An Empirical Study of Convergence Rates and Resampling-based Confidence Interval Methods For Step Threshold Linear Regression Models

Shuangcheng Hua

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2020

Reading Committee:

Youyi Fong, Chair

Chongzhi Di

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

An Empirical Study of Convergence Rates and Resampling-based Confidence Interval Methods For Step Threshold Linear Regression Models

Shuangcheng Hua

Chair of the Supervisory Committee:
Associate Professor Youyi Fong
Department of Biostatistics

This thesis studies the convergence rates of maximum likelihood estimators and coverage probabilities of resampling-based confidence intervals of parameters in step threshold linear regression models through Monte Carlo experiments under five different data generating models. The results suggest that when the data is not generated from a step threshold model, the convergence rate of the threshold estimator is cubic root n at large sample sizes as theory indicates. In terms of coverage probabilities of the confidence intervals, the results show that the optimal block size for m-out-of-n bootstrap or subsampling depends on whether the model is correctly specified.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Number Page

ACKNOWLEDGMENTS

I would like to express my gratitude to my committee chair, Professor Youyi Fong, who has constantly and convincingly conveyed a spirit of adventure in regard to research and expertly guided me through my graduate education. Without his guidance and persistent help, this thesis would not have been possible.

My appreciation also goes to my second reader, Professor Chongzhi Di, who monitored and reviewed my thesis. His valuable insights and comments launched an indispensable part of this thesis.

And I would like to give my special thank to Professor Jon Wellner in the statistics department of University of Washington for his help on explaining the observed convergence rates.

Last but not the least, I would like to thank people who have supported me throughout my education. Thank you for the guidance, strength, inspiration, skills, and making me see this adventure through to the end.

DEDICATION

This thesis is dedicated to my beloved parents, who have continuously gave me their emotional support throughout my graduate study.

Chapter 1

INTRODUCTION

We are interested in the large sample behavior of the maximum likelihood estimator (MLE) in a step threshold linear regression model. Under this model, the mean of an outcome variable is a step function of the covariate of interest. The threshold at which the jump in the mean occurs is a parameter of the model.

1.1 Convergence rate for step model estimation

Suppose that $X_1, ..., X_n$ are $i.i.d$ random variables from some distribution F, and that we are interested in estimating some parameter θ. Consider an estimator, $\hat{\theta}_n$, calculated based on $X_1, ..., X_n$. Let us assume $\tau_n(\hat{\theta}_n - \theta_0) = \mathcal{O}_{\boldsymbol{p}}(1)$, where θ_0 is the true parameter value and τ_n denotes the convergence rate of the estimator. In standard regression models, we have $\tau_n = \sqrt{n}$ for regression coefficient parameters, while in some nonstandard problems, we may have τ_n either larger or smaller than n.

Seijo et al. (2011) pointed out the least square estimator of the change point location in a step model is n^{-1} consistent to a minimizer of a two-sided compound Poisson process, while that of other regression coefficients parameters are $n^{-\frac{1}{2}}$ consistent.

According to Banerjee and McKeague (2007), the estimates of both the threshold and the slope parameters are $n^{-1/3}$ convergent in a binary decision tree approximation to a smooth regression curve. It implies that if we use step linear regression model to fit data generated by a sigmoid model, the least square estimators of step linear regression model would converge at $n^{-1/3}$ rate.

1.2 Confidence intervals methods

Efron bootstrap (referred to here as the n-bootstrap) (Efron, 1979; Efron and Tibshirani, 1993) is powerful, general and reliable in $\sqrt{n}$-convergence problems, but it often fails in

non-standard problems. For example, Seijo et al. (2011) pointed out that the least squares estimator of threshold parameter converges at a rate of n^{-1} with a nonstandard asymptotic distribution, and the confidence intervals (CIs) calculated by Efron bootstrap do not have the correct coverage. Two alternative resampling-based confidence intervals exist according to Bickel et al. (1997) about how to resample fewer than n observations: m-out-of-n bootstrap (referred to here as the m-bootstrap) and subsampling bootstrap.

1.2.1 Efron bootstrap

Suppose $(\boldsymbol{Z}_n, \boldsymbol{X}_n, \boldsymbol{Y}_n) = \{(Z_1, X_1, Y_1), ..., (Z_n, X_n, Y_n)\}$ are $i.i.d.$ random vectors from a step linear regression model $Y = \alpha_1 + \alpha_2 Z + \beta I(X > e) + \epsilon$ with parameter $\theta := (\alpha_1, \alpha_2, \beta, e)$. The procedures of Efron bootstrap for one Monte Carlo run are as follows:

Step 1: draw a random bootstrap sample $(Z_1^*, X_1^*, Y_1^*), ..., (Z_n^*, X_n^*, Y_n^*)$ with replacement from $(\boldsymbol{Z}_n, \boldsymbol{X}_n, \boldsymbol{Y}_n)$ of size n, and replicate 1000 times.

Step 2: calculate the least square estimator $\hat{\theta}_n^*$ for each bootstrap samples.

Step 3: construct sampling distribution with 1000 bootstrap statistics $\hat{\theta}_n^*$s and use it to make further inferences, such as confidence interval for θ, and so on.

To further explore the performance of the three types of bootstrap confidence intervals, we consider its following formulae. Let $\hat{\theta}_n$ denote the least square estimator of interest from the original data. Let $\hat{\theta}^*$ be the least square estimator of interest from a bootstrap sample and we took k bootstrap replicates to approximate the sampling distribution.

The symmetric percentile bootstrap confidence intervals have the form:

$$[\hat{\theta}_n - q^*, \hat{\theta}_n + q^*],$$

where q^* is the $(1-\alpha)$ quantile of $|\hat{\theta}^* - \hat{\theta}_n|$.

The percentile $100(1-\alpha)\%$ confidence interval has the form:

$$[\hat{\theta}^*_{(\frac{\alpha}{2})}, \hat{\theta}^*_{(1-\frac{\alpha}{2})}],$$

where $\hat{\theta}^*_{(\frac{\alpha}{2})}$ and $\hat{\theta}^*_{(1-\frac{\alpha}{2})}$ are the $\frac{\alpha}{2}100\%$ and $(1-\frac{\alpha}{2})100\%$ percentiles of k $\hat{\theta}^*$s.

The basic $100(1-\alpha)\%$ bootstrap confidence interval has the form:

$$[2\hat{\theta}_n - \hat{\theta}^*_{(1-\frac{\alpha}{2})}, 2\hat{\theta}_n - \hat{\theta}^*_{(\frac{\alpha}{2})}].$$

1.2.2 M-out-of-n bootstrap

The m-bootstrap is a general alternative to the n-bootstrap in nonstandard problems. According to Seijo et al. (2011), in the step linear regression model with no additional covariates, m-bootstrap is valid when model is correctly specified. In addition, the proof can be extended to allow additional covariates, but it is not quite clear how it can be extended to show the consistency of m-bootstrap when the model is misspecified.

Let $\{m_n\}_{n=1}^{\infty}$ be an arbitrary nondecreasing sequence of natural numbers such that $m_n = o(n)$ and $m_n \to \infty$. The procedure for m-bootstrap is as follows:

Step 1: for a given block size m_n, draw a random bootstrap sample (Z^*_1, X^*_1, Y^*_1), ..., $(Z^*_{m_n}, X^*_{m_n}, Y^*_{m_n})$ with replacement from $(\boldsymbol{Z}_n, \boldsymbol{X}_n, \boldsymbol{Y}_n)$ of size m_n, and replicate 1000 times.

Step 2: calculate the least square estimator $\hat{\theta}^*_{m_n}$ for each bootstrap samples.

Step 3: construct sampling distribution with 1000 bootstrap statistics $\hat{\theta}^*_{m_n}$s and use it to make further inferences, such as confidence interval for θ, and so on.

1.2.3 Subsampling bootstrap

According to Delgado et al. (2001) and Bickel and Sakov (2008), subsampling bootstrap is also a consistent bootstrap method to rectify the nonsmooth and nonstandard problems like change points. The subsampling bootstrap procedure is exactly the same as m-bootstrap except that subsampling is done without replacement. Politis and Romano (1994) showed that the m out of n bootstrap without replacement (equivalent to subsampling) with $m = o(n)$ typically works to first order both in the situations where bootstrap works and where bootstrap fails. Bickel and Sakov (2008) pointed out that subsampling is more general than the m-bootstrap since fewer assumptions are required, however, the m-bootstrap enjoys the second order properties of the n-bootstrap.

The procedure for subsampling is as follows:

Step 1: for a given block size m_n, draw a random bootstrap sample (Z_1^*, X_1^*, Y_1^*), ..., $(Z_{m_n}^*, X_{m_n}^*, Y_{m_n}^*)$ without replacement from $(\boldsymbol{Z}_n, \boldsymbol{X}_n, \boldsymbol{Y}_n)$ of size m_n, and replicate 1000 times.

Step 2: calculate the least square estimator $\hat{\theta}_{m_n}^*$ for each bootstrap samples.

Step 3: construct sampling distribution with 1000 bootstrap statistics $\hat{\theta}_{m_n}^*$s and use it to make further inferences.

For both m-bootstrap and subsampling bootstrap, the choice of the block size m_n, or m for short, has a strong effect on the coverage of the confidence interval and is critical and challenging in practical applications. Chakraborty et al. (2013) proposed a double bootstrap method for selecting m in m-bootstrap, and compared it to the method for selecting an adaptive m in Bickel and Sakov (2008). In addition, Delgado et al. (2001) proposed an algorithm for selecting m in subsampling bootstrap that is comparable to double bootstrap.

In this thesis, we try to propose a data-driven prediction rule of selecting m when the data generating models are unknown. We performed simulations to find out the best block size m satisfying that the estimated coverage probabilities are close to the nominal level 0.95 to investigate the relationship between the block size m and the samples size n, and then we will see if the rule works under different simulation setups.

The rest of the thesis is organized in the following manner. In Chapter 2, we propose a fast grid search algorithm for finding the MLE in step linear regression models. In Chapter 3, we list five data generating models that will be used in the Monte Carlo studies. In Chapter 4, we study the rates of convergence for each of the five data generating models empirically. In Chapter 5, we study the estimated coverage probabilities of the n-bootstrap confidence intervals. In Chapter 6, we study the coverage probabilities of m-bootstrap and subsampling and propose a general data-adaptive rule for selecting the block size m. In Chapter 7, we compare the performance of n-bootstrap, m-bootstrap, subsampling confidence intervals.

Chapter 2

FAST ESTIMATION OF STEP LINEAR REGRESSION MODELS

We use the approach from Fong (2019) and Elder and Fong (2019) to develop a fast grid search algorithm for step linear regression models. Consider the following model for a step linear regression model:

$$Y = \alpha_1 + \alpha_2^T z + \beta I(x > e) + \epsilon,$$

where e denotes the threshold parameter; Y is the response; x is the predictor with threshold effect; z denotes p additional predictors (a $n \times p$ matrix) and ϵ is a normally distributed error term independent of the predictors.

To estimate the step point e, maximizing the log likelihood of the step linear regression models is equivalent to minimize the sum of squares of the residuals, which is expressed as: $\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}} = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = [(\boldsymbol{I} - \boldsymbol{H_e})\boldsymbol{Y}]^T[(\boldsymbol{I} - \boldsymbol{H_e})\boldsymbol{Y}] = \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{H_e}\boldsymbol{Y}$, where $\boldsymbol{H_e} \equiv \boldsymbol{X}_e(\boldsymbol{X}_e^T\boldsymbol{X}_e)^{-1}\boldsymbol{X}_e^T$ denotes the projection matrix; $\boldsymbol{X}_e$ is the design matrix of the step linear regression models.

As $\boldsymbol{Y}^T\boldsymbol{Y}$ is independent of step point e, it is sufficient to compare $\boldsymbol{Y}^T\boldsymbol{H_e}\boldsymbol{Y}$ for a grid of e's close to $\boldsymbol{Y}^T\boldsymbol{Y}$. To speed up the computation, Fong (2019) considered the recursive relationship between design matrices $\boldsymbol{X}_{e_t}$ and $\boldsymbol{X}_{e_{t+1}}$ based on two neighboring points e_t and e_{t+1} respectively, and proposed to update $\boldsymbol{X}_{e_{t+1}}$ in terms of $\boldsymbol{X}_{e_t}$ with no need to refit the model. Furthermore, Elder and Fong (2019) divided $\boldsymbol{X}_e$ into two parts according to the independence of e, and replace the inverse matrix $(\boldsymbol{X}_e^T\boldsymbol{X}_e)^{-1}$ in $\boldsymbol{Y}^T\boldsymbol{H_e}\boldsymbol{Y}$ with a series of multiplications for accelerating the computation. In the step linear regression scenario, we consider similar formulations as follows:

Let $\boldsymbol{X}_e \equiv [\boldsymbol{1}, \boldsymbol{Z}, \boldsymbol{v}_e] \equiv [\boldsymbol{X}, \boldsymbol{v}_e]$, where $\boldsymbol{1}$ is a n-dimension vector of ones, $\boldsymbol{Z}$ is a $n \times p$ matrix, $\boldsymbol{v}_e \equiv \boldsymbol{I}(x > e)$ is a n-dimensional vector of ones and zeros ($I(x > e) = 1$ when $x > e$ and 0 otherwise).

$$(\boldsymbol{X}_e^T\boldsymbol{X}_e)^{-1} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{v}_e \\ \boldsymbol{v}_e^T\boldsymbol{X} & \boldsymbol{v}_e^T\boldsymbol{v}_e \end{bmatrix}^{-1} = c_e^{-1} \begin{bmatrix} c_e\boldsymbol{A} + \boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}_e\boldsymbol{v}_e^T\boldsymbol{X}\boldsymbol{A}^T & -\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}_e \\ -\boldsymbol{v}_e^T\boldsymbol{X}\boldsymbol{A}^T & 1 \end{bmatrix}$$

where $\boldsymbol{A} \equiv (\boldsymbol{X}^T\boldsymbol{X})^{-1}$; $c_e \equiv \boldsymbol{v}_e^T\boldsymbol{v}_e - \boldsymbol{v}_e^T\boldsymbol{H}\boldsymbol{v}_e$; $\boldsymbol{H} \equiv \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$

$$\begin{aligned}
\boldsymbol{H_e} &= \begin{bmatrix} \boldsymbol{X} & \boldsymbol{v}_e \end{bmatrix} c_e^{-1} \begin{bmatrix} c_e\boldsymbol{A} + \boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}_e\boldsymbol{v}_e^T\boldsymbol{X}\boldsymbol{A}^T & -\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}_e \\ -\boldsymbol{v}_e^T\boldsymbol{X}\boldsymbol{A}^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{X}^T \\ \boldsymbol{v}_e^T \end{bmatrix} \\
&= \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T + c_e^{-1}\left[\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}_e\boldsymbol{v}_e^T\boldsymbol{X}\boldsymbol{A}^T\boldsymbol{X}^T - \boldsymbol{v}_e\boldsymbol{v}_e^T\boldsymbol{X}\boldsymbol{A}^T\boldsymbol{X}^T - \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{v}_e\boldsymbol{v}_e^T + \boldsymbol{v}_e\boldsymbol{v}_e^T\right] \\
&= \boldsymbol{H} + c_e^{-1}\left[\left[\boldsymbol{H}\boldsymbol{v}_e\right]\left[\boldsymbol{H}\boldsymbol{v}_e\right]^T - \boldsymbol{v}_e\boldsymbol{v}_e^T\boldsymbol{H} - \boldsymbol{H}\boldsymbol{v}_e\boldsymbol{v}_e^T + \boldsymbol{v}_e\boldsymbol{v}_e^T\right] \\
&= \boldsymbol{H} + c_e^{-1}\left[\left(\boldsymbol{H} - \boldsymbol{I}\right)\boldsymbol{v}_e\right]\left[\left(\boldsymbol{H} - \boldsymbol{I}\right)\boldsymbol{v}_e\right]^T \\
\boldsymbol{Y}^T\boldsymbol{H_e}\boldsymbol{Y} &= \boldsymbol{Y}^T\left\{\boldsymbol{H} + c_e^{-1}\left[\left(\boldsymbol{H} - \boldsymbol{I}\right)\boldsymbol{v}_e\right]\left[\left(\boldsymbol{H} - \boldsymbol{I}\right)\boldsymbol{v}_e\right]^T\right\}\boldsymbol{Y} \\
&= \boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y} + c_e^{-1}\boldsymbol{Y}^T\left[\left(\boldsymbol{H} - \boldsymbol{I}\right)\boldsymbol{v}_e\right]\left[\left(\boldsymbol{H} - \boldsymbol{I}\right)\boldsymbol{v}_e\right]^T\boldsymbol{Y} \\
&= \boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y} + c_e^{-1}\left[\boldsymbol{v}_e^T\left(\boldsymbol{H} - \boldsymbol{I}\right)\boldsymbol{Y}\right]^2 \\
&\equiv \boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y} + c_e^{-1}\left(\boldsymbol{v}_e^T\boldsymbol{r}\right)^2 \quad \text{where } \boldsymbol{r} \equiv (\boldsymbol{H} - \boldsymbol{I})\boldsymbol{Y}
\end{aligned} \tag{2.1}$$

We find that the first part of equation (2.1) is independent of e whereas the second part consisting of $\boldsymbol{r}$ independent of e, c_e and $\boldsymbol{v}_e$ dependent of e. To find the recursive relationship of $\boldsymbol{Y}^T\boldsymbol{H_e}\boldsymbol{Y}$, we consider two successive (different) values of e: e_t and e_{t+1}. Suppose e_t and e_{t+1} correspond to the k^{th} and $k+1^{th}$ ascending ordered values of x, then $\boldsymbol{v}_{e_{t+1}} - \boldsymbol{v}_{e_t} = -\boldsymbol{\delta}_t$, where $\boldsymbol{\delta}_t$ ia a vector of size n with the $k+1^{th}$ entry equal to 1 and 0 otherwise. Let $\boldsymbol{B} \equiv \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-\frac{1}{2}}$. We find that:

$$\boldsymbol{v}_{e_{t+1}}^T\boldsymbol{v}_{e_{t+1}} = \boldsymbol{v}_{e_t}^T\boldsymbol{v}_{e_t} - 2\boldsymbol{\delta}_t^T\boldsymbol{v}_{e_t} + \boldsymbol{\delta}_t^T\boldsymbol{\delta}_t = \boldsymbol{v}_{e_t}^T\boldsymbol{v}_{e_t} - 1 \tag{2.2}$$

$$\boldsymbol{v}_{e_{t+1}}^T\boldsymbol{B} = \boldsymbol{v}_{e_t}^T\boldsymbol{B} - \boldsymbol{\delta}_t^T\boldsymbol{B} \tag{2.3}$$

$$\boldsymbol{v}_{e_{t+1}}^T\boldsymbol{r} = \boldsymbol{v}_{e_t}^T\boldsymbol{r} - \boldsymbol{\delta}_t^T\boldsymbol{r} \tag{2.4}$$

where $\boldsymbol{\delta}_t^T \boldsymbol{v}_e = 1$; $\boldsymbol{\delta}_t^T \boldsymbol{\delta}_t = 1$; $\boldsymbol{\delta}_t^T \boldsymbol{B}$ is the $k+1^{th}$ row vector of $\boldsymbol{B}$ and $\boldsymbol{\delta}_t^T \boldsymbol{r}$ is the $k+1^{th}$ element of $\boldsymbol{r}$. Accordingly,

$$c_{e_{t+1}}^{-1} = [\boldsymbol{v}_{e_{t+1}}^T \boldsymbol{v}_{e_{t+1}} - \boldsymbol{v}_{e_{t+1}}^T \boldsymbol{B} (\boldsymbol{v}_{e_{t+1}}^T \boldsymbol{B})^T]^{-1}$$

With the formulations above, we are capable of updating $\boldsymbol{Y}^T \boldsymbol{H}_{\boldsymbol{e}} \boldsymbol{Y}$ conditional on e_{t+1} by storing the previous values based on e_t. We write the procedures formally as follows:

Algorithm 1 The fast grid search algorithm for step linear regression models

1: Sort the samples by the ascending order of x_i
2: Compute and store the initial values: $\boldsymbol{v}_{e_1}^T \boldsymbol{v}_{e_1}$, $\boldsymbol{v}_{e_1}^T \boldsymbol{B}$, $\boldsymbol{v}_{e_1}^T \boldsymbol{r}$
3: Compute and store initial value of $\boldsymbol{Y}^T \boldsymbol{H}_{e_1} \boldsymbol{Y}$
4: For t in 1 to $n-1$: update $\boldsymbol{v}_{e_{t+1}}^T \boldsymbol{v}_{e_{t+1}}$, $\boldsymbol{v}_{e_{t+1}}^T \boldsymbol{B}$, $\boldsymbol{v}_{e_{t+1}}^T \boldsymbol{r}$ based on (2.2), (2.3), and (2.4)
5: For t in 1 to $n-1$: update $\boldsymbol{Y}^T \boldsymbol{H}_{e_{t+1}} \boldsymbol{Y}$ based on (2.1)

Chapter 3

DATA GENERATING MODELS

In this chapter, we list five different models including a step linear regression model, three sigmoid models with varying steepness, and a quadratic model. For each of the non-step model, we use step linear regression model $Y = \alpha_1 + \alpha_2 Z + \beta I(X > e)$ to fit the data simulated in large samples in order to obtain the limit values of step regression coefficient estimates.

3.1 Step linear regression model

The data generating model is:

$$Y = \alpha_1 + \alpha_2 Z + \beta I(X > e) + \epsilon. \tag{3.1}$$

Here, Y is the outcome, X is a predictor with threshold effect that follows a uniform distribution $X \sim Unif(1.5, 7.9)$, Z is an additional predictor independent of X that has a normal distribution $Z \sim N(0, \sigma = 1)$, ϵ is independent of the predictors and follows a normal distribution $\epsilon \sim N(0, \sigma = 0.3)$, α_1 is the intercept (set to be 1), α_2 is the effect of Z on Y (set to be $log(1.4)$), β represents the effect of X on Y (set to be $-log(.67)$), and the threshold parameter e is set to be 4.7.

