筑波大学

博士（医学）学位論文

Modified likelihood ratio test in accelerated failure time models for small-sample data

（加速モデルによる小標本データ解析のための尤度比検定の改良）

２０２０

筑波大学大学院博士課程人間総合科学研究科

石井　亮太

Contents

List of Tables

List of Figures

Preface

Survival data analysis is a standard method to evaluate the effect of a treatment on the survival time defined by the time until an event (death, recurrence, etc.) happens. In particular, the Cox proportional hazards (PH) model[1] is the most popular regression model in the survival data analysis. However, the Cox PH model has two important limitations: 1) the treatment effect is expressed as a hazard ratio and is difficult to interpret on the scale of survival time and 2) the model assumes that the hazard ratio between two treatment groups is constant over time (proportional hazard assumption).

The accelerated failure time (AFT) model[2,3] is an attractive alternative to the Cox PH model, since the treatment effect in the AFT model is interpreted as a ratio of survival time between treatment groups and the AFT model does not require the proportional hazard assumption. To apply the AFT model, we have to include all important covariates in the analysis model and specify the true distribution of the survival time. If we omit an important covariate and/or misspecify the distribution, the model misspecification problem occurs and wrong analysis results would be provided. In practice, it is impossible to specify the true distribution of the survival time, because the true distribution is unknown. Furthermore, we cannot include uncollected covariates and unknown prognostic factors in the analysis model. Thus, we cannot avoid the model misspecification problem. As a serious effect of the model misspecification problem, the Type-I error rate is not controlled at the nominal level (usually 0.05). Here, the Type-I error rate is the probability of the rejection of a null hypothesis when it is actually true, that is, false-positive rate.

Many researchers provided asymptotic corrections of various test statistics to control the Type-I error rate even under model misspecification. However, their corrected statistical tests do not have good performance in small samples, since the performance of these tests is ensured only when the sample size is sufficiently large. Actually, the use of these corrected tests in small-sample clinical trials cannot control the Type-I error rate at the nominal level. Hence, corrected statistics which work well in small samples are required.

The Bartlett adjustment is a popular approach for small-sample correction of the likelihood ratio statistic under the null hypothesis. The Bartlett-adjusted likelihood ratio test can control the Type-I error rate at the nominal level even under model misspecification and small samples. However, the Bartlett adjustment factor is defined by the expectation with respect to the unknown true distribution. Thus, it is impossible to derive the adjustment factor analytically under model misspecification.

In this study, we propose a novel likelihood ratio test which controls the Type-I error rate at the nominal level even under model misspecification and small samples. Our proposed method is based on the Bartlett adjustment and uses the non-parametric bootstrap method to estimate the adjustment factor. We apply the proposed method to the AFT model when the distribution of the survival time is misspecified and/or an important covariate is omitted in small samples. Our simulation results show that the Type-I error rate for the proposed method is close to the nominal level, although the existing methods result in substantial inflation of the Type-I error rate.

This dissertation is based on Ishii et al.[4] and organized as follows. Chapter

1 provides a background of this study. We explain the AFT model in Chapter 2. In Chapter 3, we outline the likelihood ratio test and adjustments of the likelihood ratio test, as well as the ordinary (naive) Wald test and robust Wald test. In Chapter 4, we describe our proposed method which applies the non-parametric bootstrap method to estimate the expectation of the likelihood ratio statistic in the finite sample size. In Chapter 5, we show the performances of the proposed method through simulation studies when the AFT model is used. In Chapter 6, we apply the proposed method to a real dataset. Finally, Chapter 7 discusses the results and concludes. We provide a sample code of our proposed method in Appendix A.

Chapter 1

Introduction

The Cox proportional hazards (PH) model[1] is the most popular regression model in the survival data analysis. The accelerated failure time (AFT) model[2,3] is a flexible alternative model and is better than the Cox PH model in some respects.[5–7] For example, the Cox PH model assumes that covariates affect the hazard function, making it difficult to interpret the parameter estimates. In contrast, the AFT model is a linear regression model for the logarithm of the event time, and its parameter estimate is intuitive. Furthermore, the exponential of the group-effect parameter in a two-group comparison can be interpreted as the ratio of expected survival time between two groups. Hence, the AFT model is widely used in actual clinical trials.[8–10] However, the AFT model requires specifying the distribution of an error term and runs the risk of model misspecification. Gosho et al.[11] shows that the Type-I error rate cannot be controlled at the nominal level under model misspecification in a two-group comparison based on the AFT model.

To address the issue of model misspecification, corrections of the asymptotic distribution of various statistics have been established. For example, the robust (empirical) covariance estimator[12,13] is a popular approach to provide a consistent estimator for the asymptotic variance-covariance matrix of the maximum likelihood estimator under model misspecification, and the estimator can be applied to Wald statistics. In fact, the function `survreg` in the package `survival` of the statistical software `R`[14] has an option to provide a robust variance estimate in the analysis using the AFT model. On the other hand, in the context of the likelihood ratio test,[15] Kent[16] shows the asymptotic distribution of the likelihood ratio statistic under model misspecification and provides the correction factor to make the likelihood ratio test asymptotically valid. However, these corrected statistical tests do not have good performance in practice, because these corrections are based only on the asymptotic theory and do not ensure the performance in small samples.[17] In the context of generalized estimating equations,[18] some researchers have proposed covariance estimators adjusted for small-sample bias;[19,20] however, these estimators cannot be universally applied, for example, to survival data with censoring. Viraswami and Reid[21] and Lunardon[22] improve the accuracy of chi-square approximation of the likelihood ratio statistic under model misspecification. To apply the methods of Viraswami and Reid[21] and Lunardon,[22] we have to calculate fourth derivatives of the log-likelihood function. In practice, nuisance parameters often exist and the profile likelihood function is used instead of the ordinary likelihood function. In this case, higher order derivatives of profile log-likelihood function are complicated and it is difficult to calculate even if we use numerical differentiation, in particular, when the maximum likelihood estimator cannot be described explicitly.

We now focus on the case of a one-dimensional parameter of interest, which is seen in the typical randomized clinical trial comparing two groups. Our consideration is applicable to many situations since clinical trials comparing two groups are widely conducted. In this case, Kent[16] suggests the adjustment of the asymptotic distribution of the likelihood ratio statistic by its asymptotic expectation, where the asymptotic expectation indicates the limit of the expectation as the sample size approaches infinity. In addition, Kent[16] shows how to estimate the adjustment factor using information and Hessian matrices. On the other hand, dividing the likelihood ratio statistic by its expectation in the finite sample size is called Bartlett adjustment[23] and allows for highly accurate chi-square approximation.[22, 24, 25] Applying Slutsky's theorem, these two facts show that the likelihood ratio statistic divided by its expectation in the finite sample size is asymptotically valid under model misspecification and have more accurate approximation to the chi-squared distribution. Hence, it is needed to estimate the expectation in the finite sample size, and not the asymptotic expectation. To estimate the expectation, Loose et al.[26] and Cordeiro and Cribari-Neto[27] use the parametric bootstrap method, while Rocke[28] uses the bootstrap method for residuals. These two bootstrap methods require the correct model specification for resampling. However, we cannot know the correct model under model misspecification; therefore, we can apply neither of the two bootstrap methods.

In practice, owing to resource limitations and the limited size of the study population, many clinical trials have small sample sizes. In addition, it is sometimes difficult to specify the distribution of the error term correctly due to censored data and small samples. Both the problems of model misspecification and small samples arise quite naturally in various survival datasets; we will later focus on an acute myelogenous leukemia dataset[29, 30] as an example. This dataset of 23 patients is from the preliminary analysis of a clinical trial to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukemia. Patients were randomly assigned to receive maintenance chemotherapy consisting of cytarabine and 6-thioguanine for two days each month or to receive no maintenance therapy. The objective of the trial was to see if maintenance chemotherapy increased the length of remission. The small sample size does not provide a good approximation of the chi-squared distribution. Furthermore, we may be unable to build the correct mean structure, as the dataset includes only survival or censoring time, censoring status, and treatment group. A second example uses a randomized double-blind trial on 64 patients with severe aplastic anemia.[3,31] Patients were randomized to cyclosporine and methotrexate (CSP + MTX) or methotrexate alone (MTX). An endpoint was the time from assignment until the diagnosis of a life-threatening stage of acute graft versus host disease. The events of interest were observed only in the early period of the trial and many patients were censored. Hence, researchers are likely to fit an incorrect model to this dataset.

In this study, we propose a robust test to model misspecification in small samples by adjusting the likelihood ratio statistic by its expectation in the finite sample size. To estimate the expectation, the non-parametric bootstrap method is used in our proposed method. We evaluate the Type-I error rate for our proposed method through simulation studies in the case of a two-group comparison based on the AFT model. In addition, we provide the sample code of `R` to apply our proposed test in the appendix.

Chapter 2

AFT model

We introduce the Cox PH model before the AFT model, since the Cox PH model is routinely applied to survival data. The Cox PH model is expressed as

$$\log h(t) = \boldsymbol{\alpha}^T \boldsymbol{x}, \tag{2.1}$$

where $h(t)$ is a hazard function, $\boldsymbol{x}$ is a vector of explanatory variables, and $\boldsymbol{\alpha}$ is a vector of regression parameters. Hence, the regression parameters are interpreted in the scale of the hazard. For example, if α_g is a coefficient of the treatment group indicator, its exponential $\exp(\alpha_g)$ shows the hazard ratio between treatments. The important assumption in the Cox PH model is proportional hazard assumption which means the hazard ratio between two treatment groups is constant over time.

The AFT model is a linear model for the logarithm of the survival time T,

$$\log T = \boldsymbol{\alpha}^T \boldsymbol{x} + \eta, \tag{2.2}$$

where the error term η is assumed a specific parametric distribution. Unlike the Cox PH model, the AFT model assumes a direct relationship between survival time and explanatory variables, and thus, it is easy for physicians to interpret the estimator of regression parameters. In fact, if α_g is a coefficient of the treatment group indicator, its exponential $\exp(\alpha_g)$ shows the ratio of survival time between treatments.

We can deal with many types of hazard ratios between treatments through the specification of the distribution of $\exp(\eta)$ in the AFT model, while the hazard ratio is always constant in the Cox PH model. We explain three popular distributions used in the AFT model.[32] The three distributions can be applied in the `survreg` function in the `R` software[14] and `LIFEREG` procedure in the `SAS` software.[33]

Weibull distribution If $\exp(\eta)$ follows the Weibull distribution with shape parameter κ, the survival function $S(t)$, density function $f(t)$, and hazard function $h(t)$ of the AFT model (2.2) are expressed as follows:

$$\begin{aligned}
S(t) &= \exp\left\{-\left(\frac{t}{\exp(\boldsymbol{\alpha}^T\boldsymbol{x})}\right)^{1/\kappa}\right\}, \\
f(t) &= \frac{1}{\kappa t}\left(\frac{t}{\exp(\boldsymbol{\alpha}^T\boldsymbol{x})}\right)^{1/\kappa}\exp\left\{-\left(\frac{t}{\exp(\boldsymbol{\alpha}^T\boldsymbol{x})}\right)^{1/\kappa}\right\}, \\
h(t) &= \frac{1}{\kappa t}\left(\frac{t}{\exp(\boldsymbol{\alpha}^T\boldsymbol{x})}\right)^{1/\kappa}.
\end{aligned}$$

The Weibull distribution includes the exponential distribution as a special case $\kappa = 1$. It is clear that the Weibull distribution satisfies the proportional hazard assumption. In addition, the Weibull distribution is the only distribution that has both a proportional hazards representation and accelerated failure-time representation.[34] The default setting in the `survreg` function in the `R` software[14] and `LIFEREG` procedure in the `SAS` software[33] is the Weibull distribution.

Log-logistic distribution If $\exp(\eta)$ follows the log-logistic distribution with shape parameter κ (i.e., η follows the logistic distribution), the survival function $S(t)$, density function $f(t)$, and hazard function $h(t)$ of AFT model (2.2) are expressed as follows:

$$\begin{aligned}
S(t) &= \frac{1}{1+(t\exp(-\boldsymbol{\alpha}^T\boldsymbol{x}))^{1/\kappa}},\\
f(t) &= \frac{1}{\kappa t}\left(\frac{t}{\exp(\boldsymbol{\alpha}^T\boldsymbol{x})}\right)^{1/\kappa}\frac{1}{\left(1+(t\exp(-\boldsymbol{\alpha}^T\boldsymbol{x}))^{1/\kappa}\right)^2},\\
h(t) &= \frac{1}{\kappa t}\left(\frac{t}{\exp(\boldsymbol{\alpha}^T\boldsymbol{x})}\right)^{1/\kappa}\frac{1}{1+(t\exp(-\boldsymbol{\alpha}^T\boldsymbol{x}))^{1/\kappa}}.
\end{aligned}$$

The log-logistic distribution has a hazard function which is hump-shaped, that is, it increases initially and, then, decreases,[35] while the hazard function in the Weibull distribution is monotone function of t.

Log-normal distribution If $\exp(\eta)$ follows the log-normal distribution with shape parameter κ (i.e., η follows the normal distribution), the survival function $S(t)$, density function $f(t)$, and hazard function $h(t)$ of AFT model (2.2) are expressed as follows:

$$\begin{aligned}
S(t) &= 1-\Phi\left(\frac{\log t-\exp(\boldsymbol{\alpha}^T\boldsymbol{x})}{\kappa}\right),\\
f(t) &= \frac{1}{\sqrt{2\pi}\kappa t}\exp\left\{-\frac{(\log t-\exp(\boldsymbol{\alpha}^T\boldsymbol{x}))^2}{2\kappa^2}\right\},\\
h(t) &= \frac{1}{\sqrt{2\pi}\kappa t}\left[1-\Phi\left(\frac{\log t-\exp(\boldsymbol{\alpha}^T\boldsymbol{x})}{\kappa}\right)\right]^{-1}\exp\left\{-\frac{(\log t-\exp(\boldsymbol{\alpha}^T\boldsymbol{x}))^2}{2\kappa^2}\right\}.
\end{aligned}$$

Here, Φ is the cumulative distribution function of the standard normal distribution. If κ is large then the hazard function reaches maximum value early in life. Thus, the log-normal distribution is used to model situations when the risk of event is decreasing.[36]

To apply the AFT model, we have to correctly specify the distribution of the error term η and include all important covariates. Covariate omission and wrong specification of the error distribution cause the model misspecification problem.[11] We consider these two types of misspecifications.

Of course, the covariate omission problem also occurs in the Cox PH model (2.1). However, covariate omission in the Cox PH model does not yield a bias in the treatment effect estimator under the null hypothesis of no difference between treatments[37–39] and inflation of the Type-I error rate.[40] Hence, statistical tests using the Cox PH model are valid even under covariate omission.

In this study, we assume a randomized clinical study with two treatment groups and we consider the simple AFT model of the form $\log T = \alpha_0 + x_g \alpha_g + \eta$, where α_0 is an intercept parameter. The exponential $\exp(\eta)$ of the error term follows a distribution with shape parameter vector $\boldsymbol{\kappa}$. In this case, α_g is a parameter of interest and α_0 and $\boldsymbol{\kappa}$ are nuisance parameters, since our aim is two-group comparison.

We let $\log T = \alpha_0 + x_g \alpha_g + \eta$ be the true model. To differentiate models, we write the fitting model as $\log T = \beta_0 + x_g \beta_g + \varepsilon$. As is the case with the true model, β_0 is an intercept, β_g is a group effect parameter, and $\exp(\varepsilon)$ has various distribution with shape parameter vector $\boldsymbol{\sigma}$.

Let T_i be the observed time for subject $i = 1, \dots, N$. Let δ_i be the event indicator that takes one if T_i is an survival time and zero if T_i is a censoring time. The density function and survival function under the fitting model are denoted by $f(t; \boldsymbol{\theta})$ and $S_f(t; \boldsymbol{\theta})$, respectively, where $\boldsymbol{\theta} = (\beta_0, \beta_g, \boldsymbol{\sigma})$ is the parameter vector. Then, the log-likelihood function is defined by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left\{ \delta_i \log f(T_i; \boldsymbol{\theta}) + (1 - \delta_i) \log S_f(T_i; \boldsymbol{\theta}) \right\}. \tag{2.3}$$

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is defined through the maximization of $\ell(\boldsymbol{\theta})$.

Chapter 3

Likelihood ratio test and Wald test

3.1 Ordinary likelihood ratio test

Let $(t_1, \delta_1), \ldots, (t_N, \delta_N)$ be independent observations and $\ell(\boldsymbol{\theta})$ be the log-likelihood function defined by (2.3). Assume that the true density function of the survival time is $g(t)$. Then, the fitting model is misspecified if $f(t; \boldsymbol{\theta}) \neq g(t)$ for all $\boldsymbol{\theta}$. We consider a null hypothesis $H_0 : \beta_g = 0$, which means that there is no difference between the two treatment groups. Thus, we partition parameter $\boldsymbol{\theta}$ into $\boldsymbol{\theta} = (\beta_g, \boldsymbol{\lambda})$ and consider β_g and $\boldsymbol{\lambda} = (\beta_0, \boldsymbol{\sigma})$ as a parameter of interest and a nuisance parameter vector, respectively. Let $\hat{\boldsymbol{\theta}} = (\hat{\beta}_g, \hat{\boldsymbol{\lambda}})$ be the unrestricted maximum likelihood estimator, and let $\hat{\boldsymbol{\theta}}_0 = (0, \hat{\boldsymbol{\lambda}}_0)$ be the maximum likelihood estimator under the null hypothesis $\beta_g = 0$. Then, the likelihood ratio statistic[15] $w = 2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_0)\}$ follows the chi-squared distribution χ_1^2 with one degree of freedom asymptotically under the null hypothesis when the fitting model is correctly specified. The ordinary likelihood ratio test is a chi-squared test performed by approximately fitting χ_1^2 to w. The reference distribution χ_1^2 is derived by assuming correct model specification; hence, the chi-square approximation would not be valid under model misspecification. Furthermore, as χ_1^2 is the asymptotic distribution of w, the distribution of w in small samples might differ from χ_1^2.

3.2 Adjustment of the likelihood ratio test under model misspecification

Let $g_0(t)$ be the true density function under the null hypothesis, and let $E_{g_0}[w]$ be the expectation of w under g_0. Assume that the expectation $E_{g_0}[w]$ converges to μ as N approaches infinity. According to Theorem 3.1 of Kent,[16] w/μ follows χ_1^2 asymptotically under the null hypothesis even when the fitting model is misspecified. This theorem means that the reference distribution of w should be a constant μ multiple of χ_1^2. The asymptotic expectation μ is often larger than one under model misspecification, while μ is one under correct model specification. Hence, when the fitting model is misspecified, the ordinary likelihood ratio test cannot control the Type-I error rate at the nominal level.

Kent[16] also shows another expression of μ using the information and Hessian matrices and the method to estimate μ. The negative multiple H of Hessian matrix and information matrix J are defined as follows:

$$H = -\left.\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \quad \text{and} \quad J = \sum_i \left.\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

where $\ell_i(\boldsymbol{\theta})$ is the log-likelihood function for subject i (i.e., $\ell(\boldsymbol{\theta}) = \sum_i \ell_i(\boldsymbol{\theta})$). Then, Kent[16] shows that μ is estimated by $[(H^{-1})_{\beta_g\beta_g}]^{-1}(H^{-1}JH^{-1})_{\beta_g\beta_g}$, where a matrix with subscript $\beta_g\beta_g$ indicates the diagonal element corresponding to

β_g. The adjusted likelihood ratio statistic obtained by the results of Kent[16] is $w_K = w/\{[(H^{-1})_{\beta_g\beta_g}]^{-1}(H^{-1}JH^{-1})_{\beta_g\beta_g}\}$. This adjusted statistic w_K is asymptotically follows χ^2_1 under the null hypothesis even when the fitting model is misspecified. Hence, the adjusted likelihood ratio test is performed by approximately fitting χ^2_1 to w_K.

