects. The WEE approach has intuitive properties that can be understood by the hospital performance measure stakeholders.

This chapter compares the WEE approach with the NYSDOH and CMS current practices to estimate a present surgical performance measure from a stream of patient outcome data across hospitals with various surgical process and post-surgical care quality and patient risk factors. The objective is to reduce uncertainty in the estimates when some sample sizes are small and manage the added bias caused by slowly changing surgical quality over time. In Section 6.1, we introduce a realistic dataset that has similar properties to the PCI patient outcomes in NYS over the period from 2004 to 2012 and give the mathematical formulations of the CMS, NYSDOH, and WEE approaches. In Section 6.2, we look at results for the performance measure estimates across

hospitals by the various approaches based on the dataset. In Section 6.3, we discuss considerations for implementing the WEE approach in practice.

6.1. Mortality rate following percutaneous coronary intervention in New York State

Data

The realistic data arising from the patients undergoing percutaneous coronary intervention (PCI) at New York State hospitals are described in Section 1.1. The actual NYSDOH data are inaccessible to the public and so a realistic dataset is created with similar properties to the actual data that are available. In particular, the sample sizes, mix of patient covariate levels, observed mortality rates, and logistic regression estimates of the covariate effects match closely to those in the NYSDOH reports (e.g., New York State Department of Health, 2015). The dataset contains realistic test outcomes for 467,401 patients at 60 hospitals over the nine-year period from 2004 to 2012. Figure 6-1 gives the number of patients who underwent PCI by year, the observed mortality rate over time, and the linear trend line in mortality rate.

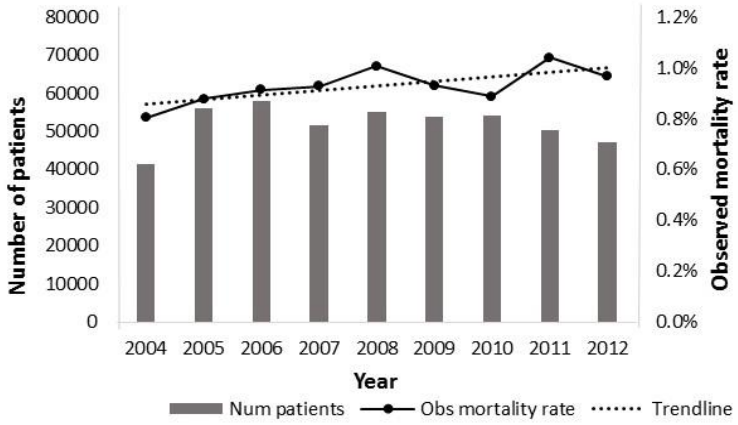


Figure 6-1. Observed mortality rate and sample size of PCI patients over time

Figure 6-1 shows that mortality rate increases slowly over time. This naïve analysis based on sample proportions in individual time periods is not a useful indicator of surgical performance over time since the distribution of risk factors for the patient population changes over time. An increase in the relative risk of patients at admission or an increase in the relative number of patients at a poorer performing hospital would result in an increase in the observed mortality rate over time, often even in the case where general surgical performance is improving. Note that the number of

patients who underwent PCI varies over time and the number of patients treated in the latest time period is smaller than almost all of the previous time periods. When we restrict our attention to the latest year's data (2012), then the similar graph by hospital is given in Figure 6-2.

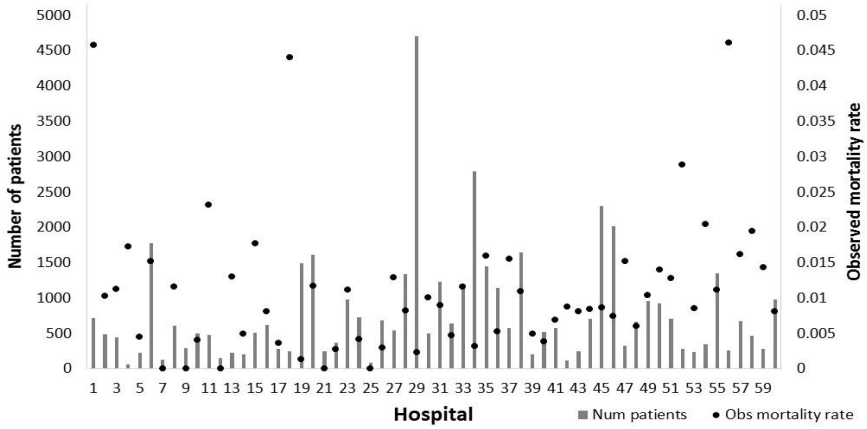


Figure 6-2. Observed mortality rate and sample size of PCI patients in 2012 by hospital

Figure 6-2 shows that there are large differences between the numbers of patients who underwent PCI across the various hospitals. Hospitals 4 and 25 treated 58 and 80 patients, respectively, whereas other hospitals treated as many as 4708 patients. Note that five hospitals reported no deaths among their patients.

NYSDOH risk-adjusted mortality rate

The risk-adjusted mortality rate approach in use by the New York State Department of Health (NYSDOH) estimates the hospital's mortality rate among PCI patients for a mix of patients at that hospital identical to the statewide mix. To get the hospital-specific, risk-adjusted mortality rate, the observed mortality rate at a particular hospital is first divided by the hospital-specific expected mortality rate. The ratio is then multiplied by the statewide (NYS) observed mortality rate. The hospital-specific expected mortality rate based on a fixed effects regression model is an estimate of the hospital's mortality rate given that the hospital's performance is the same as the average performance of all hospitals statewide. The likelihood function for the fixed effects regression model is based on a generalized linear model with linear predictor $\eta_{jm} = \alpha + \beta^T x_{jm}$, response distribution $Y_{jm} \sim binomial(1, \pi_{jm})$, and link function $\eta_{jm} = \log\left[\frac{\pi_{jm}}{1-\pi_{jm}}\right]$ for $j = 1, \dots, n_m$ where n_m is the number of patients who undergo PCI surgery at hospital m over the three-year period. The definitions of the parameters are the same as in Table 1-1. Estimation of α and β is based on

data from all patient outcomes observed in NYS in the latest three-year time period. With estimates $\hat{\alpha}$ and $\hat{\beta}$, we estimate the probability of mortality for patient j at hospital $m = 1, \dots, M$ having covariate vector x_{jm} as

$$\hat{\pi}_{jm} = \frac{\exp(\hat{\alpha}+\hat{\beta}^T x_{jm})}{1+\exp(\hat{\alpha}+\hat{\beta}^T x_{jm})}$$

The estimates $\hat{\pi}_{jm}$ are combined for all patients $j = 1, \dots, n_m$ at a particular hospital m to estimate its expected mortality rate. Then, the NYSDOH risk-adjusted, hospital-specific mortality rate estimate is

$$\hat{\pi}_m = \frac{OMR_m}{\frac{1}{n_m}\sum_{j=1}^{n_m}\hat{\pi}_{jm}} \times OMR_{NYS} \tag{21}$$

where $OMR_m = \frac{\sum_{j=1}^{n_m} y_{jm}}{n_m}$ and $OMR_{NYS} = \frac{\sum_{m=1}^{M}\sum_{j=1}^{n_m} y_{jm}}{\sum_{m=1}^{M} n_m}$ are the naïve estimates of observed mortality rates observed at hospital m and across NYS, respectively. The annual reports do not specify the methodology for estimating the uncertainty of $\hat{\pi}_m$. The stated confidence intervals are close to the 95% Agresti-Coull binomial confidence intervals for OMR_m $\left(LCL(OMR_m) \text{ and } UCL(OMR_m)\right)$ and fixed values for $\hat{\pi}_m$ and OMR_{NYS} so

$$LCL(\hat{\pi}_m) = \frac{LCL(OMR_m)}{\frac{1}{n_m}\sum_{j=1}^{n_m}\hat{\pi}_{jm}} \times OMR_{NYS}$$

$$UCL(\hat{\pi}_m) = \frac{UCL(OMR_m)}{\frac{1}{n_m}\sum_{j=1}^{n_m}\hat{\pi}_{jm}} \times OMR_{NYS}$$

The NYSDOH mortality rate estimate in (21) adjusts for the population of patients treated at each particular hospital. The hospital-specific ratio $\frac{OMR_m}{\frac{1}{n_m}\sum_{j=1}^{n_m}\hat{\pi}_{jm}}$ represents the performance of the particular hospital relative to the performance of the state as a whole. If the resulting ratio is larger (smaller) than one, the hospital has a higher (lower) mortality rate than expected on the basis of its patient mix. The hospital-specific ratio is converted to a mortality rate by multiplying the ratio by the observed mortality rate across all NYS PCI patients.

CMS hierarchical random effects model

The current practice used by the Center for Medicare and Medicaid Services (CMS) to address the challenge of estimation with small samples is a hierarchical random effects model which accounts for patient-level risk factors and hospital-level variation (COPPS-CMS White Paper Committee, 2012). The prediction of the hospital-specific mortality rate through the hierarchical

random effects model takes the place of the observed mortality rate in the NYSDOH approach. The likelihood function for the problem under consideration is based on a generalized linear mixed model (GLMM) with linear predictor $\eta_{jm} = \alpha + \delta_m + \beta^T x_{jm}$, response conditional distribution $Y_{jm}|\, \delta_m \overset{ind}{\sim} binomial(1, \pi_{jm})$, hospital-specific effects distribution $\delta_m \overset{iid}{\sim} N(0, \tau^2)$, and link function $\eta_{jm} = \log\left[\frac{\pi_{jm}}{1-\pi_{jm}}\right]$. The definitions of the parameters are the same in Table 1-1 except that $\delta = (\delta_1, \delta_2, \dots, \delta_M)^T$ is the random effect of hospitals $m = 1, \dots, M$ on the mean. Estimation of α, β, and τ^2 follow usual procedures for maximizing the likelihood within a GLMM. The effect δ is predicted through the estimate of the between–hospital variation, τ^2, and observed hospital level means. With estimates $\hat{\alpha}$ and $\hat{\beta}$ and prediction $\hat{\delta}$, we predict the mortality rate for patient *j* having covariate vector x_j at hospital $m = 1, \dots, M$ as

$$\hat{\pi}_{jm}(\hat{\theta}) = \frac{\exp(\hat{\alpha}+\hat{\delta}_m+\hat{\beta}^T x_{jm})}{1+\exp(\hat{\alpha}+\hat{\delta}_m+\hat{\beta}^T x_{jm})}$$

for $j = 1, \dots, n_m$. The hospital-specific estimate of risk-adjusted mortality rate is

$$\hat{\pi}_m(\hat{\theta}) = \frac{\sum_{j=1}^{n_m} \hat{\pi}_{jm}(\hat{\theta})}{\sum_{j=1}^{n_m} \frac{exp(\hat{\alpha}+\hat{\beta}^T x_{jm})}{1+exp(\hat{\alpha}+\hat{\beta}^T x_{jm})}} \times OMR_{NYS} \tag{22}$$

where $OMR_{NYS} = \frac{\sum_{m=1}^{M}\sum_{j=1}^{n_m} y_{jm}}{\sum_{m=1}^{M} n_m}$ is the naïve estimate of observed mortality rate observed at NYS.