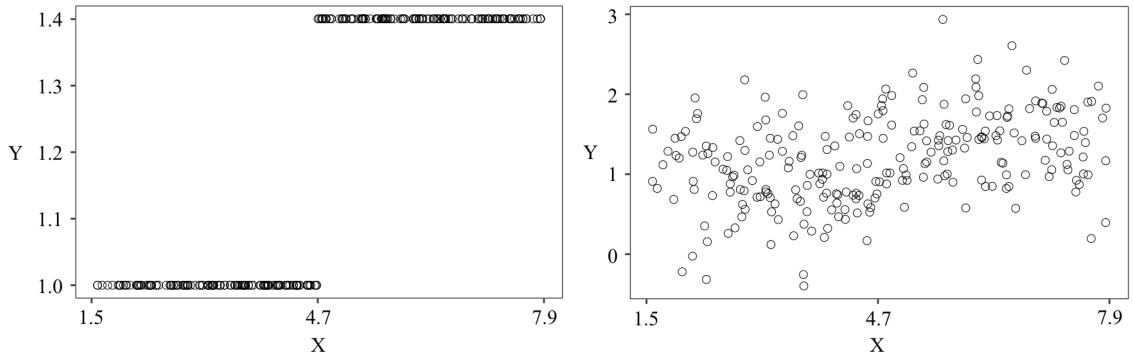

Figure 3.1: Datasets of sample size 250 simulated from the step model (3.1). Left: data simulated from a step linear regression model without the predictor Z and the error term. Right: data simulated from step linear regression model with the predictor Z and the error term.

Figure 3.1 shows a dataset of sample size 250 simulated from the step linear regression model above. As we can see in the picture, when the predictor Z and the error term ϵ are not included in the model, the scatter plot (left) shows a clear jump at the threshold. When both Z and ϵ are included, the jump is obscured.

3.2 Sigmoid model with steepness=1

The data generating model is:

$$Y = \alpha_1 + \alpha_2 Z + \beta \frac{e^{(X-4.7)}}{1 + e^{(X-4.7)}} + \epsilon. \tag{3.2}$$

Here, Y is the outcome, the predictor X follows an uniform distribution $X \sim Unif(1.5, 7.9)$, an additional predictor Z independent of X has a normal distribution $Z \sim N(0, \sigma = 1)$, the error term ϵ independent of the predictors follows a normal distribution $\epsilon \sim N(0, \sigma = 0.3)$, α_1 is the intercept (set to be 1), α_2 is the effect of Z on Y (set to be $log(1.4)$), β represents the effect of X on Y (set to be $-log(.67)$) and the steepness, which is the coefficient in front of $(X - 4.7)$ is equal to 1.

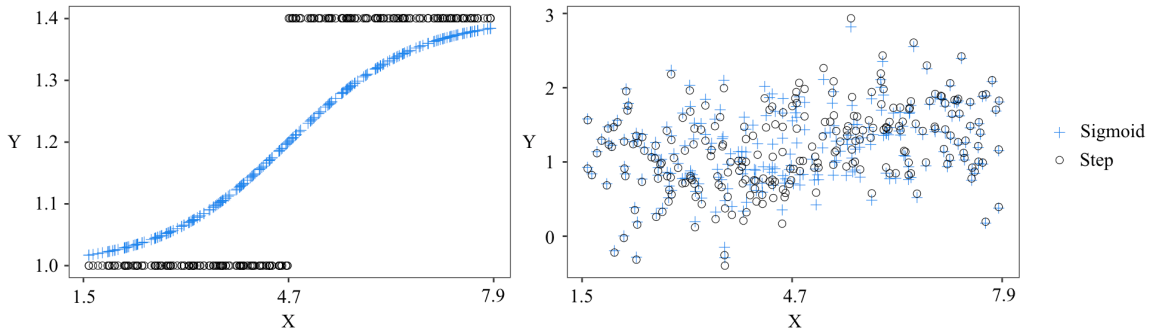

Figure 3.2: Datasets of sample size 250 simulated from sigmoid models with steepness equals to 1 (3.2). Left: data simulated without Z and the error term. Right: data simulated with Z and the error term. Blue points are the simulated data. Black points are the fitted values from step linear regression models.

Figure 3.2 shows datasets of sample size 250 simulated from the sigmoid model with steepness=1. As we can see in the picture (left), the increasing S-shaped function approximate but does not reach the "step" of step linear regression model in the left and right sides. Similarly, in the picture (right), when Z and ϵ are included, we notice that the points from sigmoid and step models do not overlap, although some of them seem to be very close.

To obtain the limit values of the MLE, we generate 10 random samples of size 10^5 and 10^6, respectively. We use step linear regression model to fit the datasets and obtain the estimates of regression coefficients and threshold parameter. As shown in Table 3.1, the estimates of β, α_1 and α_2 are almost the same between $n = 10^5$ and $n = 10^6$. Thus, we take the average of these estimates derived from the 10 random samples of size 10^6 as the limit value of step regression coefficients. As for the threshold parameter e, although there is still some variability around 4.7 in random samples of size 10^6, by symmetry of the models, we know 4.7 would be the limit.

Table 3.1: Estimated coefficients of step regression models from 10 Monte Carlo runs when data generating models are sigmoid models with steepness equals to 1 of size 10^5 and 10^6.

shape=1	e		β		α_2		α_1	
seed	n=10^5	n=10^6	n=10^5	n=10^6	n=10^5	n=10^6	n=10^5	n=10^6
1	4.841	4.615	0.236	0.238	0.336	0.336	1.087	1.078
2	4.654	4.716	0.240	0.237	0.337	0.337	1.078	1.082
3	4.571	4.710	0.242	0.238	0.337	0.336	1.072	1.082
4	4.654	4.734	0.238	0.237	0.336	0.336	1.079	1.083
5	4.639	4.751	0.241	0.237	0.336	0.337	1.079	1.084
6	4.670	4.669	0.237	0.237	0.336	0.336	1.082	1.080
7	4.599	4.673	0.238	0.237	0.338	0.336	1.077	1.080
8	4.570	4.662	0.240	0.238	0.336	0.336	1.076	1.080
9	4.780	4.645	0.236	0.237	0.335	0.336	1.084	1.079
10	4.743	4.674	0.239	0.237	0.338	0.336	1.083	1.081
Average	4.672	4.685	0.239	0.237	0.337	0.336	1.080	1.081
$\boldsymbol{\theta}_0$		4.700		0.237		0.336		1.081

3.3 Sigmoid model with steepness=5

Similarly, the data generating model is:

$$Y = \alpha_1 + \alpha_2 Z + \beta \frac{e^{5(X-4.7)}}{1 + e^{5(X-4.7)}} + \epsilon. \tag{3.3}$$

Here, Y is the outcome, the predictor X follows an uniform distribution $X \sim Unif(1.5, 7.9)$, an additional predictor Z independent of X has a normal distribution $Z \sim N(0, \sigma = 1)$, the error term ϵ independent of the predictors follows a normal distribution $\epsilon \sim N(0, \sigma = 0.3)$, α_1 is the intercept (set to be 1), α_2 is the effect of Z on Y (set to be $log(1.4)$), β represents the effect of X on Y (set to be $-log(.67)$) and the steepness is equal to 5.

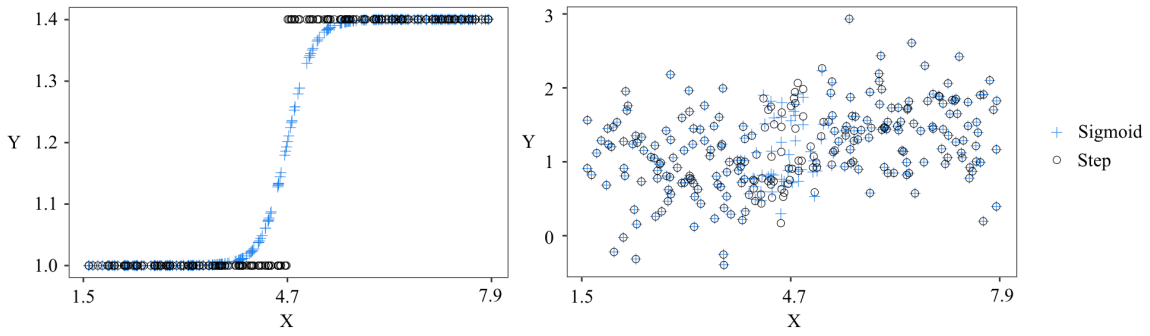

Figure 3.3: Datasets of sample size 250 simulated from sigmoid models with steepness=5 (3.3). Left: data simulated without Z and the error term. Right: data simulated with Z and the error term. Blue points are the simulated data. Black points are the fitted values from step linear regression models.

Figure 3.3 shows datasets of sample size 250 simulated from the sigmoid model with steepness=5. As we can see on the left, the increasing S-shaped function rises steeply between some interval in the middle, but is relatively flat otherwise and becomes overlapped with the step model in the left and right sides. Comparing the right hand side of Figure 3.3 with the right hand side of Figure 3.2, we see that the fitted values are closer to the data points for more of the data when the steepness is 5.

We then generate 10 random samples of size 10^5 and 10^6 respectively. Following the similar procedures as sigmoid model with steepness=1, we obtain $\boldsymbol{\theta}_0$ in Table 3.2 as the true values of the parameters of step linear regression model as the working model when the data generating model is sigmoid with steepness=5.

Table 3.2: Estimated coefficients of step regression models from 10 Monte Carlo runs when data generating models are sigmoid models with steepness equals to 5 of size 10^5 and 10^6.

shape=5	e		β		α_2		α_1	
seed	n=10^5	n=10^6	n=10^5	n=10^6	n=10^5	n=10^6	n=10^5	n=10^6
1	4.687	4.706	0.363	0.366	0.336	0.336	1.017	1.017
2	4.654	4.716	0.367	0.366	0.337	0.337	1.014	1.018
3	4.668	4.710	0.370	0.366	0.338	0.336	1.011	1.018
4	4.674	4.711	0.366	0.366	0.336	0.336	1.015	1.018
5	4.656	4.701	0.369	0.366	0.336	0.337	1.015	1.017
6	4.670	4.706	0.366	0.366	0.336	0.336	1.017	1.018
7	4.657	4.681	0.366	0.366	0.338	0.336	1.015	1.016
8	4.697	4.691	0.368	0.366	0.336	0.336	1.016	1.016
9	4.693	4.713	0.365	0.366	0.335	0.336	1.016	1.018
10	4.698	4.706	0.367	0.366	0.338	0.336	1.017	1.018
Average	4.675	4.704	0.367	0.366	0.340	0.336	1.015	1.017
$\boldsymbol{\theta}_0$		4.700		0.366		0.336		1.017

3.4 Sigmoid model with steepness=15

Similarly, the data generating model is:

$$Y = \alpha_1 + \alpha_2 Z + \beta \frac{e^{15(X-4.7)}}{1 + e^{15(X-4.7)}} + \epsilon. \tag{3.4}$$

Here, Y is the outcome, the predictor X follows an uniform distribution $X \sim Unif(1.5, 7.9)$, an additional predictor Z independent of X has a normal distribution $Z \sim N(0, \sigma = 1)$, the error term ϵ independent of the predictors follows a normal distribution $\epsilon \sim N(0, \sigma = 0.3)$, α_1 is the intercept (set to be 1), α_2 is the effect of Z on Y (set to be $log(1.4)$), β represents the effect of X on Y (set to be $-log(.67)$) and the steepness is equal to 15.

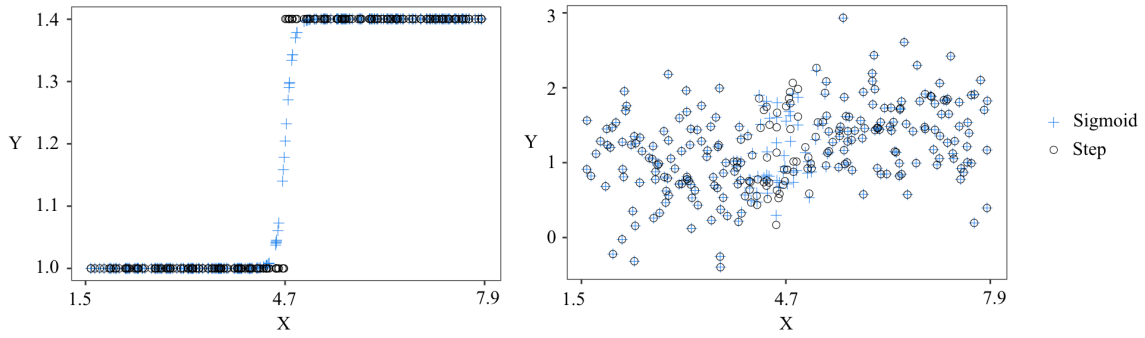

Figure 3.4: Datasets of sample size 250 simulated from sigmoid models with steepness=15 (3.4). Left: data simulated without Z and the error term. Right: data simulated with Z and the error term. Blue points are the simulated data. Black points are the fitted values from step linear regression models.

Figure 3.4 shows datasets of sample size 250 simulated from sigmoid model with steepness=15. As we can see on the left, the increasing S-shaped function rises steeply between interval $(4.5, 4.9)$ of X, but is flat otherwise and becomes overlapped with the step model in the left and right sides. Comparing the right hand side of Figure 3.4 with the right hand side of Figure 3.4, we see that the fitted values are closer to the data points for more of the data when the steepness is 15.

We then generate 10 random samples of size 10^5 and 10^6 respectively. Following the similar procedures as sigmoid models with steepness=1 and 5, we obtain $\boldsymbol{\theta}_0$ in Table 3.3 as the true values of the parameters of step linear regression model as the working model when the data generating model is sigmoid with steepness=15.

Table 3.3: Estimated coefficients of step regression models from 10 Monte Carlo runs when data generating models are sigmoid models with steepness equals to 15 of size 10^5 and 10^6.

shape=15	e		β		α_2		α_1	
seed	n=10^5	n=10^6	n=10^5	n=10^6	n=10^5	n=10^6	n=10^5	n=10^6
1	4.687	4.706	0.386	0.389	0.336	0.336	1.005	1.006
2	4.698	4.694	0.390	0.389	0.337	0.337	1.005	1.005
3	4.704	4.702	0.393	0.389	0.338	0.336	1.002	1.006
4	4.682	4.698	0.389	0.389	0.336	0.336	1.004	1.006
5	4.677	4.701	0.391	0.389	0.336	0.337	1.005	1.006
6	4.708	4.701	0.389	0.389	0.335	0.336	1.008	1.006
7	4.694	4.695	0.389	0.389	0.337	0.336	1.006	1.005
8	4.697	4.694	0.391	0.389	0.336	0.336	1.005	1.005
9	4.694	4.697	0.388	0.389	0.335	0.336	1.005	1.005
10	4.698	4.706	0.390	0.389	0.338	0.336	1.006	1.006
Average	4.694	4.699	0.390	0.389	0.336	0.336	1.005	1.006
$\boldsymbol{\theta}_0$		4.700		0.389		0.336		1.006

3.5 Quadratic Model

The data generating model is:

$$Y = \alpha_1 + \alpha_2 Z + \gamma_1 X + \gamma_2 X^2 + \epsilon, \tag{3.5}$$

where Y is the outcome, X is the predictor that follows a uniform distribution $X \sim Unif(1.9, 7.5)$, Z is an additional predictor independent of X that has a normal distribution $Z \sim N(0, \sigma = 1)$, ϵ is the error term independent of the predictors that follows a normal distribution $\epsilon \sim N(0, \sigma = 0.3)$, α_1 is the intercept (set to be -1), α_2 is the effect of Z on Y (set to be $log(1.4)$), γ_1 is the effect of X on Y (set to be -1) and γ_2 represents the effect of X^2 on Y (set to be 0.3).

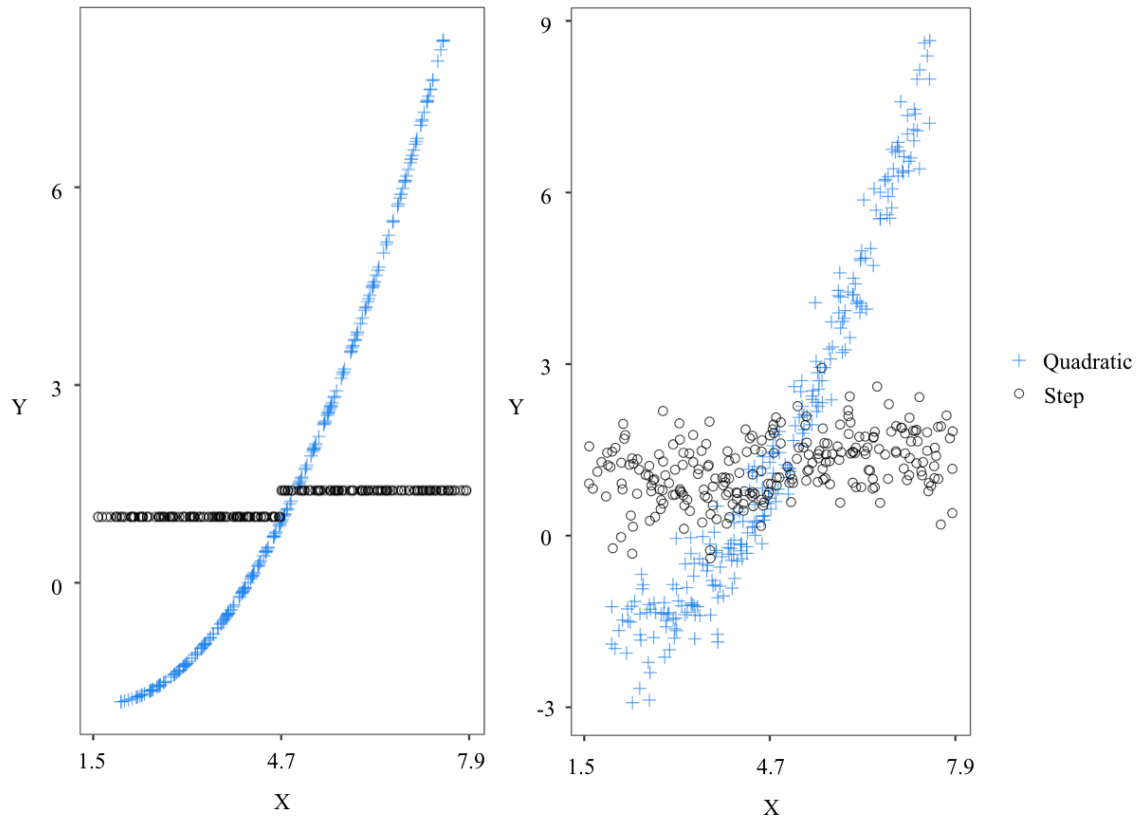

Figure 3.5: Datasets of sample size 250 simulated from the quadratic model (3.5). Left: data simulated without Z and and the error term. Right: data simulated with Z and the error term. Blue points are the simulated data. Black points are the fitted values from step linear regression models.

Figure 3.5 shows datasets of sample size 250 generated from the quadratic model. As we see in the figure, unlike sigmoid models in Figure 3.2 and 3.3, the quadratic model has a completely different trend from the step models.

Again, we generate 10 random samples of size 10^5 and 10^6 respectively and obtain the $\boldsymbol{\theta}_0$ in Table 3.4 as the true values of the parameters of step linear regression model as the working model when the data generating model is quadratic model.

Table 3.4: Estimated coefficients of step regression models from 10 Monte Carlo runs when simulating quadratic models of size 10^5 and 10^6 respectively.

quadratic	e		β		α_2		α_1	
seed	n=10^5	n=10^6	n=10^5	n=10^6	n=10^5	n=10^6	n=10^5	n=10^6
1	5.437	5.437	5.516	5.506	0.336	0.337	-0.322	-0.318
2	5.438	5.450	5.522	5.518	0.331	0.335	-0.321	-0.309
3	5.435	5.436	5.509	5.510	0.337	0.337	-0.322	-0.320
4	5.438	5.441	5.507	5.508	0.331	0.337	-0.320	-0.317
5	5.418	5.448	5.533	5.522	0.338	0.337	-0.345	-0.314
6	5.461	5.452	5.530	5.519	0.340	0.335	-0.300	-0.307
7	5.463	5.431	5.526	5.505	0.336	0.337	-0.300	-0.323
8	5.425	5.441	5.506	5.515	0.339	0.336	-0.327	-0.318
9	5.430	5.441	5.489	5.510	0.337	0.336	-0.321	-0.315
10	5.428	5.436	5.503	5.508	0.339	0.334	-0.333	-0.319
Average	5.437	5.441	5.514	5.512	0.336	0.336	-0.321	-0.316
$\boldsymbol{\theta}_0$		5.441		5.512		0.336		-0.316

Chapter 4

EMPIRICAL STUDY OF RATE OF CONVERGENCE

In the previous chapter, we simulated data from five data generating models, and obtained the limit values of step regression coefficients estimates when the data generating models do not match the step regression model. In this section, we are interested in estimating the convergence rates of threshold parameter e and regression coefficients β, α_2 and α_1 through Monte Carlo studies.

To estimate the rate of convergence, we fit the Monte Carlo standard deviation σ_n to the following model:

$$\sigma_n = a \times n^b, \tag{4.1}$$

where n is the size of data generated, a is the effect of n^b on σ_n and b denotes the convergence rate.

For each data simulating model, we generated data of size 1000, 2000, 4000, 8000, 16000, 32000, and 64000. And for each specific size, we conducted 10000 Monte Carlo runs and obtained the corresponding Monte Carlo standard deviation estimate. Thus, for each data generating model, we have seven data points of $(\hat{\sigma}_n, n)$ for each step model parameters shown in Table 4.1, and estimating b is of interest.

