Chapter 5

Simulation

This chapter contains discussion of a simulation study of estimates in the log hazard regression model. Our main aim is to check the bias and variability of our estimates for the log regression model. We consider different censoring rates and different correlation levels of the paired failure times. We also investigate the three test procedures: for exponential or Weibull marginal distributions and for proportional hazards, see Section 3.5.

Our study consists of two parts: the univariate case and the bivariate case. In the univariate case, presented in Section 5.2, we investigate the estimates of the marginal log hazard functions and the pointwise standard errors. We also test that the marginals follow exponential or Weibull distributions. In the bivariate case, presented in Section 5.3, we examine the estimated log hazards ratios and their estimated standard errors. We also test that the paired failure times follow the proportional hazards model.

In Section 5.1 we give a brief description of our data generation and model fitting.

We present the simulation results in Sections 5.2 and 5.3 for the marginal and bivariate data, respectively.

In this simulation study we find that, in general, the estimates for the log hazards perform well except in the tails of the failure time distribution or when the censoring is high. The estimated standard errors slightly underestimate the true variability, which is fairly small. The estimated standard errors for the log hazards ratios do not depend on the correlation level too much. All of the estimates depend on the censoring rate. The lower the censoring, the better the results. We will give a summary of this study in Section 5.4.

5.1 Description of the Simulation Study

In this study for each distribution of T considered, we generate 200 pairs Ts and Cs in each simulation, where the Ts are failure times, the Cs are censoring times, and the Ts and Cs are independent. We choose two censoring rates $c = 0.25$ or 0.50, i.e.,

$$P(T > C) = 0.25 \qquad \text{or} \qquad P(T > C) = 0.50.$$

We generate the failure time T from distributions with log hazards of the form

$$\alpha(t|\boldsymbol{\beta}) = \beta_1 b_1(t) + \cdots + \beta_p b_p(t), \tag{5.1}$$

where $\{b_1, \cdots, b_{p-1}\}$ is a B-spline basis and $b_p(t) \equiv \log(t)$. We use three distributions for the marginal failure time T:

Exponential :

$$\alpha_T(t) \equiv -\log(20). \tag{5.2}$$

In this case, we choose the censoring time C with

$$\alpha_C(t) \equiv -\log(20)$$

for a 50% censoring rate or

$$\alpha_C(t) \equiv -\log(60)$$

for 25% censoring.

Weibull :

$$\alpha_T(t) = -3.68 - 0.18\log(t). \tag{5.3}$$

The corresponding censoring time C has the log hazard

$$\alpha_C(t) = -3.68 - 0.18\log(t)$$

for 50% censoring or

$$\alpha_C(t) = -4.69 - 0.18\log(t)$$

for 25% censoring.

B-spline :

$$\alpha_T(t) = 0.00017\, b_1(t) - 8.13 + 0.69\log(t), \tag{5.4}$$

where b_1 is linear on $[0, 1.5)$, cubic on $[1.5, 6.17)$ and $[6.17, 63.33)$, and zero on $[63.33, \infty)$. See Sections 3.2 and 4.2 for the exact definition of b_1. The censoring time follows a Weibull with log hazard

$$\alpha_C(t) = -13.39 + 2\log(t)$$

for 50% censoring or

$$\alpha_C(t) = -25.24 + 4\log(t)$$

for a 25% censoring rate.

In all three cases, we choose the parameters for T's distribution to fit the eye data in Chapter 4. We choose the exponential distribution with mean close to the empirical mean of the non-censored failure times of the treated eyes; the Weibull which is the distribution of the estimated distribution of the failure time of the untreated eye based on the given data. The B-spline model (5.4) is the estimated distribution of the failure times of the treated eye as calculated in Section 4.2. We use S-plus to generate data from an exponential or Weibull distribution. For the data as in the B-spline model (5.4), we first generate a random variable $u \sim U(0, 1)$ then solve $S(t) = u$ for t numerically. Since t is the $(1-u)$th quantile of T's distribution, t can be calculated via heft's quantile function.

We use the following three methods to estimate the log hazard function.

True Model : Fit the data assuming it follows the true model. That is, estimate the parameters in (5.1) using the true b_j's. In this case both the number and the locations of the knots do not depend on the data set.