3.3 Adjustment of the likelihood ratio test under model misspecification and small samples

Lunardon[22] extends the results of the likelihood ratio to the marginal composite likelihood ratio and provides the explicit formula of the extended likelihood ratio statistic. In particular, Lunardon[22] provides a Bartlett-corrected version of the marginal composite likelihood ratio under model misspecification. On the other hand, Viraswami and Reid[21] also improves the chi-square approximation of the likelihood ratio statistic under model misspecification. To apply the methods of Lunardon[22] and Viraswami and Reid,[21] we have to calculate fourth derivatives of the log-likelihood function. In practice, nuisance parameters often exist, and the profile likelihood function is considered instead of the ordinary likelihood function. In this case, from Leibniz's rule for differentiation, $\boldsymbol{\lambda}_{\beta_g}$ needs to be derived with respect to β_g, where $\boldsymbol{\lambda}_{\beta_g}$ is the maximum likelihood estimator of $\boldsymbol{\lambda}$ for a given value of β_g. However, such derivatives are complicated, and higher-order derivatives of the profile log-likelihood function are difficult to calculate even if we use numerical differentiation, particularly, when the maximum likelihood estimator cannot be described explicitly.

3.4 Bartlett adjustment

The expectation of w often differs from one owing to small samples or model misspecification. In such cases, approximating χ_1^2 to w is incorrect, since the expectation of χ_1^2 is exactly one. The adjustment $w/E_{g_0}[w]$ from the aspect of the expectation is called the Bartlett adjustment,[23] and the distribution of $w/E_{g_0}[w]$ is χ_1^2 up to an error term of order N^{-2} under the null hypothesis.[22] Hence, adjustment by $E_{g_0}[w]$ improve the chi-square approximation of w even under model misspecification and small samples. However, the adjustment factor cannot be calculated exactly under model misspecification, because we cannot know the true underlying distribution g_0.

To estimate the expectation, Loose et al.[26] and Cordeiro and Cribari-Neto[27] use the parametric bootstrap method, while Rocke[28] uses the bootstrap method for residuals. These two bootstrap methods require the correct model specification for resampling. However, we cannot know the correct model under model misspecification; therefore, we can apply neither of the two bootstrap methods.

3.5 Naive Wald test and robust Wald test

The Wald test is performed by approximating the test statistic $\hat{\beta}_g/\widehat{\mathrm{SE}}(\hat{\beta}_g)$ to the standard normal distribution $N(0,1)$. We present formulas of the two types of Wald test. The only difference between the two Wald test is the standard error estimator. Let H and J be the matrices defined above. Then, H^{-1} is called the naive variance estimator for $\hat{\boldsymbol{\theta}}$, while $H^{-1}JH^{-1}$ is called the robust variance estimator.[12,13] Let $\widehat{\mathrm{SE}}_{\mathrm{Naive}}(\hat{\beta}_g)$ and $\widehat{\mathrm{SE}}_{\mathrm{Robust}}(\hat{\beta}_g)$ be the square roots of $(H^{-1})_{\beta_g\beta_g}$ and $(H^{-1}JH^{-1})_{\beta_g\beta_g}$, respectively. The naive standard error estimator $\widehat{\mathrm{SE}}_{\mathrm{Naive}}(\hat{\beta}_g)$ is asymptotically valid only under correct model specification, while the robust standard error estimator $\widehat{\mathrm{SE}}_{\mathrm{Robust}}(\hat{\beta}_g)$ is asymptotically valid even under model misspecification. However, the robust variance estimator does not work well in small samples,[17] since $\widehat{\mathrm{SE}}_{\mathrm{Robust}}(\hat{\beta}_g)$ would have small-sample bias.

Chapter 4

Proposed adjustment of likelihood ratio test

We consider the adjustment $w/E_{g_0}[w]$ to address the model misspecification and small-sample problems simultaneously. It is generally impossible to obtain the unknown true distribution g_0 under the null hypothesis and we cannot calculate $E_{g_0}[w]$ exactly. Hence, we estimate the adjustment factor $E_{g_0}[w]$ using the non-parametric bootstrap method,[41] that is, the expectation is approximated as

$$E_{g_0}[w] = \int w g_0(y)dy \approx \int w\hat{g}_0(y)dy \approx \frac{1}{B}\sum_{j=1}^{B} w^{*j}, \tag{4.1}$$

where w^{*j} is calculated by a resampling from the empirical distribution $\hat{g}_0$ of g_0. We define the estimator $\hat{E}_{g_0}[w]$ by this formula.

$\hat{E}_{g_0}[w]$ would be a consistent estimator of $E_{g_0}[w]$, owing to the consistency of the bootstrap sampling mean.[42] Thus, $w/\hat{E}_{g_0}[w]$ asymptotically follows χ^2_1 from Slutsky's theorem.

We now explain our proposed procedure. In our proposed method, we use the bootstrap from the pooled sample to estimate $\hat{E}_{g_0}[w]$, where the pooled sample means the union of data from each group. Obviously, $w/\hat{E}_{g_0}[w]$ obtained from bootstrapping the pooled sample approximately follows χ_1^2 under the null hypothesis so that the Type-I error rate can be controlled. Note that, although bootstrapping by group can also provide the approximation under the null hypothesis, this resampling method under the alternative hypothesis makes the estimated expectation of w large owing to a difference between groups and leads an overcorrection of w decreasing power. On the other hand, even when the alternative hypothesis is true, the estimated expectation of w by bootstrapping the pooled sample would be near one; hence, we can avoid the overcorrection of w. Let $y_i = (t_i, \delta_i)$ be an observation from subject i. We denote the sample size in groups 0 and 1 by n_0 and n_1, respectively. Let $\{y_1, \ldots, y_{n_0}\}$ and $\{y_{n_0+1}, \ldots, y_{n_0+n_1}\}$ be observations from groups 0 and 1, respectively. Then, the pooled sample $\{y_1, \ldots, y_{n_0+n_1}\}$ is $\{y_1, \ldots, y_{n_0}, y_{n_0+1}, \ldots, y_{n_0+n_1}\}$ and the procedure can be described by the following steps:

1. Calculate the likelihood ratio statistic w for null hypothesis $\beta_g = 0$ from the original sample.

2. Resample $\{y_1^{*j}, \ldots, y_{n_0+n_1}^{*j}\}$ from the pooled sample $\{y_1, \ldots, y_{n_0+n_1}\}$, B times ($j = 1, 2, \ldots, B$).

3. Calculate the likelihood ratio statistic w^{*j} for null hypothesis $\beta_g = 0$ while considering $\{y_1^{*j}, \ldots, y_{n_0}^{*j}\}$ and $\{y_{n_0+1}^{*j}, \ldots, y_{n_0+n_1}^{*j}\}$ as groups 0 and 1, respectively.

4. Estimate $E_{g_0}[w]$ by $\hat{E}_{g_0}[w] = \sum_{j=1}^{B} w^{*j}/B$ and obtain the corrected likelihood ratio statistic $w^* = w/\hat{E}_{g_0}[w]$.

5. Perform the proposed test by approximately fitting the chi-squared distribution with one degree of freedom to w^*.

In addition, we provide a sample code of `R` to apply our proposed test in the appendix.

Chapter 5

Simulation study

In this chapter, we examine the performance of the chi-squared test based on the proposed test statistic, $w^* = w/\hat{E}_{g_0}[w]$, under model misspecification and small samples through simulation study. We assumed a randomized clinical study with two treatment groups, that evaluate the survival time extension in an active group (group 1) compared with a placebo group (group 0).

5.1 Simulation 1

5.1.1 Simulation design

This simulation aims to compare the performance of the proposed method with existing methods under misspecification of the distribution of the error term. The survival time T assumed a model of the form $\log T = \alpha_0 + \alpha_g x_g + \eta$, where $x_g = 0, 1$ is a group indicator. We let the distribution of $\exp(\eta)$ be the Weibull, log-logistic, or log-normal distribution with shape parameter κ. We let the true parameters be $\alpha_0 = 1, \kappa = 0.5, 1, 2$. Sample sizes were the same in each group and we let sample size n in each group be $n = 20, 50, 100, 200, 500$. We assumed that the censoring time in each group followed a common exponential distribution with parameter ρ. Parameter ρ was calculated so that the censoring rate in group 0 was about q. We considered two scenarios for q: $q = 0.2$ and $q = 0.4$. We let the true group-effect parameter α_g be $\alpha_g = 0$ when we evaluated the Type-I error rate. On the other hand, we let $\alpha_g = 0.5, 1$ when we evaluated performances under $\alpha_g > 0$.

The fitting model was $\log T = \beta_0 + \beta_g x_g + \varepsilon$ and we assumed that the distribution of $\exp(\varepsilon)$ was the Weibull distribution with shape parameter σ. We let the number of trials be 100,000 if $\alpha_g = 0$ or 10,000 if $\alpha_g > 0$. We used the different numbers of trials between null and alternative hypotheses because we have to evaluate the Type-I error rate with high precision. Under each condition, we calculated the maximum likelihood estimator $\hat{\beta}_g$ of β_g and conducted the robust Wald test, naive Wald test, chi-squared tests based on ordinary likelihood ratio statistic w, adjusted statistic w_K of w by Kent,[16] and adjusted statistic $w^* = w/\hat{E}_{g_0}[w]$ using the non-parametric bootstrap resampling, and a test based on the percentile bootstrap confidence interval for the null hypothesis $\beta_g = 0$ at a significance level of 0.05. Furthermore, we calculated the 95% confidence intervals corresponding to each test statistic as follows. The 95% confidence intervals based on the naive and robust Wald test statistics are defined by $[\hat{\beta}_g - z_{0.975}\widehat{\mathrm{SE}}_{\mathrm{Naive}}(\beta_g), \hat{\beta}_g + z_{0.975}\widehat{\mathrm{SE}}_{\mathrm{Naive}}(\beta_g)]$ and $[\hat{\beta}_g - z_{0.975}\widehat{\mathrm{SE}}_{\mathrm{Robust}}(\beta_g), \hat{\beta}_g + z_{0.975}\widehat{\mathrm{SE}}_{\mathrm{Robust}}(\beta_g)]$, respectively, where $z_{0.975}$ is the 97.5th percentile of the standard normal distribution. The three likelihood ratio test statistics are commonly denoted as $\xi^{-1}2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_0)\}$, where ξ is an adjustment factor (e.g., $\xi = 1$ for w). Let $\hat{\boldsymbol{\theta}}_d = (d, \hat{\boldsymbol{\lambda}}_d)$ be the maximum likelihood estimator under the constraint $\beta_g = d$. Then, the confidence interval based on the likelihood ratio statistic is defined as $\{d \mid \xi^{-1}2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_d)\} \leq \chi^2_{1,0.95}\}$,[43] where $\chi^2_{1,0.95}$ is the 95th percentile of χ^2_1. We let the number of resampling be 1,000 to calculate $\hat{E}_{g_0}[w]$ and bootstrap confidence intervals.

We calculated the bias of $\hat{\beta}_g$ (= mean of $\hat{\beta}_g - \alpha_g$) and coverage probability of true value α_g for each confidence interval. In addition, we calculated the Type-I error rate of the six tests when $\alpha_g = 0$, while we calculated the relative bias of $\hat{\beta}_g$ (= $100 \times$ (mean of $\hat{\beta}_g - \alpha_g)/\alpha_g$) when $\alpha_g > 0$.

5.1.2 Simulation result

Figures 5.1 – 5.6 show the results for Type-I error rate. Figures 5.1 – 5.2, 5.3 – 5.4, and 5.5 – 5.6 illustrate the results when the true distributions are the Weibull, log-logistic and log-normal, respectively.

In Figures 5.1 – 5.2, we correctly specified the true distribution. Type-I error rates for the five existing methods were inflated in the small sample even when the distribution is correctly specified. On the other hand, the Type-I error rates for our proposed method were close to the nominal level even in small samples.

In Figures 5.3 – 5.6, we misspecified the true distribution. Type-I error rates for the naive Wald test and ordinary likelihood ratio test were not controlled at the nominal level regardless of sample size. Type-I error rates for the robust Wald test, chi-squared test based on w_K, and test based on the percentile bootstrap confidence interval were close to the nominal level when n was large, but they were inflated when n was small. In terms of the Type-I error rate, the five existing tests did not have sufficient performance under model misspecification and small sample. In contrast, Type-I error rates of our proposed chi-squared test based on $w^* = w/\hat{E}_{g_0}[w]$ were close to the nominal level for all parameter settings.

Figures 5.7 – 5.15 show the results of coverage probability for each test statistic and bias of $\hat{\beta}_g$. Figures 5.7 – 5.9, 5.10 – 5.12, and 5.13 – 5.15 illustrate the results when the true distributions are the Weibull, log-logistic and log-

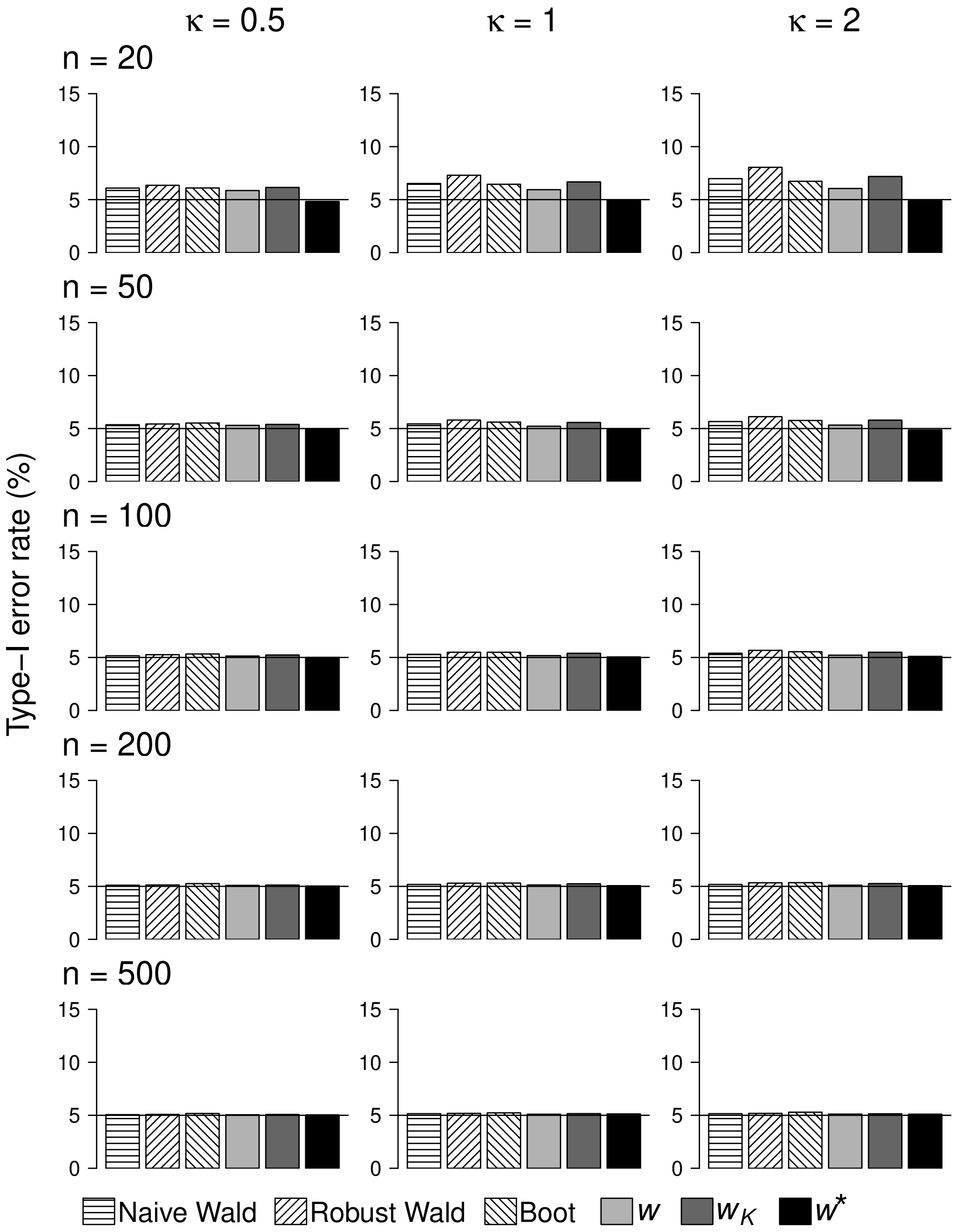

Figure 5.1: Type-I error rate (%) in simulation 1 when the true distribution is Weibull (i.e., correctly specified). The parameter q of the censoring mechanism is 0.2. κ is the shape parameter in the true model. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

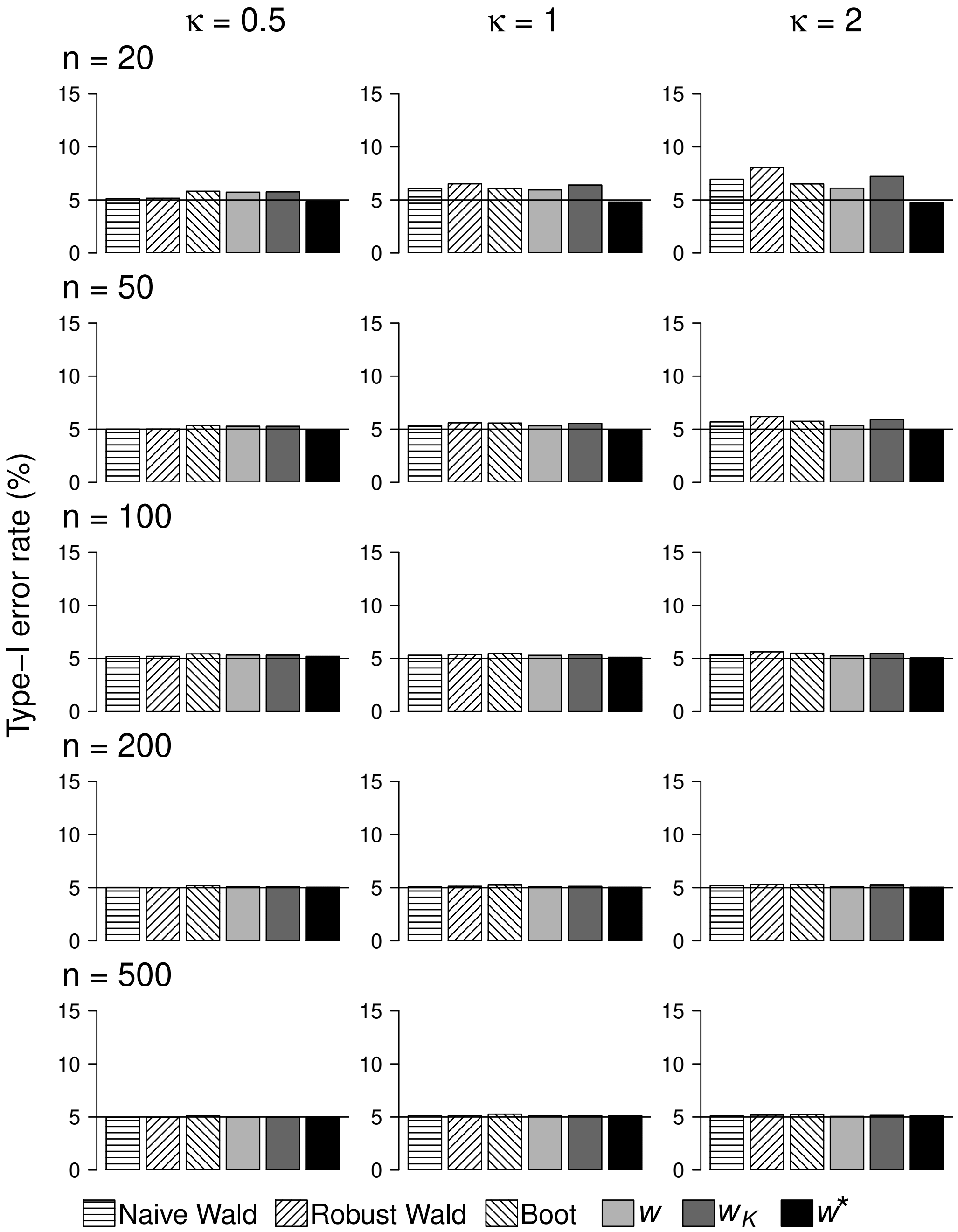

Figure 5.2: Type-I error rate (%) in simulation 1 when the true distribution is Weibull (i.e., correctly specified). The parameter q of the censoring mechanism is 0.4. κ is the shape parameter in the true model. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