We consider the following two methods to estimate the rates of convergence for the threshold regression coefficients and threshold parameter e of step linear regression models:

(1) Take log of both sides of model 4.1, and then fit a linear regression model using the seven data points for a data generating model. (Referred to here as log-linear-7).

(2) Follow the same steps as (1) except for using only the four data points of larger sizes, as we are interested in the large sample behavior of convergence rate. (Referred to here as log-linear-4).

Table 4.1: Monte Carlo standard deviation of threshold parameter and regression coefficients from 10000 Monte Carlo runs for each of the 5 data generating models. e: threshold, β: slope of $I(x > e)$, α_2: slope of z, α_1: intercept.

	Step model				Sigmoid model (steepness=1)				Sigmoid model (steepness=5)				Sigmoid model (steepness=15)				Quadratic model			
n	e	β	α_2	α_1	e	β	α_2	α_1	e	β	α_2	α_1	e	β	α_2	α_1	e	β	α_2	α_1
1000	0.036	0.019	0.010	0.014	0.503	0.020	0.010	0.025	0.146	0.019	0.010	0.016	0.075	0.019	0.010	0.014	0.081	0.126	0.054	0.090
2000	0.018	0.013	0.007	0.010	0.392	0.014	0.007	0.018	0.112	0.014	0.007	0.012	0.054	0.013	0.007	0.010	0.062	0.089	0.037	0.067
4000	0.009	0.009	0.005	0.007	0.311	0.010	0.005	0.014	0.088	0.010	0.005	0.008	0.042	0.009	0.005	0.007	0.048	0.066	0.025	0.050
8000	0.004	0.007	0.003	0.005	0.246	0.007	0.003	0.011	0.069	0.007	0.003	0.006	0.032	0.007	0.003	0.005	0.037	0.047	0.017	0.037
16000	0.002	0.005	0.002	0.003	0.196	0.005	0.002	0.008	0.055	0.006	0.002	0.005	0.025	0.005	0.002	0.004	0.029	0.033	0.012	0.028
32000	0.001	0.003	0.002	0.002	0.156	0.003	0.002	0.006	0.043	0.003	0.002	0.003	0.020	0.003	0.002	0.003	0.022	0.024	0.008	0.021
64000	0.001	0.002	0.001	0.002	0.124	0.002	0.001	0.005	0.034	0.002	0.001	0.003	0.016	0.002	0.001	0.002	0.018	0.018	0.006	0.016

Table 4.2: Estimated convergence rates and its confidence intervals calculated by 2 different methods for each of the 5 data generating models. e: threshold, β: slope of $I(x > e)$, α_2: slope of z, α_1: intercept.

	Estimated rate of convergence (95% Confidence interval)			
	e	β	α_2	α_1
Step linear regression model				
log-linear-7	-1.00 (-1.01,-0.99)	-0.50 (-0.50,-0.50)	-0.50 (-0.50,-0.50)	-0.50 (-0.51,-0.50)
log-linear-4	-1.00 (-1.02,-0.98)	-0.50 (-0.50,-0.50)	-0.50 (-0.51,-0.50)	-0.50 (-0.51,-0.50)
Sigmoid model with steepness=1				
log-linear-7	-0.34 (-0.34,-0.33)	-0.51 (-0.51,-0.51)	-0.50 (-0.50,-0.50)	-0.38 (-0.40,-0.37)
log-linear-4	-0.33 (-0.33,-0.33)	-0.50 (-0.51,-0.50)	-0.50 (-0.50,-0.50)	-0.36 (-0.37,-0.36)
Sigmoid model with steepness=5				
log-linear-7	-0.35 (-0.36,-0.34)	-0.50 (-0.50,-0.50)	-0.50 (-0.50,-0.50)	-0.44 (-0.45,-0.43)
log-linear-4	-0.34 (-0.36,-0.33)	-0.50 (-0.50,-0.50)	-0.50 (-0.51,-0.49)	-0.42 (-0.43,-0.41)
Sigmoid model with steepness=15				
log-linear-7	-0.37 (-0.39,-0.35)	-0.50 (-0.50,-0.50)	-0.50 (-0.50,-0.50)	-0.48 (-0.48,-0.47)
log-linear-4	-0.35 (-0.36,-0.33)	-0.50 (-0.50,-0.49)	-0.50 (-0.50,-0.50)	-0.47 (-0.48,-0.46)
Quadratic model				
log-linear-7	-0.37 (-0.37,-0.36)	-0.48 (-0.49,-0.47)	-0.53 (-0.54,-0.52)	-0.41 (-0.42,-0.41)
log-linear-4	-0.36 (-0.38,-0.34)	-0.47 (-0.52,-0.43)	-0.52 (-0.53,-0.51)	-0.40 (-0.43,-0.38)

Table 4.2 shows the estimated convergences rates and its corresponding confidence intervals calculated for the four parameters of step linear regression model in each data generating model. Figure 4.1, 4.2, 4.3, 4.4 and 4.5 depict the linear regression model fitted after

log-transformation for each data generating model, respectively.

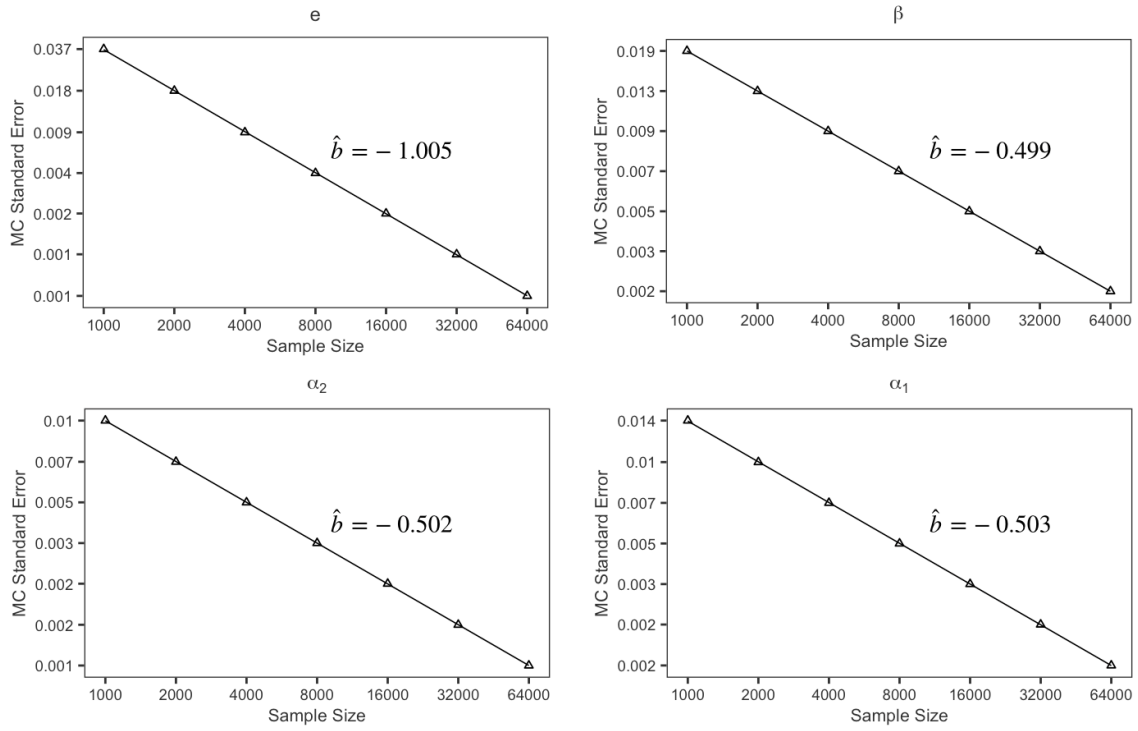

Figure 4.1: Estimated convergence rate for the 4 coefficients of step regression model when the true model is the step linear regression model

When the data is generated from the step model, as we can see in the Figure 4.1, after log-transformation, the linear regression model fits the seven data points well and there is no discernible second order trends (test p values not shown). For threshold parameter e, the slope $\hat{b} = -1.00$ by the log-linear-4 method, and the confidence interval is $(-1.02, -0.98)$, which includes -1. Thus, our results validate that when model is correctly specified, the least square estimator of the threshold parameter e converge at a n-rate. For the slope parameters β, α_1 and α_2, the estimated convergence rates are around -0.5, thus validating the theoretical results of Seijo et al. (2011).

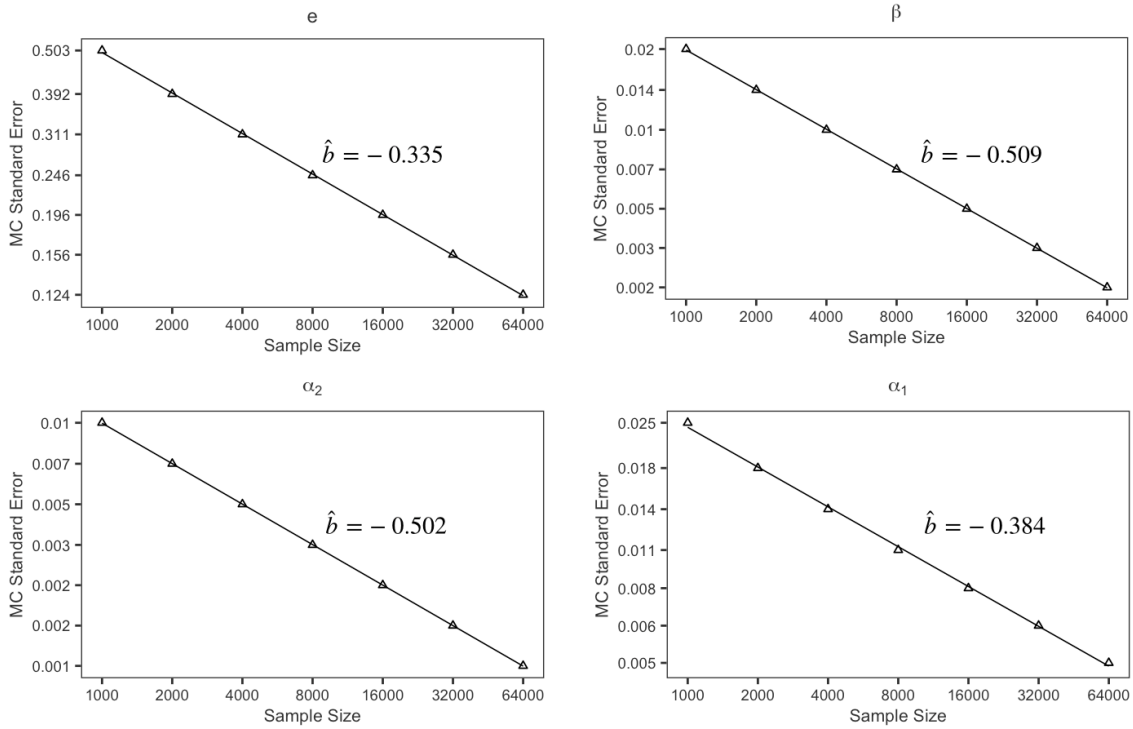

Figure 4.2: Estimated convergence rate for the 4 coefficients estimates of step regression model when the true model is a sigmoid model with steepness equals to 1.

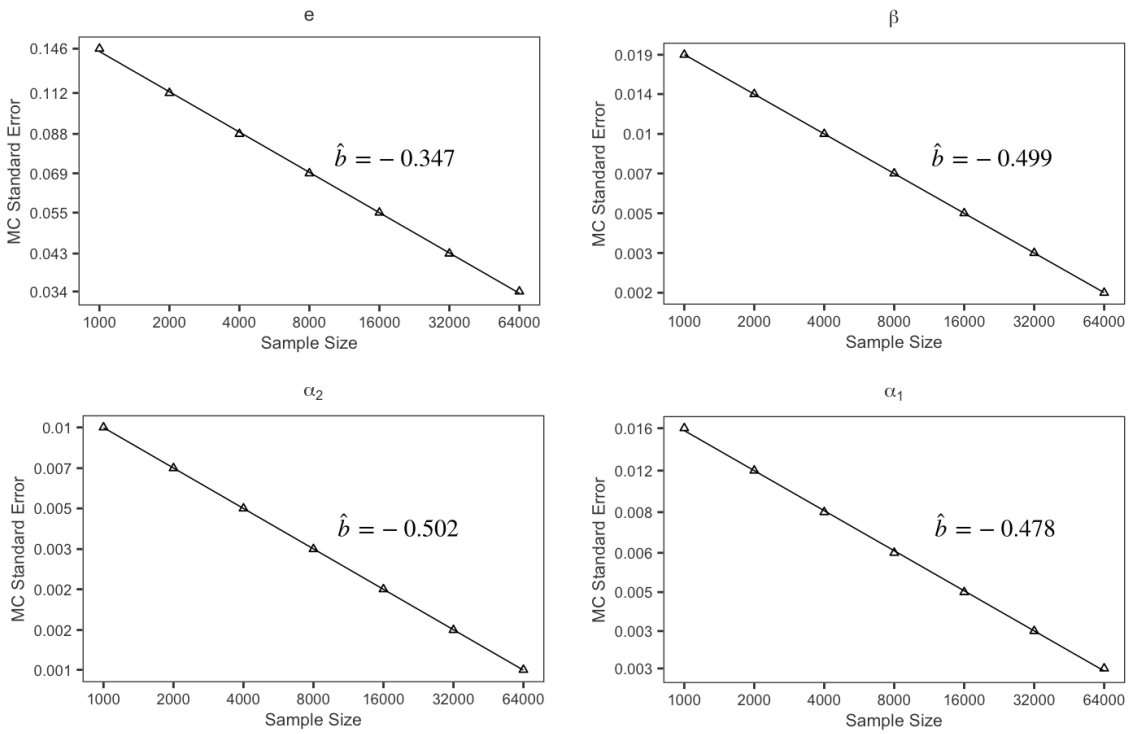

Figure 4.3: Estimated convergence rate for the 4 coefficients estimates of step regression model when the true model is a sigmoid model with steepness equals to 5.

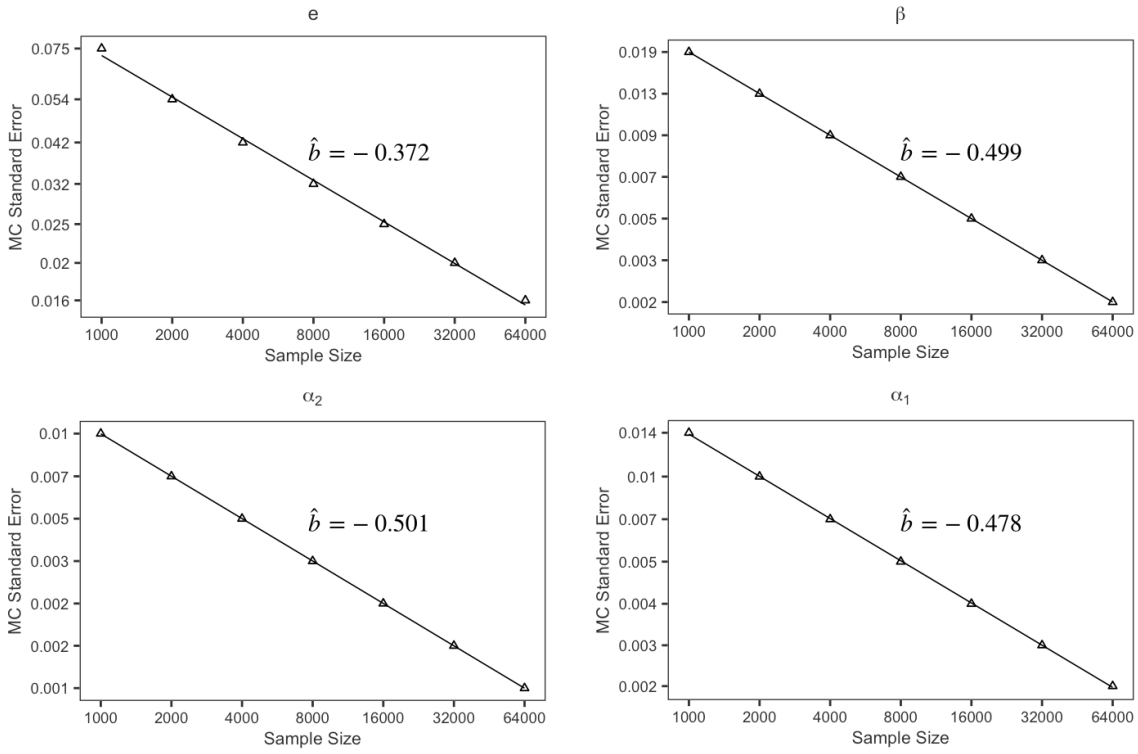

Figure 4.4: Estimated convergence rate for the 4 coefficients estimates of step regression model when the true model is a sigmoid model with steepness equals to 15.

Figure 4.2, 4.3 and 4.4 show how the convergence rates are estimated when the data generating model is sigmoid with steepness=1, 5 and 15, respectively. The linear fits appear satisfactory. By the log-linear-4 method, the estimated convergence rate for e is -0.33 (95% CI -0.33, -0.33) when steepness=1, -0.34 (95% CI -0.36, -0.33) when steepness=5, and -0.35 (95% CI -0.36, -0.33) when steepness=15, which validates the theoretical results of Banerjee and McKeague (2007). In both simulations, the estimated convergence rates for β and α_2 are around -0.5. The estimated convergence rates for α_1 are between -0.5 and -0.33, though we expect larger sample sizes will cause these rates estimates to pull towards -0.33.

It must be noted that according to Banerjee and McKeague (2007), $\alpha_1 + \beta$ converges at a $\sqrt[3]{n}$ rate, however, our simulation results show that β is $\sqrt{n}$-convergence. This is because in the particular case of the logistic functions and the uniform distribution of X on $(0, 1)$, c_1 equals to c_2 in Theorem 2.1 in Banerjee and McKeague (2007), and hence the limiting

distribution of β, which equals to $n^{1/3}((\hat{\beta}_l - \beta_l^0) - (\hat{\beta}_u - \beta_u^0)) \to (c_1 - c_2)\,\mathrm{argmax}_t\, Q(t)$ in Banerjee and McKeague (2007), is degenerate at 0. Thus in this special case there is room for an additional factor, and the simulation results show that $\hat{\beta}$ converges at the $n^{1/2}$ rate in this model.

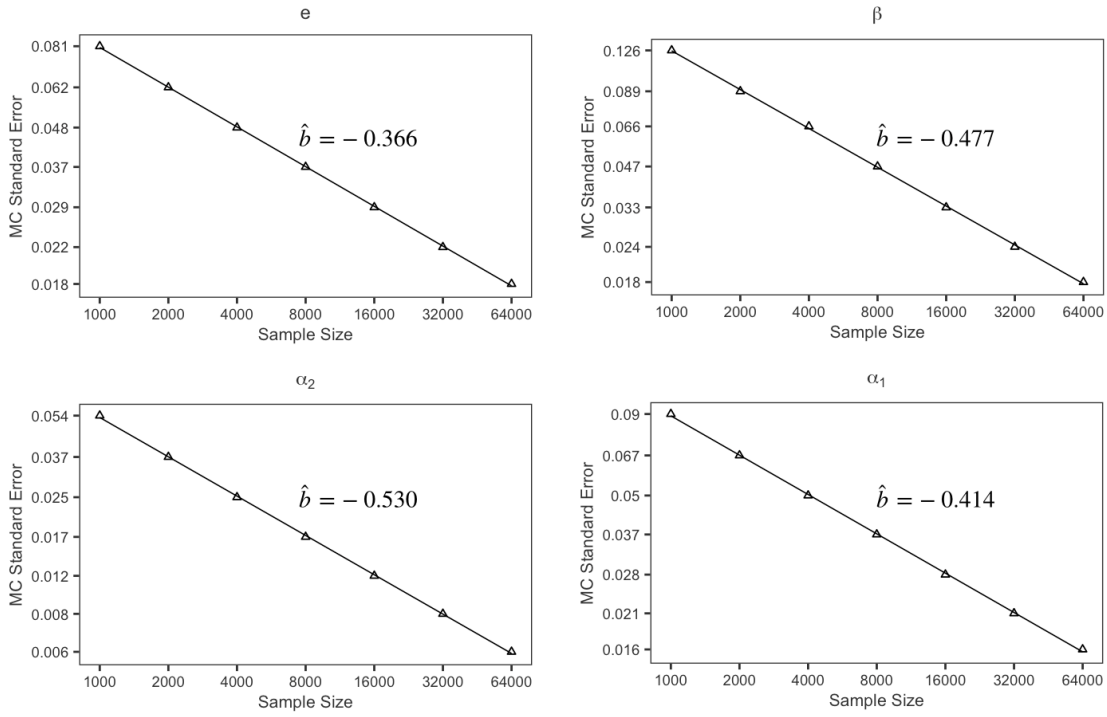

Figure 4.5: Estimated convergence rate for the 4 coefficients estimates of step regression model when the true model is a quadratic model.

In Figure 4.5, the data generating model is quadratic model. The linear regression model again fits the data well and no signs of second order effect are shown. For the threshold parameter e, the estimated convergence rate is $-0.37(-0.37, -0.36)$ by log-linear-4 method. The estimated convergence rates vary for β, α_2 and α_1, though we expect larger sample sizes will cause these rates estimates to pull towards -0.33 according to theory.

In conclusion, our numerical results suggest that when the model is misspecified, the estimate of e and α_1 converge at $n^{1/3}$ but the convergence rate of estimate of β, the size of the jump at the threshold, varies according to the true data generating model. These

results are consistent with the theoretical results of Banerjee and McKeague (2007), which claim $n^{-1/3}$ convergence for e, α_1, and $\alpha_1 + \beta$. The estimated $\sqrt{n}$ rates of convergence for β under some simulation models come as a surprise and warrants further study.