In order to estimate the uncertainty of $\hat{\pi}_m$, the current CMS approach determines a hospital-specific estimate $\hat{\pi}_m$ and 95% confidence intervals (LCL and UCL) of this estimate through the following bootstrap algorithm (COPPS-CMS White Paper Committee, 2012).

1. Sample M hospitals with replacement.
2. Fit the GLMM with all cases among the M sample hospitals. Estimate α, β, and τ^2 and predict δ.
3. Predict a hospital random effect $\hat{\delta}_m^b, m = 1, \dots, M$ by sampling from the distribution of the hospital-specific distribution $\hat{\delta}_m^b \sim N(\hat{\delta}_m, \hat{\tau}^2)$ for the unique set of M hospitals. If a hospital is sampled more than once, randomly select one random effect prediction.
4. Estimate adjusted mortality rate by hospital for bootstrap sample b,

$$\hat{\pi}_m^b = \frac{\sum_{j=1}^{n_m} \frac{exp(\hat{\alpha}+\hat{\delta}_m^b+\hat{\beta}^T x_{jm})}{1+exp(\hat{\alpha}+\hat{\delta}_m^b+\hat{\beta}^T x_{jm})}}{\sum_{j=1}^{n_m} \frac{exp(\hat{\alpha}+\hat{\beta}^T x_{jm})}{1+exp(\hat{\alpha}+\hat{\beta}^T x_{jm})}} \times OMR_{all} \text{ for } m = 1, \dots, M$$

5. Iterate steps 1-4 for B bootstrap samples. Estimate adjusted mortality rate by hospital,

$$\hat{\pi}_m = \frac{\sum_{b=1}^{n_m^b} \hat{\pi}_m^b}{B},\ LCL(\hat{\pi}_m) = \hat{\pi}^b_{m_{(0.025\times n_m^b)}},\ UCL(\hat{\pi}_m) = \hat{\pi}^b_{m_{(0.975\times n_m^b)}}$$

where n_m^b is the number of bootstrap samples that are generated for hospital m.

We run this bootstrap algorithm with $M = 60$ and $B = 810$. The value of B is selected so that we observe $n_m^b \geq 500$ for each hospital $m = 1, \dots, M$.

Like the NYSDOH performance measure, the CMS mortality rate estimate in (22) adjusts for the population of patients treated at each particular hospital. The numerator of the performance measure is the predicted total number of events for the particular hospital and is determined through estimates of the risk coefficients (stage 1), prediction of the hospital-specific intercept (stage 2), and the hospital-specific patient covariate values. The denominator of the performance measure reflects the expected total number of events for the particular hospital given its actual patient mix as in the numerator but without any hospital-specific intercept. The hospital-specific ratio $\frac{\sum_{j=1}^{n_m} \hat{\pi}_{jm}(\hat{\theta})}{\sum_{j=1}^{n_m} \frac{exp(\hat{\alpha}+\hat{\beta}^T x_{jm})}{1+exp(\hat{\alpha}+\hat{\beta}^T x_{jm})}}$ represents the performance of the particular hospital relative to the performance of the state as a whole and is interpreted in the same fashion as the ratio $\frac{OMR_m}{\frac{1}{n_m}\sum_{j=1}^{n_m} \hat{\pi}_{jm}}$ in (21). Through the prediction of the hospital-specific random effect, the hospital-specific ratio in (22) is closer to one and has lower standard error than the hospital-specific ratio in (21) for each $m = 1, \dots, M$. The difference is greater for low volume hospitals compared to high volume hospitals.

Kalbfleisch and Wolfe (2013) suggest a modification to the CMS approach whereby the standardized mortality rate is based on stage 1 estimates of the hospital-specific estimate rather than the stage 2 prediction of δ_m. To date, this work has been done for linear models only and so is not applicable for the problem at hand.

Estimation by weighted estimating equations

We use the WEE approach since some sample sizes by hospital are small and we expect that the mean mortality rates by hospital drift over time in an unpredictable way. In Section 1.3, we list two possible standard populations of interest for this application. The population of patients across all hospitals at the present time is relevant for comparing estimates across hospitals. In 2012, there are 47,045 patients across the 60 hospitals with the number of patients by hospital given in Figure

6-2. Because of the number of covariates, it is not practical to display all of their covariate values here. Figure 6-3 gives the distribution of the age of the patients in this standard population.

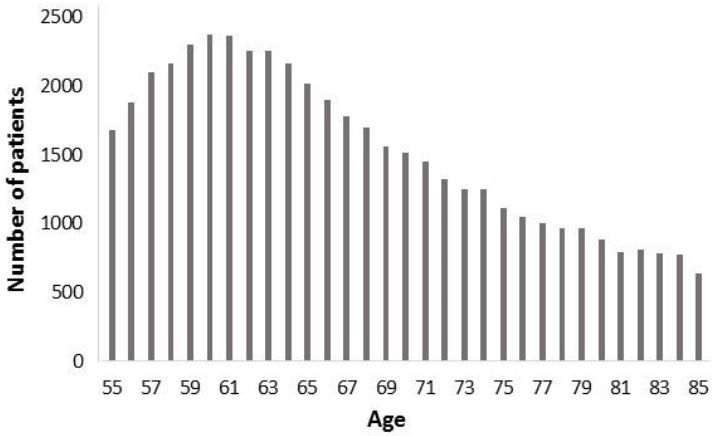


Figure 6-3. Age distribution of PCI patients in 2012

Given the covariate values of the PCI patients in 2012, we define the standard population $\left\{x_{j^*} = \left(x_{1,j^*}, \dots, x_{15,j^*}\right)^T \text{ for } j^* = 1, \dots, 47{,}045\right\}$.

Table 1-1 introduces the GLM that is selected for this problem based on π_m, the binomial mortality rate following surgery in hospital m for $m = 1, \dots, 60$. We assume that the random variables are independent across $t = 1, \dots, T$, conditional on the values of the covariates. The mean mortality rate at the baseline hospital (hospital 1) for a patient with the baseline level of the covariates (see Section 1.4) is modelled by α. The mortality rates of the other hospitals relative to the baseline hospital are modelled by $\delta_1, \delta_2, \dots, \delta_{59}$. The effects of the values of the covariates relative to the baseline values are modelled by $\beta_1, \beta_2, \dots, \beta_{15}$. For patient j tested at hospital m at time t, the binomial mortality rate relates to the model parameter $\theta_t = \left(\alpha_t, \delta_{1,t}, \delta_{2,t}, \dots, \delta_{59,t}, \beta_{1,t}, \beta_{2,t}, \dots, \beta_{15,t}\right)^T$ through the inverse link function

$$\pi_{jmt}(\theta_t; d_t) = \frac{\exp\left(\alpha_t + \delta_{1,t} I_m[1] + \delta_{2,t} I_m[2] + \cdots, \delta_{60,t} I_m[59] + \beta^{\mathrm{T}} x_{jmt}\right)}{1 + \exp\left(\alpha_t + \delta_{1,t} I_m[1] + \delta_{2,t} I_m[2] + \cdots, \delta_{60,t} I_m[59] + \beta^{\mathrm{T}} x_{jmt}\right)} \tag{23}$$

where I_m is a 59-dimensional vector with elements that are either 0 or 1 depending on the hospital that the patient attended and $I_m[i]$ is the i^{th} element of I_m. We expect that levels α_t and δ_t may change slowly over $t = 1, \dots, T$ since there may be a drift in surgical quality at the hospitals. The observed outcome for patient j at hospital m at time t is $y_{jmt} = 1$ if the patient experiences death during the same hospital stay in which he/she underwent PCI or after hospital discharge but within

30 days of surgery and $y_{jmt} = 0$ otherwise. The log-likelihood function describing the probability of data $d_t = \{y_{jmt}, j = 1, \dots, n_{mt}, m = 1, \dots, 60\}$ including all observations at time period t is

$$l_t(\theta_t; d_t) = \sum_{m=1}^{60} \sum_{j=1}^{n_{mt}} \left(I\left[y_{jmt} = 1\right] \log \pi_{jmt} + I\left[y_{jmt} = 0\right] \log\left(1 - \pi_{jmt}\right) \right) \qquad (24)$$

for indicator variables $I\left[y_{jmt} = 0\right]$ and $I\left[y_{jmt} = 1\right]$.

We select weights $\{w_t, t = 1, \dots, T\}$ by (10) with the value of the weight parameter $\lambda = 0.5$. We choose a higher value of λ for this application since the exploratory analysis in Figure 6-1 shows that there is some noticeable change in the observed mortality rate over the yearly time intervals. The effect of alternatives for λ is discussed in Section 6.2. The WEE estimates will be compared to the estimates of the naïve approach using the two special cases of the weights described in Section 3.1.