Quantile Knots : Use the quantiles of the non-censored observations as the knots that define the b_j's in (5.1). We consider two cases: three knots and six knots. In the three knots case, our knots are the quartiles. In the six knots case, our knots are the $1/6, \cdots, 5/6$ quantiles. Thus, the locations of the knots depend on the data set, but the number of the knots does not.

Flexible Knots : Choose the knots that define the b_j's in (5.1) by Kooperberg's *heft* algorithm. In this case both the number and locations of the knots depend on the data set.

Thus we have six models for (T, C), from the three distributions for the failure times and the two censoring times for each distribution. For each of these six models we run 500 simulations and calculate the above four estimates of the log hazard functions.

In the bivariate case, we will use the Clayton method to generate (T_1, T_2) from marginals for T_1 and T_2 and with a parameter θ, (see Section 5.3.1). We consider two types of paired data:

Proportional Hazards : The failure time T_1 and censoring time C_1, say for the treatment, are as in the univariate B-spline model. Then the censoring rates for T_1 are 0.5 and 0.25. The failure time T_2, say for the control, is from the distribution with the log hazard

$$\alpha_2(t) = 0.00017\, b_1(t) - 7.13 + 0.69 \log(t). \tag{5.5}$$

Then T_1 and T_2 have the proportional hazards. We assume that the treatment and control have the same censoring time $C = C_1 = C_2$.

Non-proportional Hazards : The failure time T_1 and censoring time C_1 of the treatment are as in the univariate B-spline model. The failure time T_2 of the control is from the exponential model (5.2). Thus T_1 and T_2 do not have proportional hazards. Again we also assume that the treatment and the control have the same censoring time C. So the censoring rates of T_1 are 0.5 and 0.25.

The correlation level of T_1 and T_2 is determined by a parameter θ. We will give more details about θ in Section 5.3.1. In this study we consider three correlation levels between T_1 and T_2: $\theta = 1$, in which case T_1 and T_2 are independent; $\theta = 1.5$ and $\theta = 2.5$. The bigger θ, the more T_1 and T_2 are correlated.

We use the following three methods to estimate the log hazards ratios of the paired data.

Cox : Assume T_1 and T_2 are independent and have proportional hazards. Use the Cox proportional hazards model, as implemented in Splus's *coxph*.

Same Knots : Use the Quantile Knots method with the same six knots for T_1 and T_2 to estimate each marginal log hazard. The knots are the $1/7, 2/7, 3/7, 4/7, 5/7$, $6/7$ quantiles of the union of the non-censored observations of T_1 and T_2 which are less than 80. See our explanation in Section 3.4.

Different Knots : Use the Quantile Knots method with six knots for estimating T_1's log hazard and six knots for T_2's when T_1 and T_2 have proportional hazards, and three knots for T_2's when T_1 and T_2 do not have proportional hazard. Then the knots are different for T_1 and T_2 in each simulation.

Hence in the bivariate case we have 12 models for (T_1, T_2, C): two pairs of marginals for T_1, T_2, two censoring rates, and three correlation levels. For each of the 12 models we run 500 simulations and use the above three methods to estimate the log hazards ratios.

In all of these simulations, we will look at:

- plots of the pointwise averages and quantiles of the estimates to study the bias and variability of the estimates.
- plots of the pointwise standard deviations of the estimates to study the variability of the estimates.
- plots of the quantiles of the pointwise z values of the estimates to assess the reliability of the estimates. The pointwise z values are constructed as
$$z(t) = \frac{\text{estimate - true}}{\text{estimated SE}}.$$
- plots of the empirical distribution of the p-values for testing that the failure times have exponential or Weibull marginal distributions and for testing that the two failure times have proportional hazards.

Plots appear at the end of this chapter and in the appendix.

5.2 The Univariate Case

We present our simulation results for the models (5.2), (5.3), and (5.4) below in Sections 5.2.1, 5.2.2, and 5.2.3, respectively.

5.2.1 Exponential Model

In this section we apply the True Model, Quantile Knots, and Flexible Knots estimation methods to the data generated from the exponential distribution (5.2). Figure 5.1 shows the histograms of the failure times, the censoring times, the non-censored data and the censored data for those distributions. Since 60 is the 98th percentile of non-censored observations under 25% censoring and the 99.7th percentile of the non-censored observations under 50% censoring, we only plot estimates for values of t between 0 and 60. See Figures 5.2 to 5.4.