Chapter 5

COVERAGE OF EFRON BOOTSTRAP CONFIDENCE INTERVALS IN STEP LINEAR REGRESSION MODELS

In the previous chapter, our simulation results showed that the least square estimator of threshold parameter e in the step linear regression model has a nonstandard convergence rate. In this section, we conduct Monte Carlo experiments to assess the coverage probabilities for the threshold parameter e and regression coefficient estimates α_1, α_2, β in step linear regression models by Efron bootstrap.

We first generate data from five different models including a step linear regression model, three sigmoid models with steepness equal to 1, 5 and 15 and a quadratic model. Then we use step linear regression model $Y = \alpha_1 + \alpha_2 Z + \beta I(X > e)$ to fit the data simulated. For each of the five data generating models, we considered 10000 Monte Carlo runs to generate 10000 random samples of size $n = 1000, 2000, 4000, 8000, 16000, 32000, 64000$, respectively. And for each sample, we took 1000 bootstrap replicates to approximate the sampling distributions of estimates and calculate the corresponding percentile, basic and symmetric confidence intervals by Efron bootstrap.

5.1 Model Correctly Specified

When the data generating model and the working model are both step linear regression models, we investigate the performance of percentile, basic and symmetric confidence intervals derived by Efron bootstrap and its corresponding estimated coverage probabilities in simulation studies.

5.1.1 Step linear regression model

We use a step linear regression model to fit the data generated by the step linear regression model 3.1 (Figure 3.1) of size $1000, 2000, 4000, 8000, 16000, 32000, 64000$ in 10000 Monte

Carlo runs, respectively. Table 5.1 shows the mean step model estimates, bias, Monte Carlo standard deviation, estimated coverage probabilities and mean length of three types of n-bootstrap confidence intervals.

Table 5.1: Simulation results from 10,000 Monte Carlo runs when the true model is a step linear regression model. Estimate, bias, Monte Carlo standard deviation, coverage probabilities and mean width of the 3 types of bootstrap confidence intervals (percentile, basic and symmetric) of the parameter estimates are given from Efron bootstrap by fastgrid algorithm.

	Monte Carlo results		Coverage(Width)		
n	Estimate(Bias)	MC SD	Percentile	Basic	Symmetric
e					
1000	4.696 (-0.004)	0.036	0.827 (0.116)	0.807 (0.116)	0.963 (0.143)
2000	4.698 (-0.002)	0.018	0.820 (0.058)	0.812 (0.058)	0.963 (0.071)
4000	4.699 (-0.001)	0.009	0.815 (0.029)	0.813 (0.029)	0.964 (0.035)
8000	4.699 (-0.001)	0.004	0.816 (0.014)	0.806 (0.014)	0.962 (0.018)
16000	4.700 (0.000)	0.002	0.820 (0.007)	0.805 (0.007)	0.965 (0.009)
32000	4.700 (0.000)	0.001	0.815 (0.004)	0.814 (0.004)	0.968 (0.004)
64000	4.700 (0.000)	0.001	0.814 (0.002)	0.806 (0.002)	0.964 (0.002)
β					
1000	0.401 (0.001)	0.019	0.943 (0.074)	0.943 (0.074)	0.945 (0.074)
2000	0.401 (0.001)	0.013	0.948 (0.052)	0.947 (0.052)	0.949 (0.052)
4000	0.401 (0.001)	0.009	0.951 (0.037)	0.950 (0.037)	0.952 (0.037)
8000	0.401 (0.001)	0.007	0.947 (0.026)	0.948 (0.026)	0.948 (0.026)
16000	0.401 (0.001)	0.005	0.953 (0.019)	0.953 (0.019)	0.954 (0.019)
32000	0.400 (0.000)	0.003	0.947 (0.013)	0.946 (0.013)	0.947 (0.013)
64000	0.400 (0.000)	0.002	0.953 (0.009)	0.953 (0.009)	0.954 (0.009)
α_2					
1000	0.337 (0.001)	0.010	0.951 (0.037)	0.950 (0.037)	0.950 (0.037)
2000	0.336 (0.000)	0.007	0.947 (0.026)	0.949 (0.026)	0.949 (0.026)
4000	0.337 (0.001)	0.005	0.949 (0.019)	0.947 (0.019)	0.949 (0.019)
8000	0.336 (0.000)	0.003	0.951 (0.013)	0.950 (0.013)	0.954 (0.013)
16000	0.336 (0.000)	0.002	0.947 (0.009)	0.948 (0.009)	0.948 (0.009)
32000	0.336 (0.000)	0.002	0.949 (0.007)	0.949 (0.007)	0.950 (0.007)
64000	0.336 (0.000)	0.001	0.949 (0.005)	0.949 (0.005)	0.950 (0.005)
α_1					
1000	1.000 (0.000)	0.014	0.947 (0.053)	0.947 (0.053)	0.949 (0.053)
2000	1.000 (0.000)	0.010	0.949 (0.037)	0.949 (0.037)	0.951 (0.037)
4000	1.000 (0.000)	0.007	0.948 (0.026)	0.950 (0.026)	0.950 (0.026)
8000	1.000 (0.000)	0.005	0.950 (0.019)	0.951 (0.019)	0.952 (0.019)
16000	1.000 (0.000)	0.003	0.949 (0.013)	0.950 (0.013)	0.951 (0.013)
32000	1.000 (0.000)	0.002	0.950 (0.009)	0.950 (0.009)	0.952 (0.009)
64000	1.000 (0.000)	0.002	0.949 (0.007)	0.948 (0.007)	0.949 (0.007)

As shown by Table 5.1, the percentile, basic and symmetric bootstrap confidence intervals have coverage probabilities close to nominal level 0.95 for regression coefficient estimates β, α_1 and α_2. As for threshold parameter e, both the percentile and basic bootstrap confidence intervals suffer from an undercoverage, which do not improve with larger sample sizes. This provides empirical evidence that Efron bootstrap does not work for the n^{-1}-convergent $\hat{e}$. The symmetric bootstrap confidence interval gives coverage probabilities above the nominal level 0.95, with wider confidence intervals, but there is no theoretical basis for why this should work in general.

To get more intuition about the failure of Efron bootstrap, we divide the bootstrap distributions of e from 10^4 MC replicates into eight categories based on the coverage of the three types of CI. Table 5.2 describes the proportion of the eight categories and the distribution of skewness within each category. Here, we use Fisher-Pearson's coefficient to measure skewness. We regards those calculated less than -0.5, greater than 0.5, and between -0.5 and 0.5 as left-skewed, right-skewed and symmetric (middle), respectively.

Table 5.2: 10^4 samples of size 2000 divided into 8 categories based on the coverage scenarios of three types of bootstrap confidence intervals for threshold parameter e. Within each category, skewness of the sampling distributions is tabulated.

Covered?				Skewed?		
Percentile	Basic	Symmetric	Proportion (%)	Left-skewed (%)	Middle (%)	Right-skewed (%)
no	no	no	2	23	27	50
yes	no	no	1	28	13	59
no	yes	no	1	72	3	25
yes	yes	no	0	NA	NA	NA
no	no	yes	0	NA	NA	NA
yes	no	yes	16	45	20	35
no	yes	yes	15	68	11	21
yes	yes	yes	65	32	22	45
			100	40	20	40

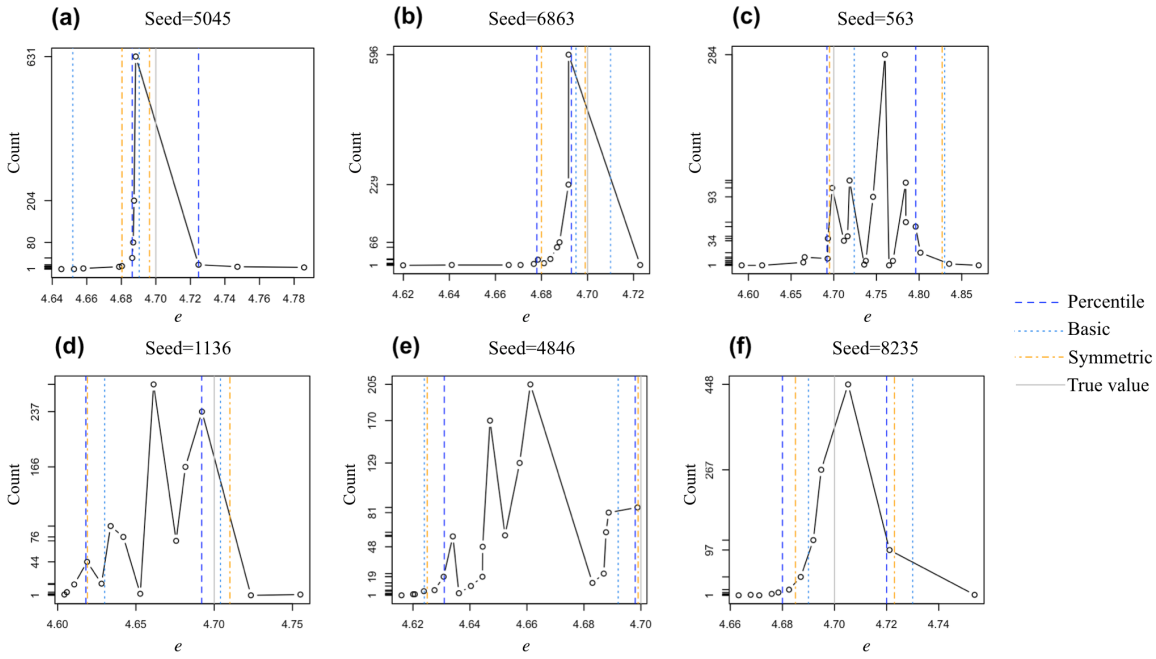

Figure 5.1: Bootstrap distributions of e from 6 datasets of size 2000 when the data generating model is a step linear regression model. Top left: right-skewed; percentile cover; basic not cover; symmetric not cover. Top middle: left-skewed; percentile not cover; basic cover; percentile not cover. Top right: percentile cover; basic not cover; symmetric cover. Bottom left: percentile not cover; basic cover; symmetric cover. Bottom middle: percentile not cover; basic not cover; symmetric not cover. Bottom right: percentile cover; basic cover; symmetric cover.

The table shows that

- Overall the majority (80%) of the sampling distributions of the threshold parameter are skewed, either left-skewed (40%) or right-skewed (40%).

- When the percentile CI covers but the basic and symmetric CIs do not, most (59%) bootstrap distributions are also right-skewed (Figure 5.1(b)).

- When the basic CI covers but the percentile CI does not, most (68%) of the bootstrap distributions are left-skewed (Figure 5.1(b)).

In addition, the results show that

- When the symmetric CI does not cover the truth, almost all (92%) $\hat{e}$ is less than the truth 4.7.

- When the basic CI does not cover the truth, most (56%) of $\hat{e}$ are greater than 4.7 (Figure 5.1(c)).

- When the percentile CI does not cover the truth, almost all (99%) $\hat{e}$ are less than 4.7 (Figure 5.1(d)).

Finally, Figure 5.1 (e) and (f) show the typical bootstrap distributions of e when none of the bootstrap CIs cover the truth and all of the 3 bootstrap CIs cover the truth, respectively.

Based on Table 5.1 and Figure 5.2, we find the majority of the sampling distributions of the threshold parameter are skewed. The percentile CI tends to cover the truth when the bootstrap distributions are right-skewed whereas the basic CI tends to perform better when the bootstrap distributions are left-skewed. Note that the symmetric CI perform well when the bootstrap distributions are either right-skewed or left-skewed, thus has the coverage probabilities closest to the nominal level 0.95.

5.2 Model misspecified

In this section, we examine the performance of percentile, basic and symmetric confidence intervals calculated by Efron bootstrap when the working model is a step linear regression model whereas the true data generating models are not.

5.2.1 Sigmoid model with steepness=1

We use a step linear regression model to fit the data generated by the sigmoid model 3.2 (Figure 3.2 right) of size $1000, 2000, 4000, 8000, 16000, 32000, 64000$, respectively. Table 5.3 shows the mean step model estimates, bias with respect to θ_0 defined in Table 3.1, Monte Carlo standard deviation, estimated coverage probabilities and mean length of the three types of confidence intervals for the regression coefficients β, α_2, α_1 and the threshold parameter e calculated by Efron bootstrap.

Table 5.3: Simulation results from 10,000 Monte Carlo runs when the true model is a sigmoid model with steepness equals to 1. Estimate, %bias, Monte Carlo standard deviation, coverage probabilities and mean width of the 3 types of n-bootstrap confidence intervals (percentile, basic and symmetric) of the parameter estimates are given.

	Monte Carlo results		Coverage(Width)		
n	Estimate(Bias)	MC SD	Percentile	Basic	Symmetric
e					
1000	4.699 (-0.001)	0.503	0.972 (1.567)	0.656 (1.567)	0.974 (1.971)
2000	4.697 (-0.003)	0.392	0.973 (1.225)	0.653 (1.225)	0.969 (1.536)
4000	4.707 (0.007)	0.311	0.968 (0.966)	0.645 (0.966)	0.968 (1.220)
8000	4.699 (-0.001)	0.246	0.972 (0.767)	0.657 (0.767)	0.970 (0.963)
16000	4.699 (-0.001)	0.196	0.972 (0.609)	0.647 (0.609)	0.972 (0.768)
32000	4.700 (0.000)	0.156	0.970 (0.485)	0.652 (0.485)	0.972 (0.611)
64000	4.699 (-0.001)	0.124	0.971 (0.386)	0.652 (0.386)	0.969 (0.485)
β					
1000	0.249 (0.012)	0.020	0.844 (0.074)	0.923 (0.074)	0.921 (0.079)
2000	0.245 (0.008)	0.014	0.865 (0.052)	0.931 (0.052)	0.927 (0.055)
4000	0.242 (0.005)	0.010	0.881 (0.037)	0.935 (0.037)	0.932 (0.039)
8000	0.240 (0.003)	0.007	0.893 (0.026)	0.941 (0.026)	0.935 (0.027)
16000	0.239 (0.002)	0.005	0.906 (0.019)	0.945 (0.019)	0.943 (0.019)
32000	0.238 (0.001)	0.003	0.914 (0.013)	0.943 (0.013)	0.942 (0.014)
64000	0.238 (0.001)	0.002	0.921 (0.009)	0.945 (0.009)	0.945 (0.010)
α_2					
1000	0.337 (0.001)	0.010	0.951 (0.038)	0.945 (0.038)	0.952 (0.038)
2000	0.336 (0.000)	0.007	0.951 (0.027)	0.945 (0.027)	0.951 (0.027)
4000	0.337 (0.001)	0.005	0.951 (0.019)	0.947 (0.019)	0.951 (0.019)
8000	0.336 (0.000)	0.003	0.953 (0.013)	0.949 (0.013)	0.954 (0.013)
16000	0.336 (0.000)	0.002	0.948 (0.009)	0.945 (0.009)	0.949 (0.009)
32000	0.336 (0.000)	0.002	0.950 (0.007)	0.948 (0.007)	0.950 (0.007)
64000	0.336 (0.000)	0.001	0.949 (0.005)	0.946 (0.005)	0.948 (0.005)
α_1					
1000	1.076 (-0.005)	0.025	0.965 (0.086)	0.841 (0.086)	0.969 (0.097)
2000	1.078 (-0.003)	0.018	0.973 (0.063)	0.830 (0.063)	0.971 (0.072)
4000	1.080 (-0.001)	0.014	0.975 (0.048)	0.809 (0.048)	0.972 (0.055)
8000	1.080 (-0.001)	0.011	0.982 (0.036)	0.803 (0.036)	0.975 (0.042)
16000	1.081 (0.000)	0.008	0.981 (0.028)	0.777 (0.028)	0.975 (0.032)
32000	1.081 (0.000)	0.006	0.980 (0.021)	0.752 (0.021)	0.976 (0.025)
64000	1.081 (0.000)	0.005	0.982 (0.016)	0.747 (0.016)	0.977 (0.020)

Compared to Table 5.1, there are two differences in Table 5.3: (1) The percentile CIs perform well for threshold parameter e, although there is some overcoverage especially at large sample sizes. The basic CIs suffer from a even more severe undercoverage compared to Table 5.1. (2) All of the three types of CIs do not show good coverage probabilities for β, the magnitude of jump at the threshold in Table 5.3, but the coverage improves with larger sample size.

As for α_1, the coverage is close to the nominal level 0.95, which is consistent with Table 5.1. As for α_2, although there is a slight overcoverage at large sample size compared to Table 5.1, both of the percentile and symmetric CIs provide valid coverage, while the basic CIs perform poorly due to the undercoverage phenomenon.

To further investigate the discrepancies of coverage between percentile and basic CIs for threshold parameter e, we search for the data samples simulated satisfying that percentile and symmetric CIs cover the truth whereas the basic ones do not. Figure 5.2 illustrated three typical scenarios of the sampling distribution of the threshold parameter e: (i) When the sampling distribution is left-skewed (Figure 5.2 left), the lower bound of the percentile CIs is smaller than that of basic CIs, which means that the percentile CIs adapt to the skewness in the sampling distributions, and thus, are more likely to cover the truth. (ii) When the sampling distribution is right-skewed (Figure 5.2 right), the upper bound of the percentile CIs is greater than that of the basic CIs. (iii) When the sampling distribution is not skewed, as shown by Figure 5.2 middle, the percentile CIs seem more likely to cover the truth than the basic CIs.

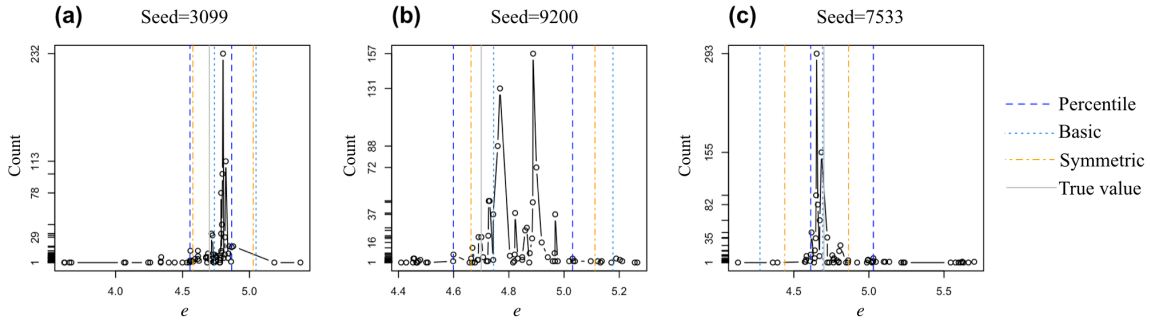

Figure 5.2: Bootstrap distributions of $\hat{e}$ from 3 datasets of size 2000 when the data generating model is a sigmoid model with steepness=1. Left: left-skewed sampling distribution; Middle: not-skewed sampling distribution; Right: right-skewed sampling distribution. In all three cases percentile and symmetric CIs cover and basic CIs does not.

5.2.2 Sigmoid model with steepness=5

We use a step linear regression model to fit the data generated by the sigmoid model 3.3 (Figure 3.3 right) of size $1000, 2000, 4000, 8000, 16000, 32000, 64000$, respectively. Table 5.4 showed simulation results of the four parameters when using step linear regression model as working model and sigmoid model with steepness=5 as data generating model.

Table 5.4: Simulation results from 10,000 Monte Carlo runs when the true model is a sigmoid model with steepness equals to 5. Estimate, bias, Monte Carlo standard deviation, coverage probabilities and mean width of the 3 types of n-bootstrap confidence intervals.

	Monte Carlo results		Coverage(Width)		
n	Estimate(Bias)	MC SD	Percentile	Basic	Symmetric
e					
1000	4.698 (-0.002)	0.146	0.960 (0.458)	0.658 (0.458)	0.966 (0.572)
2000	4.698 (-0.002)	0.112	0.965 (0.350)	0.656 (0.350)	0.969 (0.441)
4000	4.699 (-0.001)	0.088	0.966 (0.273)	0.646 (0.273)	0.966 (0.344)
8000	4.699 (-0.001)	0.069	0.969 (0.215)	0.644 (0.215)	0.969 (0.272)
16000	4.700 (0.000)	0.055	0.971 (0.170)	0.655 (0.170)	0.972 (0.215)
32000	4.700 (0.000)	0.043	0.968 (0.134)	0.657 (0.134)	0.968 (0.168)
64000	4.700 (0.000)	0.034	0.970 (0.106)	0.651 (0.106)	0.973 (0.134)
β					
1000	0.371 (0.005)	0.019	0.927 (0.074)	0.938 (0.074)	0.941 (0.075)
2000	0.369 (0.003)	0.013	0.935 (0.052)	0.948 (0.052)	0.948 (0.053)
4000	0.368 (0.002)	0.010	0.939 (0.037)	0.947 (0.037)	0.949 (0.037)
8000	0.367 (0.001)	0.008	0.940 (0.026)	0.947 (0.026)	0.948 (0.027)
16000	0.367 (0.001)	0.005	0.948 (0.019)	0.953 (0.019)	0.956 (0.019)
32000	0.366 (0.000)	0.003	0.944 (0.013)	0.945 (0.013)	0.949 (0.013)
64000	0.366 (0.000)	0.003	0.950 (0.009)	0.951 (0.009)	0.953 (0.009)
α_2					
1000	0.337 (0.001)	0.010	0.951 (0.038)	0.946 (0.038)	0.951 (0.038)
2000	0.336 (0.000)	0.007	0.949 (0.027)	0.945 (0.027)	0.949 (0.027)
4000	0.337 (0.001)	0.005	0.949 (0.019)	0.945 (0.019)	0.950 (0.019)
8000	0.337 (0.001)	0.003	0.950 (0.013)	0.948 (0.013)	0.952 (0.013)
16000	0.336 (0.000)	0.002	0.943 (0.009)	0.943 (0.009)	0.944 (0.009)
32000	0.336 (0.000)	0.002	0.942 (0.007)	0.941 (0.007)	0.944 (0.007)
64000	0.336 (0.000)	0.001	0.934 (0.005)	0.932 (0.005)	0.935 (0.005)
α_1					
1000	1.015 (-0.002)	0.016	0.956 (0.060)	0.916 (0.060)	0.955 (0.063)
2000	1.015 (-0.002)	0.012	0.962 (0.043)	0.914 (0.043)	0.959 (0.045)
4000	1.016 (-0.001)	0.008	0.964 (0.032)	0.903 (0.032)	0.960 (0.033)
8000	1.017 (0.000)	0.006	0.970 (0.023)	0.898 (0.023)	0.966 (0.025)
16000	1.017 (0.000)	0.005	0.970 (0.017)	0.891 (0.017)	0.965 (0.018)
32000	1.017 (0.000)	0.003	0.974 (0.013)	0.880 (0.013)	0.968 (0.014)
64000	1.017 (0.000)	0.003	0.976 (0.009)	0.863 (0.009)	0.970 (0.010)

Similar to the results from the last simulate scenario (Table 5.3), the percentile and symmetric CIs perform well for threshold parameter e, albeit at a slight overcoverage, while the basic CIs perform poorly. Compared to Table 5.3, the mean width of CIs for threshold parameter is relatively narrower. This is not surprising because the change point depends on the features of sigmoid models, and the sigmoid model with steepness=5 increase sharply than sigmoid model with steepness=1. As for other regression coefficients, all of the three types of CIs outperform those in Table 5.3 but still suffer from some undercoverage for β. This is because when the middle part of the sigmoid model increases more quickly, the step model tend to fit the data better. It is intuitive that we can get better estimator results if the fitting model is more like the true data generating model. As for α_1 and α_2, the basic CIs are prone to a slight undercoverage, while the percentile and the basic CIs provide valid coverage.

