## Essays on risk and econometrics

Yue, Y.

**Publication date**
2022
**Document Version**
Final published version

Yuan Yue

# Essays on Risk and Econometrics

This thesis contains the outcome of two projects: risk and property prices, and computational aspects of model averaging. In the first project, we investigate the effect of objective and subjective earthquake risk embedded in Japanese property prices. We employ a multivariate error components regression model to exploit a rich dataset containing transaction prices and various characteristics of residential properties. While the earthquake probabilities are seen as objective measures of earthquake risk, we elicit a subjective measure of risk from the data by means of a parametric family of probability weighting functions. The estimated shape of the probability weighting function provides insight on how people's perception of earthquake probabilities is reflected in property prices.

In the second project we study the properties of the weighted-average least squares (WALS) estimator. The idea of model averaging emerges from the insight that model selection and estimation should not be seen as two separate steps, but rather as one integrated procedure. Model averaging estimators do not select one best-fitting candidate model but estimate a whole range of candidate models and assigns weights to each of the candidate estimates. We explore the computational properties of WALS and develop statistical packages that enable the computation of WALS estimates.

Essays on Risk and Econometrics

Yuan Yue

793

Universiteit van Amsterdam

# Essays on Risk and Econometrics

Yuan Yue

Essays on Risk and Econometrics

# ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. ir. K.I.J. Maex

ten overstaan van een door het College voor Promoties ingestelde commissie,

in het openbaar te verdedigen in de Agnietenkapel

op donderdag 28 april 2022, te 10.00 uur

door Yuan Yue

geboren te Anhui, China

***Promotiecommissie***

| | | |
|---|---|---|
| *Promotores:* | Prof. dr. R.J.A. Laeven | Universiteit van Amsterdam |
| | Prof. dr. J.R. Magnus | Vrije Universiteit Amsterdam |
| *Overige leden:* | Prof. dr. R.J.M. Alessie | Universiteit van Groningen |
| | Prof. dr. K. Antonio | Universiteit van Amsterdam |
| | Prof. dr. H.P. Boswijk | Universiteit van Amsterdam |
| | Prof. dr. F.J.G.M. Klaassen | Universiteit van Amsterdam |
| | Prof. dr. S.J. Koopman | Vrije Universiteit Amsterdam |

Faculteit Economie en Bedrijfskunde

# Acknowledgment

First and foremost I would like to thank my supervisors, Professors Jan Magnus and Roger Laeven for their invaluable supervision and patience during my PhD research. Both Jan and Roger are brilliant scholars and while working with them I have gained not only academic knowledge and research skills but more importantly, the ability of concentration and dedication. Although I have decided not to pursue a career in academia, the things I have learned from my supervisors will continue to benefit me in my career and personal life which I am sincerely grateful for.

The completion of my PhD thesis would not be possible without the tutelage and encouragement from Jan Magnus. Through my experience of working closely with Jan I have been impressed and inspired by his unstoppable quest for knowledge and everlasting curiosity for research. Jan is not only a mentor to me in research but also taught me important lessons in life by example. He has shown me the power of simplicity when it comes to building and explaining complicated theories. In the final stage of my research project I have learned from Jan how to set feasible goals and the correct priorities.

While working on the research projects I had the pleasure of meeting my co-authors, Prof. Masako Ikefuji and Dr. Giuseppe De Luca, who are accomplished researchers in their respective fields. I would like to thank Masako for treating me like a friend and I am deeply inspired by her dedication to research and attention to detail. I am grateful to Giuseppe for his generous help in the study of WALS and the STATA programme and for his insightful comments and discussions during the project.

I am grateful to my fellow PhD students Hao Fang, Lingwei Kong, Merrick Li, and Junze Sun for daily interactions and table football games that made the day-to-day life of a PhD more enjoyable. My ex-roommate Zhenzhen Fan has shown remarkable self-discipline and intellectual curiosity which are important traits for success in academia. I also thank my colleagues Katrien Antonio, Umut Can, Jan de Kort, Jitze Hooijsma, Rob Kaas, Andrei Lalu, Hans Schumacher, Rob Sperna Weiland, Frank van Berkum, Servaas van Bilsen, Leendert van Gastel, Michel Vellekoop, Tim Boonen, Xiye Yang from the University of Amsterdam and Tinbergen Institute who contributed fruitful discussions and interesting clashes of thoughts in the brown bag seminars, workshops, and various other occasions.

Finally, I would like to thank my parents and my husband, without the support of whom I would never be able to complete my PhD studies. They have shown unconditional love and trust in me even when I had doubts in myself, and it is their love and trust that enabled me to wade through the hardships along the journey.

Yuan Yue
Amsterdam, November 2021

# Declaration of author's contribution

Chapters 2–4 contain the research that has eventually led to the paper "Earthquake risk embedded in property prices: Evidence from five Japanese cities," coauthored with Masako Ikefuji, Roger Laeven, and Jan Magnus, and published in the *Journal of the American Statistical Association*. All four authors have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

Chapter 5 is a single-authored chapter, supervised by Jan Magnus and Giuseppe De Luca.

Financial support of Van Ameyde BV is gratefully acknowledged.

# Table of contents

# Chapter 1

# Introduction

This thesis contains the outcome of two unrelated projects: risk and property prices, and computational aspects of model averaging.

In the first project, we investigate the effect of objective and subjective earthquake risk embedded in Japanese property prices. This is measured within the framework of a hedonic pricing model, which is the benchmark model for analyzing property prices. In hedonic pricing models, the characteristics of a property are seen as components that each independently contribute to a part of the property price.

We collect a rich dataset containing transaction prices of residential properties and various characteristics that are relevant to property prices. We distinguish between three types of residential properties: residential land (land only), residential land (land and building), and pre-owned condominiums. Each type has different attributes but also shares many characteristics. Among the property characteristics, there are cross-sectional data such as information on the attractiveness of the district where the property is located, time-series data such as macroeconomic variables, and also individual characteristics such as square footage, building coverage ratio, or distance to the nearest station.

To exploit the available dataset, we employ a multivariate error components regression model. The error terms are the sum of three independent components, capturing the time-specific, cross-section-specific, and individual-specific effects. Furthermore, each component is a vector instead of a scalar, enabling the pooling of equations of closely related error structures while maintaining a relatively small number of parameters. This vector has three elements, each corresponding to one of the three property types. The dimension of the huge variance matrix caused by the vector form can be drastically reduced thanks to the error components structure.

We introduce earthquake risk measured as the probability of an earthquake exceeding a certain magnitude or intensity threshold over a certain time period. Since the occurrence of earthquakes is frequent in Japan and varies both spatially and temporally, earthquake risk is a non-negligible characteristic in the valuation of property prices. We distinguish between long-run risk and short-run risk. The long-run earthquake risk data is provided by the Japan Seismic Hazard Information Station, and is defined as the probability of an earthquake exceeding certain intensity thresholds in the next thirty years in a given area. We take the average of these long-run probabilities over the entire sample period to create a time-invariant measure of the overall riskiness corresponding to a given area. On the other hand, short-run

probabilities are ninety-day probabilities which vary per time period and per city, simulated by a temporal epidemic-type aftershock sequence (ETAS) model. The ETAS model is a path-dependent marked point process commonly used for modelling seismic activities. The idea behind the ETAS model is that each earthquake can trigger aftershocks like epidemics and that the intensity of the impact of each trigger event diminishes over time.

While the long-run and short-run earthquake probabilities are seen as objective measures of earthquake risk, we also try to elicit a subjective measure of risk from the data. This is achieved by using a parametric family of probability weighting functions, which is widely used in economic analysis and decision theory. The idea is that, by entering the weighted (subjective) probabilities instead of the original (objective) probabilities in the regression function, we can estimate (by maximum likelihood) from the data the unknown parameter by doing a grid search. The corresponding variance of this estimator needs to be derived because the situation is nonstandard in that one of the regressors depends on the parameter of interest. The estimated parameter sheds light on the shape of the probability weighting function and thus provides insight on how people's perception of small and large probabilities is reflected in the property prices. When the probability weighting function is inverse-$S$ shaped, it means that people overweight small probabilities and underweight large probabilities. When the probability weighting function is $S$-shaped, it means that people underweight small probabilities and overweight large probabilities. When the parameter equals 1, the function degenerates to the identity function, which means there are no subjective distortions of the probabilities.

We found that long-run objective earthquake risk has a significantly negative impact on property prices. The additional impact of objective short-run earthquake risk is not significantly different from zero. However, the distorted short-run earthquake probabilities (allowing for probability weighting) do have a significantly negative impact on property prices. We found this probability weighting function to be $S$-shaped, thus underweighting small probabilities and overweighting larger probabilities. This finding is contrary to conventional wisdom in decision theory where probability weighting functions are commonly found to be inverse-$S$ shaped, which may be explained by the fact that the background earthquake intensity is positive, so that people do not perceive temporary deviations of short-run earthquake risk with a reference probability of zero but with a positive reference probability.

In the second project we study the properties of a model-averaging estimator, namely the weighted-average least squares (WALS) estimator. The idea of model averaging emerges from the insight that model selection and estimation should not be seen as two separate steps, but rather as one integrated procedure. Model averaging does not select one best-fitting candidate model but estimates a whole range of candidate models and assigns weights to each of the candidate estimates.

The common and naive use of $t$-ratios in applied econometrics as a diagnostic statistic goes like this. When the $t$-ratio of a regressor is above a certain threshold (usually 1.96 at the 5% significance level), the regressor is deemed "significant" and is kept in the model; and when the $t$-ratio is below the threshold, it is removed from the model. This approach thus ignores the fact that the same data have been used for diagnostic testing and estimation, so that inference obtained from the second step is likely to be misleadingly precise because it ignores the uncertainty generated from the first step.

The above procedure is called "pretesting", and it leads to estimators which are not differentiable and

hence not admissible. It is the simplest form of a WALS estimator, namely the case where the weights of candidate models can only be 0 or 1. The WALS procedure generalizes this discrete version to a continuous version, where the weights are now continuous functions of the $t$-ratio. Literature on model averaging diverges into frequentist model averaging (FMA) and Bayesian model averaging (BMA). We explore the properties of WALS which is a Bayesian combination of frequentist estimators. This estimator has advantages over the traditional BMA estimators in terms of interpretation and computational efficiency.

The framework of WALS is the linear regression model with independent and identically distributed normal error terms. We distinguish between focus regressors, which we want to keep in the model regardless of the outcome of diagnostic testing, and auxiliary regressors, which may or may not be in the model.

We develop statistical packages that enable the computation of WALS estimates, standard errors, bias and mean squared errors, confidence intervals, and predictions. The estimation hinges on a choice of prior and prior parameters, which come from a reflected generalized Gamma family — the Weibull, Subbotin, and Laplace priors. We show that the Laplace prior leads to estimates with higher bias, while for Weibull and Subbotin priors the theoretically obtained minimax regret prior parameters, which minimize the maximum regret over all possible values, can lead to more bias than other choices of the prior parameter.

WALS estimation makes use of numerical integration results except for the case of the Laplace prior. We show the effect of the choice between two alternative integration routines, the Gauss-Laguerre quadrature and the adaptive quadrature, on the precision and computational efficiency of the program. We also explore the limits to the WALS estimation procedure by using simulation set-ups where the matrix of regressors is nearly singular and when the number of auxiliary regressors is large. We found that, as long as the input data are of full column rank, the estimator is able to produce estimation results, but the bias increases exponentially when the correlation between regressors increases. We establish a relationship between the number of required observations and the number of auxiliary regressors under the same targeted bias level, and found this relationship to be approximately linear.

The project focuses on the computational properties of the WALS estimator. Apart from the findings listed above, we show some other aspects of the WALS estimation procedures, such as the effect of Monte Carlo replications on the precision of confidence intervals, and the comparison of the computational speed of different packages (`R`, `Python`, or `Stata`). By exploring these aspects we aim to provide insight in the performance of the WALS estimator and the various available estimation options, so that a typical user can make informed decisions when using WALS in empirical applications.

The thesis is structured as follows. Chapter 2 sets up a model with a multiple error components structure and derives the associated maximum likelihood estimation procedure. It also designs a grid search procedure and derives the variance of the parameter of interest when one of the regressors depend on this parameter. Chapter 3 explains the data collection process for the empirical study of earthquake risk embedded in property prices. Chapter 4 shows the full picture of the empirical study and presents the empirical results. Chapter 5 explores the (computational) properties of the WALS estimator.

# Chapter 2

# Estimation of panel data models with multivariate error components

## 2.1 Introduction

The combination of cross-sectional data and time-series data has become widespread and can be useful in the economic analysis of phenomena that involve cross-sectional differences, temporal fluctuations, or both.

A standard single error regression model would be insufficient to unveil the complexity in the relationship within and between equations. Regression models with error terms being the sum of two or more independent components, widely known as error components models, have been extensively used for the aforementioned purposes ever since the seminal work of Balestra and Nerlove (1966). Two-error components models involve one time-specific or cross-section-specific component, and one individual-specific component. Three-error components models involve both the time-specific and cross-section-specific components, and the individual-specific part.

Multivariate error components models have error structures where the components are vectors instead of scalars. The vector form enables pooling equations of closely related error structures together while maintaining a relatively small number of parameters, but also gives rise to a huge variance matrix that is computationally cumbersome to estimate directly. Exploiting the independence assumption of the error components it is possible to decompose the variance matrix and drastically reduce its dimensions. In this chapter we introduce a multivariate three-error components model for which we develop associated maximum likelihood estimation and variance computation procedures. Another contribution of this chapter is the introduction of a regressor that depends on a parameter of interest, the value and variance of which needs to be estimated. We derive the variance of such a parameter and design a procedure to find its estimate based on a grid search.

Multivariate two-error components were first employed by Chamberlain and Griliches (1975) using maximum likelihood techniques. Multivariate three-error components were first considered by Avery (1977) who derived a feasible Aitken estimator, which is however not maximum likelihood and turns out to be asymptotically inefficient. Baltagi (1980) derived an alternative estimator, also not maximum likelihood, which is asymptotically efficient. Magnus (1982) discussed the estimation and testing of the

multivariate two- and three-error components models in a maximum likelihood context.

The rest of this chapter proceeds as follows. Section 2.2 lays out the model. Section 2.3 specifies the error structure. Section 2.4 decomposes the large matrix into smaller ones using the assumptions on the error structure. Section 2.5 derives the concentrated likelihood. Section 2.6 decomposes the concentrated likelihood function into matrices of smaller dimensions. Section 2.7 derives the variances of the parameters and describes the estimation method using concentrated likelihood. Section 2.8 concludes.

## 2.2 The model

We describe the model in a hedonic pricing context, where the dependent variable is log-property price and the independent variables are property characteristics. There are different types of properties and each type corresponds to different characteristics and total price levels. There is overlap in the sets of characteristics for each type, such as variables related to the location and/or time but indifferent to types. Each property is located in a certain district and transaction takes place at a certain time.

Suppose there are $p$ types of property, $N$ different districts and $T$ time periods in the data set. We denote the $h$-th observation of type $k$ in district $i$ during quarter $t$ as $y_{it}^{(h,k)}$. The number of observations varies per district, type and quarter, and this affects the precision. We let $H_{it}^{(k)}$ denote the number of observations on each type in district $i$ during quarter $t$.

Our general model can be written as

$$y_{it}^{(h,k)} = x_{it}^{(h,k)\prime}\beta_0 + x_{it}(\psi)'\beta_1 + u_{it}^{(h,k)}, \tag{2.1}$$

where $x_{it}^{(h,k)}$ includes constant terms, time-invariant characteristics, district-invariant characteristics, and individual-specific characteristics. $x_{it}(\psi)$ is the variable that depends on an additional parameter.

In order to obtain a (balanced) panel we average over $h$, and obtain

$$\bar{y}_{it}^{(k)} = \bar{x}_{it}^{(k)\prime}\beta_0 + x_{it}(\psi)'\beta_1 + \bar{u}_{it}^{(k)}, \tag{2.2}$$

where we average over $H_{it}^{(k)}$ items, which thus depends on how many properties of type $k$ there are in a given district.

Next we combine the three types of property into one $p \times 1$ vector:

$$\bar{y}_{it} = X_{it}^{*}\beta_0 + \iota x_{it}(\psi)'\beta_1 + \bar{u}_{it}, \tag{2.3}$$

where $\iota$ is a $p \times 1$ vector of ones, which we write more succinctly as

$$\bar{y}_{it} = \bar{X}_{it}\beta + \bar{u}_{it} \qquad (i = 1,\ldots,N;\ t = 1,\ldots,T), \tag{2.4}$$

where $\bar{y}_{it}$ is a $p \times 1$ vector of random observations, explained by (non-random) regressors $\bar{X}_{it} = \bar{X}_{it}(\psi)$, an unknown parameter vector $\beta$, and random errors $\bar{u}_{it}$ $(p \times 1)$. The estimation of $\beta$ is done by first fixing $\psi$ to find the conditional estimates and then runnig a grid search over a set of values for $\psi$ to maximize the likelihood; see Sections 2.5 and 2.7.

## 2.3 Specification of the error term

The errors are assumed to follow a $p$-variate three-error components structure,

$$\bar{u}_{it} = \zeta_i + \eta_t + \epsilon_{it}, \tag{2.5}$$

a sum of three independent components each of which is iid with zero means and variances

$$\operatorname{var}(\zeta_i) = \Sigma_\zeta, \qquad \operatorname{var}(\eta_t) = \Sigma_\eta, \qquad \operatorname{var}(\epsilon_{it}) = \Sigma_\epsilon, \tag{2.6}$$

where $\Sigma_\zeta$ and $\Sigma_\eta$ are positive semidefinite, and $\Sigma_\epsilon$ is positive definite, all of order $p \times p$.

The Cholesky decomposition of $\Sigma_\zeta$, $\Sigma_\eta$, and $\Sigma_\epsilon$ can be written as

$$\Sigma_\zeta = L_\zeta L_\zeta', \; \Sigma_\eta = L_\eta L_\eta', \; \Sigma_\epsilon = L_\epsilon L_\epsilon', \tag{2.7}$$

where $L_\zeta$, $L_\eta$, and $L_\epsilon$ are $p \times p$ lower triangular matrices.

Our error structure implies that

$$\mathrm{E}(\bar{u}_{it}\, \bar{u}_{js}') = \begin{cases} \Sigma_\zeta + \Sigma_\eta + \Sigma_\epsilon & \text{if } i = j \text{ and } t = s, \\ \Sigma_\zeta & \text{if } i = j \text{ and } t \neq s, \\ \Sigma_\eta & \text{if } i \neq j \text{ and } t = s, \\ 0 & \text{if } i \neq j \text{ and } t \neq s. \end{cases} \tag{2.8}$$

Let

$$Y = \begin{pmatrix} \bar{y}_{11} & \bar{y}_{12} & \dots & \bar{y}_{1T} \\ \bar{y}_{21} & \bar{y}_{22} & \dots & \bar{y}_{2T} \\ \vdots & \vdots & & \vdots \\ \bar{y}_{N1} & \bar{y}_{N2} & \dots & \bar{y}_{NT} \end{pmatrix}, \qquad U = \begin{pmatrix} \bar{u}_{11} & \bar{u}_{12} & \dots & \bar{u}_{1T} \\ \bar{u}_{21} & \bar{u}_{22} & \dots & \bar{u}_{2T} \\ \vdots & \vdots & & \vdots \\ \bar{u}_{N1} & \bar{u}_{N2} & \dots & \bar{u}_{NT} \end{pmatrix}, \tag{2.9}$$

and

$$\bar{X}_{(t)} = \begin{pmatrix} \bar{X}_{1t} \\ \bar{X}_{2t} \\ \vdots \\ \bar{X}_{Nt} \end{pmatrix}, \qquad X = \begin{pmatrix} \bar{X}_{(1)} \\ \bar{X}_{(2)} \\ \vdots \\ \bar{X}_{(T)} \end{pmatrix}. \tag{2.10}$$

Then we can write (2.4) in stacked form as

$$y = X\beta + u, \tag{2.11}$$

where $y = \operatorname{vec} Y$ and $u = \operatorname{vec} U$. We shall assume that $y$ is normally distributed with mean $\mu = X\beta$ and variance $\Omega(\theta)$, so that $\beta$ refers to the mean parameters and $\theta$ to the variance parameters. More specifically, $\theta$ contains the non-zero elements of $L_\zeta$, $L_\eta$, and $L_\epsilon$, thus $(p+1) \times p/2 \times 3$ parameters in the three-error components model and $(p+1) \times p/2 \times 2$ parameters in the two-error components model.

## 2.4 Decomposition and properties of $\Omega$

Given the error components structure proposed in Section 2.3, we show that the $(NTp) \times (NTp)$ variance matrix of the error term $u$ in (2.11) takes a particularly convenient form, allowing an easy way to calculate its inverse and determinant:

**Proposition 2.4.1.** *Let $\imath_T$ and $\imath_N$ denote vectors containing only ones, of orders $T$ and $N$, respectively, and let $J_T = \imath_T \imath_T' / T$ and $J_N = \imath_N \imath_N' / N$. Then,*

$$\Omega = \text{var}(u) = V_1 \otimes \Delta_1 + V_2 \otimes \Delta_2 + V_3 \otimes \Delta_3 + V_4 \otimes \Delta_4,$$

*where*

$$V_1 = J_T \otimes J_N, \qquad V_2 = J_T \otimes (I_N - J_N),$$
$$V_3 = (I_T - J_T) \otimes J_N, \qquad V_4 = (I_T - J_T) \otimes (I_N - J_N),$$

*and*

$$\Delta_1 = \Sigma_\epsilon + T\Sigma_\zeta + N\Sigma_\eta, \qquad \Delta_2 = \Sigma_\epsilon + T\Sigma_\zeta,$$
$$\Delta_3 = \Sigma_\epsilon + N\Sigma_\eta, \qquad \Delta_4 = \Sigma_\epsilon.$$

*In addition,*

$$\Omega^{-1} = V_1 \otimes \Delta_1^{-1} + V_2 \otimes \Delta_2^{-1} + V_3 \otimes \Delta_3^{-1} + V_4 \otimes \Delta_4^{-1}$$

*and*

$$|\Omega| = |\Delta_1| \, |\Delta_2|^{N-1} \, |\Delta_3|^{T-1} \, |\Delta_4|^{(N-1)(T-1)}.$$

**Proof:** First, we note that the $V_i$ are idempotent matrices, that $V_i V_j = 0$ $(i \neq j)$, and that $\sum_i V_i = I_{NT}$. This follows from the mixed-product property of Kronecker products: for matrices $A$, $B$, $C$, and $D$ such that $AC$ and $BD$ exists, $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. In fact, $J_T$ and $J_N$ are both idempotent, thus

$$V_1 V_1 = (J_T \otimes J_N)(J_T \otimes J_N) = (J_T J_T) \otimes (J_N J_N) = J_T \otimes J_N = V_1,$$
$$V_1 V_2 = (J_T \otimes J_N)(J_T \otimes (I_N - J_N)) = (J_T J_T) \otimes (J_N (I_N - J_N)) = J_T \otimes (J_N - J_N) = 0,$$

and for all other $i$ and $j$ it follows similarly. Also, rules of the Kronecker product gives

$$V_1 + V_2 = J_T \otimes J_N + J_T \otimes (I_N - J_N) = J_T \otimes I_N,$$
$$V_1 + V_3 = J_T \otimes J_N + (I_T - J_T) \otimes J_N = I_T \otimes J_N,$$
$$V_1 + V_2 + V_3 + V_4 = J_T \otimes I_N + (I_T - J_T) \otimes I_N = I_T \otimes I_N = I_{NT}.$$

The ranks of the matrices are given by

$$\begin{aligned}
rank(V_1) &= rank(J_T)rank(J_N) = 1, \\
rank(V_2) &= rank(J_T)rank(I_N - J_N) = N - 1, \\
rank(V_3) &= rank(I_T - J_T)rank(J_N) = T - 1, \\
rank(V_4) &= rank(I_T - J_T)rank(I_N - J_N) = (N-1)(T-1).
\end{aligned}$$

We write

$$\begin{aligned}
\Omega = \text{var}(u) &= \iota_T \iota_T' \otimes I_N \otimes \Sigma_\zeta + I_T \otimes \iota_N \iota_N' \otimes \Sigma_\eta + I_T \otimes I_N \otimes \Sigma_\epsilon \\
&= J_T \otimes I_N \otimes T\Sigma_\zeta + I_T \otimes J_N \otimes N\Sigma_\eta + I_T \otimes I_N \otimes \Sigma_\epsilon \\
&= (V_1 + V_2) \otimes T\Sigma_\zeta + (V_1 + V_3) \otimes N\Sigma_\eta + (V_1 + V_2 + V_3 + V_4) \otimes \Sigma_\epsilon \\
&= V_1 \otimes \Delta_1 + V_2 \otimes \Delta_2 + V_3 \otimes \Delta_3 + V_4 \otimes \Delta_4.
\end{aligned}$$

The results now follow from Baltagi (1980), Magnus (1982, Lemma 2.1), and Abadir and Magnus (2005, Exercise 8.73).

To verify the inverse, we directly check the product:

$$\begin{aligned}
&\Omega(V_1 \otimes \Delta_1^{-1} + V_2 \otimes \Delta_2^{-1} + V_3 \otimes \Delta_3^{-1} + V_4 \otimes \Delta_4^{-1}) \\
&= \sum_i (V_i \otimes \Delta_i)(V_i \otimes \Delta_i^{-1}) + \sum_{i \neq j} (V_i \otimes \Delta_i)(V_j \otimes \Delta_j^{-1}) \\
&= \sum_i (V_i V_i) \otimes (\Delta_i \Delta_i^{-1}) + \sum_{i \neq j} (V_i V_j) \otimes (\Delta_i \Delta_j^{-1}) \\
&= \sum_i V_i + \sum_{i \neq j} 0 = I_{NT}.
\end{aligned}$$

For the determinant, since the eigenvalues of $\Omega$ are the eigenvalues of $\Delta_1$, $\Delta_2$, $\Delta_3$ and $\Delta_4$ with multiplicities of 1, $N-1$, $T-1$ and $(N-1)(T-1)$, respectively, and the determinant is the product of the eigenvalues, we have

$$|\Omega| = |\Delta_1| \, |\Delta_2|^{N-1} \, |\Delta_3|^{T-1} \, |\Delta_4|^{(N-1)(T-1)}.\|$$

In the special case where $\Sigma_\zeta = 0$ we have

$$\Delta_1 = \Delta_3 = \Sigma_\epsilon + N\Sigma_\eta, \qquad \Delta_2 = \Delta_4 = \Sigma_\epsilon, \tag{2.12}$$

and

$$\Omega = I_T \otimes J_N \otimes \Delta_1 + I_T \otimes (I_N - J_N) \otimes \Delta_2. \tag{2.13}$$

In the special case where $\Sigma_\eta = 0$ we have

$$\Delta_1 = \Delta_2 = \Sigma_\epsilon + T\Sigma_\zeta, \qquad \Delta_3 = \Delta_4 = \Sigma_\epsilon, \tag{2.14}$$

and

$$\Omega = J_T \otimes I_N \otimes \Delta_1 + (I_T - J_T) \otimes I_N \otimes \Delta_3. \tag{2.15}$$

Both are examples of a multivariate two-error components structure. Notice that we employ two idempotent matrices when there are two components, but that we need four (rather than three) when there are three components.

## 2.5 The concentrated likelihood

Under normality, the loglikelihood takes the form

$$L(\beta, \theta, \psi) = \text{constant} - (1/2)\log|\Omega| - (1/2)(y - X\beta)'\Omega^{-1}(y - X\beta). \tag{2.16}$$

Maximizing $L$ with respect to $\beta$ and $\theta$ is assumed to be (relatively) easy, while maximization with respect to $\psi$ is more difficult. We write

$$\mu = X\beta. \tag{2.17}$$

Upon differentiating $\mu$ we obtain

$$d\mu = Xd\beta + (dX)\beta = Xd\beta + (\beta' \otimes I_n)Zd\psi, \tag{2.18}$$

where

$$Z = \partial \operatorname{vec} X / \partial \psi'. \tag{2.19}$$

Differentiating the loglikelihood then gives

$$\begin{aligned} dL = &-(1/2)\operatorname{tr}(\Omega^{-1}d\Omega) + (1/2)(y - X\beta)'\Omega^{-1}(d\Omega)\Omega^{-1}(y - X\beta) \\ &+ (y - X\beta)'\Omega^{-1}Xd\beta + (y - X\beta)'\Omega^{-1}(dX)\beta. \end{aligned} \tag{2.20}$$

It follows from (2.20) that the first-order conditions are

$$\begin{aligned} (y - X\beta)'\Omega^{-1}Xd\beta &= 0, \\ (y - X\beta)'\Omega^{-1}(d\Omega)\Omega^{-1}(y - X\beta) &= \operatorname{tr}(\Omega^{-1}d\Omega), \\ (y - X\beta)'\Omega^{-1}(dX)\beta &= 0, \end{aligned} \tag{2.21}$$

for $\beta$, $\theta$, and $\psi$, respectively. This implies that $\hat{\beta}$ takes the simple form

$$\hat{\beta}(\theta, \psi) = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y, \tag{2.22}$$

so that we can concentrate the likelihood with respect to $\beta$. The concentrated loglikelihood is

$$L^* = L(\theta, \psi) = \text{constant} - (1/2)\log|\Omega| - (1/2)\hat{u}'\Omega^{-1}\hat{u}, \tag{2.23}$$

where

$$\hat{u} = y - X\hat{\beta} = y - X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y.$$

For fixed $\psi$ we have $d\psi = 0$ and

$$\begin{aligned} dL^* = &-(1/2)\,\mathrm{tr}(\Omega^{-1}d\Omega) + (1/2)\hat{u}'\Omega^{-1}(d\Omega)\Omega^{-1}\hat{u} \\ &- \hat{u}'\Omega^{-1}X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}(d\Omega)\Omega^{-1}\hat{u}, \end{aligned} \tag{2.24}$$

using the fact that

$$\begin{aligned} d\hat{\beta} &= [d(X'\Omega^{-1}X)^{-1}]X'\Omega^{-1}y + (X'\Omega^{-1}X)^{-1}d(X'\Omega^{-1}y) \\ &= -(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}(d\Omega)\Omega^{-1}\hat{u}. \end{aligned} \tag{2.25}$$

## 2.6 Estimation using the concentrated likelihood

Given (2.23), we can obtain the ML estimates of the unknown parameters under normality by minimizing

$$L^* = \log|\Omega| + (y - X\beta)'\Omega^{-1}(y - X\beta). \tag{2.26}$$

Given the special structure of $\Omega$ this function also takes a convenient form:

**Proposition 2.6.1.** *We have*

$$\begin{aligned} L^* = &\log|\Delta_1| + (N-1)\log|\Delta_2| + (T-1)\log|\Delta_3| + (N-1)(T-1)\log|\Delta_4| \\ &+ (1/N)(1/T)(\sum_{i,t} v_{it})'(\Delta_1^{-1} - \Delta_2^{-1} - \Delta_3^{-1} + \Delta_4^{-1})(\sum_{i,t} v_{it}) \\ &+ (1/T)\sum_i (\sum_t v_{it})'(\Delta_2^{-1} - \Delta_4^{-1})(\sum_t v_{it}) \\ &+ (1/N)\sum_t (\sum_i v_{it})'(\Delta_3^{-1} - \Delta_4^{-1})(\sum_i v_{it}) + \sum_{i,t} v_{it}'\Delta_4^{-1}v_{it}, \end{aligned}$$

*where $v_{it} = \bar{y}_{it} - \bar{X}_{it}\beta$. In addition, we have*

$$\begin{aligned} X'\Omega^{-1}X = &(1/N)(1/T)(\sum_{i,t} X_{it})'(\Delta_1^{-1} - \Delta_2^{-1} - \Delta_3^{-1} + \Delta_4^{-1})(\sum_{i,t} X_{it}) \\ &+ (1/T)\sum_i (\sum_t X_{it})'(\Delta_2^{-1} - \Delta_4^{-1})(\sum_t X_{it}) \\ &+ (1/N)\sum_t (\sum_i X_{it})'(\Delta_3^{-1} - \Delta_4^{-1})(\sum_i X_{it}) + \sum_{i,t} X_{it}'\Delta_4^{-1}X_{it}. \end{aligned}$$

**Proof:** Let $e_i^{(N)}$ denote the $i$th column of $I_N$ and let $e_t^{(T)}$ denote the $t$th column of $I_T$, where $I_N$ and $I_T$

are identity matrices of orders $N$ and $T$, respectively. Then, we can write

$$v = \sum_{i=1}^{N}\sum_{t=1}^{T} e_t^{(T)} \otimes e_i^{(N)} \otimes v_{it}, \qquad X = \sum_{i=1}^{N}\sum_{t=1}^{T} e_t^{(T)} \otimes e_i^{(N)} \otimes X_{it},$$

and

$$\begin{aligned}\Omega^{-1} = {} & J_T \otimes J_N \otimes \Delta_1^{-1} + J_T \otimes (I_N - J_N) \otimes \Delta_2^{-1} \\ & + (I_T - J_T) \otimes J_N \otimes \Delta_3^{-1} + (I_T - J_T) \otimes (I_N - J_N) \otimes \Delta_4^{-1}.\end{aligned}$$

Noting that the transpose of a Kronecker product is the Kronecker product of transposes: $(A \otimes B)' = A' \otimes B'$, and the mixed-product property of Kronecker products, we obtain

$$\begin{aligned}(a_1 \otimes b_1 \otimes C_1)'(D \otimes E \otimes F)(a_2 \otimes b_2 \otimes C_2) &= (a_1' \otimes b_1' \otimes C_1')(D \otimes E \otimes F)(a_2 \otimes b_2 \otimes C_2) \\ &= \big((a_1' \otimes b_1')(D \otimes E)\big) \otimes (C_1'F)(a_2 \otimes b_2 \otimes C_2) \\ &= (a_1'D \otimes b_1'E \otimes C_1'F)(a_2 \otimes b_2 \otimes C_2) \\ &= (a_1'Da_2) \otimes (b_1'Eb_2) \otimes (C_1'FC_2) \\ &= (a_1'Da_2)(b_1'Eb_2)(C_1'FC_2),\end{aligned}$$

where $a_1$ and $a_2$ are vectors of order $T \times 1$, $b_1$ and $b_2$ are vectors of order $N \times 1$, $D$, $E$, and $F$ are square matrices of orders $T$, $N$, and $p$, respectively. $C_1$ and $C_2$ can be vectors of order $p \times 1$ or matrices of order $p \times k$. The last equality comes from the observation that $a_1'Da_2$ and $b_1'Eb_2$ are both scalars.

We substitute $a_1$, $a_2$ with $e_t^{(T)}$, $e_s^{(T)}$; $b_1$, $b_2$ with $e_j^{(N)}$, $e_i^{(N)}$; $D$ with $J_T$ or $I_T - J_T$, and $E$ with $J_N$ or $I_N - J_N$. This gives

$$\begin{aligned}e_t^{(T)'} J_T e_s^{(T)} &= 1/T, \\ e_t^{(T)'}(I_T - J_T)e_s^{(T)} &= \delta_{st} - 1/T, \\ e_i^{(N)'} J_N e_j^{(N)} &= 1/N, \\ e_i^{(N)'}(I_N - J_N)e_j^{(N)} &= \delta_{ij} - 1/N,\end{aligned}$$

where $\delta_{ij}$ and $\delta_{st}$ denote the Kronecker $\delta$, that is, $\delta_{ij} = 1$ if $i = j$ and zero otherwise; and $\delta_{st} = 1$ if $s = t$ and zero otherwise.

We obtain

$$\begin{aligned}v'\Omega^{-1}v = {} & \sum_{i,j,s,t}(1/T)(1/N)v_{it}'\Delta_1^{-1}v_{js} + \sum_{i,j,s,t}(1/T)(\delta_{ij} - 1/N)v_{it}'\Delta_2^{-1}v_{js} \\ & + \sum_{i,j,s,t}(\delta_{st} - 1/T)(1/N)v_{it}'\Delta_3^{-1}v_{js} \\ & + \sum_{i,j,s,t}(\delta_{st} - 1/T)(\delta_{ij} - 1/N)v_{it}'\Delta_4^{-1}v_{js}.\end{aligned}$$

Reorganizing terms, we have

$$\begin{aligned} v'\Omega^{-1}v = (1/T)(1/N)\sum_{i,j}\sum_{t,s} v_{it}'\left(\Delta_1^{-1}-\Delta_2^{-1}-\Delta_3^{-1}+\Delta_4^{-1}\right)v_{js} \\ + (1/T)\sum_{i}\sum_{t,s} v_{it}'\left(\Delta_2^{-1}-\Delta_4^{-1}\right)v_{is} \\ + (1/N)\sum_{i,j}\sum_{t} v_{it}'\left(\Delta_3^{-1}-\Delta_4^{-1}\right)v_{jt} + \sum_{i}\sum_{t} v_{it}'\Delta_4^{-1}v_{it}. \end{aligned}$$

Similarly,

$$\begin{aligned} X'\Omega^{-1}X = &\sum_{i,j,s,t}(1/T)(1/N)X_{it}'\Delta_1^{-1}X_{js} + \sum_{i,j,s,t}(1/T)(\delta_{ij}-1/N)X_{it}'\Delta_2^{-1}X_{js} \\ &+ \sum_{i,j,s,t}(\delta_{st}-1/T)(1/N)X_{it}'\Delta_3^{-1}X_{js} \\ &+ \sum_{i,j,s,t}(\delta_{st}-1/T)(\delta_{ij}-1/N)X_{it}'\Delta_4^{-1}X_{js} \\ = &(1/N)(1/T)(\sum_{i,t}X_{it})'(\Delta_1^{-1}-\Delta_2^{-1}-\Delta_3^{-1}+\Delta_4^{-1})(\sum_{i,t}X_{it}) \\ &+ (1/T)\sum_{i}(\sum_{t}X_{it})'(\Delta_2^{-1}-\Delta_4^{-1})(\sum_{t}X_{it}) \\ &+ (1/N)\sum_{t}(\sum_{i}X_{it})'(\Delta_3^{-1}-\Delta_4^{-1})(\sum_{i}X_{it}) + \sum_{i,t}X_{it}'\Delta_4^{-1}X_{it}. \end{aligned}$$

The decomposition of the concentrated likelihood comes from the determinant of $\Omega$ shown in Proposition 2.4.1 and the derivation of $v'\Omega^{-1}v$. ‖

## 2.7 The variance

It follows from (2.20) that

$$\begin{aligned} d^2L = &(1/2)\operatorname{tr}(\Omega^{-1}d\Omega)^2 - (y-X\beta)'\Omega^{-1}(d\Omega)\Omega^{-1}(d\Omega)\Omega^{-1}(y-X\beta) \\ &- (d\mu)'\Omega^{-1}(d\mu) - 2(y-X\beta)'\Omega^{-1}(d\Omega)\Omega^{-1}(d\mu) + (y-X\beta)'\Omega^{-1}(d^2\mu) \\ &- (1/2)\operatorname{tr}(\Omega^{-1}d^2\Omega) + (1/2)(y-X\beta)'\Omega^{-1}(d^2\Omega)\Omega^{-1}(y-X\beta). \end{aligned}$$

Minus the expectation of the second differential takes the simple form

$$-\operatorname{E}(d^2L) = (1/2)\operatorname{tr}(\Omega^{-1}d\Omega)^2 + (d\mu)'\Omega^{-1}(d\mu), \tag{2.27}$$

which implies that the information matrix will be block-diagonal in $(\beta,\psi)$ and $\theta$. This shows that we don't have to take the variance of the maximum likelihood (ML) estimator $\hat{\theta}$ into account when

calculating the variance of the ML estimators $(\hat{\beta}, \hat{\psi})$. Thus, writing

$$(d\mu)'\Omega^{-1}(d\mu) = \begin{pmatrix} d\beta \\ d\psi \end{pmatrix}' \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \begin{pmatrix} d\beta \\ d\psi \end{pmatrix}, \tag{2.28}$$

where

$$V_{11} = X'\Omega^{-1}X, \qquad V_{12} = V_{21}' = X'\left(\beta' \otimes \Omega^{-1}\right) Z, \tag{2.29}$$

and

$$V_{22} = Z'\left(\beta\beta' \otimes \Omega^{-1}\right) Z, \tag{2.30}$$

we obtain estimators of the variances of $\hat{\beta}$ and $\hat{\psi}$ as

$$\widehat{\text{var}}(\hat{\beta}) = V_{11}^{-1} + V_{11}^{-1}V_{12}\left(V_{22} - V_{21}V_{11}^{-1}V_{12}\right)^{-1} V_{21}V_{11}^{-1} \tag{2.31}$$

and

$$\widehat{\text{var}}(\hat{\psi}) = \left(V_{22} - V_{21}V_{11}^{-1}V_{12}\right)^{-1}, \tag{2.32}$$

where the parameters in the $V_{ij}$ matrices are replaced by their estimators.

The variance matrix $\Omega = \text{var}(u)$ is of a very large dimension, but the error components structure allows us to write it in a convenient form, allowing simple expressions for its inverse and determinant; see Proposition 2.4.1. We also need simple expressions for quadratic forms like $v'\Omega^{-1}v$ and $X'\Omega^{-1}X$. These are provided in Proposition 2.6.1.

Estimation of the parameters then proceeds as follows. For given $\psi$ we maximize the concentrated likelihood (2.23) with respect to the variance parameters $\theta$, where using the explicit expression (2.24) for the gradient will speed up the optimization. Performing a grid search on $\psi$ we obtain the ML estimates $\hat{\theta}$ and $\hat{\psi}$. Then we find $\hat{\beta}$ from (2.22). Finally, the estimated variances of $\hat{\beta}$ and $\hat{\psi}$ are obtained from (2.31) and (2.32).

## 2.8 Conclusion

In this chapter we have introduced an estimation procedure for combined cross-sectional and time-series data models with a multivariate error components structure and independent variables depending on additional parameters. The assumption of mutually independent error components facilitates the decomposition of the huge variance matrix into sums of Kronecker products of much smaller matrices. The estimation of the additional parameter can be conducted by a combination of grid search and concentrated likelihood. The variance of such a parameter is derived so that statistical tests become feasible.

# Chapter 3

# Data collection

## 3.1 Introduction

In the development of an economic model, the availability and quality of data often plays an important role. Various choices have to be made and various obstacles have to be overcome. In this chapter we provide more details of the data set used in Chapter 4, introduce the data collection procedure, detailing the complexity of integrating data from multiple sources and the motivation behind our choices involved in the process.