Note that in the Flexible Knots estimation method, if no knots are chosen and the coefficient of the log tail is zero, then an exponential model is fit. This occurs 468 out of 500 times for 25% censoring and 481 out of 500 times for 50% censoring. Hence the estimates from the Flexible Knots method are usually the same as from the True Model. So we do not include the results from the Flexible Knots estimation method in these figures.

Figure 5.2 shows the quartiles and empirical mean of the estimated log hazards from the True Model, Quantile Knots, and Flexible Knots methods and the true log hazard function for the exponential model (5.2). As expected, we find that the estimates from the True Model method are the closest to the true log hazard. For the Quantile Knots method, using three knots seems better than using six knots, which is somewhat surprising. In general, we expect that the more knots are used, the smaller the bias of the estimates but the larger the variability. The six knots estimates are more variable but they are also more biased. This may be since the true distribution is exponential and so no knots are needed. Using more knots would not reduce the bias but it would increase the variability of the estimates. It is not surprising that the higher the censoring rate, the higher the bias of the estimates.

Figure 5.3 shows the pointwise standard deviations of the estimated log hazards and the quartiles and the empirical mean of the estimated standard errors of the estimated log hazards. It is normal that the True Model method has the smallest pointwise standard deviations. The standard deviations using six knots are bigger than when using three knots. The higher the censoring rate, the bigger the standard deviations. Comparing the the pointwise standard deviations of the estimated log hazards with the pointwise quantiles of the estimated standard errors, we find that the bias of the estimated standard errors is the smallest with True Model method and the largest when using six knots. The higher censoring rate causes a bigger bias of the estimated standard errors except with the True Model method.

We give the 97.5 and 2.5 percentiles and the quartiles of the pointwise z-values of the 500 estimated log hazards in Figure 5.4. We expect that they are between -2 and 2. We find that they are not out of this range too much for all estimates. Thus, pointwise confidence intervals based on the estimates would be reliable.

Figure 5.5 presents the histograms and the qq-plots of

$$z_i = \frac{\hat{\beta}_i - \beta_i}{\hat{\text{se}}_{\beta_i}},$$

$i = 1, 2, 3$, the standardized estimates of $\boldsymbol{\beta}$ with the Quantile method using three knots.

Those figures show that the estimated $\boldsymbol{\beta}$s are approximately normal. For the method using six knots, the graphs look the same. We do not show them here.

We also check three test procedures that the failure times follow an exponential distribution. Each procedure involves fitting the log hazard by same regression model and then using a χ^2 test that some regression parameters are equal to zero. First we fit the data from model 5.2 to a Weibull model, that is,

$$\alpha(t|\boldsymbol{\beta}) = \beta_1 \log(t) + \beta_2.$$

Then we use $(\hat{\beta}_1/\hat{se}(\beta_1))^2$ as the test statistic to test $\beta_1 = 0$. This is the usual parametric test of exponential versus Weibull. When using the Quantile Knots method, equation (5.1) can be written as

$$\alpha(t|\boldsymbol{\beta}) = \beta_1 b_1(t) + \beta_2 b_2(t) + \beta_3 b_3(t)$$

when using three knots and

$$\alpha(t|\boldsymbol{\beta}) = \beta_1 b_1(t) + \cdots + \beta_5 b_5(t) + \beta_6 b_6(t)$$

when using six knots. As discussed in Section 3.5.1, the test is equivalent to test that

$$\beta_1 = \beta_3 = 0$$

when using three knots and

$$\beta_1 = \cdots = \beta_4 = \beta_6 = 0$$

when using six knots. Using the χ^2 statistics discussed in Section 3.5.1, we can calculate p-values of the test for each simulation.

In Figure 5.6, we give plots of the empirical distributions of the p-value to test the hypothesis that the failure times have an exponential distribution. The plots in the top row in Figure 5.6 show the p-values for the test of exponential versus Weibull, and those in the second and third rows are the p-values for the test of exponential versus B-spline model as in (5.4) when using three and six knots, respectively. Since the null hypothesis is true, the p-value should be uniformly distributed. Thus we expect that each plot would be a straight line. We find that the test works well and it does not depend on the knot selection or the censoring rates too much. However, the test rejects a little too often when the censoring rate is high.