To further investigate the discrepancies of coverage between percentile and basic CIs for threshold parameter e, we search for the data samples simulated satisfying that percentile and symmetric CIs cover the truth whereas the basic ones do not. Figure 5.3 illustrates the similar sampling distribution as Figure 5.2. To sum up, the ability of adapting to any skewness in the sampling distribution makes the percentile CIs perform better than the basic CIs.

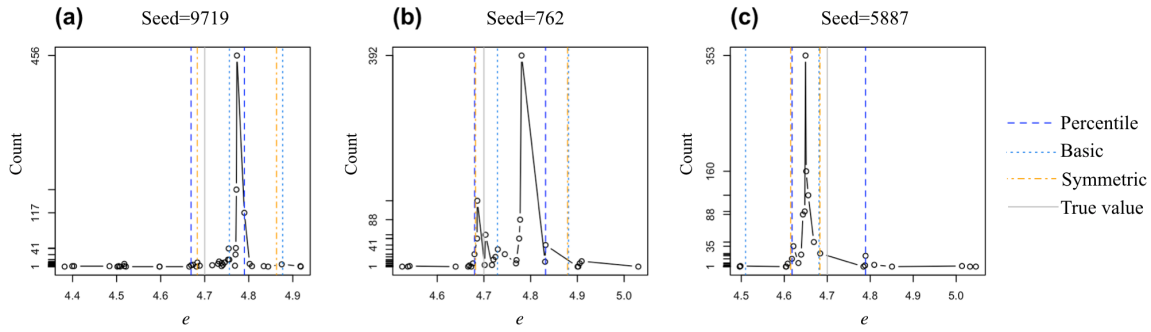

Figure 5.3: Bootstrap distributions of $\hat{e}$ from 3 datasets of size 2000 when the data generating model is a sigmoid model with steepness=5. Left: left-skewed sampling distribution; Middle: not-skewed sampling distribution; Right: right-skewed sampling distribution. In all three cases percentile and symmetric CIs cover and basic CIs does not.

5.2.3 Sigmoid model with steepness=15

We use a step linear regression model to fit the data generated by the sigmoid model 3.4 (Figure 3.4 right) of size $1000, 2000, 4000, 8000, 16000, 32000, 64000$, respectively. Table 5.5 showed simulation results of the four parameters when using step linear regression model as working model and sigmoid model with steepness=15 as data generating model.

Table 5.5: Simulation results from 10,000 Monte Carlo runs when the true model is a sigmoid model with steepness equals to 15. Estimate, bias, Monte Carlo standard deviation, coverage probabilities and mean width of the 3 types of n-bootstrap confidence intervals.

	Monte Carlo results		Coverage(Width)		
n	Estimate(Bias)	MC SD	Percentile	Basic	Symmetric
e					
1000	4.697 (-0.003)	0.075	0.950 (0.234)	0.674 (0.234)	0.965 (0.293)
2000	4.698 (-0.002)	0.055	0.956 (0.171)	0.665 (0.171)	0.964 (0.214)
4000	4.699 (-0.001)	0.042	0.965 (0.130)	0.660 (0.130)	0.972 (0.163)
8000	4.699 (-0.001)	0.032	0.964 (0.101)	0.655 (0.101)	0.966 (0.126)
16000	4.700 (0.000)	0.025	0.968 (0.078)	0.655 (0.078)	0.969 (0.098)
32000	4.700 (0.000)	0.020	0.969 (0.061)	0.649 (0.061)	0.969 (0.077)
64000	4.700 (0.000)	0.016	0.967 (0.049)	0.649 (0.049)	0.970 (0.061)
β					
1000	0.392 (0.003)	0.019	0.937 (0.074)	0.940 (0.074)	0.942 (0.074)
2000	0.391 (0.002)	0.013	0.945 (0.052)	0.949 (0.052)	0.949 (0.053)
4000	0.390 (0.001)	0.009	0.948 (0.037)	0.948 (0.037)	0.951 (0.037)
8000	0.390 (0.001)	0.007	0.946 (0.026)	0.947 (0.026)	0.948 (0.026)
16000	0.389 (0.000)	0.005	0.951 (0.019)	0.953 (0.019)	0.954 (0.019)
32000	0.389 (0.000)	0.003	0.945 (0.013)	0.944 (0.013)	0.946 (0.013)
64000	0.389 (0.000)	0.002	0.950 (0.009)	0.952 (0.009)	0.953 (0.009)
α_2					
1000	0.337 (0.001)	0.010	0.950 (0.037)	0.948 (0.037)	0.950 (0.038)
2000	0.336 (0.000)	0.007	0.948 (0.026)	0.948 (0.026)	0.949 (0.026)
4000	0.337 (0.001)	0.005	0.949 (0.019)	0.947 (0.019)	0.950 (0.019)
8000	0.336 (0.000)	0.003	0.950 (0.013)	0.948 (0.013)	0.950 (0.013)
16000	0.336 (0.000)	0.002	0.943 (0.009)	0.943 (0.009)	0.945 (0.009)
32000	0.336 (0.000)	0.002	0.941 (0.007)	0.941 (0.007)	0.942 (0.007)
64000	0.336 (0.000)	0.001	0.933 (0.005)	0.931 (0.005)	0.934 (0.005)
α_1					
1000	1.004 (-0.002)	0.014	0.950 (0.055)	0.939 (0.055)	0.952 (0.056)
2000	1.005 (-0.001)	0.010	0.952 (0.039)	0.942 (0.039)	0.956 (0.039)
4000	1.005 (-0.001)	0.007	0.951 (0.028)	0.935 (0.028)	0.953 (0.028)
8000	1.005 (-0.001)	0.005	0.955 (0.020)	0.937 (0.020)	0.956 (0.020)
16000	1.005 (-0.001)	0.004	0.954 (0.014)	0.933 (0.014)	0.954 (0.015)
32000	1.006 (0.000)	0.003	0.956 (0.010)	0.928 (0.010)	0.955 (0.010)
64000	1.006 (0.000)	0.002	0.957 (0.007)	0.919 (0.007)	0.956 (0.008)

Similar as Table 5.3 and Table 5.4, the percentile and symmetric CIs perform well for threshold parameter e while the basic CIs perform poorly. As for other regression coefficients, all of the three types of CIs perform similarly as those in Table 5.4 for β. As for α_1 and α_2, the basic CIs are prone to a slight undercoverage, especially at large sample size, while the percentile and the basic CIs provide valid coverage.

To further investigate the discrepancies of coverage between percentile and basic CIs for threshold parameter e, we search for the data samples simulated satisfying that percentile and symmetric CIs cover the truth whereas the basic ones do not. Figure 5.4 illustrates the similar sampling distribution as Figure 5.2 and Figure 5.3, and again validates that the ability of adapting to any skewness in the sampling distribution makes the percentile CIs perform better than the basic CIs.

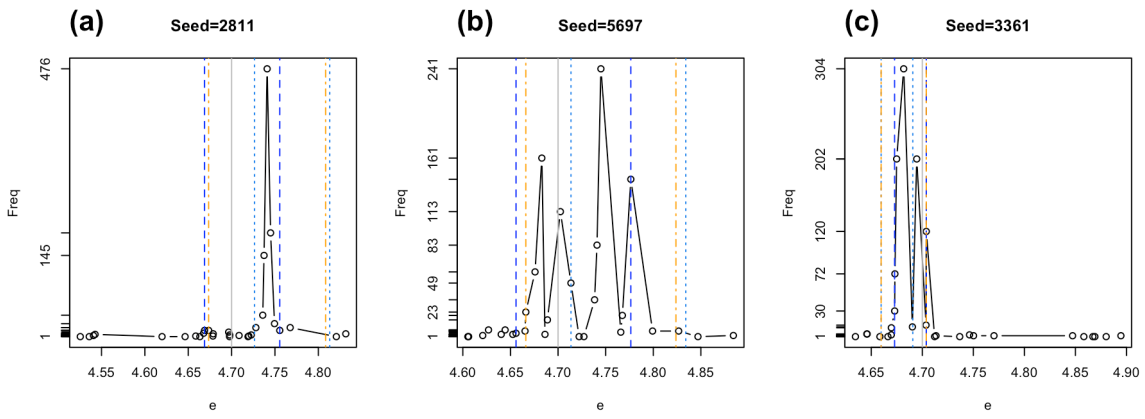

Figure 5.4: Bootstrap distributions of $\hat{e}$ from 3 datasets of size 2000 when the data generating model is a sigmoid model with steepness=15. Left: left-skewed sampling distribution; Middle: not-skewed sampling distribution; Right: right-skewed sampling distribution. In all three cases percentile and symmetric CIs cover and basic CIs does not.

5.2.4 Quadratic model

We use a step linear regression model to fit the data generated by the quadratic model 3.5 (Figure 3.5 right) of size $1000, 2000, 4000, 8000, 16000, 32000, 64000$, respectively. Table 5.6 shows the Monte Carlo simulation results when using a step linear regression model as working model and quadratic model as data generating model.

The results are listed in Table 5.6 and show that both percentile and symmetric CIs provide good coverage for the regression coefficients and threshold parameter in step models, while the basic ones do not. The poor coverage probabilities of basic CIs are similar to those in Table 5.3 and Table 5.4, which again shows that the basic CIs do not perform well when the data generating models are far away from the fitting model.

Again, to explore the discrepancies of coverage between percentile and basic CIs for threshold parameter e, we search for the data samples simulated satisfying that percentile and symmetric CIs cover the truth whereas the basic ones do not. Figure 5.5 depicted the three typical scenarios of the sampling distribution: right-skewed, not-skewed and left-skewed.

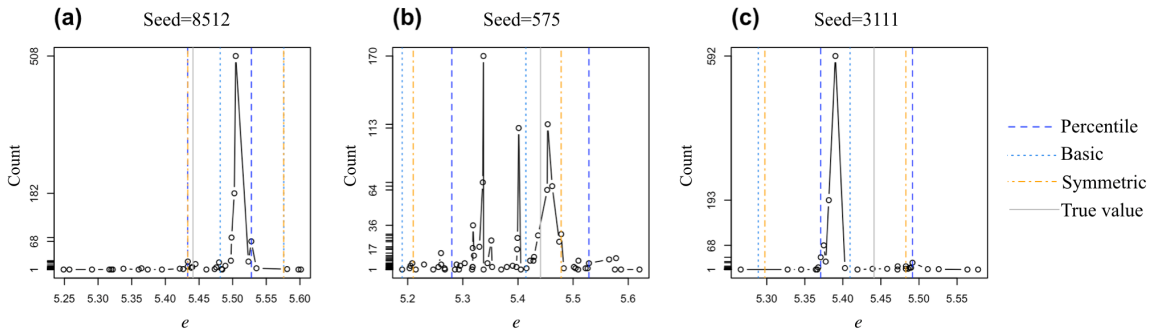

Figure 5.5: Bootstrap distributions of $\hat{e}$ from 3 datasets of sample size 2000 when the data generating model is a quadratic model. Left: left-skewed sampling distribution; Middle: not-skewed sampling distribution; Right: right-skewed sampling distribution. In all three cases percentile and symmetric CIs cover and basic CIs does not.

Table 5.6: Simulation results from 10,000 Monte Carlo runs when the true model is a quadratic model. Estimate, %bias, Monte Carlo standard deviation, coverage probabilities and median width of the 3 types of bootstrap confidence intervals (percentile, basic and symmetric) of the parameter estimates are given from Efron bootstrap by fastgrid algorithm.

	Monte Carlo results		Coverage(Width)		
n	Estimate(Bias)	MC SD	Percentile	Basic	Symmetric
e					
1000	5.436 (-0.005)	0.081	0.955 (0.258)	0.691 (0.258)	0.965 (0.317)
2000	5.439 (-0.002)	0.062	0.956 (0.197)	0.681 (0.197)	0.963 (0.244)
4000	5.440 (-0.001)	0.048	0.958 (0.151)	0.670 (0.151)	0.964 (0.189)
8000	5.441 (0.000)	0.037	0.964 (0.118)	0.676 (0.118)	0.967 (0.147)
16000	5.441 (0.000)	0.029	0.965 (0.091)	0.661 (0.091)	0.966 (0.114)
32000	5.441 (0.000)	0.022	0.965 (0.071)	0.675 (0.071)	0.966 (0.088)
64000	5.441 (0.000)	0.016	0.963 (0.056)	0.662 (0.056)	0.964 (0.070)
β					
1000	5.517 (0.005)	0.126	0.956 (0.487)	0.936 (0.487)	0.955 (0.494)
2000	5.514 (0.002)	0.089	0.958 (0.341)	0.931 (0.341)	0.956 (0.349)
4000	5.513 (0.001)	0.066	0.962 (0.253)	0.936 (0.253)	0.957 (0.259)
8000	5.512 (0.000)	0.047	0.962 (0.178)	0.927 (0.178)	0.960 (0.184)
16000	5.512 (0.000)	0.033	0.956 (0.125)	0.917 (0.125)	0.954 (0.130)
32000	5.511 (-0.001)	0.024	0.954 (0.089)	0.903 (0.089)	0.949 (0.093)
64000	5.511 (-0.001)	0.018	0.952 (0.023)	0.948 (0.023)	0.953 (0.023)
α_2					
1000	0.336 (0.000)	0.054	0.962 (0.207)	0.947 (0.207)	0.964 (0.212)
2000	0.337 (0.001)	0.037	0.962 (0.142)	0.948 (0.142)	0.961 (0.144)
4000	0.337 (0.001)	0.025	0.954 (0.097)	0.941 (0.097)	0.954 (0.099)
8000	0.337 (0.001)	0.017	0.955 (0.068)	0.949 (0.068)	0.955 (0.068)
16000	0.336 (0.000)	0.012	0.958 (0.047)	0.950 (0.047)	0.958 (0.048)
32000	0.337 (0.001)	0.008	0.953 (0.033)	0.946 (0.033)	0.952 (0.033)
64000	0.336 (0.000)	0.006	0.952 (0.023)	0.948 (0.023)	0.953 (0.023)
α_1					
1000	-0.318 (-0.002)	0.090	0.962 (0.325)	0.864 (0.325)	0.957 (0.353)
2000	-0.317 (-0.001)	0.067	0.963 (0.238)	0.854 (0.238)	0.958 (0.261)
4000	-0.316 (0.000)	0.050	0.965 (0.176)	0.844 (0.176)	0.960 (0.195)
8000	-0.316 (0.000)	0.037	0.967 (0.131)	0.829 (0.131)	0.963 (0.147)
16000	-0.315 (0.001)	0.028	0.970 (0.097)	0.815 (0.097)	0.965 (0.109)
32000	-0.315 (0.001)	0.021	0.970 (0.072)	0.811 (0.072)	0.965 (0.082)
64000	-0.316 (0.000)	0.016	0.969 (0.055)	0.787 (0.055)	0.967 (0.063)

Chapter 6

SELECTING M FOR M-BOOTSTRAP AND SUBSAMPLING CONFIDENCE INTERVAL METHODS

In the previous chapter, we examined the performance of percentile, basic, and symmetric n-bootstrap confidence intervals. In this chapter, we conduct Monte Carlo experiments to investigate how the coverage of m-bootstrap and subsampling bootstrap confidence intervals depends on the block size m and seek to develop data-adaptive rules to select m.

For each of the four data generating models including a step model, two sigmoid models with steepness=1 and 5 and a quadratic model, and for each $n \in \{250, 500, 1000, 1500, 2000\}$, we

(1) Conduct simulate 10000 samples of data of size n;

(2) For a series of $m \leq n$, calculate the coverage probabilities of percentile, basic and symmetric confidence intervals;

(3) Find a data-driven rule for m so that the percentile CI coverage of e is closest to the nominal level 0.95.

In step (3) of the above, to investigate the the rule of choosing the best block size m for threshold parameter e in a specific data generating scheme, we fit the model:

$$m = k \times n^p, \tag{6.1}$$

where m is the block size of m-bootstrap, $k \neq 0$ is a constant denoting the effect of n^p on m, and $p < 1$ is a positive real number denoting the power. For each data generating model, we have five data points of m and n. We consider the following two methods based on the above model:

Method 1: Take log of both sides of model 6.1, and then fit a linear regression model with intercept ($k \neq 1$) using the five data points of m and n. (Referred to here as log-linear-k$\neq$1)

Method 2: Follow the same steps as Method 1 except for fitting a linear regression model

without intercept ($k = 1$). (Referred to here as log-linear-k=1)

Since we may obtain different rules for different data generating schemes, we first compare the rules of the four data generating models, and our ultimate goals are to:

(1). Derive a rule using only the three mis-specification scenarios (referred to here as mis-rule) by fitting a linear mixed model with the 3×5 data points.

(2). Search for a general rule suitable for the four data generating models (referred to here as mixed-rule) by fitting a linear mixed model with all the 4×5 data points.

6.1 M-out-of-n bootstrap

6.1.1 Model correctly specified

We followed the procedures above to calculate the coverage probabilities of threshold parameters e by m-bootstrap. For each $n = 250, 500, 1000, 1500, 2000$, the relationship between the estimated coverage probabilities of threshold parameter e by m-bootstrap and different block size m are presented in Figure 6.1.

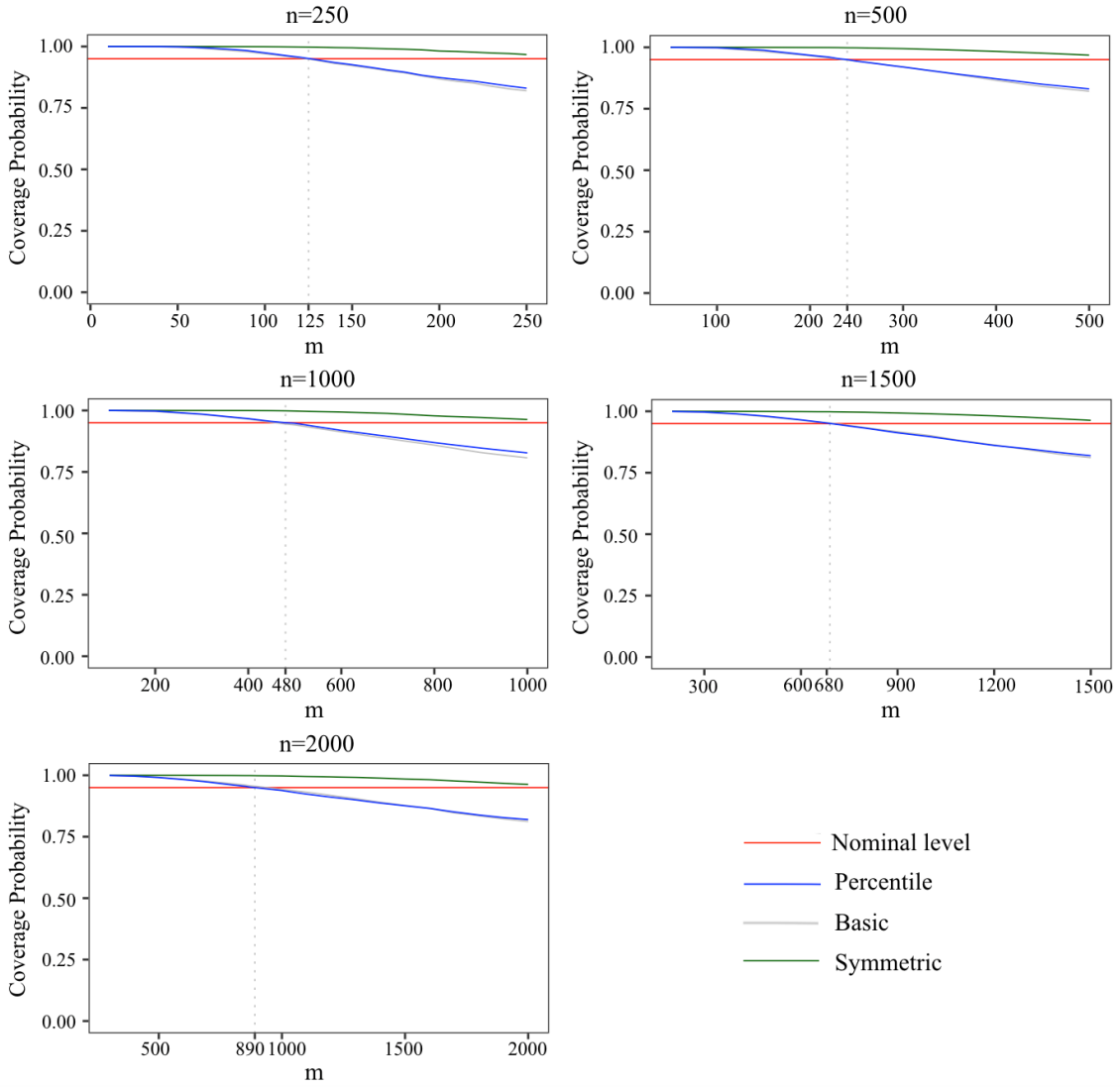

Figure 6.1: Estimated coverage probabilities of percentile, basic and symmetric confidence intervals given by m-bootstrap method from 10000 Monte Carlo runs when the true model is a step linear regression model.