We are interested in the effect of earthquake risk on property prices in major cities in Japan. We select five Japanese cities/areas for our purpose (see Section 3.2): Tokyo Metropolitan Area (23 special wards), Osaka City, Nagoya City, Fukuoka City and Sapporo City. We shall refer to the Tokyo Metropolitan Area as a city, although officially it is an area, not a city.

Each city is divided into wards and each ward is divided into districts. Certain information is available per ward, which can affect the attractiveness of buying a property in that ward. For example, population characteristics, information about schools and medical facilities, shopping, safety, etc. These 'attractiveness' characteristics are described in Section 3.3.

We distinguish between three types of properties: 'residential land (land and building)', 'residential land (land only)', and condominium. Sales prices and property characteristics are available for each of these types in each of the five cities. These are described in Section 3.4. We do not know the exact location of a property, but we do know in which district the property lies and we also know the distance to the nearest station and the name of that station.

Some macro variables are relevant and affect house prices nationally. These variables are described in Section 3.5.

We next come to the earthquake and risk data. Historical earthquake data are described in Section 3.6. The estimation and simulation of ETAS models based on the historical earthquake data are introduced in Section 3.7.

Japan is geographically split up in meshes of varying size. The largest (first mesh) is $80 \times 80$ km, the smallest (quarter mesh) is $250 \times 250$ meter. The data on these meshes are described in Section 3.8. While historical earthquake data are described in Section 3.6, earthquake *risk* data are described in Section 3.9. Finally, we describe how to link stations and districts to these meshes in Section 3.10.

## 3.2 Cities

Japan has twelve cities with populations of more than one million people. Almost 100 million Japanese, or 78% of the country's total population of 127.4 million, live in urban areas. The total population of Japan's largest 103 cities amounts to 63.9 million or just over half of all the country's residents. Tokyo, with almost nine million inhabitants, is often referred to as a city, but is officially known and governed as a 'metropolitan prefecture'. With a population of 3.7 million, Yokohama, south of Tokyo, is Japan's second largest city. It is the country's largest port and a manufacturing and ship building centre. Japan's third-largest city is Osaka with 2.7 million inhabitants. It is the country's third most important seaport and home to many leading Japanese manufacturers. Nagoya (2.3 million inhabitants) is the center of the Chukyo Metropolitan Area and is home to the Mitsubishi Aircraft Company and the Toyota factory. Eight cities have between one and two million inhabitants: Sapporo, Kobe, Fukuoka, Kyoto, Kawasaki, Saitama, Hiroshima, and Sendai.

From these twelve cities we selected five: Sapporo, Tokyo, Nagoya, Osaka, and Fukuoka. These five cities provide a good representation of the major cities in Japan, in terms of geographical spread (Sapporo in the North, Fukuoka in the South) and earthquake risk (Tokyo highest, then Osaka and Nagoya, then Sapporo and Fukuoka).

We excluded Yokohama, Kawasaki, and Saitama, because they are located in the same metropolitan area as Tokyo, the 'Kanto' area. Similarly, we excluded Kobe and Kyoto, because they are located in the same metropolitan area as Osaka, the 'Keihanshin' area. Thus, each of the three major metropolitan areas is represented: the greater Tokyo area (Tokyo, Yokohama, Kawasaki, Saitama) by Tokyo, the Kansai region (Osaka, Kobe, Kyoto) by Osaka, and the Chukyo metropolitan area by Nagoya.

To obtain a representative geographical spread we added Sapporo, the largest city in the North, and Fukuoka, the second largest city in the West after Osaka.

Hiroshima was excluded because the metro system is not sufficiently dense to identify the properties, and Sendai because it lies in the 2011 Fukushima disaster area and property prices there are completely distorted.

## 3.3 Wards and attractiveness characteristics

A designated city is a Japanese city that has a population greater than 500,000 and has been designated as such by order of the Cabinet of Japan. Designated cities are delegated many of the tasks normally performed by prefectural governments, such as public education, social welfare, sanitation, business licensing, and urban planning. Designated cities are required to subdivide themselves into wards ('ku'), each of which has a ward office conducting various administrative functions for the city government. The 23 special wards of Tokyo are not part of this system, as Tokyo is a prefecture, and its wards are effectively independent cities. The five cities together contain 80 wards: 24 in Osaka, 23 in Tokyo, 16 in Nagoya, 10 in Sapporo, and 7 in Fukuoka.

When considering to buy a property in a given city we are likely to be interested in certain characteristics of these wards, in particular characteristics that make one ward more or less attractive than another. Information about wards can be downloaded from

www.e-stat.go.jp/SG1/estat/eStatTopPortalE.do,

the portal site of official statistics of Japan, under the category 'Regional Statistics'. The English version of this website provides the statistics for one year (not the same year for each variable), while the Japanese version provides 'time-series' data. Data are available in 11 categories: A. Population and households; B. Natural environment; C. Economic base; D. Administrative base; E. Education; F. Labor; G. Culture and sports; H. Dwelling; I. Health and medical care; J. Welfare and social security; and K. Safety. More detailed information on these categories (in Japanese) is available from

www.e-stat.go.jp/SG1/chiiki/FileStream.do?file=koumoku.html.

Some variables (such as unemployment and the number of traffic accidents) are updated (or adjusted) annually, while others (such as population and households) are only updated after each 5-year census. The data cover the period 2007–2015, so in order to obtain a general indicator that reflects the attractiveness over the whole sample period of our housing price data (2006–2015), a simple average is calculated. In case of a missing record the nearest data available are used as a proxy; for example, data for year 2006 are assumed to be the same as data for 2007. In this way we construct *one* (time-independent) value for each item in each ward.

Table 3.1: Attractiveness characteristics by ward

| Category | Item |
|---|---|
| Population | % younger than 15 years |
| | % older than 65 years |
| | % immigrants |
| | % emigrants |
| | % foreigners |
| | % private households |
| | % nuclear families |
| | % one-person households |
| Schools, culture, & welfare | number of daycare nurseries |
| | number of schools (kindergarten) |
| | number of schools (primary) |
| | number of schools (junior/senior high) |
| | number of students per teacher (primary school) |
| | number of homes for the aged |
| | number of community halls |
| | number of libraries |
| Medical facilities | number of general hospitals |
| | number of physicians |
| | number of dentists |
| | number pharmacists |
| Safety | number of traffic accidents |
| | number of criminal offenses |
| Shopping | number of large-scale retail stores |
| | number of department stores |
| | annual sales of commercial goods |
| Housing | % privately owned houses |
| | habitable land area (% of total area) |
| Employment | unemployment ratio |
| | % of self-employed |
| | % of executives |

For our purpose we selected the variables listed in Table 3.1. These are the ward characteristics that might have explanatory power on property prices, representative of each category. Note that not all the listed variables are actually used in the base model or the sensitivity analysis; instead of adding all variables with significant regression coefficients in the model, we choose a relatively small subset of variables with low correlation between each other. Our goal is not to find as much variables as possible to explain property prices, but to focus on the risk variables with some auxiliary variables in the model.

Below we explain how these items were constructed from the available data.

### 3.3.1 Population

**% younger than 15 years:**

The percentage of population younger than 15 years is obtained as the ratio of 'population younger than 15 years old' to 'total population' of each ward. The data are constant for years 2007–2011 and for years 2012–2015. Since we only want a time-independent indicator, simple averaging and extrapolation is used, so that for each ward

$$\text{percentage U15 population} = [6 \times \text{PctU15}_{2007} + 4 \times \text{PctU15}_{2012}]/10,$$

where $\text{PctU15}_t$ denotes the percentage in year $t$, the first term represents the percentage for the first 6 years (2006–2011) and the second for the last 4 years (2012–2015).

**% older than 65 years:**

Obtained as the ratio of 'population aged 65 years or more' to 'total population'. The data is constant for years 2007–2011 and for years 2012–2015. The indicator is thus averaged in the same pattern as the item '% younger than 15 years'.

**% immigrants:**

Obtained as the ratio of 'number of immigrants from other municipalities' to 'total population'. The numbers are varying each year. We assume that data for years 2005 and 2006 are the same as year 2007.

**% emigrants:**

Obtained as the ratio of 'number of emigrants to other municipalities' to 'total population'. The numbers are varying each year. The data is missing in 2007–2011 for wards in Sapporo and in 2008–2011 for wards in Nagoya. For Sapporo we use year 2012 data as a substitute for the missing values. For Nagoya we use year 2007 data to substitute missing values in 2008–2009 and the year 2012 data to substitute those in 2010–2011.

**% foreigners:**

Obtained as the ratio of 'number of foreigners' to 'total population'. The numbers are constant for years 2007–2011 and for years 2012–2015.

**% private households:**

Obtained as the ratio of 'number of private households' to 'number of households'. The numbers are constant for years 2007–2011 and for years 2012–2015.

**% nuclear families:**

Obtained as the ratio of 'number of nuclear family households' to 'number of households'. The numbers are constant for years 2007–2011 and for years 2012–2015.

**% one-person households:**

Obtained as the ratio of 'number of one-person households' to 'number of households'. The numbers are constant for years 2007–2011 and for years 2012–2015.

### 3.3.2 Schools, culture, welfare

**number of daycare nurseries (per inhabitable area):**

The variable is obtained as the number of daycare nurseries per square km of inhabitable area.

**number of schools (kindergarten) (per inhabitable area):**

The numbers of kindergartens per square km of inhabitable area are missing in years 2006–2007 for all wards in Sapporo, so we approximate numbers of these years with those of 2008. The numbers are missing in year 2010 for all wards in Nagoya, so we approximate this as the mean numbers from years 2009 and 2011.

**number of schools (primary) (per inhabitable area):**

The number of primary schools is divided by the area of inhabitable land.

**number of schools (junior/senior high)**
**(per inhabitable area):**

The total number of junior and senior high schools is divided by the area of inhabitable land. For senior high schools, the numbers for Nagoya and Fukuoka are missing in year 2010; we use the average of 2009 and 2011 data as an proxy.

**number of students per teacher (primary school):**

The ratio of students to teachers for primary schools.

**number of homes for the aged (per 65+ population):**

The number of homes for the aged divided by 1/1000 times the population over 65 years old.

**number of community halls (per capita):**

The number of community halls is divided by 1/10000 of the total population. The data is missing for Sapporo in 2007, which we approximate using 2008 data.

**number of libraries (per capita):**

The number of libraries is divided by 1/10000 of the total population. The data is missing for Sapporo in 2007, which we approximate using 2008 data.

### 3.3.3 Medical facilities

**number of general hospitals (per capita):**

The number of general hospitals is divided by 1/10000 times the total population.

**number of physicians (per capita):**

The number of physicians is divided by 1/100 times the total population.

**number of dentists (per capita):**

The number of dentists is divided by 1/100 times the total population.

**number pharmacists (per capita):**

The number of pharmacists is divided by 1/100 times the total population.

### 3.3.4 Safety

**number of traffic accidents (per capita):**

The number of traffic accidents is divided by 1/100 times the total population. The numbers are unknown for years 2012–2015, which we approximate with the numbers from year 2011.

**number of criminal offenses (per capita):**

The number of criminal offenses is divided by 1/100 times the total population. The numbers are unknown for years 2012–2015, which we approximate with the numbers from year 2011.

### 3.3.5 Shopping

**number of large-scale retail stores (per capita):**

The number of large-scale retail stores is divided by 1/1000 times the total population.

**number of department stores (per capita):**

The number of department stores is divided by 1/10000 times the total population.

**annual sales of commercial goods (per capita):**

The annual sales of commercial goods are divided by the total population. The unit is in 1000 yen.

### 3.3.6 Housing

**% privately owned houses:**

Obtained as the ratio of privately owned houses per dwelling building. It is constant for years 2011–2015 so we use this number to approximate the missing data for years before 2011.

**habitable land area (% of total area):**

Obtained as the ratio of inhabitable land area to total area. It is constant for years 2007–2012 and for years 2013–2015.

### 3.3.7 Employment

**unemployment ratio (per labor supply population):**

Obtained as the ratio of unemployed population to the population of labor supply.

**% of self-employed (per labor supply population):**

Obtained as the ratio of self-employed (including those with and without employee) to the population of labor supply.

**% of executives (per labor supply population):**

Obtained as the ratio of number of executives to the population of labor supply.

### 3.3.8 Variable selected

The summary statistics of the variables finally selected for the estimation and sensitivity analysis are shown below.

Table 3.2: Summary statistics of key ward characteristics

| city | mean | min | 25% | 50% | 75% | max | sd |
|---|---|---|---|---|---|---|---|
| **PctImmi** | | | | | | | |
| Tokyo | 0.077 | 0.043 | 0.059 | 0.075 | 0.085 | 0.133 | 0.024 |
| Osaka | 0.063 | 0.031 | 0.047 | 0.051 | 0.071 | 0.138 | 0.028 |
| Nagoya | 0.063 | 0.039 | 0.049 | 0.059 | 0.070 | 0.112 | 0.019 |
| Fukuoka | 0.079 | 0.065 | 0.066 | 0.068 | 0.087 | 0.110 | 0.019 |
| Sapporo | 0.065 | 0.050 | 0.054 | 0.062 | 0.072 | 0.103 | 0.016 |
| **NCrime** | | | | | | | |
| Tokyo | 2.544 | 1.275 | 1.447 | 1.671 | 3.453 | 11.267 | 2.124 |
| Osaka | 3.600 | 1.817 | 2.242 | 2.649 | 3.088 | 14.794 | 2.927 |
| Nagoya | 3.098 | 1.979 | 2.050 | 2.621 | 3.106 | 9.615 | 1.872 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Fukuoka | 2.354 | 1.743 | 1.864 | 1.992 | 2.733 | 3.546 | 0.765 |
| Sapporo | 1.415 | 0.842 | 1.221 | 1.338 | 1.517 | 2.537 | 0.454 |
| **PctUnemploy** | | | | | | | |
| Tokyo | 0.054 | 0.029 | 0.047 | 0.054 | 0.061 | 0.072 | 0.012 |
| Osaka | 0.102 | 0.074 | 0.085 | 0.098 | 0.111 | 0.202 | 0.026 |
| Nagoya | 0.055 | 0.045 | 0.047 | 0.055 | 0.060 | 0.074 | 0.009 |
| Fukuoka | 0.068 | 0.063 | 0.066 | 0.067 | 0.067 | 0.078 | 0.005 |
| Sapporo | 0.071 | 0.064 | 0.069 | 0.070 | 0.073 | 0.079 | 0.004 |
| **PctExec** | | | | | | | |
| Tokyo | 0.094 | 0.061 | 0.070 | 0.080 | 0.105 | 0.181 | 0.033 |
| Osaka | 0.064 | 0.032 | 0.047 | 0.056 | 0.072 | 0.127 | 0.024 |
| Nagoya | 0.072 | 0.048 | 0.060 | 0.067 | 0.082 | 0.116 | 0.017 |
| Fukuoka | 0.052 | 0.044 | 0.045 | 0.048 | 0.053 | 0.075 | 0.011 |
| Sapporo | 0.058 | 0.050 | 0.052 | 0.055 | 0.057 | 0.084 | 0.010 |
| **PctForeign** | | | | | | | |
| Tokyo | 0.029 | 0.014 | 0.020 | 0.025 | 0.030 | 0.064 | 0.014 |
| Osaka | 0.037 | 0.015 | 0.018 | 0.024 | 0.035 | 0.208 | 0.040 |
| Nagoya | 0.023 | 0.014 | 0.016 | 0.022 | 0.027 | 0.050 | 0.009 |
| Fukuoka | 0.011 | 0.006 | 0.007 | 0.009 | 0.016 | 0.019 | 0.005 |
| Sapporo | 0.003 | 0.002 | 0.002 | 0.003 | 0.004 | 0.006 | 0.001 |
| **Nhosp** | | | | | | | |
| Tokyo | 0.643 | 0.257 | 0.372 | 0.460 | 0.685 | 3.707 | 0.690 |
| Osaka | 0.779 | 0.326 | 0.566 | 0.670 | 0.937 | 1.412 | 0.300 |
| Nagoya | 0.627 | 0.212 | 0.443 | 0.592 | 0.740 | 1.664 | 0.332 |
| Fukuoka | 0.739 | 0.544 | 0.578 | 0.640 | 0.902 | 1.027 | 0.200 |
| Sapporo | 0.991 | 0.738 | 0.825 | 0.932 | 1.032 | 1.769 | 0.298 |
| **Ndaycare** | | | | | | | |
| Tokyo | 1.868 | 0.541 | 1.572 | 1.827 | 2.251 | 2.844 | 0.569 |
| Osaka | 1.667 | 0.360 | 1.212 | 1.764 | 2.087 | 2.924 | 0.626 |
| Nagoya | 0.994 | 0.501 | 0.775 | 0.982 | 1.117 | 1.658 | 0.308 |
| Fukuoka | 0.815 | 0.511 | 0.681 | 0.897 | 0.958 | 1.019 | 0.198 |
| Sapporo | 0.492 | 0.136 | 0.349 | 0.518 | 0.652 | 0.748 | 0.203 |
| **Nkindergtn** | | | | | | | |
| Tokyo | 1.419 | 0.811 | 1.057 | 1.333 | 1.658 | 2.476 | 0.441 |
| Osaka | 1.072 | 0.244 | 0.676 | 0.936 | 1.384 | 2.504 | 0.569 |
| Nagoya | 0.713 | 0.197 | 0.506 | 0.743 | 0.899 | 1.226 | 0.273 |
| Fukuoka | 0.704 | 0.281 | 0.397 | 0.693 | 1.009 | 1.140 | 0.353 |
| Sapporo | 0.399 | 0.176 | 0.291 | 0.343 | 0.501 | 0.625 | 0.153 |
| **Nagedhome** | | | | | | | |
| Tokyo | 0.178 | 0.122 | 0.153 | 0.174 | 0.197 | 0.290 | 0.040 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Osaka | 0.211 | 0.072 | 0.176 | 0.216 | 0.244 | 0.329 | 0.063 |
| Nagoya | 0.231 | 0.098 | 0.193 | 0.222 | 0.272 | 0.395 | 0.082 |
| Fukuoka | 0.294 | 0.176 | 0.279 | 0.292 | 0.312 | 0.407 | 0.068 |
| Sapporo | 0.227 | 0.175 | 0.179 | 0.203 | 0.219 | 0.428 | 0.078 |
| **Ndepstore** | | | | | | | |
| Tokyo | 0.239 | 0.074 | 0.112 | 0.175 | 0.222 | 1.286 | 0.264 |
| Osaka | 0.179 | 0.000 | 0.078 | 0.109 | 0.174 | 1.112 | 0.223 |
| Nagoya | 0.276 | 0.076 | 0.180 | 0.249 | 0.307 | 0.741 | 0.162 |
| Fukuoka | 0.191 | 0.023 | 0.098 | 0.213 | 0.242 | 0.419 | 0.133 |
| Sapporo | 0.198 | 0.079 | 0.128 | 0.181 | 0.216 | 0.474 | 0.113 |
| **Nlargeretail** | | | | | | | |
| Tokyo | 0.305 | 0.094 | 0.112 | 0.119 | 0.248 | 2.132 | 0.451 |
| Osaka | 0.208 | 0.070 | 0.109 | 0.141 | 0.181 | 1.022 | 0.207 |
| Nagoya | 0.196 | 0.100 | 0.126 | 0.155 | 0.213 | 0.638 | 0.130 |
| Fukuoka | 0.176 | 0.082 | 0.106 | 0.135 | 0.228 | 0.344 | 0.108 |
| Sapporo | 0.204 | 0.159 | 0.175 | 0.184 | 0.208 | 0.353 | 0.056 |

It can be seen from Table 3.2 that there is sufficient variation in each variable. The distribution pattern between different cities is also very different. For example, the percentage of immigration (PctImmi) for the districts in Tokyo is on average higher than Osaka, but the standard deviation is smaller.

## 3.4 Property prices and determinants

### 3.4.1 The MLIT data set

In our study we shall work with sales prices rather than with rental prices, because we believe sales are more permanent than rentals and therefore the effect of earthquake risk on choosing the property will be more informative.

Nakagawa *et al.* (2009) use land prices over various years (from 1980 onwards) and describe the data in their Section 3 (for the Tokyo area). Their data are based on the Koji-Chika data set published by the Ministry of Land, Infrastructure, Transport, and Tourism. The well-known Koji-Chika set provides *fictional* sales prices (as produced by 'experts') and they are only available at annual intervals, which we consider to be too long.

Thus we shall use a different data set, which provides self-reported transaction prices at three-months intervals. This data set known as the 'Real estate transaction-price information' and is provided by the Ministry of Land, Infrastructure, Transport and Tourism (MLIT); see

`www.land.mlit.go.jp/webland_english/servlet/MainServlet`,

The information in this data set is based on the results of a questionnaire survey of persons involved in real estate transactions conducted by MLIT, compiled and published quarterly.

The Real Estate Transaction Questionnaire Survey was conducted for government ordinance-designed major cities of three metropolitan areas (including the 23 special wards of Tokyo, Osaka and Nagoya)

starting at the 3rd quarter of 2005. The survey region expanded to cover prefectural capitals (including Sapporo and Fukuoka) starting at the 2nd quarter of 2006. After 2007 2nd quarter, the prefectural office location cities of the whole country were included in the survey region.

In our analysis we use the data from 2nd Quarter 2006 to 3rd Quarter 2015, where all five cities are surveyed. Thus, in total, we have 38 quarters of observations.

We distinguish between three types of properties: (1) 'residential land (land and building)', hereafter 'land & building'; (2) 'residential land (land only)', hereafter 'land only'; and (3) 'pre-owned condominiums', hereafter 'condos'. We have data on 362658 properties of which approximately 44% are condo's, 34% are land & buildings, and 22% are land only.

Table 3.3: List of variables, LandMLIT data set: Name and description

| Variable name | Description |
|---|---|
| Type | condos/ land only/ land & building |
| Region | residential/ commercial/ industrial... |
| City/Town/Ward/Village code | postcode |
| Prefecture | |
| City/Town/Ward/Village | which ward is the property in |
| District | which district is the property in |
| Nearest station name | |
| Nearest station distance minutes | 80m/min, accurate to minute for 0–30 min |
| Transaction price, total | price in 10,000 yen |
| Layout | number of rooms, stories etc. |
| Area $m^2$ | (floor) area of the property, accurate to 5 or 10 $m^2$ |
| Transaction price unit price $m^2$ | |
| Land shape | |
| Frontage | length of land in contact of front road |
| Total floor area $m^2$ | |
| Year of construction | |
| Building structure | steel / concrete / wood... |
| Building use | family / office / factory... |
| Purpose of use | similar as above |
| Frontage road direction | |
| Frontage road classification | city road / prefectural road... |
| Frontage road breadth | |
| City planning | plans for the district |
| Max bldg coverage ratio | |
| Max floor area ratio | |
| Transaction period | date of contract |
| Remarks | other transaction-related issues |

In Table 3.3 we provide a list of the available variables in the housing data set, together with a short description.

### 3.4.2 Included/excluded variables and availability

Not all these variables are selected to be included in our study.

Table 3.4: Included variables, LandMLIT data set

| Variable name | Availability |
|---|---|
| Type | all types |
| City/Town/Ward/Village | all records |
| District | all records |
| Nearest station name | most records |
| Nearest station distance (minutes) | most records |
| Transaction price, total | all records |
| Area $m^2$ | all records |
| Total floor area $m^2$ | 'land & building' only |
| Year of construction | unknown for 'land only' |
| Building structure | unknown for 'land only' |
| Max bldg coverage ratio | all records |
| Max floor area ratio | all records |
| City planning for land use | all records |
| Transaction period | all records |

The selected variables are provided in Table 3.4. Obviously the transaction price is included since its logarithm is our dependent variable. The variable 'Max bldg coverage ratio' is important in the literature.

Table 3.5: Not included variables, LandMLIT data set

| Variable name | Availability |
|---|---|
| Region | unknown for condos |
| City/Town/Ward/Village code | all records |
| Prefecture | all records |
| Layout | condos only |
| Transaction price unit price $m^2$ | 'land only' only |
| Land shape | unknown for condos |
| Frontage | unknown for condos |
| Building use | unknown for 'land only' |
| Purpose of use | only after 2013 |
| Frontage road direction | unknown for condos |
| Frontage road classification | unknown for condos |
| Frontage road breadth | unknown for condos |
| Remarks | some records |

The variables that are excluded from our data set are listed in Table 3.5. In the variables 'Layout' and 'Building use' there are too many categories, and in 'Purpose of use' the data are only available for too short a period.

Table 3.6: Summary of number of wards, districts, properties per type, and stations

| City | Wards | Areas | Buildings | Land | Condos | Metro stations |
|---|---|---|---|---|---|---|
| Tokyo | 23 | 898 | 57568 | 33991 | 92518 | 482 |
| Osaka | 24 | 564 | 21064 | 6901 | 21855 | 220 |
| Nagoya | 16 | 1379 | 14640 | 13110 | 11029 | 159 |
| Fukuoka | 7 | 318 | 7847 | 5660 | 12475 | 75 |
| Sapporo | 10 | 551 | 11763 | 9461 | 11461 | 86 |
| Total | 80 | 3710 | 112882 | 69123 | 149338 | 1022 |

### 3.4.3 Sample selection

For all five cities and the whole sample period (2006Q2–2015Q3), there are 362,658 records (before sample selection). In choosing the sample, the following criteria are applied:

- We exclude all records where walking time to nearest station is longer than 30 minutes or nearest station is unknown.
- We exclude records with living area larger than 2000 square meters.
- In cases of 'pre-owned condominiums' and 'residential Land (land and building)', we exclude properties built before the war (1945).

After selection we are left with 91.4% of the original data, that is, 331,390 records. In addition, 47 records are apparently wrongly coded because the location information given by the 'district' and 'nearest station' do not match. We manually checked these records and decided that the information may not be accurate, so we exclude these from our sample. More information on how we verified the location information can be found in Subsection 3.10.5.

This leads to the summary statistics provided in Table 3.6. We emphasize that we do not know the exact location of a property. We only know two things about the location, namely the district in which the property lies and the name of and distance to the nearest station. In the five cities together there are 3710 districts and 1022 stations after applying the sample selection criteria mentioned above. So, in order to identify the location of a property, the district information is more accurate than the station information.

### 3.4.4 Property prices

For all records in our data, the total transaction value (unit: 10,000 yen) excluding overhead costs (such as agent's commission) is provided. Figures are rounded to two decimal places by the provider, but no other numerical adjustments were made.

The main quantiles of the distribution of the total transaction price per city and per type are given in Table 3.7. Property prices are highly skewed with the median well below the mean. Not surprisingly, Tokyo is the most expensive city, followed by Osaka and Nagoya. Cheapest are Fukuoka and Sapporo.

The cheapest property is a condo built in 1984 in the Nagayoshinagahara district of Hirano Ward, Osaka. It is a 1DK room of $40m^2$ and it was sold in 2012 for 530 yen (about 5 dollars and 30 cents).

Table 3.7: Total price, quantiles ($\times$ 10 million yen)

| City | 5% | 25% | 50% | 75% | 95% | $n$ |
|---|---|---|---|---|---|---|
| *Land & Building* | | | | | | |
| Tokyo | 1.5 | 3.7 | 5.0 | 7.5 | 34.0 | 57568 |
| Osaka | 0.5 | 1.6 | 3.1 | 4.5 | 26.0 | 21064 |
| Nagoya | 1.0 | 2.7 | 3.7 | 4.8 | 16.0 | 14640 |
| Fukuoka | 1.0 | 2.1 | 3.2 | 5.3 | 26.0 | 7847 |
| Sapporo | 0.7 | 1.6 | 2.6 | 3.8 | 14.0 | 11763 |
| *Land only* | | | | | | |
| Tokyo | 1.2 | 3.1 | 4.9 | 8.3 | 27.0 | 33991 |
| Osaka | 0.6 | 1.6 | 3.0 | 6.4 | 25.0 | 6901 |
| Nagoya | 0.8 | 1.8 | 2.7 | 4.5 | 13.0 | 13110 |
| Fukuoka | 0.6 | 1.5 | 2.2 | 4.5 | 16.0 | 5660 |
| Sapporo | 0.4 | 0.9 | 1.3 | 2.3 | 7.4 | 9461 |
| *Condo* | | | | | | |
| Tokyo | 0.7 | 1.6 | 2.5 | 3.8 | 7.0 | 92518 |
| Osaka | 0.4 | 1.0 | 1.6 | 2.2 | 3.6 | 21855 |
| Nagoya | 0.3 | 0.9 | 1.5 | 2.3 | 3.6 | 11029 |
| Fukuoka | 0.2 | 0.5 | 1.1 | 1.8 | 3.1 | 12475 |
| Sapporo | 0.2 | 0.6 | 1.1 | 1.7 | 2.7 | 11461 |

This is obviously a symbolic prize and one may make up an explanatory story, but we don't know the background. Such extremely cheap properties are rare in our sample. Out of our 331,343 sample records, there are 524 properties (0.16%) with a sales price under one million yen (about $10,000). This subsample of 524 properties are mostly small properties, but there are no apparent patterns in terms of location, distance to nearest station, region, city zone, or transaction period.

We attempted to find out a little more about these 'outliers'. The 'Land Economy and Construction and Engineering Industry Bureau' of the Ministry of Land, Infrastructure, Transport and Tourism (MLIT) told us that their questionnaire involves people involved in real estate transactions, not real estate agencies or organizations. The information totally relies on the answers in the questionnaire. The National Tax Agency (Osaka Region) told us that it is legally possible for properties to be sold for such low prices. But there are fiscal restrictions: if the estimated value of a property and the realized deviate too much (according to the tax authority), then this sale may be subject to gift or inheritance tax. Finally, two private real-estate agencies (one in Fukuoka, one in Osaka) told us that there might be special issues with these properties. For example, the owner went bankrupt and the creditor placed a mortgage on the property; or the property suffers from a psychologically defect (such as criminal homicide or suicide, in Japan such information must be provided by the seller); or that the deal includes the right of property (house) or land lease; or renovation is very expensive so that the previous owner sold the house at low price possibly to the real estate broker who would then renovate and sell for a higher price. We marked these 524 properties in our data set, so that we are able to do the analysis with and without these 'outliers'.

The most expensive property is a building in Ginza, Chuo Ward, Tokyo. It has a land area of $1200m^2$ and a total floor area larger than $2000m^2$. This property was sold in 2013 for 24,000 million yen (about

$240 million).

Table 3.8: Total log-price, quantiles

| City | 5% | 25% | 50% | 75% | 95% | Skew | $n$ |
|---|---|---|---|---|---|---|---|
| *Land & Building* | | | | | | | |
| Tokyo | 16.5 | 17.4 | 17.7 | 18.1 | 19.6 | 0.1 | 57568 |
| Osaka | 15.4 | 16.6 | 17.2 | 17.6 | 19.4 | −0.3 | 21064 |
| Nagoya | 16.1 | 17.1 | 17.4 | 17.7 | 18.9 | −0.1 | 14640 |
| Fukuoka | 16.1 | 16.9 | 17.3 | 17.8 | 19.4 | 0.1 | 7847 |
| Sapporo | 15.8 | 16.6 | 17.1 | 17.5 | 18.8 | −0.1 | 11763 |
| *Land only* | | | | | | | |
| Tokyo | 16.3 | 17.2 | 17.7 | 18.2 | 19.4 | 0.1 | 33991 |
| Osaka | 15.5 | 16.6 | 17.2 | 18.0 | 19.3 | 0.1 | 6901 |
| Nagoya | 15.8 | 16.7 | 17.1 | 17.6 | 18.7 | 0.1 | 13110 |
| Fukuoka | 15.6 | 16.5 | 16.9 | 17.6 | 18.9 | 0.3 | 5660 |
| Sapporo | 15.1 | 16.0 | 16.4 | 17.0 | 18.1 | 0.1 | 9461 |
| *Condo* | | | | | | | |
| Tokyo | 15.7 | 16.6 | 17.0 | 17.5 | 18.1 | −0.0 | 92518 |
| Osaka | 15.1 | 16.1 | 16.6 | 16.9 | 17.4 | −0.2 | 21855 |
| Nagoya | 14.9 | 16.0 | 16.5 | 17.0 | 17.4 | −0.1 | 11029 |
| Fukuoka | 14.6 | 15.5 | 16.2 | 16.7 | 17.2 | −0.2 | 12475 |
| Sapporo | 14.5 | 15.5 | 16.2 | 16.6 | 17.1 | −0.3 | 11461 |

The distribution of the property prices is clearly highly skewed. Hence, we also present the log-prices in Table 3.8. The log-prices are more symmetric, which is one reason why we choose log-prices to be our dependent variable. To analyze the symmetry of the log-prices we also present the 'skew' in Table 3.8. The skew is defined in terms of the quartiles as: $(Q3 - Q2) - (Q2 - Q1)$, where $Q1$, $Q2$, and $Q3$ denote the first, second, and third quartile, respectively. If this number is positive, then we say there is positive skew. Of the 15 items there is 1 without skew. Of the remaining 14, 7 have negative skew and 7 positive skew. The assumption of symmetry of log-prices therefore does not seem unreasonable.

We emphasize one other point. The legality of property sales works differently in different countries. In many countries, there are two contracts: the first when you agree on a price, the second when you actually exchange. Once the second contract is signed, the first becomes obsolete. There may be several months between the two contracts. In the first contract it would say that A intends to buy from B the following property for such and such a price, but under condition that a mortgage can be obtained, a property inspector will not find major faults, etc. So, it is binding under certain conditions. The buyer typically pays a percentage (like 10%) of the price as a guarantee. The second contract is signed when all is in order, the money is with the solicitor, and the house is empty. It is the second contract which is the official document and its date is the official date, even though the actual price has been negotiated and decided (much) earlier.

Fortunately, in Japan it works differently. There is only *one* purchase contract, which is signed after the price has been agreed on. If the buyer cancels the purchase after signing the contract, he/she loses the deposit, which is typically 10% of the price but can be lower (sometimes negotiable). In the case of a

condo that is typically sold before completion, the deposit is usually much lower (less than 5% typically). Banks provide preliminary review services before signing a purchase contract. There are non-negligible cases where they eventually decide not to provide loans, but the probability of this happening is low.

Important in our case is that the purchase date given in our data set corresponds to the moment when the price was agreed, not to the moment that the exchange of property/money takes place.

### 3.4.5 Housing characteristics as explanatory variables

#### Type

According to the LandMLIT website:

> *Real estate is divided into the following types: residential land, pre-owned condominiums, etc., agricultural land, and forest land. Residential land is further divided into two types of residential land (land only) and residential land (land and building). Transactions for residential land (land only) indicate the transactions for land only. Transactions for residential land (land and building) mean the package transactions for the land and buildings, etc.. Transactions for pre-owned condominiums, etc. are the transactions for condominium units (apartments, etc.).*

For pre-owned condos, the ward, district, nearest station, distance to nearest station, floor area, year of construction, building structure, building use, city planning, building coverage ratio, and floor area ratio are provided.

For residential land (land only), the region, ward, district, nearest station, distance to nearest station, area in square meters, unit land price, land shape, frontage of land, frontage road width/direction/classification, city planning, building coverage ratio, and floor area ratio are provided.

For residential land (land and building), the region, ward, district, nearest station, distance to nearest station, area in square meters, total floor area of building, frontage road width/direction/classification, year of construction, building structure, building use, city planning, building coverage ratio, and floor area ratio are provided.

#### Location information

For each record, the city, ward, and district where the record is located are specified. In addition, the name and walking distance of the nearest station are given. These are the only two measures of location of a given record available to us.

#### Time distance to nearest station

According to the LandMLIT website:

> *For the residential land (land only), residential land (land and building), and pre-owned condominiums, etc., the name of the nearest train station and the time distance (minute) from the location of the property to the nearest train station (for subway, to the ground entrance) are displayed. A time distance less than 30 minutes is displayed in minutes. A*

*time distance greater than 30 minutes is displayed in the following time periods: 30 minutes to 59 minutes, 1 hour to 1 hour 29 minutes, 1 hour 30 minutes to 1 hour 59 minutes, and 2 hours or more.*

The time distance by walking is calculated following laws and regulations concerning advertisement. Ordinance for 'Enforcement of the fair competition codes concerning indication of real estate', Chapter 5, Article 10, in accordance with the 'Fair competition codes', Article 15, assigns a walking rate formula. This formula states that one-minute walking on a road is equal to a distance of 80 meters. There are 1955 properties that are '0' minutes walking distance to the nearest station.

Table 3.9: Summary statistics distance (0–29min)

| City | mean | 25% | 50% | 75% | sd | *n* |
|---|---|---|---|---|---|---|
| Tokyo | 8.2 | 4 | 7 | 11 | 5.04 | 184077 |
| Osaka | 6.8 | 4 | 6 | 9 | 4.21 | 49820 |
| Nagoya | 10.8 | 6 | 9 | 15 | 6.73 | 38779 |
| Fukuoka | 11.0 | 6 | 9 | 15 | 6.73 | 25982 |
| Sapporo | 11.2 | 6 | 10 | 15 | 7.00 | 32685 |

Table 3.9 summarizes the distance in minutes for each city. We provide the mean, three quantiles, and the standard deviation. Osaka has the densest railway structure, followed by Tokyo. Fukuoka, Nagoya, and Sapporo have a somewhat less dense railway/metro system.

**Area and total floor area (in square meters)**

From the official description:

*For each of the residential land (land only), residential land (land and building), agricultural land, and forest land, the surveyed area ($m^2$) obtained from a survey of persons involved in transactions or the registered area ($m^2$) specified in a register if the surveyed area is unknown is provided. For pre-owned condominiums, etc., the floor area ($m^2$) of the exclusively owned area registered in a register (the area measured inside walls or other partitions) is provided. For all land types, data for small properties with an area less than 10 $m^2$ are not published. For properties with an area of less than 200 $m^2$, the area data are displayed in 5 $m^2$ intervals, while for properties with an area of 200 $m^2$ or greater, the data are displayed after rounding the figures to the first two digits from the left. For transactions for land with an area of 2,000 $m^2$ or greater, the data are displayed as '2,000 $m^2$ or greater'.*

For 'land only' types, the variable 'area' refers to the area of the land; for condos it refers to the floor area.

Another variable is the total floor area of the building ($m^2$). We have:

*For buildings on residential land (land and building), the total floor area ($m^2$) is provided. For cases where the floor area is less than 200 $m^2$, the data are displayed in 5 $m^2$ intervals, and for cases where the floor area is 200 $m^2$ or greater, figures are rounded to two decimal*

*places. For large transactions where the building floor area is 2,000 $m^2$ or greater, the data are displayed as '2,000 $m^2$ or greater' whereas for small transactions where the floor area is less than 10 $m^2$, the data are displayed as 'less than 10 $m^2$'.*

We set the total floor area of properties where the building floor area is 2,000 $m^2$ or greater to 2000 and those where the floor area is less than 10 $m^2$ to 10.

Table 3.10: Summary statistics area ($m^2$)

| City | mean | 25% | 50% | 75% | sd | *n* |
|---|---|---|---|---|---|---|
| *Land & Building* | | | | | | |
| Tokyo | 128.5 | 65.0 | 90.0 | 125.0 | 148.0 | 57568 |
| Osaka | 134.8 | 55.0 | 75.0 | 130.0 | 180.7 | 21064 |
| Nagoya | 186.1 | 110.0 | 135.0 | 180.0 | 178.0 | 14640 |
| Fukuoka | 259.4 | 140.0 | 180.0 | 270.0 | 240.6 | 7847 |
| Sapporo | 252.5 | 160.0 | 200.0 | 260.0 | 206.3 | 11763 |
| *Land Only* | | | | | | |
| Tokyo | 165.7 | 70.0 | 105.0 | 175.0 | 193.0 | 33991 |
| Osaka | 218.0 | 75.0 | 125.0 | 250.0 | 257.9 | 6901 |
| Nagoya | 240.6 | 120.0 | 170.0 | 270.0 | 224.1 | 13110 |
| Fukuoka | 325.4 | 150.0 | 220.0 | 370.0 | 302.7 | 5660 |
| Sapporo | 287.8 | 165.0 | 220.0 | 300.0 | 257.2 | 9461 |
| *Condo* | | | | | | |
| Tokyo | 46.7 | 20.0 | 45.0 | 65.0 | 27.9 | 92518 |
| Osaka | 53.8 | 30.0 | 60.0 | 70.0 | 26.9 | 21855 |
| Nagoya | 65.6 | 60.0 | 70.0 | 80.0 | 23.7 | 11029 |
| Fukuoka | 53.4 | 25.0 | 60.0 | 75.0 | 27.7 | 12475 |
| Sapporo | 69.1 | 60.0 | 70.0 | 85.0 | 22.8 | 11461 |

Table 3.11: Summary statistics total floor area ($m^2$)

| City | mean | 25% | 50% | 75% | sd | *n* |
|---|---|---|---|---|---|---|
| *Land & Building* | | | | | | |
| Tokyo | 195.0 | 85.0 | 95.0 | 145.0 | 302.2 | 57568 |
| Osaka | 255.6 | 90.0 | 105.0 | 200.0 | 384.6 | 21064 |
| Nagoya | 214.1 | 100.0 | 110.0 | 155.0 | 312.8 | 14640 |
| Fukuoka | 282.2 | 100.0 | 125.0 | 230.0 | 399.9 | 7847 |
| Sapporo | 286.3 | 110.0 | 140.0 | 300.0 | 351.5 | 11763 |

**Year of construction and age**

For properties built before 1945, the construction year data are displayed as 'before the war'. We discard these data and further categorize the year of construction into '1946–1981', '1982–2000', and '2001–now'. This categorization is due to time points of major changes in the Building Standards Act; see

`www.uncrd.or.jp/hyogo/hesi/pdf/expmeeting/otani.pdf`,

from the United Nations website. In 1981 the establishment of 'The shin-taishin, or New Earthquake Resistant Building Standard Amendment' came into effect following the disaster caused by the 1978 earthquake off the shore of Miyagi. The new standard stipulates that buildings must able to resist an earthquake of at least JMA seismic scale upper 6 instead of scale 5. In 2000 the act was revised with a more stringent standard for wooden buildings, and also new houses must provide a 10-Year warranty against defects. These changes might lead to structural breaks in the quality of houses built around these time points.

Table 3.12: Summary statistics building age

| City | mean | min | 25% | 50% | 75% | max | sd | *n* |
|---|---|---|---|---|---|---|---|---|
| Tokyo | 14.4 | −2 | 1 | 11 | 24 | 67 | 13.5 | 184077 |
| Osaka | 18.1 | −2 | 4 | 17 | 30 | 69 | 14.8 | 49820 |
| Nagoya | 14.9 | −1 | 0 | 13 | 25 | 69 | 13.9 | 38779 |
| Fukuoka | 17.2 | −2 | 7 | 18 | 25 | 66 | 12.3 | 25982 |
| Sapporo | 18.1 | −1 | 8 | 18 | 27 | 65 | 12.4 | 32685 |

We also include the numerical 'age' of the property, i.e. transaction year minus the year of construction. This is displayed in Table 3.12. This can be a negative number, namely if the property was sold before construction.