Under (3), the weighted estimating function vector of length 75 is

$$Q(\theta; d, w) = \sum_{t=1}^{T} w_t \psi_t(\theta; d_t) = \begin{bmatrix} \sum_{t=1}^{T} w_t \sum_{m=1}^{60} \sum_{j=1}^{n_{mt}} I\left[y_{jmt} = 1\right] - \pi_{jmt} \\ \sum_{t=1}^{T} w_t \sum_{j=1}^{n_{2t}} I\left[y_{j2t} = 1\right] - \pi_{\mathrm{j2t}} \\ \vdots \\ \sum_{t=1}^{T} w_t \sum_{j=1}^{n_{60,t}} I\left[y_{j,60,t} = 1\right] - \pi_{\mathrm{j,60,t}} \\ \sum_{t=1}^{T} w_t \sum_{m=1}^{61} \sum_{j=1}^{n_{mt}} x_{1,jmt} \left(I\left[y_{jmt} = 1\right] - \pi_{jmt} \right) \\ \vdots \\ \sum_{t=1}^{T} w_t \sum_{m=1}^{61} \sum_{j=1}^{n_{mt}} x_{15,jmt} \left(I\left[y_{jmt} = 1\right] - \pi_{jmt} \right) \end{bmatrix} \qquad (25)$$

given the present time value of the parameter, $\theta = \theta_T = (\alpha, \delta_1, \dots, \delta_{59}, \beta_1, \dots, \beta_{15})^T$, data $d = \{d_t, t = 1, \dots, T\}$, weights $w = \{w_t, t = 1, \dots, T\}$, and inverse link function $\pi_{jmt}(\theta_t; d_t)$.

The WEE estimate $\hat{\theta}$ is the solution of $Q\left(\hat{\theta}; d, w\right) = 0$. Through (11), we estimate the weighted information estimate of variance, $\widehat{var}\left(\hat{\theta}\right)$ involving $I_t\left(\hat{\theta}\right)$, the expected information function at each time period evaluated at the WEE estimate. With estimate $\hat{\theta}$, we compute the estimate of the mortality rate $\hat{\pi}_{j^*m}$ for each of the standard population patients $j^* = \{1, \dots, 47{,}045\}$, given $\{x_{j^*}\}$, at each of the hospitals $m = 1, \dots, 60$ as in Table 2-1. Similarly, with estimate $\widehat{var}\left(\hat{\theta}\right)$, we compute estimate $\widehat{var}\left(\hat{\pi}_{j^*m}\right)$ through (5) and estimate $\widehat{var}(\hat{\pi}_m)$ through (4).

Results

The exploratory analysis in Figure 6-1 shows that PCI patient mortality rates change over time and sample size in the latest time period is smaller than most other time periods. There are

differences in the mortality rates across hospitals since their patients have different risk factors and the surgical quality varies across hospitals. The objective is to estimate the present time mortality rate by hospital for a standard population of patients with a bias/variance trade-off so that stakeholders can use the measure for the functions described in Section 6.1.

Figure 6-4, Figure 6-5, and Figure 6-6 give the estimates of mortality rate by hospital and 95% confidence intervals of these estimates based on the WEE approach assuming normality and the two industry practices discussed previously. Note that the estimates in Figure 6-4 are based on data $\{d_t, t = 2004, \dots, 2012\}$ over a nine-year period and the estimates in Figure 6-5 and Figure 6-6 are based on data $\{d_t, t = 2010, \dots, 2012\}$ over a three-year period as per industry practices. The format of the graphs is like those in the NYSDOH annual reports which show the estimates and 95% confidence intervals assuming normality. The horizontal line on each graph is the overall observed mortality rate for all patients in NYS over the time period of the data.

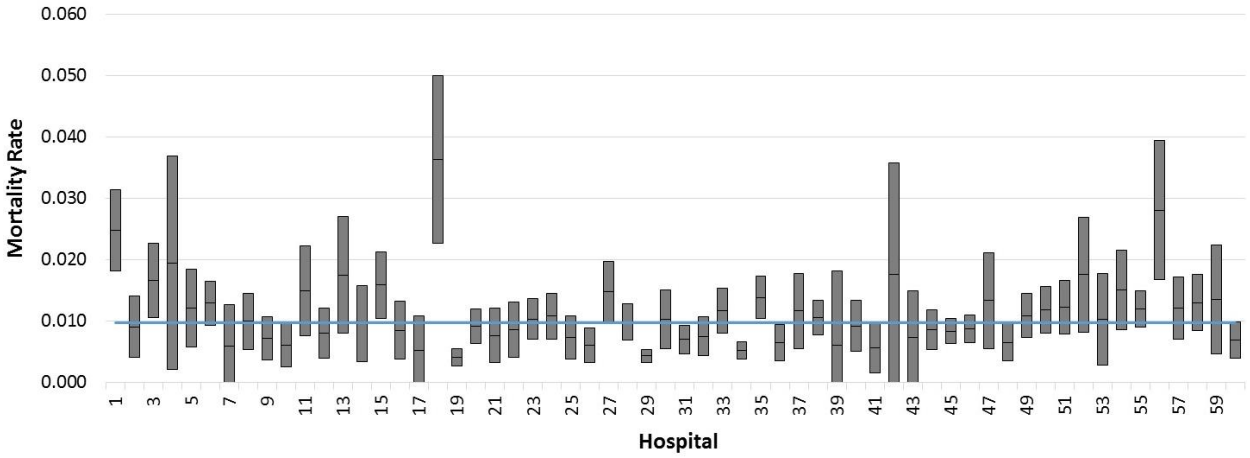


Figure 6-4. WEE estimates of 2012 mortality rates by hospital (λ=0.5)

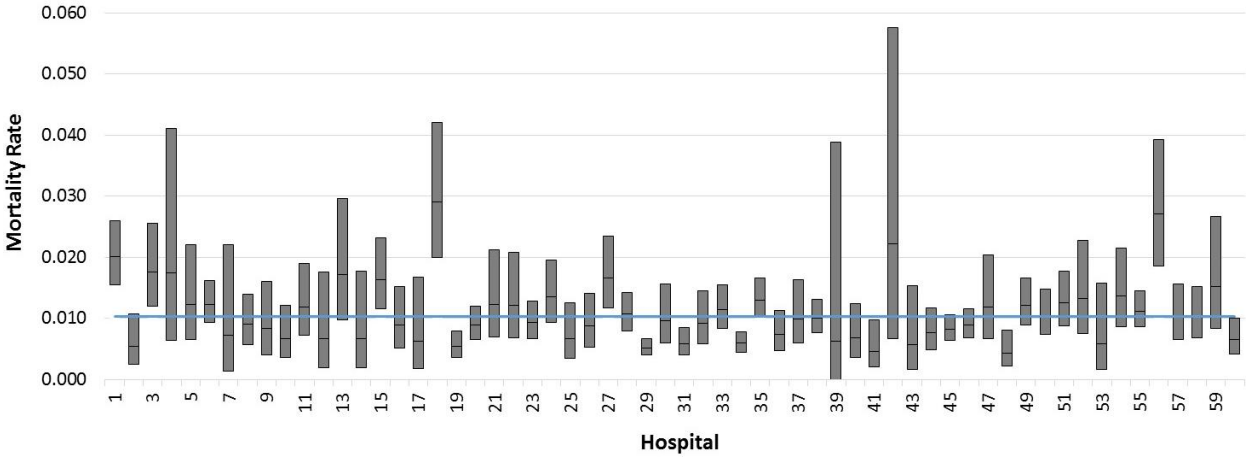


Figure 6-5. NYSDOH estimates of 2012 mortality rates by hospital

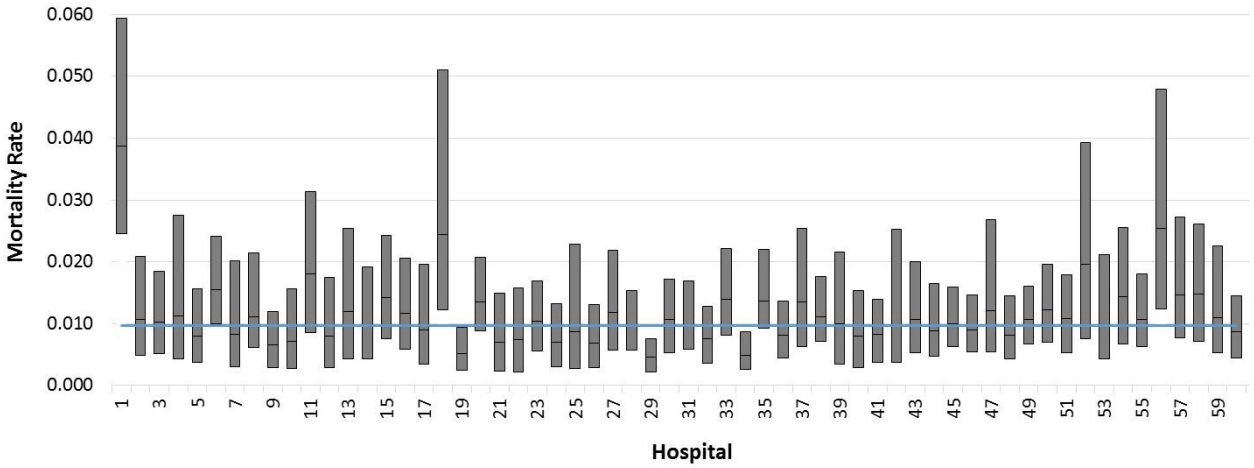


Figure 6-6. CMS estimates of 2012 mortality rates by hospital

Figure 6-4, Figure 6-5, and Figure 6-6 show some important differences between the estimates by the various approaches. We compare the precision, bias, and the suitability of these estimates to serve their intended functions.

Comparison of precision

We compare the widths of the confidence intervals in Figure 6-4, Figure 6-5, and Figure 6-6 by hospital. For 32 of the 60 hospitals, the widths of the confidence intervals based on the WEE approach are narrower than those based on the NYSDOH approach. For 54 of the 60 hospitals, the widths of the confidence intervals based on the WEE approach are narrower than those based on the CMS approach. In particular, note the differences in the widths of the confidence intervals for those hospitals with exceptional performance. For each hospital with no deaths in 2012 (ref. 7, 9, 12, 21, 25), the width of the confidence interval of the estimate by the NYSDOH or CMS approaches is around 38% larger than that by the WEE approach. For the three hospitals having the highest mortality rates based on the WEE approach (ref. 1, 18, 56), the widths of the confidence intervals for these hospitals in Figure 6-6 support the Kalbfleisch and Wolfe (2013) claim that CMS estimates of exceptional performance have poor precision. The effect of borrowing strength from the historical data through the WEE approach when there is little statistical information in the present time period data is a more precise estimate of present performance.