The results are shown in Figure 6.1. Note that when $m = n$, m-bootstrap is the equivalent to n-bootstrap, and the estimated coverage probabilities are equal to those listed in Table 5.1. The estimated coverage probabilities decrease from 1 as the block size m increases. The lines indicating percentile and basic coverage probabilities nearly overlap. In addition, as m tends towards n, the coverage of the symmetric line never dips under the

nominal level. As for the percentile and the basic lines, the estimated coverage probabilities become the nominal level 0.95 at some m.

Based on Figure 6.1, we estimate the optimal block sizes $m = 125$ for $n = 250$; $m = 240$ for $n = 500$; $m = 480$ for $n = 1000$; $m = 690$ for $n = 1500$ and $m = 890$ for $n = 2000$ by spline interpolation. Using these as input, we proceed to estimate the rules for selecting m. Table 6.1 shows the estimated k and p and the corresponding confidence intervals calculated by the two methods described previously, and Figure 6.2 shows the corresponding regression lines with confidence bands.

Table 6.1: Estimated k and p and its confidence intervals calculated by 2 different methods.

	k	p
Method 1: log-linear-k≠1	0.659 (0.522,0.831)	0.950 (0.916,0.985)
Method 2: log-linear-k=1	1	0.889 (0.879,0.899)

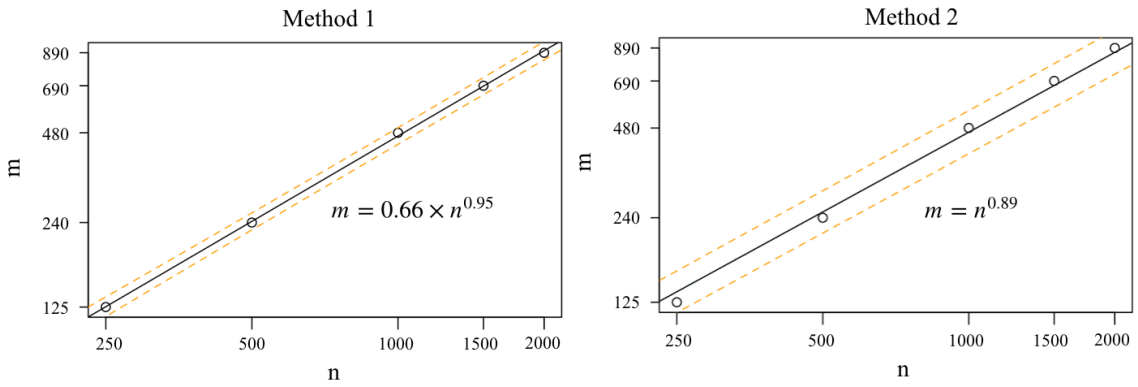

Figure 6.2: Fitted linear regression models of the relationship between the lock size m and the sample size n by m-bootstrap for threshold parameter e when the true model is the step linear regression model. Left: Method 1 ($k \neq 1$). Right: Method 2 ($k = 1$).

6.1.2 Model misspecified

In Table 5.3, 5.4 and 5.6, we saw that the coverage probabilities of both percentile and symmetric CIs calculated by n-bootstrap for threshold parameter e were close to the nom-

inal level 0.95 or prone to slight overcoverage. Figure 6.3 shows the estimated coverage probabilities of percentile, basic and symmetric confidence intervals by m-bootstrap when the data generating model is a sigmoid model with steepness=1. We found that the fitted lines of estimated coverage probabilities for the percentile and symmetric CI have a decreasing trend but are always above the red line indicating the nominal level 0.95. Similar trends happen when the data generating models are sigmoid model with steepness=5 and quadratic model (results and plots not shown).

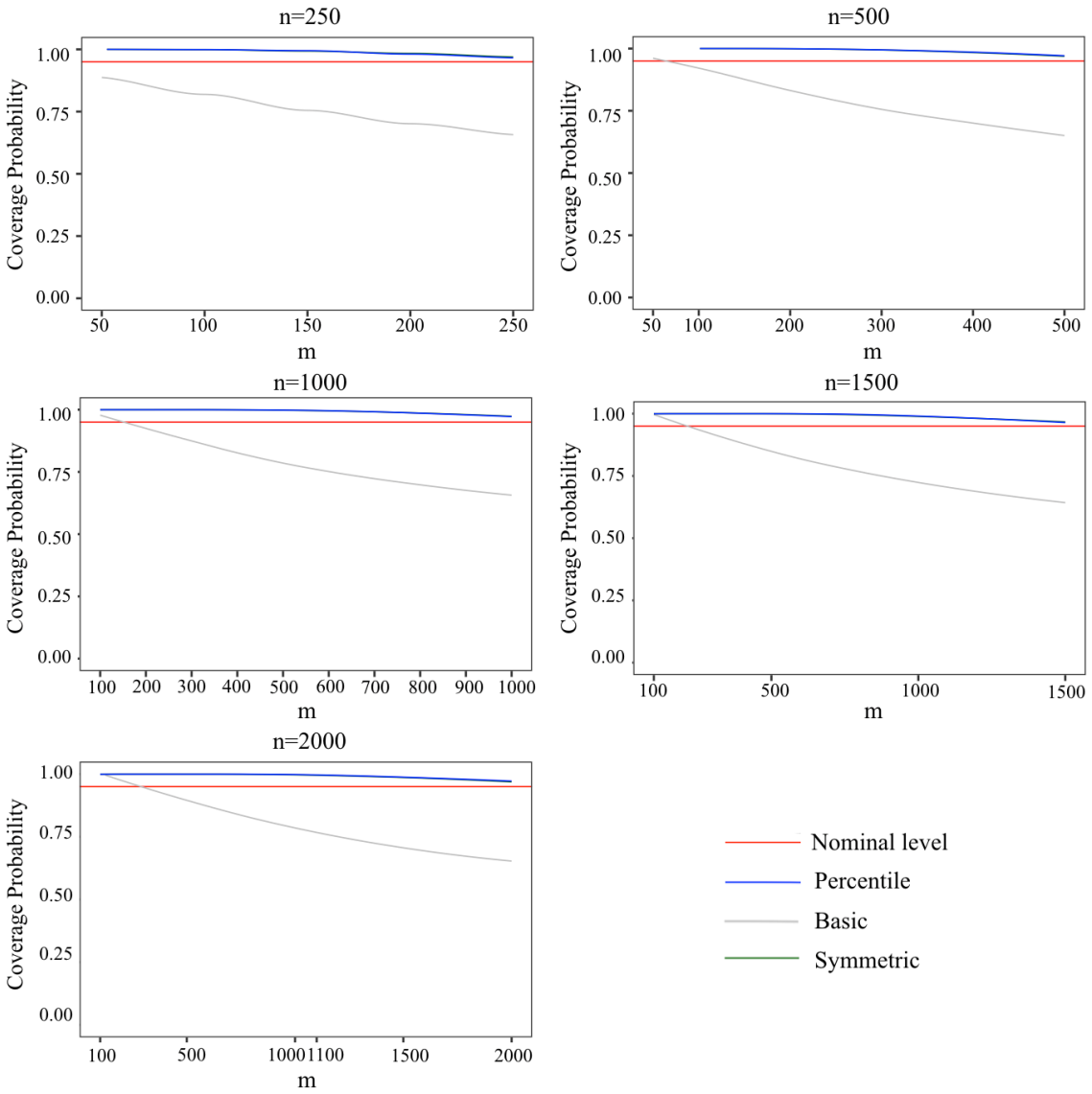

Figure 6.3: Estimated coverage probabilities of percentile, basic and symmetric confidence intervals given by m-bootstrap method from 10000 Monte Carlo runs when the true model is a sigmoid model with steepness=1.

6.1.3 A General Rule

In this section, we are interested in looking for a general rule of selecting the block size m for m-bootstrap when the data generating model is unknown.

Table 6.2 summarizes the value of block size ms with estimated coverage probabilities of percentile confidence intervals closest to the nominal level 0.95. When the working model and the data generating model are misspecified (sig1, sig5, qua), the estimated coverage probabilities of percentile confidence intervals are closest to the nominal level 0.95 when the block size m is equal to the sample size n. When the working model and the data generating model are both step linear regression models, the block size ms chosen are substantially smaller than those chosen when models are misspecified.

Table 6.2: Block size m satisfying that the estimated coverage probabilities of percentile confidence intervals by m-bootstrap are close to the nominal level 0.95 under certain sample size n.

	m	n	m	n	m	n	m	n	m	n
step	125	250	240	500	480	1000	690	1500	890	2000
sig1	250	250	500	500	1000	1000	1500	1500	2000	2000
sig5	250	250	500	500	1000	1000	1500	1500	2000	2000
qua	250	250	500	500	1000	1000	1500	1500	2000	2000

We then fit a linear mixed model with random intercept and random slope based on Model 6.1 to investigate the relationship between the block size m and the sample size n considering all the four data generating models. We regard the four data generating models as four different conditions, and then use linear mixed model to fit the data consisting of the block size m and the sample size n summarized in Table 6.2.

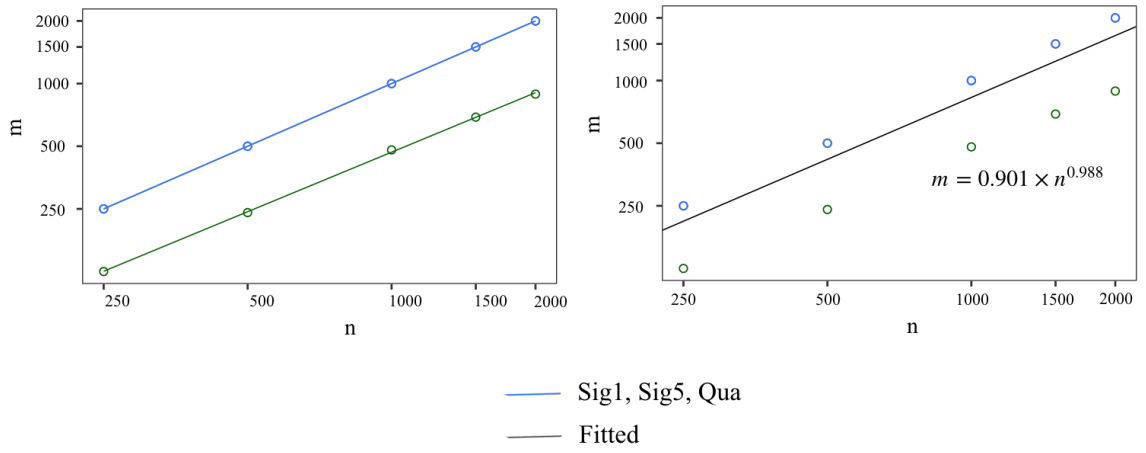

Figure 6.4: m-bootstrap. Left: Fitted linear regression models by Method 1 depicting the relationship between the block size m and the sample size n. Right: Fitted linear mixed model by Method 1 depicting the relationship between the block size m and the sample size n considering all the data points of the four data generating models.

According to Model 6.1 Method 1, the estimated k is 0.901, with 95% confidence interval $(0.746, 1.090)$, and the estimated p is 0.988, with 95% confidence interval $(0.965, 1.010)$. Figure 6.4 shows the linear regression lines fitted for each data generating model by Method 1 in the left, while the linear mixed model fitted based on all the four data generating models in the right. The estimated prediction rule of m- bootstrap for selecting the block size m when the data generating models are unknown is:

$$m = 0.901 \times n^{0.988}. \tag{6.2}$$

Clearly, this result is a compromise between having the model correctly specified and having the model misspecified. The particular compromise reached here is a result of having 1 scenario where the model is correctly specified and 3 scenarios where the model is misspecified. In addition, we derive another prediction rule of m-bootstrap for selecting the block size m considering the three misspecification scenarios only. Since the estimated coverage probabilities are closest to the nominal level 0.95 when the block size m is equal to the sample size n, then the prediction rule should be: $m = n$. More sophisticated rules should

be considered, but unfortunately are out of scope of this thesis.

6.2 Subsampling bootstrap

6.2.1 Step linear regression model

We follow similar procedures to derive data-adaptive rules for m for subsampling. For each $n = 250, 500, 1000, 1500, 2000$, the relationship between the estimated coverage probabilities of threshold parameter e and block size m is presented in Figure 6.5.

As shown by Figure 6.5, the estimated coverage probabilities decrease from 1 to 0 as the block size m increases towards n because resampling is without replacement. The lines for percentile and basic CIs nearly overlap, while the symmetric CIs generally have higher coverage. Based on these results, we estimate that the best block size $m = 90$ for $n = 250$; $m = 175$ for $n = 500$; $m = 355$ for $n = 1000$; $m = 515$ for $n = 1500$ and $m = 680$ for $n = 2000$.

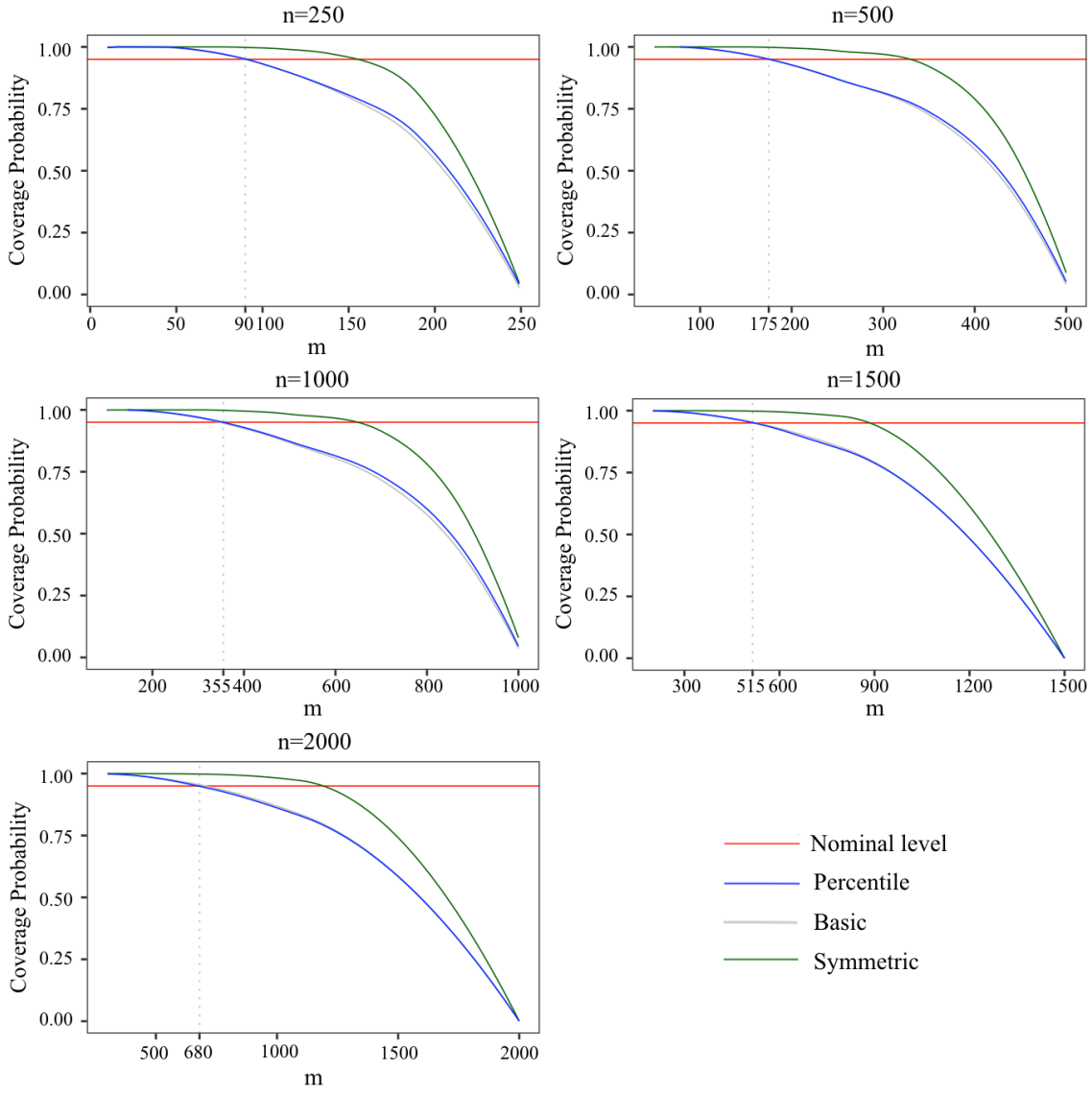

Figure 6.5: Estimated coverage probabilities of percentile, basic and symmetric confidence intervals given by subsampling bootstrap method from 10000 Monte Carlo runs when the true model is a step linear regression model.

To obtain data-adaptive rules for choosing m, we fit the model 6.1. Table 6.3 shows the estimated k and p and the corresponding confidence intervals calculated by the two methods described previously. Figure 6.6 shows the linear regression lines fitted by Method 1 and Method 2, respectively. As shown by Figure 6.6, the linear regression model fit the five data

points well and there is no visible second order trend. The confidence bands are shown by the orange dotted lines in the Figure 6.6.

Table 6.3: Estimated k and p and its confidence intervals calculated by 2 different methods when the true model is the step linear regression model.

	k	p
Method 1: log-linear-k≠1	0.410 (0.344,0.489)	0.976 (0.950,1.002)
Method 2: log-linear-k=1	1	0.845 (0.824,0.866)

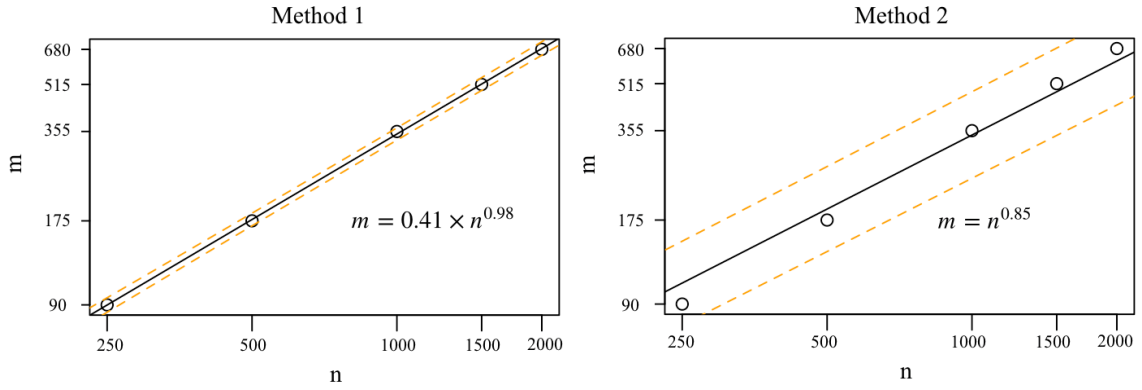

Figure 6.6: Fitted linear regression lines depicting the relationship between the block size m and the sample size n calculated by subsampling bootstrap for threshold parameter e when the true model is the step linear regression model. Left: Method 1 ($k \neq 1$). Right: Method 2 ($k = 1$).

6.2.2 Sigmoid model with steepness=1

As shown by Figure 6.7, the estimated coverage probabilities decrease similarly as Figure 6.5 from 1 to 0 as the block size m increases towards n. Unlike in Figure 6.5, the percentile and symmetric CI coverage lines nearly overlap, while the basic CIs generally have smaller coverage.