5.2.2 Weibull Model

As in Section 5.2.1, we first present the histograms of the data in Figure 5.7. From the histograms we can see that, for both censoring rates, most non-censored failure times were less than 150. In fact 150 is about the 98th percentile for 50% censoring and the 95th percentile for 25% censoring of the observations. So in Figures 5.8 to 5.10, we end our plots at time 150.

As we mentioned in Section 3.4, the large range of the distribution of non-censored failure times may cause numerical problems for the Quantiles Knots and the Flexible Knots methods. This happens in our simulations for model (5.3) when we use the Flexible Knots method. Using the knots chosen by the *heft* algorithm we get four warning messages under 25% censoring and ten warning messages under 50% censoring. Moreover, we can not carry out the calculation for the estimated standard error for any simulated data sets due to the big ranges of the knots. Hence with the Flexible Knots methods we can only get the results of the estimated log hazards for simulated data sets that did not produce warning messages. We can not get any results related to the estimated standard errors. However, the Quantile Knots Method works well for the data as in model (5.3).

Figure 5.8 shows summary plots of the estimated log hazards compared to the true log hazard of the model (5.3). We find that under 25% censoring the True Model and Quantile Knots estimation methods perform well. The estimates with the Flexible Knots method are highly biased when time is large, especially for high censoring.

When we look at the pointwise standard deviations in Figure 5.9, as expected, the True Model method gives the smallest standard deviations. The Quantile Knots method using six knots has the largest standard deviations.

When we compare the pointwise quartiles of our estimated standard error with the pointwise standard deviations of the estimated log hazards, we find that all the estimates look fine, except that the estimates are a little bit too small after time 40. The censoring rate effect is the same as in the exponential model: the higher the censoring rate, the bigger the bias and variability of the estimates of the standard errors, particularly for small values of time.

Figure 5.10 shows the pointwise 2.5 and 97.5 percentiles and quartiles of the empirical z values of the estimated log hazards. They are as expected for standard normal random variables.

In Figure 5.11, we give the histograms and qq-plots of the estimates of $\boldsymbol{\beta}$ from the

Quantile Knots method with six knots under 50% censoring. From the plots we can see that the estimates of $\boldsymbol{\beta}$ are approximately normal. For the True Model and three knots estimates, the graphs look the same. We do not present them here.

For test procedures, we check not only the tests that data follow an exponential distribution as in Section 5.2.1, but also test that data follow a Weibull distribution. For the Weibull test we fit the data using b_j's from the Quantile Knots estimation method. We use the test statistics discussed in Section 3.5.1 to test the hypothesis that the data follow a Weibull distribution. The test is equivalent to testing that $\beta_1 = 0$ when using three knots and that $\beta_1 = \cdots = \beta_4 = 0$ when using six knots. Figures 5.12 and 5.13 show the plots of the empirical distribution functions of the p-values for testing the null hypotheses that the distribution is exponential and that the distribution is Weibull, respectively. We expect a high proportion of small p-values in Figure 5.12 and a straight line in Figure 5.13.

From Figure 5.12, we see that under 25% censoring the True Model method and the Quantile Knots method with six knots has the lowest power. Increasing censoring to 50% results in lower power for all methods.

From Figure 5.13, we find that the test for Weibull works very well and it depends on neither the number of the knots nor the censoring rates.

5.2.3 The B-spline Model

We first look at the histograms of the failure and censoring distributions, the censored failure times, and the non-censored failure times presented in Figure 5.14. From the histograms for the non-censored failure times, we see a high proportion of times on the interval $[0, 20]$, with the remaining times almost uniformly distributed on $[20, 150]$ for censoring rate 25% and on $[20, 100]$ for censoring rate 50%.

As in the Weibull model (5.3) case, the large range of the distribution of non-censored failure times causes some numerical problems for the B-spline model when we use the Quantiles Knots and the Flexible Knots methods. Using the Flexible Knots method, that is, using the knots chosen by the *heft* algorithm, we can calculate only 113 estimated standard errors out of 500 simulations under 25% censoring and 306 under 50% censoring. The same problem occurs when we use six knots with the Quantile Knots method. To avoid those problems, when we use six knots we choose the knots which are the quantiles of the non-censored failure times less than 80. We present the simulation results in Figures 5.2.15 to 5.2.19. The results for the Flexible Knots method are based on the estimates that we can get.