Comparison of bias

The trade-off for improved precision in WEE estimates is added bias when performance changes over time. We are unable to quantify bias in this analysis since we do not know the true

value of the parameter, but we presume that the WEE estimate based on present time data only is the closest WEE estimate to the present true value. The WEE weight parameter $\lambda = 0.5$ is used in this application and so the data from the most recent time period provides roughly 50% of the weight within the estimating function. The influence of data from previous time periods reduces quickly for time periods further in the past. We look at the mortality rate estimate for hospital 1. The naïve estimate of mortality rate at this hospital using present data only (WEE with $\lambda \to 1$) is $\hat{\pi}_{1(t=T)} = 0.042$ (LCL=0.028, UCL=0.056) and using all data from 2004 to 2012 weighted equally (WEE with $\lambda \to 0$) is $\hat{\pi}_{1(t\leq T)} = 0.012$ (LCL=0.0095, UCL=0.014). The WEE estimate with $\lambda = 0.5$ is $\hat{\pi}_{1(WEE)} = 0.025$ (LCL=0.018, UCL=0.031). Under the presumption stated previously, $\hat{\pi}_{1(WEE)}$ has less bias than $\hat{\pi}_{1(t\leq T)}$ since it is closer to $\hat{\pi}_{1(t=T)}$. The influence of past data results in a significantly lower WEE estimate of mortality rate compared to the naïve estimate using present data only since mortality rate at hospital 1 increases over time. Note from Figure 6-5 and Figure 6-6 that there are significant differences between the estimates by the NYSDOH and CMS approaches for this hospital and the WEE estimate is between the other two. For slower changes over time, the differences between the estimates by the various approaches will be smaller. In Section 6.2, we discuss the impact of the choice of the weight parameter on the trade-off between bias and variance.

Comparison by performance measure function

When the mortality rate estimate is used to compare performance to target and to decide which hospitals to inspect, the differences in precision and bias of the estimates by the various approaches affect the outcomes. We see in Figure 6-4, Figure 6-5, and Figure 6-6 that the WEE approach identifies four hospitals with significantly worse performance than the overall mean that are not identified by the CMS approach (ref. 3, 15, 28, 36). The CMS approach is the least sensitive approach for identifying outlying hospitals because of the shrunken hospital effect predictions.

We consider the suitability of the estimates by the various approaches to monitor performance over time. Figure 6-7 gives the estimates and 95% confidence intervals of mortality rate made each year by the three approaches for a particular hospital (ref. 3). The WEE estimates over time are mean mortality rate estimates for the same standard population which is the population of patients in 2012.

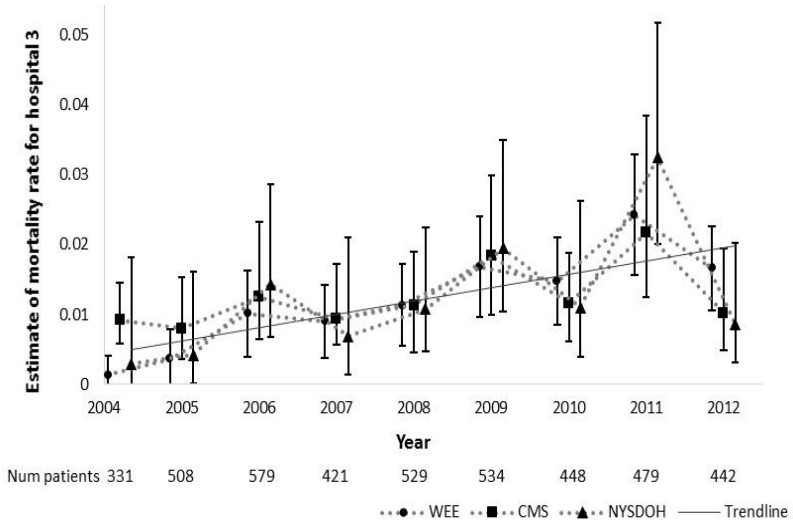


Figure 6-7. Estimates of mortality rate for hospital 3 by various approaches over time

Figure 6-7 shows that there is an increasing trend in mortality rate estimates over the period from 2006 to 2012 at hospital 3. Around this trend, the estimates by the NYSDOH approach fluctuate the least across time periods. The estimates by the CMS and WEE approaches highlight a significantly lower mortality rate in 2012 compared to 2011 that is not detected by the NYSDOH approach. Based on this example, we see that the WEE is a more sensitive approach than the NYSDOH approach at detecting changes over time.

Considering precision and bias of the estimates, the ability to identify hospitals with exceptional performance, and sensitivity to changes over time, the WEE approach has advantages over the CMS and NYSDOH approaches for the realistic PCI in NYS dataset. The adoption of the WEE approach could have an important impact on the functions of the hospital surgical quality performance stakeholders.

6.2. Implementation of the WEE approach

Through a realistic dataset, we demonstrate that the WEE approach to estimate the present surgical quality performance measure has advantages over the two current industry practices. The WEE approach offers a trade-off between estimation using present time data only or all historical data weighted equally in the analysis of temporal data. This trade-off is especially important when sample sizes at some time periods may be small and a parameter describing the mean outcome changes slowly over time. There are numerous other applications where the WEE approach should be considered for improving estimates. In so doing, deliberate thought is required on certain

aspects of this approach where characteristics of the particular dataset and available knowledge are important. Three aspects are discussed: selecting time subgroups and the weight parameter, estimating covariate effects, and handling missing data and sampling zeros.

Selecting time subgroups and weight parameter

In Section 1.2, we present the data $d = \{d_t\}$ as observations over $t = 1, \dots, T$ time periods. Depending on the application, the data may arise in an ongoing manner or in collections of observations at distinct time intervals. For example, a company that studies data from a customer survey may receive on-line customer responses on a daily basis. Choosing an appropriate time interval to define the subgroups is important; for example, responses may be grouped by day, week, or month. Principles of rational subgrouping from statistical quality control literature (Montgomery, 2013) intend to minimize within subgroup variation and maximize between subgroup variation. Similar objectives should be considered when subgrouping the data for implementing the WEE approach so that the true value of the parameter to be estimated changes slowly across the defined time periods. Naturally, sample sizes by time period depend on the choice of the time interval for subgroups and in general, we expect that small samples in some time periods may occur.

In (10), we present a formula to calculate exponentially declining weights as a function of a weight parameter λ taking a value between 0 and 1. A larger λ value increases the weight given to the most recent data in the weighted estimating functions. The choice of λ regulates the bias-variance trade-off. In general, a larger λ reduces bias and a smaller λ reduces uncertainty. The appropriate selections of time subgroups and weight parameter λ are related. For example, we may be able to subgroup a set of data by week or by month. The sample sizes in subgroups by week are smaller and so uncertainty in the estimate is more of a concern. We should select a smaller λ to reduce uncertainty. By contrast, if we subgroup the data monthly rather than weekly, then uncertainty is less of a concern and we can increase the value of λ somewhat.

We demonstrate the relationship between the selections of time intervals and λ and their effect on the WEE estimates through the realistic PCI dataset. In Section 6.1, we present the results for the data in yearly subgroups. Next, we consider the situation where the month of the surgery is also available and it is possible to update the analyses at monthly intervals. To illustrate the impact of this alternative, since we don't have the actual data, we assign months randomly for each patient within the year that his/her surgery took place. Figure 6-8 gives the number of patients who underwent PCI in each of the latest 15 months, the observed mortality rate over this time period, and the linear trend line in mortality rate.

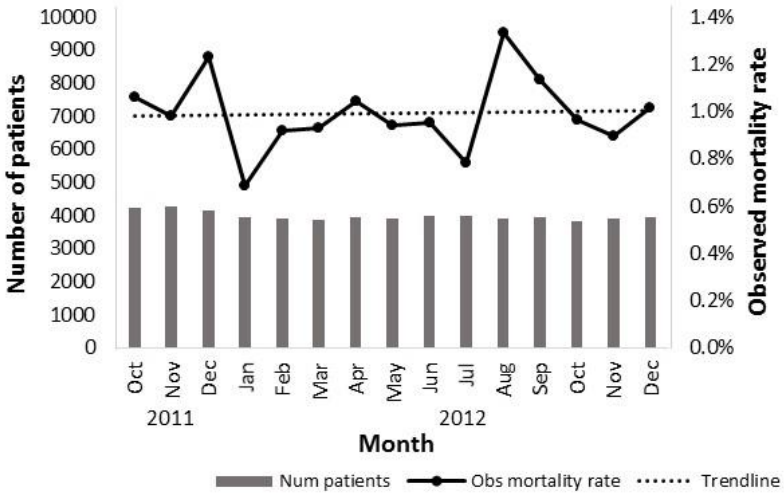


Figure 6-8. Observed mortality rate of PCI patients over latest 15 months

Comparing Figure 6-8 to Figure 6-1 we see that the average rate of change by time period in observed mortality rate based on data in monthly subgroups is slower than that based on data in yearly subgroups. A smaller value of λ is appropriate when implementing the WEE approach based on monthly data since the uncertainty resulting from a small sample in the latest time period is more of a concern than the bias resulting from combining data across time periods. In Section 3.2, we discuss the notion of effective sample size and show for a binomial model without covariates that $N_{eff} = \frac{\left(\sum_{t=1}^{T} w_t n_t\right)^2}{\sum_{t=1}^{T} w_t^2 n_t}$ for the WEE estimator. Figure 6-9 gives N_{eff} versus the value of λ for the monthly and yearly sample sizes in the realistic PCI dataset.