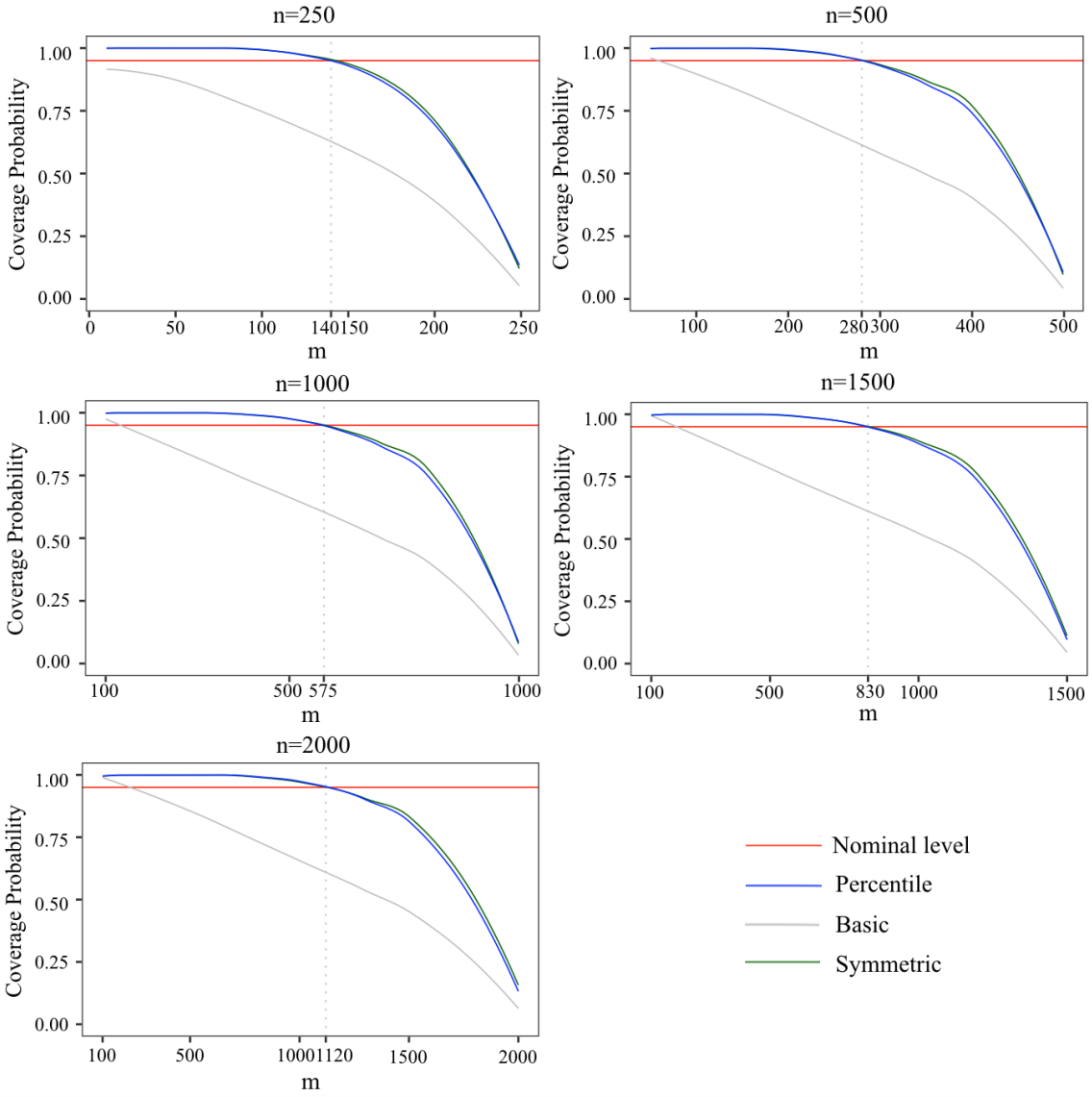

Figure 6.7: Estimated coverage probabilities of percentile, basic and symmetric confidence intervals given by subsampling bootstrap method from 10000 Monte Carlo runs when the true model is a sigmoid model with steepness=1.

To investigate the best block size m with coverage probabilities of percentile CIs close to the nominal level 0.95, we fit the Model 6.1. Based on Figure 6.7, the block size $m = 140$ for $n = 250$; $m = 280$ for $n = 500$; $m = 575$ for $n = 1000$; $m = 830$ for $n = 1500$ and $m = 1120$ for $n = 2000$. We then fit Model 6.1 to find a rule of selecting m when the data generating

model is the sigmoid model with steepness=1, which is misspecified as the working model.

Table 6.4 shows the estimated k and p and the corresponding confidence intervals calculated by two methods differing by whether $k = 1$. The fitted value of p bt Method 1 is nearly 1. Figure 6.8 shows the linear regression lines fitted by Method 1 and Method 2, respectively.

Table 6.4: Estimated k and p and its confidence intervals calculated by 2 different methods when the data generating model is the sigmoid model with steepness=1.

	k	p
Method 1: log-linear-k≠1	0.564 (0.458,0.694)	0.999 (0.968,1.030)
Method 2: log-linear-k=1	1	0.915 (0.902,0.928)

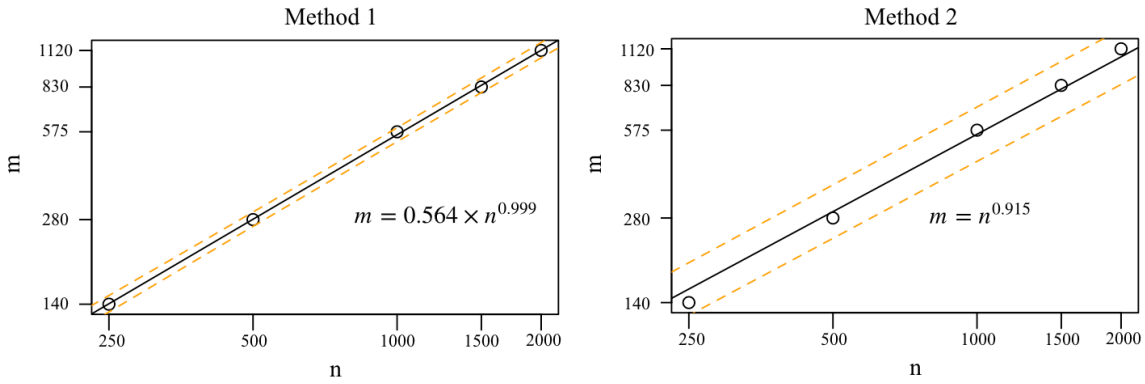

Figure 6.8: Fitted linear regression lines depicting the relationship between the block size m and the sample size n calculated by subsampling bootstrap for threshold parameter e when the data generating model is the sigmoid model with steepness=1. Left: Method 1 ($k \neq 1$). Right: Method 2 ($k = 1$).

6.2.3 Sigmoid model with steepness=5

Following the same procedure as above, we find the best block size $m = 135$ for $n = 250$; $m = 280$ for $n = 500$; $m = 565$ for $n = 1000$; $m = 830$ for $n = 1500$ and $m = 1120$ for $n = 2000$ satisfy the coverage probabilities of percentile CIs for threshold parameter

e are close to the nominal level 0.95 (shown by Figure 6.9). Actually, the block size ms found for sigmoid model with steepness=5 are very similar to those for sigmoid model with steepness=1. In the next subsection we will see that if the optimal ms when the true model is the quadratic model are also very similar.

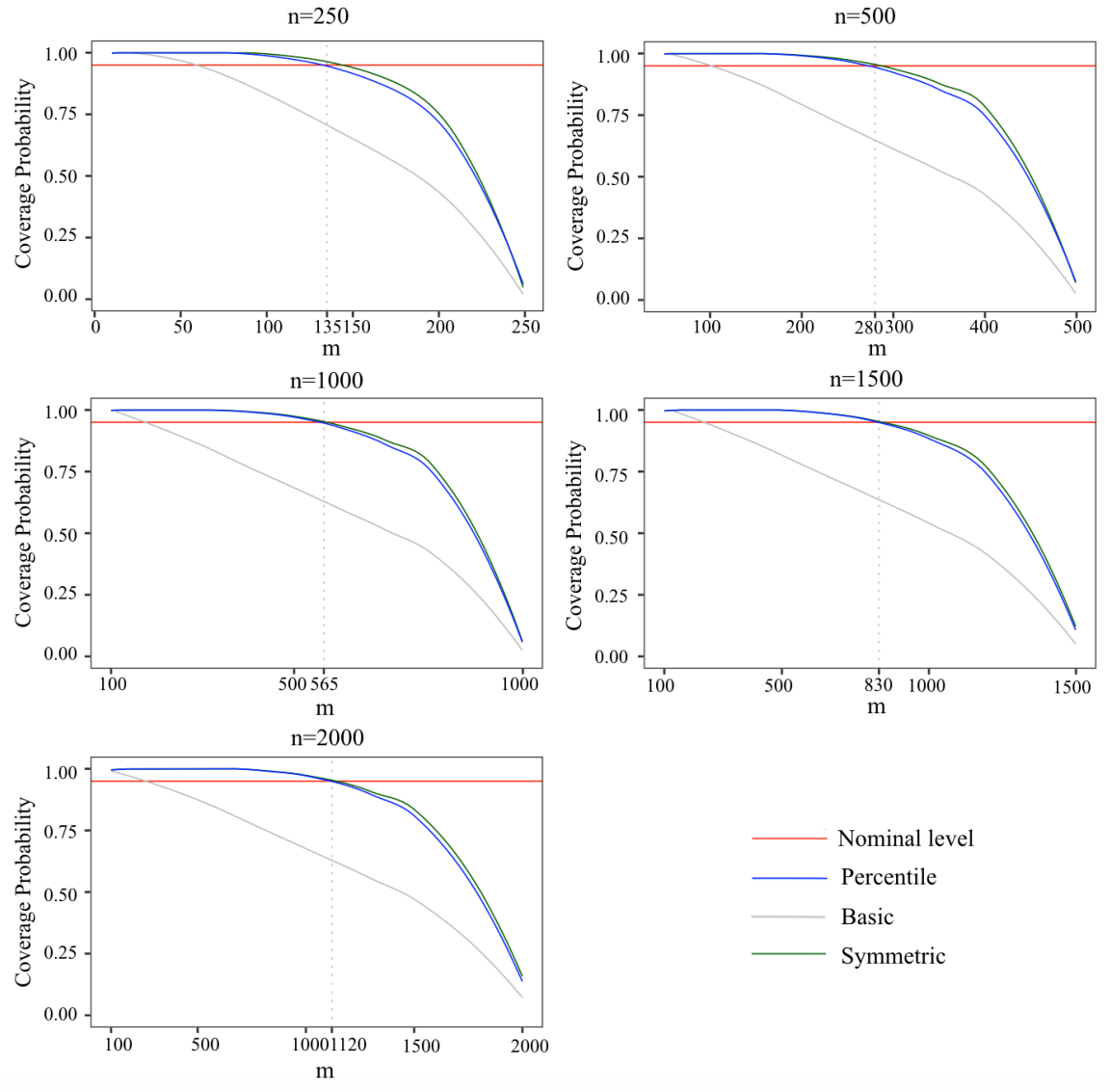

Figure 6.9: Estimated coverage probabilities of percentile, basic and symmetric confidence intervals given by subsampling bootstrap method from 10000 Monte Carlo runs when the true model is a sigmoid model with steepness=5.

Table 6.5 shows the estimated k and p and the corresponding confidence intervals calculated by the two methods described previously. We note that the p estimated by Method 1 is greater than 1, which seemingly contradicts the requirement that $m = o(n)$.

Table 6.5: Estimated k and p and its confidence intervals calculated by 2 different methods when the data generating model is the sigmoid model with steepness=5.

	k	p
Method 1: log-linear-k≠1	0.508 (0.417,0.620)	1.013 (0.984,1.043)
Method 2: log-linear-k=1	1	0.914 (0.898,0.929)

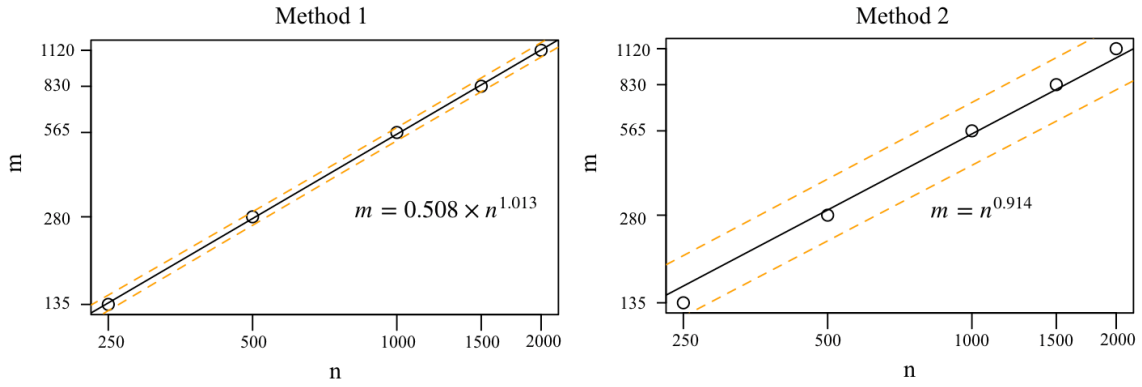

Figure 6.10: Fitted linear regression lines depicting the relationship between the block size m and the sample size n calculated by subsampling bootstrap for threshold parameter e when the data generating model is the sigmoid model with steepness=5. Left: Method 1 ($k \neq 1$). Right: Method 2 ($k = 1$)

6.2.4 Quadratic model

As shown by Figure 6.11, the block size $m = 130$ for $n = 250$; $m = 275$ for $n = 500$; $m = 575$ for $n = 1000$; $m = 830$ for $n = 1500$ and $m = 1100$ for $n = 2000$ satisfy that the coverage probabilities of percentile CIs for threshold parameter e are close to the nominal level 0.95. Again, the block size ms found are very similar to those for sigmoid models.

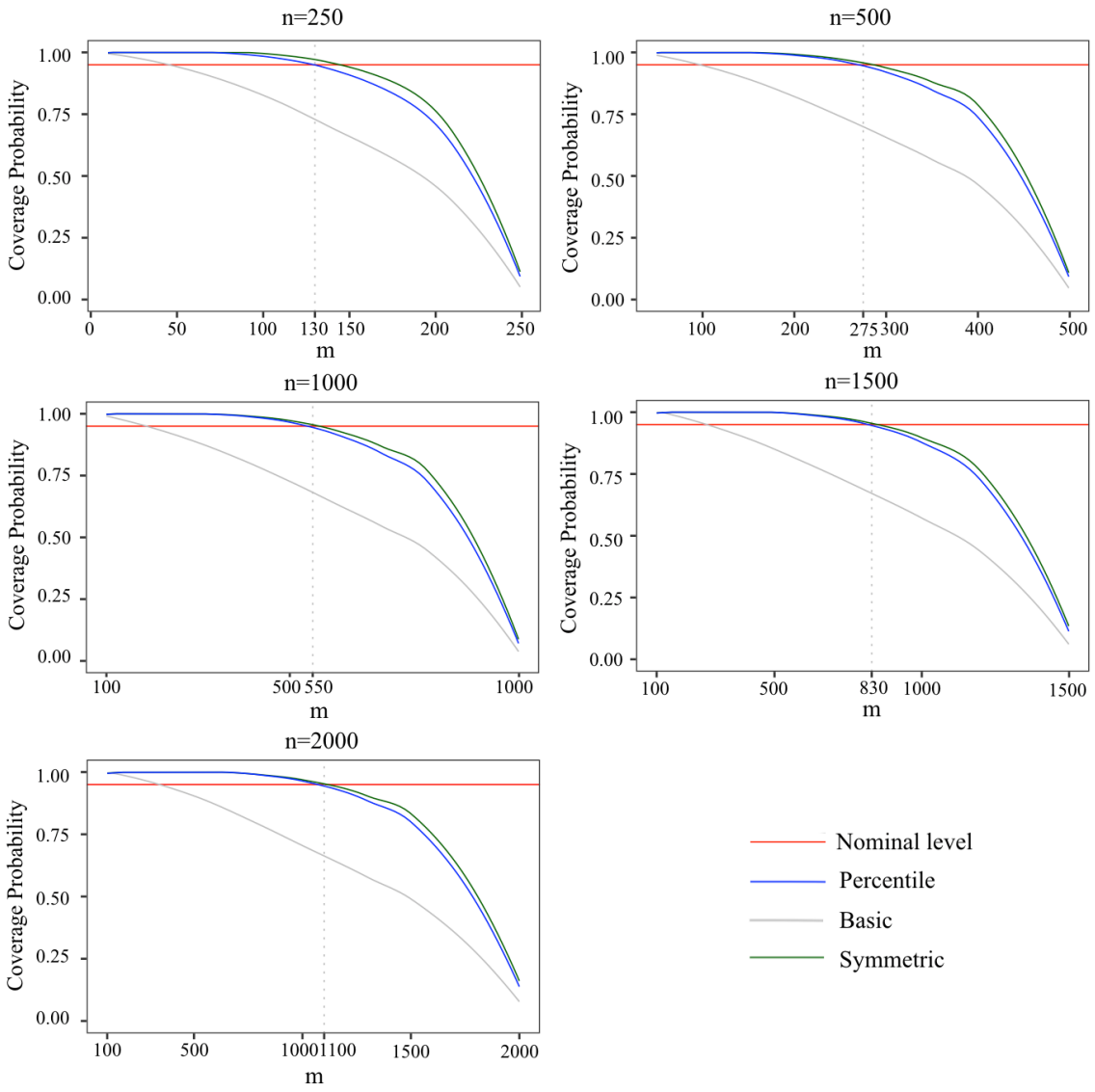

Figure 6.11: Estimated coverage probabilities of percentile, basic and symmetric confidence intervals given by subsampling bootstrap method from 10000 Monte Carlo runs when the true model is a quadratic model.

Table 6.6 shows the estimated k and p and the corresponding confidence intervals calculated by the two methods described previously. And Figure 6.12 shows the linear regression lines fitted.

Table 6.6: Estimated k and p and its confidence intervals calculated by 2 different methods when the data generating model is the quadratic model.

	k	p
Method 1: log-linear-k≠1	0.461 (0.368,0.578)	1.025 (0.991,1.058)
Method 2: log-linear-k=1	1	0.911 (0.893,0.929)

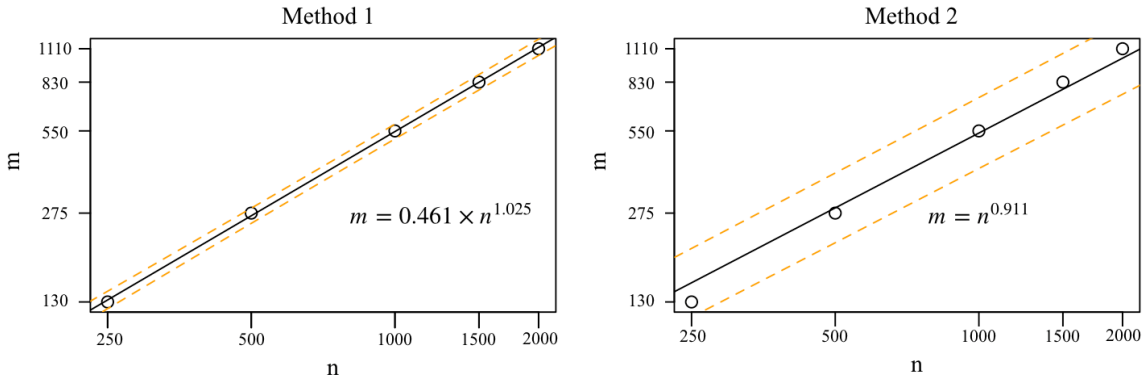

Figure 6.12: Fitted linear regression models depicting the relationship between the block size m and the sample size n by subsampling bootstrap for threshold parameter e when the data generating model is the quadratic model. Left: Method 1 ($k \neq 1$). Right: Method 2 ($k = 1$)

6.2.5 A General Rule

In this section, we seek to develop a general rule for selecting the block size m in subsampling when the data generating models are unknown. Table 6.7 summarizes the value of block size ms with estimated coverage probabilities of percentile confidence intervals closest to the nominal level 0.95. We found that when the working model and the data generating model are misspecified (sig1, sig5, qua), the selected block size m is very similar under a certain sample size n. When the working model and the data generating model are both step linear regression models, the block size ms chosen are substantially smaller than those chosen when models are misspecified.

Table 6.7: Block size m satisfying that the estimated coverage probabilities of percentile confidence intervals by subsampling bootstrap are close to the nominal level 0.95 under certain sample size n.

	m	n	m	n	m	n	m	n	m	n
step	90	250	175	500	355	1000	515	1500	680	2000
sig1	140	250	280	500	575	1000	830	1500	1120	2000
sig5	135	250	280	500	565	1000	830	1500	1120	2000
qua	130	250	275	500	575	1000	830	1500	1100	2000

In the previous sections, for each of the four data generating models, we have used two methods to develop data-adaptive rules for choosing m. Since Method 1 generally gives more accurate m, we consider only Method 1 here. We fit a linear mixed model with random intercept and random slope based on Model 6.1 to investigate the relationship between the block size m and the sample size n considering all the four data generating models. We consider the four data generating models as four different conditions, and then use linear mixed model with random slope and random intercept to fit the data consisting of the block size m and the sample size n summarized in Table 6.7.

Figure 6.13 shows the linear regression lines fitted for each data generating model by Method 1 in the left, while the linear mixed model fitted based on all the four data generating models in the right. The estimated prediction rule of subsampling bootstrap for selecting

the block size m when the data generating models are unknown is:

$$m = 0.481 \times n^{1.004}. \tag{6.3}$$

The 95% confidence interval for k is (0.424,0.547), and the 95% confidence interval for p is (0.984, 1.02). In the next chapter we will use this rule as the general rule for selecting m when the data generating models are unknown. But we note that this rule violates the condition that $m = o(n)$. Furthermore, as noted for m-bootstrap, this result comes from a simple compromise between having the model correctly specified and having the model misspecified. Both issues warrant further investigation and are out of scope for this thesis.

Again, we consider another prediction rule for selecting the block size e by fitting a linear mixed model considering the three misspecification scenarios only:

$$m = 0.508 \times n^{1.013}. \tag{6.4}$$

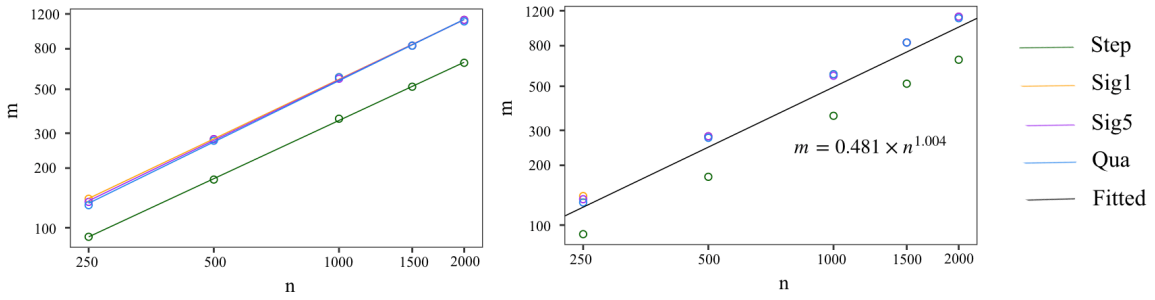

Figure 6.13: Left: Fitted linear regression models by Method 1 depicting the relationship between the block size m and the sample size n by subsampling bootstrap for threshold parameter e when the data generating models are step model, sigmoid models and quadratic model, respectively. Right: Fitted linear mixed model by Method 1 depicting the relationship between the block size m and the sample size n considering all the data points of the four data generating models.