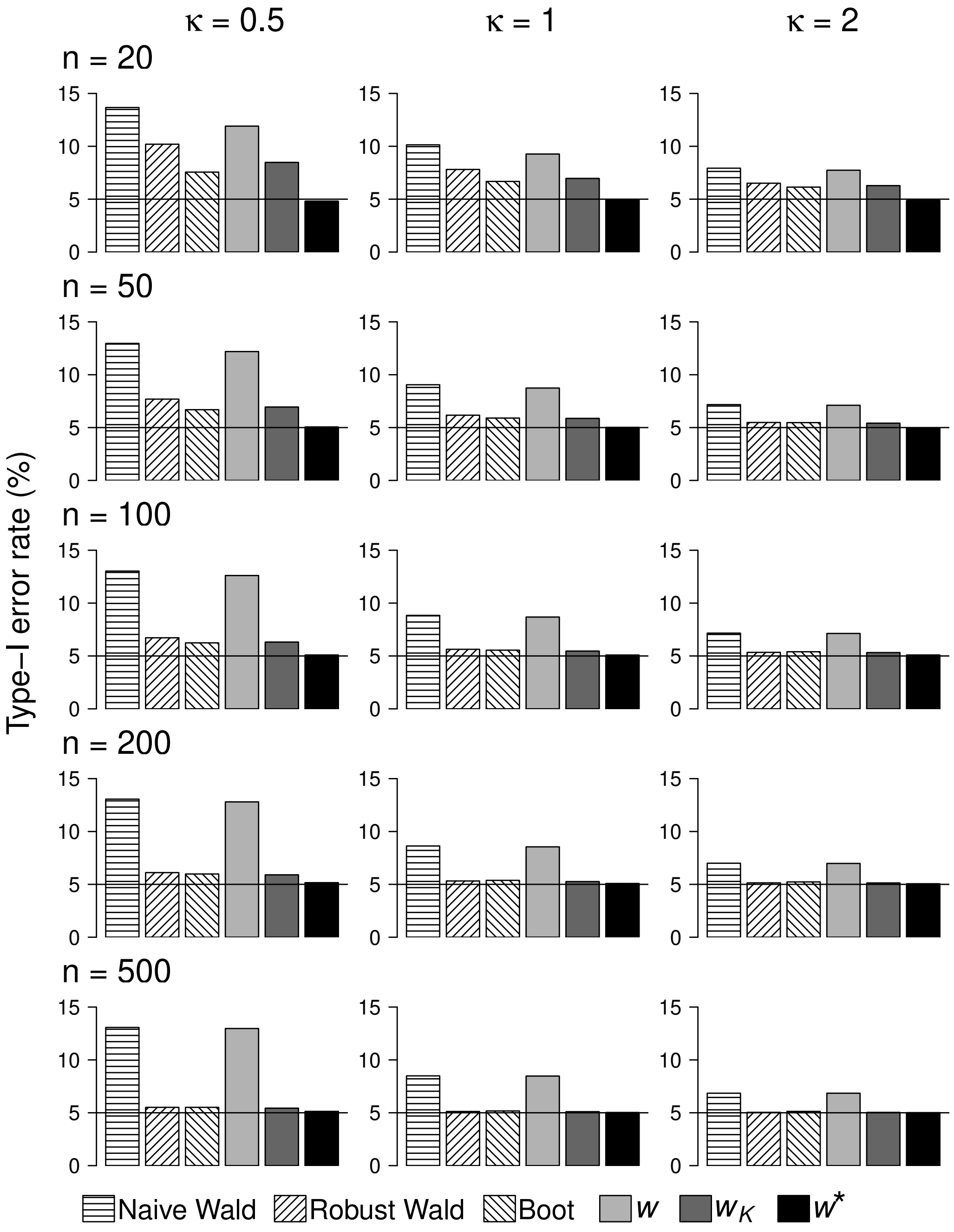

Figure 5.3: Type-I error rate (%) in simulation 1 when the true distribution is log-logistic (i.e., misspecified). The parameter q of the censoring mechanism is 0.2. κ is the shape parameter in the true model. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

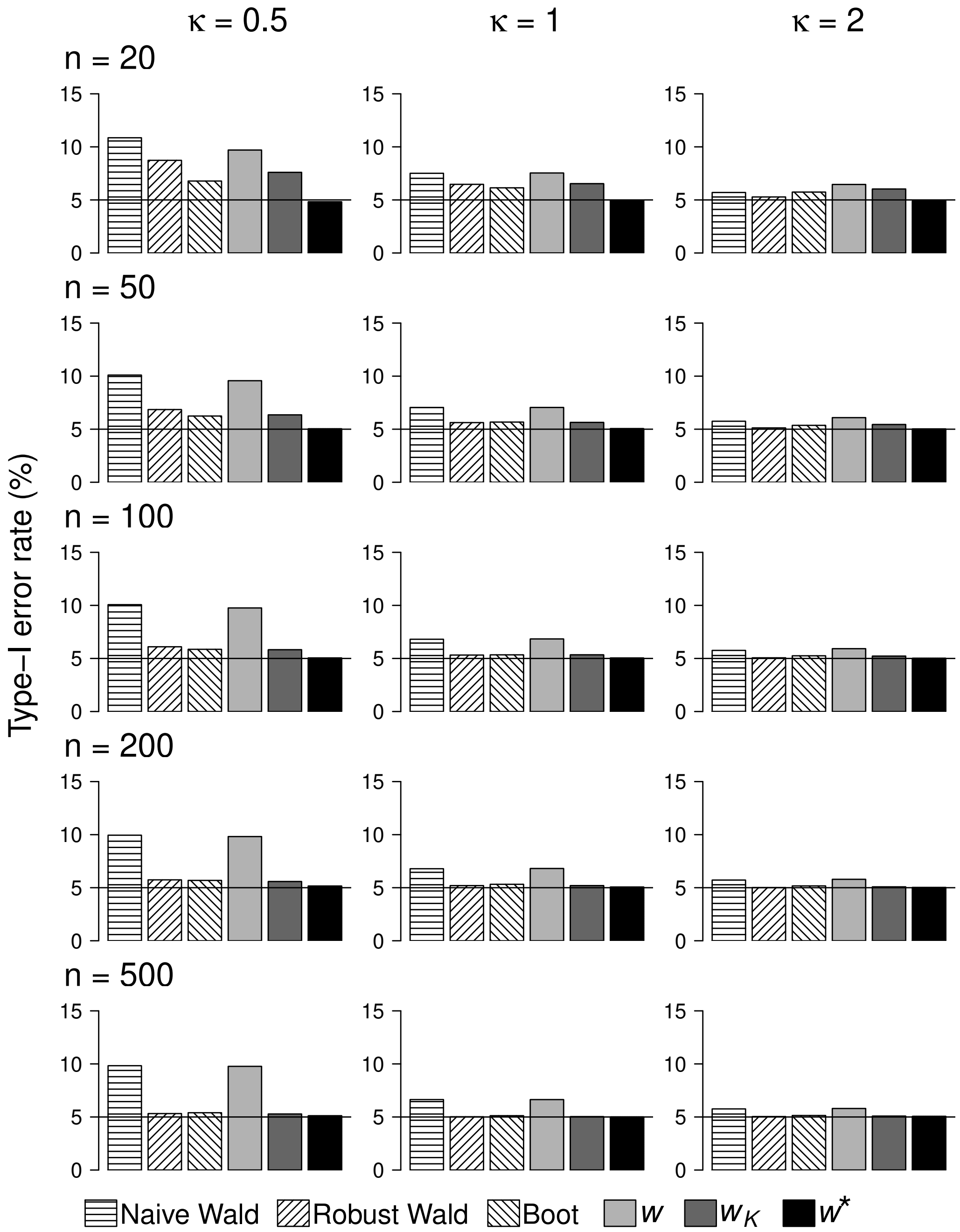

Figure 5.4: Type-I error rate (%) in simulation 1 when the true distribution is log-logistic (i.e., misspecified). The parameter q of the censoring mechanism is 0.4. κ is the shape parameter in the true model. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

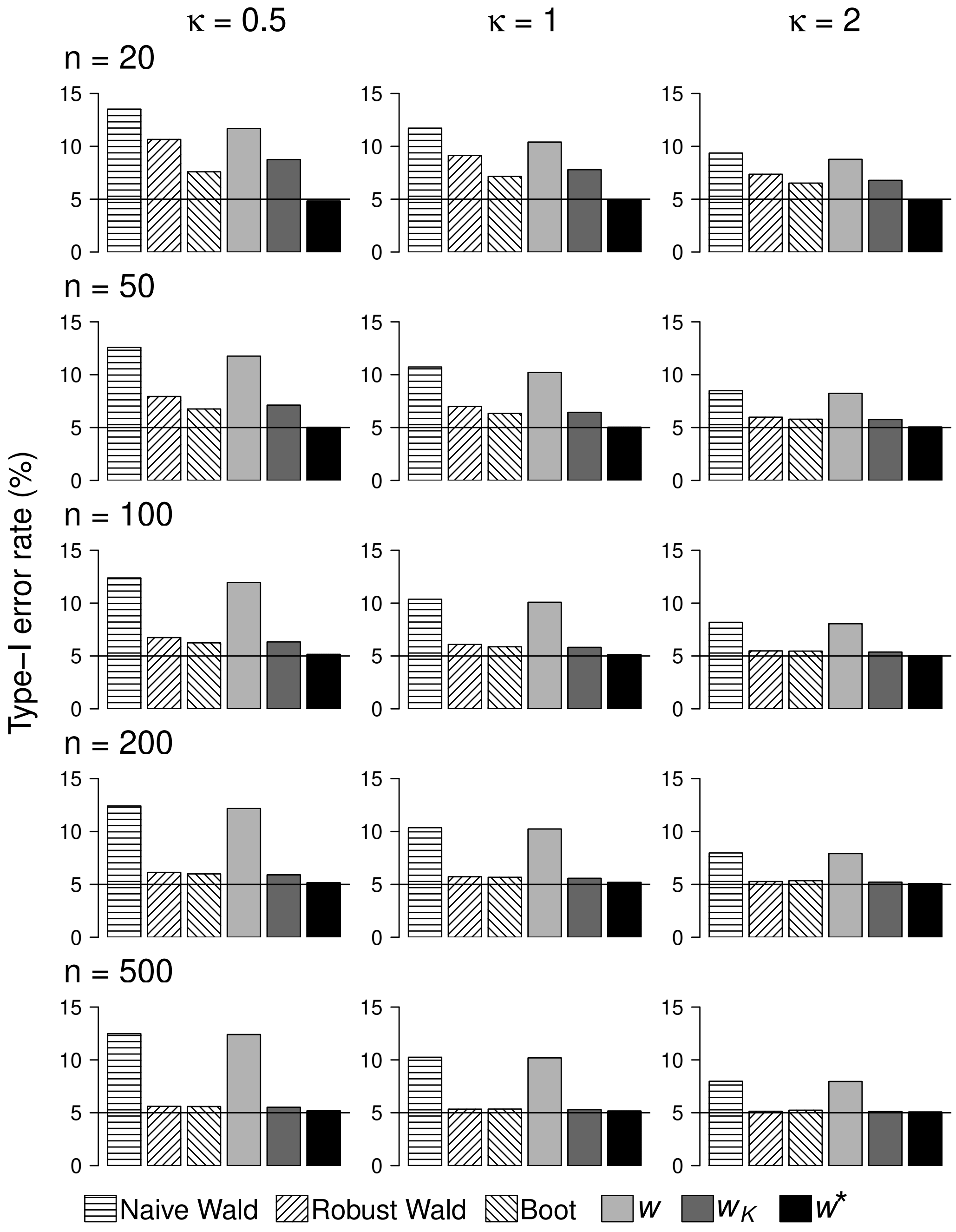

Figure 5.5: Type-I error rate (%) in simulation 1 when the true distribution is log-normal (i.e., misspecified). The parameter q of the censoring mechanism is 0.2. κ is the shape parameter in the true model. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

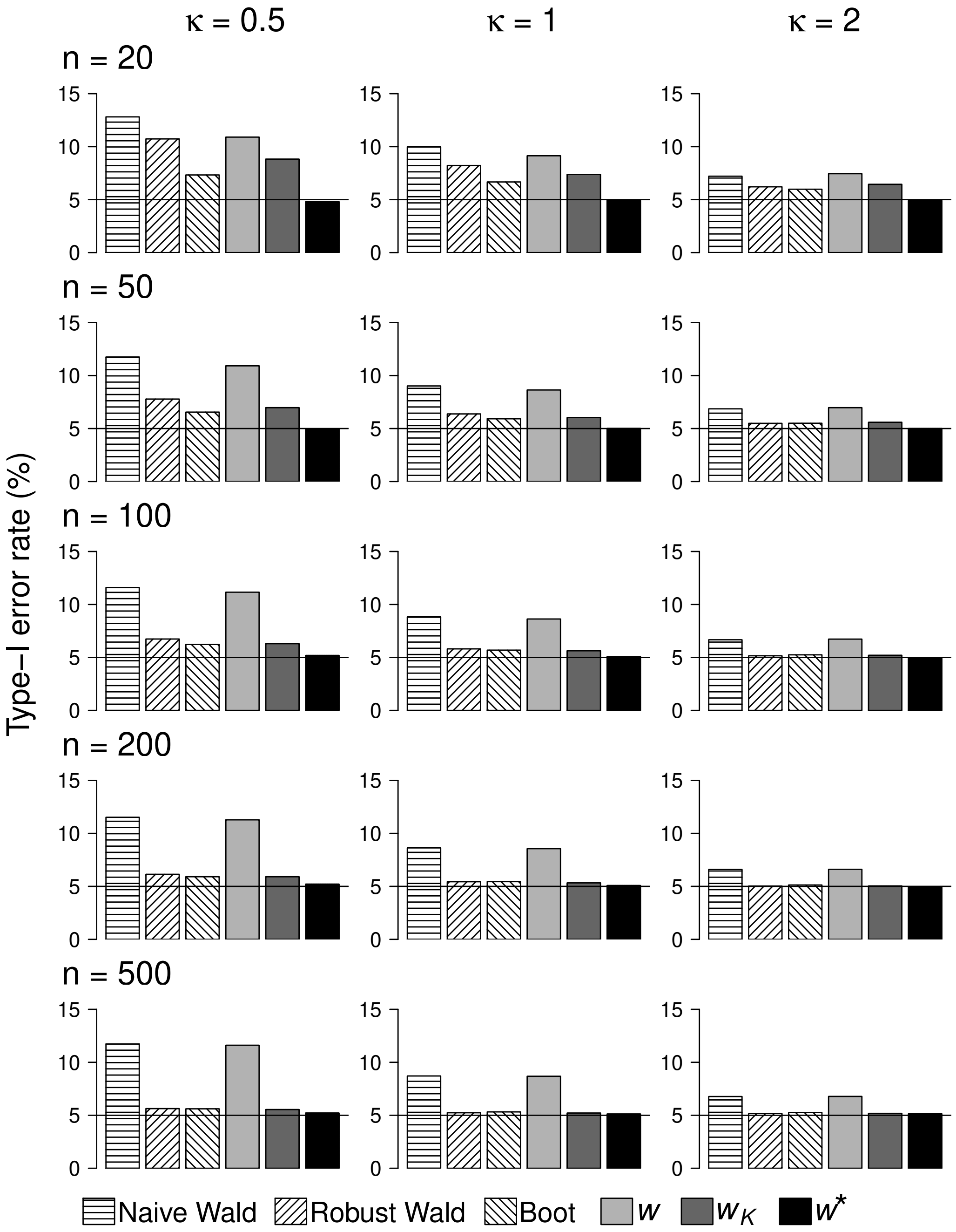

Figure 5.6: Type-I error rate (%) in simulation 1 when the true distribution is log-normal (i.e., misspecified). The parameter q of the censoring mechanism is 0.2. κ is the shape parameter in the true model. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

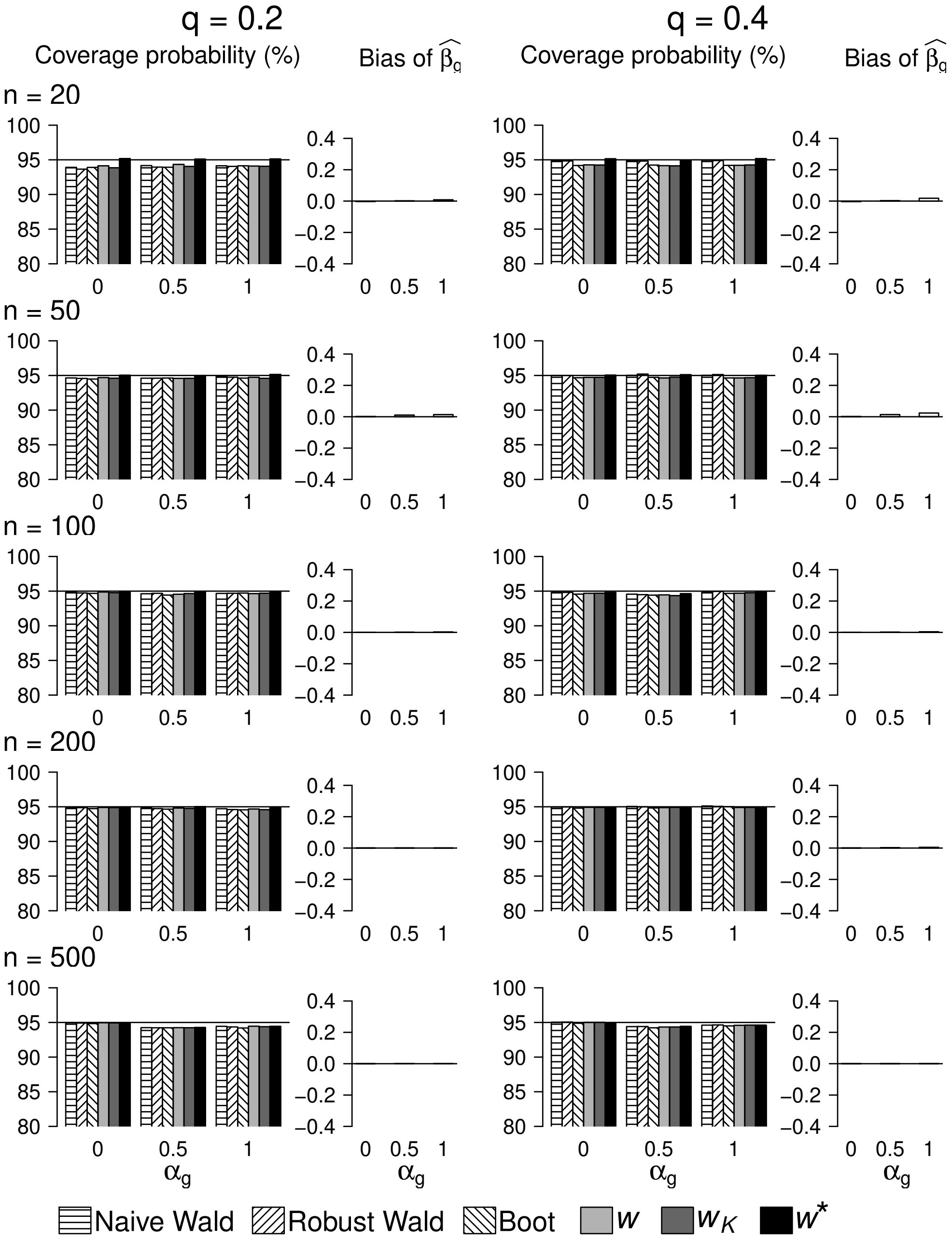

Figure 5.7: Coverage probability (%) and bias of $\hat{\beta}_g$ in simulation 1 when the true distribution is Weibull with shape parameter $\kappa = 0.5$ (i.e., correctly specified). q is the parameter of the censoring mechanism. α_g and β_g are group effects in the true model and fitting model, respectively. $\hat{\beta}_g$ is the maximum likelihood estimator of β_g. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