**Building structure**

Building structure can be 'Steel frame reinforced concrete', 'Reinforced concrete', 'Steel frame', 'Light steel structure', 'Concrete block', 'Wooden', or combinations of these structures. This leads to a large number of building structures. We summarize these in five categories: 'contains steel frame reinforced concrete' (SRC), 'contains reinforced concrete but not steel frame reinforced concrete' (RC), 'contains steel but not reinforced concrete' (S), 'contains wood but not steel or reinforced concrete' (W), and 'NA'. This categorization is to some extent arbitrary but represents a general way to distinguish between building structures regarding fire- and earthquake-resistant features. The SRC and RC types are deemed most resistant to earthquake and fire; The S types and W types less so.

Table 3.13: Number of properties per building structure

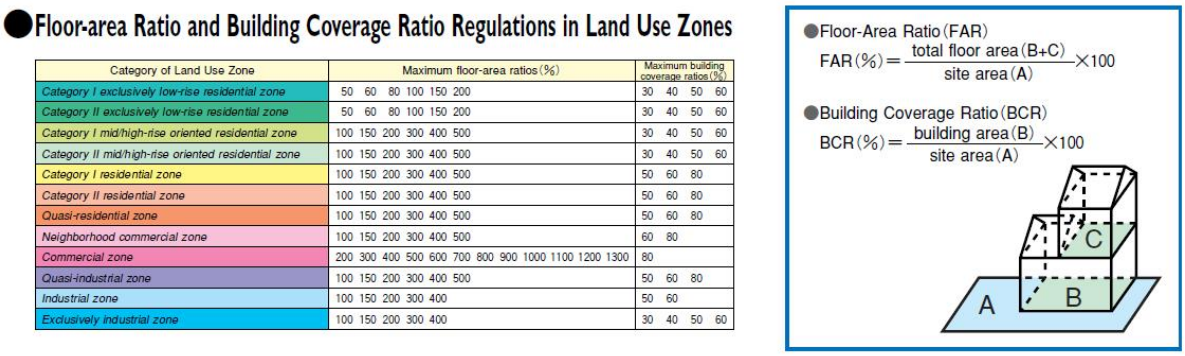
| Building structure | *Land & Building* | *Condo* |
|---|---|---|
| SRC | 2288 | 54904 |
| RC | 12581 | 92804 |
| S | 17619 | 1135 |
| W | 75017 | 14 |
| NA | 5377 | 481 |
| Total | 112882 | 149338 |

Table 3.13 contains a summary of the available information.

**Building coverage ratio and floor area ratio**

For all three types the *designated* maximum building coverage ratio (%) and maximum floor-area ratio (%) are provided. These ratios are legally allowed maxima, different for each piece of land. Usually buildings with larger designated ratios are more expensive.

The building coverage ratio is the percentage of the site area to the building area. The floor area ratio is the percentage of the total floor area to the site area.



●Floor-area Ratio and Building Coverage Ratio Regulations in Land Use Zones

| Category of Land Use Zone | Maximum floor-area ratios (%) | Maximum building coverage ratios (%) |
|---|---|---|
| Category I exclusively low-rise residential zone | 50 60 80 100 150 200 | 30 40 50 60 |
| Category II exclusively low-rise residential zone | 50 60 80 100 150 200 | 30 40 50 60 |
| Category I mid/high-rise oriented residential zone | 100 150 200 300 400 500 | 30 40 50 60 |
| Category II mid/high-rise oriented residential zone | 100 150 200 300 400 500 | 30 40 50 60 |
| Category I residential zone | 100 150 200 300 400 500 | 50 60 80 |
| Category II residential zone | 100 150 200 300 400 500 | 50 60 80 |
| Quasi-residential zone | 100 150 200 300 400 500 | 50 60 80 |
| Neighborhood commercial zone | 100 150 200 300 400 500 | 60 80 |
| Commercial zone | 200 300 400 500 600 700 800 900 1000 1100 1200 1300 | 80 |
| Quasi-industrial zone | 100 150 200 300 400 500 | 50 60 80 |
| Industrial zone | 100 150 200 300 400 | 50 60 |
| Exclusively industrial zone | 100 150 200 300 400 | 30 40 50 60 |



Figure 3.1: Urban land use planning system in Japan (Ministry of Land, Infrastructure and Transport)

For different city planning zones, there are different limits on these ratios, as shown Figure 3.1.

Table 3.14: Summary statistics Building Coverage Ratio

| City | mean | min | 25% | 50% | 75% | max | sd | *n* |
|---|---|---|---|---|---|---|---|---|
| Tokyo | 65 | 30 | 60 | 60 | 80 | 80 | 11 | 184077 |
| Osaka | 75 | 40 | 80 | 80 | 80 | 80 | 9 | 49820 |
| Nagoya | 63 | 30 | 60 | 60 | 80 | 80 | 12 | 38779 |
| Fukuoka | 64 | 30 | 60 | 60 | 80 | 80 | 12 | 25982 |
| Sapporo | 62 | 30 | 60 | 60 | 80 | 80 | 13 | 32685 |

Table 3.15: Summary statistics Floor Area Ratio

| City | mean | min | 25% | 50% | 75% | max | sd | *n* |
|---|---|---|---|---|---|---|---|---|
| Tokyo | 302 | 60 | 200 | 300 | 400 | 1300 | 157 | 184077 |
| Osaka | 341 | 80 | 200 | 300 | 400 | 1300 | 172 | 49820 |
| Nagoya | 240 | 50 | 200 | 200 | 200 | 1100 | 130 | 38779 |
| Fukuoka | 238 | 50 | 150 | 200 | 400 | 1000 | 134 | 25982 |
| Sapporo | 214 | 50 | 200 | 200 | 200 | 900 | 104 | 32685 |

The relevant data in our data set are summarized in Table 3.14 and Table 3.15.

#### City planning

For all three types of records, the use of districts designated by the City Planning Act is provided. A detailed explanation can be found at

`www.mlit.go.jp/crd/city/plan/tochiriyou/pdf/reaf_e.pdf`.

The planning can be: Category 1 Exclusive Low Rise Residential District, Category 1 Exclusive Mid-high Rise Residential District, Category 1 Residential District, Category 2 Exclusive Low Rise Residential District, Category 2 Exclusive Mid-high Rise Residential District, Category 2 Residential District, Commercial District, Exclusive Industrial District, Industrial District, Near-commercial District, Outside Urban Planning Area, Semi-industrial District, Semi-residential District, Semi-urban Planning Area, or Urbanization Control Area.

We may further categorize them into subclasses: Residential, Commercial, Industrial, and Other ('Urbanization control area', 'Non-divided city planning area', 'Quasi-city planning area', and 'Outside city planning area'). The number of properties for each land use category is summarized in Table 3.16.

Table 3.16: Number of properties for each land use category

| Land Use | *Land & Building* | *Land Only* | *Condo* |
|---|---|---|---|
| Residential area | 76303 | 48011 | 50178 |
| Commercial area | 20979 | 12324 | 71237 |
| Industrial area | 15339 | 8467 | 26260 |
| Urbanization control area | 140 | 231 | 11 |
| NA | 121 | 90 | 1652 |
| Total | 112882 | 69123 | 149338 |

### 3.4.6 Information ignored in our analysis

#### Region

For all condos, the region information is unknown. For 'land only' and 'land & building', the region can be one of the following: residential area, commercial area, potential residential area, or industrial area.

#### Layout

The layout information is only known for condos. This can take the following values: NA, 3LDK, 4LDK, 5LDK, 2LDK, 1K, 2DK, 1R, 3DK, 1LDK, 1DK, Open Floor, 2LDK+S, 2K, 3LDK+S, 2DK+S, 3LD, 1LDK+S, 1DK+S, 4DK, 4LDK+S, 2L, 3LK, 5LDK+S, Studio Apartment, 2LK, Duplex, 3K, 3DK+S, 4K, 2K+S, 1R+S, 3K+S, 1LK, 5DK, 1L, 6DK, 1K+S, 7LDK, 6LDK, 4DK+S, 2LK+S, 3LD+S, 2LD+S, 8LDK, 4L, 4LDK+K, 4L+K, 3D, 6LDK+S.

We do not use this information since it is difficult to categorize and already we have floor area in our selected data.

**Unit land price**

According to the MLIT website, the unit price ($10,000$ yen) per $m^2$ is provided only for 'land only'. This unit price is obtained by dividing the total transaction value of each plot of land by the land area ($m^2$). The variable is not available for 'land & building' and condos, so we do not use this information.

**Land shape**

Only for 'land only' and 'land & building' the general shape of land is provided. The land shape can take the following forms: NA, rectangular shaped, semi-square shaped, semi-rectangular shaped, irregular Shaped,semi-trapezoidal shaped, trapezoidal shaped, semi-shaped, square-shaped, flag-shaped, etc.

**Frontage of land / frontage road**

For 'land only' and 'land & building', the frontage/width of land (in $m$) is provided, that is, the length of land in contact with a frontage road, as well as the width (in $m$), type, and direction of the road in contact with the land.

**Building use**

For the buildings on residential land (land and building) and exclusively owned areas of pre-owned condominiums, the current usage is provided. This can be: house, shop, other, office, housing complex, parking lot, factory, warehouse, workshop, or any combination of these. We do not include this variable since there are too many interactions and it would be difficult to categorize.

**Purpose of use**

Purpose of use is provided only for records where the transaction period is after 1st quarter 2013. For these records, purpose of use can be: NA, house, shop, other, office, warehouse, factory, among which 'house' is the majority (45,373 among 53,596 records where purpose of use is known). We do not use this information because of its limited availability.

**Additional remarks that might impact housing price**

According to the MLIT website, remarks are provided when there is additional information that may have impact on transaction prices. These are provided only when relevant additional information is obtained via a questionnaire survey. Out of 158,474 records where remarks are provided, 137,490 are condos. The remarks can be: NA, dealings of non-redecorating real estate, dealings of redecorated real estate, dealings with auction or arbiter participation, dealings including private road, dealings between related objects, dealings including special circumstances, dealings of adjacent land, dealings of real estate that includes damage, dealings of real estate with mortgage issues, dealings including a valueless house, or a combination of these items.

## 3.5 Macro-economic variables

Housing prices are affected by general economic conditions. In order to incorporate possible effects of these economic conditions, we include the following macro-economic indicators as explanatory variables: GDP, CPI, interest rate and stock price.

Table 3.17: Macro-economic variables

| Name | Description | Frequency | Source |
|---|---|---|---|
| GDP | Nominal (not seas. adj.) | Quarterly | Cabinet office |
| CPI | All items, 2015-base | Monthly | Statistics Bureau |
| Interest rate | Basic discount rate | Quarterly | Bank of Japan |
| TOPIX | Tokyo Stock Price Index | Monthly | Cabinet Office |

GDP figures are provided by the cabinet office website

`www.esri.cao.go.jp/index-e.html`.

We use the nominal GDP series, not seasonally adjusted. The reason that we not adjust for quarter is that we include many other controls in the analysis that vary over quarters. If we would take seasonally adjusted GDP series, then the determinants of the seasonal adjustment and our own control variables would become confounded.

The monthly CPI data can be downloaded from

`www.stat.go.jp/english/data/cpi/index.htm`.

We use the version released on August 26, 2016, which is 2015-based. Since our housing price data are per quarter, we integrate the monthly data into quarterly data using simple averages.

Interest rate is one of the most important factors that have an impact on house prices. We use quarterly time-series data of the 'basic discount rate',

`www.stat-search.boj.or.jp/index_en.html#`,

provided by the Bank of Japan.

Stock prices reflect business conditions, which might affect the housing market. We download monthly data of the Tokyo Exchange Tokyo Price Index (TOPIX) from the ESRI website

`www.esri.cao.go.jp/en/stat/di/di-e.html`.

## 3.6 Historical earthquakes

### 3.6.1 Data source

The Japan Meteorological Agency (JMA) employs a seismic intensity scale to measure the intensity of earthquakes. It is measured in units of 'shindo' (seismic intensity). The JMA scale differs from the more common Richter scale (and the Moment Magnitude Scale), which measure magnitude, that is, the energy

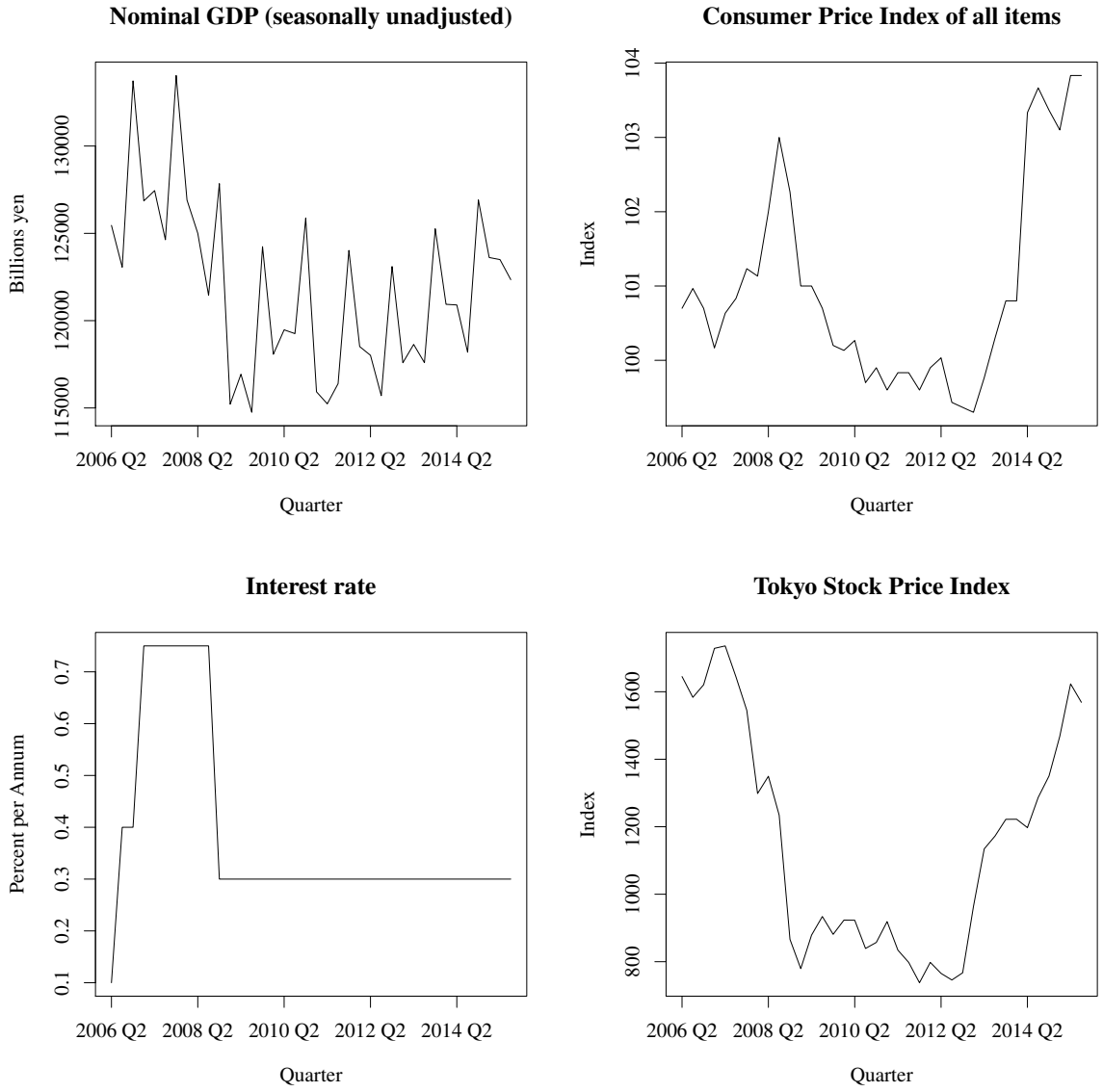



Figure 3.2: Japan macro-economic data series, 2006 Q2 – 2015 Q3

released by the earthquake. In contrast, the JMA scale describes the degree of shaking. The intensity of an earthquake is not completely determined by its magnitude, but varies with the event's depth and the distance from the event. For example, an earthquake may be described as shindo 4 in Tokyo, shindo 3 in Yokohama, and shindo 2 in Shizuoka. The JMA operates a network of 180 seismographs and 627 seismic intensity meters and provides real-time earthquake reports to the media and on the Internet.

The JMA data can be downloaded from

`www.data.jma.go.jp/svd/eqev/data/bulletin/shindo_e.html`.

Each year's earthquake data are stored in separate `.dat` files, which provide record entries of two possible types:

(1) Hypocenter record; and

(2) Seismic intensity and acceleration data record.

A type (1) record contains the information we need for modeling the earthquake process: date, time, exact location, depth, magnitude, intensity, etc. Following each type (1) record there are one or more type (2) records, which contain descriptions from seismic detection stations about this earthquake.

The data are available for the years 1923–2015. A subset of the data can be selected with dropdown menus from

`www.data.jma.go.jp/svd/eqdb/data/shindo/index.php`.

These interactive tables display the hypocenter records that are natural earthquakes (specified as '1' in column 'Subsidiary information') with known maximum intensity and known hypocenter (specified in column 'Identifiers', which cannot be 'N: Hypocenter unknown').

### 3.6.2 Description

The description of all the earthquake parameters can be found on

`www.data.jma.go.jp/svd/eqev/data/bulletin/data/shindo/format_e.txt`.

In each type (1) record, the following information is available:

- Record type identifier (A: hypocenter record, B: hypocenter record (for two or more spatio-temporally close earthquakes whose seismic intensity data cannot be separated), D: hypocenter record (for two or more temporally close earthquakes whose seismic intensity data cannot be separated)
- Year, month, day, hour, minute, second of origin time (Japan Standard Time = UTC + 9 hours)
- Standard error (of origin time, in seconds)
- Latitude of hypocenter (degrees and minutes)

- Standard error for latitude (minutes)
- Longitude of hypocenter (degrees and minutes)
- Standard error for longitude (minutes)
- Depth in kilometers
- Standard error for depth (kilometers)
- Magnitude 1
- Magnitude type (1),

  JMA magnitudes: J: MJ - Local Meteorological Office magnitude; D: MD - Displacement magnitude; d: Md - As per MD, but for two stations; V: MV - Velocity magnitude; v: Mv - As per MV, but for two or three stations.

  Moment magnitudes: W: MW - Moment magnitude based on the JMA centroid moment tensor solution.

  Other organizations' magnitudes: B: mb - USGS body wave magnitude; S: MS - USGS surface wave magnitude.
- Magnitude 2
- Magnitude type 2 (see magnitude type 1)
- Travel time table type
- Hypocenter location precision (1: Depth-free method; 2: Depth-slice method; 3: Fixed depth; 4: Based on depth phase; 5: Based on S-P time; 7: Poor solution; 8: Undetermined or not accepted)
- Subsidiary information on event (1: Natural earthquake; 2: Insufficient number of JMA stations; 3: Artificial event; 4: Noise; 5: Low-frequency earthquake)
- Maximum intensity (1: One; 2: Two; 3: Three; 4: Four; 5: Five (until September 1996); 6: Six (until September 1996); 7: Seven; A: Five lower; B: Five upper; C: Six lower; D: Six upper; R: Remarkable earthquake (shock felt over 300 km away) (until 1977); M: Moderate earthquake (shock felt over 200 km away but not over 300 km away) (until 1977); S: Small earthquake (shock felt over 100 km away but not over 200 km away) (until 1977); L: Local earthquake (shock felt less than 100 km away) (until 1977); F: Felt earthquake (until 1984); X: Shock felt by some people but not by JMA observers (until September 1996))
- Damage class (after Utsu) (1: Slight damage (cracks on walls and ground); 2: Light damaged (broken houses, roads, etc.); 3: 2–19 fatalities or 2–999 houses destroyed; 4: 20–199 fatalities or 1,000–9,999 houses destroyed; 5: 200–1,999 fatalities or 10,000–99,999 houses destroyed; 6: 2,000–19,999 fatalities or 100,000–999,999 houses destroyed; 7: 20,000+ fatalities or 1,000,000+ houses destroyed; X: Injury or damage of unclear scale (until 1988); Y: Injury and damage included in the grade for the preceding or following event (until 1988))

- 1929–1988 Tsunami class (after Utsu)
- Number of epicenter location district
- Number and name of epicenter location region
- Number of shocks felt
- Identifiers (K: JMA hypocenter identified with high precision; S: JMA hypocenter identified with low precision; N: Hypocenter unknown (first observation point used); U: USGS hypocenter; I: ISC hypocenter; R: Preliminary hypocenter (included only in district observatory databases); H,D,M: Exact observation time unknown)

### 3.6.3 JMA intensity scale

The tables detailing the JMA intensity scales can be found in

`www.jma.go.jp/jma/en/Activities/inttable.html`.

An excerpt of the description of each intensity level is shown below.

**Human perception and reaction, indoor situation, outdoor situation**

| Seismic intensity | Human perception and reaction | Indoor situation | Outdoor situation |
|---|---|---|---|
| 0 | Imperceptible to people, but recorded by seismometers. | - | - |
| 1 | Felt slightly by some people keeping quiet in buildings. | - | - |
| 2 | Felt by many people keeping quiet in buildings. Some people may be awoken. | Hanging objects such as lamps swing slightly. | - |
| 3 | Felt by most people in buildings. Felt by some people walking. Many people are awoken. | Dishes in cupboards may rattle. | Electric wires swing slightly. |
| 4 | Most people are startled. Felt by most people walking. Most people are awoken. | Hanging objects such as lamps swing significantly, and dishes in cupboards rattle. Unstable ornaments may fall. | Electric wires swing significantly. Those driving vehicles may notice the tremor. |
| 5 Lower | Many people are frightened and feel the need to hold onto something stable. | Hanging objects such as lamps swing violently. Dishes in cupboards and items on bookshelves may fall. Many unstable ornaments fall. Unsecured furniture may move, and unstable furniture may topple over. | In some cases, windows may break and fall. People notice electricity poles moving. Roads may sustain damage. |
| 5 Upper | Many people find it hard to move; walking is difficult without holding onto something stable. | Dishes in cupboards and items on bookshelves are more likely to fall. TVs may fall from their stands, and unsecured furniture may topple over. | Windows may break and fall, unreinforced concrete-block walls may collapse, poorly installed vending machines may topple over, automobiles may stop due to the difficulty of continued movement. |
| 6 Lower | It is difficult to remain standing. | Many unsecured furniture moves and may topple over. Doors may become wedged shut. | Wall tiles and windows may sustain damage and fall. |
| 6 Upper | It is impossible to remain standing or move without crawling. People may be thrown through the air. | Most unsecured furniture moves, and is more likely to topple over. | Wall tiles and windows are more likely to break and fall. Most unreinforced concrete-block walls collapse. |
| 7 | | Most unsecured furniture moves and topples over, or may even be thrown through the air. | Wall tiles and windows are even more likely to break and fall. Reinforced concrete-block walls may collapse. |

**Wooden houses**

| Seismic intensity | Wooden houses | |
|---|---|---|
| | High earthquake resistance | Low earthquake resistance |
| 5 Lower | - | Slight cracks may form in walls. |
| 5 Upper | - | Cracks may form in walls. |
| 6 Lower | Slight cracks may form in walls. | Cracks are more likely to form in walls.<br>Large cracks may form in walls.<br>Tiles may fall, and buildings may lean or collapse. |
| 6 Upper | Cracks may form in walls. | Large cracks are more likely to form in walls.<br>Buildings are more likely to lean or collapse. |
| 7 | Cracks are more likely to form in walls.<br>Buildings may lean in some cases. | Buildings are even more likely to lean or collapse. |

**Reinforced-concrete buildings**

| Seismic intensity | Reinforced-concrete buildings | |
|---|---|---|
| | High earthquake resistance | Low earthquake resistance |
| 5 Upper | - | Cracks may form in walls, crossbeams and pillars. |
| 6 Lower | Cracks may form in walls, crossbeams and pillars. | Cracks are more likely to form in walls, crossbeams and pillars. |
| 6 Upper | Cracks are more likely to form in walls, crossbeams and pillars. | Slippage and X-shaped cracks may be seen in walls, crossbeams and pillars.<br>Pillars at ground level or on intermediate floors may disintegrate, and buildings may collapse. |
| 7 | Cracks are even more likely to form in walls, crossbeams and pillars.<br>Ground level or intermediate floors may sustain significant damage. Buildings may lean in some cases. | Slippage and X-shaped cracks are more likely to be seen in walls, crossbeams and pillars.<br>Pillars at ground level or on intermediate floors are more likely to disintegrate, and buildings are more likely to collapse. |

Figure 3.3: Explanation of the JMA Seismic Intensity Scale, source: JMA website

### 3.6.4 Sample selection

We only extract the type (1) records. In order to use the records for modeling the earthquake process, we may choose samples based on time period, hypocenter location, magnitude or intensity threshold, and depth.

We select only the records that are 'natural earthquakes'. In doing so we discard the records labeled as 'Insufficient number of JMA stations', 'Noise', or 'Low-frequency earthquake' since these records may not be reliable.

We discard the records with unknown hypocenters since the location information is inaccurate; furthermore the magnitudes for these records are also unknown.

We discard the records with unknown 'maximum intensity' since these records are earthquakes that are spatio-temporally close to another earthquake so that they cannot be separated.

Note that even with the same sample selection criteria as described in the literature, the resulting sample catalog can be quite different. This is because the JMA catalog has been updated (modified) many times as technology and knowledge have improved; see

`www.data.jma.go.jp/svd/eqev/data/bulletin/data/hypo/relocate.html`

for further discussion (Japanese only).

### 3.6.5 Summary statistics

There are 194,882 records in the entire 1923–2015 period, among which 105,685 records are natural earthquake with known hypocenter and known maximum intensity. Table 3.18 contains a summary of these records.

Table 3.18: Summary of JMA earthquake records for 1923–2015

| Magnitude | Intensity | $n$ |
|---|---|---|
| Magnitude<3 | <=4 | 26015 |
| 3<=Magnitude<4 | <=4 | 33464 |
| | 5/5-lower/5 upper | 4 |
| 4<=Magnitude<5 | <=4 | 22845 |
| | 5/5-lower/5 upper | 57 |
| 5<=Magnitude<6 | <=4 | 7085 |
| | 5/5-lower/5 upper | 157 |
| | 6/6-lower/6 upper | 7 |
| 6<=Magnitude<7 | <=4 | 1283 |
| | 5/5-lower/5 upper | 131 |
| | 6/6-lower/6 upper | 24 |
| | 7 | 1 |
| Magnitude>=7 | <=4 | 109 |
| | 5/5-lower/5 upper | 44 |
| | 6/6-lower/6 upper | 22 |
| | 7 | 2 |
| Unknown | <=4 | 14432 |
| | 5/5-lower/5 upper | 3 |

In some records two magnitudes (and two magnitude types) are reported. The second magnitude is supplementary information and the difference between the two magnitudes recorded for the same earthquake is usually small. In Table 3.18 and in other places where magnitudes are needed, we use the first magnitude (Magnitude 1) without further clarification.

## 3.7 ETAS estimation and simulation

The Epidemic Type Aftershock Sequence (ETAS) model was introduced by Ogata (1988) and has been widely used to capture the quiescence and activation of seismic activities. The basic idea is that each earthquake may trigger a sequence of aftershocks like 'epidemics' and that the severity of influence diminishes over time (and distance). Despite its many space-time extensions, we choose the temporal version of this model as described in the following for simplicity of estimation and simulation.

Given the observations of earthquake occurrences at time $t_1, t_2, ..., t_n$ on an interval $[0, T]$ $(T \geq t_n)$, the associated counting process is defined as $N_t = \sum_{i=1}^{n} \mathbb{1}_{t_i \leq t}$.

The corresponding left-continuous $\mathscr{F}_t$-conditional jump intensity process $\lambda_t$ describes the mean jump rate per unit of time,

$$\lambda_t = \lambda(t|\mathscr{F}_t) = \lim_{h \downarrow 0} \frac{1}{h} \Pr\left[N_{t+h} - N_t > 0 | \mathscr{F}_t\right].$$

In a temporal ETAS model, the conditional intensity function may be written as

$$\lambda_t = \lambda_\infty + \sum_{t_i<t} c(m_i, m_c) g(t - t_i),$$

where $\lambda_\infty$ (shocks per unit time) is the background seismicity with $\lambda_\infty > 0$. The aftershock decay (time response function) takes the form of the Modified Omori function, $g(t - t_i) = \frac{K}{(t-t_i+C)^p}$. The weight assigned to the aftershock decay is an exponential function of the difference between the magnitude of the earthquake and the cut-off magnitude $m_c$: $c(m_i, m_c) = \exp(\beta(m_i - m_c))$. The intensity consists of the background intensity and a weighted sum of all the aftershock decays, where the sum is taken over all earthquakes before time $t$.

For the estimation of the ETAS model, we take the earthquake catalog of areas around the five Japanese cities in the period 1970-1-1 to 2015-12-31. The space windows and cut-off magnitudes for each city is shown below.

Table 3.19: Space window of the earthquake catalog

| city | latMin | latMax | lngMin | lngMax | codeMin | codeMax |
|---|---|---|---|---|---|---|
| Tokyo | 34 | 37 | 138 | 141 | 13101 | 13123 |
| Osaka | 33.5 | 36.5 | 134 | 137 | 27102 | 27128 |
| Nagoya | 33.5 | 36.5 | 135.5 | 138.5 | 23101 | 23116 |
| Fukuoka | 32 | 35 | 129 | 132 | 40131 | 40137 |
| Sapporo | 41.5 | 45.5 | 138.5 | 143.5 | 1101 | 1110 |

The space windows and magnitude thresholds are chosen such that 1) the properties in our data set are located around the center of space windows in each cities; 2) the number of observations within each space window is moderate, in the sense that they are comparable across cities and that they yield meaningful results for the ETAS estimation routine; 3) the estimation of ETAS models for each city is appropriate, in the sense that they converge and pass the test of residuals. We use the test described by Berman (1983), where we consider whether the transformed inter-arrival times are iid exponential random variables with unit mean.

We show below the p-values for the Kolmogorov-Smirnov (K-S) test, number of observations as well as model parameter estimates for each city. The estimation threshold is chosen to be magnitude 4.5 for Fukuoka, Nagoya, Osaka, Sapporo and 5 for Tokyo. This means that in each city, only earthquake records above the corresponding thresholds are used for estimation.

Table 3.20: Estimated with threshold 4.5 for Osaka, Nagoya, Fukuoka, Sapporo and 5 for Tokyo

| city | K-S | $N$ | $\lambda_\infty$ | $K$ | $C$ | $p$ | $\beta$ |
|---|---|---|---|---|---|---|---|
| Tokyo | 0.764 | 370 | 0.0076 | 0.0351 | 0.0155 | 1.0072 | 0.9253 |
| Osaka | 0.738 | 154 | 0.0073 | 0.0014 | 0.0064 | 1.1913 | 2.4854 |
| Nagoya | 0.940 | 177 | 0.0071 | 0.0066 | 0.0009 | 0.9616 | 1.7250 |
| Fukuoka | 0.982 | 102 | 0.0037 | 0.0098 | 0.0035 | 1.0330 | 1.4390 |
| Sapporo | 0.874 | 486 | 0.0193 | 0.0039 | 0.1044 | 1.2091 | 2.4424 |

The K-S test is one way of analyzing the residuals in an ETAS model; see Ogata (1988). Suppose the integral of the conditional intensity function is

$$\Lambda(t) = \int_0^t \lambda(s)ds,$$

and the random time change $\tau = \Lambda(t)$. The $\Lambda(\cdot)$ function transforms the event time series $\{t_i\}$ to $\{\tau_i\}$ which is a stationary Poisson process of intensity 1. The differences $Y_k = \tau_k - \tau_{k-1}$ should be iid exponential random variables with mean 1, and $U_k = 1 - \exp(-Y_k)$ should be iid standard uniform random variables. We test here whether $U_k$ comes from the standard uniform distribution, where a p-value of lower than 0.05 rejects the null hypothesis that $U_k$ comes from the standard uniform distribution at a 0.05 significance level.

The estimated intensity and actual earthquake events are used to simulate 90-day probabilities of an earthquake exceeding the magnitude threshold of 5.5 for each city. The simulation method follows Ogata (1981).

Figure 3.4 shows the short-run risk series. As shown in the figure, the short-run probabilities spike up immediately after a large earthquake and dies out gradually until another earthquake occurs.

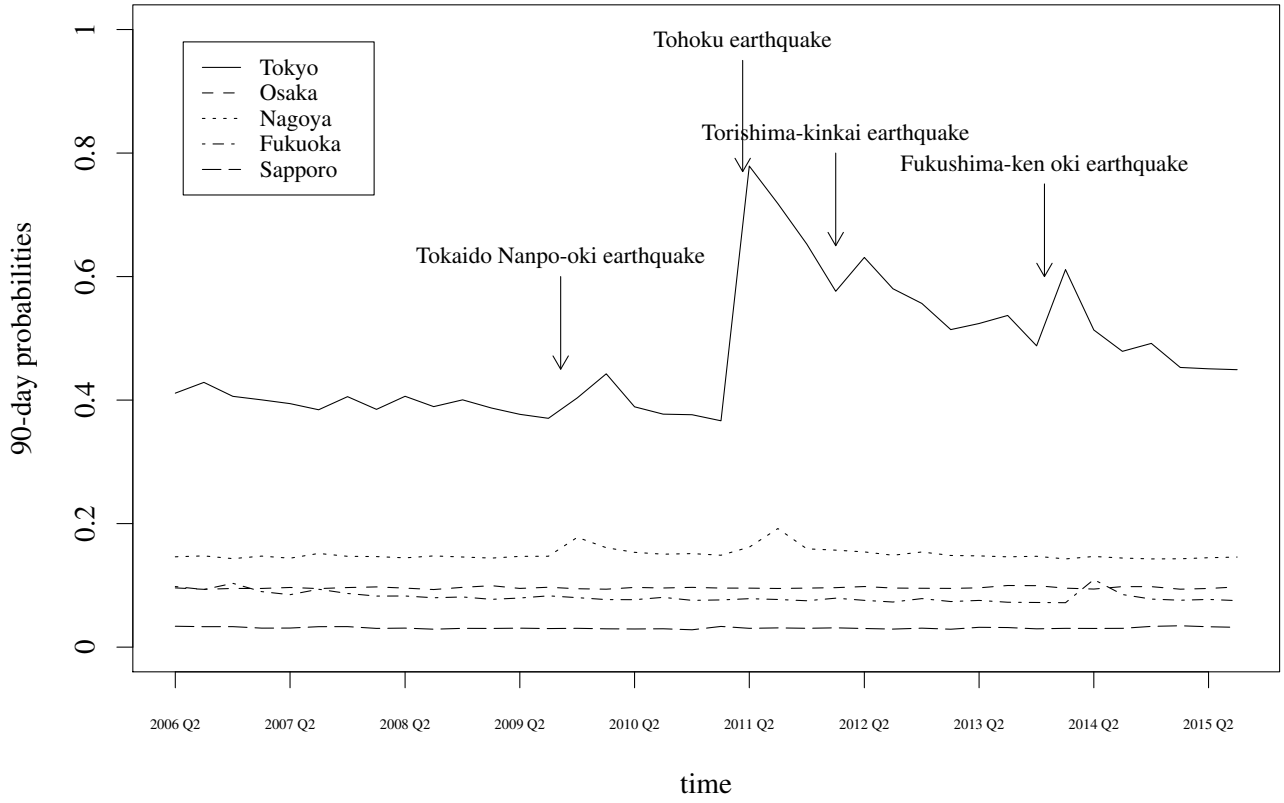



Figure 3.4: Simulated short run risk

## 3.8 Mesh codes

The Statistics Bureau of Japan explains the definition of mesh code in 'Standard Grid Square and Grid Square Code Used for the Statistics'; see the following webpage

`www.stat.go.jp/english/data/mesh/02.htm`.

There are five levels of precision:

- First mesh ('Primary Area Partition') is obtained by dividing the whole area of Japan into blocks measuring 1 degree of longitude and 2/3 degree of latitude.
- Second mesh ('Secondary Area Partition') is obtained by dividing first mesh areas into $8 \times 8$ squares.
- Third mesh ('Third Area Partition / Basic Grid Square') is obtained by dividing second mesh areas into $10 \times 10$ squares.

- Half/quarter mesh is obtained by dividing the third mesh grid into $2 \times 2$ or $4 \times 4$ equal parts to get half or quarter grid squares, respectively.

Table 3.21: Levels of mesh codes

| Level | No. digits | Format | Scale |
|---|---|---|---|
| First mesh | 4 | XXXX | 80km |
| Second mesh | 6 | XXXX–XX | 10km |
| Third mesh | 8 | XXXX–XXXX | 1km |
| Half mesh | 9 | XXXX–XXXX–X | 500m |
| Quarter mesh | 10 | XXXX–XXXX–X–X | 250m |

The coding and scale of the five levels is given in Table 3.21.

In Figure 3.5 we provide a map of all first meshes.

In Figure 3.6 we show how mesh code numbers are calculated from the coordinates, based on

`www.stat.go.jp/english/data/mesh/05-1.htm`.

## 3.9 Predicted earthquake risks

### 3.9.1 The J-SHIS data set

While we considered data on actual earthquakes in the previous section, the current section deals with data on earthquake risk. Earthquake risk has several dimensions. What matters is not only how likely it is that there is an earthquake, but also how much damage the resulting fire will cause. Houses (since 1981) are essentially earthquake-proof, but they may not be fire-proof. When in a Japanese text it says 'earthquake risk' the meaning can be ambiguous. It may mean only earthquake, but more likely it means 'earthquake and related risks combined'.

The Japan Seismic Hazard Information Station (J-SHIS) was established to help prevent and prepare for earthquake disasters The seismic hazard maps serve as a sharing platform of seismic hazard information by regarding the maps as a group of information incorporating the underlying data used in the evaluation process, such as the seismic activity models, seismic source models, and subsurface structure models, rather than mere maps as final products. Operations started in May 2005.

Four years later, in 2009, it was decided to incorporate the latest technology and a new J-SHIS system was launched. The new J-SHIS manages various data in an integrated manner. It includes the new National Seismic Hazard Maps for Japan which consist of the Probabilistic Seismic Hazard Maps (PSHM) for Japan with a 250 m mesh resolution and the Scenario Earthquake Shaking Maps (SESM) based on detailed strong motion estimation of earthquakes occurring at major active fault zones, as well as the deep subsurface structure models for Japan and 250 m mesh geomorphological land classification models used for the required calculations. The new J-SHIS also provides these data in a user-friendly manner by superposing them on background maps. The new J-SHIS is a web mapping system based on open source software which allows general users to easily view various data on their Internet browsers.

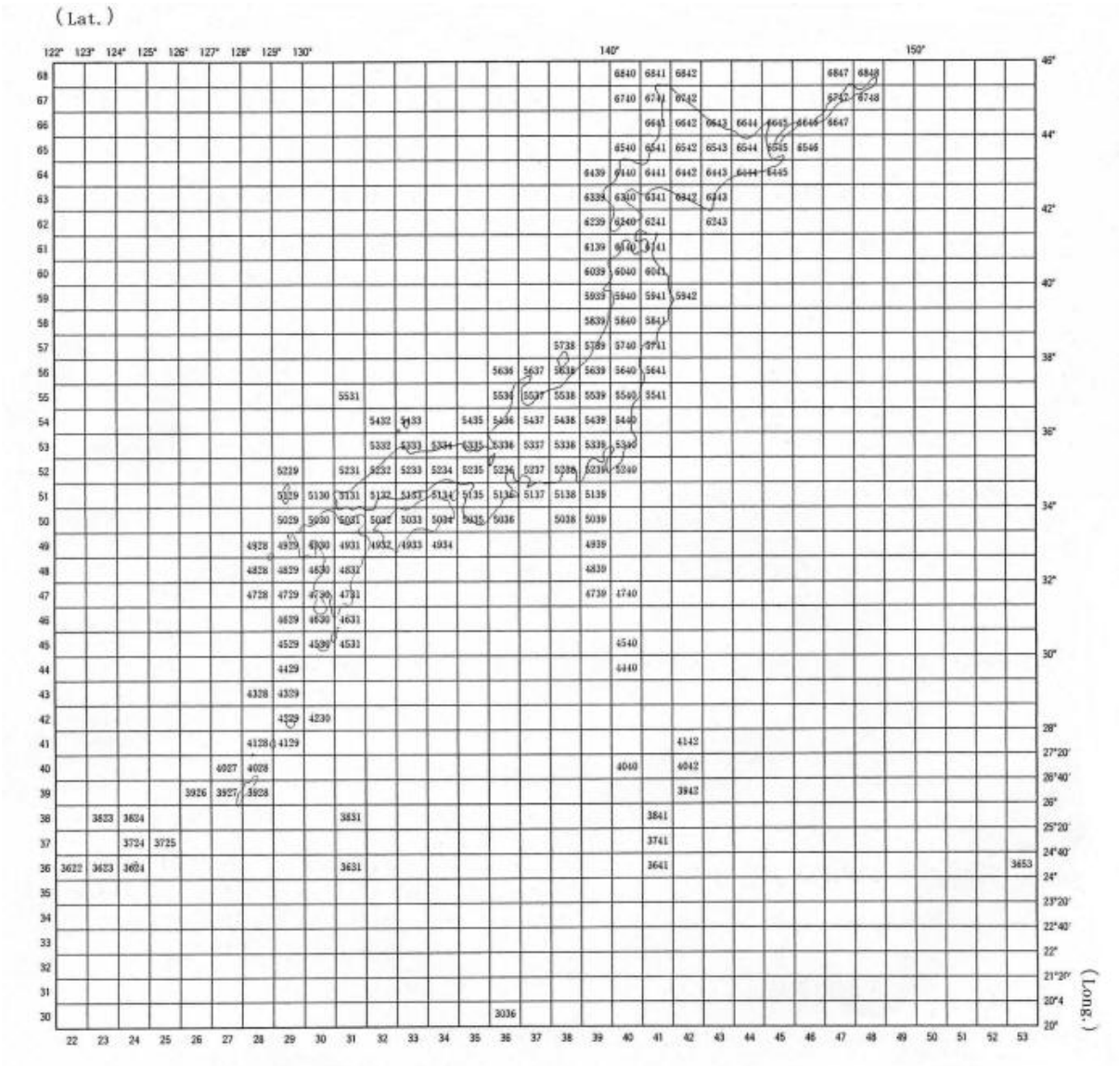



Figure 3.5: Map of first meshs

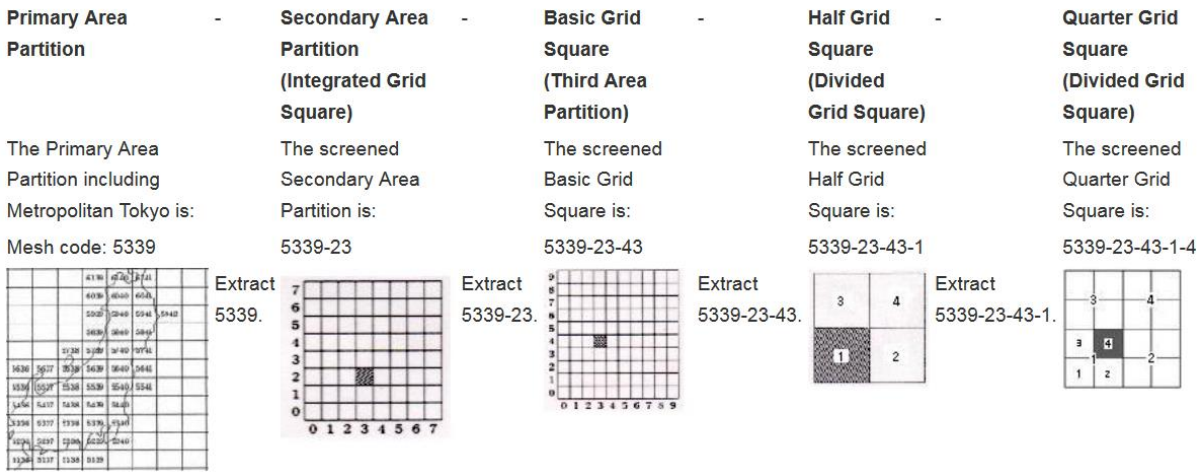



Figure 3.6: Method of calculating mesh code numbers from coordinates.