Chapter 7

COMPARISON OF N-BOOTSTRAP, M-BOOTSTRAP AND SUBSAMPLING CONFIDENCE INTERVAL METHODS

In this final chapter we will use the five data generating models listed in Chapter 3 to compare the coverage probabilities and width of n-bootstrap percentile and symmetric CI, m-bootstrap percentile CI using the general data-adaptive rule

$$m = 0.901 \times n^{0.988},$$

and subsampling percentile CI using the general data-adaptive rule

$$m = 0.481 \times n^{1.004}.$$

Note that both rules are compromises between the case where the working model is correctly specified and cases where the working model is misspecified, and that the compromises favor the case where the working is misspecified. In addition, the rule were optimized for the coverage of $\hat{e}$. We will also use the prediction rules derived only using the misspecification scenarios: $m = n$ for m-bootstrap, and $m = 0.508 \times n^{1.013}$ for subsampling bootstrap, respectively.

Table 7.1, 7.2, 7.3 and 7.4 show the performance of n-bootstrap, m-bootstrap and subsampling confidence intervals when sample sizes are 2000, 1000, 500, and 200, respectively. And for each of the m-bootstrap and subsampling cases, we consider a general prediction rule derived by all generating models (referred to here as mixed-rule) and a prediction rule using only the misspecification scenarios (referred to here as mis-rule).

As shown by mis-rule in Tables 7.1, 7.2, 7.3, 7.4, the m-bootstrap is equivalent to the n-bootstrap, and the subsampling bootstrap CI only gives coverage probabilities close to the nominal level 0.95 for e but undercoverage for other parameters. This is because when

Table 7.1: Estimated coverage probabilities, width and ratio (in brackets) of percentile and symmetric confidence intervals calculated by n-bootstrap, and only percentile confidence intervals by m-bootstrap and subsampling bootstrap given by the prediction rule when the data generating models are step model, sigmoid models and quadratic model of size 2000 from 10000 Monte Carlo runs. The width of symmetric CIs is compared with the width of the percentile CIs (ref), and the width of percentile CIs calculated by m-bootstrap and subsampling bootstrap based on the prediction rule is compared with those calculated by n-bootstrap.

			Mixed-Rule		Mis-Rule	
	n-bootstrap		m-bootstrap	subsampling	m-bootstrap	subsampling
	perc	symm	perc	perc	perc	perc
e						
step	0.82 (ref:0.06)	0.96 (1.23)	0.86 (1.18)	0.87 (1.17)	0.82 (1.00)	0.82 (0.96)
sig1	0.97 (ref:1.23)	0.97 (1.25)	0.98 (1.08)	0.98 (1.01)	0.97 (1.00)	0.95 (0.91)
sig5	0.96 (ref:0.35)	0.97 (1.26)	0.98 (1.08)	0.96 (1.03)	0.96 (1.00)	0.95 (0.93)
sig15	0.96 (ref:0.17)	0.96 (1.25)	0.97 (1.09)	0.97 (1.06)	0.96 (1.00)	0.95 (0.94)
qua	0.96 (ref:0.20)	0.96 (1.24)	0.97 (1.08)	0.97 (1.05)	0.96 (1.00)	0.94 (0.91)
β						
step	0.95 (ref:0.05)	0.95 (1.00)	0.97 (1.10)	0.95 (1.01)	0.95 (1.00)	0.92 (0.89)
sig1	0.87 (ref:0.05)	0.93 (1.06)	0.89 (1.10)	0.87 (1.01)	0.87 (1.00)	0.83 (0.89)
sig5	0.94 (ref:0.05)	0.95 (1.01)	0.96 (1.10)	0.94 (1.01)	0.94 (1.00)	0.90 (0.89)
sig15	0.94 (ref:0.05)	0.95 (1.01)	0.97 (1.10)	0.95 (1.01)	0.94 (1.00)	0.91 (0.89)
qua	0.96 (ref:0.34)	0.96 (1.02)	0.97 (1.08)	0.97 (1.04)	0.96 (1.00)	0.93 (0.90)
α_2						
step	0.95 (ref:0.03)	0.95 (1.00)	0.97 (1.10)	0.95 (1.01)	0.95 (1.00)	0.92 (0.89)
sig1	0.95 (ref:0.03)	0.95 (1.01)	0.97 (1.10)	0.95 (1.01)	0.95 (1.00)	0.92 (0.89)
sig5	0.95 (ref:0.03)	0.95 (1.00)	0.97 (1.10)	0.95 (1.01)	0.95 (1.00)	0.92 (0.89)
sig15	0.95 (ref:0.03)	0.95 (1.00)	0.97 (1.10)	0.95 (1.01)	0.95 (1.00)	0.92 (0.89)
qua	0.96 (ref:0.14)	0.96 (1.02)	0.98 (1.10)	0.96 (1.01)	0.96 (1.00)	0.93 (0.89)
α_1						
step	0.95 (ref:0.04)	0.95 (1.00)	0.97 (1.10)	0.95 (1.01)	0.95 (1.00)	0.92 (0.89)
sig1	0.97 (ref:0.06)	0.97 (1.13)	0.98 (1.08)	0.97 (1.01)	0.97 (1.00)	0.95 (0.91)
sig5	0.96 (ref:0.04)	0.96 (1.05)	0.98 (1.09)	0.95 (1.01)	0.96 (1.00)	0.94 (0.90)
sig15	0.95 (ref:0.04)	0.96 (1.02)	0.97 (1.10)	0.95 (1.01)	0.95 (1.00)	0.92 (0.89)
qua	0.96 (ref:0.24)	0.96 (1.10)	0.98 (1.10)	0.97 (1.03)	0.96 (1.00)	0.94 (0.91)

the subsampling tries to decrease coverage for e to 0.95, it ends up decreasing coverage for other parameters as well.

As shown by mixed-rule in Tables 7.1, 7.2, 7.3, 7.4, for the coverage of e, m-bootstrap and subsampling have similar performances. They are both wider and have better coverage probabilities than n-bootstrap percentile CI. The subsampling method is somewhat better

Table 7.2: Estimated coverage probabilities, width and ratio (in brackets) of percentile and symmetric confidence intervals calculated by n-bootstrap, and only percentile confidence intervals by m-bootstrap and subsampling bootstrap given by the prediction rule when the data generating models are step model, sigmoid models and quadratic model of size 1000 from 10000 Monte Carlo runs. The width of symmetric CIs is compared with the width of the percentile CIs (ref), and the width of percentile CIs calculated by m-bootstrap and subsampling bootstrap based on the prediction rule is compared with those calculated by n-bootstrap.

			Mixed-Rule		Mis-Rule	
	n-bootstrap		m-bootstrap	subsampling	m-bootstrap	subsampling
	perc	symm	perc	perc	perc	perc
e						
step	0.83 (ref:0.12)	0.96 (1.23)	0.87 (1.18)	0.88 (1.18)	0.83 (1.00)	0.84 (0.99)
sig1	0.97 (ref:1.57)	0.97 (1.26)	0.98 (1.08)	0.98 (1.02)	0.97 (1.00)	0.96 (0.92)
sig5	0.96 (ref:0.48)	0.97 (1.25)	0.98 (1.09)	0.97 (1.05)	0.96 (1.00)	0.95 (0.94)
sig15	0.95 (ref:0.23)	0.97 (1.25)	0.97 (1.11)	0.97 (1.09)	0.95 (1.00)	0.95 (0.97)
qua	0.96 (ref:0.26)	0.97 (1.23)	0.97 (1.10)	0.97 (1.06)	0.96 (1.00)	0.95 (0.97)
β						
step	0.94 (ref:0.07)	0.94 (1.00)	0.96 (1.10)	0.95 (1.01)	0.94 (1.00)	0.91 (0.90)
sig1	0.84 (ref:0.07)	0.92 (1.08)	0.87 (1.09)	0.85 (1.01)	0.84 (1.00)	0.81 (0.90)
sig5	0.93 (ref:0.07)	0.94 (1.01)	0.95 (1.10)	0.93 (1.01)	0.93 (1.00)	0.89 (0.90)
sig15	0.94 (ref:0.07)	0.94 (1.01)	0.96 (1.10)	0.94 (1.01)	0.94 (1.00)	0.91 (0.90)
qua	0.96 (ref:0.49)	0.96 (1.01)	0.97 (1.09)	0.95 (0.97)	0.96 (1.00)	0.92 (0.85)
α_2						
step	0.95 (ref:0.04)	0.95 (1.00)	0.97 (1.10)	0.95 (1.02)	0.95 (1.00)	0.92 (0.90)
sig1	0.95 (ref:0.04)	0.95 (1.01)	0.97 (1.10)	0.96 (1.02)	0.95 (1.00)	0.93 (0.90)
sig5	0.95 (ref:0.04)	0.95 (1.01)	0.97 (1.10)	0.96 (1.02)	0.95 (1.00)	0.92 (0.90)
sig15	0.95 (ref:0.04)	0.95 (1.00)	0.97 (1.10)	0.96 (1.02)	0.95 (1.00)	0.92 (0.90)
qua	0.96 (ref:0.21)	0.96 (1.02)	0.98 (1.10)	0.97 (1.02)	0.96 (1.00)	0.94 (0.90)
α_1						
step	0.95 (ref:0.05)	0.95 (1.01)	0.96 (1.10)	0.95 (1.01)	0.95 (1.00)	0.92 (0.90)
sig1	0.97 (ref:0.09)	0.97 (1.13)	0.98 (1.08)	0.97 (1.01)	0.97 (1.00)	0.95 (0.91)
sig5	0.96 (ref:0.06)	0.95 (1.04)	0.97 (1.09)	0.96 (1.01)	0.96 (1.00)	0.93 (0.91)
sig15	0.95 (ref:0.05)	0.94 (1.02)	0.97 (1.10)	0.95 (1.01)	0.95 (1.00)	0.92 (0.90)
qua	0.96 (ref:0.33)	0.96 (1.09)	0.98 (1.11)	0.97 (1.02)	0.96 (1.00)	0.94 (0.93)

than the m-bootstrap method because the confidence intervals are in general shorter.

For the coverage of β, symmetric CI has the best performance. It is the only method that provides coverage above 90% when the data generating models is sigmoid with steepness=1. Under the other data generating models, it has the same width as the other methods.

For the coverage of α_1, the subsampling CI and the n-bootstrap percentile CI have good

Table 7.3: Estimated coverage probabilities, width and ratio (in brackets) of percentile and symmetric confidence intervals calculated by n-bootstrap, and only percentile confidence intervals by m-bootstrap and subsampling bootstrap given by the prediction rule when the data generating models are step model, sigmoid models and quadratic model of size 500 from 10000 Monte Carlo runs. The width of symmetric CIs is compared with the width of the percentile CIs (ref), and the width of percentile CIs calculated by m-bootstrap and subsampling bootstrap based on the prediction rule is compared with those calculated by n-bootstrap.

			Mixed-Rule		Mis-Rule	
	n-bootstrap		m-bootstrap	subsampling	m-bootstrap	subsampling
	perc	symm	perc	perc	perc	perc
e						
step	0.83 (ref:0.24)	0.97 (1.23)	0.86 (1.16)	0.87 (1.17)	0.83 (1.00)	0.84 (1.00)
sig1	0.97 (ref:2.02)	0.97 (1.26)	0.98 (1.08)	0.97 (1.03)	0.97 (1.00)	0.96 (0.94)
sig5	0.95 (ref:0.63)	0.97 (1.25)	0.97 (1.09)	0.97 (1.09)	0.95 (1.00)	0.95 (0.96)
sig15	0.93 (ref:0.36)	0.96 (1.23)	0.95 (1.12)	0.95 (1.12)	0.93 (1.00)	0.94 (0.98)
qua	0.95 (ref:0.35)	0.96 (1.22)	0.97 (1.08)	0.97 (1.08)	0.95 (1.00)	0.94 (0.96)
β						
step	0.95 (ref:0.10)	0.95 (1.00)	0.97 (1.09)	0.95 (1.02)	0.95 (1.00)	0.92 (0.91)
sig1	0.81 (ref:0.11)	0.91 (1.10)	0.84 (1.09)	0.82 (1.02)	0.81 (1.00)	0.78 (0.91)
sig5	0.93 (ref:0.10)	0.94 (1.02)	0.95 (1.09)	0.95 (1.09)	0.93 (1.00)	0.90 (0.91)
sig15	0.94 (ref:0.10)	0.95 (1.01)	0.96 (1.09)	0.96 (1.09)	0.94 (1.00)	0.92 (0.91)
qua	0.95 (ref:0.68)	0.95 (1.01)	0.97 (1.10)	0.97 (1.10)	0.95 (1.00)	0.93 (0.90)
α_2						
step	0.95 (ref:0.05)	0.95 (1.00)	0.96 (1.09)	0.95 (1.03)	0.95 (1.00)	0.92 (0.91)
sig1	0.95 (ref:0.05)	0.95 (1.01)	0.97 (1.09)	0.96 (1.03)	0.95 (1.00)	0.93 (0.92)
sig5	0.95 (ref:0.05)	0.95 (1.01)	0.97 (1.09)	0.97 (1.09)	0.95 (1.00)	0.92 (0.92)
sig15	0.95 (ref:0.05)	0.95 (1.00)	0.97 (1.09)	0.97 (1.09)	0.95 (1.00)	0.92 (0.92)
qua	0.97 (ref:0.31)	0.96 (1.03)	0.98 (1.10)	0.98 (1.10)	0.97 (1.00)	0.95 (0.92)
α_1						
step	0.95 (ref:0.08)	0.95 (1.01)	0.97 (1.10)	0.95 (1.02)	0.95 (1.00)	0.92 (0.91)
sig1	0.96 (ref:0.12)	0.96 (1.12)	0.97 (1.09)	0.96 (1.02)	0.96 (1.00)	0.94 (0.92)
sig5	0.95 (ref:0.08)	0.95 (1.04)	0.97 (1.09)	0.97 (1.09)	0.95 (1.00)	0.93 (0.92)
sig15	0.95 (ref:0.08)	0.95 (1.02)	0.96 (1.09)	0.96 (1.09)	0.95 (1.00)	0.92 (0.91)
qua	0.96 (ref:0.45)	0.95 (1.07)	0.97 (1.06)	0.97 (1.06)	0.96 (1.00)	0.93 (0.89)

coverage and are the shortest.

Finally, for the coverage of α_2, all methods perform similarly.

In conclusion, if one looks for a method that is robust under all scenarios, we recommend the subsampling percentile CI using our general data-adaptive rule for picking block size due to its performance and relatively well-developed theory behind the methods. On the

Table 7.4: Estimated coverage probabilities, width and ratio (in brackets) of percentile and symmetric confidence intervals calculated by n-bootstrap, and only percentile confidence intervals by m-bootstrap and subsampling bootstrap given by the prediction rule when the data generating models are step model, sigmoid models and quadratic model of size 200 from 10000 Monte Carlo runs. The width of symmetric CIs is compared with the width of the percentile CIs (ref), and the width of percentile CIs calculated by m-bootstrap and subsampling bootstrap based on the prediction rule is compared with those calculated by n-bootstrap.

			Mixed-Rule		Mis-Rule	
	n-bootstrap		m-bootstrap	subsampling	m-bootstrap	subsampling
	perc	symm	perc	perc	perc	perc
e						
step	0.84 (ref:0.64)	0.97 (1.22)	0.88 (1.18)	0.89 (1.18)	0.84 (1.00)	0.86 (1.01)
sig1	0.97 (ref:2.81)	0.97 (1.27)	0.98 (1.06)	0.98 (1.03)	0.97 (1.00)	0.96 (0.96)
sig5	0.94 (ref:1.07)	0.97 (1.23)	0.96 (1.11)	0.96 (1.10)	0.94 (1.00)	0.95 (0.98)
sig15	0.91 (ref:0.75)	0.96 (1.22)	0.94 (1.16)	0.94 (1.15)	0.91 (1.00)	0.92 (1.00)
qua	0.94 (ref:0.52)	0.97 (1.23)	0.96 (1.08)	0.96 (1.12)	0.94 (1.00)	0.94 (1.00)
β						
step	0.94 (ref:0.16)	0.95 (1.01)	0.96 (1.09)	0.95 (1.03)	0.94 (1.00)	0.93 (0.93)
sig1	0.77 (ref:0.18)	0.91 (1.14)	0.80 (1.08)	0.78 (1.04)	0.77 (1.00)	0.75 (0.94)
sig5	0.92 (ref:0.16)	0.94 (1.03)	0.94 (1.09)	0.93 (1.04)	0.92 (1.00)	0.90 (0.94)
sig15	0.94 (ref:0.16)	0.94 (1.01)	0.96 (1.09)	0.95 (1.03)	0.94 (1.00)	0.92 (0.94)
qua	0.95 (ref:1.07)	0.95 (1.02)	0.97 (1.10)	0.96 (1.06)	0.95 (1.00)	0.93 (0.91)
α_2						
step	0.95 (ref:0.08)	0.95 (1.01)	0.96 (1.09)	0.96 (1.05)	0.95 (1.00)	0.93 (0.95)
sig1	0.95 (ref:0.09)	0.95 (1.01)	0.97 (1.09)	0.96 (1.06)	0.95 (1.00)	0.94 (0.96)
sig5	0.95 (ref:0.09)	0.95 (1.01)	0.97 (1.09)	0.96 (1.05)	0.95 (1.00)	0.93 (0.95)
sig15	0.95 (ref:0.08)	0.95 (1.01)	0.96 (1.09)	0.96 (1.05)	0.95 (1.00)	0.93 (0.95)
qua	0.97 (ref:0.52)	0.96 (1.04)	0.98 (1.09)	0.97 (1.05)	0.97 (1.00)	0.95 (0.95)
α_1						
step	0.95 (ref:0.12)	0.96 (1.02)	0.97 (1.09)	0.96 (1.03)	0.95 (1.00)	0.94 (0.93)
sig1	0.95 (ref:0.19)	0.96 (1.13)	0.97 (1.07)	0.96 (1.03)	0.95 (1.00)	0.94 (0.94)
sig5	0.95 (ref:0.14)	0.96 (1.04)	0.97 (1.08)	0.96 (1.03)	0.95 (1.00)	0.94 (0.94)
sig15	0.95 (ref:0.13)	0.96 (1.02)	0.97 (1.09)	0.96 (1.03)	0.95 (1.00)	0.94 (0.93)
qua	0.96 (ref:0.66)	0.95 (1.08)	0.97 (1.09)	0.97 (1.06)	0.96 (1.00)	0.95 (0.97)

other hand, if one would like a method that works well when the model is misspecified, we recommend percentile n-bootstrap CI for all the parameters except β and symmetric n-bootstrap CI for β. That β requires special treatment may be related to the fact that its estimated convergence rate is close to 1/2, while the other parameters have estimated convergence rates close to 1/3.

Other than the methods we have compared here, smoothed bootstrap (Peter Hall and Romano, 1989) is also worth considering. The validity of the smoothed bootstrap when the model is correctly specified was proved by Seijo et al. (2011) in a simple threshold model, and may be extended to models with covariates. Smoothed bootstrap performed well in terms of coverage in the simulations of Yu (2014) and Seijo et al. (2011), offering advantages over m-bootstrap and subsampling in some scenarios. It will be interesting to compare its performance with other methods when the model is not correctly specified. Another method that warrants further investigation is the test-inversion CI methods proposed by Banerjee and McKeague (2007), which may be shorter than the subsampling CI but are only developed for the threshold parameter. In order to further improve the coverage probabilities of m-bootstrap and subsampling for e, double bootstrap-type methods (Chakraborty et al., 2013) are also worth considering, but it will not help improve the coverage for β if m is picked based on the coverage of e. Exploration along these directions is out of scope of this thesis, but will be worth consideration in future studies.

BIBLIOGRAPHY

Banerjee, M. and McKeague, I.W. (2007), "Confidence sets for split points in decision trees," *The Annals of Statistics*, 35, 543–574.

Bickel, P.J. and Sakov, A. (2008), "On the choice of m in the m out of n bootstrap and confidence bounds for extrema," *Statistica Sinica*, pp. 967–985.

Bickel, P.J., Götze, F. and van Zwet, W.R. (1997), "Resampling fewer than n observations: gains, losses, and remedies for losses," *Statistica Sinica*, 7, 1–31.

Chakraborty, B., Laber, E.B. and Zhao, Y. (2013), "Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme," *Biometrics*, 69, 714–723.

Delgado, M.A., Rodrıguez-Poo, J.M. and Wolf, M. (2001), "Subsampling inference in cube root asymptotics with an application to Manskis maximum score estimator," *Economics Letters*, 73, 241–250.

Efron, B. (1979), "Bootstrap Methods: Another Look At The Jackknife," *The Annals of Statistics*, 7, 1–26.

Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York, NY.

Elder, A. and Fong, Y. (2019), "Estimation and Inference for Upper Hinge Regression Models," *Environmental and Ecological Statistics*, 26, 287–302.

Fong, Y. (2019), "Fast Bootstrap Confidence Intervals for Continuous Threshold Linear Regression," *Journal of Computational and Graphical Statistics*, 28, 466–470.

Peter Hall, T.J.D. and Romano, J.P. (1989), "On smoothing and the bootstrap," *The Annals of Statistics*, 17, 692–704.

Politis, D.N. and Romano, J.P. (1994), "Large sample confidence regions based on subsamples under minimal assumptions," *The Annals of Statistics*, 22, 2031–2050.

Seijo, E., Sen, B. et al (2011), "Change-point in stochastic design regression and the bootstrap," *The Annals of Statistics*, 39, 1580–1607.

Yu, P. (2014), "The bootstrap in threshold regression," *Econometric Theory*, 30, 676–714.

VITA

Shuangcheng Hua entered University of Washington in September 2018. She is a master student majoring in Biostatistics. Before that, she received the degree of Bachelor of Economics in Finance at the Renmin University of China.

She welcomes your comments to `chloehua@uw.edu`.