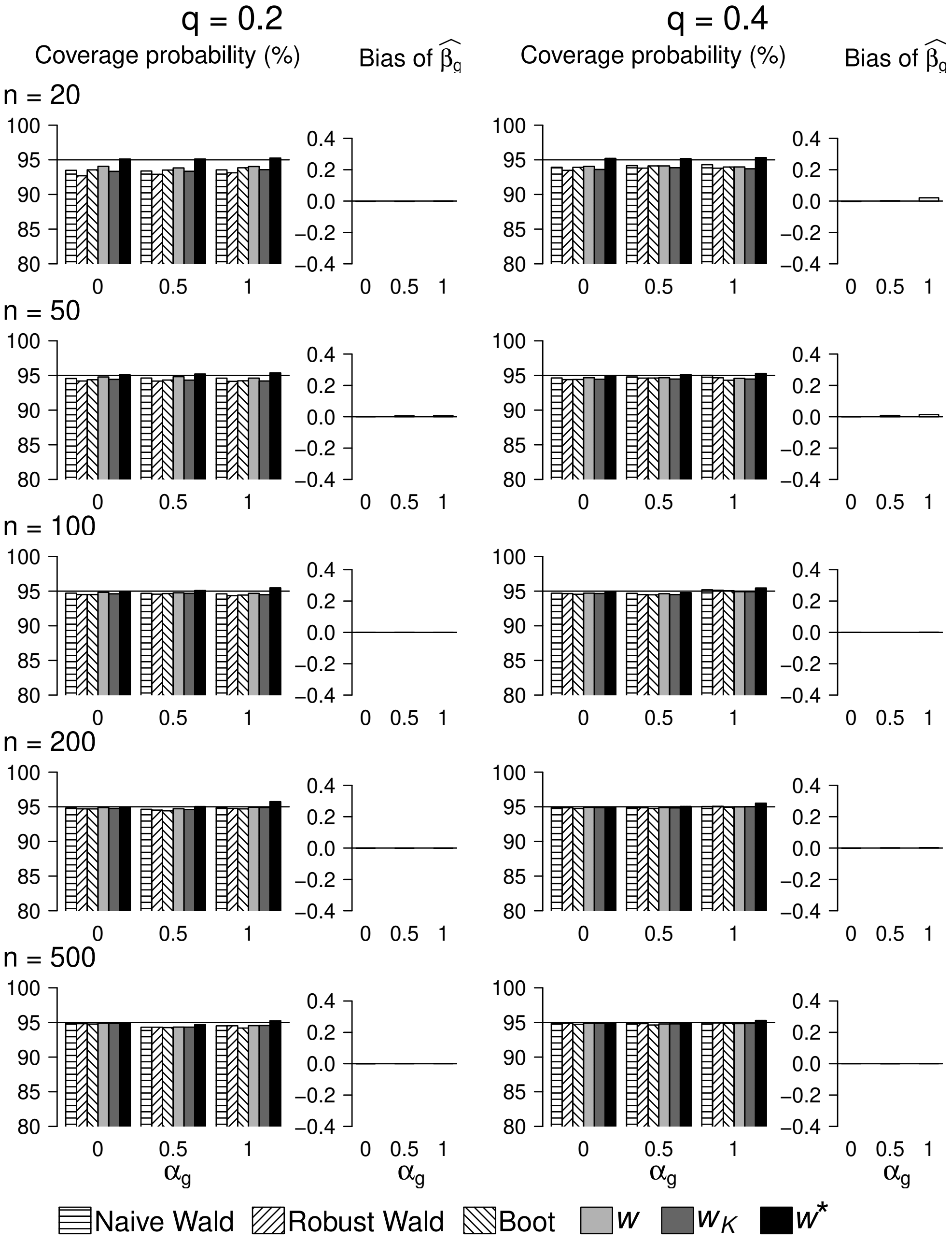

Figure 5.8: Coverage probability (%) and bias of $\hat{\beta}_g$ in simulation 1 when the true distribution is Weibull with shape parameter $\kappa = 1$ (i.e., correctly specified). q is the parameter of the censoring mechanism. α_g and β_g are group effects in the true model and fitting model, respectively. $\hat{\beta}_g$ is the maximum likelihood estimator of β_g. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

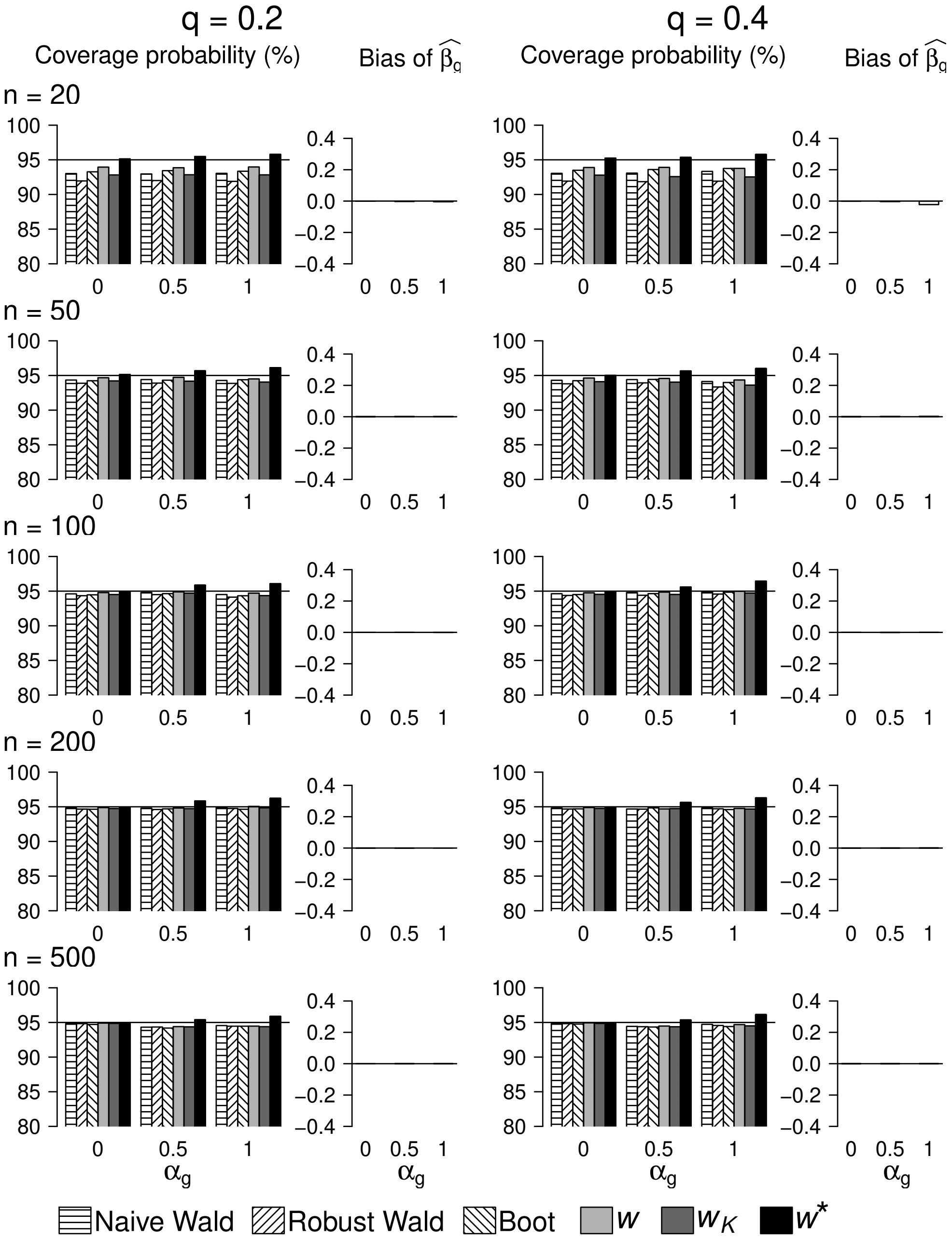

Figure 5.9: Coverage probability (%) and bias of $\hat{\beta}_g$ in simulation 1 when the true distribution is Weibull with shape parameter $\kappa = 2$ (i.e., correctly specified). q is the parameter of the censoring mechanism. α_g and β_g are group effects in the true model and fitting model, respectively. $\hat{\beta}_g$ is the maximum likelihood estimator of β_g. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

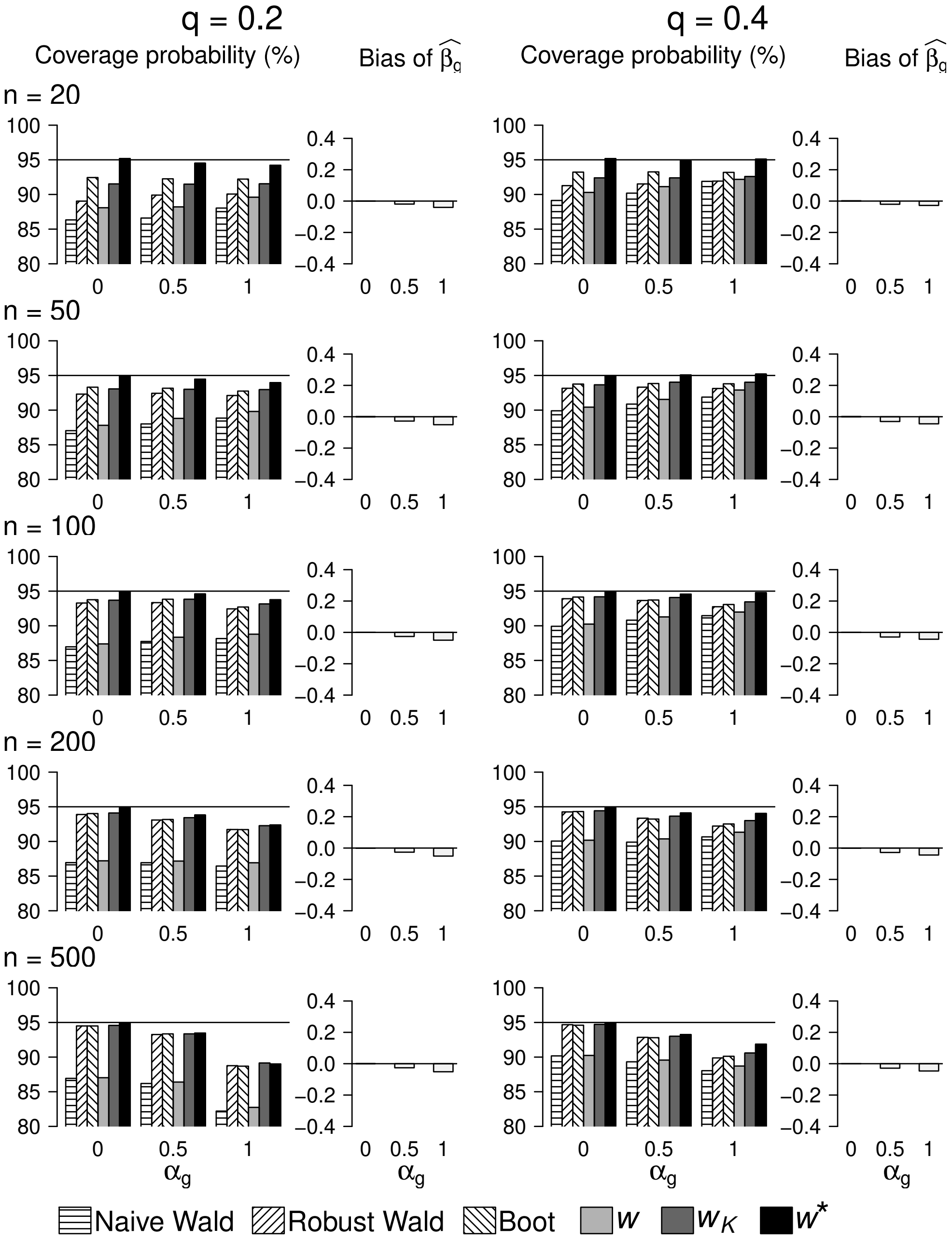

Figure 5.10: Coverage probability (%) and bias of $\hat{\beta}_g$ in simulation 1 when the true distribution is log-logistic with shape parameter $\kappa = 0.5$ (i.e., misspecified). q is the parameter of the censoring mechanism. α_g and β_g are group effects in the true model and fitting model, respectively. $\hat{\beta}_g$ is the maximum likelihood estimator of β_g. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

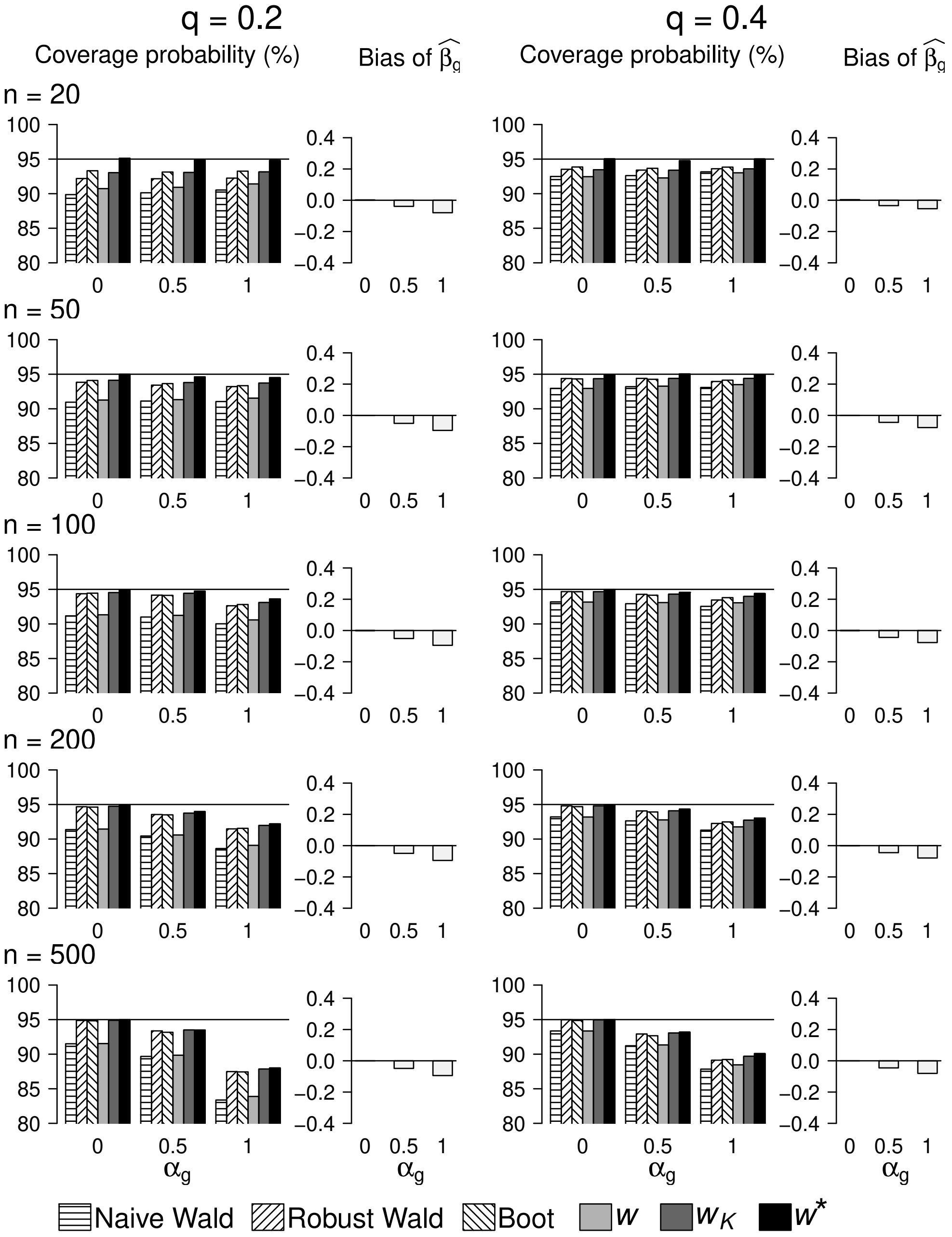

Figure 5.11: Coverage probability (%) and bias of $\hat{\beta}_g$ in simulation 1 when the true distribution is log-logistic with shape parameter $\kappa = 1$ (i.e., misspecified). q is the parameter of the censoring mechanism. α_g and β_g are group effects in the true model and fitting model, respectively. $\hat{\beta}_g$ is the maximum likelihood estimator of β_g. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

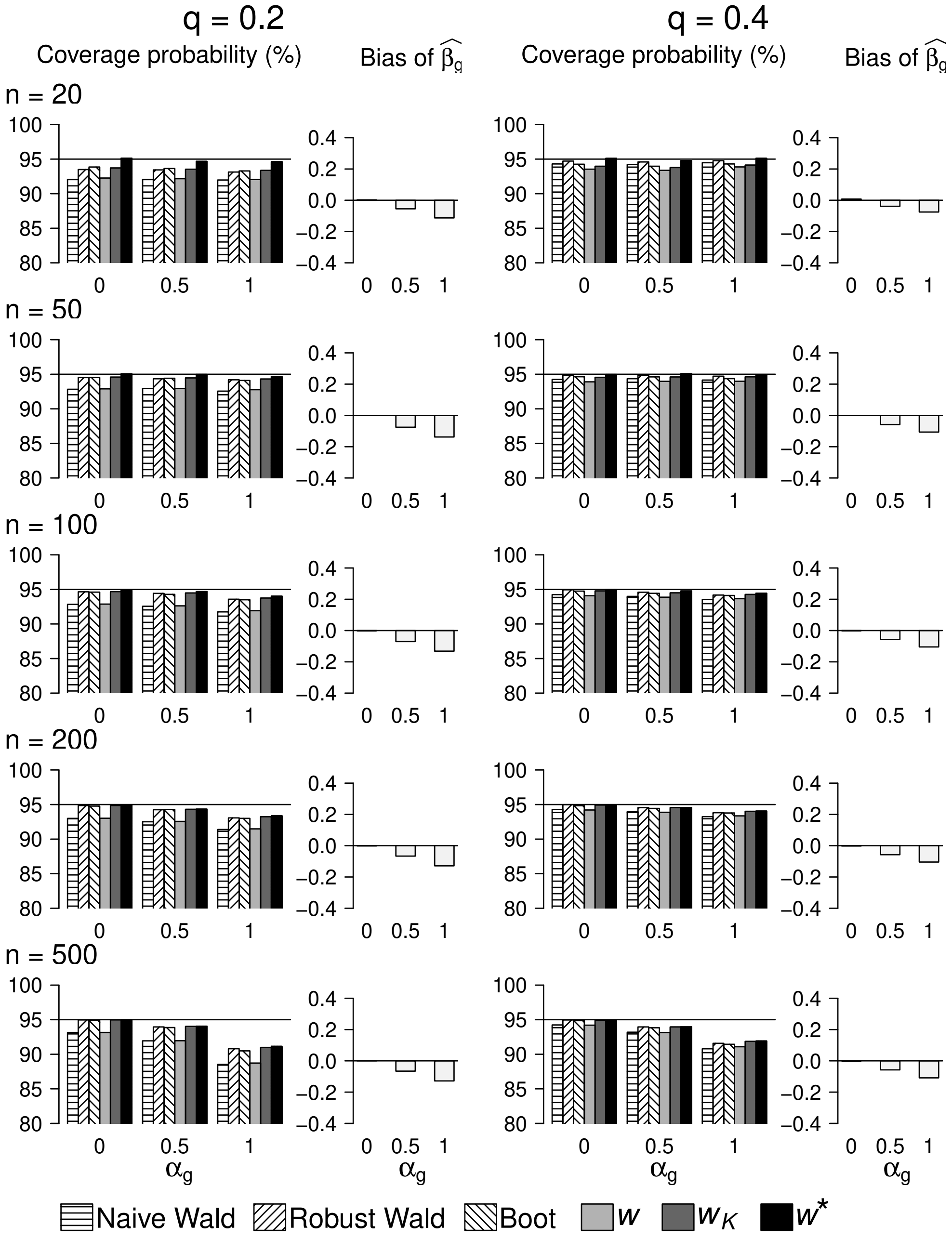

Figure 5.12: Coverage probability (%) and bias of $\hat{\beta}_g$ in simulation 1 when the true distribution is log-logistic with shape parameter $\kappa = 2$ (i.e., misspecified). q is the parameter of the censoring mechanism. α_g and β_g are group effects in the true model and fitting model, respectively. $\hat{\beta}_g$ is the maximum likelihood estimator of β_g. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