Especially, the notable new functions have enabled the users to overlap the seismic hazard maps on Google maps including the layer transparency function, to freely move and zoom in and out the maps, to view the seismic hazard maps with a 250 m mesh resolution, to search a precise location by addresses and postal codes, to select and show a source fault on the browser, and to display attribute values for each mesh. The new system has been in operation since July 2009.

Responding to the 2011 off the Pacific Coast of Tohoku Earthquake, studies are being made on improvement of J-SHIS.

### 3.9.2 Data source

The relevant J-SHIS data is the ‘probability of exceedance’. According to

`www.j-shis.bosai.go.jp/en/glossary`,

this term means:

> *The probability that shakes will exceed a certain level of intensity at a point for a certain time period (over the next 30 or 50 years, in this guidebook). For example, the ‘Map of ground motions of seismic intensity for a 3% probability of exceedance occurring within 30 years from the present’ means the probability that each point is affected by shakes exceeding its seismic intensity shown on the map is 3% within 30 years from the present.*

The probabilities are either ‘average case’ or ‘maximum case’.

‘Average case’ means:

> *In the long-term evaluation for the 98 major active fault zones, there are many cases where both mean recurrence interval and the latest event have been evaluated with a range of values. In the preparation of the seismic hazard map, the result of calculating the probability of occurrence by using the mid-values of individual ranges is called ‘Average case’*,

while ‘Maximum case’ means:

*The result of calculating the probability of occurrence by using the smallest value of mean recurrence interval and the oldest value of the latest event is the highest probability.*

Two analytical periods are typically given: 30 years (T30) and 50 years (T50). In our data set we confine ourselves to the 'average case' and an analytical period of 30 years.

The J-SHIS data are available in two different formats for two subperiods. For 2008–2014, the probability of earthquakes happening in the coming 30 or 50 years exceeding certain intensity thresholds can be downloaded using `Python` from

`www.j-shis.bosai.go.jp/map/JSHIS2/download.html?lang=en`

under tab 'Dataset/Probablistic Seismic Hazard Maps/Seismic Hazard Map'. The records come in separate `.csv` files for each year and each first mesh. We downloaded 8 first meshes: Tokyo `5339`; Osaka `5135` and `5235`; Nagoya `5235`, `5236`, and `5237`; Fukuoka `4930` and `5030`; Sapporo `6441`. Each entry corresponds with a 1/4 mesh area (250m), and the thresholds are 5-lower (denoted as I45), 5 upper (denoted as I50), 6-lower (I55) and 6 upper (I60). There can be multiple models in one year. In 2012 there are data for two models and in 2013 there are data for three models. We use model 1 by default in both years.

For 2005–2008, the PSHM map data can be downloaded from

`wwwold.j-shis.bosai.go.jp/j-shis/index_en.html`,

one zip file each year for the whole of Japan. The thresholds can be chosen as 6-lower or 5-lower. Each entry corresponds with a third mesh area (1km). The probabilities are only available for an analytical period of thirty years. There is some overlapping in the two sources: both the new and the old system provides data for year 2008. Although in different formats, these data coincide (maximum difference is 5e-7, which is just rounding error). For further analysis we use the new data source for the year 2008.

### 3.9.3 Summary statistics

In summary, there are annual data for two subperiods: 2005–2008 and 2008–2016. (The year 2008 appears in both sets.) In the first period each mesh is about 1 square km (third mesh); in the second period each mesh is about 250 square meters (quarter mesh).

Seismic intensity is given on a scale from 0 to 7, where 5 and 6 are further divided into 5-lower, 5-upper, 6-lower, and 6-upper. In the second period the data provide:

- probability of exceedance larger than 5-lower (30 years);
- probability of exceedance larger than 5-upper (30 years);
- probability of exceedance larger than 6-lower (30 years);
- probability of exceedance larger than 6-upper (30 years).

In the first period, only two of these probabilities are provided, namely 5-lower and 6-lower. Summary statistics of the hazard probabilities for all the stations in the housing sample are provided below.

Table 3.22: Summary statistics of seismic hazard probabilities (for each unique district in the housing sample), averaged over 2005–2014

| City | mean | min | 25% | 50% | 75% | max | sd | #districts |
|---|---|---|---|---|---|---|---|---|
| *Exceeding intensity level '5 lower'* | | | | | | | | |
| Tokyo | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 0.00 | 898 |
| Osaka | 0.93 | 0.90 | 0.92 | 0.94 | 0.95 | 0.97 | 0.02 | 564 |
| Nagoya | 0.96 | 0.91 | 0.94 | 0.97 | 0.98 | 0.98 | 0.02 | 1379 |
| Fukuoka | 0.39 | 0.06 | 0.30 | 0.42 | 0.48 | 0.56 | 0.12 | 318 |
| Sapporo | 0.33 | 0.05 | 0.21 | 0.33 | 0.44 | 0.51 | 0.12 | 551 |
| *Exceeding intensity level '6 lower'* | | | | | | | | |
| Tokyo | 0.35 | 0.16 | 0.22 | 0.28 | 0.49 | 0.59 | 0.13 | 898 |
| Osaka | 0.37 | 0.22 | 0.30 | 0.39 | 0.44 | 0.52 | 0.09 | 564 |
| Nagoya | 0.56 | 0.21 | 0.41 | 0.61 | 0.67 | 0.77 | 0.14 | 1379 |
| Fukuoka | 0.03 | 0.00 | 0.02 | 0.03 | 0.03 | 0.05 | 0.01 | 318 |
| Sapporo | 0.01 | 0.00 | 0.01 | 0.01 | 0.02 | 0.03 | 0.01 | 551 |

Generally speaking, Tokyo, Nagoya and Osaka are all highly prone to small earthquakes. The probability of an earthquake in Tokyo (in a 30 year period) with intensity more severe than '5-lower' is close to 1. Nagoya is more likely to have larger earthquakes than Tokyo and Osaka. The variation of probabilities of severe earthquakes within each of these three big cities is also much larger than the other two. Fukuoka and Sapporo seem less risky to have larger earthquakes, but there is still considerable probability of small earthquakes. These tables show that there is sufficient variation in the 'riskiness' of the five cities and that the risk measures under both thresholds (intensities 5-lower and 6-lower) are important in characterizing a distribution of earthquake risk.

## 3.10 Districts versus stations

### 3.10.1 Two options

From the property price data set (as described in Section 3.4), the district, ward, city and the nearest station of each property is provided. Since we do not know the exact location of a property, we need to use a proxy. There are two options. The first option is to identify a property with the district in which it lies. A city consists of wards, and wards consist of districts. So there are many districts within one city.

The second option is to identify a property with its nearest station. In our analysis we confine the data to properties within 30 minutes of a station (walking distance) and we know the name of this nearest station.

We shall investigate both options.

### 3.10.2 Station coordinates

Since we confine our analysis to properties that are within 30 minutes walking distance from a station, it is not unreasonable to identify the location of the property by the location of that metro station. There are fewer stations than there are districts, so this division will be somewhat less accurate.

We find the location information of each station, given the name of the station and the city it is located in. It is necessary to specify the city as well since in some cases the same station name appears in different cities (such as Nakanoshima Station in Sapporo and Osaka).

We use Google Maps to find the coordinates of each station. In the search result page on Google Maps, geographic coordinates appear in the url and can be collected for further usage. In cases where station name is ambiguous and direct requesting yields zero results, manual correction is needed, for example inserting spaces in the long station name or using the Japanese translation instead.

The station coordinates are double-checked by comparing the output with those found on the Wikipedia pages of Japanese railway stations.

`en.wikipedia.org/wiki/List_of_railway_stations_in_Japan`.

We calculate the distance between coordinates found with the two different sources. Out of the 1022 unique stations in our sample, the coordinates discrepancies for 95% are less than 150 m, and those for only two records are more than 1 km (but both are less than 1.6 km).

### 3.10.3 District coordinates

Using Google Maps,the location information for each district can also be found. Since districts are often irregularly shaped, the coordinates are approximate and not necessarily at the exact geometric center of a district. Postcodes can be obtained from the formatted address in the output.

### 3.10.4 From coordinates to mesh codes

Using either the district of the nearest station, we are able to obtain the rough location of each property. The location information can then be used to find the time-varying seismic risk probabilities associated with each property. In the JSHIS risk data set, the seismic hazard information is stored for each mesh grid.

Given the coordinates, we calculate the 10 digit mesh codes using the method described in Figure 3.6.

A brief summary of the number of districts, stations and the number of sample records for each first mesh code is shown below. We can see that the number of districts is much larger than the number of stations, so using the district center of a property as its location proxy is more accurate. Also the ward characteristics are district related. Therefore we have chosen to use the district information in the main paper. The results for using stations as location proxy are included as a part of the sensitivity analysis.

Table 3.23: Number of districts and stations for each first mesh code

| City | first mesh | #districts | #stations | #samples |
|---|---|---|---|---|
| Tokyo | 5339 | 898 | 482 | 184077 |
| Osaka | 5235 | 304 | 115 | 30107 |
| | 5135 | 260 | 105 | 19713 |
| Nagoya | 5236 | 1295 | 150 | 36552 |
| | 5237 | 84 | 9 | 2227 |
| Fukuoka | 5030 | 318 | 75 | 25982 |
| Sapporo | 6441 | 551 | 86 | 32685 |

### 3.10.5 Cross validation

In the procedures of finding coordinates using location names and finding mesh codes using coordinates, we used multiple sources in order to minimize the risk of associating properties with wrong locations. However, it is still possible that either or both of the location proxies were wrongly reported since the property price data set is obtained from surveys. We thus need to check the validity of the location information of each record.

For a given record in the sample, we have obtained the coordinates of the nearest station; we also know the approximate center of the district where the property is located in. Since we have chosen the walking distance to nearest stations to be less than 30 minutes, the distance from the nearest station to the district center should not be too large. If the distance is above a certain threshold, then this record is suspicious and manual check is in order. We have chosen this threshold to be 6.22 km which is the 99% quantile of the station to district center distances of all the unique station-district pairs.

In the end we narrowed down the list of questionable records to 45 station-district combinations that seemed invalid. These correspond to 47 records in the sample and have been excluded.

# Chapter 4

# Earthquake risk embedded in property prices: Evidence from five Japanese cities

## 4.1 Introduction

Using the data set described in Chapter 3, we employ a hedonic property price model with a multivariate error components structure to analyze the subjective evaluation of both short-run and long-run earthquake risk embedded in Japanese property prices[1]. It is well-known that earthquakes tend to occur in clusters rather than in isolation. These seismic clusters may take the form of foreshocks and aftershocks anticipating and following a major earthquake or of a collection of major earthquakes triggering one another by causing frictions that put strain on neighboring faults. There is therefore objective predictive content embedded in the occurrence of earthquakes. This phenomenon is known as seismic excitation and there exists a large literature in statistics aimed at capturing it.

In a different strand of the literature in economics, several papers analyze the impact of natural catastrophes on property prices. Most commonly, this literature incorporates the prevailing binary state of the world, depending on whether or not a catastrophe has occurred, into a hedonic house price model of the Rosen (1974) type, which has become the benchmark model in analyzing property prices. Within a typical hedonic price model, the characteristics of a property are viewed as detachable components that each contribute to a part of the property price. The selection of components range from traditional house attributes such as square footage, location and building age, to external factors such as macroeconomic effects. The negative effect coming from hazardous environmental events, such as flood, hurricane and earthquake, has been addressed by various researchers; see, among others, Brookshire et al. (1985); Kawawaki and Ota (1996); Beron et al. (1997); Yamaga et al. (2002); Bin and Polasky (2004); Nakagawa et al. (2007, 2009); Daniel et al. (2009); Naoi et al. (2009); Gu et al. (2011); Bin and Landry (2013); Hanaoka et al. (2018); Hidano et al. (2015).

In recent years, a large body of literature has documented empirically that people do typically not

[1] An updated version of this chapter has been published (Ikefuji et al., 2021). Note: The published paper comes with three supplementary files: First, the Appendix ( `https://in05.hostcontrol.com/resources/bc83e0a2cccc29/39400b566c/file-object/ILMY-Earthquake-JASA-accepted-Supp.pdf`) which is an intrinsic part of the paper; second, the Data Documentation (`https://bit.ly/3qHcTQ3`) which contains a description of the data, but not the actual data set; and third the actual data and the R codes (`https://github.com/yy112/earthquake-risk`).

treat objectively given probabilities linearly, but rather tend to overweight small probability events and underweight large probability events. This is particularly relevant when evaluating catastrophic events that are often of a low-probability high-impact nature. Various modern theories of decision under risk, such as rank-dependent utility theory and prospect theory, feature a probability weighting function that 'distorts' objective probabilities.

We contribute to this literature by introducing into a hedonic price model an objective measure of seismic excitation, next to a more conventional measure of long-run earthquake risk, while allowing for probability weighting in the spirit of the non-expected utility theories of rank-dependent utility and prospect theory. We use a hedonic price model with the multivariate error component structure described in Chapter 2, which enables us to estimate the model while pooling properties of different types together, in spite of the very large dimension of the variance matrix and the fact that each property type corresponds to different features and total price levels. Our approach allows to isolate the total compensation for earthquake risk embedded in Japanese property prices, and to decompose this into pieces stemming from short-run risk and long-run risk, and a further decomposition into objective and distorted risk components.

We can summarize our main findings as follows. First, we find that objective long-run earthquake risk has a significant negative impact on property prices, and increasingly so at higher risk levels. Second, given that long-run risk matters for property prices, we find that the additional impact of objective short-run earthquake risk on property prices, while estimated at negative values, is not significantly different from zero. Upon allowing for probability weighting, however, the distorted short-run earthquake probabilities do have a significantly negative effect on property prices. Third, the probability weighting function for short-run earthquake risk is found to be S-shaped, thus underweighting small probabilities and overweighting larger probabilities, contrary to the inverse S-shaped probability weighting function found in many experiments. This remarkable finding may be explained by the fact that the background arrival rate of earthquakes is positive rather than zero, in particular in Tokyo where the short-run earthquake probabilities never drop below 35% in the period that we analyze. Therefore, people may tend to evaluate and overweight temporary deviations of the short-run earthquake probabilities from the background seismicity caused by seismic excitation not with respect to zero but with respect to a positive reference probability level. In an extension of our base model, we also analyze probability distortions of long-run time-invariant earthquake probabilities. In this case, we find that small probabilities tend to be overweighted and large probabilities tend to be underweighted, in accordance with conventional wisdom.

Most of the studies on the interplay between property prices and environmental hazards investigate the risks of floods or earthquakes in the USA or Japan. In the case of the USA, Brookshire et al. (1985) analyze a hedonic house price model in an expected utility framework, examine self-insurance for earthquake hazards in Los Angeles and San Francisco, and show that buyers pay less for houses within a relatively risky area if they possess adequate information about earthquake hazards. Bin and Polasky (2004) estimate and compare the effects of flood hazards on property prices before and after Hurricane Floyd (the 1999 flooding in North Carolina), and show that the market price of a property located within a flood plain gets discounted by more than a property located outside the flood plain. Re-examining these findings, Bin and Landry (2013) estimate hedonic property prices for the same location

with two major flooding events, and show that the implicit risk premia disappear rapidly.

In the case of Japan, Nakagawa et al. (2009), using the 1998 Tokyo hazard map, show strong negative impacts of earthquake risks on land prices. Gu et al. (2011), using an updated Tokyo hazard map, find that in previously safe areas, a decrease in risk rankings (even more safety) has a positive impact on relative land prices, while in previously dangerous areas, an increase in risk rankings (even more risk) has a negative effect. Naoi et al. (2009) estimate individuals' valuation of earthquake risk, based on nation-wide panel data of earthquake hazard information and records of observed earthquakes. They show that after a big earthquake people discount house prices and house rents within the earthquake area. Hidano et al. (2015) examine the effect of seismic hazard risk information on properties in Tokyo, and find that the price of properties in low-risk zones is higher than the prices in high-risk zones, but that for new more earthquake-resistant properties the influence of seismic hazard risk information is limited.

We also mention two survey-data studies on how risk preferences of households have changed after the Tohoku earthquake (the Great East Japan earthquake) in 2011. Naoi et al. (2012) find that although respondents seemed to be more prepared for natural disasters after the Tohoku experience, actual (costly) mitigation activities depend on household income. Hanaoka et al. (2018) examine whether risk preferences of men and women have changed, and if so whether they changed in a different way, after the Tohoku earthquake. There is some evidence that men have become more risk tolerant, while women have become more risk averse. Finally, our work is also related to the financial econometrics literature on the estimation of risk and financial excitation premia embedded in asset and derivative prices; see Aït-Sahalia et al. (2014, 2015), and Boswijk et al. (2016).

The rest of this chapter proceeds as follows. Section 4.2 explains our treatment of objective seismic excitation and of probability weighting. Section 4.3 describes the data set. Section 4.4 lays out the model. Section 4.5 presents the estimation results. Section 4.6 analyzes the influence of each component to the total property prices and calculates the implied premia for earthquake risk. Section 4.7 discusses the robustness of our estimation results. Section 4.8 concludes.

## 4.2 Seismic excitation and probability weighting

In this section we consider short-run earthquake probabilities as objective measures of seismic excitation, and develop a regression design that allows for probability weighting.

### 4.2.1 Short-run earthquake probabilities

Our approach estimates an Epidemic Type Aftershock Sequence (ETAS) model and generates a panel of model-implied short-run earthquakes probabilities which vary per quarter and per city to be used in our regression design. These probabilities can be viewed as objective measures of short-run earthquake risk, summarizing publicly available information per time period and per city.

The occurrence of major earthquakes have served previously in hedonic price models with regression discontinuity design as natural exogenous events to elicit causal pricing effects. Limitations of this conventional approach include the somewhat rudimentary binary nature of this treatment, which does not reflect the multiplicity of the events, the time elapsed since the last event, and the severity of the events.

By contrast, our approach relies on a continuous-time predictive earthquake intensity that depends on all previous earthquakes, with recent ones being more important than older ones, and explicitly accounts for the severity of the events. Furthermore, the earthquake intensity can be translated into objective short-run probabilities enabling us to analyze probability weighting.

The ETAS model was introduced by Ogata (1988) and has since been widely used to capture the quiescence and activation of seismic dynamics. The basic idea of the model is that each earthquake can trigger a sequence of aftershocks like 'epidemics' in that the occurrence of an earthquake makes future earthquakes more likely and that the impact of the trigger event diminishes over time (and distance). Despite the existence of several space-time extensions, we choose the temporal version of the ETAS model as described in the following, which we estimate separately for each of the five cities. Because we consider five cities this treatment is natural and simpler than first estimating a space-time version to a large area that covers all five cities and next isolating the city effects.

Formally, the ETAS model is a path-dependent marked point process and a special case of a Hawkes self-exciting process. Given observations of earthquake occurrences at times $t_1, t_2, \ldots, t_n$ over an interval $[0, T]$ $(T \geq t_n)$, the associated counting process $N_t$ is defined as $N_t = \sum_{i=1}^{n} \mathbb{1}_{t_i \leq t}$. Denoting by $\mathscr{F}_t$ the information filtration up to time $t$, the corresponding left-continuous $\mathscr{F}_t$-conditional jump intensity process $\lambda_t$ describes the mean jump rate per unit of time,

$$\lambda_t = \lambda(t|\mathscr{F}_t) = \lim_{h \downarrow 0} \frac{1}{h} \Pr\left[N_{t+h} - N_t > 0 | \mathscr{F}_t\right].$$

In the temporal ETAS model, the conditional intensity function may be written as

$$\lambda_t = \lambda_\infty + \sum_{t_i < t} c(m_i, m_c) g(t - t_i),$$

where $\lambda_\infty > 0$ (measured in number of jumps per time unit) is the background seismicity, $g(t - t_i)$ is the aftershock decay (i.e., time response) function, and the weight assigned to the aftershock decay is a function $c(m_i, m_c)$ of the magnitude of the earthquake $m_i$ and a cut-off (i.e., threshold) magnitude $m_c$. Thus, the earthquake intensity depends on the background intensity and a weighted sum of all aftershock decays, where the sum is taken over all earthquakes that have occurred before time $t$. (In the ETAS model, $g$ takes the form of a so-called modified Omori law and $c$ takes an exponential form.)

We estimate the ETAS model for each of the five cities that we consider, based on the earthquake catalog of five areas covering the five cities, over the period January 1, 1970, to December 31, 2015. Next, the estimated intensities are used to generate by simulation 90-days probabilities of an earthquake exceeding a magnitude threshold of 5.5, for each city. Our simulation method follows Ogata (1981). Further details about the parameterization, estimation and simulation within the ETAS model are explained in Chapter 3.

In Figures 4.1 and 4.2 we plot the earthquake intensities along with the corresponding short-run probability series for two of the five cities, Tokyo and Nagoya. The probabilities spike up immediately after a large earthquake and die out gradually until another major earthquake occurs. The Tohoku earthquake of Friday 11 March 2011 was the most powerful earthquake ever recorded in Japan. The spike is visible in 2011/Q2 (rather than in 2011/Q1) because the short-run probabilities are simulated based on

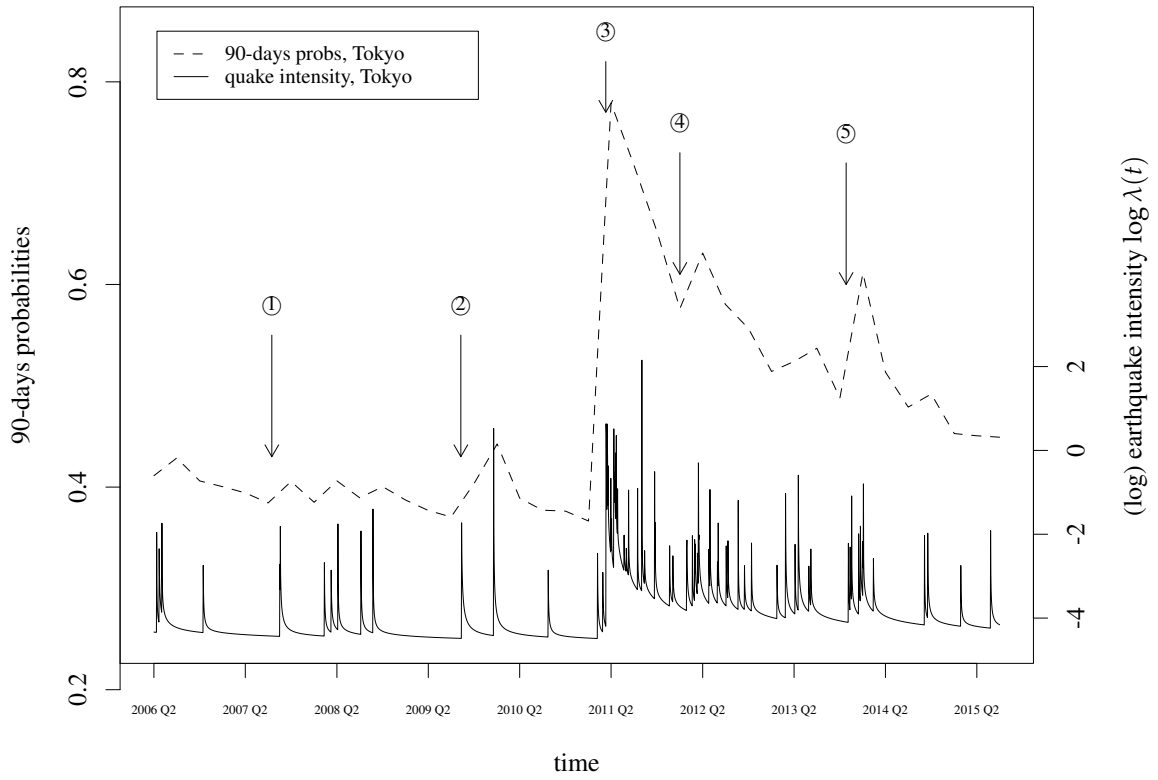



Figure 4.1: Short-run earthquake risk: Simulated short-run earthquake probabilities and the logarithm of the earthquake intensity series for Tokyo. Events marked in the graph: ①: 2007-07-16 Chuetsu Offshore earthquake, $M6.8$. ②: 2009-08-09 Izu Islands earthquake, $M6.8$ and 2009-08-11 Shizuoka earthquake, $M6.5$. ③: 2011-03-11 Tohoku earthquake, $M9.0$. ④: 2012-01-01 Izu Islands, $M7.0$. ⑤: 2013-10-26 Fukushima-ken oki earthquake, $M7.1$. (Source: Japan Meteorological Agency.)

actual earthquakes up to and including the previous quarter.

The objective measure of seismic excitation given by the 90-days earthquake probabilities is included in our regression design. The rationale is that, in addition to the long-run earthquake risk that people may take into consideration when purchasing a property, news from a recent nearby earthquake may also temporarily affect property prices. Just like objective seismic excitation generated by a self-exciting process, the impact of such bad news on people's perception of risk peaks right after the event and dies out as time proceeds.

### 4.2.2 Probability weighting

To account for probability weighting our regression design furthermore allows for a parametric probability weighting function. There is a large literature on probability weighting. Probability weighting is an important ingredient of prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), and of the related decision theories given by the dual theory of choice under risk (Yaari, 1987) and rank-dependent utility (Quiggin, 1982), which are building blocks of prospect theory.

We shall consider two canonical one-parameter families of probability weighting functions, pro-

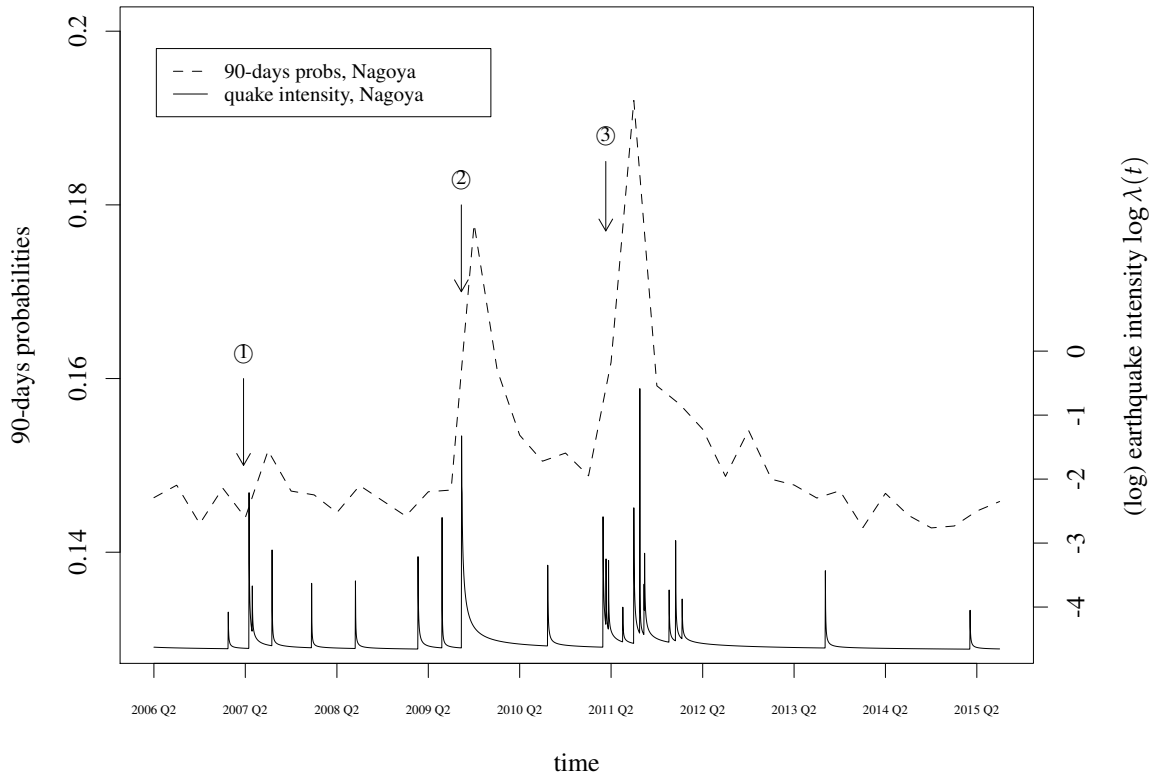



Figure 4.2: Short-run earthquake risk: Simulated short-run earthquake probabilities and the logarithm of the earthquake intensity series for Nagoya. Events marked in the graph: ①: 2007-03-25 Noto Hanto earthquake, $M6.9$. ②: 2009-08-11 Shizuoka earthquake, $M6.5$. ③: 2011-03-11 Tohoku earthquake, $M9.0$. (Source: Japan Meteorological Agency.)

posed by Tversky and Kahneman (1992) and Prelec (1998), respectively. The Tversky-Kahneman function — see also Wu and Gonzalez (1996) — is given by

$$w(p) = \frac{p^\psi}{(p^\psi + (1-p)^\psi)^{1/\psi}}, \tag{4.1}$$

while the Prelec function is given by

$$w(p) = e^{-(-\log p)^\psi}. \tag{4.2}$$

The parameter $\psi$ is restricted to be positive. When $0.279 < \psi < 1$ the Tversky-Kahneman function is inverse S-shaped, while the Prelec function is inverse S-shaped for $0 < \psi < 1$. The lower bound in the parameter for the Tversky-Kahneman function ensures monotonicity. When $\psi = 1$ both functions reduce to $w(p) = p$; and when $\psi > 1$ both functions are initially S-shaped, but (only) the Tversky-Kahneman function becomes convex for large values of $\psi$.

In laboratory experiments (see Wu and Gonzalez (1996), and Abdellaoui (2000)) the probability weighting function is often found to be inverse S-shaped, first concave and then convex. An inverse

S-shape captures the phenomenon that people tend to become less sensitive to changes in objective probabilities as these probabilities move further away from the reference point 0 and become more sensitive as they get closer to the reference point 1. The inverse S-shape is consistent with a positive third derivative of the probability weighting function. The interpretation of the signs of the successive derivatives of the probability weighting function was recently provided by Eeckhoudt et al. (2020). Note that contrary to the Tversky-Kahneman function the Prelec function has an invariant fixed point and inflection point at $p = 1/e = 0.37$, which implies that it can never be globally convex or concave.

## 4.3 The data

The data collection process for this project has been complex and elaborate, and in this section we provide a brief summary. Full details and references to sources are available in Chapter 3. We are interested in the impact of earthquake risk on property prices in major cities in Japan, and we have selected five cities for our purpose. Each city is divided into wards and each ward is divided into districts. (In the original data set the word 'area' is used. We prefer 'district' to avoid confusion with other uses of the word 'area'.) Certain information that can affect (and explain/predict) the attractiveness of buying a property is available per ward. For example, population characteristics, information about schools and medical facilities, shopping, safety, etc. We distinguish between three types of properties: 'residential land (land and building)', 'residential land (land only)', and 'pre-owned condominiums' (hereafter, condos). Sales prices and property characteristics are available for each of these types in each of the five cities. We do not know the exact location of a property, but we do know in which district the property lies and we also know the distance to the nearest station and the name of that station. Some macro variables are relevant and affect house prices nationally. Finally, we have information on historical earthquake data and on earthquake risk data.

*Cities.* Japan has twelve cities with a population of more than one million people. Almost 100 million people, or 78% of the country's total population of 127.4 million, live in urban areas. The total population of Japan's largest 103 cities amounts to 63.9 million or just over half of all the country's residents. Tokyo, with almost nine million inhabitants, is by far the largest Japanese city. (Strictly speaking, Tokyo is not a city — it is a prefecture, but we shall call it a city.) With a population of 3.7 million, Yokohama, south of Tokyo, is Japan's second largest city. Osaka and Nagoya are Japan's third and fourth cities, each with a population of over two million. Eight cities have between one and two million inhabitants: Sapporo, Kobe, Fukuoka, Kyoto, Kawasaki, Saitama, Hiroshima, and Sendai.

From these twelve cities we selected five: Tokyo, Osaka, Nagoya, Fukuoka, and Sapporo. This choice guarantees that each of the three major metropolitan areas is represented: the greater Tokyo area (Tokyo, Yokohama, Kawasaki, Saitama) by Tokyo, the Kansai region (Osaka, Kobe, Kyoto) by Osaka, and the Chukyo metropolitan area by Nagoya. To obtain a representative geographical spread we added Sapporo, the largest city in the North, and Fukuoka, the second largest city in the West after Osaka. Data limitations prevented us from including Hiroshima, while Sendai was not included because it is too close to Fukushima where the 2011 nuclear disaster took place following the Tohoku earthquake.

*Wards.* A designated city is a Japanese city that has a population greater than 500,000 and has been designated as such by order of the Cabinet of Japan. Designated cities are delegated many of the tasks normally performed by prefectural governments, such as public education, social welfare, sanitation, business licensing, and urban planning. Designated cities are required to subdivide themselves into wards ('ku'), each of which has a ward office conducting various administrative functions for the city government. The 23 special wards of Tokyo are not part of this system, as Tokyo is a prefecture, and its wards are effectively independent cities. The five cities together contain 80 wards (regular and special together): 23 in Tokyo, 24 in Osaka, 16 in Nagoya, 7 in Fukuoka, and 10 in Sapporo.

When considering to buy a property in a given city, one is likely to be interested in certain characteristics of these wards. The original data set contains one hundred characteristics divided into eleven categories. Since many of these are highly correlated we first select eleven of these divided into six categories: two from population; three from schools, culture and welfare; one from medical facilities; one from safety; two from shopping facilities; and two from employment. Only four of these appear in our base model, but extensive sensitivity analyses will be conducted in Section 4.7 to assess how adding more characteristics may affect the results.

*Districts.* Within each city there are wards, and within each ward there are districts (usually 'cho', sometimes 'machi'). An average ward in Nagoya contains 86 districts, an average ward in Osaka only 23. The number of districts ranges from 318 in Fukuoka to 1383 in Nagoya (1379 after prescreening). In total there are 3714 districts (3710 after prescreening) in the five cities together.

*Property types.* In a given district $i$ we have observations on three types of (residential) properties: land and buildings, land only, and condos. Most properties are condos (45.1%), followed by land and buildings (34.1%) and land only (20.8%). We have observations over $T = 38$ quarters, from 2006/Q2 to 2015/Q3.

Table 4.1: Distribution of properties over cities, wards, and districts

| City | Ward | District | Land & building | Land only | Condo | Station |
|---|---|---|---|---|---|---|
| Tokyo | 23 | 898 | 57,568 | 33,991 | 92,518 | 482 |
| Osaka | 24 | 564 | 21,064 | 6,901 | 21,855 | 220 |
| Nagoya | 16 | 1,379 | 14,640 | 13,110 | 11,029 | 159 |
| Fukuoka | 7 | 318 | 7,847 | 5,660 | 12,475 | 75 |
| Sapporo | 10 | 551 | 11,763 | 9,461 | 11,461 | 86 |
| Total | 80 | 3,710 | 112,882 | 69,123 | 149,338 | 1,022 |

Records with obvious errors have been excluded. Also excluded are records where the walking time to the nearest station is longer than thirty minutes or the nearest station is unknown; records with a living area larger than 2000 square meters; and properties built before the war (1945). After applying the above criteria we are left with $N = 3710$ districts in total. The number of wards, districts, properties of each type, and stations in each city is displayed in Table 4.1.

*Property prices and characteristics.* We work with sales prices rather than with rental prices, because sales are more permanent than rentals and we would therefore expect that the effect of earthquake risk on choosing a property will be more informative.

Nakagawa et al. (2009) use land prices over various years (from 1980 onwards) and describe the data in their Section 3 (for the Tokyo area). Their data are based on the Koji-Chika data set published by the Ministry of Land, Infrastructure, Transport, and Tourism (MLIT). The Koji-Chika data set provides *fictional* sales prices (as produced by 'experts') and they are only available at annual intervals, which we consider to be too long for our purpose. We use a different data set, which provides self-reported *transaction* prices at three-months intervals. This data set, also provided by the MLIT, is known as the 'real estate transaction-price information'; see

`www.land.mlit.go.jp/webland_english/servlet/MainServlet`.

The information in this data set is based on the results of a questionnaire survey of persons involved in real estate transactions conducted by MLIT, compiled and published quarterly. We thus know the transaction price and the transaction date (quarter), and also in which district the property lies and the name of the nearest station. In addition, many property characteristics are provided, of which we shall only consider: total area in square meters, total floor area in square meters, distance to nearest station measured in walking minutes, age of the building (if applicable), building structure (reinforced concrete, steel, or wood), purpose of city planning in the urban control area, maximum building coverage ratio, and maximum floor area ratio. Different types may have different regressors. For example, the equation for land only does not have 'building structure' or 'building age' as a regressor; and the equation for condos does not use 'building structure' as a regressor.

*Economic indicators.* Property prices are affected by general economic conditions. In order to incorporate possible effects of these economic conditions, we have selected two national macroeconomic indicators: GDP and CPI.

*Long-run earthquake risk.* We consider two measures of earthquake risk: short-run risk (i.e., seismic excitation; see Section 4.2.1) and long-run risk. Long-run earthquake risk is defined as the probability of an earthquake exceeding certain intensity thresholds in the next 30 years in a given area, provided by the Japan Seismic Hazard Information Station (JSHIS). We select the threshold intensities '5-lower' (medium risk) and '6-lower' (high risk) in our analysis. The JSHIS probabilities are provided in various mesh sizes, varying from one square km to 250 square meters. For each district we identify its center and then define the risk of that district as the JSHIS risk associated with the smallest available mesh in which this center lies. Although the JSHIS exceedance probabilities are updated every one or two years, we take the average of the JSHIS risk data over all available years, thus obtaining a time-invariant measure of long-run risk for each district. These probabilities are included as objective measures of long-run earthquake risk in our regression design, at the district level. Choosing a district of relative safety may be viewed as a form of self-insurance. Therefore, provided this information, which is publicly available, is

known among consumers, we would expect higher property prices in relatively safe areas all else being equal.

If the intensity is '5 lower' then, according to the Japan Meteorological Agency, many people will be frightened and feel the need to hold on to something stable. Hanging objects (such as lamps) will swing violently, books may fall from bookshelves, and unstable furniture may topple over. Windows may break and fall down, electricity poles may move, and roads may sustain damage. There may be cracks in the walls of wooden properties. If the intensity is '6 lower' then the effects will be more severe. It will be difficult to remain standing, unsecured furniture will move and topple over, and cracks in walls, crossbeams, and pillars will appear not only in wooden properties but also in properties built from reinforced concrete.

Summary statistics are shown in Table 3.22. It is clear from Table 3.22 that Tokyo, Nagoya, and Osaka are high-risk cities with respect to 'small' earthquakes. In fact, it is almost certain that an earthquake will occur in Tokyo with an intensity more severe than '5 lower' within the next 30 years. Regarding the occurrence of 'severe' earthquakes ('6 lower'), Nagoya is more exposed than Tokyo and Osaka, and much more exposed than Fukuoka and Sapporo. The variation in probabilities of severe earthquakes in Tokyo, Osaka, and Nagoya is also much larger than in the other two cities. Fukuoka and Sapporo are not likely to have severe earthquakes, but there is still considerable probability (and variation) of smaller earthquakes. This suggests that it is important to use both thresholds, 5-lower and 6-lower, in characterizing the distribution of long-run earthquake risk for our purpose. In particular, this will guarantee sufficient variation of long-run probabilities in the hedonic price model discussed in Section 4.4.

## 4.4 The model

The dependent variable is log-property price, and we denote the $h$-th observation of type $k$ in district $i$ during quarter $t$ as $y_{it}^{(h,k)}$. The most common method of modeling the property market is hedonic pricing, pioneered by Rosen (1974) who argued that an item's total price can be thought of as the sum of the price of each of its homogeneous characteristics, so that the effect of each characteristic on the price can be determined by regressing (log)price on these characteristics.

We shall follow the hedonic approach. In our case the (log)price is determined by characteristics of the property itself (size, age, etc.), the surrounding environment (location, crime rate, schools, etc.), earthquake risk factors, and macroeconomic influences.

The district $i$ determines the city $c(i)$, which takes values $1,\ldots,5$ depending on the city in which district $i$ is situated. Also, the time variable $t$ determines in which quarter $q(t)$ the transaction took place, taking values $1,\ldots,4$ depending on whether $t$ refers to the first, second, third, or fourth quarter. The number of observations varies per district, type and quarter, and this affects the precision. We let $H_{it}^{(k)}$ denote the number of observations on each type $k = 1,2,3$ in district $i$ during quarter $t$.

We model the difference between cities by a shift $\alpha_{c(i)}$ in the intercept term, but we assume that all other parameters are the same between cities. The difference between cities is thus completely captured by the $\alpha_{c(i)}$.