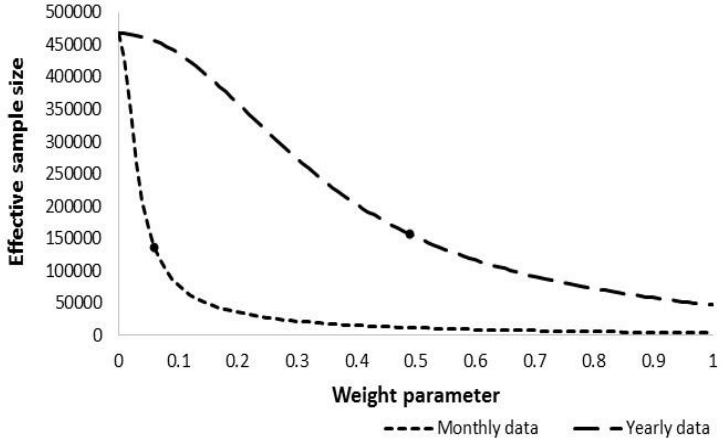


Figure 6-9. Effective sample size depending on λ and data subgrouping for the PCI dataset

Figure 6-9 shows that the maximum value of N_{eff} is 467,401 which is the total number of observations in the dataset and occurs when $\lambda \to 0$ for data in both monthly and yearly subgroups. At the other extreme, when $\lambda \to 1$, $N_{eff} = 47{,}045$ for the data in yearly subgroups which is the sample size in the latest year (2012) and $N_{eff} = 3931$ for the data in monthly subgroups which is the sample size in the latest month (December 2012). The marker on the yearly N_{eff} curve is the value at $\lambda = 0.5$ as selected in Section 6.1. The marker on the monthly N_{eff} curve has a similar value of N_{eff} which occurs when $\lambda = 0.06$. We select a value of $\lambda = 0.06$ for WEE analysis of the data in monthly subgroups which is considerably smaller than $\lambda = 0.5$ that is used in the WEE analysis of the data in yearly subgroups. Since the two analyses have similar values of N_{eff}, then we expect the uncertainties of the resulting estimates to be similar. Figure 6-10 gives the WEE estimates of mortality rate by hospital and 95% confidence intervals of these estimates assuming normality based on the analysis of data in monthly subgroups.

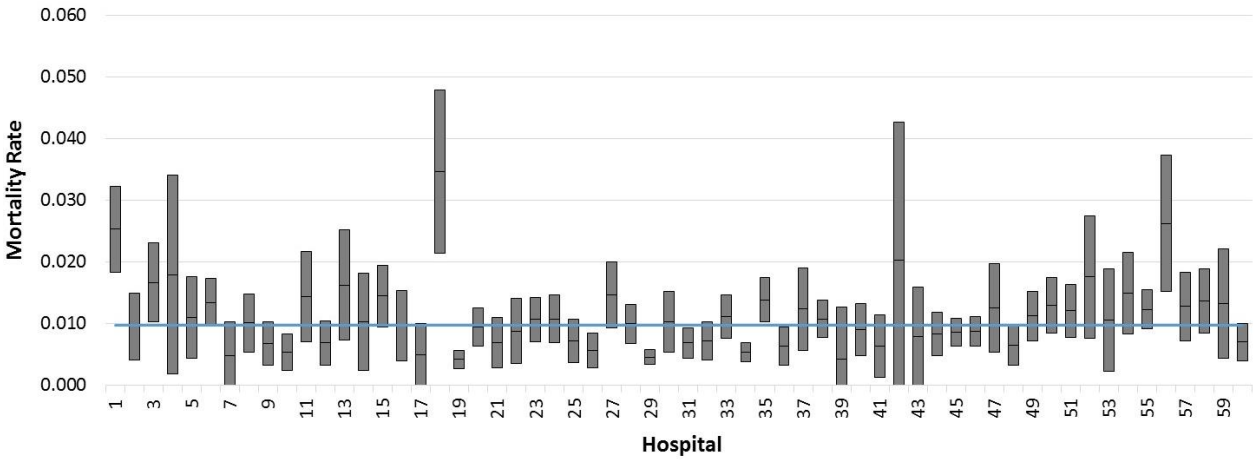


Figure 6-10. WEE estimates of 2012 mortality rate by hospital based on monthly data ($\lambda = 0.06$)

Figure 6-10 shows that the WEE estimates and confidence intervals based on the data in monthly subgroups are comparable to those based on the data in yearly time intervals in Figure 6-4. In a monitoring problem, we could make a graph such as Figure 6-7 based on WEE estimates updated monthly. Updating the WEE estimates more frequently should detect performance changes sooner.

As discussed, the weight parameter λ regulates the bias-variance trade-off in the WEE estimate of the performance measure. We demonstrate this trade-off through a particular hospital (ref. 1) where mortality rate changes are relatively fast over time compared to that of the hospitals on average. Figure 6-11 gives the number of patients and observed mortality rates at hospital 1 over time based on yearly data.

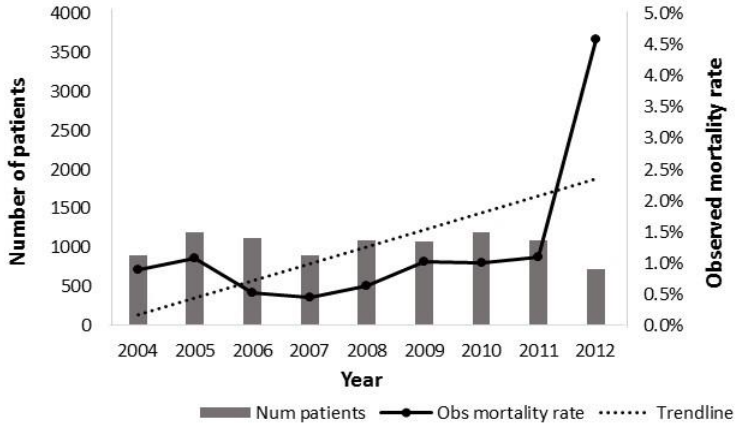


Figure 6-11. Sample sizes and observed mortality rates of PCI patients at hospital 1 over time

Figure 6-11 shows that a large change occurred at this hospital in the most recent time period. Note also that there were the fewest number of patients at this hospital in 2012. Figure 6-12 gives the WEE estimates and 95% confidence intervals based on this data with various values of weight parameter λ.

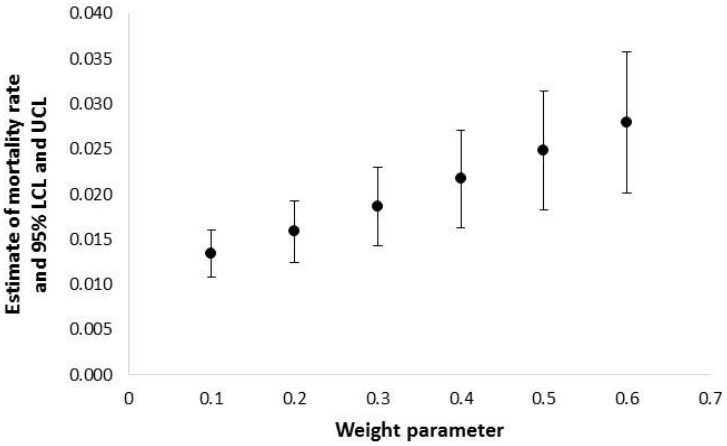


Figure 6-12. WEE estimates of 2012 mortality rate for hospital 1 with various λ

Figure 6-12 shows that the bias-variance trade-off in the WEE estimate of 2012 mortality rate at hospital 1 relates to the weight parameter. As λ increases, there is more uncertainty in the WEE estimate but we presume that the estimate is closer to the true present performance. For a particular problem of interest, similar analyses across various selections of λ provide an understanding of the bias/variance trade-off. Where WEE estimates are similar for various selections of λ, as they are when performance changes slowly over time, then a relatively small value of λ should be

chosen to minimize uncertainty. We have found the value of $\lambda = 0.1$ to be appropriate under this condition.

6.3. Summary and discussion

Estimation of a surgical performance measure is important for healthcare monitoring and regulation. Current industry practices pool patient outcome data over three years to reduce uncertainty since the sample size in the current year may be small. They adjust for observed incoming patient health characteristics since various hospitals treat patients with different surgical risks. The estimates involve a generalized linear binomial model with fixed or random covariate and hospital effects. We propose the WEE approach as an alternative that similarly incorporates past data, adjusts for patient risk, and involves a binomial model. The key advantage of the WEE approach is that similar data from time periods further in the past are used to improve performance estimates while managing the added bias when performance changes slowly over time. When the precision and bias in the estimates of the performance measure are improved, we can compare performance to target, screen performance to decide which hospitals to inspect, and monitor performance for arising problems with more sensitivity and reliability.

We discuss considerations relative to the implementation of the WEE approach for the hospital performance measure. The ability to specify a weight parameter within the WEE approach allows a choice to be made in selecting the time span of the data subgroups. If the data are available monthly, estimation through data in monthly intervals is practical through the WEE approach so that changes in performance are detected much quicker than through analyses based on a pooled three-year set of observations. Further, estimates of covariate effects are more precise since they are based on a larger set of data. In the WEE approach, covariate effects need not be updated at every time period which can simplify the analyses. With any approach, considerations need to be made for missing data and sampling zeros, but their impact will have less effect on estimates by the WEE approach than the other approaches discussed.

Estimating covariate effects

In this work, we assume that the true value of parameter vector θ_t may change slowly over time; however, it is possible that the covariate effects may either be known or assumed to be fixed over time. In the case where elements of θ_t are known, we can substitute the known values and reduce the number of estimating functions appropriately. In the case where elements of θ_t are assumed to be fixed over time ("fixed parameters"), then the lowest uncertainty estimates of these

elements are based on data across all time periods weighted equally. It is reasonable in this case to separate the estimation of the fixed parameters from the estimation of the remaining parameters that change slowly over time through a two-stage approach to estimation. In Section 7.2, we discuss the two-stage approach and two other alternatives to the WEE formulation when there are fixed parameters which we intend to explore as future work. Under the two-stage approach, the estimation of the fixed parameters may happen less frequently than the estimation of the remaining parameters since we expect that changes in the fixed parameter estimates are small. For example, in the monthly update of WEE estimates, the estimates of the fixed parameters may be updated annually.

Missing data and sampling zeros

Minimum levels of quality and completeness of the data are important in order that the estimates of hospital performance are useful. At the NYSDOH, data are verified through review of unusual reporting frequencies, cross-matching of data with other Department of Health databases, and a review of medical records for a selected sample of cases. These activities are extremely important to ensure consistent interpretation of data elements across hospitals and across time periods. There are various reasons that some patient cases may be excluded from analysis; for example, patients that reside outside of NYS and those who are determined to be at extremely high-risk for death preceding surgery are removed from the PCI in NYS analysis. A further requirement is that covariate and outcome data are reported consistently over time. In the NYSDOH analysis, a change was made in 2004 to include deaths that occurred outside of the hospital and within a 30-day period from the date of surgery in addition to in-hospital deaths. Further, a patient's pre-procedural myocardial infarction was reported as one of five possible levels in 2007, whereas in 2012 there were eight possible levels and only two of these also had appeared in 2007. Adjustments need to be made so that there are consistent definitions of the outcome and the covariates within the dataset used to estimate present mortality by any of the three approaches discussed in this paper. These are important practical considerations.