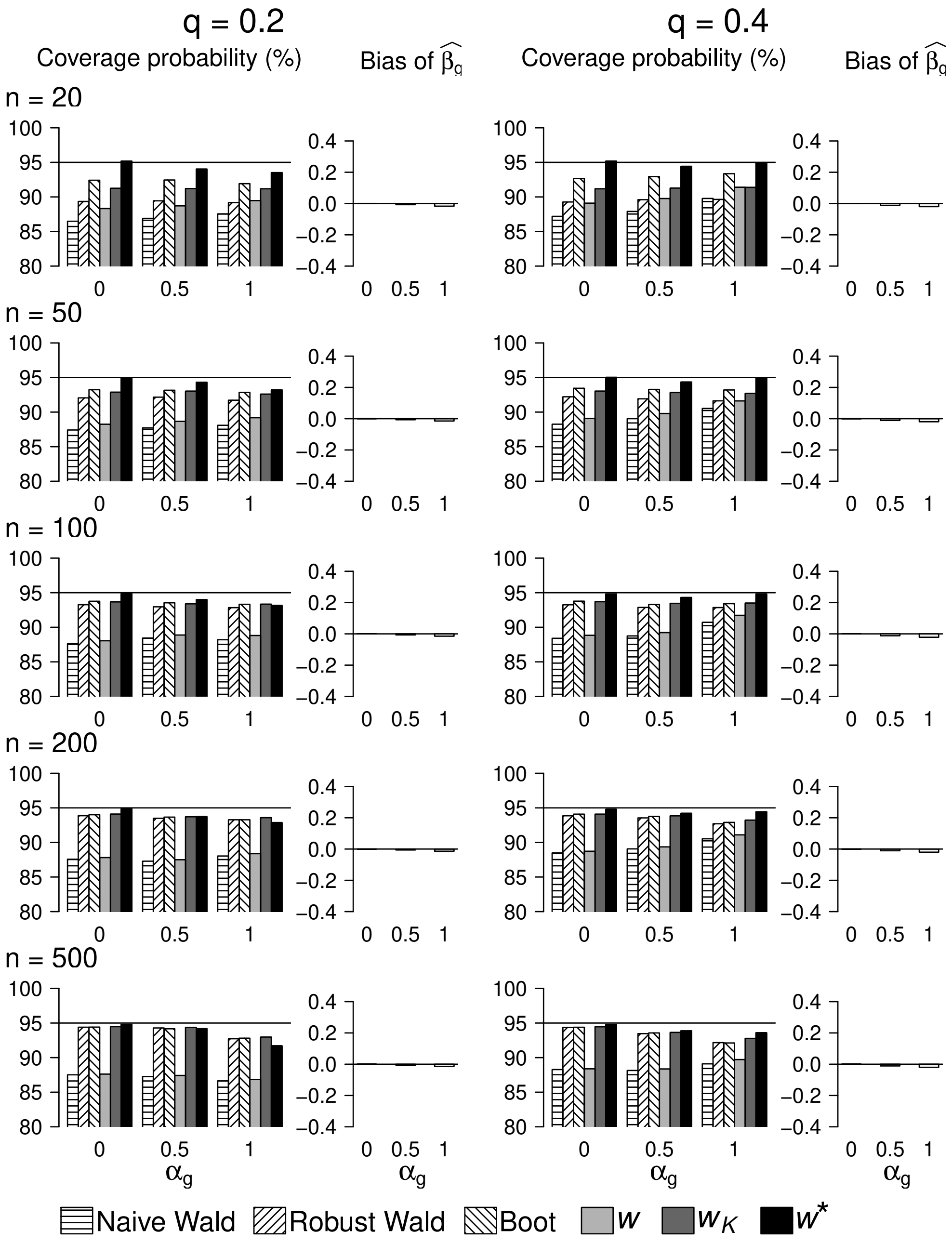

Figure 5.13: Coverage probability (%) and bias of $\hat{\beta}_g$ in simulation 1 when the true distribution is log-normal with shape parameter $\kappa = 0.5$ (i.e., misspecified). q is the parameter of the censoring mechanism. α_g and β_g are group effects in the true model and fitting model, respectively. $\hat{\beta}_g$ is the maximum likelihood estimator of β_g. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

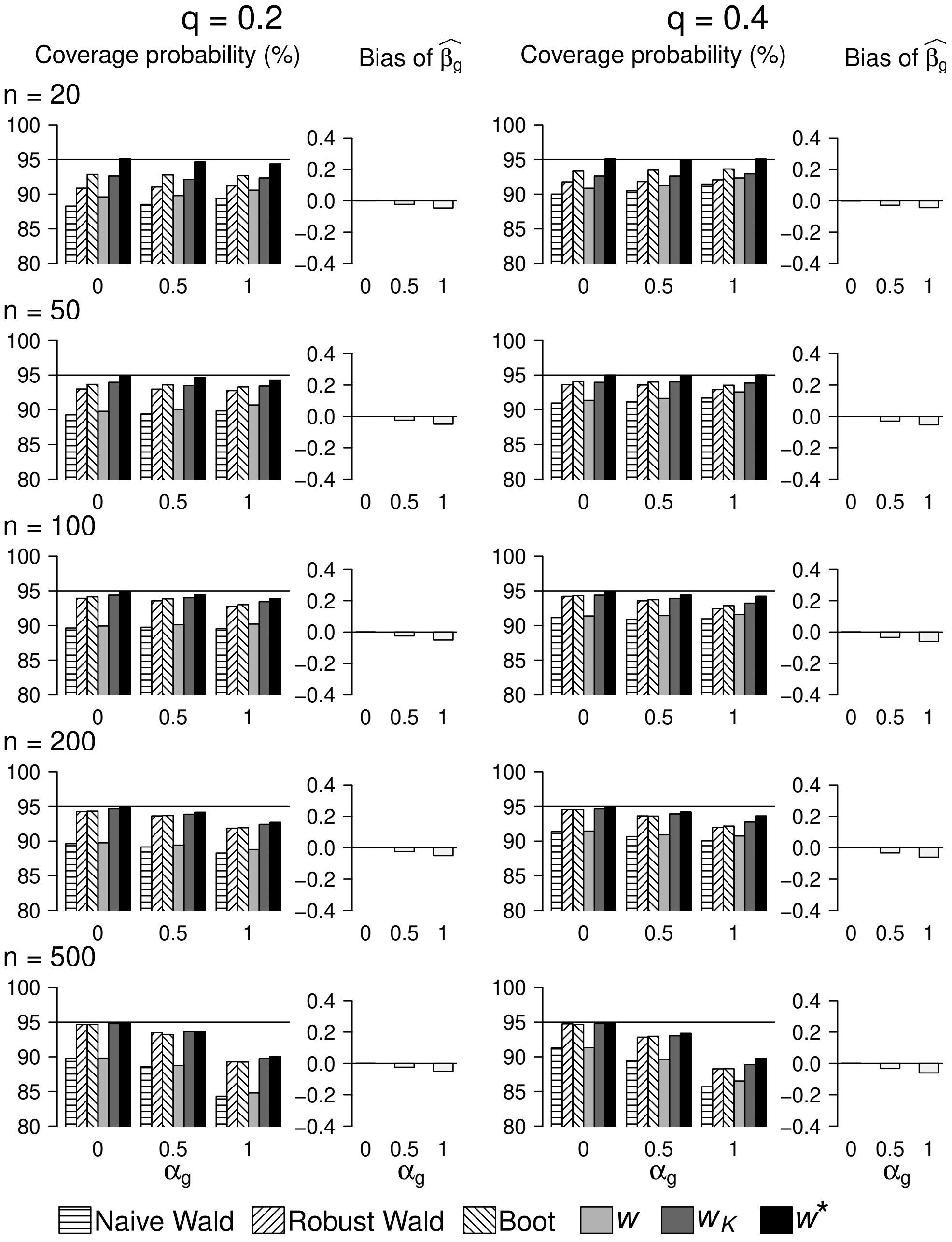

Figure 5.14: Coverage probability (%) and bias of $\hat{\beta}_g$ in simulation 1 when the true distribution is log-normal with shape parameter $\kappa = 1$ (i.e., misspecified). q is the parameter of the censoring mechanism. α_g and β_g are group effects in the true model and fitting model, respectively. $\hat{\beta}_g$ is the maximum likelihood estimator of β_g. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

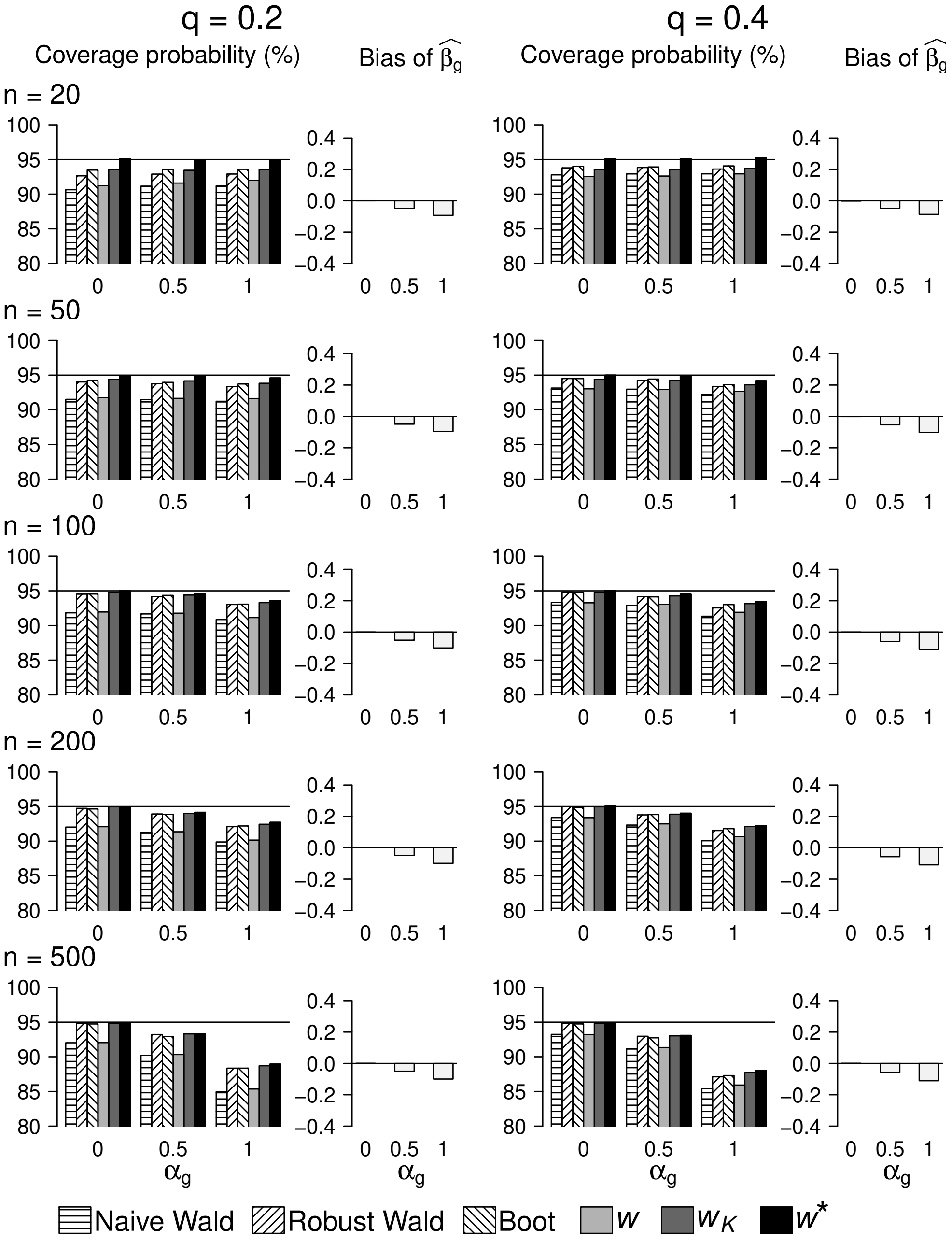

Figure 5.15: Coverage probability (%) and bias of $\hat{\beta}_g$ in simulation 1 when the true distribution is log-normal with shape parameter $\kappa = 2$ (i.e., misspecified). q is the parameter of the censoring mechanism. α_g and β_g are group effects in the true model and fitting model, respectively. $\hat{\beta}_g$ is the maximum likelihood estimator of β_g. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

Table 5.1: Relative bias (%) of $\hat{\beta}_g$ in simulation 1. The parameter of the censoring mechanism is $q = 0.2$

			True distribution		
κ	n	α_g	Weibull	Log-logistic	Log-normal
0.5	20	0.5	0.4	−3.8	−1.5
0.5	20	1.0	0.8	−4.0	−1.7
0.5	50	0.5	2.1	−5.5	−1.4
0.5	50	1.0	1.4	−5.1	−1.5
0.5	100	0.5	0.3	−5.2	−1.3
0.5	100	1.0	0.3	−5.0	−1.5
0.5	200	0.5	0.2	−5.3	−1.2
0.5	200	1.0	0.1	−5.1	−1.4
0.5	500	0.5	0.2	−5.2	−1.3
0.5	500	1.0	0.1	−5.1	−1.4
1.0	20	0.5	−0.3	−7.8	−4.5
1.0	20	1.0	0.1	−8.0	−4.6
1.0	50	0.5	1.0	−10.2	−4.9
1.0	50	1.0	0.7	−9.6	−4.9
1.0	100	0.5	0.2	−10.2	−4.9
1.0	100	1.0	0.1	−9.5	−5.1
1.0	200	0.5	0.0	−9.9	−4.8
1.0	200	1.0	0.0	−9.4	−5.0
1.0	500	0.5	0.1	−9.7	−4.9
1.0	500	1.0	0.1	−9.4	−5.1
2.0	20	0.5	−0.6	−10.9	−9.8
2.0	20	1.0	−0.4	−11.3	−9.3
2.0	50	0.5	0.4	−15.1	−9.8
2.0	50	1.0	0.2	−13.8	−9.6
2.0	100	0.5	0.0	−14.1	−10.4
2.0	100	1.0	−0.1	−13.2	−10.1
2.0	200	0.5	0.0	−13.3	−9.9
2.0	200	1.0	0.0	−12.9	−10.0
2.0	500	0.5	0.0	−13.2	−9.9
2.0	500	1.0	0.0	−12.8	−10.0

Note: α_g and β_g are group effects in the true model and fitting model, respectively. κ is the shape parameter in the true model. Relative bias (%) is $100 \times (\text{mean of } \hat{\beta}_g - \alpha_g)/\alpha_g$, where $\hat{\beta}_g$ is the maximum likelihood estimator of β_g.

Table 5.2: Relative bias (%) of $\hat{\beta}_g$ in simulation 1. The parameter of the censoring mechanism is $q = 0.4$

			True distribution		
κ	n	α_g	Weibull	Log-logistic	Log-normal
0.5	20	0.5	0.6	−4.1	−2.4
0.5	20	1.0	1.8	−2.8	−2.0
0.5	50	0.5	2.8	−6.1	−2.3
0.5	50	1.0	2.4	−4.5	−2.0
0.5	100	0.5	0.4	−5.9	−2.4
0.5	100	1.0	0.4	−4.4	−2.2
0.5	200	0.5	0.5	−5.6	−2.1
0.5	200	1.0	0.5	−4.4	−2.0
0.5	500	0.5	0.0	−5.7	−2.2
0.5	500	1.0	0.0	−4.6	−2.0
1.0	20	0.5	0.5	−6.9	−5.6
1.0	20	1.0	2.1	−5.4	−4.3
1.0	50	0.5	1.7	−9.0	−6.0
1.0	50	1.0	1.4	−7.8	−5.4
1.0	100	0.5	0.0	−8.8	−6.8
1.0	100	1.0	0.1	−7.7	−6.0
1.0	200	0.5	0.3	−8.9	−6.5
1.0	200	1.0	0.2	−7.9	−6.0
1.0	500	0.5	0.0	−9.1	−6.5
1.0	500	1.0	0.1	−8.1	−6.1
2.0	20	0.5	−0.7	−7.7	−9.6
2.0	20	1.0	−2.3	−7.5	−8.7
2.0	50	0.5	0.5	−11.6	−10.7
2.0	50	1.0	0.3	−10.7	−10.2
2.0	100	0.5	−0.2	−11.4	−12.0
2.0	100	1.0	0.0	−10.5	−11.0
2.0	200	0.5	0.1	−11.5	−11.4
2.0	200	1.0	0.1	−10.4	−10.9
2.0	500	0.5	0.0	−11.5	−11.5
2.0	500	1.0	0.0	−10.9	−11.1

Note: α_g and β_g are group effects in the true model and fitting model, respectively. κ is the shape parameter in the true model. Relative bias (%) is $100 \times (\text{mean of } \hat{\beta}_g - \alpha_g)/\alpha_g$, where $\hat{\beta}_g$ is the maximum likelihood estimator of β_g.

normal, respectively. Tables 5.1 – 5.2 show the results of the relative bias.

There was no bias for parameter estimates $\hat{\beta}_g$ for all the settings under the null hypothesis since both distributions of survival time and censoring time are common in groups under the null hypothesis. When $\alpha_g > 0$, the parameter estimates $\hat{\beta}_g$ had a downward bias under model misspecification, while $\hat{\beta}_g$ had little bias under the correct model specification. When the true distribution was Weibull (i.e., correctly specified), coverage probabilities of the five existing methods were close to the nominal level in large samples, but they were lower than the nominal level in small samples. In contrast, the coverage probabilities of our proposed method were close to the nominal level in many cases even if n was small; however, in other cases, the coverage probabilities were somewhat larger than the nominal level. When the true distribution were log-logistic and log-normal (i.e., misspecified), coverage probabilities of the five existing methods were lower than the nominal level for all parameter settings. On the other hands, coverage probabilities of our proposed method were larger than those of the existing methods and were close to the nominal level in many cases.

5.2 Simulation 2

5.2.1 Simulation design

This simulation aims to compare the performance of the proposed methods with existing methods under misspecification of both the error distribution and mean structure. The survival time assumed a model of the form $\log T = \alpha_0 + \alpha_g x_g + \alpha_1 x_1 + \eta$, where $x_1 \sim Bernoulli(0.3)$ is a binary covariate. We let the distribution of $\exp(\eta)$ be the Weibull or log-normal distribution with shape parameter κ. We let the true parameters be $\alpha_0 = 1, \alpha_g = 0, 0.5, 1$, and

$\alpha_1 = \pm 1, \pm 2$, and let $\kappa = 2$ and 0.5 for Weibull distribution and log-normal distribution, respectively. The sample sizes were the same in each group and we let sample size n in each group be $n = 20, 200$. The fitting model was $\log T = \beta_0 + \beta_g x_g + \varepsilon$ and we assumed that the distribution of $\exp(\varepsilon)$ was the Weibull distribution with shape parameter σ. Hence, we misspecified the error distribution and/or omitted the important covariate. The other settings were the same as simulation 1.