Our model can now be written as

$$y_{it}^{(h,k)} = \alpha_0^{(k)} + \alpha_{c(i)} + \gamma_{q(t)} + x_{i\cdot}^{(k)\prime}\beta_1 + x_{\cdot t}^{(k)\prime}\beta_2 + x_{it}^{(h,k)\prime}\beta_3 + r_{it}(\psi)'\beta_4 + u_{it}^{(h,k)}, \tag{4.3}$$

where $x_{i\cdot}$ denotes a variable that is constant over time, but varies over districts (attractiveness variables), $x_{\cdot t}$ denotes a variable that is constant over districts, but varies over time (economic indicators), $x_{it}$ denotes a variable that varies over districts and over time (property characteristics), and $r_{it}$ denotes the risk data (same for each type $k$) given by the (distorted) short- and long-run earthquake probabilities. The reference dummies are the city dummy for Tokyo and the quarter dummy for Q4; these are set to zero.

In order to obtain a (balanced) panel we average over $h$, and obtain

$$\bar{y}_{it}^{(k)} = \alpha_0^{(k)} + \alpha_{c(i)} + \gamma_{q(t)} + x_{i\cdot}^{(k)\prime}\beta_1 + x_{\cdot t}^{(k)\prime}\beta_2 + \bar{x}_{it}^{(k)\prime}\beta_3 + r_{it}(\psi)'\beta_4 + \bar{u}_{it}^{(k)}, \tag{4.4}$$

where we average over $H_{it}^{(k)}$ items, which thus depends on how many properties of type $k$ there are in a given district.

Next we combine the three types of property into one $3 \times 1$ vector:

$$\bar{y}_{it} = \alpha_0 + (\alpha_{c(i)} + \gamma_{q(t)})\,\iota + X_{i\cdot}^{*}\beta_1 + X_{\cdot t}^{*}\beta_2 + X_{it}^{*}\beta_3 + \iota r_{it}(\psi)'\beta_4 + \bar{u}_{it}, \tag{4.5}$$

where $\iota = (1,1,1)'$, which we write more succinctly as

$$\bar{y}_{it} = \bar{X}_{it}\beta + \bar{u}_{it} \qquad (i = 1,\ldots,N;\ t = 1,\ldots,T), \tag{4.6}$$

where $\bar{y}_{it}$ is a $p \times 1$ vector of random observations, explained by (non-random) regressors $\bar{X}_{it} = \bar{X}_{it}(\psi)$, an unknown parameter vector $\beta$, and random errors $\bar{u}_{it}$ ($p \times 1$). In our case $p = 3$.

The errors are assumed to follow a $p$-variate three-error components structure,

$$\bar{u}_{it} = \zeta_i + \eta_t + \epsilon_{it}, \tag{4.7}$$

a sum of three independent components each of which is iid with zero means and variances

$$\mathrm{var}(\zeta_i) = \Sigma_\zeta, \qquad \mathrm{var}(\eta_t) = \Sigma_\eta, \qquad \mathrm{var}(\epsilon_{it}) = \Sigma_\epsilon, \tag{4.8}$$

where $\Sigma_\zeta$ and $\Sigma_\eta$ are positive semidefinite, and $\Sigma_\epsilon$ is positive definite, all of order $p \times p$.

Although the model appears to be linear in the parameters this is not completely the case, because the risk variable $r_{it}$ is a non-linear function of one or more $\psi$'s which appear in the probability weighting function $w(p)$. The estimation procedure taking this issue into account has been discussed in Chapter 2.

## 4.5 Estimation results

Our primary interest is in earthquake risk and its impact on property prices. More specifically, we wish to answer three questions:

(1) Do objective long-run earthquake probabilities have an effect on property prices?

(2) If so, do objective short-run earthquake probabilities have an effect on property prices, in addition to the effect of long-run probabilities?

(3) And do potentially distorted short-run earthquake probabilities have an effect on property prices, in addition to the effect of long-run probabilities?

Table 4.2: Results under various risk assumptions

| | variable | LR only | LR and objective SR | Base model |
|---|---|---|---|---|
| | land & building | 3.7592 | 4.5593 | 4.3812 |
| *intercepts* | land only | 3.5949 | 4.3940 | 4.2155 |
| | condo | 3.1025 | 3.9024 | 3.7244 |
| | Osaka | −0.2273 | −0.2625 | −0.2615 |
| *city dummies* | Nagoya | −0.3801 | −0.4100 | −0.4139 |
| | Fukuoka | −0.8770 | −0.9133 | −0.9108 |
| | Sapporo | −1.2050 | −1.2458 | −1.2388 |
| | immigrants | 6.7245 | 6.7224 | 6.7218 |
| *ward* | crime | −0.0437 | −0.0436 | −0.0436 |
| *attractiveness* | unemployment | −4.3360 | −4.3395 | −4.3399 |
| | executives | 3.3426 | 3.3447 | 3.3464 |
| *economic* | log(GDP) | 0.5606 | 0.5220 | 0.5229 |
| *indicators* | log(CPI) | 1.5347 | 1.4687 | 1.5030 |
| | area ($m^2$) | 0.0025 | 0.0025 | 0.0025 |
| | floor area ($m^2$) | 0.0006 | 0.0006 | 0.0006 |
| | distance to nearest station | −0.0145 | −0.0145 | −0.0145 |
| | age | −0.0121 | −0.0121 | −0.0121 |
| | built 1981–2000 | 0.1674 | 0.1658 | 0.1652 |
| *property* | built after 2000 | 0.4136 | 0.4126 | 0.4123 |
| *characteristics* | structure: reinf. concrete | 0.4348 | 0.4344 | 0.4343 |
| | structure: steel | 0.1867 | 0.1867 | 0.1867 |
| | structure: wood | −0.1264 | −0.1266 | −0.1266 |
| | urban control | −0.8972 | −0.8967 | −0.8967 |
| | max building coverage ratio | −0.0019 | −0.0019 | −0.0019 |
| | max floor area ratio | 0.0004 | 0.0004 | 0.0004 |
| | long run 45–55 | −0.1433 | −0.1427 | −0.1427 |
| *risk* | long run 55+ | −0.5037 | −0.5039 | −0.5041 |
| | short run | — | $-0.0915^{\ddagger}$ | −0.0514 |
| | $\hat{\psi}$ | — | — | $3.74^{\dagger}$ |
| $\Delta \log L$ | | −68.5 | −15.8 | — |

Before we answer these questions and comment on our estimates in Table 4.2, we explain our econometric modelling strategy. This strategy is based on two ingredients. First, we aim for parsimony. We want the smallest model that captures the essence of our story. This means that sometimes regressors have been deleted from our model even when the associated parameters are 'significant'. Significance does not imply importance, and importance is what interests us. Second, we make a distinction between focus and auxiliary regressors. The focus regressors are the effects that we are interested in or are part

of the minimum set that would make up a credible model, while the auxiliary regressors are only in the model because they improve the estimation of the focus parameters.

Since we have many observations, most estimates are likely to be significant at the usual 1.96 level. We provide more information about the results by strengthening the significance requirement on the $t$-values. Thus, a $^{\ddagger}$ will indicate that $|t| \leq 1.96$, which we interpret as not significant, while $^{\dagger}$ indicates significance with $1.96 < |t| \leq 4.00$. Estimates without superscript are therefore significant with $|t| > 4.0$. The choice of 4.0 is somewhat arbitrary and chosen *a posteriori* in order to provide more information about the precision of our estimates, in particular our parameter estimates pertaining to the risk variables. (All $t$-values test the null hypothesis that the parameter of interest equals zero, except the $t$-value of $\hat{\psi}$ which tests the null that $\psi = 1$.)

Now consider the first question. The results are presented in Table 4.2 under the heading 'LR only' and we see that all estimates are significant, that is, their $t$-value (in absolute terms) exceeds 4.0. Regarding the long-run risk effects, we remark that *long run 45–55* (medium risk) indicates the JSHIS probability that an earthquake occurs in the next thirty years of higher intensity than 5-lower and lower intensity than 6-lower; and that *long run 55+* (high risk) indicates the JSHIS probability that in the next thirty years an earthquake occurs of intensity 6-lower or higher. Both medium risk and high risk appear to have a significant negative impact on property prices. The higher risk level has a more severe impact, which is intuitively reasonable. Hence, long-run risk matters. This answers the first question.

Next, we consider the second question: given that long-run risk plays a role, do objective short-run probabilities also have an effect on property prices? The results are presented in the next column of Table 4.2 under the heading 'LR and objective SR'. Apparently they don't: the effect of the risk variable *short run*, while negative as we would expect, is not significantly different from zero.

Finally, we consider the third question: given that long-run risk plays a role, do potentially *distorted* short-run probabilities also have an effect on property prices? The results are displayed in the final column of Table 4.2 under the heading 'Base model'. Apparently they do: after distortion, short-run probabilities have a significant negative effect on property prices.

The difference between distorted and objective short-run risk is that short-run probabilities are now allowed to be distorted using a probability weighting function, in this case the one-parameter weighting function (4.2) proposed by Prelec (1998), which yields the highest likelihood. The parameter $\psi$ in the Prelec function is estimated to be 3.74 and is significantly different from unity, since the absolute value of its $t$-value lies between 1.96 and 4.00; in fact $|t| = 2.91$.

As shown in Figure 4.3, the estimated probability weighting function has an S-shaped pattern where small probabilities are underweighted and large probabilities are overweighted, which is in contrast to the inverse S-shaped probability weighting function often found in experiments. This contrast may be explained by the fact that with a positive background intensity of earthquakes, temporary deviations of short-run earthquake probabilities caused by seismic excitation are not evaluated (and overweighted) with respect to a reference probability of zero but with respect to a positive reference probability level. This applies in particular to Tokyo where the 90-day probability of an earthquake exceeding magnitude threshold of 5.5 never drops below 35% in the period that we analyze.

In summary: long-run risk matters, objective short-run risk does not, but distorted short-run risk

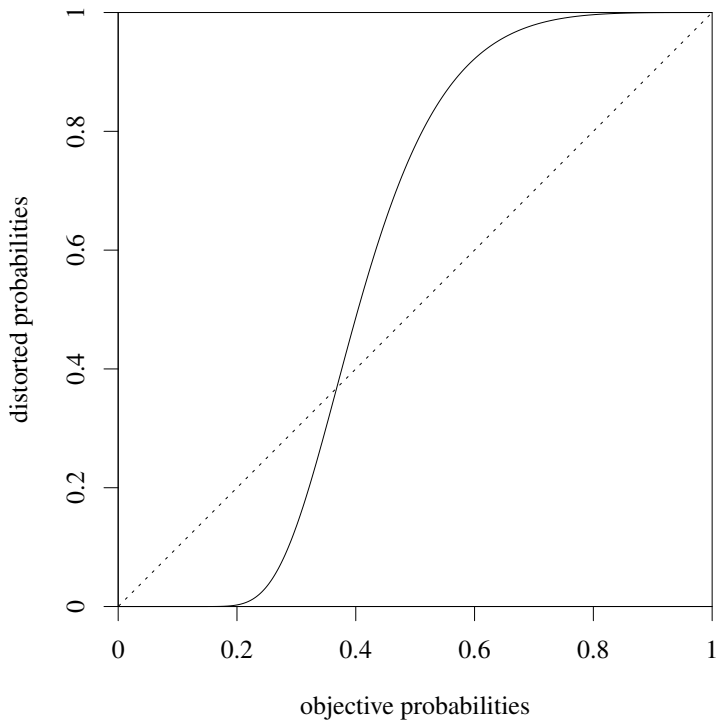



Figure 4.3: Estimated probability weighting of short-run probabilities, Prelec probability weighting function, $\hat{\psi} = 3.74$

does. In addition, all non-risk parameter estimates in the second and third columns are similar to the ones in the first column and all are significant (with a $t$-value larger than 4.0 in absolute value).

We briefly comment on these other (non-risk) parameters in the base model.

*Intercept and city dummies.* Tokyo, of course, is the most expensive city to buy property. If we set the property price level of Tokyo at 1.00, then the average property price levels of the other cities are 0.77 in Osaka, 0.66 in Nagoya, 0.40 in Fukuoka, and 0.29 in Sapporo. (Recall that we don't regress price but log-price on these dummies.) Also, if we set the price of land and building at 1.00, then the average price of the other types of property are 0.85 for land only and 0.52 for condos.

*Quarterly effects.* Estate agents sometimes tell customers that some months are better to buy or sell than others. Our results (in quarters, not months) are ambiguous, which is why we have omitted the quarter dummies from our regression. We return to this issue in our sensitivity analysis section.

*Ward attractiveness.* As discussed in Section 4.3, we selected eleven characteristics for each ward, divided into six categories. Only four of these eleven characteristics appear in our base model: percentage

of immigrants (representing population); number of criminal offenses (representing safety); and unemployment ratio and percentage of executives (representing employment). Executives make a ward more attractive, while crime and unemployment make it less attractive. Immigrants too make a ward more attractive, which makes sense if we realize that the word 'immigrant' refers to somebody moving into the ward from another municipality, usually within Japan. Hence, the more people move in from other areas in Japan, the more attractive the ward apparently is.

*Economic indicators.* Property prices are affected by general economic conditions, and two indicators appear in Table 4.2 and in our base model: $\log(\text{GDP})$ and $\log(\text{CPI})$, both of which have a positive effect on property prices. The inclusion of $\log(\text{CPI})$ has the additional advantage that if we wish to explain real (rather than nominal) property prices, then all results remain the same except that the effect of $\log(\text{CPI})$ is now 0.503 rather than 1.503. Hence, CPI has a positive effect not only on nominal but also on real property prices.

*Property characteristics.* A large (floor) area and proximity to the nearest station contribute positively to the price. New buildings are preferred to old ones, where we have included two dummies because major changes occurred in the regulations on earthquake-resistance standards in 1981 and 2000. As a result, buyers prefer a house built between 1981 and 2000 over a house built before 1981, and they like a house built after 2000 even better. Regarding the structure, wood is not desirable, steel is desirable, but reinforced concrete is preferred. Urban control signifies restrictions on development possibilities, and this has a negative effect on prices.

For all three property types the designated maximum building coverage ratio (BCR) and the maximum floor-area ratio (FAR) are provided. These ratios are legally allowed maxima, different for each piece of land. The BCR is the percentage of the building area to the site area; the FAR is the percentage of the total floor area to the site area. We use both ratios in our regression and find a negative effect of BCR and a positive effect of FAR. Shimizu and Nishimura (2006) and Nakagawa et al. (2009) use only FAR and find mixed effects and a positive effect, respectively. Hidano et al. (2015) use both ratios (as we do) and find a negative effect of BCR and a mixed effect of FAR.

*Error components.* We estimated the coefficients using the multivariate three-error components structure, as described in Section 4.4. It turns out that

$$\text{tr}(\Sigma_\epsilon) = 0.496 > \text{tr}(\Sigma_\zeta) = 0.129 \gg \text{tr}(\Sigma_\eta) = 0.002$$

and as a result we set $\Sigma_\eta = 0$, so that we end up with a two-error components structure. The effect of this is negligible and will be discussed further in our sensitivity analysis in Section 4.7.

## 4.6 Importance ordering and premia for earthquake risk

Next we wish to determine an ordering of importance of the explanatory variables, in particular the importance of the risk variables, and to calculate the premia for earthquake risk embedded in property

prices.

Table 4.3: Influences of each component for each type and city, real prices. Interquartile range between brackets.

| | intercept | ward(+) | ward(-) | macro | property(+) | property(-) | long-run risk | short-run risk |
|---|---|---|---|---|---|---|---|---|
| *Type* | | | | | | | | |
| land & building | 0.3254 | 0.0497 | -0.0285 | 0.6533 | 0.0600 | -0.0354 | -0.0187 | -0.0014 |
| | (0.0320) | (0.0162) | (0.0104) | (0.0436) | (0.0273) | (0.0224) | (0.0085) | (0.0025) |
| land only | 0.3152 | 0.0514 | -0.0284 | 0.6637 | 0.0360 | -0.0191 | -0.0180 | -0.0000 |
| | (0.0394) | (0.0159) | (0.0095) | (0.0404) | (0.0296) | (0.0087) | (0.0085) | (0.0024) |
| condo | 0.2936 | 0.0613 | -0.0302 | 0.6911 | 0.0403 | -0.0334 | -0.0196 | -0.0018 |
| | (0.0217) | (0.0226) | (0.0108) | (0.0490) | (0.0216) | (0.0219) | (0.0076) | (0.0031) |
| *City* | | | | | | | | |
| Tokyo | 0.3119 | 0.0582 | -0.0253 | 0.6598 | 0.0452 | -0.0285 | -0.0186 | -0.0026 |
| | (0.0316) | (0.0177) | (0.0079) | (0.0382) | (0.0270) | (0.0194) | (0.0076) | (0.0015) |
| Osaka | 0.3095 | 0.0457 | -0.0464 | 0.6899 | 0.0491 | -0.0323 | -0.0228 | -0.0000 |
| | (0.0354) | (0.0306) | (0.0102) | (0.0479) | (0.0302) | (0.0247) | (0.0050) | (0.0000) |
| Nagoya | 0.3035 | 0.0498 | -0.0269 | 0.6808 | 0.0518 | -0.0313 | -0.0280 | -0.0000 |
| | (0.0266) | (0.0152) | (0.0103) | (0.0437) | (0.0332) | (0.0208) | (0.0087) | (0.0000) |
| Fukuoka | 0.2519 | 0.0535 | -0.0324 | 0.7044 | 0.0487 | -0.0363 | -0.0061 | -0.0000 |
| | (0.0292) | (0.0243) | (0.0074) | (0.0590) | (0.0352) | (0.0240) | (0.0021) | (0.0000) |
| Sapporo | 0.2442 | 0.0534 | -0.0318 | 0.7145 | 0.0532 | -0.0362 | -0.0033 | -0.0000 |
| | (0.0308) | (0.0116) | (0.0040) | (0.0556) | (0.0375) | (0.0255) | (0.0032) | (0.0000) |

We write the prediction based on our original model (4.4) as

$$\hat{\hat{y}}_{it}^{(k)} = \hat{\alpha}_0^{(k)} + \hat{\alpha}_{c(i)} + \hat{\gamma}_{q(t)} + x_{i\cdot}^{(k)\prime}\hat{\beta}_1 + x_{\cdot t}^{(k)\prime}\hat{\beta}_2 + \bar{x}_{it}^{(k)\prime}\hat{\beta}_3 + r_{it}(\hat{\psi})'\hat{\beta}_4. \tag{4.9}$$

In order to determine an ordering of importance of the explanatory variables, we note that the size of an estimated parameter gives no indication of the size of its influence, because this influence depends also on how the associated regressor is measured. We write (4.9) symbolically as

$$\log(\text{price}) = \text{intercept} + W_+ - |W_-| + M + P_+ - |P_-| - |R_{lr}| - |R_{sr}|, \tag{4.10}$$

where the intercept comprises the (combined) constant term $\hat{\alpha}_0^{(k)} + \hat{\alpha}_{c(i)} + \hat{\gamma}_{q(t)}$ (positive); $W_+$ and $W_-$ contain the two positive and two negative ward regressors in $x_{i\cdot}^{(k)\prime}\hat{\beta}_1$; $M$ contains the two macro regressors in $x_{\cdot t}^{(k)\prime}\hat{\beta}_2$ (both positive); $P_+$ and $P_-$ contain the seven positive and five negative property regressors in $\bar{x}_{it}^{(k)\prime}\hat{\beta}_3$, respectively; and $R_{lr}$ and $R_{sr}$ contain the long-run and short-run risk regressors in $r_{it}(\hat{\psi})'\hat{\beta}_4$ (all negative).

Some categories (the ward characteristics $W$ and the property characteristics $P$) contain both positive and negative influences. Simple addition would then be misleading since two opposite forces would hide possibly important influences. Hence we calculate the influences by first defining

$$A = \text{intercept} + W_+ + |W_-| + M + P_+ + |P_-| + |R_{lr}| + |R_{sr}|, \tag{4.11}$$

where all items are positive (by construction). Influences can then be decomposed into contributions from various categories by using $A$ as the common denominator, that is, by computing $\text{intercept}/A$, $(W_+ + |W_-|)/A$, *et cetera*.

Table 4.3 presents the median of the relative influences for each component, by type and by city,

using log *real* property prices as the dependent variable. Macroeconomic indicatorsare very important, contributing around 67%. The intercepts for type and city are also important, contributing around 31%. Location matters as well, as the two subsets of ward attractiveness regressors take up around 5% and -3% of the influence, while the two sets of individual property characteristics add up to another 5% and -3%. This leaves around -2% for long-run and (distorted) short-run risk. The influence of long-run risk is almost the same for all property types, but it differs substantially among different cities. Fukuoka and Sapporo, where earthquakes are relatively rare, are not much influenced by long-run risk,while Nagoya is the most influenced. Regarding short-run risk, only Tokyo is influenced and the importance of short-run risk in Tokyo is about one-seventh of the influence of long-run risk. The joint influence of long-run and distorted short-run earthquake risk, on average -2.0% of log property prices, translates in monetary terms into a marginal effect of around -7 million Japanese yen per property, slightly more than the average annual income of a middle-income Japanese household in the period 2006/Q2 to 2015/Q3 that we analyze (Source: e-Stat Portal Site of Official Statistics Japan).

While the macro variables are by far the most relevant in explaining *median* house prices, they may be less relevant in explaining the *dispersion around* the median. To consider this aspect, Table 4.3 also displays the interquartile ranges (in brackets) of the relative influences. They reveal that the macro variables are still important, but all other variables (including the risk variables) are also quite important. More specifically, we see that individual property characteristics and intercepts for type and city are relevant in explaining dispersion in property prices (2.6%, 1.8% and 3.1% on average, respectively), still surpassed by macroeconomic variables (4.4%), and quite closely followed by the two sets of ward characteristics (1.8% and 1.0%) and risk variables (1.1%). Remarkably, the risk variables thus almost stand on equal footing with ward characteristics in explaining dispersion in property prices.

We can also compute these influences per quarter, in particular the quarter after the Tohoku earthquake (2011/Q2). The median influences of each component are essentially the same in that quarter with the exception of short-run risk in Tokyo, which is -0.26% overall but -0.40% in 2011/Q2. Large earthquakes have an important short-run effect in Tokyo. The influence of long-run risk remains the same.

We now investigate the influence of long-run and short-run risk in more detail, by decomposing the premia for earthquake risk. More precisely, we calculate and compare the predictions from four models. In model $\mathscr{M}_0$ there are no risk variables, either long-run or short-run; in model $\mathscr{M}_1$ we only have the two (objective) long-run risk variables; in model $\mathscr{M}_2$ we have long-run plus objective short-run risk variables; and in model $\mathscr{M}_3$ we have long-run plus distorted short-run risk variables (our base model).

Table 4.4 contains the results of this experiment. Let us denote the median of the log-price predictions in the four models by $m_0$, $m_1$, $m_2$, and $m_3$, respectively. Then the column $m_1 - m_0$ contains the premium of including (objective) long-run risk compared to not including any risk variable; the column $m_2 - m_1$ contains the premium of including objective short-run risk (in addition to long-run risk) compared to not including short-run risk; and the column $m_3 - m_2$ contains the premium of including distorted short-run risk (in addition to long-run risk) compared to including objective short-run risk.

We see that there is not much difference between different types of property and that the premium for long-run risk (compared to no risk) is much larger than the additional premium for short-run risk. Tokyo, Osaka, and Nagoya have a substantial premium for (objective) long-run risk of about 24–34%, while in

Table 4.4: Decomposition of the premia for earthquake risk per type and city

| type | city | median log-price | median premium | | |
|---|---|---|---|---|---|
| | | | $m_1 - m_0$ | $m_2 - m_1$ | $m_3 - m_2$ |
| *land & building* | Tokyo | 17.7275 | −0.2620 | −0.0246 | −0.0092 |
| | Osaka | 17.2495 | −0.2783 | −0.0049 | 0.0049 |
| | Nagoya | 17.4264 | −0.3393 | −0.0076 | 0.0076 |
| | Fukuoka | 17.2812 | −0.0630 | −0.0043 | 0.0043 |
| | Sapporo | 17.0736 | −0.0558 | −0.0016 | 0.0016 |
| *land only* | Tokyo | 17.7073 | −0.2409 | −0.0241 | −0.0087 |
| | Osaka | 17.2167 | −0.2691 | −0.0048 | 0.0048 |
| | Nagoya | 17.1113 | −0.3293 | −0.0077 | 0.0077 |
| | Fukuoka | 16.9066 | −0.0658 | −0.0046 | 0.0046 |
| | Sapporo | 16.3805 | −0.0517 | −0.0016 | 0.0016 |
| *condo* | Tokyo | 17.0344 | −0.2621 | −0.0246 | −0.0093 |
| | Osaka | 16.5881 | −0.2677 | −0.0051 | 0.0051 |
| | Nagoya | 16.5236 | −0.3175 | −0.0079 | 0.0079 |
| | Fukuoka | 16.2134 | −0.0740 | −0.0042 | 0.0042 |
| | Sapporo | 16.2134 | −0.0457 | −0.0015 | 0.0015 |

Fukuoka and Sapporo this premium is 5–7%, thus much smaller. This is consistent with their different long-run risk profile. All long-run premia are negative, which means that long-run risk is compensated for through an adjustment in property prices in all cities.

Regarding short-run risk, there is a big difference between Tokyo and the other cities. In Tokyo, property prices are compensated for objective short-run risk with a median premium of about 2.5%, and there is an additional median compensation for distorted short-run risk of about 1%, because people tend to overweight large short-run earthquake probabilities in the Tokyo property market. In the quarter after the Tohoku earthquake these median premia rise to 3.0% and 1.7%, respectively.

Outside Tokyo we see that $(m_3 - m_2) \approx -(m_2 - m_1)$, which implies that the overall effect $(m_3 - m_1)$ is almost zero. This is caused by the shape of the estimated probability weighting function. The short-run probabilities outside Tokyo are relatively small, and after probability weighting they become even smaller (bottom part of the S-curve). People thus underweight small short-run probabilities; in fact they almost ignore them altogether. This effect (or lack of effect) can be decomposed into a 'compensation' $(m_2 - m_1 < 0)$ for objective short-run risk and a 'reward' $(m_3 - m_2 > 0)$ for underweighting short-run risk.

The power of econometrics is well-illustrated by the fact that, while property prices are the highest in Tokyo, the largest compensation (that is, reduction) for short-run risk (objective and distorted) and a sizeable compensation for long-run risk is in Tokyo.

## 4.7 Sensitivity analysis

Our base model depends on assumptions regarding which variables to include and which not, how to measure or group certain variables, the choice of functional forms, and the stochastic specification. We wish to show that our results are robust, and we shall do so by deviating from our base model in various directions. (Of course, the selected base model was, in fact, itself the result of extensive sensitivity analyses.) In each case we are interested to find out whether our focus parameters are affected by these deviations. We are less interested to find out whether the deviations themselves are 'significant' or

not, since these deviations typically represent auxiliary variables and are not the primary focus of our investigation.

Our focus variables are the risk variables and, in addition, four key characteristics of the property: area ($m^2$), floor area ($m^2$), distance to the nearest station, and age of the property. We have chosen the location (distance to nearest station) and the size (area and floor area) as our focus variables, and one characteristic of the property (age).

*Ward attractiveness.* Our base model contains four variables which measure the attractiveness of a ward. We extend this list by adding seven ward characteristics: the percentage of foreigners, and the number of hospitals, daycare centers, kindergartens, homes for the aged, department stores, and large retail stores.

Table 4.5: Sensitivity — ward attractiveness and economic indicators

| | Base | +Attr. | −GDP |
|---|---|---|---|
| area ($m^2$) | 0.0025 | 0.0025 | 0.0025 |
| floor area ($m^2$) | 0.0006 | 0.0006 | 0.0006 |
| distance to nearest station | −0.0145 | −0.0142 | −0.0145 |
| age | −0.0121 | −0.0121 | −0.0122 |
| long run 45–55 | −0.1427 | −0.1961 | −0.1411 |
| long run 55+ | −0.5041 | −0.5706 | −0.5024 |
| short run | −0.0514 | −0.0519 | −0.0839 |
| $\hat{\psi}$ | $3.74^{\dagger}$ | $3.75^{\dagger}$ | $2.63^{\dagger}$ |
| $\Delta \log L$ | — | 472.9 | −407.8 |

If we compare the column '+Attr.' with the base model ('Base') in Table 4.5 we see that very little changes, thus showing the robustness with regard to these ward characteristics. These additional ward characteristics are therefore omitted in view of parsimony and the fact that, while they may be significant, they are not important.

*Economic indicators.* In the same Table 4.5 we also experiment with deleting $\log(\text{GDP})$, so that the only economic indicator is $\log(\text{CPI})$. This has some (although not a large) effect in particular on short-run risk, so that we keep GDP in the model as a general plausible indicator of economic activity.

*Property characteristics.* Next we experiment with the property characteristics. We consider three deviations from the base model, reported in Table 4.6.

In the first column we remove the urban control variable; in the second column we remove the three building structure dummies; and in the third column we add, in addition to urban control, three further land-use variables ('residential', 'commercial', and 'industrial'), which describe the city's intentions of the usage of the land. Again, the estimated parameters appear to be robust to these changes; inclusion of urban control and, in particular, building structure dummies appears to substantially increase the loglikelihood, which makes sense because building a property costs more when steel is used instead of wood, and even more when reinforced concrete is used.

*Cities.* In our base model we have selected five Japanese cities. Although our selection is based on

Table 4.6: Sensitivity — property characteristics

| | Urban control | Build. Struct. | Land use |
|---|---|---|---|
| area ($m^2$) | 0.0025 | 0.0025 | 0.0025 |
| floor area ($m^2$) | 0.0006 | 0.0009 | 0.0006 |
| distance to nearest station | −0.0147 | −0.0159 | −0.0146 |
| age | −0.0121 | −0.0119 | −0.0121 |
| long run 45–55 | −0.1060 | −0.1685 | −0.1387 |
| long run 55+ | −0.4661 | −0.5263 | −0.4767 |
| short run | −0.0516 | −0.0508 | −0.0515 |
| $\hat{\psi}$ | $3.72^{\dagger}$ | $3.89^{\dagger}$ | $3.76^{\dagger}$ |
| $\Delta$ log L | −786.6 | −5824.4 | 33.9 |

careful considerations (geographical spread and risk variation, in particular) as discussed in Section 4.3, this is still somewhat arbitrary. Suppose we only had four cities. How would this affect our estimates? This is shown in Table 4.7. In the first column we delete Tokyo, in the second column we delete Osaka,

Table 4.7: Sensitivity — four cities

| | Tokyo | Osaka | Nagoya |
|---|---|---|---|
| area ($m^2$) | 0.0023 | 0.0024 | 0.0025 |
| floor area ($m^2$) | 0.0006 | 0.0006 | 0.0006 |
| distance to nearest station | −0.0152 | −0.0145 | −0.0147 |
| age | −0.0126 | −0.0127 | −0.0115 |
| long run 45–55 | −0.2427 | −0.1124 | −0.1571 |
| long run 55+ | −0.4302 | −0.4759 | −0.6160 |
| short run | $-0.1873^{\ddagger}$ | −0.0627 | −0.0525 |
| $\hat{\psi}$ | $1.9^{\ddagger}$ | $4.04^{\dagger}$ | $4.11^{\dagger}$ |

and in the third column we delete Nagoya. The effect on the non-risk parameters (area, distance, age) is small, but the effect on the risk parameters is not so small. Deleting Tokyo has quite a large effect on the risk parameters, because the short-run risk of Osaka, Nagoya, Fukuoka and Sapporo is relatively small compared to Tokyo, and estimation is less accurate when there is less variation in the risk variables. Deleting Osaka or Nagoya only affects the risk estimates marginally. Deleting Fukuoka or, in particular, Sapporo leads to unreliable results for the long-run risk parameters, probably caused by the fact that without these cities there is insufficient variation in the long-run risk variables leading to inaccurate estimation results. They are therefore omitted from the table. (Notice that we do not show the difference in loglikelihood in this table since the numbers of observations are different with different subsets of the sample.)

*Time dimension.* Our observations are per quarter and we could include quarter dummies to capture the idea that buying or selling in one quarter is more advantageous than in another.

Our base model does not include quarter dummies and in Table 4.8 we experiment with three possible extensions, namely adding three quarter dummies, adding one dummy for the fourth quarter (because there are relatively few earthquakes in the fourth quarter), and adding one dummy for the quarter following the Tohoku earthquake, respectively. In the cases Q123 and Q4 the likelihood increases substantially, but the key estimates don't change much, although the short-run risk parameters now become less signif-

Table 4.8: Sensitivity — quarters and Tohoku dummy

| | Base | Q123 | Q4 | Tohoku |
|---|---|---|---|---|
| area ($m^2$) | 0.0025 | 0.0025 | 0.0025 | 0.0025 |
| floor area ($m^2$) | 0.0006 | 0.0006 | 0.0006 | 0.0006 |
| distance to nearest station | −0.0145 | −0.0145 | −0.0145 | −0.0145 |
| age | −0.0121 | −0.0120 | −0.0120 | −0.0121 |
| long run 45–55 | −0.1427 | −0.1415 | −0.1406 | −0.1426 |
| long run 55+ | −0.5041 | −0.5033 | −0.5025 | −0.5040 |
| short run | −0.0514 | $-0.0162^{\dagger}$ | $-0.0208^{\dagger}$ | −0.0562 |
| $\hat{\psi}$ | $3.74^{\dagger}$ | $4.56^{\ddagger}$ | $3.89^{\ddagger}$ | $3.27^{\dagger}$ |
| $\Delta \log L$ | — | 1091.3 | 1007.8 | 6.3 |

icant. In the case of Tohoku even the likelihood does not increase much. Because the quarter dummies and the short-run risk are both time effects, which are likely to interact with each other, the results are ambiguous. This is why we prefer to exclude quarter dummies, thus making the interpretation easier and more transparent.

*Stochastics.* In our base model we have estimated two variance matrices:

$$\Sigma_\zeta = 0.129 \begin{pmatrix} 0.16 & 0.10 & -0.00 \\ 0.10 & 0.18 & -0.04 \\ -0.00 & -0.04 & 0.66 \end{pmatrix}, \quad \Sigma_\epsilon = 0.407 \begin{pmatrix} 0.31 & 0.01 & 0.00 \\ 0.01 & 0.33 & 0.00 \\ 0.00 & 0.00 & 0.36 \end{pmatrix},$$

while we set $\Sigma_\eta = 0$. This is because when we estimate the full three-error components model, we find

$$\Sigma_\zeta = 0.129 \begin{pmatrix} 0.16 & 0.11 & -0.00 \\ 0.11 & 0.18 & -0.04 \\ -0.00 & -0.04 & 0.66 \end{pmatrix}, \quad \Sigma_\epsilon = 0.406 \begin{pmatrix} 0.31 & 0.01 & 0.00 \\ 0.01 & 0.33 & 0.00 \\ 0.00 & 0.00 & 0.36 \end{pmatrix},$$

while

$$\Sigma_\eta = 0.002 \begin{pmatrix} 0.32 & 0.35 & 0.00 \\ 0.35 & 0.44 & -0.06 \\ 0.00 & -0.06 & 0.24 \end{pmatrix}.$$

The matrices $\Sigma_\zeta$ and $\Sigma_\epsilon$ are thus hardly affected and $\Sigma_\eta$ is about one hundred times smaller than the other two.

In Table 4.9, column 2 we see that the key parameters are also hardly affected, although the likelihood (with six additional parameters) increases substantially. A formal test (not trivial in this case) may indicate that the hypothesis $\Sigma_\eta = 0$ is rejected in favor of $\Sigma_\eta > 0$, but we opt — in line with current ideas about the theory of applied econometrics (Angrist and Pischke, 2009; Magnus, 2017) — for parsimony and importance rather than for significance.

*Station versus district.* We know a lot about each property from the data, but not its exact location. We know in which district the property lies and we also know the name of the nearest station. In our setup we use districts as our location reference and there are 3,710 districts in our data set. But we could also

Table 4.9: Sensitivity — stochasticity and station versus district

| | Base | 3-errors | station |
|---|---|---|---|
| area ($m^2$) | 0.0025 | 0.0025 | 0.0026 |
| floor area ($m^2$) | 0.0006 | 0.0006 | 0.0006 |
| distance to nearest station | −0.0145 | −0.0146 | −0.0137 |
| age | −0.0121 | −0.0121 | −0.0115 |
| long run 45–55 | −0.1427 | −0.1448 | −0.1378† |
| long run 55+ | −0.5041 | −0.5067 | −0.5742 |
| short run | −0.0514 | −0.0443 | −0.0548 |
| $\hat{\psi}$ | 3.74† | 3.52† | 3.56† |
| $\Delta \log L$ | — | 735.2 | |

use the nearest station as our location reference. There are 1,022 stations, so the district measure should be more precise. In fact, as Table 4.9, column 3 shows, the results are amazingly similar, demonstrating that the precise method of approximating the location is not so important.

*Probability weighting functions: an extension.* In our base model we use objective long-run probabilities and distorted short-run probabilities based on the Prelec probability weighting function. This raises various questions. First, one could argue that we should allow long-run probabilities to be distorted too; and second, we could experiment with different probability weighting functions.

Table 4.10: Sensitivity and extension — probability weighting functions

| | Base | dist. SR TK | dist. LR Prelec | dist. LR TK |
|---|---|---|---|---|
| area ($m^2$) | 0.0025 | 0.0025 | 0.0025 | 0.0025 |
| floor area ($m^2$) | 0.0006 | 0.0006 | 0.0006 | 0.0006 |
| distance to nearest station | −0.0145 | −0.0145 | −0.0143 | −0.0143 |
| age | −0.0121 | −0.0121 | −0.0121 | −0.0121 |
| long run 45–55 | −0.1427 | −0.1427 | −0.8644 | −0.4856 |
| long run 55+ | −0.5041 | −0.5039 | −1.3838 | −1.5028 |
| short run | −0.0514 | −0.0733‡ | −0.0517 | −0.0518 |
| $\hat{\psi}$ | 3.74† | 1.40‡ | 3.78† | 3.77† |
| $\hat{\gamma}$ | — | — | 0.17 | 0.32 |
| $\Delta \log L$ | — | −14.6 | 152.8 | 167.6 |

In Table 4.10 we experiment with an alternative functional form for the short-run risk variable, namely the weighting function (4.1) introduced by Tversky and Kahneman (1992). In particular, in column 2 (dist. SR, TK) we replace the Prelec function applied to the short-run earthquake probabilities with the Tversky-Kahneman probability weighting function. The estimation results are similar to the base model, but somewhat less precise, and the loglikelihood decreases. The Tversky-Kahneman probability weighting function is found to be S-shaped, just like the Prelec function, confirming the robustness of this finding.

Next we also allow long-run risk to be distorted using both the Prelec and the Tversky-Kahneman weighting functions. The model contains two related time-invariant long-run probabilities and we quite naturally assume that these two probabilities share the same weighting function with the same parameter $\gamma$. (In particular, *distorted long run 45–55* is computed as *distorted long run 45+* minus *distorted long*

*run 55+*, consistent with Choquet integration.) In columns 3 and 4 of Table 4.10 we allow both long-run risk and short-run risk to be distorted.

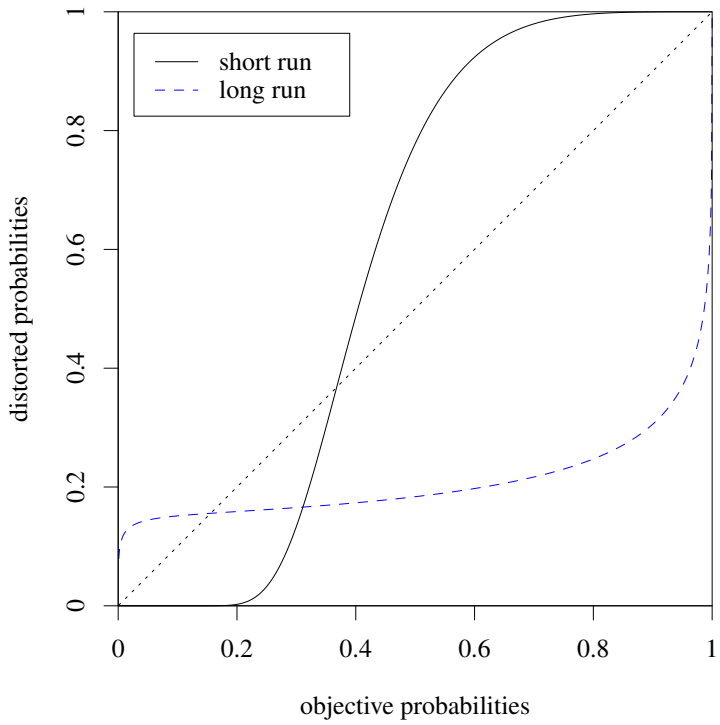



Figure 4.4: Implied probability weighting functions of long-run and short-run earthquake risk

The model with the higher likelihood is the one with an inverse S-shaped Tversky-Kahneman weighting function for long-run risk and an S-shaped Prelec weighting function for short-run risk, as shown in Figure 4.4. We note that the Prelec function for long-run risk, although yielding a lower log-likelihood than the Tversky-Kahneman weighting function, is also found to be inverse S-shaped, which is reassuring for the robustness of our results. Thus, in an extension of our base model that allows for distortion of time-invariant long-run earthquake probabilities we find evidence of a conventional inverse S-shaped probability weighting function. This means that when purchasing property in Japan, people tend to overweight small long-run probabilities and underweight large long-run probabilities.

Summarizing, we have conducted extensive sensitivity analyses on our base model, always moving *one* step away from our base model. The base model proved to be remarkably robust in most directions. In some cases, however, one could argue that the base model should have been adjusted. The reason why we have not done so and prefer the current base model is twofold. First, we aim for parsimony; we prefer a simpler model over a more complex model. Second, if we were to change our base model, we would need to do (and we have done) the sensitivity analysis again for all cases, now based on the new base model. Then there will be other directions that prove to be sensitive. It is unlikely that there exists

a model that is insensitive in every direction.

## 4.8 Conclusion

We have studied the impact of earthquake risk on Japanese property prices using a rich panel data set. We have not only allowed for time-invariant long-run earthquake probabilities to impact property prices, but we have also analyzed the impact of short-run earthquake probabilities generated from a seismic excitation model.

We have shown that long-run earthquake probabilities negatively impact property prices and increasingly so at higher risk levels. We have also shown that short-run earthquake probabilities have a negative impact on property prices, and that this effect becomes statistically significant only after we allow for probability weighting.