By comparing the estimated log hazards with the true hazard function in Figure 5.15, we fined that the biases of all estimates are fairly small when time is less than 70 under 25% censoring and when time is less than 40 under 50% censoring. We also find that those two numbers, 70 and 40, are close to the medians, 70.82 and 36.89, of the non-censored failure times in the two models. The estimates with the Flexible method perform well until time 100 under 25% censoring. We expect that using larger knots would get better estimates at tail part if there were not numerical calculation problems.

Figure 5.16 gives the pointwise standard deviations of the estimated log hazards and the quantiles and the empirical mean of estimated standard errors. From those plots we find that the True Model method and using three knots give the similar pointwise standard deviations which are smaller than the pointwise standard deviations with using six knots. But the the difference between using six knots and using three knots is no as big as with data generated according to the Exponential and Weibull models. All of the estimates have the biggest variability at time zero. The censoring rate effect is as the same as for data generated as in models (5.2) and (5.3). The biases of the estimated standard error are very small except with the Flexible Knots method.

Next we look at the plot of the 2.5 and 97.5 quantiles and the quartiles of the pointwise z values of the estimated log hazards from the True Model and Quantile Knots methods in Figure 5.17. We find that under 25% censoring the estimates look reliable for time is less than 70. Under 50% censoring the reliability of the estimates is not too bad as the time is less than 40, which match what we see from Figure 5.2.14. Different from in the Exponential model and Weibull model cases, for the B-spline model, the estimates with using six knots look more reliable than those with three knots. It does not surprise us since we expect using six knots would obtain better estimates than using three knots. To see if we can rely on the number of knots used by *heft* algorithm, we also check the numbers of knots used by *heft* algorithm. The average of the numbers of knots used by *heft* in the 500 simulations is 5.18, which is close to 6.

The histograms and the qq-plots of the estimated parameters $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$, in Figure 5.18, show that these estimates look normally distributed but not with mean zero.

We also test the hypothesis that the data follow a Weibull distribution. For the Quantile Knots estimation method, the test procedures are the same to those we used for the data from the Weibull distribution (5.3) in Section 5.2.2. For the True Model method, we test $\beta_1 = 0$. For the Flexible Knots method we test $\beta_1 = \cdots = \beta_{K-2} = 0$, where K is the number of the knots chosen by the *heft* for each simulation. The test statistics are the same as we discussed in Section 3.5.1. Figure 5.19 shows that, under 25% censoring, by all the four estimation methods the test procedures for that the failure times have a Weibull distribution have a high power except that by using six knots.

5.3 The Bivariate Case

In this section we generate the paired data (T_1, C_1, T_2, C_2). T_1 and T_2 are from (5.2), (5.3) or (5.4). We assume $C_1 = C_2$, and the censoring times are defined according to the distribution of T_1 from (5.2), (5.3) or (5.4). We are interested in investigating our estimated log hazards ratios and the estimates of their standard errors. We will also test for proportional hazards.

In Section 5.3.1 we introduce the method we use to generate dependent data for given marginal distributions. Then we present the simulation results in Sections 5.3.2 to 5.3.4.

5.3.1 Generation of Paired Dependent Data

Clayton (1978) proposed a family of bivariate distributions for survival times. Let S_1 and S_2 denote the marginal survivor functions for each member of a pair of failure times

(T_1, T_2). The joint survivor function for the Clayton model with parameter θ is

$$S(t_1, t_2, \theta) = \begin{cases} \left[S_1(t_1)^{1-\theta} + S_2(t_2)^{1-\theta} - 1\right]^{-\frac{1}{\theta-1}}, & t_1, t_2 > 0,\ \theta > 1 \\ S_1(t_1)S_2(t_2), & t_1, t_2 > 0.\ \theta = 1 \end{cases} \tag{5.6}$$

When $\theta = 1$, the failure times are independent.

The parameter θ can be written in terms of two conditional hazard functions. If we denote the hazard for the conditional distribution of T_1 given $T_2 = t_2$ and the hazard for T_1 given $T_2 > t_2$ by $h_{T_1|T_2=t_2}(t_1)$ and $h_{T_1|T_2>t_2}(t_1)$, respectively, then in this model $h_{T_1|T_2=t_2}(t_1)/h_{T_1|T_2>t_2}(t_1) = \theta$. See Clayton (1978).