5.2.2 Simulation result

Figure 5.16 shows the Type-I error rate in simulation 2. In the upper half of Figure 5.16, the true distribution is correctly specified, but the important covariate x_1 is omitted from the mean structure. Type-I error rates for the naive Wald test and the ordinary likelihood ratio test were not controlled at the nominal level regardless of the sample size, owing to the misspecification of the mean structure. Type-I error rates for the robust Wald test, chi-squared test based on w_K, and test based on the percentile bootstrap confidence interval were inflated in the small sample as in simulation 1. In contrast, Type-I error rates for the proposed method were controlled at the nominal level for all simulation settings.

In the lower half of Figure 5.16, the true distribution was misspecified and the important covariate x_1 is omitted from the mean structure. The result was similar to those of the upper half, but the degree of the inflation of the existing methods were larger, owing to the misspecification of both the error distribution and the mean structure. On the other hand, Type-I error ratess for the proposed method were controlled at the nominal level.

Figures 5.17 – 5.20 show the results of coverage probability for each test statistic and bias of $\hat{\beta}_g$. Figures 5.17 – 5.18 and 5.19 – 5.20 illustrate the results

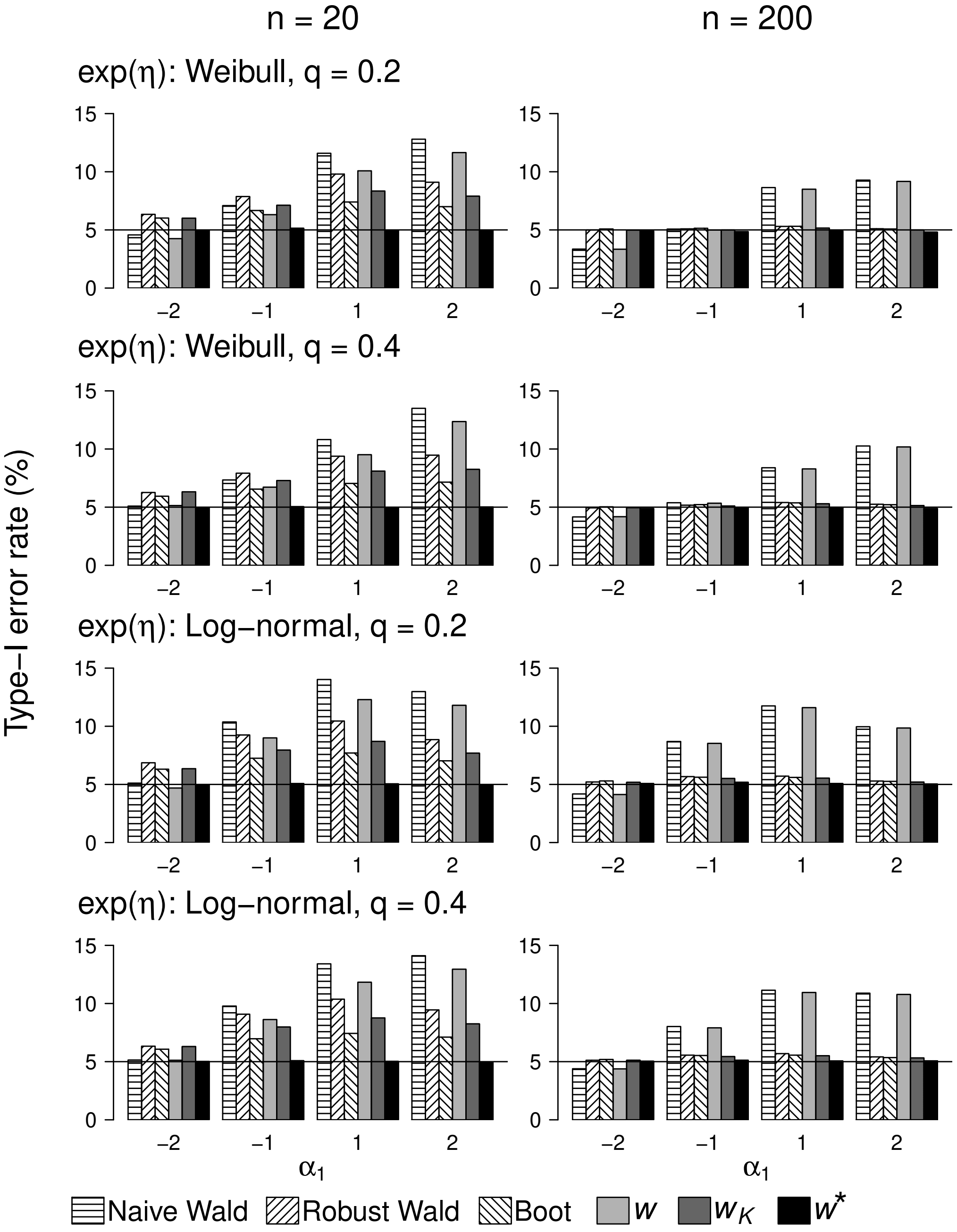

Figure 5.16: Type-I error rate (%) in simulation 2. The parameter q of the censoring mechanism is 0.2. α_1 is the coefficient of the omitted covariate x_1. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

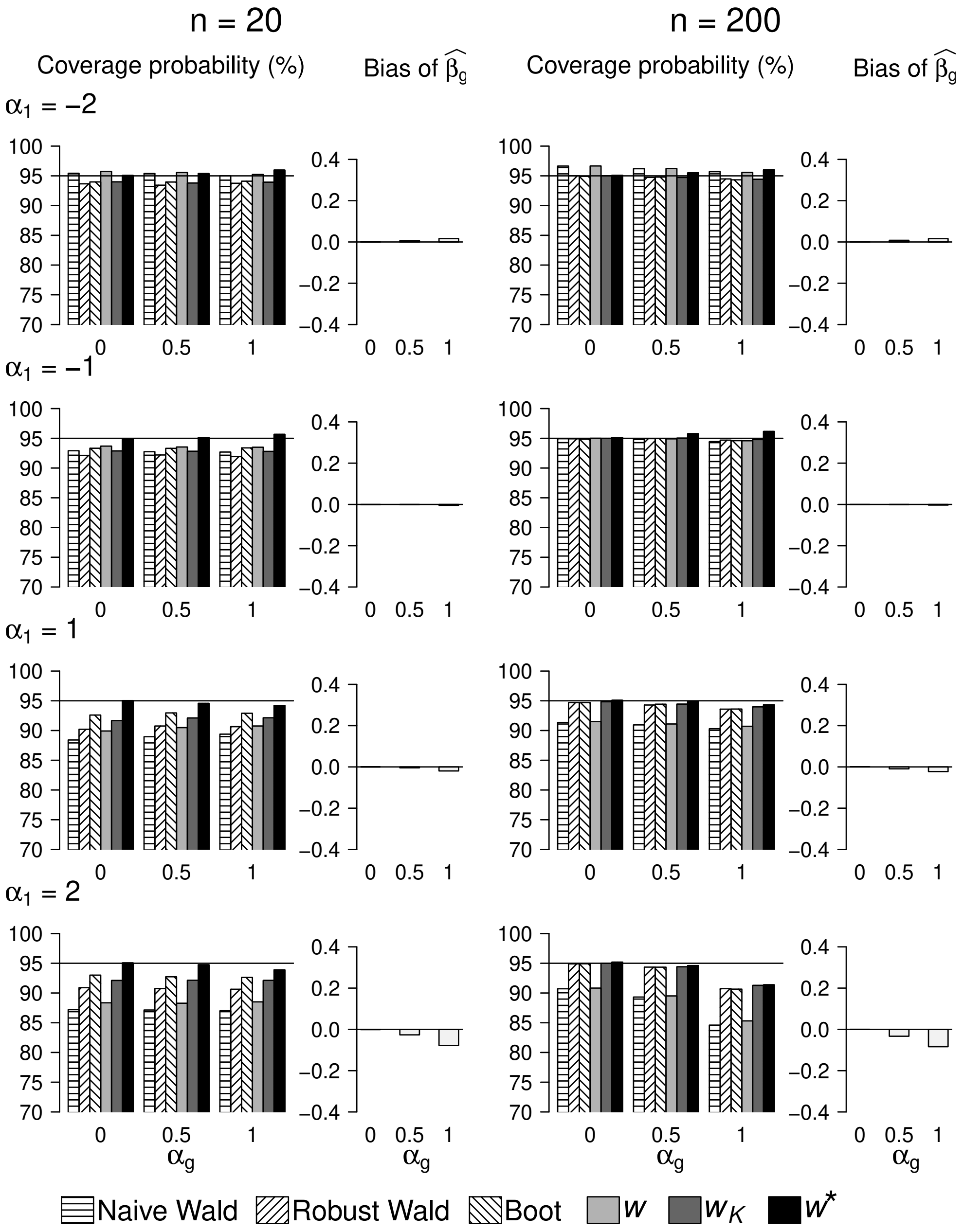

Figure 5.17: Coverage probability (%) and bias of $\hat{\beta}_g$ in simulation 2 when the true distribution is Weibull. The parameter q of the censoring mechanism is 0.2. α_g and β_g are group effects in the true model and fitting model, respectively. κ is the shape parameter in the true model. $\hat{\beta}_g$ is the maximum likelihood estimator of β_g. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

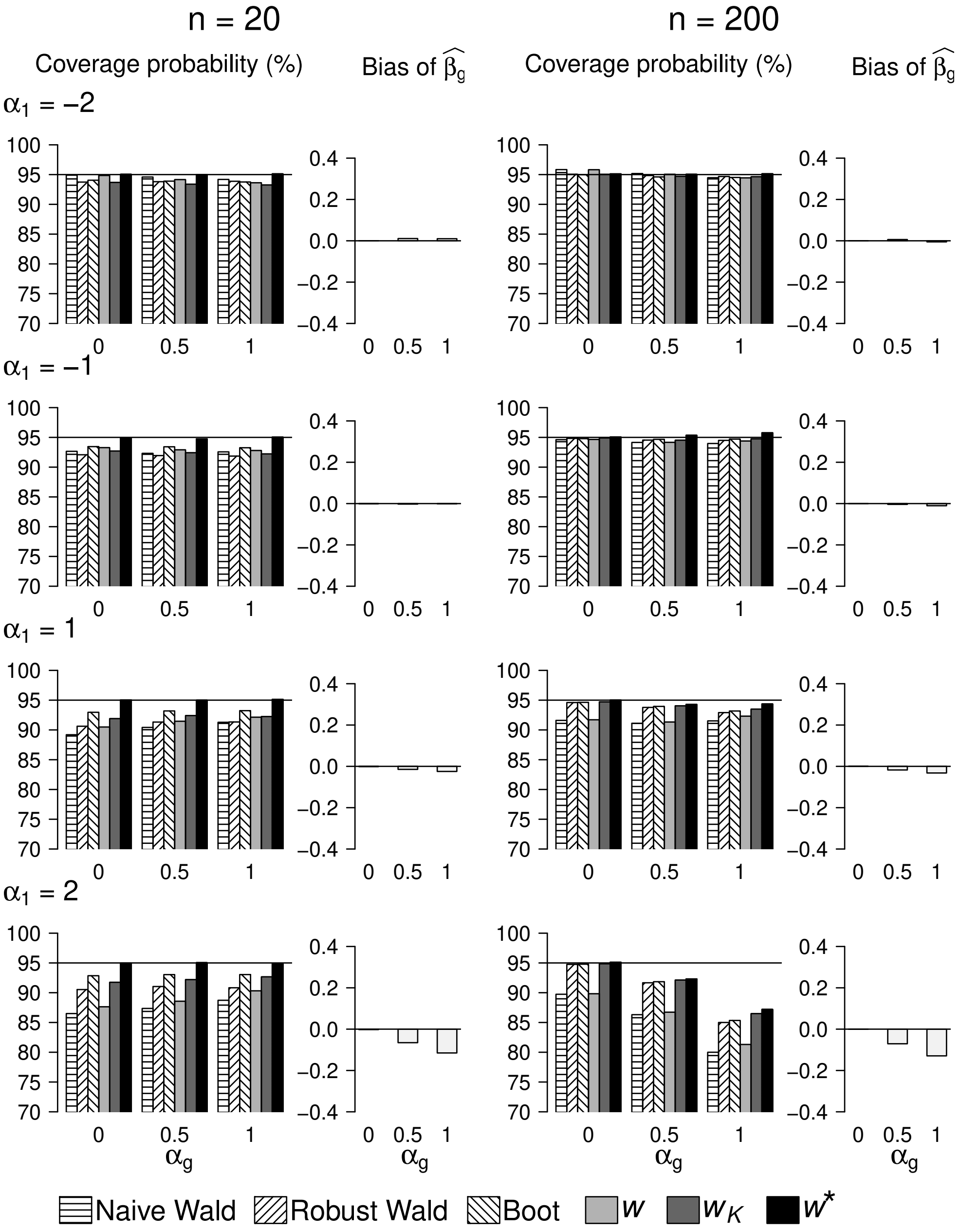

Figure 5.18: Coverage probability (%) and bias of $\hat{\beta}_g$ in simulation 2 when the true distribution is Weibull. The parameter q of the censoring mechanism is 0.4. α_g and β_g are group effects in the true model and fitting model, respectively. κ is the shape parameter in the true model. $\hat{\beta}_g$ is the maximum likelihood estimator of β_g. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

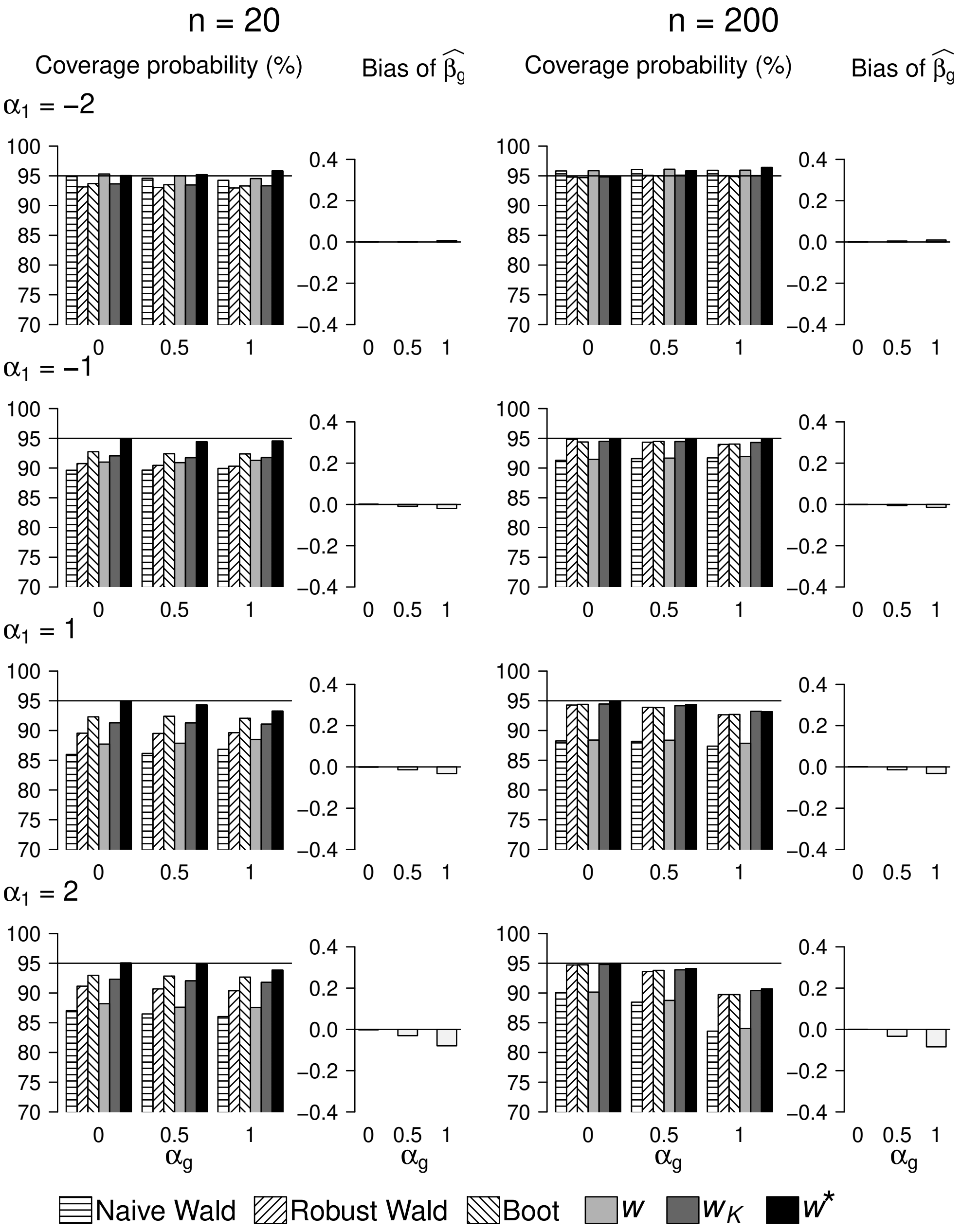

Figure 5.19: Coverage probability (%) and bias of $\hat{\beta}_g$ in simulation 2 when the true distribution is log-normal. The parameter q of the censoring mechanism is 0.2. α_g and β_g are group effects in the true model and fitting model, respectively. κ is the shape parameter in the true model. $\hat{\beta}_g$ is the maximum likelihood estimator of β_g. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

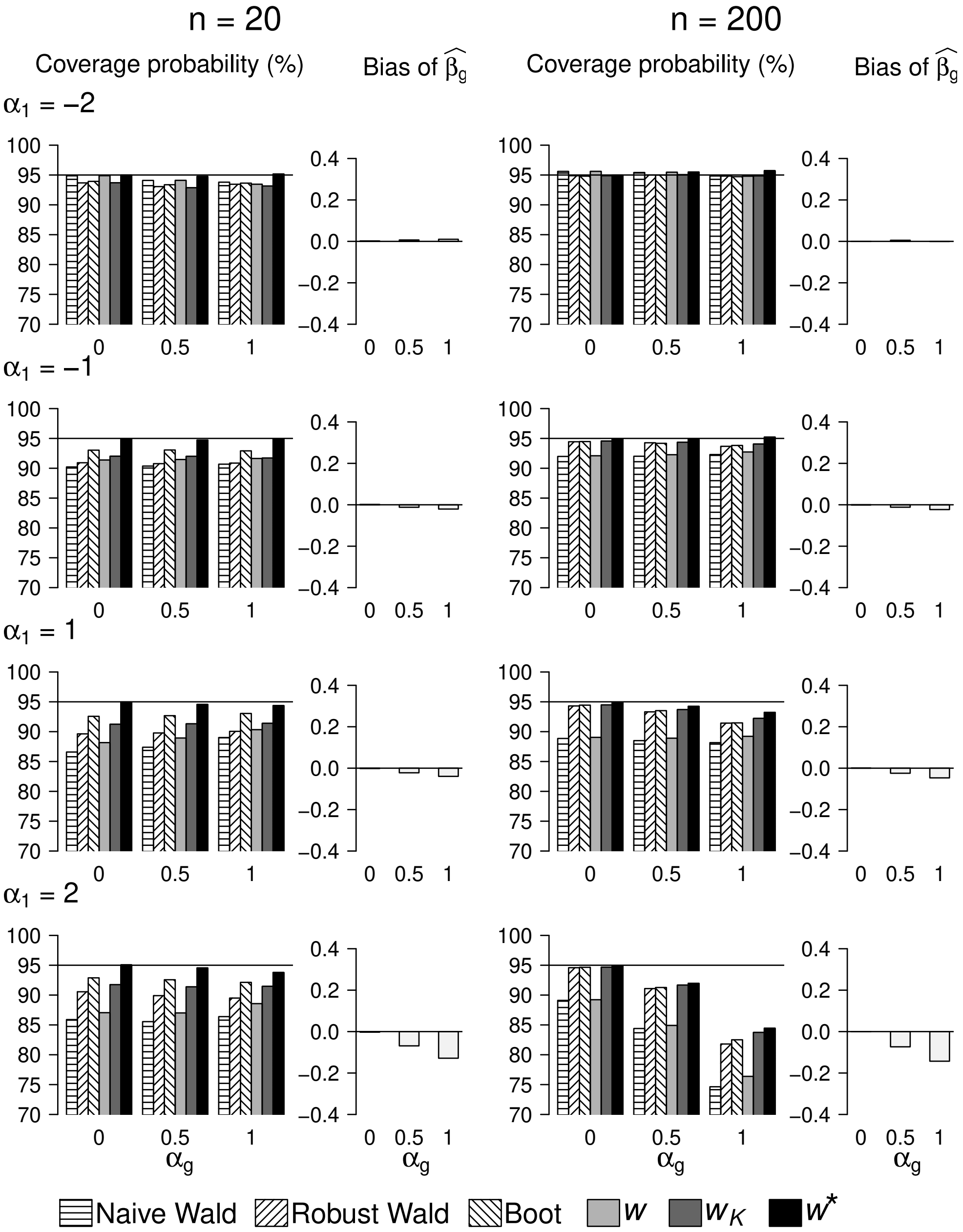

Figure 5.20: Coverage probability (%) and bias of $\hat{\beta}_g$ in simulation 2 when the true distribution is log-normal. The parameter q of the censoring mechanism is 0.4. α_g and β_g are group effects in the true model and fitting model, respectively. κ is the shape parameter in the true model. $\hat{\beta}_g$ is the maximum likelihood estimator of β_g. w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and Boot is the test based on the percentile bootstrap confidence interval.