The probability weighting function associated with short-run earthquake probabilities is found to be S-shaped. That stands in contrast to the familiar inverse S-shaped probability weighting functions predominantly found in experiments. The shape we find may be explained by the fact that in our setting there is a non-negligible positive background arrival rate of earthquakes. People may therefore tend to evaluate earthquake probabilities, and overweight their temporary deviations under seismic excitation, not with respect to zero but with respect to a positive reference probability level. This remarkable finding calls for the development of reference-dependent models for *probabilities* to augment the large literature on reference-dependent models for changes in *wealth* levels.

# Chapter 5

# Computational properties of the WALS estimator

## 5.1 Introduction

The t-statistic is commonly used by econometricians to determine the statistical significance of a parameter estimate in linear regression models. More specifically, its use can be seen as two-fold. First, it is used for hypothesis testing. Under certain assumptions the t-statistic follows a Student's t-distribution, thus by checking its value against the critical values under a certain significance level, the null hypothesis that the estimated coefficient of a regressor of interest is zero can be tested. Second, it is used as a diagnostic, where the econometrician is not sure about whether to include a regressor in the model and chooses to include it only when the t-statistic is above a certain threshold or when including the regressor would improve the estimator of other regressor(s) in the model.

In the second approach, the t-statistic is used as a criterion for model selection. The process of using the t-statistic as a criterion for model selection is called pretesting. The resulting estimator, the pretest estimator, is defined as

$$b_i = w\hat{\beta}_{iu} + (1-w)\hat{\beta}_{ir}, \tag{5.1}$$

where

$$w = \begin{cases} 1 & \text{if } |t_j| > c, \\ 0 & \text{if } |t_j| \leq c, \end{cases} \tag{5.2}$$

and $\hat{\beta}_{iu}$ and $\hat{\beta}_{ir}$ are estimators from the unrestricted and restricted model, while $t_j$ is the t-statistic corresponding to $\beta_j$. In the restricted model, $\beta_j = 0$. The value of $c$ is usually chosen to be 1.96 if the significance level is 5%. As has been discussed in, among others, Magnus and Durbin (1999) and Danilov and Magnus (2004), the pretest estimator is non-differentiable, inadmissible, and is not robust to small changes in the level of $c$ or small perturbations in the data.

Pretesting is commonly employed by applied econometricians but often without proper scrutiny into the implications of using the same dataset for selecting a model and estimating the parameters in the selected model. Danilov and Magnus (2004) shows that model selection and estimation should not be viewed as two separate steps but should be seen as one integrated procedure, since ignoring the

uncertainty in the model selection would lead to misleadingly precise estimates.

The pretest estimator is a simple special case of model-averaging estimators, which combine model selection and estimation. The literature on model averaging diverges into two major schools: Frequentist Model Averaging (FMA) and Bayesian Model Averaging (BMA). In this chapter we focus on the weighted-average least squares (WALS) estimator, which is a Bayesian combination of frequentist estimators developed in Magnus et al. (2010). A Bayesian flavor is introduced in the estimator because the weights and conditional estimates of each model are determined by the input data and the choice of priors. As shown by Magnus et al. (2010), the WALS estimator has a number of advantages over the BMA estimator from both the theoretical and practical standpoints.

Magnus and De Luca (2016) presents a comprehensive review on the WALS method from its inception until 2014. In this survey paper, the theory related to WALS was summarized and a consistent framework was introduced. After the publishing of Magnus and De Luca (2016), a number of extensions and improvements to the WALS estimator have been proposed. De Luca et al. (2018) extends the WALS estimator from Gaussian linear models to generalized linear models. De Luca et al. (2021b) derives estimators of the finite-sample bias and variance of WALS. In particular, two plug-in estimators of the bias and variance of the posterior mean are proposed and analyzed, namely the frequentist Maximum Likelihood (ML) estimator and the Bayesian Double Shrinkage (DS) estimator. These estimators can be embedded in the WALS estimation procedure to obtain bias-corrected WALS estimation results. De Luca et al. (2021c) proposes a simulation-based approach for estimating WALS confidence and prediction intervals, and compares the performance of WALS with several competing estimators in a Monte Carlo simulation study. The competing estimators include the unrestricted and restricted least-squares estimator, two post-selection estimators based on the Akaike and Bayesian information criteria, various frequentist model averaging estimators, and a popular shrinkage estimator (the adaptive LASSO). The paper found that the bias-corrected WALS estimator leads to better confidence and prediction intervals.

Empirical applications of the WALS estimator have been studied in various papers. Liski et al. (2010) compares WALS with alternative model selection methods in a application to hip fracture treatment costs. Poghosyan and Magnus (2012) uses WALS to estimate and forecast Armenian real GDP growth and inflation. Seya et al. (2012) employs WALS in spatial hedonic land price models. Xu (2014) applies the WALS estimator to investigate the robustness of the cross-country relationship between anti-self-dealing rules and proxies for stock market development. Clarke (2017) extends the WALS framework from OLS to two stage least squares (2SLS) and applies WALS to two examples, one on the returns to schooling and the other on the effect of religion in explaining differences in cross-country economic growth. Tumala et al. (2018) uses BMA and WALS to investigate predictors of inflation in Nigeria. Afonso and Jalles (2019) analyzes the determinants of bond spreads considering non-conventional monetary policy using BMA and WALS. Furceri and Ostry (2019) applies WALS to determine a robust set of determinants of income inequality. Comunale and Mongelli (2020) uses WALS to select variables when investigating which variables have consistently supported growth in euro area countries in the past thirty years. Rahman et al. (2020) and Rahman and Shang (2020) apply WALS on a Pakistanian ensemble multi-satellite precipitation dataset and evaluates the performance of the WALS estimator against alternative estimators. Mignamissi and Kuete (2020) uses both BMA and WALS to analyze the key determinants of African well-being under model uncertainty. Furceri et al. (2021) uses BMA and WALS

to analyze robust determinants of initial output losses from the Covid-19 pandemic. Aller et al. (2021) investigates robust determinants of CO2 emissions using BMA, Cluster-LASSO and WALS.

The rest of this chapter is structured as follows. Section 5.2 lays out the theoretical framework of WALS and introduces relevant mathematical notations. Section 5.3 introduces the three packages that implement the WALS procedures, and presents a comparison between the available options of the three packages so that the user can choose the package most suitable for her purposes. Section 5.4 shows an example of the output from all three packages using a small example dataset under a common model specification. Section 5.5 compares the theoretical properties and performance of different choices in the prior distributions and prior parameters, so that the user can make an informed decision about which prior to choose apart from the default options. Section 5.6 compares the efficiency of two alternative integration routines, Gauss-Laguerre Quadrature and Adaptive Quadrature. This section also shows the effects of a cut-off point where the program switches between two alternative methods in calculating posterior moments. Section 5.7 demonstrates the simulated bias and variances under different prior distribution and parameters, the values of which come from a large number of replications of Monte Carlo simulations and are pre-stored so that bias-corrected WALS estimators can be obtained in the programs without having to simulate the biases and variances on the go. We compute what is the minimum number of Monte Carlo replication needed in order to achieve a certain degree of accuracy in the final estimated biases. Section 5.8 analyzes the relationship between the distribution of WALS estimates and confidence intervals when the number of Monte Carlo simulations used in the estimation of confidence intervals changes. Section 5.9 analyzes the performance of the programs under two extreme cases with simulated data: the case when the number of auxiliary regressors $k_2$ becomes large, and the case when the input data is nearly singular (correlation between regressors is high). Section 5.10 compares the three packages in terms of computation time. Finally, 5.11 concludes.

## 5.2 The WALS framework

We adopt the framework of a linear regression model

$$y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon \tag{5.3}$$

where $y(n \times 1)$ is the endogenous variable, $X_1(n \times k_1)$ are the focus regressors that the econometrician is certain to include in the model, $X_2(n \times k_2)$ are the auxiliary regressors that may or may not be included in the model. We assume $k_1 \geq 1, k_2 \geq 0, k_1 + k_2 \leq n$, and that the disturbance $\epsilon$ has mean zero and a diagonal positive definite variance matrix.

The model-averaging estimators take the form

$$\hat{\beta}_1 = \sum_{j=1}^{2^{k_2}} \lambda_j \hat{\beta}_{1j}, \qquad \hat{\beta}_2 = \sum_{j=1}^{2^{k_2}} \lambda_j \hat{\beta}_{2j}, \tag{5.4}$$

where the $\lambda_j$ are the model weights.

The WALS estimator proposed by Magnus et al. (2010) reduces the dimension of the model averaging estimator from order $2^{k_2}$ to $k_2$ by exploiting a semi-orthogonal transformation of the auxiliary

regressors.

The steps of the WALS estimator is briefly explained as follows:

Step 1: Determine the $k_1$ focus regressors ($X_1$) and $k_2$ auxiliary regressors ($X_2$). We require $0 < k_1, k_2 < n$. When $k_2 = 0$, the model reduces to unrestricted OLS. Furthermore the matrix $X = [X_1, X_2]$ is required to have full column rank.

Step 2: Scale the input data by diagonal matrices $\Delta_1$ ($k_1 \times k_1$) and $\Delta_2$ ($k_2 \times k_2$) to ensure that all diagonal elements of the scaled data is equal to 1. Equivariance is imposed for numerical stability.

Step 3: Semi-orthogonalization transformation. This step reduces the dimension of the problem from $2^{k_2}$ to $k_2$. After the transformation, the original regressors $X_1, X_2$ are transformed into $Z_1, Z_2$ such that $Z_1\gamma_1 = X_1\beta_1$, $Z_2\gamma_2 = X_2\beta_2$, and the data generating process (DGP) can be rewritten as

$$y = Z_1\gamma_1 + Z_2\gamma_2 + \epsilon.$$

Step 4: Solving the unrestricted and restricted models. $\gamma_1$, $\gamma_2$ and $\sigma^2$ can be estimated using OLS.

Step 5: Compute the mean and variance in posterior distribution. Under a given prior distribution (we assume Weibull, Subbotin, or the Laplace prior), the mean $m = (m_1, ..., m_{k_1})$ and variance $V = diag(v_1, ..., v_{k_2})$ of the posterior distribution is computed for each of the $k_2$ components of $x$. In this step numerical integration is required for Weibull and Subbotin priors, while theoretical derivations exist for the Laplace prior.

Step 6: Calculate the bias and variance of the posterior mean by using plug-in estimators. The bias and variances of the posterior mean $\eta$ is simulated and pre-stored for each prior and each value of $\eta$ between 0 to 30, with a step size of 0.01. For values within this range, the bias and variances can be calculated through linear interpolation of the stored values. For values outside of this range, asymptotic approximation is used.

Step 7: Calculate the WALS point estimates by transforming the estimated parameters $\hat{\gamma}_1$ and $\hat{\gamma}_2$ to $\hat{\beta}_1$ and $\hat{\beta}_2$.

Step 8: Calculate the bias and variance of the WALS estimates using the plug-in estimators from Step 6.

Step 9: Estimate confidence intervals of the WALS estimator. The confidence intervals are simulation-based where we draw from the estimated sampling distribution of the bias-corrected WALS estimators of $\beta_1$ and $\beta_2$. The draws are stored for calculating prediction intervals when WALS prediction is required.

Step 10: Estimate the MSE, skewness and kurtosis of the WALS estimator.

Embedded in the theory of WALS is the seemingly trivial Normal location problem, where we need to estimate one parameter $\eta$ given one observation $x$, generated by the Normal distribution $N(\eta, 1)$:

$$x|\eta \sim N(\eta, 1).$$

Combined with a prior $\pi(\eta)$, the posterior density $p(\eta|x)$ can be written as

$$p(\eta|x) = \frac{\phi(x-\eta)\pi(\eta)}{\int_{-\infty}^{\infty}\phi(x-\eta)\pi(\eta)d\eta}$$

where $\phi$ is the standard normal density function. We denote the mean and variance of $\eta$ in the posterior distribution as $m(x)$ and $v(x)$.

## 5.3 Packages

We present three statistical packages that implement the WALS procedures for linear regression models (assuming normality and homoskedasticity in the disturbances). In this section we discuss also the input/output options for the user. We describe which choices the user can make and how to implement these choices in the different packages.

### 5.3.1 `Stata`

The WALS routine in `Stata` is called `wals`[1].

The syntax of the package is:

**Usage**

```
wals depvar foc([varlist]), aux([varlist]) [options]
```

where the options are explained as follows.

**Arguments**

| | |
|---|---|
| depvar | Name of the dependent variables. |
| `foc(varlist)` | A list of dim_X1 focus regressors. If `nocons` is not specified in the options, a constant term is added to the (beginning of) focus regressors, and the number of focus regressors is $k_1 = dim\_X1 + 1$. If `nocons` is specified in the options, $k_1 = dim\_X1$. |
| `aux(varlist)` | A list of dim_X2 auxiliary regressors. If `auxcons` is specified in the options, a constant term is added to the auxiliary regressors (but not to the focus regressors), and the number of auxiliary regressors becomes $k_2 = dim\_X2 + 1$. Otherwise, $k_2 = dim\_X2$. |
| `nocons` | If this is specified, do not add a constant to the focus regressors. |
| `auxcons` | If this is specified, do not add a constant term to the focus regressors but add a constant to the auxiliary regressors. |
| `consname(...)` | Custom name of the constant term, for example can be "_cons" or "constant". |
| `prior(...)` | The chosen prior distribution, can be one of `prior(weibull)`, `prior(subbotin)`, or `prior(laplace)`. Default value is "weibull". |

[1]The `Stata` package `wals` is kindly provided by Dr. Giuseppe De Luca and Prof. Jan Magnus. An earlier version of this package has been published in De Luca and Magnus (2011).

`priorpar(...)` Choice of the parameter $q$ for Weibull and Subbotin priors. This is ignored for Laplace priors. if option "priorpar(.5)" is specified, use $q = 0.5$; otherwise use the minimax regret parameter.

`quadp(...)` The number of quadrature points to use in the Gauss-Laguerre quadrature method. Default value is `quadp(1000)`. This is ignored for Laplace priors since closed-form solution is available and numerical integration is not necessary for Laplace priors.

`estmom(...)` The type of plug-in estimation to use, can be `estmom(ds)` (double shrinkage) or `estmom(ml)` (maximum likelihood). Default option is "ds".

`cint(...)` The type of the confidence interval, can be `cint(mc)` or `cint(naive)`. Default option is "mc".

`level(...)` An integer (or a sequence of integers) between 0 to 100, the significance level of the confidence intervals of the parameter estimates. Default option is `level(90)`.

`mcreps(...)` An integer, the number of repetitions to use for Monte Carlo draws used in the calculation of bias-corrected posterior moments and construction of confidence intervals. Default option is `mcreps(5000)`.

`mcseed(...)` An integer, the random seed to use for Monte Carlo draws. Default option is `mcseed(1)`.

`mcsaving(...)` Specify the name of the data to store the Monte Carlo simulated draws of the bias correction. `mcsave(...)` must be specified if prediction intervals may be needed for WALS prediction.

`sigma(...)` If not specified, the standard deviation of the error term is inferred from the input data. Otherwise, the user may specify a positive real number as the presumed standard deviation of the error term.

## Examples

```
* WALS regression of y on focus variables X1 (adding the constant term) and auxiliary variables X2,
* under the Laplace prior; the name of the constant term is "_cons"
wals `y', foc(`X1') aux(`X2') prior(laplace) consname(_cons)

* all auxiliary variables except the constant term
wals `y', foc() aux(`X1' `X2')

* focus variables X1 and auxiliary variables X2 (adding the constant term)
wals `y', foc(`X1') aux(`X2') auxcons

* focus variables X1 (without constant term) and auxiliary variables X2
wals `y', foc(`X1') aux(`X2') nocons

* subbotin prior with 500 quadrature points and maximum likelihood estimators of the sampling moments
```

```
wals `y', foc(`X1') aux(`X2') prior(subbotin) quadp(500) estmom(ml)

* multiple levels for confidence intervals
wals `y', foc(`X1') aux(`X2') prior(laplace) lev(90 95 99) mcseed(1234) estmom(ml) mcreps(1000000)

* naive and simulation-based approaches with ds and ml plug-in estimators of the bias
wals `y', foc(`X1') aux(`X2') prior(laplace) level(90) cint(naive)
wals `y', foc(`X1') aux(`X2') prior(laplace) level(90) cint(naive) estmom(ml)


* save the dataset with draws of the bias-corrected estimates
wals `y', foc(`X1') aux(`X2') lev(90) mcseed(1234) estmom(ml) mcsav(WALS_bc_draws)


* WALS (in-sample) predictions: weibull prior
wals `y', foc(`X1') aux(`X2') mcsav(WALS_bc_draws_W) mcseed(1234)
predict wals_lp_W, xb
```

### 5.3.2 R

The WALS routine in R is collected in an R package called walsR.

---

wals *The main WALS estimation procedure.*

---

**Usage**

```
wals(
y,
X_focus = NULL,
X_aux = NULL,
no_cons = FALSE,
aux_cons = FALSE,
prior = "weibull",
quad_pts = 1000,
plugin_type = "ds",
conf_int_type = "mc",
conf_levels = 95,
mc_reps = 5000,
mc_seed = 1,
mc_save = FALSE,
sigma = NULL,
```

```
choice_q = "minimax"
)
```

## Arguments

| | |
|---|---|
| y | An $n * 1$ vector, matrix or data frame containing the dependent variables. |
| X_focus | An $n * dim\_X1$ vector, matrix or data frame containing the focus regressors. If no_cons == FALSE, a constant term is added to the focus regressors, and the number of focus regressors is $k1 = dim_X1+1$. If no_cons == TRUE, $k1 = dim_X1$. Default value of X_focus is NULL, which cannot occur at the same time as the option no_cons=TRUE since the minimum number of focus regressors is 1. |
| X_aux | An $n * dim\_X2$ matrix or data frame containing the auxiliary regressors. If aux_cons == TRUE, a constant term is added to the (end of) auxiliary regressors, and the number of auxiliary regressors is $k2 = dim\_X2 + 1$. If aux_cons == FALSE, $k2 = dim\_X2$. Default value of X_aux is NULL, which cannot occur at the same time as the option aux_cons=FALSE since the minimum number of auxiliary regressors is 1. |
| no_cons | A boolean. Whether to add a constant to the focus regressors and include in the model. Default value is FALSE. |
| aux_cons | A boolean. Whether to add a constant to the auxiliary regressors and include in the model. Default value is FALSE. The options no_cons = FALSE and aux_cons = TRUE cannot occur at the same time since only one constant term is allowed in the model. |
| prior | The chosen prior distribution, can be one of "weibull", "subbotin", "laplace". Default value is "weibull". |
| quad_pts | The number of quadrature points to use in the Gauss-Laguerre quadrature method. Default value is 1000. This is ignored for Laplace priors since closed-form solution is available and numerical integration is not necessary for Laplace priors. |
| plugin_type | The type of plug-in estimation to use, can be "ds" (double shrinkage) or "ml" (maximum likelihood). Default value is "ds". |
| conf_int_type | The type of the confidence interval, can be "mc" or "naive". Default value is "mc". |
| conf_levels | An integer (or a sequence of integers) between 0 to 100, the significance level of the confidence intervals of the parameter estimates. Default value is 95. |
| mc_reps | An integer, the number of repetitions to use for Monte Carlo draws used in the calculation of bias-corrected posterior moments and construction of confidence intervals. Default value is 5000. |
| mc_seed | An integer, the random seed to use for Monte Carlo draws. Default value is 1. |

| | |
|---|---|
| mc_save | A boolean. If TRUE, save the Monte Carlo simulated draws of the bias correction. mc_save = TRUE is necessary if prediction intervals are needed for WALS prediction. Default value is FALSE. |
| sigma | If NULL, the standard deviation of the error term is inferred from the input data. Otherwise, the user may specify a positive real number as the presumed standard deviation of the error term. Default value is NULL. |
| choice_q | A string, choice of the parameter $q$ for Weibull and Subbotin priors. Can be "minimax" or "0.5". This is ignored for Laplace priors. Default value is "minimax". |

## Value

`wals(...)` returns an object of class `wals`. The function `summary_wals` can be used to obtain and print a summary table of the results. The function `predict_wals` can be used to perform in-sample or out-of-sample prediction based on a `wals` object. An object of class `wals` is a list containing at least the following components:

| | |
|---|---|
| beta_hat | Coefficient estimates. |
| beta_bias | The estimated bias of the coefficient estimates. |
| beta_var | The variance covariance matrix of the coefficient estimates. |
| beta_MSE | The MSE of the model. |
| beta_RMSE | The RMSE of the model. |
| beta_VARRMSE | The variance to MSE ratio of the model. |
| std_error | The standard error of the coefficient estimates. |
| t | The t-statistic of the coefficient estimates. |
| beta_CI | The confidence intervals of the coefficient estimates. |
| skewnewss | The skewness of the (bias-corrected) WALS estimator. |
| kurtosis | The kurtosis of the (bias-corrected) WALS estimator. |
| kappa | The condition number of the data. |
| sigma_hat | The estimated sigma (if not provided ex ante) or user-specified sigma of the model. |

Other information such as the prior parameters, arguments passed to the function call, and the input data are also stored in this object. To obtain a complete list of the object (for example if the `wals` object is stored in `res`), run `attributes(res)$names`.

## Examples

```
res <- wals(growth_data["growth"],
growth_data[c("gdp60", "equipinv", "school60", "life60", "dpop")],
growth_data[c("law", "tropics",  "avelf","confuc")])
```

```
res <- wals(growth_data["growth"],
growth_data[c("gdp60", "equipinv")],
growth_data[c("law", "tropics",  "avelf","confuc")],
no_cons=TRUE, aux_cons=TRUE)

res <- wals(growth_data["growth"],
growth_data[c("gdp60", "equipinv")],
growth_data[c("law", "tropics",  "avelf","confuc")],
prior="laplace")
```

---

`predict_wals` *The WALS prediction procedure.*

---

## Usage

```
predict_wals(
object,
PI_level = 95,
out_of_sample = FALSE,
X_focus = NULL,
X_aux = NULL
)
```

## Arguments

| | |
|---|---|
| `object` | An object of class `wals`, obtained from running the WALS estimation using the `wals(...)` function |
| `PI_level` | An integer or a sequence of integers between 0 to 100, the confidence level(s) of the prediction. Default value is 95, corresponding to [2.5%, 97.5%] prediction intervals. |
| `out_of_sample` | A boolean indicating whether the prediction is performed for the same dataset used for estimation. If FALSE, no additional input data needs to be provided. Default value is FALSE. If TRUE, additional data `X_focus` and `X_aux` should be provided. |
| `X_focus` | Additional input data for the focus regressors, only needed if out_of_sample == TRUE. This should be provided in the same order as the focus regressors used for estimation. If a constant term is added during estimation, only the regressors excluding the constant term need to be provided. Default value is NULL. |

X_aux — Additional input data for the auxiliary regressors, only needed if out_of_sample == TRUE. This should be provided in the same order as the auxiliary regressors used for estimation. If a constant term is added during estimation, only the regressors excluding the constant term need to be provided. Default value is NULL.

### Value

Note that to enable the calculation of prediction intervals, the Monte Carlo draws simulated in the estimation procedure must be stored by setting the option mc_save=TRUE. This function returns a list containing the following items:

y_pred — Predicted values of y.

pi_y_pred — Prediction intervals corresponding to the levels specified in PI_level.

### Examples

```
res <- wals(growth_data["growth"],
growth_data[c("gdp60", "equipinv", "school60", "life60", "dpop")],
growth_data[c("law", "tropics", "avelf","confuc")], mc_save=TRUE)
## in sample prediction
predict_wals(res)
predict_wals(res, c(90, 95))

## out of sample prediction
predict_wals(res, out_of_sample=TRUE,
X_focus=growth_data_pred[c("gdp60", "equipinv", "school60", "life60", "dpop")],
X_aux=growth_data_pred[c("law", "tropics", "avelf","confuc")])
```

---

## summary_wals — *The summary function which provides post-estimation analysis.*

---

### Usage

```
summary_wals(object, save_df = FALSE, digits = 6)
```

### Arguments

object — An object of class wals, obtained from running the WALS estimation using the wals(...) function

save_df — A boolean. Whether to save the summary as data frames. Default value is FALSE, in which case the summary is displayed on screen.

digits — An integer, the number of digits to be used for displaying the data frame of results.

**Value**

If save_df == TRUE, return a list containing the following items:

| | |
|---|---|
| params_df | A data frame with the chosen options: `no_cons`, `aux_cons`, `prior`, `quad_pts`, `plugin_type`, `conf_int_type`, `mc_reps` and the parameter values: `n`, `k1`, `k2`, `a`, `b`, `c`, `sigma`. |
| res_df | A data frame with the estimation results: variable names, coefficients, bias, standard errors, RMSE, t-statistics, and the confidence intervals of the coefficient estimates. |

**Examples**

```
res <- wals(growth_data["growth"],
growth_data[c("gdp60", "equipinv", "school60", "life60", "dpop")],
growth_data[c("law", "tropics", "avelf","confuc")], mc_save=TRUE)

## print summary to screen
summary_wals(res, digits=4)

## alternatively, save summary to data frame
summary <- summary_wals(res, save_df=TRUE)
res_df <- summary$res_df
```

### 5.3.3 Python

The WALS package in Python is called WALS and is written within the framework of the popular statistical package statsmodels.

The WALS estimator is defined in a class called WALS in similar fashion as the popular statsmodels.OLS class, where class methods like .fit(), .summary(), .predict() provides an easy interface to inspect and use the model.

**Usage**

```
model = WALS(endog, exog_focus, exog_auxiliary,
no_cons=False, aux_cons=False,
prior='weibull', choice_q='minimax')

res = model.fit(plugin_type='ds',
conf_int_type='mc', conf_levels=95,
quad_pts=1000,
mc_reps =5000,
mc_seed=1,
mc_save=False,
```

```
sigma=None)

res.predict(PI_level = 95,
exog_focus = None,
exog_auxiliary = None)

res.summary()
```

**Arguments**

| | |
|---|---|
| endog | An $n * 1$ matrix or data frame containing the dependent variables. |
| exog_focus | An $n * dim\_X1$ matrix or data frame containing the focus regressors. If no_cons == False, a constant term is added to the focus regressors, and the number of focus regressors is $k_1 = dim\_X1 + 1$. If no_cons == True, $k_1 = dim\_X1$. Default value of X_focus is None, which cannot occur at the same time as the option no_cons=True since the minimum number of focus regressors is 1. |
| exog_aux | An $n * dim\_X2$ matrix or data frame containing the auxiliary regressors. If aux_cons == True, a constant term is added to the auxiliary regressors, and the number of auxiliary regressors is $k_2 = dim\_X2 + 1$. If aux_cons == False, $k_2 = dim\_X2$. Default value of X_aux is None, which cannot occur at the same time as the option aux_cons=False since the minimum number of auxiliary regressors is 1. |
| no_cons | A boolean. Whether to add a constant to the focus regressors and include in the model. Default value is False. |
| aux_cons | A boolean. Whether to add a constant to the auxiliary regressors and include in the model. Default value is False. The options no_cons = False and aux_cons = True cannot occur at the same time since only one constant term is allowed in the model. |
| prior | The chosen prior distribution, can be one of "weibull", "subbotin", "laplace". Default value is "weibull". |
| choice_q | A string, choice of the parameter $q$ for Weibull and Subbotin priors. Can be "minimax" or "0.5". This is ignored for Laplace priors. Default value is "minimax". |
| plugin_type | The type of plug-in estimation to use, can be "ds" (double shrinkage) or "ml" (maximum likelihood). Default value is "ds". |
| conf_int_type | The type of the confidence interval, can be "mc" or "naive". Default value is "mc". |
| quad_pts | The number of quadrature points to use in the Gauss-Laguerre quadrature method. Default value is 1000. This is ignored for Laplace priors since closed-form solution is available and numerical integration is not necessary for Laplace priors. |
| conf_levels | An integer (or a sequence of integers) between 0 to 100, the significance level of the confidence intervals of the parameter estimates. Default value is 95. |

| | |
|---|---|
| mc_reps | An integer, the number of repetitions to use for Monte Carlo draws used in the calculation of bias-corrected posterior moments and construction of confidence intervals. Default value is 5000. |
| mc_seed | An integer, the random seed to use for Monte Carlo draws. Default value is 1. |
| mc_save | A boolean. If TRUE, save the Monte Carlo simulated draws of the bias correction. mc_save = TRUE is necessary if prediction intervals are needed for WALS prediction. Default value is FALSE. |
| sigma | If sigma==None, the standard deviation of the error term is inferred from the input data. Otherwise, the user may specify a positive real number as the presumed standard deviation of the error term. Default value is None. |

**Value**

WALS returns an object of class WALS. Applying the .fit() method on a WALS object returns an object of class WALSResults. The functions summary can be used on WALSResults object to obtain and print a summary table of the results. An object of class WALSResults is a list containing at least the following components:

| | |
|---|---|
| params | Coefficient estimates. |
| bias | The estimated bias of the coefficient estimates. |
| variance | The variance covariance matrix of the coefficient estimates. |
| mse | The MSE of the model. |
| rmse | The RMSE of the model. |
| varmse | The variance to MSE ratio of the model. |
| std_error | The standard error of the coefficient estimates. |
| t | The t-statistic of the coefficient estimates. |
| ci | The confidence intervals of the coefficient estimates. |
| skew | The skewness of the (bias-corrected) WALS estimator. |
| kurt | The kurtosis of the (bias-corrected) WALS estimator. |
| condnum | The condition number of the data. |
| s | The estimated sigma (if not provided ex ante) or user-specified sigma of the model. |

## 5.4 Example

In this section we demonstrate an example of the WALS estimation, prediction, and post-estimation analyses using the example datasets provided together with the packages.

We provide two example datasets, growth_data and growth_data_pred. Both are small datasets as analyzed in Magnus et al. (2010). growth_data contains 72 observations and 11 columns, while growth_data_pred contains 2 observations and 11 columns, which can be used for the illustration of out-of-sample prediction.

## Examples

In Stata:

```
clear all
version 11.1
set mem 500m
set more off
set linesize 255
set seed 123456789

cd "path_to_working_directory"
adopath ++  "path_to_working_directory"

import delimited using "growth_data_pred.csv",  delimiters(",") clear asdouble
drop growth
saveold "growth_data_pred", replace

import delimited using "growth_data.csv",  delimiters(",") clear asdouble
local y "growth"
local X1 "gdp60 equipinv school60 life60 dpop"
local X2 "law tropics avelf confuc"
noi sum `y' `X1' `X2'
saveold "growth_data", replace

* WALS estimation

wals `y', foc(`X1') aux(`X2') mcseed(1234)


* WALS (in-sample) predictions & 95% prediction intervals
wals `y', foc(`X1') aux(`X2') mcsav(WALS_bc_draws_W) mcseed(1234)
predict wals_lp, xb pi(wals_lp_low wals_lp_upp)
noi sum wals_lp*
drop wals_lp*



* WALS (out-of-sample) predictions & 95% prediction intervals
use "growth_data", clear
wals `y', foc(`X1') aux(`X2') mcsav(WALS_bc_draws_W) mcseed(1234)
append using "growth_data_pred.dta"
```

```
predict wals_lp if e(sample)!=1, xb pi(wals_lp_low wals_lp_upp)
format %7.3f `X1' `X2' wals_lp*
format %7.5f wals_lp*
list `X1' `X2' wals_lp* if e(sample)!=1, noobs abbr(20)
drop wals_lp*
```

In R:

```
growth_data <- read.csv(file="data/growth_data.csv")
growth_data_pred <- read.csv(file="data/growth_data_pred.csv")

res <- wals(growth_data["growth"],
growth_data[c("gdp60", "equipinv", "school60", "life60", "dpop")],
growth_data[c("law", "tropics", "avelf","confuc")], mc_save=TRUE)

# post estimation summary
summary_wals(res, digits=4)

# in-sample prediction
pred <- predict_wals(res)
y_pred <- pred$y_pred
pi_y_pred <- pred$pi_y_pred

# out-of-sample prediction
X_focus_oos <- growth_data_pred[c("gdp60", "equipinv", "school60", "life60", "dpop")]
X_aux_oos <- growth_data_pred[c("law", "tropics", "avelf","confuc")]
pred_oos <- predict_wals(res, out_of_sample=TRUE,
X_focus=X_focus_oos,
X_aux=X_aux_oos)
y_pred_oos <- pred_oos$y_pred
pi_y_pred_oos <- pred_oos$pi_y_pred
cbind(X_focus_oos, X_aux_oos, y_pred_oos, pi_y_pred_oos)
```

In Python:

```
from wals import WALS
import pandas as pd
```

```
data = pd.read_csv('data/growth_data.csv', engine='python')
y = data.growth
X_focus = data[["gdp60", "equipinv", "school60", "life60", "dpop"]]
X_aux = data[["law", "tropics",  "avelf","confuc"]]
wals_model = WALS(y, X_focus, X_aux)
res = wals_model.fit(mc_save=True)
res.summary()
# in-sample prediction
y_pred, pi_y_pred = res.predict()

# out-of-sample prediction
data_oos = pd.read_csv('data/growth_data_pred.csv')
X_focus_oos = data_oos[["gdp60", "equipinv", "school60", "life60", "dpop"]]
X_aux_oos = data_oos[["law", "tropics",  "avelf","confuc"]]
y_pred_oos, pi_y_pred_oos = res.predict(exog_focus=X_focus_oos, exog_auxiliary=X_aux_oos)

pred_oos_summary = pd.DataFrame(data=np.hstack((res.exog_pred, y_pred_oos, pi_y_pred_oos)),
columns = res.model.coef_names + ['wals_pred'] + pi_y_pred_oos.columns.tolist())
pred_oos_summary
```

The results are as follows.

Table 5.1: Estimation results for the growth data example using `Stata`

```
. wals `y', foc(`X1') aux(`X2') mcseed(1234)

WALS estimates - weibull prior                  Number of obs =        72
                                                k1            =         6
                                                k2            =         4
                                                a             =    0.1124
                                                b             =    0.6931
                                                c             =    0.8876
                                                quad. pts.    =      1000
                                                sigma         =    0.0109
```

| growth | PM Coef. | DS Bias | DS Std.Err. | DS RMSE | t | MC-DS [95% Conf. | Interval] |
|---|---|---|---|---|---|---|---|
| constant | .0547959 | -.0018708 | .0224677 | .0225455 | 2.44 | .0109039 | .1007654 |
| gdp60 | -.0152994 | .0001484 | .0033239 | .0033272 | -4.60 | -.0221466 | -.008729 |
| equipinv | .1615355 | .0194095 | .0548098 | .058145 | 2.95 | .035573 | .2548099 |
| school60 | .0175747 | .0000819 | .009804 | .0098044 | 1.79 | -.002079 | .0372147 |
| life60 | .0008947 | .0000372 | .0003552 | .0003571 | 2.52 | .0001633 | .001564 |
| dpop | .2853201 | -.0813453 | .2474629 | .2604898 | 1.15 | -.1571793 | .8649583 |
| law | .0129196 | -.0034302 | .0060667 | .0069693 | 2.13 | .0022953 | .0293774 |
| tropics | -.0054075 | .0016561 | .0032177 | .0036189 | -1.68 | -.0147355 | .0007958 |
| avelf | -.0055352 | .0019424 | .0045763 | .0049715 | -1.21 | -.0190627 | .0038515 |
| confuc | .0478828 | -.0080158 | .0166339 | .0184646 | 2.88 | .0208375 | .0880513 |

Table 5.2: Out-of-sample prediction results for the growth data example using `Stata`

| gdp60 | equipinv | school60 | life60 | dpop | law | tropics | avelf | confuc | wals_lp | wals_lp_low | wals_lp_upp |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 8.960 | 0.095 | 1.000 | 70.700 | 0.016 | 1.000 | 0.386 | 0.113 | 0.000 | 0.02880 | 0.02168 | 0.03567 |
| 7.119 | 0.028 | 0.320 | 41.000 | 0.023 | 0.667 | 1.000 | 0.803 | 0.000 | -0.00182 | -0.01095 | 0.00672 |

Table 5.3: Estimation results for the growth data example using R

```
WALS estimates - weibull prior
--------------------------------------------------------------------------------
Input:

       no_cons =   FALSE                 Number of obs =     72
      aux_cons =   FALSE                           k1 =      6
         prior = weibull                           k2 =      4
      quad_pts =    1000                          a = 0.1124
   plugin_type =      ds                          b = 0.6931
 conf_int_type =      mc                          c = 0.8876
       mc_reps =    5000                      sigma = 0.0109

--------------------------------------------------------------------------------

Cofficients:
   (focus)     Coef.    Bias Std.Err   RMSE      t    2.5%   97.5%
    constant  0.0548 -0.0019  0.0225 0.0225  2.439  0.0100  0.1010
       gdp60 -0.0153  0.0001  0.0033 0.0033 -4.603 -0.0219 -0.0087
    equipinv  0.1615  0.0194  0.0548 0.0581  2.947  0.0325  0.2544
    school60  0.0176  0.0001  0.0098 0.0098  1.793 -0.0016  0.0370
      life60  0.0009  0.0000  0.0004 0.0004  2.519  0.0002  0.0016
        dpop  0.2853 -0.0813  0.2475 0.2605  1.153 -0.1660  0.8718

--------------------------------------------------------------------------------
 (auxiliary)   Coef.    Bias Std.Err   RMSE      t    2.5%  97.5%
         law  0.0129 -0.0034  0.0061 0.0070  2.130  0.0014 0.0297
     tropics -0.0054  0.0017  0.0032 0.0036 -1.681 -0.0152 0.0007
       avelf -0.0055  0.0019  0.0046 0.0050 -1.210 -0.0187 0.0034
      confuc  0.0479 -0.0080  0.0166 0.0185  2.879  0.0219 0.0910

--------------------------------------------------------------------------------
```

Table 5.4: Out-of-sample prediction for the growth data example using R

| gdp60 | equipinv | school60 | life60 | dpop | law | tropics | avelf | confuc | y_pred_oos | 2.5% | 97.5% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 8.959569 | 0.0950 | 1.00 | 70.7 | 0.01645300 | 1.0000000 | 0.3857 | 0.1127971 | 0 | 0.028800518 | 0.02189328 | 0.035657617 |
| 7.118826 | 0.0281 | 0.32 | 41.0 | 0.02344454 | 0.6666667 | 1.0000 | 0.8027273 | 0 | -0.001820124 | -0.01084107 | 0.006889407 |

Table 5.5: Estimation results for the growth data example using `Python`

WALS estimates - weibull prior

| | | | |
|---|---|---|---|
| **no_cons** | False | **Number of obs.** | 72 |
| **aux_cons** | False | **k1** | 6 |
| **prior** | weibull | **k2** | 4 |
| **quad_pts** | 1000 | **a** | 0.1124 |
| **plugin_type** | ds | **b** | 0.6931 |
| **conf_int_type** | mc | **c** | 0.8876 |
| **mc_reps** | 5000 | **d** | 1.0000 |
| **condition number** | 1.76 | **sigma** | 0.0109 |

WALS parameter estimates

| | Coef. | Bias | Std. Err. | RMSE | t. stat. | 2.5% | 97.5% |
|---|---|---|---|---|---|---|---|
| **const** | 0.0548 | -0.0019 | 0.0225 | 0.0225 | 2.4389 | 0.0073 | 0.1061 |
| **gdp60** | -0.0153 | 0.0001 | 0.0033 | 0.0033 | -4.6028 | -0.0219 | -0.0089 |
| **equipinv** | 0.1615 | 0.0194 | 0.0548 | 0.0581 | 2.9472 | 0.0380 | 0.2529 |
| **school60** | 0.0176 | 8.188e-05 | 0.0098 | 0.0098 | 1.7926 | -0.0014 | 0.0359 |
| **life60** | 0.0009 | 3.719e-05 | 0.0004 | 0.0004 | 2.5189 | 0.0002 | 0.0016 |
| **dpop** | 0.2853 | -0.0813 | 0.2475 | 0.2605 | 1.1530 | -0.1163 | 0.8072 |
| **law** | 0.0129 | -0.0034 | 0.0061 | 0.0070 | 2.1296 | 0.0042 | 0.0275 |
| **tropics** | -0.0054 | 0.0017 | 0.0032 | 0.0036 | -1.6805 | -0.0145 | 0.0003 |
| **avelf** | -0.0055 | 0.0019 | 0.0046 | 0.0050 | -1.2095 | -0.0191 | 0.0035 |
| **confuc** | 0.0479 | -0.0080 | 0.0166 | 0.0185 | 2.8786 | 0.0194 | 0.0891 |

Table 5.6: Out-of-sample prediction for the growth data example using `Python`

| | const | gdp60 | equipinv | school60 | life60 | dpop | law | tropics | avelf | confuc | wals_pred | 2.5% | 97.5% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.0 | 8.959569 | 0.0950 | 1.00 | 70.699997 | 0.016453 | 1.000000 | 0.3857 | 0.112797 | 0.0 | 0.028801 | 0.021529 | 0.03572 |
| 1 | 1.0 | 7.118826 | 0.0281 | 0.32 | 41.000000 | 0.023445 | 0.666667 | 1.0000 | 0.802727 | 0.0 | -0.001820 | -0.015020 | 0.01106 |

We show the differences in the parameter estimates and standard errors in the following table.

Table 5.7: Comparison of estimation results for the growth data example using `Stata`, `R`, and `Python`

| var | coef. (`Stata`) | diff. in coef. (R-`Stata`) | diff. in coef. (`Python`-`Stata`) | std. error (`Stata`) | diff. in s.e. (R-`Stata`) | diff. in s.e. (`Python`-`Stata`) |
|---|---|---|---|---|---|---|
| 0.054796 | 0.054796 | 0.000000 | 0.000004 | 0.022468 | 0.000000 | 0.000032 |
| -0.015299 | -0.015299 | 0.000000 | -0.000001 | 0.003324 | 0.000000 | -0.000024 |
| 0.161536 | 0.161536 | 0.000000 | -0.000036 | 0.054810 | 0.000000 | -0.000010 |
| 0.017575 | 0.017575 | 0.000000 | 0.000025 | 0.009804 | 0.000000 | -0.000004 |
| 0.000895 | 0.000895 | 0.000000 | 0.000005 | 0.000355 | 0.000000 | 0.000045 |
| 0.285320 | 0.285320 | 0.000000 | -0.000020 | 0.247463 | 0.000000 | 0.000037 |
| 0.012920 | 0.012920 | 0.000000 | -0.000020 | 0.006067 | 0.000000 | 0.000033 |
| -0.005408 | -0.005408 | 0.000000 | 0.000008 | 0.003218 | 0.000000 | -0.000018 |
| -0.005535 | -0.005535 | 0.000000 | 0.000035 | 0.004576 | 0.000000 | 0.000024 |
| 0.047883 | 0.047883 | 0.000000 | 0.000017 | 0.016634 | 0.000000 | -0.000034 |

## 5.5 Choice of prior

We analyze three priors: Weibull, Subbotin, and Laplace. All three are of the reflected generalized Gamma family, and have the general form

$$\pi(\theta) = \frac{qc^{\delta}}{2\Gamma(\delta)}|\theta|^{-\alpha}e^{-c|\theta|^{q}}, \qquad \delta = \frac{1-\alpha}{q}. \tag{5.5}$$

The Weibull, Subbotin, and Laplace priors are special cases of the general form, where $\alpha + q = 1$ for Weibull, $\alpha = 0$ for Subbotin, and $\alpha = 0$ and $q = 1$ for Laplace. Closely related to the choice of prior is the choice of the parameter $q$, which is relevant for Weibull and Subbotin priors. When $q$ is known for a given prior, the entire prior distribution is also known.