In any particular application, there may be missing data or sampling zeros that should be considered. Relative to the hospital performance problem, we discuss implementation of the WEE approach under the following conditions:

- a new hospital is added or removed from the dataset during the time period of the data
- some patient level covariate data are missing
- outcome is not reported for some patients

- there are zero patient deaths observed at a particular level of the categorical covariates

In the PCI in NYS dataset, 13 new hospitals began performing PCI and one hospital closed during the study period from 2004 to 2012. Data on patients from these hospitals do not appear for one or more of the time intervals either at the beginning or the end the study period. The statistical literature refers to this type of data as monotone missing data (Little and Rubin, 2002). We consider the simpler case where there are no covariates to be estimated in the model. When there are no data on patients at a particular hospital past time t_0, as in the case where a hospital closes or ceases the surgery of interest, the WEE estimate of the particular hospital effect does not change for any $T \geq t_0$. This observation holds under exponentially declining weights and assuming that the beginning time period of the data remains fixed. Similarly, when there are no data on patients at a particular hospital before time t_0, as in the case where a new hospital is added, the WEE estimate of that particular hospital effect based on the dataset $\{d_t; t \in [t_o, T]\}$ is the same as that based on any of the larger datasets $\{d_t; t \in [t', T], t' < t_0\}$. The same properties hold in the case where there are covariates but only one hospital effect to estimate. In the more general case where more than one hospital effect and covariate effects are estimated, the WEE estimate for a hospital effect differs depending on the number of time periods where there are no data from that hospital. This occurs because the estimates for the covariates and their uncertainties change as we include more time periods and the hospital effects are estimated simultaneously with covariate effects. The WEE estimates for a particular hospital change as the amount of monotone missing data for that hospital increases, but are better estimates since the estimates of the covariate effects under the fixed effects assumption are based on more data. In the case where a hospital no longer performs the particular surgery, then the data from this hospital may be included in the analysis in order to improve the estimates of the covariate effects; however, since the WEE approach intends to provide an estimate of present performance, then estimates of this hospital effect are no longer relevant.

Another type of missing data occurs when some data are missing on patient-level values of the covariates or patient-level outcomes. Typically this occurs if the data collection processes are inconsistent across hospitals or over time. In SAS and other analysis procedures, the default approach to deal with this missing data is to delete the incomplete cases from the analysis. There is a large body of literature on missing data mechanisms and strategies for dealing with datasets collected over time having missing values (Little and Rubin, 2002; Colosimo, Fausto, Freitas, and Pinto, 2012; Jansen, Beunckens, Molengerghs, Verbeke, and Mallinckrodt, 2006). At the onset of an analysis, it is important to investigate missing-data patterns and mechanisms that lead to missing data. In the context of the hospital problem, it is important to investigate whether

missingness is related to the performance of the hospital where the patient attended and apply the appropriate classification: missingness at random (MAR), completely at random (MCAR), or not at random (NMAR). The classification guides the selection of a procedure to deal with the missingness which are broadly grouped into imputation-based methods, model-based methods, and weighting procedures. In the case of MAR data, then an imputation-based method is recommended within the WEE approach since the imputation procedure occurs separately from the parameter estimation procedure. Imputation is preferred over ignoring incomplete cases. Yuan (2000) presents SAS procedures for creating multiple imputations for incomplete multivariate data. Under the MCAR assumption, the missing data values are a simple random sample of all data values. Here, analysing only the complete cases is an acceptable approach. Further work is needed to recommend a procedure for handling NMAR data within the WEE approach.

As in broader categorical data analysis, there are difficulties when there are sampling zeros in the observed data (Agresti, 2007). A sampling zero occurs when all patients having a particular level of a categorical covariate are observed to have the same outcome. Infinite WEE estimates occur for that covariate level. In the hospital application, this occurs when there are no deaths among patients at any hospital at one particular level of the categorical covariate across time. Some software programs (such as PROC GENMOD in SAS) provide warnings that the fitting process fails when infinite estimates occur. Agresti (2007) asserts that grouping the data into bins by categorical covariate levels and by time and adding a small constant (such as 10^{-8}) to the sampling zero cell count may be adequate for ensuring convergence. One can then estimate parameters for which the true estimates are finite and are not affected by the sampling zeros. Sensitivity analysis is recommended to investigate the impact of this change to the data. Another approach is to combine levels of the covariate to obtain non-zero counts by outcome value. This is tenable when the covariate data are ordinal or if there is another natural way to combine levels. In the PCI in NYS dataset, the categorical covariates having more than two levels are ordinal. It would be natural to collapse levels of ventricular ejection fraction, pre-procedural myocardial infarction, or renal failure creatinine if necessary. Note that information is lost in defining the variable more crudely but is less detrimental than removing the parameter representing the covariate level effect from the model completely.

The concerns related to missing patient-level covariate or outcome data and sampling zeros exist among all three approaches to the hospital performance problem including the current industry practices. It is important to note however, that the WEE approach involves a larger dataset than the other two approaches and so the instances of sampling zeros are reduced. Further, the

imputation methods for MAR data are more reliable when based on a larger dataset. The WEE approach is less impacted by missing data and sampling zeros than the current industry practices.

Future work

The observations of this chapter are based on a realistic example dataset created to have similar properties as actual outcomes among patients undergoing percutaneous coronary intervention in New York State during the period 2004 to 2012. The limitation of this work is that the observations are based solely on one dataset. Further work is recommended to apply the WEE approach to other health care performance datasets and compare the estimates to current industry practices. We discuss one particular opportunity. The U.S. Scientific Registry of Transplant Recipients (SRTR) provides data on transplant patient outcomes (Scientific Registry of Transplant Recipients Home Page, n.d.). Stakeholders may want to compare risk-adjusted outcomes across transplant centers or donation service areas, across groups of patients with different risk factors, or across time. We expect that transplant outcomes change slowly over time due to factors that are not observed and the number of patients treated in some transplant centers, donation service areas, time periods, or patient risk groups may be small. The current industry practice (Scientific Registry of Transplant Recipients, 2016a) is estimation based on a Cox proportional hazards model of the time to an event such as removal from the waiting list, post-listing death, graft failure, and post-transplant death. The model adjusts for patient, donor, and transplant characteristics. The observed number of events at a particular center is compared to an expected number of events among similar patients based on the model fit to all data available nationally. The analysis is based on data from the most recent year only. The SRTR states that "estimates become unstable as fewer patients are being followed" (Scientific Registry of Transplant Recipients, 2016b) but in their annual reports they offer no statement of uncertainties or discussion of sample size (U.S. Department of Health and Human Services, 2014). Through its inclusion of historical data, the WEE approach could have an important impact on the estimates and the ability to detect differences among groups and changes in outcomes over time. In general, the WEE approach for measuring health care performance deserves further attention.

Chapter 7: Summary, Discussion, and Future Work

This research is motivated by three problems requiring present time estimates of performance. The problems span marketing, diagnostic testing, and healthcare applications. We have a stream of data from different subjects collected over time. In the particular problems under study, there are two or three possible outcomes and we may have data on many covariates in multiple streams. We may want to monitor an estimate of a performance parameter of interest over time or compare the estimates across streams. We expect that the parameter may drift slowly over time in an unpredictable way. Additionally, some sample sizes may be small. Through study of real and simulated datasets, we extend the weighted estimating equations (WEE) approach originally proposed by Steiner and MacKay (2014) to these new application areas and show its benefit for regulating the bias/variance trade-off in the present time estimate of a parameter relative to current industry practices and other alternative approaches.

To meet the objectives of the motivating applications, we require estimates of uncertainty of the WEE estimate and the distribution of a hypothesis test statistic based on the WEE estimate. We derive approximations for these quantities based on asymptotic properties of the score and information functions. Through the motivating applications and a simple analytic example we demonstrate that these approximations are useful under various conditions. We provide SAS code that is convenient for computing the WEE estimate, the estimate of uncertainty, and the hypothesis test statistic.

Within the context of the various applications, we discuss implementation considerations such as selecting the time subgroups, the time window of historical data, and the covariates and considerations for some large sample sizes. We discuss a more precise WEE estimate of the parameter when some covariate effects are known or assumed to be fixed over time. We consider the impact of missing data and sampling zeros and give an argument that the instances of sampling zeros are reduced and imputation methods for missing data are more reliable for the WEE approach relative to current industry practices and other alternative approaches.

We compare estimates based on the WEE approach to current industry practice within each motivating application as well as naïve and EWMA approaches. We suspect that mean squared error of the estimate is not uniformly lower for one approach relative to another, but the quantitative results show that are certainly circumstances where the WEE approach provides better estimates. Through simulation, we show that the WEE estimates have less bias than the naïve

estimate based on all of the historical data and are more precise than the naïve estimate using present time data only. Qualitatively, we highlight that EWMA estimates have more uncertainty, do not use all the information in the data, and are not possible in cases where instances of some covariate levels are not present in some data by time period. We give evidence to show that the WEE can have a substantial impact on the stakeholder's abilities to use the estimates to meet their objectives relative to current industry practices.

7.1. Alternative approaches

Other methods of analyses are possible. Specifically, we could add a temporal component to model the changing nature of the parameter over time. This is a feasible alternative when the change in the parameter follows a regular pattern over time. However, in the applications that motivate this work, the slow drift in the parameter may arise due to changes in many contributing factors and a fixed form of a model to capture its temporal behaviour limits its applicability.

Two reviewers of our work have suggested that the Kalman Filter introduced in Section 2.4 is an alternative to the WEE approach. We consider a qualitative comparison of the two. Both approaches seek to produce an estimate of $\theta = \theta_T$ with greater precision by using both current and past data. Each sacrifices unbiasedness for additional precision if the parameter changes over time.