Table 5.3: Relative bias (%) of $\hat{\beta}_g$ in simulation 2.

	n	α_1	α_g	$\exp(\eta)$: Weibull	$\exp(\eta)$: Log-normal
$q = 0.2$	20	−2	0.5	1.3	0.1
	20	−2	1.0	1.6	0.6
	20	−1	0.5	−0.1	−1.8
	20	−1	1.0	−0.4	−1.9
	20	1	0.5	−0.9	−2.7
	20	1	1.0	−2.0	−3.2
	20	2	0.5	−5.3	−6.1
	20	2	1.0	−7.8	−8.0
	200	−2	0.5	1.6	0.9
	200	−2	1.0	1.6	0.9
	200	−1	0.5	−0.1	−1.1
	200	−1	1.0	−0.3	−1.4
	200	1	0.5	−1.8	−2.7
	200	1	1.0	−2.3	−3.2
	200	2	0.5	−6.7	−6.7
	200	2	1.0	−8.4	−8.5
$q = 0.4$	20	−2	0.5	2.2	1.3
	20	−2	1.0	1.0	1.0
	20	−1	0.5	−0.5	−2.4
	20	−1	1.0	−0.1	−2.1
	20	1	0.5	−2.9	−4.4
	20	1	1.0	−2.5	−4.0
	20	2	0.5	−13.1	−13.8
	20	2	1.0	−11.5	−12.9
	200	−2	0.5	1.2	1.1
	200	−2	1.0	−0.5	−0.1
	200	−1	0.5	−0.8	−2.3
	200	−1	1.0	−1.0	−2.3
	200	1	0.5	−3.5	−4.9
	200	1	1.0	−3.2	−4.7
	200	2	0.5	−14.1	−14.7
	200	2	1.0	−12.9	−14.3

Note: α_g and β_g are group effects in the true model and fitting model, respectively. q is the parameter of the censoring mechanism. α_1 is the coefficient of the omitted covariate x_1. Relative bias (%) is $100 \times (\text{mean of } \hat{\beta}_g - \alpha_g)/\alpha_g$, where $\hat{\beta}_g$ is the maximum likelihood estimator of β_g.

when the true distributions are the Weibull and log-normal, respectively. Table 5.3 shows the results of the relative bias.

There was no bias for parameter estimates $\hat{\beta}_g$ for all the settings under the null hypothesis as is the case with simulation 1. When $\alpha_g > 0$, the parameter estimates $\hat{\beta}_g$ had a downward bias, since the fitting model was always misspecified in simulation 2.

Coverage probabilities of the five existing methods were not close to the nominal level owing to the covariate omission. In contrast, the coverage probabilities of our proposed method were closer to the nominal level than those of the existing methods.

From the above results, the proposed method can provide a robust test for the misspecification of the error distribution and mean structure.

Chapter 6

Case study

6.1 Acute myelogenous leukemia data

The acute myelogenous leukemia data are from the preliminary analysis of a clinical trial to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukemia.[29,30] Patients were randomly assigned receive either maintenance chemotherapy consisting of cytarabine and 6-thioguanine for two days each month or to receive no maintenance therapy. The objective of the trial was to examine if maintenance chemotherapy increased the length of remission. The sample sizes in the maintained group and the unmaintained group were 11 and 12, respectively. Figure 6.1 shows the Kaplan-Meier curves for each treatment group.

In our examination, the fitting model was $\log T = \beta_0 + \beta_g x_g + \varepsilon$, where x_g is a treatment group indicator ($1 =$ maintenance chemotherapy; $0 =$ no maintenance therapy). We fit the exponential, Weibull, log-logistic, and log-normal model to $\exp(\varepsilon)$. To test the effects of maintenance chemotherapy, we conducted the robust and naive Wald tests, chi-squared tests based on ordinary likelihood ratio statistic w, adjusted statistic w_K of w by Kent,[16] and adjusted

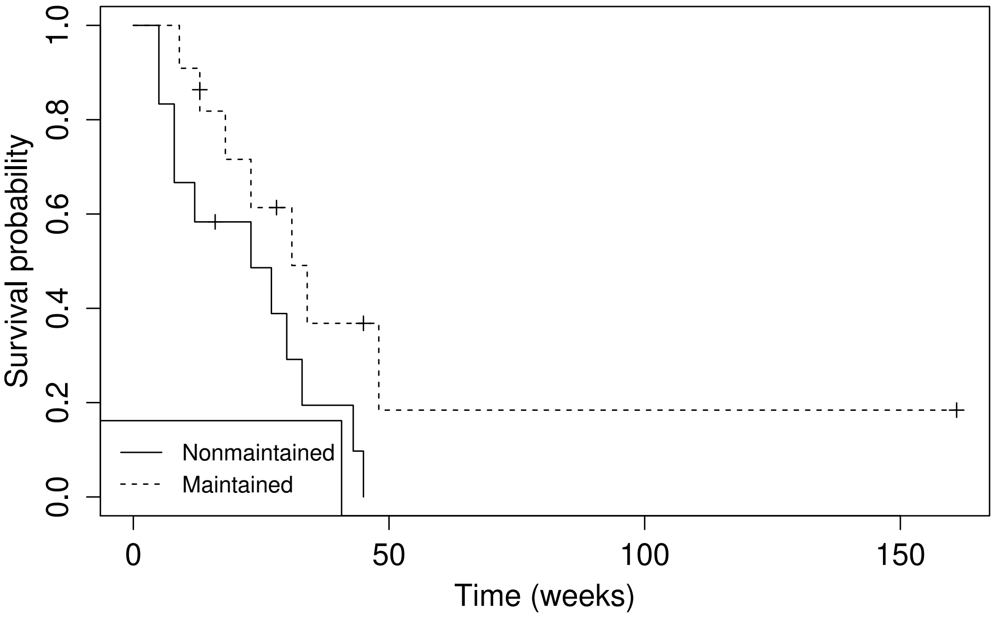

Figure 6.1: The Kaplan-Meier curves for each treatment group for acute myelogenous leukemia data.

statistic $w^* = w/\hat{E}_{g_0}[w]$ using the non-parametric bootstrap resampling for the null hypothesis $\beta_g = 0$ and calculated the p-values. To compare the goodness of fit, we calculated the maximum values of the log-likelihood function as a reference.

Table 6.1 presents the results for the analysis of acute myelogenous leukemia data. The p-values of the naive and robust Wald tests and the chi-squared tests based on w and w_K were dependent on the specified models. In particular, a choice of fitting model had an effect on whether the p-values for the naive Wald test and unadjusted likelihood ratio test are below the significance level 0.05. On the other hand, the p-values of our proposed method were uniformly larger than the significance level. Although the p-values of the robust Wald test and the chi-squared test based on w_K were also larger than the significance

Table 6.1: Results for acute myelogenous leukemia data.

	P-value					
	Wald		LR			
Model	Naive	Robust	w	w_K	w^*	LL
Exponential	0.0475	0.0579	0.0439	0.0538	0.0794	−81.3
Weibull	0.0151	0.0562	0.0212	0.0700	0.0798	−80.5
Log-logistic	0.1243	0.1578	0.1208	0.1540	0.1357	−79.4
Log-normal	0.0568	0.0705	0.0618	0.0762	0.0804	−78.9
Range of p-values	0.1092	0.1016	0.0996	0.1002	0.0563	

Note: LR is the likelihood ratio test, w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and LL is the maximum value of the log-likelihood function.

level, they varied much more than those of our proposed method. The reason why the p-values from the log-logistic model were much larger than those from other models would be the small sample and outlier. The range of p-values for our proposed method was the narrowest; hence, our proposed method provided stable results. Further, the p-values of our proposed method were larger than the unadjusted likelihood ratio test, and in most cases, larger than the robust Wald test and the chi-squared test based on w_K. The results of the simulation study in this article support these results. In addition, we expect from the simulation study that the results of the adjusted likelihood ratio test are robust to misspecification of mean structure, although we might omit an important covariate due to the paucity of available variables in the dataset.

6.2 Severe aplastic anemia data

The severe aplastic anemia data are from a randomized clinical trial on 64 patients with severe aplastic anemia.[3,31] Patients were randomized to cyclosporine and methotrexate (CSP + MTX) or methotrexate alone (MTX). An endpoint was the time from assignment until the diagnosis of a life-threatening stage of acute graft versus host disease. The sample sizes in each group were the same. Figure 6.2 shows the Kaplan-Meier curves for each treatment group.

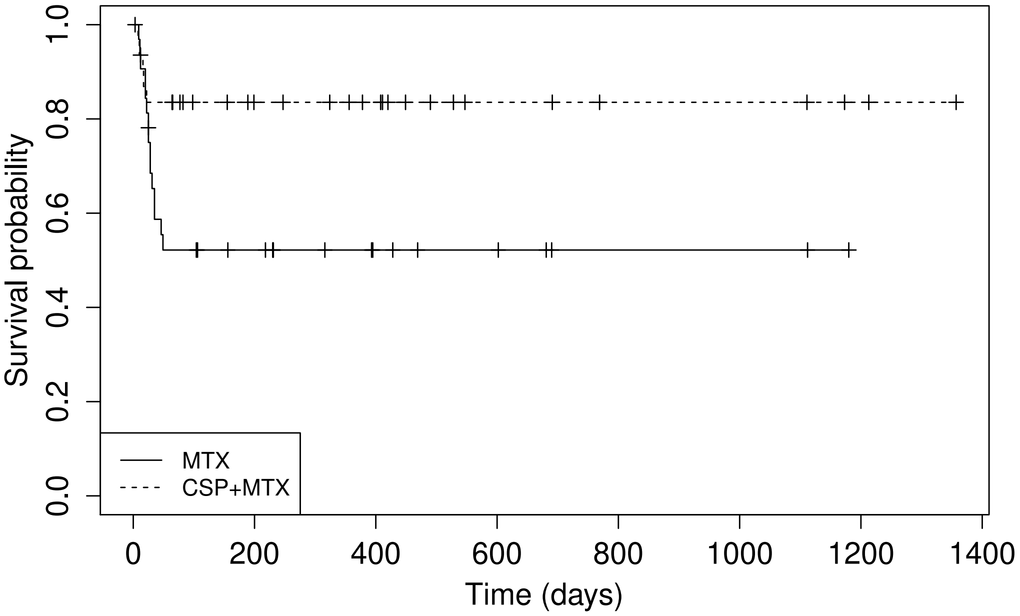

Figure 6.2: The Kaplan-Meier curves for each treatment group for severe aplastic anemia data.

In our examination, the fitting model was $\log T = \beta_0 + \beta_g x_g + \varepsilon$, where x_g is a treatment group indicator (1 = CSP + MTX; 0 = MTX). The other settings were the same as those for acute myelogenous leukemia data.

Table 6.2 presents the results for the severe aplastic anemia data. Similar to results for acute myelogenous leukemia data, the p-values of naive and robust Wald tests and the chi-squared tests based on w and w_K were dependent on the specified models. It is difficult to choose an appropriate model for the severe aplastic anemia data, as the events of interest were observed only in the early period of the trial and many patients were censored. In fact, as seen

Table 6.2: Results for severe aplastic anemia data.

	P-value					
	Wald		LR			
Model	Naive	Robust	w	w_K	w^*	LL
Exponential	0.0031	0.0162	0.0012	0.0087	0.0140	-152.5
Weibull	0.0190	0.0087	0.0059	0.0020	0.0112	-136.1
Log-logistic	0.0106	0.0063	0.0056	0.0031	0.0128	-134.8
Log-normal	0.0215	0.0112	0.0137	0.0065	0.0156	-133.8
Range of p-values	0.0184	0.0049	0.0125	0.0067	0.0044	

Note: LR is the likelihood ratio test, w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w by Kent,[16] w^* is the adjusted statistic by $\hat{E}_{g_0}[w]$ (proposal), and LL is the maximum value of the log-likelihood function.

in the maximum value of the likelihood function, the exponential model was not fitted well. Hence, model misspecification has affected the results from the exponential model. On the other hand, the p-values of our proposed method were slightly dependent on the fitting models. In addition, even when the exponential model was fitted, the p-values of our proposed method were larger than the unadjusted likelihood ratio test and, in most cases, larger than the robust Wald test and the chi-squared test based on w_K.

Chapter 7

Discussion

We proposed the procedure for estimating the the expectation of the likelihood ratio statistic in the finite sample size using the non-parametric bootstrap resampling in the two-group comparison. The simulation study indicated that our proposed method could control Type-I error rate in the two-group comparison based on the AFT model. In addition, we applied our proposed method to two actual time-to-event datasets with small-sample sizes and concluded that our proposed method could provide results that were robust to model misspecification.

Type-I error rates of existing methods (naive and robust Wald tests, ordinary likelihood ratio test, likelihood ratio test proposed by Kent,[16] and the test based on the percentile bootstrap confidence interval) could be controlled at the nominal level when the model is correctly specified and the sample size is sufficiently large. However, Type-I error rates of existing methods were not controlled in a small sample even when the model is correctly specified. Furthermore, when we misspecified the model, the naive Wald test and ordinary likelihood ratio test yielded the inflation of the Type-I error rate even in a large sample due to model misspecification. The robust Wald test and likelihood ratio test proposed by Kent[16] controlled the Type-I error rate in a large sample even under model misspecification. However, these two methods did not control the Type-I error rate in a small sample, because their correction factors are based on asymptotic properties. The test based on the percentile bootstrap confidence interval was conducted by estimating percentiles in tail areas; therefore, this estimation performs poorly in a small sample. In fact, this test yielded inflation in a small sample even under correct model specification. On the other hand, in our proposed method, we used the bootstrap method to estimate the mean of the likelihood ratio statistic, and not a percentile in the tail areas; thus, it is expected that the bootstrap method performs well even in a small sample. From the above results, the existing methods resulted in substantial inflation of the type 1 error rate under small-sample size and model misspecification; therefore, the existing methods are not practically useful for analysis with AFT models. In contrast, the adjusted likelihood ratio test we proposed had a Type-I error rate near the nominal level even under small-sample size and model misspecification. Hence, in terms of Type-I error rate, we could propose a practical test statistic for the AFT model.

However, our methods do have limitations. The point estimates of the treatment effect remain biased under the alternative hypothesis and model misspecification. Thus, the confidence interval based on our proposed method did not have sufficient performance under model misspecification owing to such bias. In addition, when the treatment effect was excessively large, the confidence interval based on our proposed method showed coverage probabilities larger than the nominal level. Hutton and Monaghan[44] proves the asymptotic unbiasedness of the treatment effect for the uncensored case, but not for the censored case. Hence, some additional corrections are required in future work. A flexible model that includes many models is useful to avoid such bias, since simulation results under the correct model specification were mostly unbiased. For example, the use of generalized gamma distribution,[45–47] which includes Weibull distribution and log-normal distribution as special cases, could be considered. In addition, diagnostic statistics for the distribution such as the Cox-Snell residual[48] might be useful to choose an appropriate distribution and reduce the bias of the point estimate. As an alternative strategy, the semi-parametric AFT model,[2,7] which does not need to specify the error distribution, provides an unbiased estimator for many underlying distributions.[49] However, the semi-parametric AFT model has not been used widely because of difficulties in computing the estimators and lack of efficient and reliable computational methods.[50,51] Furthermore, a small-sample problem for the semi-parametric AFT model has not been studied sufficiently. Hence, some small-sample corrections for the semi-parametric AFT model are required in future work.

Acknowledgments

I would like to express my sincere appreciation to my supervisor Professor Masahiko Gosho and vice-supervisor Associate Professor Kazushi Maruo for the invaluable support of my research. They provided not only beneficial comments about my PhD research but also important things about research activity and life in general. Without their encouragement and guidance, this dissertation would not have been possible.

I am deeply indebted to all of my referees Professor Koichi Hashimoto, Professor Tomoko Sankai, Professor Kenichi Koike, and Assistant Professor Masao Iwagami, who provided instructive comments to improve this dissertation.

I greatly appreciate all the members of Biostatistics Unit, Clinical and Translational Research Center, Keio University Hospital for their substantial support.

I wish to thank all the members of Biostatistics Research Organization for invaluable comments and encouragement. Discussions in forums and monthly meetings are essential for this dissertation.

I would like to thank all the members of the Gosho laboratory, Dr. Masashi Shimura, Mr. Keisuke Tada, Mr. Kenichi Takahashi, Mr. Tomohiro Ohigashi, and Mr. Satoshi Yoshida, for useful discussions.

I would like to thank Professor Tadashi Taniguchi, National Institute of Technology, Gunma College. Thanks to his interesting lecture, I became in-

terested in study and research of mathematics.

Finally, I am grateful to my parents and my wife for supporting and encouraging my research activity.

Source

This contents previously published in Statistics in Biopharmaceutical Research (doi: 10.1080/19466315.2020.1752297) are re-used in this dissertation based on the approval from Taylor & Francis.

Bibliography

[1] Cox DR. Regression Models and Life-Tables. J R Stat Soc Series B Stat Methodol. 1972;34(2):187–220.

[2] Miller RG. Least squares regression with censored data. Biometrika. 1976;63(3):449–464.

[3] Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. 2nd ed. New York: John Wiley and Sons, Inc.; 2002.

[4] Ishii R, Maruo K, Noma H, Gosho M. Statistical inference based on accelerated failure time models under model misspecification and small samples. Stat Biopharm Res. In press 2020.

[5] Kay R, Kinnersley N. On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: a case study in influenza. Drug Inf J. 2002;36(3):571–579.

[6] Patel K, Kay R, Rowell L. Comparing proportional hazards and accelerated failure time models: an application in influenza. Pharm Stat. 2006;5(3):213–224.

[7] Wei LJ. The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. Stat Med. 1992;11(14-15):1871–1879.

[8] Hay AD, Little P, Harnden A, Thompson M, Wang K, Kendrick D, et al. Effect of oral prednisolone on symptom duration and severity in nonasthmatic adults with acute lower respiratory tract infection: a randomized clinical trial. J Am Med Assoc. 2017;318(8):721–730.