One natural question is how to choose the prior and its parameter $q$ when performing WALS estimation. In our WALS estimation packages, we provide 5 options regarding this choice: Weibull prior with $q = 0.887630085544086$ (minimax regret parameter); Weibull prior with $q = 0.5$; Subbotin prior with $q = 0.799512530172489$ (minimax regret parameter); Subbotin prior with $q = 0.5$; and the Laplace prior. The minimax regret solutions, as their names suggest, minimize the maximum regret over all possible values of $x$, where regret is defined as the risk subtracted by its theoretical lower bound $\eta^2/(1+\eta^2)$. Using a simulated dataset, we compare the bias and RMSE of the estimated coefficient of the focus regressor under different choices of prior.

### 5.5.1 Estimated Bias and RMSE under different choices of prior: an example

The simulation setting is as follows: there are 2 focus regressors (first of which is a constant term) and 8 auxiliary regressors. The model coefficient of the second focus regressor $\beta_2$ is of interest. All regressors (apart from the constant term) are multivariate normally distributed with mean 0, variance $\sigma_x^2 = 0.7$, and pairwise correlation $\rho = 0.7$. The true DGP is $y = \beta_1 + \beta_2 X_1 + \beta_3 X_3 ... + \beta_{10} X_{10} + \epsilon$ where $\epsilon$ is i.i.d. standard normally distributed. $\beta_1$ and $\beta_2$ are 1 while all other $\beta$'s are equal to $\xi = 0.5$. We generate $n = 2000$ observations for each simulation run, and perform 500 replications of each simulation. We use the box plot to illustrate the distribution of the WALS estimator of the focus parameter $\beta_2$. It can

be seen in Figure 5.1 that the mean of the $\hat{\beta}_2$ is closer to its true DGP under the Weibull and Subbotin with $q = 0.5$, and the bias is slightly larger under the Weibull and Subbotin when $q$ is equal to the minimax regret value. The bias under the Laplace prior is the largest. Similar observations is found when inspecting the RMSE in Figure 5.2.

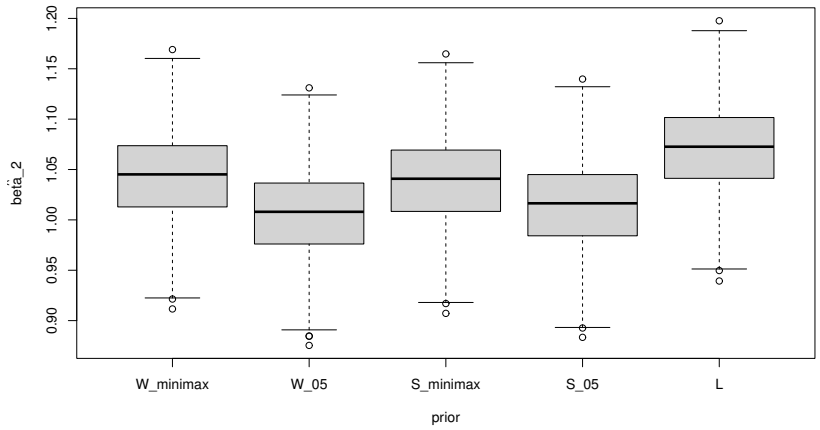



Figure 5.1: WALS estimator of the coefficient of the focus regressor from different priors

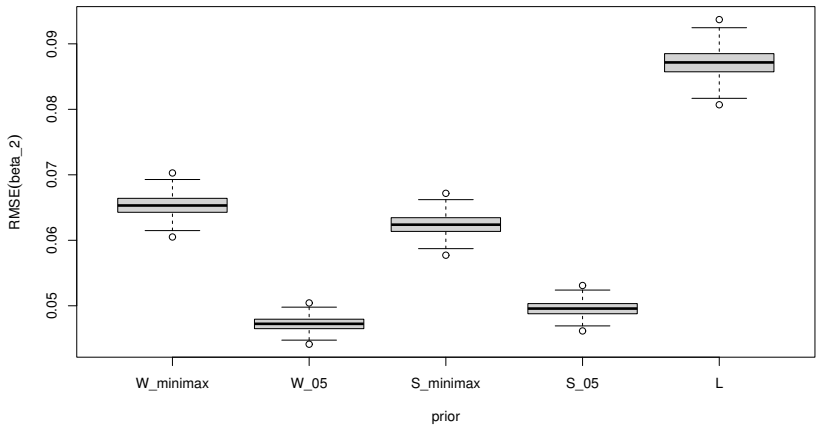



Figure 5.2: RMSE from different priors

### 5.5.2 What is the optimal prior?

The simulated example above seems to indicate that Weibull and Subbotin priors with $q = 0.5$ are more preferable than other choices of $q$, if the bias and RMSE are the focus metric for assessing the choice. Can this conclusion be generalized? This question has been partially addressed in De Luca et al. (2018). What we found in the empirical study seems to indicate that the Laplace prior, although having nice theoretical properties and much faster to compute, generally has a higher bias and RMSE then the other two choices of prior. The performance of Weibull and Subbotin are very similar in terms of bias, RMSE

and computing time.

In asymptotic studies we have found that the choice of $q$ with Subbotin and Weibull prior is not trivial since the speed of convergence does not increase fast enough as $q$ approaches the minimax regret value. The fact that the choice of $q = 0.5$ seems to perform better than the minimax regret parameter value is interesting since the minimax regret parameters are theoretically derived optimal choices, since they minimize the maximum regret for all choices of $x$.

Under the framework of the normal location problem we can analyze the sampling properties of the posterior mean under a wider choice of $q$. In Figures 5.3, 5.4 and 5.5 we show the sampling bias and variances of the posterior mean corresponding to each different $\eta$ between 0 to 30 (with a step size of 0.01). The simulation is based on 1,000,000 replications and the procedure of the Monte Carlo tabulations has been described in detail in De Luca et al. (2021b). Based on Figures 5.3, 5.4, it seems the minimax regret choice of $q$ has lower bias (compared with $q = 0.5$) for smaller values of $\eta$, and higher bias (compared with $q = 0.5$) for larger values of $\eta$. Therefore, the choice of $q$ will be data-dependent. The estimator is unbiased at $\eta = 0$ with $\delta(\eta) = 0$, while the variance $\sigma^2(\eta)$ is constant for large values of $\eta$. When $\eta$ increases, the bias seems to converge to zero for all values of $q$, but the convergence is not fast enough (i.e. the slope is not steep enough) especially for large values of $q$. For the Laplace prior, the convergence is not visible at all as the bias seems constant for large values of $\eta$.

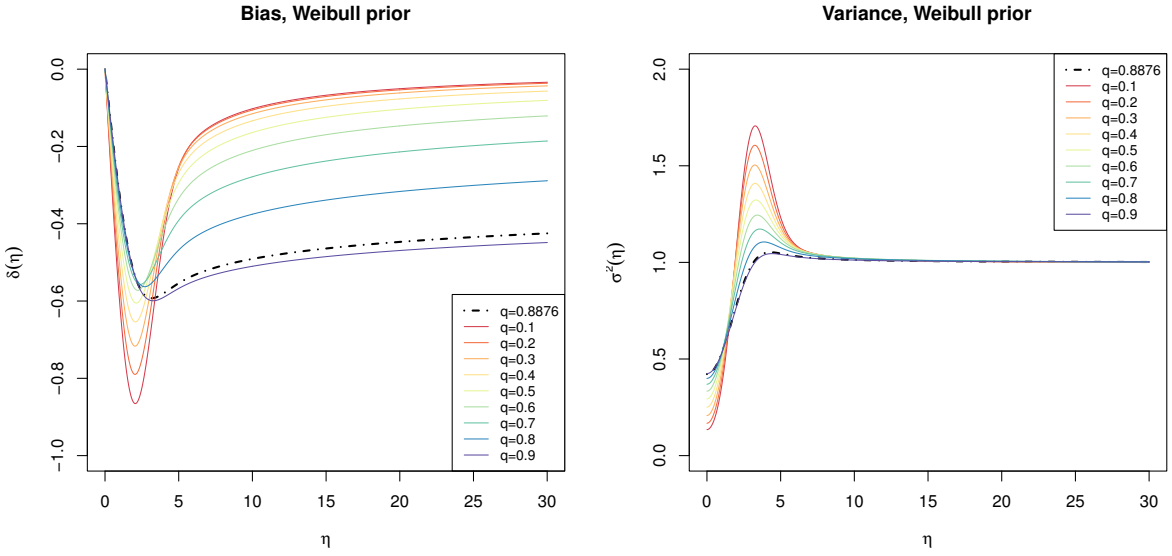



Figure 5.3: Simulated bias and variance of the Monte Carlo tabulated values, Weibull prior with different choices of the parameter q

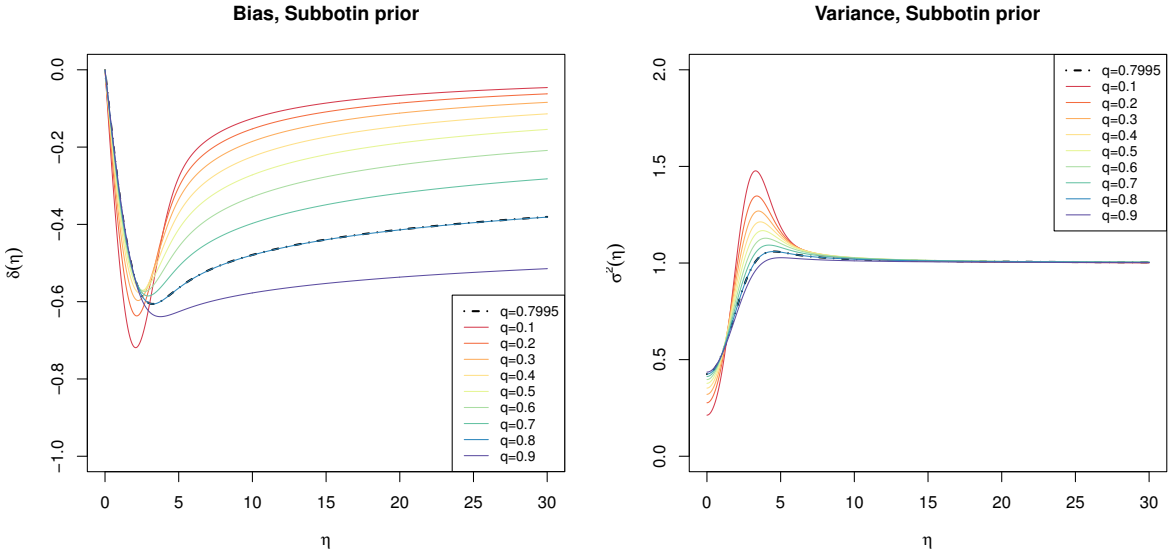



Figure 5.4: Simulated bias and variance of the Monte Carlo tabulated values, Subbotin prior with different choices of the parameter q

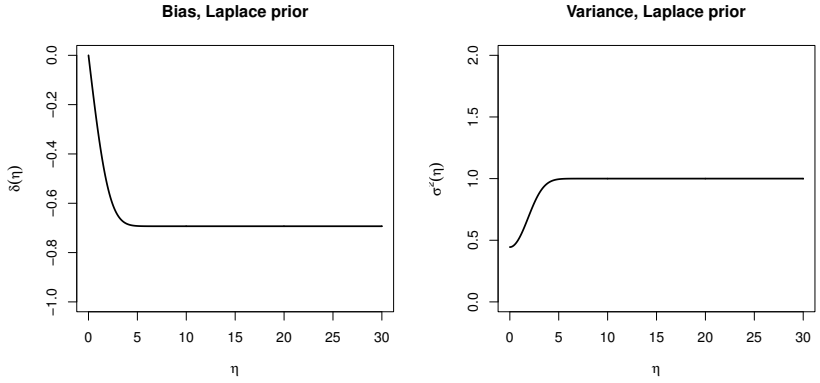



Figure 5.5: Simulated bias and variance of the Monte Carlo tabulated values, Laplace prior

Furthermore, we plot the regret as a function of $\eta$, where the regret is defined as

$$regret(\eta;\alpha,q) = \delta(\eta)^2 + \sigma^2(\eta) - \frac{\eta^2}{1+\eta^2}.$$

In can be seen from Figures 5.6 and 5.7 that for both Weibull and Subbotin, the maximum regret is achieved around $\eta = 4 \sim 5$ for all $q$'s, and the maximum regret is decreasing in $q$ as $q$ increases from 0.1 to the minimax value (0.8876 for Weibull and 0.7995 for Subbotin), and then increases in $q$ for values larger than the minimax value. This confirms that the theoretically derived minimax regret values indeed minimizes the maximum regret.

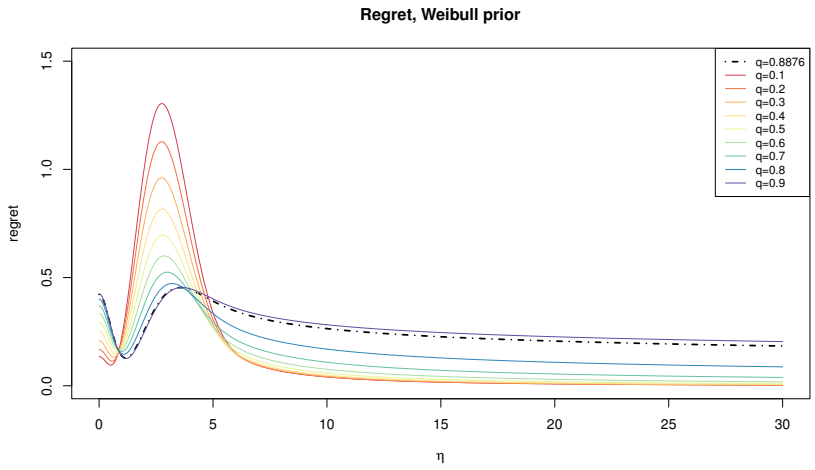



Figure 5.6: Simulated regret of the Monte Carlo tabulated values, Weibull prior with different choices of the parameter q

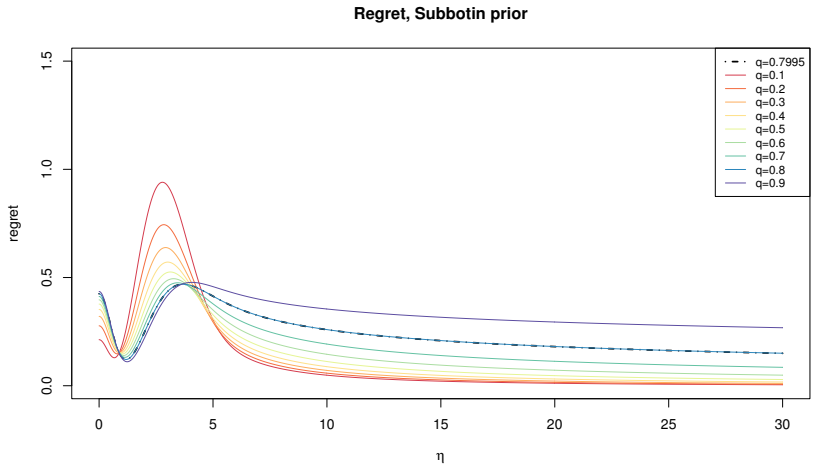



Figure 5.7: Simulated regret of the Monte Carlo tabulated values, Subbotin prior with different choices of the parameter q

## 5.6 Comparison and efficiency of integration routines

Step 5 of the WALS estimation procedure calculates the posterior mean of $\eta$ given $x$, given the assumptions on the prior density. The calculation of the posterior mean requires the evaluation of an integral. Under the Laplace prior, there exists a closed form solution to such integral. Under the Weibull and Subbotin priors, on the other hand, it involves numerical methods to calculate this integral. We introduce the considerations in implementing this numerical calculation, with special attention to the tail of the prior where an asymptotic formula is derived to approximate the integrals for large values of $x$. We present several numerical integration methods, namely the Gauss-Laguerre Quadrature method and the Adaptive Quadrature, and show that the Gauss-Laguerre quadrature method is appropriate under cer-

tain conditions. The Gauss-Laguerre quadrature method is used in all packages we provide for WALS estimation.

### 5.6.1 Integration methods

#### Gauss-Laguerre Quadrature

The Gauss-Laguerre Quadrature is a numerical method for approximating the value of integrals in the following form:

$$\int_0^\infty e^{-x} f(x) dx. \tag{5.6}$$

Using the Gauss-Laguerre method, the integral above is approximated by

$$\int_0^\infty e^{-x} f(x) dx \approx \sum_{i=1}^{n_{Qp}} w_i f(x_i) \tag{5.7}$$

where $n_{Qp}$ is the number of quadrature points, $x_i$ is the Gauss-Laguerre quadrature points and $w_i$ is the Gauss-Laguerre weights. The quadrature points and weights are obtained by solving for the eigenvalues and eigen vectors of a symmetric matrix of order $n_{Qp}$. The accuracy and efficiency of the Gauss-Laguerre method depends on the choice of the number of quadrature points $n_{Qp}$.

#### Adaptive Quadrature

The `integrate` command in `R` calculates adaptive quadrature of functions of one variable over a finite or infinite interval. For this method the relative and absolute accuracy may be specified. We set the relative accuracy at $10^{-13}$ and absolute accuracy at $10^{-14}$.

We compare the difference between the posterior means calculated under both methods, for a range of values of $x$. Under the Gauss-Laguerre method, we consider several choices of the number $n_{Qp} \in \{100, 500, 1000, 1500, 2000\}$. For each choice of $n_{Qp}$ we plot $m_{GL}(x) - m_{AQ}(x)$ against $x$. The results under both the Weibull and Subbotin priors (both using the minimax regret parameters) are shown below.

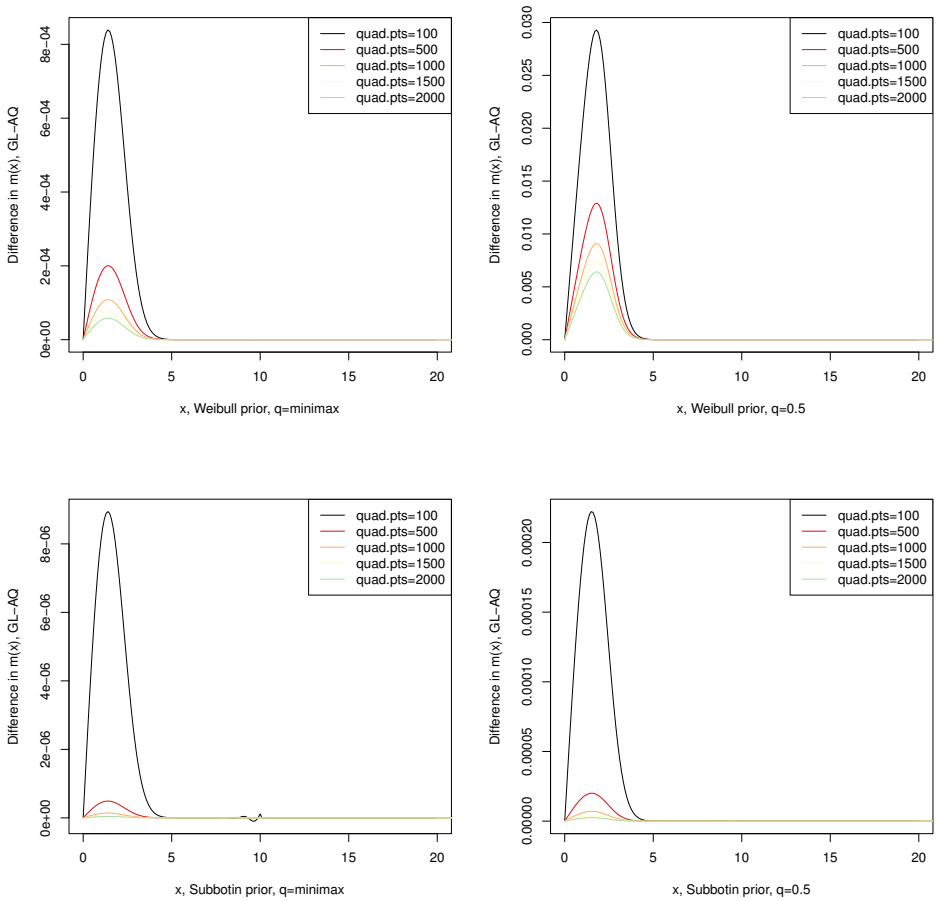



Figure 5.8: Difference between posterior means calculated under different integration routines

As can be seen from the plots above, the difference between the posterior means calculated using Gauss-Laguerre and Adaptive Quadrature decreases as the number of quadrature points under Gauss-Laguerre increases. When $x > 10$, the difference is in the order of $10^{-15}$ for both Weibull and Subbotin priors. When $x \leq 10$ and the number of quadrature points is at least 1000, the maximum difference is in the order of $10^{-4}$ for Weibull and $10^{-7}$ for Subbotin.

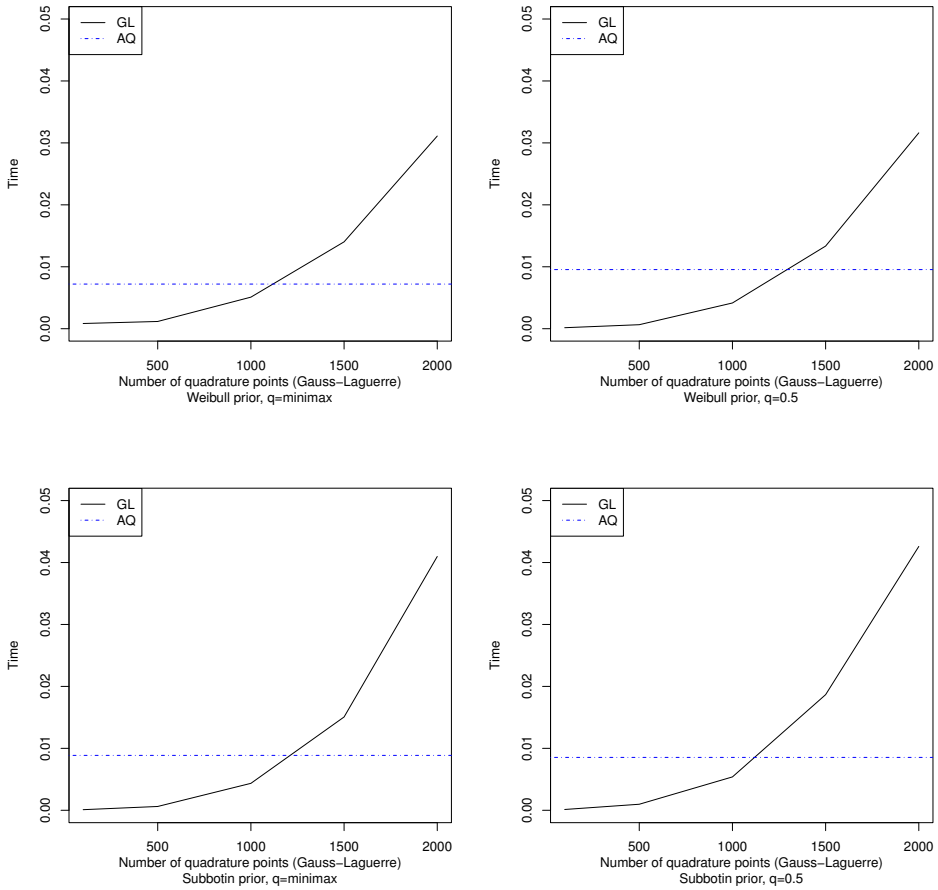



Figure 5.9: Difference in computation time under different integration routines

As can be seen from the graph, computation time under the Gauss-Laguerre increases as the number of quadrature points increases. For $n_{Qp} = 1000$, the Gauss-Laguerre approach is slightly faster than Adaptive quadrature, under both Weibull and Subbotin priors.

Based on the above comparison, we decide to use the Gauss-Laguerre quadrature method for the numerical approximation of integrals in our WALS packages. The default number of quadrature points is set at 1000 while the user has the option to change this value if a different trade-off between accuracy and efficiency is desired. The comparison we show above focuses on the posterior means of a limited number of observations $x$. In actual applications of the WALS estimation, the number of observations can be much larger, where the merit of Gauss-Laguerre method is more prominent: the calculation of Gauss-Laguerre points and weights only needs to be performed once for a given choice of prior parameters and quadrature points, which can then be re-used for the calculation of all integrals within the program. Contrarily, the Adaptive Quadrature method does not benefit from this economy of scale.

### 5.6.2 Derivation of the integral and its asymptotic approximation

Consider a general class of priors, namely the reflected generalized gamma distribution, where the density $\pi(\cdot)$ is of the form

$$\pi(\theta) = \frac{qc^{\delta}}{2\Gamma(\delta)}|\theta|^{-\alpha}e^{-c|\theta|^{q}}. \tag{5.8}$$

We consider three special cases of this class of priors, namely the Weibull ($\alpha + q = 1$), the Subbotin ($\alpha = 0$), and Laplace ($\alpha = 0, q = 1$).

According to Theorem 1 of Pericchi and Smith (1992), the posterior mean of $\eta$, i.e. the conditional distribution of $\eta$ given an observation of $x \sim N(\eta, 1)$ , is given by

$$m(x) = E[\eta|x] = x + \frac{d\log(A_0(x))}{dx} \tag{5.9}$$

where

$$A_0(x) = \int_{-\infty}^{\infty}\phi(x-\eta)\pi(\eta)d\eta = \int_{-\infty}^{\infty}\phi(u)\pi(u+x)du \tag{5.10}$$

and $\phi(\cdot)$ is the normal density. Equation 5.9 is known as the Brown-Tweedie formula.

We also define

$$A_1(x) = \int_{-\infty}^{\infty}(x-\eta)\phi(x-\eta)\pi(\eta)d\eta = -\int_{-\infty}^{\infty}u\phi(u)\pi(u+x)du \tag{5.11}$$

and we immediately have

$$A_0'(x) = \int_{-\infty}^{\infty}\phi'(x-\eta)\pi(\eta)d\eta = \int_{-\infty}^{\infty}-(x-\eta)\phi(x-\eta)\pi(\eta)d\eta = -A_1(x). \tag{5.12}$$

$A_0(x)$ and $A_1(x)$ can also be further simplified into integrals from 0 to infinity, since

$$A_0(x) = \int_{-\infty}^{\infty}\phi(x-\eta)\pi(\eta)d\eta = \int_{0}^{\infty}(\phi(x-\eta)+\phi(x+\eta))\pi(\eta)d\eta \tag{5.13}$$

$$A_1(x) = \int_{-\infty}^{\infty}(x-\eta)\phi(x-\eta)\pi(\eta)d\eta = \int_{0}^{\infty}\left((x-\eta)\phi(x-\eta)+(x+\eta)\phi(x+\eta)\right)\pi(\eta)d\eta \tag{5.14}$$

It follows from the Brown-Tweedie formula that

$$m(x) = x + \frac{d\log(A_0(x))}{dx} = x - \frac{A_1(x)}{A_0(x)}. \tag{5.15}$$

The fraction $\frac{A_1(x)}{A_0(x)}$ can be calculated by numerically solving the two integrals separately and dividing one by another, when $x$ is not too large and both integrals are moderately different from zero. However, as $x \to \infty$, both $A_0(x)$ and $A_1(x)$ converge to 0. We define a cut-off point $x_c$. For $x \leq x_c$, the posterior moments can be directly calculated from equation 5.15. For $x > x_c$, following De Luca et al. (2021a)

we scale both the numerator and the denominator with $\pi(x)$, and obtain

$$A_0(x)/\pi(x) = \int_0^\infty \phi(\eta)\,[\pi(x+\eta)/\pi(x) + \pi(x-\eta)/\pi(x)]\,d\eta \tag{5.16}$$

$$A_1(x)/\pi(x) = \int_0^\infty \eta\phi(\eta)\,[\pi(x-\eta)/\pi(x) - \pi(x+\eta)/\pi(x)]\,d\eta \tag{5.17}$$

$$m(x) = x - \frac{A_1(x)}{A_0(x)} = x - \frac{A_1(x)/\pi(x)}{A_0(x)/\pi(x)}. \tag{5.18}$$

This calculation is valid for large values of $x$ because $A_1(x)/\pi(x) \to 1$ and $A_0(x)/\pi(x) \to 0$ as $x \to \infty$.

### 5.6.3 Choice of the cut-off point

For $x \leq x_c$, we calculate the posterior moments based on equation 5.15 (method 1); for $x > x_c$, we calculate the posterior moments based on equation 5.18 (method 2). The cut-off point $x_c$ is chosen empirically. Using two different specifications of the parameter $q$, namely the minimax regret prior parameters and $q = 0.5$, we investigate the difference in posterior means for both the Weibull and Subbotin priors for a range of values of $x$. As shown in the plots below, the difference between $x$ and the posterior mean $m(x)$ diverges under method 1 when $x$ is above 80, while $x - m(x)$ remains close to 0 under method 2 even when $x$ is large. Furthermore, the difference between the two methods for small values of $x$ is only visible when $x$ is smaller than 10.

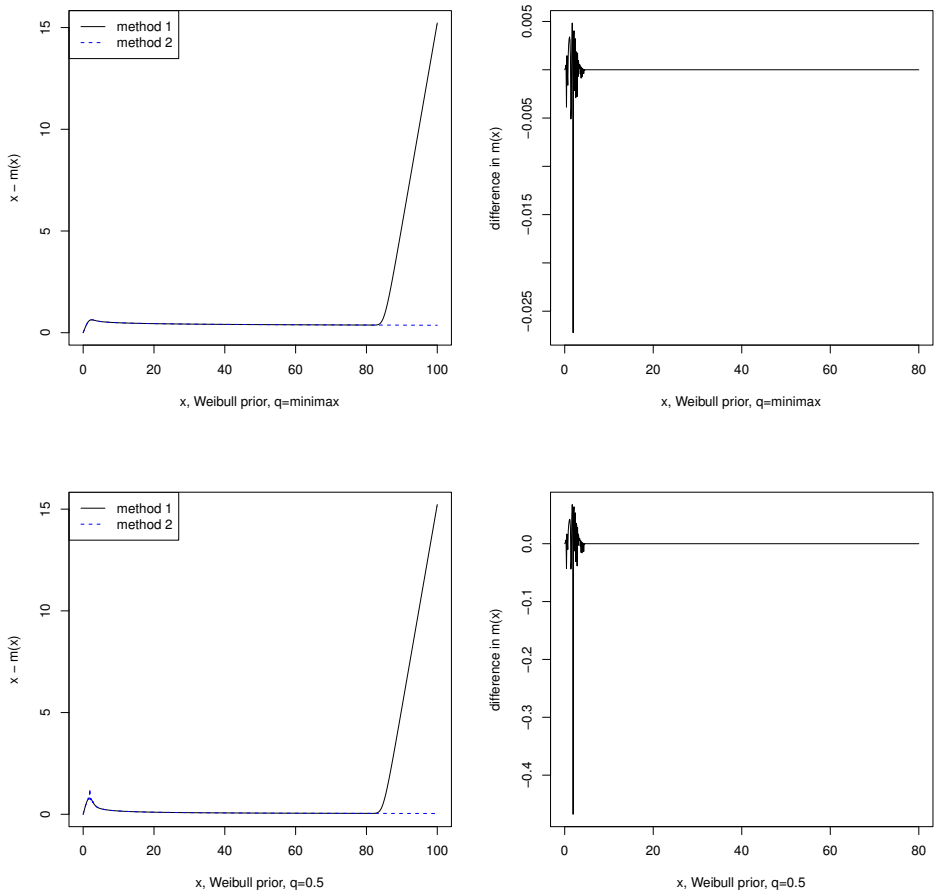



Figure 5.10: Difference between two methods for Weibull prior

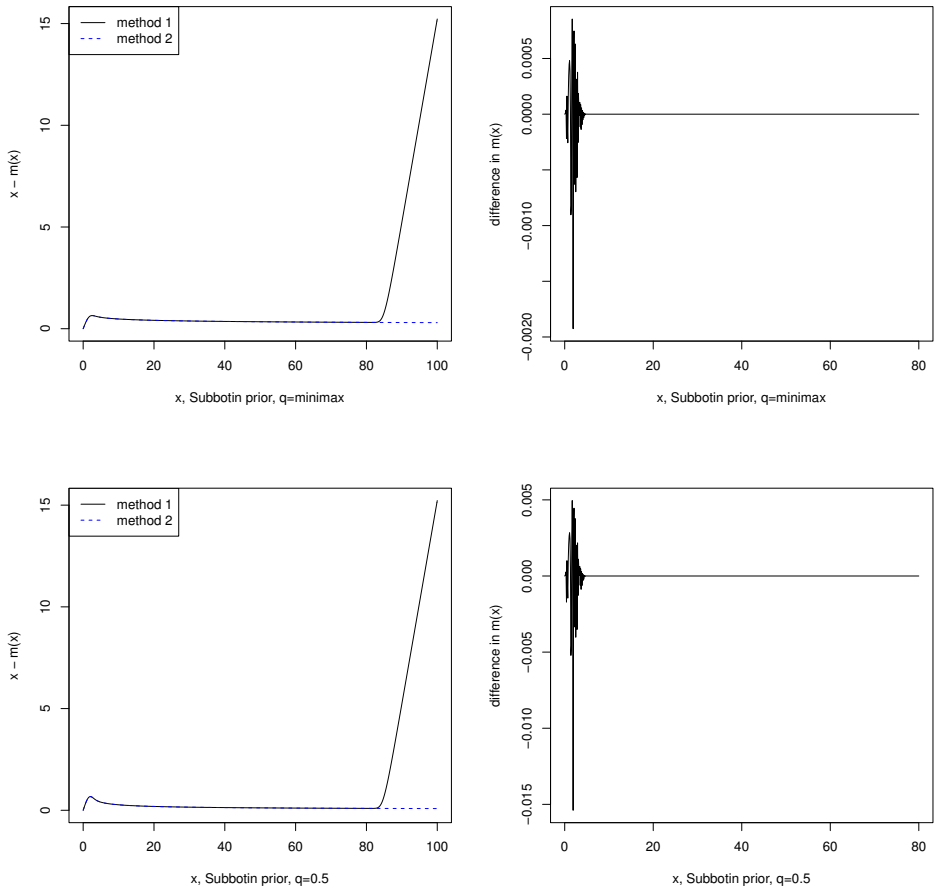



Figure 5.11: Difference between two methods for Subbbotin prior

To assess the performance of the two methods for small $x$, we calculate the posterior mean under the Laplace prior, for which the closed-form solution is available. The numerically obtained integral values are then compared with the theoretical value to see how the two methods perform when $x$ is small. It can be seen from the plot that the deviations from theoretical values from both method 1 and method 2 are close to 0, while the deviation under method 1 is more stable and that under method 2 is more volatile for $x \in [0, 8.24]$. For $8.25 \leq x \leq 82.43$, the deviation in posterior mean from the theoretical values range from $-5.7586502 \times 10^{-4}$ to $1.8913227 \times 10^{-10}$ under method 1 and $-1.4210855 \times 10^{-14}$ to $1.4210855 \times 10^{-14}$ under method 2. Therefore the cut-off point can be chosen anywhere between $[8.25, 82.43]$. We choose the cut-off point $x_c = 10$.

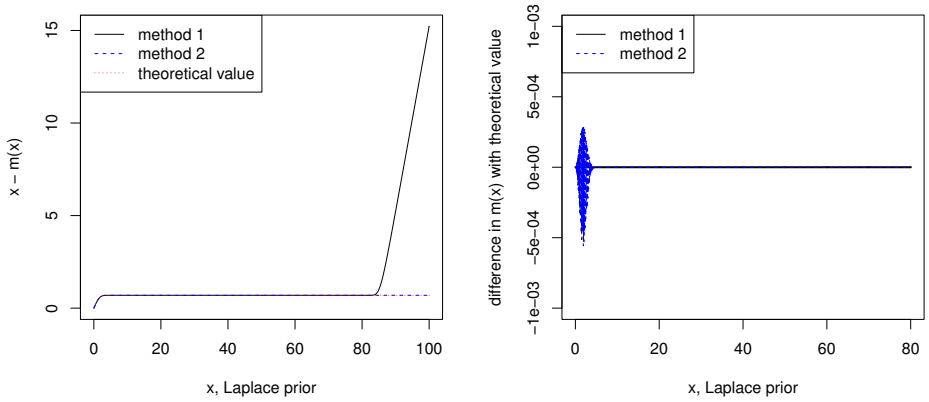



Figure 5.12: Difference between two methods for Laplace prior

## 5.7 Monte Carlo tabulations

The sampling moments of the posterior mean $m(x)$ is obtained numerically. We obtain tabulations of the bias $\delta(\eta)$ and variance $\sigma^2(\eta)$ over a range of values of $\eta$ via Monte Carlo simulation and store the tabulated values for a number of choices of priors in the packages. Detailed procedures on the Monte Carlo tabulation is explained in De Luca et al. (2021b).

We use a large number of replications ($10^6$) for values of $\eta$ between 0 to 30 with a step size of 0.01, under the Weibull/Subbotin priors. In this section we investigate whether the number of replications can be considered sufficient to achieve a certain accuracy level in the tabulated biases.

In Figure **??** we show the mean and 95% confidence interval of the simulated bias and variance corresponding to each value of $\eta$ under the Weibull prior with Mimimax regret parameter $q = 0.8876$. It can seen from the left panel that the bias (defined as $m(x) - \eta$) is always negative. As the values of $\eta$ increases, the bias drops to a minimum around -0.6 for $\eta$ around 4, then gradually reverts to 0. The variance quickly rises from around 0.4 to around 1.1 when $\eta$ increases from 0 to 4, then drops slightly to a stable level around 1.0.

The confidence intervals indicated by the shaded area are the empirical [2.5%, 97.5%] quantiles based on the simulations, where $10^6$ replications were obtained in 200 batches of simulations each with $5,000$ replications. The bias and variances within each batch instead of the individual instances were stored.

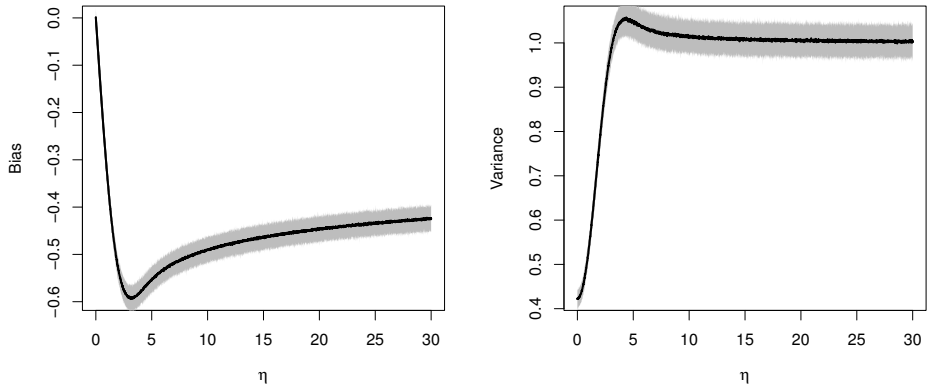



Figure 5.13: Mean and 95% confidence interval of the simulated bias and variance corresponding to each value of $\eta$

We show the absolute and relative errors under 0.01 significance level and our current choice of number of replications, $N = 1,000,000$.