The system dynamic model of the KF describes the evolution of the state vector (here the parameter) that can be used to estimate the parameter at current time T, given $\hat{\theta}_{T-1}$. In the three motivating applications, we have no such model so it is logical to use the most recent estimate $\hat{\theta}_{T-1}$ to estimate the parameter at time T. We take the weighted average of the two estimates $\hat{\theta}_{T-1}$ and $\hat{\theta}_T$ based on d_T with dynamic weights based on their precision. If the parameter changes over time, then there is a bias in $\hat{\theta}_{T-1}$. If there is a small sample size at time T, then there is large uncertainty in $\hat{\theta}_T$.

Unlike the KF, the WEE approach does not combine the current and past estimates. Instead, it creates an estimating function through the weighted average of the likelihood-based score functions across time with weights that are fixed. Note that the score functions based on d_t, $t = 1, \ldots, T$ are sufficient statistics for the data at each time period and hence contain all of the available information about the parameter. For most models including the nonlinear model used in our example, the KF estimates $\hat{\theta}_1, \ldots, \hat{\theta}_{T-1}$ are not sufficient statistics and hence information is lost by using $\hat{\theta}_{T-1}$ to summarize the historic data.

In terms of computation for the non-linear models considered in this work, both methods require the solution of estimating equations with p unknowns (presuming the KF uses the maximum likelihood estimate at time T) and similar calculations to find the standard errors. The WEE approach is motivated by applications with small samples in the latest time period. If there are insufficient data at current time T to estimate all the parameter components, then a standard implementation of the KF is not applicable. The standard KF implementation could be adapted but it is not obvious how to proceed. With small amounts of data and no system dynamic model, present data has less impact on the KF estimate as time goes by and so bias in $\hat{\theta}_{T-1}$ is important to consider. In the real customer loyalty dataset used in Chapter 4, there are insufficient data to estimate all of the parameter components in 20 of the 42 time periods where data are observed and no obvious system dynamic model. As a result, a standard implementation of the KF is not reliable for updating the customer loyalty measure estimates over time.

It is not easy to quantitatively compare the performance of the two approaches through a simulation study since there are many possible parameter and covariate values and ways that the parameter might change over time. We suspect that one approach is not uniformly better than the alternatives; however, the qualitative comparison points at WEE as the more flexible approach for the estimation problem at hand. Additionally, to implement a change to the current industry practices, decision makers need to be made aware of the reason for the change and the basic premise of the new approach. The WEE approach is an intuitive solution to the bias/variance trade-off problem.

7.2. Future work

In Section 6.3, we discuss the opportunity to apply the WEE approach to the U.S. Scientific Registry of Transplant Recipients (SRTR) data on transplant patient outcomes. The current industry practice estimates the time to an event such as post-transplant death based on data from the present year only and states that sample sizes may be small and thus estimates may be unstable. Future work will apply the WEE approach to real SRTR data in order to compute more stable estimates. This will extend the present application of the WEE approach to a time to event likelihood model involving censoring. Further, we will look for applications that involve a continuous outcome measure in order to extend the application of the WEE approach to a broader class of problems.

Weighted estimating equation alternative formulations

Through solving the weighted estimating equations relating to the weighted estimating functions in (3) we obtain estimate $\hat{\theta}$ for model parameter $\theta = \theta_T$ under a model where θ_t does not change over time $t = 1, \dots, T$. The general problem of this research considers that there may be a slow change in the parameter over time and so we know that $\hat{\theta}$ is a biased estimate which we tolerate in order to regulate a bias/variance trade-off. It may be the case that we expect some elements of θ drift over time ("time-varying parameters"), but others remain fixed ("fixed parameters"). In the hospital performance application, we may expect that the mean mortality, α, and hospital effects, δ, are time-varying parameters and a covariate effect such as the effect of age on mortality, β_1, is a fixed parameter. In the future, we intend to evaluate various alternatives to the standard WEE formulation that reduce the bias of estimates of the time-varying parameters when there is one or more fixed parameters. We outline three intuitive alternatives when some (but not all) elements of θ_t, $t = 1, \dots, T$, are fixed parameters. We consider a GLM involving a mean level parameter, α, and two parameters describing covariate effects, say β_1 and β_2. We assume that α and β_1 may change slowly and β_2 is fixed over time.

The first alternative to the vector of weighted estimating functions in (3) is to remove the weights from the estimating function(s) related to the fixed parameter(s). In the example under consideration, there are three elements in the estimating function vector which we refer to by the parameter that is involved through the partial derivative. We remove the weights from the estimating function related to β_2 and assign exponentially declining weight values as in (3) for those estimating functions related to α and β_1. The estimating function vector is then

$$Q(\theta; d, w) = \begin{bmatrix} w_1 \frac{\partial l_1(\theta;d_1)}{\partial \alpha} + w_2 \frac{\partial l_2(\theta;d_2)}{\partial \alpha} + \cdots + w_T \frac{\partial l_T(\theta;d_T)}{\partial \alpha} \\ w_1 \frac{\partial l_1(\theta;d_1)}{\partial \beta_1} + w_2 \frac{\partial l_2(\theta;d_2)}{\partial \beta_1} + \cdots + w_T \frac{\partial l_T(\theta;d_T)}{\partial \beta_1} \\ \frac{\partial l_1(\theta;d_1)}{\partial \beta_2} + \frac{\partial l_2(\theta;d_2)}{\partial \beta_2} + \cdots + \frac{\partial l_T(\theta;d_T)}{\partial \beta_2} \end{bmatrix}$$

where $\{w_t\}$ are exponentially declining weights. We solve the estimating equation for $\hat{\theta} = \left(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2\right)^T$ as usual.

A second alternative to the standard WEE formulation when we assume that β_2 is a fixed parameter is to use the history of estimates of the time-varying parameters α and β_1. As in the previous alternative, we remove the weights from the estimating function related to β_2 but here we substitute the previous estimates $\hat{\alpha}_t$ and $\hat{\beta}_{1,t}$ from time periods $t = 1, \dots, T-1$. The estimating

functions related to α and β_1 use exponentially declining weight values as in (3). Then, the vector of weighted estimating functions is

$$Q(\theta|d, w, \hat{\alpha}_1, \ldots, \hat{\alpha}_{T-1}, \hat{\beta}_{1,1}, \ldots, \hat{\beta}_{1,T-1}) = \begin{bmatrix} w_1 \frac{\partial l_1(\theta;d_1)}{\partial \alpha} + w_2 \frac{\partial l_2(\theta;d_2)}{\partial \alpha} + \cdots + w_T \frac{\partial l_T(\theta;d_T)}{\partial \alpha} \\ w_1 \frac{\partial l_1(\theta;d_1)}{\partial \beta_1} + w_2 \frac{\partial l_2(\theta;d_2)}{\partial \beta_1} + \cdots + w_T \frac{\partial l_T(\theta;d_T)}{\partial \beta_1} \\ \frac{\partial l_1(\theta;d_1)}{\partial \beta_2}\Big|_{\substack{\alpha=\hat{\alpha}_1 \\ \beta_1=\hat{\beta}_{1,1}}} + \frac{\partial l_2(\theta;d_2)}{\partial \beta_2}\Big|_{\substack{\alpha=\hat{\alpha}_2 \\ \beta_1=\hat{\beta}_{1,2}}} + \cdots + \frac{\partial l_T(\theta;d_T)}{\partial \beta_2} \end{bmatrix}$$

where $\{w_t\}$ are exponentially declining weights. We solve the estimating equation for $\hat{\theta} = \left(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2\right)^T$ as usual.

A third alternative to the standard WEE formulation when we assume that β_2 is a fixed parameter is a two-stage approach to estimation. In stage 1, we estimate β_2 with an unweighted estimating function and in stage 2, we estimate α and β_1 with a weighted estimating function using the estimates from stage 1 as known quantities. In stage 1, we need initial values of the time-varying parameters α and β_1. The previous time period estimates of these parameters are good choices. Then, the estimating functions for the two stages are

i. Fix α and β_1 and solve the unweighted estimating function

$$Q_1(\beta_2|d, \alpha, \beta_1) = \frac{\partial l_1(\theta;d_1)}{\partial \beta_2} + \frac{\partial l_2(\theta;d_2)}{\partial \beta_2} + \ldots + \frac{\partial l_T(\theta;d_T)}{\partial \beta_2} \text{ for } \hat{\beta}_2.$$

ii. Fix $\beta_2 = \hat{\beta}_2$ and solve the weighted estimating function

$$Q_2\left(\alpha, \beta_1|d, w, \hat{\beta}_2\right) = \begin{bmatrix} w_1 \frac{\partial l_1(\theta;d_1)}{\partial \alpha} + w_2 \frac{\partial l_2(\theta;d_2)}{\partial \alpha} + \cdots + w_T \frac{\partial l_T(\theta;d_T)}{\partial \alpha} \\ w_1 \frac{\partial l_1(\theta;d_1)}{\partial \beta_1} + w_2 \frac{\partial l_2(\theta;d_2)}{\partial \beta_1} + \cdots + w_T \frac{\partial l_T(\theta;d_T)}{\partial \beta_1} \end{bmatrix}_{\beta_2=\hat{\beta}_2} \text{ for } \hat{\alpha} \text{ and } \hat{\beta}_1.$$

Note that we can update estimates in the two stages with different frequencies. Since we expect that changes in the fixed parameter estimates are small, then stage 1 may occur less frequently than the estimation of the time-varying parameters in stage 2. For example, in the monthly update of WEE estimates, the estimates of the covariate effects that are assumed to be fixed may be updated annually.

Some early investigation of these alternatives for simulated data from the customer loyalty application shows that the estimates from the various alternatives have less bias than the standard WEE estimates but they are unstable. Further consideration of the approximations for the estimate

of the variance of $\hat{\theta}$ and the distribution of the hypothesis test statistic involving $\hat{\theta}$ under these alternatives is required.

References

Agresti, A. (2007). *Analysis of Ordinal Categorical Data, Second Edition.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Ahmad, O.B., Boschi-Pinto, C., Lopex, A.D., Murray, C., Lozano, R., & Inou, M. (2001). Age Standardization of Rates: A New WHO Standard (discussion paper). *World Health Organization*. Retrieved 05/27/2015, from www.who.int/healthinfo/paper31.pdf.

Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika*, 65(1), 53-59.

Box, G.E.P., Jenkins, G.M., & MacGregor, J.F. (1974). Some Recent Advances in Forecasting and Control. *Applied Statistics*, 23, 158-179.

Bursac, Z., Gauss, C.H., Williams, D.K., & Hosmer, D.W. (2008). Purposeful selection of variables in logistic regression. *Biology and Medicine*, 3(17).

Canadian Institute for Health Information (2016). CIHI's Strategic Plan, 2016 to 2021. *Canadian Institute for Health Information.* Retrieved 06/22/2016, from www.cihi.ca/en/about-cihi/corporate-strategies/strategic-plan.

Cancer Care Ontario (2008). Colon Cancer Check. *Cancer Care Ontario*. Retrieved 09/14/2015, from www.cancercare.on.ca/common/pages/UserFile.aspx?fileId=260524.

Casella, G. & Berger, R.L. (2002). *Statistical Inference, Second Edition*. Duxbury, Pacific Grove, California.

Chen, B., Yi, G.Y., & Cook, R.J. (2010). Weighted Generalized Estimating Functions for Longitudinal Response and Covariate Data That Are Missing at Random. *Journal of the American Statistical Association*, 105(489), 336-353.

Clark, D.E., Hannan, E.L., & Wu, C. (2010). Predicting Risk-Adjusted Mortality for Trauma Patients: Logistic vs. Multilevel Logistic Models. *Journal of the American College of Surgeons*, 211(2), 224-231.

Colosimo, E.A., Fausto, M.A., Freitas, M.A., & Pinto, J.A. (2012). Practical modeling strategies for unbalanced longitudinal data analysis. *Journal of Applied Statistics*, 39(9), 2005-2013.

Cooper Barfoot, P.L., Steiner, S.H., & MacKay, R.J. (2016). Bias/Variance Trade-off in Estimates of a Process Parameter based on Temporal Data. Manuscript submitted for publication to Journal of Quality Technology, original submission - April 2016, revision submission #2 - November 2016.

COPPS-CMS White Paper Committee (2012). Statistical Issues in Assessing Hospital Performance. *The Committee of the Presidents of Statistical Societies*. Retrieved 10/19/2016, from www.cms.gov/.

Geyer, C.J. (2013). 5601 Notes: The Sandwich Estimator. *The School of Statistics, University of Minnesota*. Retrieved 05/13/2015, from www.stat.umn.edu/geyer/5601/notes/sand.pdf.

Godambe, V.P. & Kale, B.K. (1991). Estimating functions: an overview. *Estimating Functions,* V.P. Godambe (ed.), Oxford University Press, Oxford, 3-20.

Grewal, M.S. & Andrews, A.P. (2008). *Kalman Filtering Theory and Practice Using MATLAB*. John Wiley & Sons, Hoboken, New Jersey.

Hardin, J.W. & Hilbe, J.M. (2013). *Generalized Estimating Equations, Second Edition*. Chapman & Hall/CRC Press, Boca Raton, Florida.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning, Second Edition*. Springer, New York, New York.

Hu, F. (1997). The Asymptotic Properties of the Maximum-Relevance Weighted Likelihood Estimators. *The Canadian Journal of Statistics*, 25(1), 45-59.

Hu, F. & Rosenberger, W.F. (2000). Analysis of time trends in adaptive designs with application to a neurophysiology experiment. *Statistics in Medicine*, 19, 2067-2075.

Hu, F & Zidek, J.V. (2002). The Weighted Likelihood. *The Canadian Journal of Statistics*, 30(3), 347-371.

Resources to help you learn and use SAS (n.d.). *Institute for Digital Research and Education UCLA* Retrieved 10/19/2016, from ats.ucla.edu/stat/sas.

Jansen, I., Beunckens, C., Molengerghs, G., Verbeke, G., & Mallinckrodt, C. (2006). Analyzing Incomplete Discrete Longitudinal Clinical Trial Data. *Statistical Science*, 21(1), 52-69.

Kalbfleisch, J.D. & Wolfe, R.A. (2013). On Monitoring Outcomes of Medical Providers. *Statistics in Biosciences*, 5(2), 286-302.

Lehmann, E.L. & Romano, J.P. (2005). *Testing Statistical Hypotheses*. Springer Science+Business Media, New York, New York.

Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika,* 73(1), 13-22.

Lipsitz, S.R., Kim, K., & Zhao, L. (1994). Analysis of Repeated Categorical Data Using Generalized Estimating Equations. *Statistics in Medicine,* 13, 1149–1163.

Little, R.J.A. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data, Second Edition*. John Wiley & Sons, Hoboken, New Jersey.

Liu, X., MacKay, R.J., & Steiner, S.H. (2008). Monitoring Multiple Stream Processes. *Quality Engineering*, 20, 296-308.

Lucas, J.M. & Saccucci, M.S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements. *Technometrics,* 32, 1-29.

Markey, R., Reichheld, F.F., & Dullweber, A. (2013). A Test of Customer Loyalty. *Bain & Company*. Retrieved 10/19/2016, from www.bain.com/publications/articles/a-test-of-customer-loyalty-smeinfo.aspx.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Press, Boca Raton, Florida.

McGregor, J. (2006). Would You Recommend Us? *Business Week*, January 30, 2006, 95-95.

Montgomery, D.C. (2013). *Introduction to Statistical Quality Control, Seventh Edition.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Muth, J.F. (1960). Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, 55(290), 299-305.

New York State Department of Health (2015). Percutaneous Coronary Interventions (PCI) in New York State 2010-2012. *New York State Department of Health.* Retrieved 06/22/2016, from www.health.ny.gov/statistics/diseases/cardiovascular/.

Nowinski, V. (2009). Net Promoter Economics: The Impact of Word of Mouth. *Satmetrix.* Retrieved 10/19/2016, from cdn2.hubspot.net/hub/268441/file-1479657669-pdf/White_Papers_(PDFs)/Aug_19/WP-NetPromoEconTheImpactOfWordOfMouth-Wireless.pdf.

NPS Benchmarks (n.d.). *CustomerGauge*. Retrieved 02/03/2015, from www.npsbenchmarks.com.

Qu, A., Yi, G.Y., Song, P.X.-K., & Wang, P. (2011). Assessing the validity of weighted generalized estimating equations. *harvard*

Reichheld, F.F. (2003). The One Number You Need to Grow. *Harvard Business Review*, 81(12), 46-54.

Reichheld, F.F. & Markey, R. (2011). *The Ultimate Question 2.0: How Net Promoter Companies* Thrive *in a Customer-driven World*. Harvard Business Press, Boston, Massachusetts.

Robins, J.M., Rotnitzky, A., & Zhao, L.P. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, 90(429), 106-121.

Scientific Registry of Transplant Recipients Home Page (n.d.). *Scientific Registry of Transplant Recipients.* Retrieved 10/10/2016, from srtr.transplant.hrsa.gov.

Scientific Registry of Transplant Recipients (2016a). Technical methods for the Program-Specific Reports. *The Scientific Registry of Transplant Recipients*. Retrieved 10/10/2016, from www.srtr.org/csr/current/tech_notes.aspx.

Scientific Registry of Transplant Recipients (2016b). The SRTR Program-Specific Reporting Tools: Key Points. *The Scientific Registry of Transplant Recipients*. Retrieved 10/10/2016, from www.srtr.org/csr/current/tech_notes.aspx.

Shenton, L.R. & Bowman, K.O. (1977). A Bivariate Model for the Distribution of $\sqrt{b_1}$ and b_2. *Journal of the American Statistical Association*, 72, 206-211.

Small, C.G. (2010). *Expansions and Asymptotics for Statistics*. Chapman & Hall/CRC Press, Boca Raton, Florida.

Spiegelhalter, D.J., Sherlaw-Johnson, C., Bardsley, M., Blunt, I., Wood, C., & Grigg, O. (2012). Statistical methods for healthcare regulation: rating, screening and surveillance. *Journal of the Royal Statistical Society A*, 175(1), 1-47.

Steiner, S.H. (1999). EWMA Control Charts with Time-Varying Control Limits and Fast Initial Response. *Journal of Quality Technology*, 31, 75-86.

Steiner, S.H. (2014). Risk-Adjusted Monitoring of Outcomes in Health Care. *Statistics in Action: A Canadian Outlook,* J.F. Lawless (ed.), Chapman & Hall/CRC Press, Boca Raton, Florida, 225-241.

Steiner, S.H. & MacKay, R.J. (2014). Monitoring Risk-adjusted Medical Outcomes Allowing for Changes over Time. *Biostatistics*, 15(4), 665-676.

U.S. Department of Health and Human Services (2014). United States Organ Transplantation OPTN/SRTR 2012 Annual Data Report. *U.S. Department of Health and Human Services*. Retrieved 10/10/2015, from srtr.transplant.hrsa.gov/annual_reports.

Wald, A. (1949). Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics*, 20(5), 595-601.

Wang, X., van Eeden, C., & Zidek, J.V. (2002). Technical report on weighted likelihood estimation and asymptotic properties of the weighted likelihood estimators. Technical Report No. 201. *Department of Statistics, University of British Columbia*. Retrieved 26/02/2015, from www.stat.ubc.ca/Research/TechReports/tr02.php.

White, H. (1982). Maximum likelihood estimation of misspecified model. *Econometrica*, 50, 1–25.

World Health Organization (2009). WHO Guidelines for Safe Surgery 2009: Safe Surgery Saves Lives. *World Health Organization.* Retrieved 19/10/2016, from apps.who.int/iris/ bitstream/10665/44185/1/9789241598552_eng.pdf.

Xie, M. & Yang, Y. (2003). Asymptotics for Generalized Estimating Equations with Large Cluster Sizes. *The Annals of Statistics*, 31(1), 310–347.

Yashchin, E. (1995). Estimating the Current Mean of a Process Subject to Abrupt Changes. *Technometrics*, 37, 311–323.

Yuan, Y.C. (2000). Multiple Imputation for Missing Data: Concepts and New Development. *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference (No. 267), SAS Institute*. Retrieved 10/19/2016, from support.sas.com/rnd/app/stat/papers/ multipleimputation.pdf.