[9] Dobson J, Whitley RJ, Pocock S, Monto AS. Oseltamivir treatment for influenza in adults: a meta-analysis of randomised controlled trials. Lancet. 2015;385 9979:1729–1737.

[10] Koti KM. Exponential Failure-Time Mixture Model Approach for Validating KRAS as a Predictive Biomarker for Panitumumab Monotherapy in the Treatment of Metastatic Colorectal Cancer. Stat Biopharm Res. 2011;3(3):425–433.

[11] Gosho M, Maruo K, Sato Y. Effect of covariate omission in Weibull accelerated failure time model: a caution. Biometrical J. 2014;56(6):991–1000.

[12] Cox DR. Tests of separate families of hypotheses. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics; 1961. p. 105–123.

[13] White H. Maximum likelihood estimation of misspecified models. Econometrica. 1982;50(1):1–25.

[14] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2018.

[15] Neyman J, Pearson ES. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. Biometrika. 1928;20A(1/2):175–240.

[16] Kent JT. Robust properties of likelihood ratio tests. Biometrika. 1982;69(1):19–27.

[17] Gosho M, Hirakawa A, Noma H, Maruo K, Sato Y. Comparison of bias-corrected covariance estimators for MMRM analysis in longitudinal data with dropouts. Stat Methods Med Res. 2017;26(5):2389–2406.

[18] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13–22.

[19] Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. J Am Stat Assoc. 2001;96(456):1387–1396.

[20] Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics. 2001;57(1):126–134.

[21] Viraswami K, Reid N. A note on the likelihood-ratio statistic under model misspecification. Can J Stat. 1998;26(1):161–168.

[22] Lunardon N. Towards a unification of second-order theory for likelihood and marginal composite likelihood. Biometrika. 2016;103(1):225–230.

[23] Bartlett MS. Properties of sufficiency and statistical tests. In: Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences; 1937. p. 268–282.

[24] Barndorff-Nielsen OE, Hall P. On the level-error after Bartlett adjustment of the likelihood ratio statistic. Biometrika. 1988;75(2):374–378.

[25] Barndorff-Nielsen OE, Cox DR. Inference and Asymptotics. London: Chapman and Hall; 1994.

[26] Loose LH, Valença DM, Bayer FM. On bootstrap testing inference in cure rate models. J Stat Comput Simul. 2018;88(17):3437–3454.

[27] Cordeiro GM, Cribari-Neto F. An Introduction to Bartlett Correction and Bias Reduction. New York: Springer; 2014.

[28] Rocke DM. Bootstrap Bartlett adjustment in seemingly unrelated regression. J Am Stat Assoc. 1989;84(406):598–601.

[29] Embury SH, Elias L, Heller PH, Hood CE, Greenberg PL, Schrier SL. Remission Maintenance Therapy in Acute Myelogenous Leukemia. West J Med. 1977;126(4):267–272.

[30] Miller RG, Gong G, Muñoz A. Survival Analysis. New York: John Wiley and Sons, Inc.; 1981.

[31] Storb R, Deeg HJ, Farewell V, Doney K, Appelbaum F, Beatty P, et al. Marrow transplantation for severe aplastic anemia: methotrexate alone compared with a combination of methotrexate and cyclosporine for prevention of acute graft-versus-host disease. Blood. 1986;68(1):119–125.

[32] Collett D. Modelling Survival Data in Medical Research. 2nd ed. London: Chapman and Hall; 2003.

[33] SAS Institute. SAS/STAT 9.4 User's Guide. SAS. Cary, NC, USA; 2019.

[34] Cox DR, Oakes D. Analysis of Survival Data. London: Chapman and Hall; 1984.

[35] Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer-Verlag; 2003.

[36] V B, Nikulin MS. Accelerated Life Models: Modeling and Statistical Analysis. London: Chapman and Hall; 2001.

[37] Gail MH, Weiand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika. 1984;71(3):431–444.

[38] Bretagnolle J, Huber-Carol C. Effects of Omitting Covariates in Cox's Model for Survival Data. Scand J Stat. 1988;15(2):125–138.

[39] Hauck WW, Anderson S, Marcus SM. Should We Adjust for Covariates in Nonlinear Regression Analyses of Randomized Trials? Control Clin Trials. 1998;19(3):249 – 256.

[40] Lagakos SW, Schoenfeld DA. Properties of Proportional-Hazards Score Tests under Misspecified Regression Models. Biometrics. 1984;40(4):1037–1048.

[41] Efron B. Bootstrap methods: another look at the jackknife. Ann Stat. 1979;7(1):1–26.

[42] Bickel PJ, Freedman DA. Some asymptotic theory for the bootstrap. Ann Stat. 1981;9(6):1196–1217.

[43] Pawitan Y. In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford: Oxford University Press; 2001.

[44] Hutton JL, Monaghan PF. Choice of parametric accelerated life and proportional hazards models for survival data: asymptotic results. Lifetime Data Anal. 2002;8(4):375–393.

[45] Stacy EW. A generalization of the gamma distribution. Ann Math Stat. 1962;33(3):1187–1192.

[46] Prentice RL. A log gamma model and its maximum likelihood estimation. Biometrika. 1974;61(3):539–544.

[47] Cox C, Chu H, Schneider MF, Muñoz A. Parametic survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Stat Med. 2007;26:4352–74.

[48] Cox DR, Snell EJ. A General Definition of Residuals. J R Stat Soc Series B Stat Methodol. 1968;30(2):248–275.

[49] Zeng D, Lin DY. Efficient Estimation for the Accelerated Failure Time Model. J Am Stat Assoc. 2007;102(480):1387–1396.

[50] Jin Z, Lin DY, Wei LJ, Ying Z. Rank-Based Inference for the Accelerated Failure Time Model. Biometrika. 2003;90(2):341–353.

[51] Huang J, Ma S, Xie H. Regularized Estimation in the Accelerated Failure Time Model with High-Dimensional Covariates. Biometrics. 2006;62(3):813–820.

Appendix A

R function for our proposed method

A.1 Details of R function

The function `AFT_Bartlett` in section A.3 provides results of our proposed method. The function `AFT_Bartlett` has arguments `indat`, `fitdist`, `B`, and `seed`. `indat` is a data frame intended to be analyzed, which has following three variables.

time: survival time or censoring time

event: indicator variable, which takes one if the subject experiences event and zero if the subject experiences censoring

group: group indicator variable (e.g., 0 = placebo group; 1 = active group)

`fitdist` specifies a fitting distribution of the error term. In the function `AFT_Bartlett`, we can specify `exponential`, `weibull`, `loglogistic`, and `lognormal`; other distributions are not supported. `B` is the number of resampling arising from the approximation (4.1). `seed` is a random seed.

The function AFT_Bartlett provides the results of naive and robust Wald tests and chi-squared tests based on w, w_K, and $w^* = w/\hat{E}_{g_0}[w]$, where w is the ordinary likelihood ratio statistic, w_K is the adjusted statistic of w proposed by Kent,[16] and w^* is the adjusted statistic using the non-parametric bootstrap method. In the Wald tests, a point estimate of the treatment effect, its standard error, z-value, and p-value are presented. On the other hands, in the likelihood ratio tests, test statistic (w, w_K, or $w^* = w/\hat{E}_{g_0}[w]$), degree of freedom, and p-value are presented.

A.2 Example

We provide an example of an analysis for the acute myelogenous leukemia dataset used in section 6.1. In the following R code, we analyze the dataset by fitting Weibull distribution. Let the number of resampling B be 1,000 and let the random seed be 5678.

```
# observed times
time <- c(9,13,13,18,23,28,31,34,45,48,161,
          5,5,8,8,12,16,23,27,30,33,43,45)
# event indicators
event <- c(1,1,0,1,1,0,1,1,0,1,0,1,1,1,1,1,0,1,1,1,1,1,1)
# group indicators
group <- c(1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0)

leukemiadat <- data.frame(time=time,event=event,group=group)

AFT_Bartlett(leukemiadat,'weibull',1000,5678)
```

The results are shown as follows. Here, Loglik is the maximum value of the log-likelihood function.

```
$`Information`
                                value
number of resampling             1000
fitting distribution          weibull
Loglik                -80.5216452034027


$Wald
        Estimate Std. Error        z          p
Naive  0.9293416  0.3825019 2.429639 0.01511385
Robust 0.9293416  0.4867046 1.909457 0.05620314


$LR
                 w df          p
Ordinary  5.314048  1 0.02115415
Kent      3.282175  1 0.07003608
Bootstrap 3.050050  1 0.08073464
```

A.3 R code

```
AFT_Bartlett <- function(indat,fitdist,B,seed){
  library(survival)
  library(MASS)
  set.seed(seed)
  n0 <- sum(indat$group == 0)
  n1 <- sum(indat$group == 1)
  b <- 1:B
```

```
for(i in 1:B){
  samp <- sample(1:(n0+n1),replace=T)
  btime <- matrix((indat$time)[samp],ncol=1)
  bevent <- matrix((indat$event)[samp],ncol=1)
  gdat <- matrix(c(numeric(n0),numeric(n1)+1),ncol=1)

  btry <- try(bres <- survreg(formula = Surv(btime,bevent)~
                              gdat,dist = fitdist))

  if(class(btry) != ’try-error’){
    b[i] <- 2*(bres$loglik[2] - bres$loglik[1])
    bscale <- bres$scale
  }else{
    b[i] <- NaN
    bscale <- NaN
  }
}
b <- b[!is.nan(b)]

resn <- survreg(formula = Surv(time, event)~group,
                data = indat,dist = fitdist)
tablen <- summary(resn)$table
resr <- survreg(formula = Surv(time, event)~group,
                data = indat,dist = fitdist,robust = T)
tabler <- summary(resr)$table

naive <- matrix(tablen[2,],1,4)
```

```
robust <- matrix(tabler[2,-3],1,4)

cname1 <- c("Estimate","Std. Error","z","p")
rname1 <- c("Naive","Robust")

Wald <- as.data.frame(rbind(naive,robust),row.names = rname1)
colnames(Wald) <- cname1

w <- 2*(resn$loglik[2] - resn$loglik[1])
adjw <- w/mean(b)
ordinaryLR <- matrix(c(w,1,1-pchisq(w,df=1)),1,3)
proposedLR <- matrix(c(adjw,1,1-pchisq(adjw,df=1)),1,3)

dat <- indat$time
xdat <- indat$group
flg <- indat$event

if(fitdist == "exponential"){
  mle <- c(resn$coefficients)
  beta0hat <- mle[1]
  betaghat <- mle[2]

  z <- log(dat) - (beta0hat+betaghat*xdat)
  ell_0 <- -flg + exp(z)
  ell_g <- (-flg + exp(z))*xdat

  # score
```

```
    j_00 <- sum(ell_0*ell_0)
    j_g0 <- sum(ell_g*ell_0)
    j_gg <- sum(ell_g*ell_g)

    j <- c(j_gg,j_g0,j_g0,j_00)
    J <- matrix(j,2,2)

    # Hesse matrix
    ell_00 <- -exp(z)
    ell_g0 <- -exp(z)*xdat
    ell_gg <- -exp(z)*xdat

    h_00 <- sum(ell_00)
    h_g0 <- sum(ell_g0)
    h_gg <- sum(ell_gg)

    h <- c(h_gg,h_g0,h_g0,h_00)
    H <- matrix(-h,2,2)
  }
  if(fitdist == "weibull"){
    mle <- c(resn$scale,resn$coefficients)
    sigmahat <- mle[1]
    beta0hat <- mle[2]
    betaghat <- mle[3]

    z <- (log(dat) - (beta0hat+betaghat*xdat))/sigmahat
    ell_s <- flg*(-1/sigmahat - z/sigmahat) +
      z*exp(z)/sigmahat
    ell_0 <- -flg/sigmahat+exp(z)/sigmahat
    ell_g <- -flg*xdat/sigmahat+exp(z)*xdat/sigmahat

    # score
    j_ss <- sum(ell_s*ell_s)
    j_0s <- sum(ell_0*ell_s)
    j_gs <- sum(ell_g*ell_s)
    j_00 <- sum(ell_0*ell_0)
    j_g0 <- sum(ell_g*ell_0)
    j_gg <- sum(ell_g*ell_g)

    j <- c(j_gg,j_g0,j_gs,j_g0,j_00,j_0s,j_gs,j_0s,j_ss)
    J <- matrix(j,3,3)

    # Hesse matrix
    ell_ss <- sigmahat^(-2)*
      (flg*(1+2*z) - 2*z*exp(z) - z^2*exp(z))
    ell_0s <- sigmahat^(-2)*(flg - exp(z) - z*exp(z))
    ell_gs <- sigmahat^(-2)*(flg - exp(z) - z*exp(z))*xdat
    ell_00 <- sigmahat^(-2)*(-exp(z))
    ell_g0 <- sigmahat^(-2)*(-exp(z))*xdat
    ell_gg <- sigmahat^(-2)*(-exp(z))*xdat

    h_ss <- sum(ell_ss)
    h_0s <- sum(ell_0s)
    h_gs <- sum(ell_gs)
```

```
  h_00 <- sum(ell_00)

  h_g0 <- sum(ell_g0)

  h_gg <- sum(ell_gg)


  h <- c(h_gg,h_g0,h_gs,h_g0,h_00,h_0s,h_gs,h_0s,h_ss)

  H <- matrix(-h,3,3)

}

if(fitdist == "loglogistic"){

  mle <- c(resn$scale,resn$coefficients)

  sigmahat <- mle[1]

  beta0hat <- mle[2]

  betaghat <- mle[3]


  z <- (log(dat) - (beta0hat+betaghat*xdat))/sigmahat

  a <- exp(z)/(1+exp(z))

  ell_s <- (-flg*z - flg + (1+flg)*z*a)/sigmahat

  ell_0 <- (-flg + (1+flg)*a)/sigmahat

  ell_g <- (-flg + (1+flg)*a)*xdat/sigmahat


  # score

  j_ss <- sum(ell_s*ell_s)

  j_0s <- sum(ell_0*ell_s)

  j_gs <- sum(ell_g*ell_s)

  j_00 <- sum(ell_0*ell_0)

  j_g0 <- sum(ell_g*ell_0)

  j_gg <- sum(ell_g*ell_g)
```

```
j <- c(j_gg,j_g0,j_gs,j_g0,j_00,j_0s,j_gs,j_0s,j_ss)
J <- matrix(j,3,3)

# Hesse matrix
z_s <- -z/sigmahat
z_0 <- -1/sigmahat
z_g <- -xdat/sigmahat

u_s <- -flg*z - flg + (1+flg)*z*a
u_0 <- flg*(1-2/(1+exp(z))) + (1-flg)*(1-1/(1+exp(z)))
u_g <- flg*xdat*(1-2/(1+exp(z))) +
  (1-flg)*xdat*(1-1/(1+exp(z)))

ell_ss <- -sigmahat^(-2)*u_s + sigmahat^(-1)*z_s*
  (-flg+(1+flg)*exp(z)*(1+z+exp(z))/(1+exp(z))^2)
ell_0s <- -sigmahat^(-2)*u_0 +
  sigmahat^(-1)*(1+flg)*z_s*exp(z)/(1+exp(z))^2
ell_gs <- -sigmahat^(-2)*u_g +
  sigmahat^(-1)*(1+flg)*xdat*z_s*exp(z)/(1+exp(z))^2
ell_00 <- sigmahat^(-1)*(1+flg)*z_0*exp(z)/(1+exp(z))^2
ell_g0 <- sigmahat^(-1)*(1+flg)*z_g*exp(z)/(1+exp(z))^2
ell_gg <- sigmahat^(-1)*
  (1+flg)*xdat*z_g*exp(z)/(1+exp(z))^2

h_ss <- sum(ell_ss)
h_0s <- sum(ell_0s)
h_gs <- sum(ell_gs)
```

```
  h_00 <- sum(ell_00)
  h_g0 <- sum(ell_g0)
  h_gg <- sum(ell_gg)

  h <- c(h_gg,h_g0,h_gs,h_g0,h_00,h_0s,h_gs,h_0s,h_ss)
  H <- matrix(-h,3,3)
}
if(fitdist == "lognormal"){
  mle <- c(resn$scale,resn$coefficients)
  sigmahat <- mle[1]
  beta0hat <- mle[2]
  betaghat <- mle[3]

  z <- (log(dat) - (beta0hat+betaghat*xdat))/sigmahat
  a <- dnorm(z)/(1-pnorm(z))
  ell_s <- (flg*(-1+z^2) + (1-flg)*z*a)/sigmahat
  ell_0 <- (flg*z + (1-flg)*a)/sigmahat
  ell_g <- (flg*z + (1-flg)*a)*xdat/sigmahat

  # score
  j_ss <- sum(ell_s*ell_s)
  j_0s <- sum(ell_0*ell_s)
  j_gs <- sum(ell_g*ell_s)
  j_00 <- sum(ell_0*ell_0)
  j_g0 <- sum(ell_g*ell_0)
  j_gg <- sum(ell_g*ell_g)
```

```
j <- c(j_gg,j_g0,j_gs,j_g0,j_00,j_0s,j_gs,j_0s,j_ss)
J <- matrix(j,3,3)

# Hesse matrix
z_s <- -z/sigmahat
z_0 <- -1/sigmahat
z_g <- -xdat/sigmahat

dphi <- -z*exp(-z^2/2)/sqrt(2*pi)

u_s <- flg*(-1+z^2) + (1-flg)*z*a
u_0 <- flg*z + (1-flg)*a
u_g <- (flg*z + (1-flg)*a)*xdat

pz <- pnorm(z)
dz <- dnorm(z)

ell_ss <- -sigmahat^(-2)*u_s + sigmahat^(-1)*z_s*
  (2*flg*z + (1-flg)*((dz+z*dphi)*(1-pz)+z*dz^2)/(1-pz)^2)
ell_0s <- -sigmahat^(-2)*u_0 + sigmahat^(-1)*z_s*
  (flg + (1-flg)*(dphi*(1-pz)+dz^2)/(1-pz)^2)
ell_gs <- -sigmahat^(-2)*u_g + sigmahat^(-1)*z_s*xdat*
  (flg + (1-flg)*(dphi*(1-pz)+dz^2)/(1-pnorm(z))^2)
ell_00 <- sigmahat^(-1)*z_0*
  (flg + (1-flg)*(dphi*(1-pz)+dz^2)/(1-pz)^2)
ell_g0 <- sigmahat^(-1)*z_g*
  (flg + (1-flg)*(dphi*(1-pz)+dz^2)/(1-pz)^2)
  ell_gg <- sigmahat^(-1)*z_g*xdat*
    (flg + (1-flg)*(dphi*(1-pz)+dz^2)/(1-pz)^2)

  h_ss <- sum(ell_ss)
  h_0s <- sum(ell_0s)
  h_gs <- sum(ell_gs)
  h_00 <- sum(ell_00)
  h_g0 <- sum(ell_g0)
  h_gg <- sum(ell_gg)

  h <- c(h_gg,h_g0,h_gs,h_g0,h_00,h_0s,h_gs,h_0s,h_ss)
  H <- matrix(-h,3,3)
}
Hinv <- ginv(H)
Kent <- Hinv[1,1]^(-1)*((Hinv%*%J%*%Hinv)[1,1])

wK <- w/Kent
KentLR <- matrix(c(wK,1,1-pchisq(wK,df=1)),1,3)



cname2 <- c("w","df","p")
rname2 <- c("Ordinary","Kent","Bootstrap")

LR <- as.data.frame(rbind(ordinaryLR,KentLR,proposedLR),
                    row.names = rname2)
colnames(LR) <- cname2
```

```
  Info <- data.frame(value = c(B,fitdist,resn$loglik[2]),
                     row.names = c("number of resampling",
                                   "fitting distribution",
                                   "Loglik"))
  Info$value <- as.character(Info$value)

  outdat <- list(Information=Info,Wald=Wald,LR=LR)

  return(outdat)
}
```