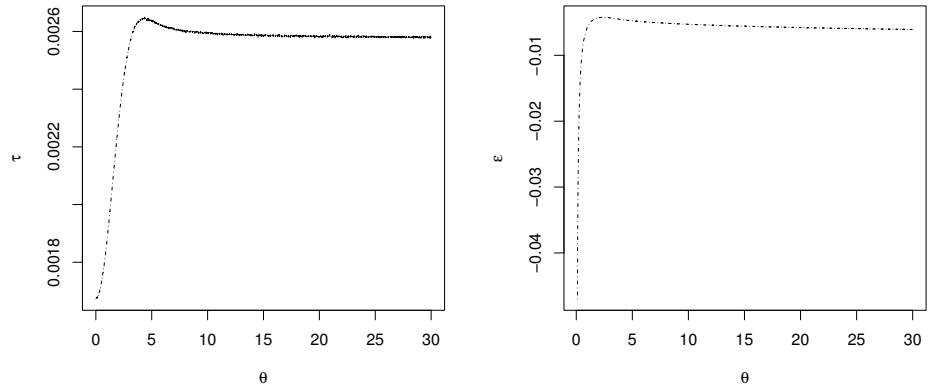



Figure 5.14: Relative and absolute error of the simulated bias when the significance level is 0.01 and total number of replications is 10e6

We calculate the number of required replications of the Monte Carlo simulation if a certain level of accuracy in the final estimates is required. We are interested in the average bias $\theta$ and its estimator $\hat{\theta}$. For each value of $\eta$ between 0 to 30 with a step size of 0.01, $B = 200$ simulation runs were performed, where each run contains $N_B$ replications. So in total $x_i$ is simulated $N = B \times N_B$ times, for each value of $\eta$. We don't have to store each $x_i$, but it is sufficient to store

$$S_1(j) = \frac{\sum_{i=1}^{N_B} x_i}{N_B}, \qquad S_2(j) = \frac{\sum_{i=1}^{N_B} x_i^2}{N_B} - \left(\frac{\sum_{i=1}^{N_B} x_i}{N_B}\right)^2$$

for $j = 1, 2, ..., B$. Next we compute

$$\hat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{B}\sum_{j=1}^{B} S_1(j)$$

and

$$s_N^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{B}\sum_{j=1}^{B}(S_2(j) + S_1(j)^2) - \left(\frac{\sum_{i=1}^{B} S_1(j)}{B}\right)^2$$

From the central limit theorem we know that

$$\frac{N^{1/2}(\hat{\theta}_N - \theta)}{s_N} \sim \mathrm{N}(0,1).$$

Hence, for given $\alpha$ (say, $\alpha = 0.01$ or $0.05$) we have

$$\Pr(-z_{\alpha/2} < \frac{N^{1/2}(\hat{\theta}_N - \theta)}{s_N} < z_{\alpha/2}) = 1 - \alpha.$$

This gives

$$\Pr\left(\hat{\theta}_N - \frac{z_{\alpha/2}s_N}{\sqrt{N}} < \theta < \hat{\theta}_N + \frac{z_{\alpha/2}s_N}{\sqrt{N}}\right) = 1 - \alpha.$$

This result is based on Central Limit Theorem and has been studied in, along others, Kiviet et al. (2012). Following the notation of Kiviet, we denote the absolute error as $\tau$ and the relative error as $\epsilon$. When $\tau$ and $\alpha$ are chosen ex ante, we choose the total number of replications $N$ such that

$$\frac{z_{\alpha/2}s_N}{\sqrt{N}} < \tau,$$

that is, we choose

$$N = z_{\alpha/2}^2 s_N^2 \tau^2.$$

Alternatively, when $\epsilon$ and $\alpha$ are chosen ex ante, we choose the total number of replications $N$ such that

$$\frac{z_{\alpha/2}s_N}{\sqrt{N}} < \epsilon\theta,$$

that is, we choose

$$N = z_{\alpha/2}^2 s_N^2 / (\epsilon^2\theta^2).$$

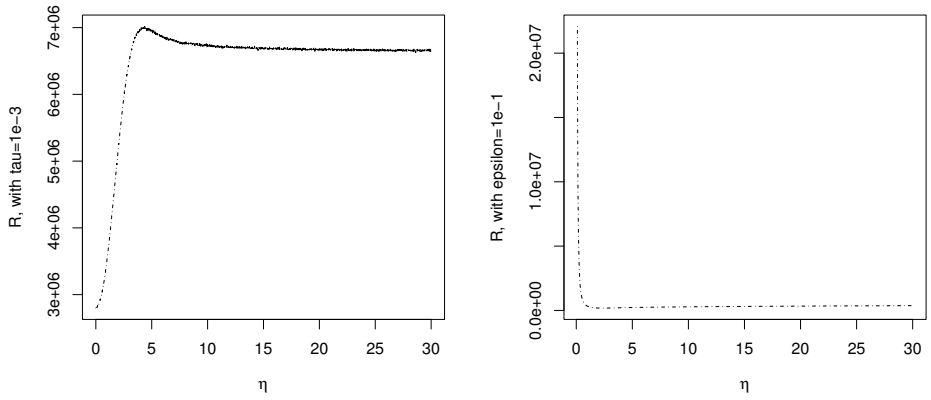



Figure 5.15: Number of required simulations under requirements on the relative or absolute error, when the significance level is 0.01

It can be seen in Figure 5.15 that the number of replication required to achieve a relative error or 1e-3 or an absolute error of 1e-1 depends on the value of $\eta$. For smaller values of $\eta$ it is more difficult to achieve the required absolute error. While in both bases the required $R$ is in the magnitude of 1e6 to 1e7 for most choices of $\eta$, we choose the number of replications N to be 1e6.

## 5.8 Number of Monte Carlo replications

In the WALS estimation procedures it is possible to choose the number of replications and the random seed to use in the Monte Carlo simulation. The choice of the number of Monte Carlo replications has an impact on the precision of the confidence intervals since the confidence intervals are simulation-based. Using a simulated example, we analyze the relationship between the spread in the boundaries of the confidence intervals.

The simulation set-up is as follows. There are 2 focus regressors (first of which is a constant term) and 8 auxiliary regressors. The model coefficient of the second focus regressor $\beta_2$ is of interest. All regressors (apart from the constant term) are multivariate normally distributed with mean 0, variance $\sigma_x^2 = 0.7$, and pairwise correlation $\rho = 0.7$. The true DGP is $y = \beta_1 + \beta_2 X_1 + \beta_3 X_3 ... + \beta_1 0 X_1 0 + \epsilon$ where $\epsilon$ is i.i.d. standard normally distributed. $\beta_1$ and $\beta_2$ are 1 while all other $\beta$'s are equal to $\xi = 0.5$. We generate $n = 2000$ observations for each simulation run, and perform 500 replications for each of the choice of Monte Carlo simulations: $N = 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, ..., 24000, 25000$. In each replication a different Monte Carlo seed is chosen for the WALS estimation.

We use box plots to illustrate the distribution of the WALS estimator of the 95% lower and upper bound of the focus parameter $\beta_2$. It can be seen in Figure 5.16 and 5.17 that as the number of the chosen simulation for WALS estimation becomes larger, the spread in the boundaries of the confidence interval under different random seeds tends to become smaller and smaller. For a sufficiently large number of simulations, the impact of random seeds on the confidence interval is almost completely eliminated. Under the default choice ($mc\_rep = 5000$), 95% of the difference in the interval boundaries between

different random seeds is within 0.01.

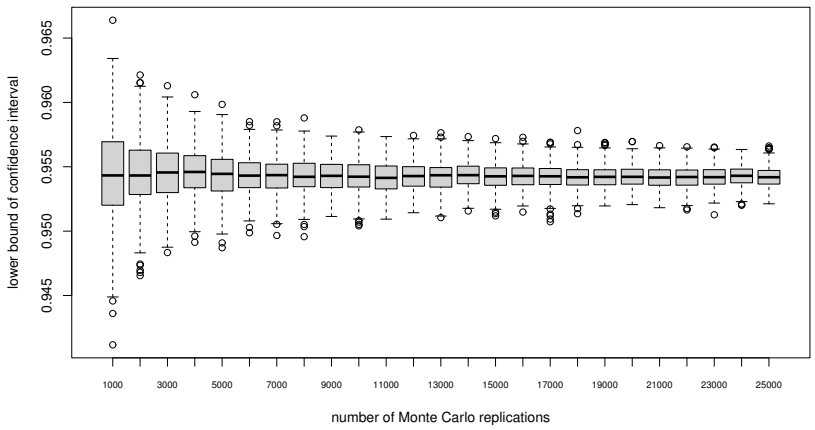



Figure 5.16: Box plot of the lower bound of the 95% confidence interval, under different choices of number of Monte Carlo replications

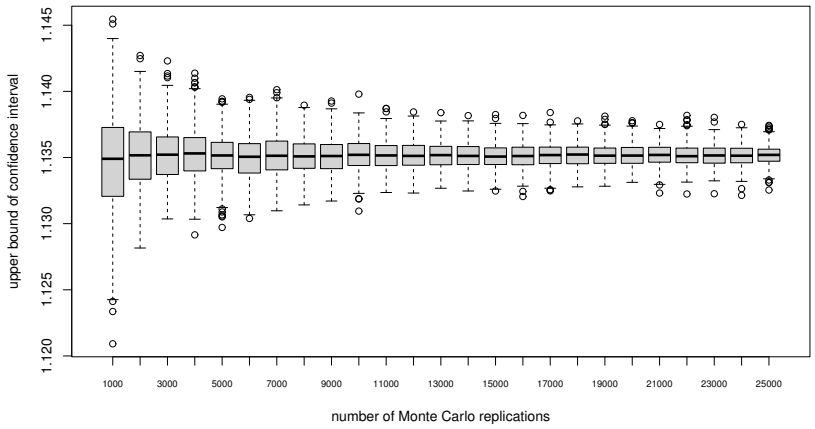



Figure 5.17: Box plot of the upper bound of the 95% confidence interval, under different choices of number of Monte Carlo replications

We also show the average computation time needed under different choices of `mc_rep`. 5.18 shows that, as expected, the computation time is linear in the chosen number of Monte Carlo replications to be used in the estimation of confidence intervals. Therefore, the choice of `mc_rep` should be based on an evaluation of the trade-off between precision of inference and computational efficiency.

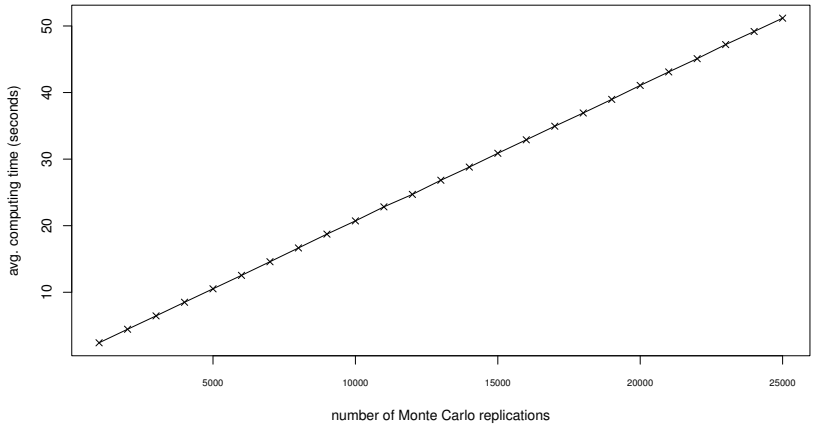



Figure 5.18: Computation time under different choices of number of Monte Carlo replications

## 5.9 Limits to the program

In this section we explore the limits to the WALS program under extreme conditions in the input data. In particular, we investigate two special cases: when the number of auxiliary regressors is large and when the input data is nearly singular.

We use simulated data under the framework in De Luca et al. (2021c). Our goal is to assess the bias and RMSE of the focus regressor in the model.

### 5.9.1 Large $k_2$

We perform the simulation with $k_1 = 2, \sigma_x^2 = \rho = 0.7, \xi = 0.5, \beta_{focus} = [1,1], \beta_{auxiliary} = [\xi, \xi, ...]$, and the error term is standard normal distributed. $k_2$ is chosen between a range of values $\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200\}$. The number of observations $n$ ranges from $\{200, 400, 600, 800, ..., 10000\}$. For each combination of $k_2$ and $n$ we perform 500 replications.

In Figure 5.19 it can be seen that the bias and RMSE increase as $k_2$ increases if $n$ is fixed, and decreases as $n$ increases if $k_2$ is fixed. For large $k_2$ and small $n$, although the program is still able to produce estimates, the bias and RMSE are much higher than with smaller $k_2$ or larger $n$.

From Figure 5.19 we can also see that the number of observations required to achieve a certain level of bias/RMSE in the WALS estimator of the focus regressor increases with the number of auxiliary regressors. For a given bias level (for example $\hat{\beta}_2 = 1 + 0.05$), the number of required observations under each $k_2$ is obtained from a linear interpolation of the known values of $n$ and their corresponding biases. For target bias $\in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ we plot the $n$ that achieves the target bias as a function of $k_2$ in 5.20. It can be seen that the relationship is approximately linear for each value of the targeted bias, but the linear relation is much steeper for smaller bias values, which can be expected. This provides some guidance on the required number of observations for different values of $k_2$, if the user aims to achieve a certain bias level. For example, if $k_2$ is doubled, to achieve roughly the same level of

bias, the number of observations to use in the estimation should also be doubled.

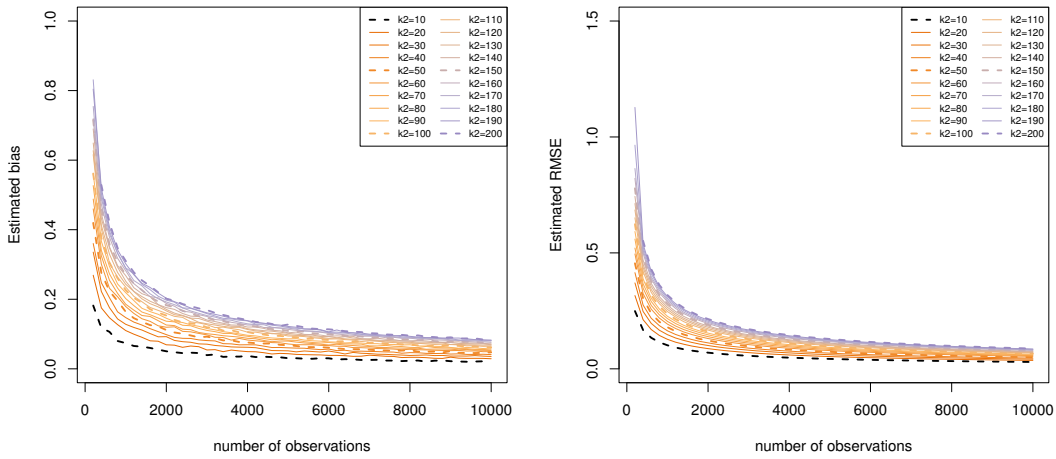



Figure 5.19: Bias and RMSE of the (bias-corrected WALS estimator) of the focus regressor when the number of auxiliary regressors increase

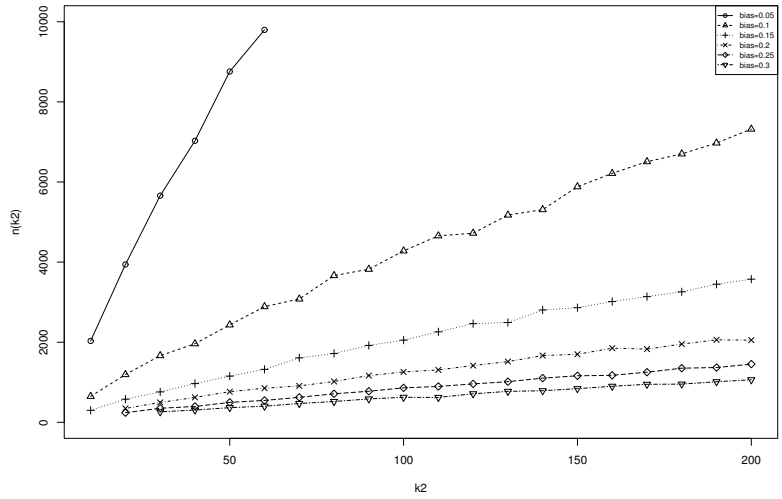



Figure 5.20: Number of observations required to achieve a certain level of bias in the estimator of focus regressor as the number of auxiliary regressors increase

### 5.9.2 Near-singularity

We perform the simulation with $n = 100, k_1 = 2, k_2 = 8, \sigma_x^2 = \rho, \xi = 0.5, \beta_{focus} = [1,1], \beta_{auxiliary} = [\xi, \xi, ...]$, and the error term is standard normal distributed. $\rho$ is chosen between a range of values $\{0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.97, 0.99, 0.999, 0.9999\}$. In Figure 5.21 it can be seen that, while the program is able to produce estimation results, the bias and RMSE increase exponentially as $\rho$ increases.

The programs have built-in checks that ensures the input data has full column rank. However as shown in this section, the check does not stop the program even when the correlation between in the regressors is extremely high. The user is therefore encouraged to perform correlation analysis and check for possible multicollinearity before performing WALS estimation, since the existence of high correlation impacts the bias and RMSE exponentially, as shown in Figure 5.21.

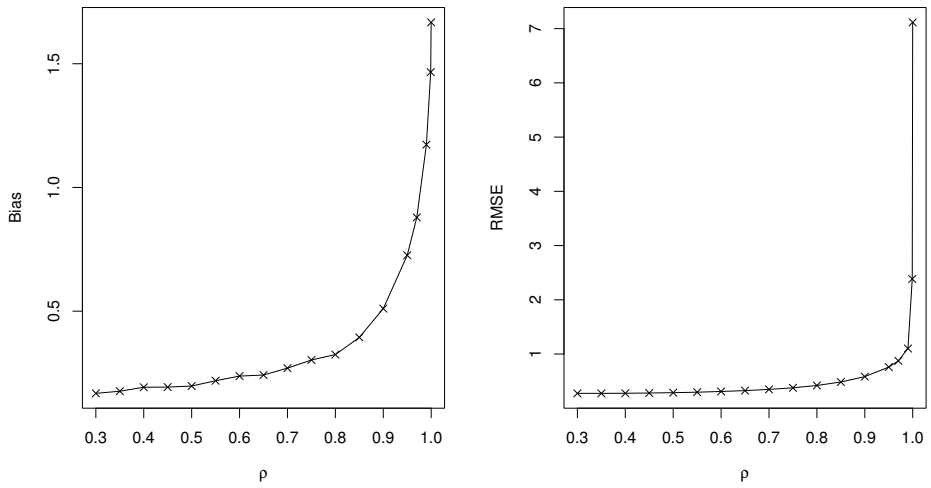



Figure 5.21: Bias and RMSE of the (bias-corrected WALS estimator) of the focus regressor when the correlation between regressors change

## 5.10 Comparison of the three packages

In this section we compare the three implementations in terms of computation time. One of the advantages of the WALS estimation is that it reduces the dimension of the model-averaging problem from $2^{k_2}$ to $k_2$. Using a simulated example we investigate whether the computation time is also reduced to linear.

The simulation set-up is as follows. There are 2 focus regressors (first of which is a constant term) and the number of auxiliary regressors $k_2$ ranges from $[10, 20, 30, ..., 100]$. The model coefficient of the second focus regressor $\beta_2$ is of interest. All regressors (apart from the constant term) are multivariate normally distributed with mean 0, variance $\sigma_x^2 = 0.7$, and pairwise correlation $\rho = 0.7$. The true DGP is $y = \beta_1 + \beta_2 X_1 + \beta_3 X_3 ... + \beta_1 0 X_1 0 + \epsilon$ where $\epsilon$ is i.i.d. standard normally distributed. $\beta_1$ and $\beta_2$ are 1 while all other $\beta$'s are equal to $\xi = 0.5$. We generate either $n = 500$ or $n = 1000$ observations for each simulation run. The computation time (in seconds) using the `R`, `Python`, and `Stata` packages are shown in Figure 5.22, where the WALS estimation uses the default settings with Weibull prior and $q$=minimax values. The calculations were performed on a Laptop with Intel(R) Core(TM) i7-8565U CPU/1.80 GHz with 4 core 8 processor and 8 GB of RAM.

From this figure we can see that the number of observations has almost no effect on the computation time. The computation time is approximately linear in the number of auxiliary regressors under all packages. The difference in computation time across different packages is quite substantial, with `Stata`

being the fastest, R being the slowest and Python in the middle. The underlying programming language of our Stata program, Mata, is a byte-compiled language with syntax similar to C/C++. On the other hand, R and Python are both interpreted high-level programming languages which can be notoriously slow when it involves for loops. Attempts at vectorization and parallelization have been tried but they did not seem to significantly improve the computation efficiency for our R and Python packages. In future improvements of the packages we will consider extending the R and Python packages with functionalities to call externally compiled functions, for example by using the Rcpp library (for R) or Cython module (for Python).

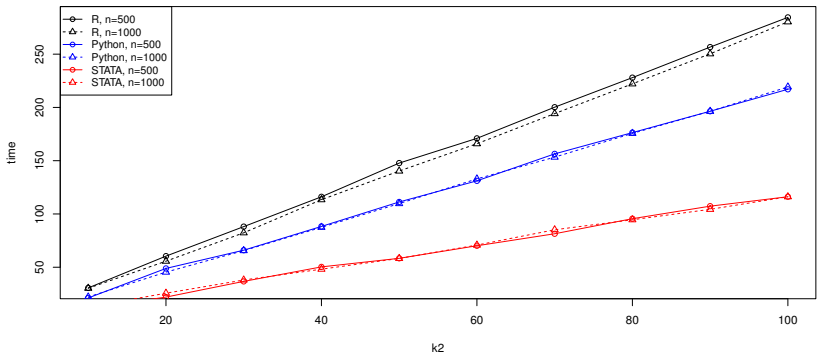



Figure 5.22: Computation time using different packages

## 5.11 Concluding remarks

In this chapter we have analyzed the properties of the WALS estimator, with a focus on its computational aspects. Three computational programs (from Stata, R, and Python) that implement WALS estimation and prediction are introduced and compared. In our analyses, we explored the choice of prior and prior parameters as well as the effect of the choice on estimation. We found that the Weibull and Subbotin priors lead to less bias than the Laplace prior, while the theoretically derived minimax regret parameter leads to more bias than the other choice of parameter, $q = 0.5$. We have also explored the effect of the choice on the numerical integration routine. We show that adaptive quadrature, compared with the Gauss-Laguerre quadrature used in the programs, is less efficient when the scale of the data is large. The program contains pre-stored Monte Carlo tabulations that are used for the plug-in estimators of posterior moments. We show the level of absolute and relative errors that may have arisen give the number of replications used in the simulation of the Monte Carlo tabulations. On the other hand, the estimation program itself contains one step with Monte Carlo simulations where random draws are necessary for the computation of confidence intervals. We show that the number of simulations used in this step impacts the precision of the boundaries calculated for the confidence intervals. Finally, we have explored the limits to the program with respect to the dimension of the model and near singularity. We found that the number of observations required to achieve a certain level of bias is approximately a linear function in the number of auxiliary regressors. On the other hand, although the program is able to generate results when the input data is nearly singular (but still has full column rank), the bias increases

exponentially as the level of correlation increases.

There are several directions for future extensions.

First, the WALS framework we use assumes homogeneity in the error components. It is possible to extend this to the heterogeneous case.

Second, the simple regression model can be extended to generalized linear models, as analyzed in De Luca et al. (2018). The results using, for example, Logit, Probit, and Poisson models, can be added to enrich the framework and the computer programs we provide.

Lastly, we have not investigated the treatment of missing values in WALS estimation. Dardanoni et al. (2012) discussed the estimation of a linear regression model using data with missing values where imputations can be used to fill the missing values. Options for imputing missing values can be embedded in future extensions of the programs we provide.

# Bibliography


Abadir, K. M. and Magnus, J. R. (2005). *Matrix Algebra*, volume 1 of *Econometric Exercises* (Eds. K. M. Abadir, J. R. Magnus, and P. C. B. Phillips). Cambridge University Press: New York.

Abdellaoui, M. (2000). Parameter-free elicitation of utility and probability weighting functions. *Management Science*, 46:1497–1512.

Afonso, A. and Jalles, J. T. (2019). Quantitative easing and sovereign yield spreads: Euro-area time-varying evidence. *Journal of International Financial Markets, Institutions and Money*, 58:208–224.

Aït-Sahalia, Y., Cacho-Diaz, J., and Laeven, R. J. (2015). Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics*, 117(3):585–606.

Aït-Sahalia, Y., Laeven, R. J., and Pelizzon, L. (2014). Mutual excitation in Eurozone sovereign CDS. *Journal of Econometrics*, 183:151–167.

Aller, C., Ductor, L., and Grechyna, D. (2021). Robust determinants of CO2 emissions. *Energy Economics*, 96:105154.

Avery, R. B. (1977). Error components and seemingly unrelated regressions. *Econometrica*, 45:199–209.

Balestra, P. and Nerlove, M. (1966). Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas. *Econometrica*, 34:585–612.

Baltagi, B. H. (1980). On seemingly unrelated regressions with error components. *Econometrica*, 48:1547–1551.

Baltagi, B. H. (2008). *Econometric Analysis of Panel Data*. John Wiley: New York.

Bauer, D. and Kramer, F. (2016). The risk of a mortality catastrophe. *Journal of Business & Economic Statistics*, 34:391–405.

Beron, K. J., Murdoch, J. C., Thayer, M. A., and Vijverberg, W. P. (1997). An analysis of the housing market before and after the 1989 Loma Prieta earthquake. *Land Economics*, 73:101–113.

Bin, O. and Landry, C. E. (2013). Changes in implicit flood risk premiums: Empirical evidence from the housing market. *Journal of Environmental Economics and Management*, 65(3):361–376.

Bin, O. and Polasky, S. (2004). Effects of flood hazards on property values: Evidence before and after Hurricane Floyd. *Land Economics*, 80(4):490–500.

Boswijk, H. P., Laeven, R. J., and Lalu, A. (2016). Asset returns with self-exciting jumps: Option pricing and estimation with a continuum of moments. *Working paper*.

Brookshire, D. S., Thayer, M. A., Tschirhart, J., and Schulze, W. D. (1985). A test of the expected utility model: Evidence from earthquake risks. *Journal of Political Economy*, 93(2):369–389.

Chamberlain, G. and Griliches, Z. (1975). Unobservables with a variance-components structure: Ability, schooling, and the economic success of brothers. *International Economic Review*, 16:422–449.

Clarke, J. A. (2017). Model averaging OLS and 2SLS: An application of the WALS procedure. *Working paper*.

Comunale, M. and Mongelli, F. P. (2020). Who did it? A European detective story. was it real, financial, monetary and/or institutional: Tracking growth in the Euro area with an atheoretical tool. *CAMA Working Paper*.

Daniel, V. E., Florax, R. J., and Rietveld, P. (2009). Flooding risk and housing values: An economic assessment of environmental hazard. *Ecological Economics*, 69(2):355–365.

Danilov, D. and Magnus, J. R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics*, 122(1):27–46.

Dardanoni, V., De Luca, G., Modica, S., and Peracchi, F. (2012). A generalized missing-indicator approach to regression with imputed covariates. *The Stata Journal*, 12(4):575–604.

De Luca, G. and Magnus, J. R. (2011). Bayesian model averaging and weighted-average least squares: Equivariance, stability, and numerical issues. *The Stata Journal*, 11(4):518–544.

De Luca, G., Magnus, J. R., and Peracchi, F. (2018). Weighted-average least squares estimation of generalized linear models. *Journal of Econometrics*, 204(1):1–17.

De Luca, G., Magnus, J. R., and Peracchi, F. (2021a). Asymptotic properties of WALS. *In progress*.

De Luca, G., Magnus, J. R., and Peracchi, F. (2021b). Sampling properties of the Bayesian posterior mean with an application to WALS estimation. *Journal of Econometrics*, to appear, https://doi.org/10.1016/j.jeconom.2021.04.008.

De Luca, G., Magnus, J. R., and Peracchi, F. (2021c). Weighted-average least squares (WALS): Confidence and prediction intervals. *Tinbergen Institute Discussion Paper 2021-038/III*.

Eeckhoudt, L. R., Laeven, R. J., and Schlesinger, H. (2020). Risk apportionment: The dual story. *Journal of Economic Theory*, 185:104971.

Furceri, D., Ganslmeier, M., Ostry, J. D., and Yang, N. (2021). Initial output losses from the Covid-19 pandemic: Robust determinants. *CEPR Discussion Paper No. DP15892*.

Furceri, D. and Ostry, J. D. (2019). Robust determinants of income inequality. *Oxford Review of Economic Policy*, 35(3):490–517.

Gu, T., Nakagawa, M., Saito, M., and Yamaga, H. (2011). On asymmetric effects of changes in regional risk rankings on relative land prices in the Tokyo Metropolitan area: A test of implications implied by prospect theory using market equilibrium prices (in Japanese). *Journal of Behavioral Economics and Finance*, 4:1–19.

Hanaoka, C., Shigeoka, H., and Watanabe, Y. (2018). Do risk preferences change? Evidence from the Great East Japan earthquake. *American Economic Journal: Applied Economics*, 10(2):298–330.

Hidano, N., Hoshino, T., and Sugiura, A. (2015). The effect of seismic hazard risk information on property prices: Evidence from a spatial regression discontinuity design. *Regional Science and Urban Economics*, 53:113–122.

Ikefuji, M., Laeven, R. J., Magnus, J. R., and Yue, Y. (2021). Earthquake risk embedded in property prices: Evidence from five japanese cities. *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2021.1928512.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292.

Kawawaki, Y. and Ota, M. (1996). The influence of the Great Hanshin-Awaji earthquake on the local housing market. *Review of Urban & Regional Development Studies*, 8(2):220–233.

Kiviet, J. F. et al. (2012). *Monte Carlo Simulation for Econometricians*, volume 5 of *Foundations and Trends® in Econometrics*. NOW publishers: Boston/Delft.

Liski, A., Liski, E., Sund, R., and Juntunen, M. (2010). A comparison of WALS estimation with pretest and model selection alternatives with an application to costs of hip fracture treatments. In *Proceedings of the Third Workshop on Information Theoretic Methods in Science and Engineering*, pages 1–6. Tampere International Center for Signal Processing.

Magnus, J. R. (1982). Multivariate error components analysis of linear and nonlinear regression models by maximum likelihood. *Journal of Econometrics*, 19(2-3):239–285.

Magnus, J. R. (2017). *Introduction to the Theory of Econometrics*. VU University Press: Amsterdam.

Magnus, J. R. and De Luca, G. (2016). Weighted-average least squares (WALS): A survey. *Journal of Economic Surveys*, 30(1):117–148.

Magnus, J. R. and Durbin, J. (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica*, 67(3):639–643.

Magnus, J. R., Powell, O., and Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics*, 154(2):139–153.

Mignamissi, D. and Kuete, Y. F. M. (2020). What makes Africans happy? *Economics Bulletin*, 40(4):2741–2754.

Nakagawa, M., Saito, M., and Yamaga, H. (2007). Earthquake risk and housing rents: Evidence from the Tokyo Metropolitan area. *Regional Science and Urban Economics*, 37(1):87–99.

Nakagawa, M., Saito, M., and Yamaga, H. (2009). Earthquake risks and land prices: Evidence from the Tokyo Metropolitan area. *The Japanese Economic Review*, 60(2):208–222.

Naoi, M., Seko, M., and Ishino, T. (2012). Earthquake risk in Japan: Consumers' risk mitigation responses after the Great East Japan earthquake. *Journal of Economic Issues*, 46(2):519–530.

Naoi, M., Seko, M., and Sumita, K. (2009). Earthquake risk and housing prices in Japan: Evidence before and after massive earthquakes. *Regional Science and Urban Economics*, 39(6):658–669.

Ogata, Y. (1981). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27.

Pericchi, L. and Smith, A. (1992). Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):793–804.

Poghosyan, K. and Magnus, J. R. (2012). WALS estimation and forecasting in factor-based dynamic models with an application to Armenia. *International Econometric Review*, 4(1):40–58.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 66:497–527.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, 3(4):323–343.

Rahman, K. U. and Shang, S. (2020). A regional blended precipitation dataset over Pakistan based on regional selection of blending satellite precipitation datasets and the dynamic weighted average least squares algorithm. *Remote Sensing*, 12(24):4009.

Rahman, K. U., Shang, S., Shahid, M., Wen, Y., and Khan, A. J. (2020). Development of a novel weighted average least squares-based ensemble multi-satellite precipitation dataset and its comprehensive evaluation over Pakistan. *Atmospheric Research*, 246:105133.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55.

Seya, H., Tsutsumi, M., Mendez, S., and Vega, A. (2012). Application of model averaging techniques to spatial hedonic land price models. *Econometrics: New Research*, pages 63–88.

Shimizu, C. and Nishimura, K. (2006). Biases in appraisal land price information: The case of Japan. *Journal of Property Investment & Finance*, 24(2):150–175.

Tumala, M. M., Olubusoye, O. E., Yaaba, B. N., Yaya, O. S., and Akanbi, O. B. (2018). Investigating predictors of inflation in Nigeria: BMA and WALS techniques. *African Journal of Applied Statistics*, 5(1):301–321.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

Wu, G. and Gonzalez, R. (1996). Curvature of the probability weighting function. *Management Science*, 42(12):1676–1690.

Xu, W. (2014). Law matters?: A Bayesian analysis of the cross-country relationship between anti-self-dealing rules and stock market outcomes. *Applied Economics Letters*, 21(5):366–371.

Yaari, M. E. (1987). The dual theory of choice under risk. *Econometrica*, 55:95–115.

Yamaga, H., Nakagawa, M., and Saito, M. (2002). Earthquake risks and land pricing: The case of the Tokyo Metropolitan area (in Japanese). *Journal of Applied Regional Science*, 7:51–62.

# Samenvatting (Summary in Dutch)

Dit proefschrift bevat de resultaten van twee ongerelateerde projecten: risico- en vastgoedprijzen, en computationele aspecten van *model averaging* (modelmiddeling).

In het eerste project onderzoeken we het effect van objectief en subjectief aardbevingsrisico ingebed in Japanse vastgoedprijzen. Dit wordt gemeten in het kader van een hedonisch prijsmodel, het referentiemodel voor het analyseren van vastgoedprijzen. In hedonische prijsmodellen worden de kenmerken van onroerend goed gezien als componenten die elk onafhankelijk bijdragen aan een deel van de prijs van het onroerend goed.

We verzamelen een rijke dataset met transactieprijzen van woningen en verschillende kenmerken die relevant zijn voor vastgoedprijzen. We maken onderscheid tussen drie soorten woningen: woongrond (alleen grond), woongrond (grond en gebouw) en appartementen in eigendom. Elk type heeft verschillende kenmerken, maar deelt ook veel kenmerken. Onder de kenmerken van het onroerend goed bevinden zich transversale gegevens, zoals informatie over de aantrekkelijkheid van de wijk waar het onroerend goed is gelegen, tijdreeksgegevens zoals macro-economische variabelen, en ook individuele kenmerken zoals oppervlakte, dekkingsgraad van gebouwen of de afstand naar het dichtstbijzijnde station.

Om de beschikbare dataset te benutten, gebruiken we een regressiemodel voor multivariate storingscomponenten (*multivariate error components*). De storingstermen zijn de som van drie onafhankelijke componenten, die de tijd-specifieke, doorsnede-specifieke en individueel-specifieke effecten vastleggen. Bovendien is elke component een vector in plaats van een scalair, waardoor vergelijkingen van nauw verwante foutstructuren kunnen worden gepoold met behoud van een relatief klein aantal parameters. Deze vector heeft drie elementen, die elk overeenkomen met één van de drie eigenschapstypen. De dimensie van deze enorme variantiematrix veroorzaakt door de vectorvorm kan drastisch worden verminderd dankzij de structuur van de storingscomponenten.

Wij introduceren aardbevingsrisico gemeten als de kans dat een aardbeving een bepaalde drempelwaarde of intensiteit over een bepaalde tijdsperiode overschrijdt. Aangezien aardbevingen frequent voorkomen in Japan en zowel ruimtelijk als temporeel variëren, is aardbevingsrisico een niet te verwaarlozen kenmerk bij de waardering van vastgoed. Wij maken onderscheid tussen lange termijn risico en korte termijn risico. De lange termijn gegevens over het aardbevingsrisico worden geleverd door het *Japan Seismic Hazard Information Station* en dit risico wordt gedefinieerd als de kans dat een aardbeving in de komende dertig jaar bepaalde intensiteitsdrempels overschrijdt in een bepaald gebied. We nemen het gemiddelde van deze lange termijn kans over de gehele steekproef periode om een tijds-

invariante maatstaf te creëren voor de algehele risicograad van een bepaald gebied. De korte termijn kans is een negentig-dagen kans die varieert per tijdsperiode en per stad, gesimuleerd door een temporeel epidemisch naschoksequentie (ETAS) model. Het ETAS-model is een pad-afhankelijk gemarkeerd puntproces dat vaak wordt gebruikt voor het modelleren van seismische activiteiten. Het idee achter het ETAS-model is dat elke aardbeving naschokken kan veroorzaken zoals bij epidemieën en dat de intensiteit van de impact van elke getriggerde gebeurtenis na verloop van tijd afneemt.

Hoewel de kans op aardbevingen op lange en korte termijn worden gezien als objectieve maatstaven voor het aardbevingsrisico, proberen we ook een subjectieve risicomaatstaf uit de gegevens te halen. Dit wordt bereikt door gebruik te maken van een parametrische familie van kanswegingsfuncties, die veel worden gebruikt in economische analyse en beslissingstheorie. Het idee is dat we, door de gewogen (subjectieve) kans in te voeren in plaats van de oorspronkelijke (objectieve) kans in de regressiefunctie, (met maximale waarschijnlijkheid) uit de gegevens de onbekende parameter kunnen schatten door een rasterzoekopdracht uit te voeren. De bijbehorende variantie van deze schatter moet worden afgeleid omdat de situatie niet standaard is in die zin dat één van de regressoren afhangt van de parameter van belang. De geschatte parameter werpt licht op de vorm van de kanswegingsfunctie en geeft zo inzicht in hoe de perceptie van mensen van een kleine en grote kans wordt weerspiegeld in de vastgoedprijzen. Wanneer de kanswegingsfunctie omgekeerd $S$-vormig is, betekent dit dat mensen een kleine kans te zwaar wegen en een grote te licht. Wanneer de kanswegingsfunctie $S$-vormig is, betekent dit dat mensen een kleine kans te licht inschatten en een grote kans te zwaar. Wanneer de parameter gelijk is aan 1, degenereert de functie tot de identiteitsfunctie, wat betekent dat er geen subjectieve vervorming van de kans is.

We ontdekten dat het objectieve aardbevingsrisico op lange termijn een aanzienlijk negatief effect heeft op de vastgoedprijzen. De extra impact van het objectieve aardbevingsrisico op korte termijn is niet significant verschillend van nul. De vertekende waarschijnlijkheden van aardbevingen op korte termijn (waarbij de waarschijnlijkheid wordt gewogen) hebben echter een aanzienlijk negatief effect op de vastgoedprijzen. We vonden dat deze kanswegingsfunctie $S$-vormig was, waardoor een kleine kans te licht werd gewogen en een grote kans te zwaar. Deze bevinding is in strijd met de conventionele wijsheid in de beslissingstheorie, waar kanswegingsfuncties gewoonlijk omgekeerd $S$-vormig zijn, hetgeen kan worden verklaard door het feit dat de intensiteit van de aardbeving op de achtergrond groter is dan nul, zodat mensen geen tijdelijke afwijkingen van een korte termijn aardbevingsrisico lopen met een referentiekans van nul maar met een positieve referentiekans.

In het tweede project bestuderen we de eigenschappen van een model-gemiddelde schatter, namelijk de gewogen gemiddelde kleinste kwadraten (WALS) schatter. Het idee van modelmiddeling komt voort uit het inzicht dat model selectie en -schatting niet als twee afzonderlijke stappen moeten worden gezien, maar als één geïntegreerde procedure. Modelmiddeling selecteert niet één best passend kandidaatmodel, maar schat een hele reeks kandidaatmodellen en kent gewichten toe aan elk van de kandidaatschattingen.

De meest voorkomende naïeve toepassing van de $t$-ratio in de toegepaste econometrie als diagnostische statistiek gaat als volgt. Wanneer de $t$-ratio van een regressor boven een bepaalde drempel ligt (meestal 1.96 op het 5% significantieniveau), wordt de regressor als "significant" beschouwd en in het model gehouden; en wanneer de $t$-ratio onder die drempel ligt, wordt deze uit het model verwijderd.

Deze benadering negeert dus het feit dat dezelfde gegevens zijn gebruikt voor diagnostische testen en schattingen, zodat gevolgtrekkingen verkregen uit de tweede stap waarschijnlijk misleidend nauwkeurig zullen zijn omdat het de onzekerheid negeert die in de eerste stap wordt gegenereerd.

De hierboven beschreven procedure wordt *pretesting* genoemd en leidt tot schatters die niet differentieerbaar zijn en dus niet toelaatbaar (*inadmissible*). Het is de eenvoudigste vorm van een WALS-schatter, namelijk het geval waarin de gewichten van kandidaatmodellen slechts 0 of 1 kunnen zijn. De WALS-procedure generaliseert deze discrete versie naar een continue versie, waarbij de gewichten nu continue functies zijn van de $t$-ratio. De literatuur over modelmiddeling bestaat vooral uit een frequentistische benadering (FMA) of een Bayesiaanse benadering (BMA). Wij onderzochten de eigenschappen van WALS, die een Bayesiaanse combinatie is van frequentistische schatters. Deze schatter heeft voordelen ten opzichte van de traditionele BMA-schatters in termen van interpretatie en rekenefficiëntie.

Het raamwerk van WALS is het lineaire regressiemodel met onafhankelijke en identiek verdeelde normale storingstermen. We maken onderscheid tussen focusregressoren, die we in het model willen behouden, ongeacht de uitkomst van diagnostische toetsen, en hulpregressoren, die al dan niet in het model kunnen voorkomen.

Wij ontwikkelden statistische pakketten die de berekening van WALS-schattingen, standaardfouten, bias, gemiddelde kwadratische fouten, betrouwbaarheidsintervallen en voorspellingen mogelijk maken. De schatting hangt af van een keuze van prior verdelingen en prior parameters, die afkomstig zijn van een gereflecteerde gegeneraliseerde Gamma familie — de Weibull, Subbotin en Laplace prior verdelingen. We laten zien dat de Laplace prior leidt tot schattingen met een hogere bias, terwijl voor Weibull en Subbotin prior de theoretisch verkregen minimax-regret-prior parameters, die de maximale regret over alle mogelijke waarden minimaliseren, tot meer vertekening kunnen leiden dan andere keuzes van de prior parameter.

WALS-schatting maakt gebruik van numerieke integratieresultaten, behalve in het geval van de Laplace prior. We laten het effect zien van de keuze tussen twee alternatieve integratieroutines, de Gauss-Laguerre kwadratuur en de adaptieve kwadratuur op de precisie en rekenkundige efficiëntie van het programma. We verkennen ook de grenzen van de WALS-schattingsprocedure door gebruik te maken van simulatie-opstellingen waarbij de matrix van regressoren bijna singulier is terwijl het aantal hulpregressoren groot is. We ontdekten dat, zolang de invoergegevens van volledige kolom rang zijn, de WALS procedure schattingen kan produceren, hoewel de bias exponentieel toeneemt wanneer de correlatie tussen de regressoren toeneemt. We hebben een relatie gelegd tussen het aantal vereiste observaties en het aantal hulpregressoren onder hetzelfde beoogde bias-niveau, en vonden dat deze relatie ongeveer lineair is.

Het project heeft zich gericht op de computationele eigenschappen van de WALS-schatter. Afgezien van de hierboven genoemde bevindingen, laten we enkele andere aspecten zien van de schattingsprocedures van WALS, zoals het effect van Monte Carlo replicaties op de precisie van betrouwbaarheidsintervallen, en de vergelijking van de rekensnelheid van verschillende pakketten (`R`, `Python`, of `Stata`). Door deze aspecten te onderzoeken, willen we inzicht verschaffen in de prestaties van de WALS-schatter en de verschillende beschikbare schattingsopties, zodat een doorsnee gebruiker weloverwogen beslissingen kan nemen bij het gebruik van WALS in empirische toepassingen.

Het proefschrift is als volgt opgebouwd. Hoofdstuk 2 zet een model op met een structuur met meerdere storingscomponenten en leidt de bijbehorende maximale aannemelijkheid schattingsprocedure af. Het ontwerpt ook een rasterzoekprocedure en leidt de variantie af van de parameter van belang wanneer een van de regressoren afhankelijk is van deze parameter. Hoofdstuk 3 legt het gegevensverzamelingsproces uit voor de empirische studie van het aardbevingsrisico dat is ingebed in vastgoedprijzen. Hoofdstuk 4 toont het volledige beeld van het empirische onderzoek en presenteert de empirische resultaten. Hoofdstuk 5 onderzoekt de (computationele) eigenschappen van de WALS-schatter.