For given S_1, S_2 and θ we generate bivariate data (T_1, T_2) satisfying the Clayton model using the following:

(A) Generate $T_1 = t_1$ according to S_1;

(B) Generate $T_2 = t_2$ according to the conditional distribution of T_2 given $T_1 = t_1$.

(A) can be carried out as in the univariate case. For (B), when $\theta > 1$, the conditional probability is (abusing notation slightly),

$$\begin{aligned} P(T_1 > t_2 | T_1 = t_1) &= \frac{P(T_2 > t_2 \text{ and } T_1 = t_1)}{P(T_1 = t_1)} = \frac{\partial S(t_1, t_2, \theta)/\partial t_1}{S_1'(t_1)} \\ &= \frac{-\frac{1}{\theta-1}\left[S_1(t_1)^{1-\theta} + S_2(t_2)^{1-\theta} - 1\right]^{-\frac{1}{\theta-1}-1}(1-\theta)S_1(t_1)^{-\theta}S_1'(t_1)}{S_1'(t_1)} \\ &= \left[S_1(t_1)^{1-\theta} + S_2(t_2)^{1-\theta} - 1\right]^{\frac{\theta}{1-\theta}} S_1(t_1)^{-\theta} \\ &= 1 - F_{T_2|T_1=t_1}(t_2). \end{aligned}$$

We will generate $U = u$, a uniform $[0, 1]$, and solve $u = F_{T_2|T_1=t_1}(t_2)$ for t_2:

$$u = 1 - P(T_1 > t_2 | T_1 = t_1)$$

$$= \; 1 - \left[S_1(t_1)^{1-\theta} + S_2(t_2)^{1-\theta} - 1\right]^{\frac{\theta}{1-\theta}} S_1(t_1)^{-\theta}.$$

Solving for $S_2(t_2)$, we have

$$S_2(t_2) = \left\{[(1-u)^{\frac{1-\theta}{\theta}} - 1]S_1(t_1)^{1-\theta} + 1\right\}^{\frac{1}{1-\theta}}. \tag{5.7}$$

Now solve Equation 5.7 for t_2. Let p_2 denote 1 minus the right part of Equation 5.7 equal p_2. Then t_2 is the p_2th quantile of the distribution function $1 - S_2$. If S_2 has a standard distribution say mean one exponential, then t_2 =*qexp* (p_2) in Splus. For nonstandard distributions, solve $S_2(t_2) = 1 - p_2$ for t_2 numerically.

Thus, for given marginal survivor functions S_1 and S_2 and the parameter θ, we can generate a random bivariate variable $(T_1, T_2) = (t_1, t_2)$ satisfying the Clayton model by the following procedure.

1. Generate a random variable $T_1 = t_1$ according to S_1;

2. Calculate $S_1(t_1)$;

3. Generate a random variable $U = u \sim U(0, 1)$;

4. Calculate

$$p_2 = 1 - \left\{[(1-u)^{\frac{1-\theta}{\theta}} - 1]S_1(t_1)^{1-\theta} + 1\right\}^{\frac{1}{1-\theta}};$$

5. Solve $S_2(t_2) = 1 - p_2$ for t_2.

Our choices of marginal distributions S_1 and S_2 follow the proportional hazards and non-proportional hazards model as defined in Section 5.1. Figure 5.20 presents the plots of the correlation coefficients of T_1 and T_2 vs θ, where T_k has survivor function S_k,

$k = 1, 2$ In our simulation studies, we consider these models with $\theta = 1, \theta = 1.5$, and 2.5. When $\theta = 1$, T_1 and T_2 are independent.

5.3.2 Proportional Hazards Model

We present our simulation results for the proportional hazards model in this section. We address the effect of the correlation level and the comparisons of the Cox, Same Knots, and Different Knots estimates of the log hazards ratio. Recall that the Cox estimate assumes that $\theta = 1$. Since we had studied the effect of the censoring rate on the estimates of log hazards in the univariate case, we only present the results for 25% censoring here and give the results for 50% censoring in the Appendix.

Figure 5.21 shows the histograms of the two marginals and the non-censored failure times. We can see that there are very few non-censored observations larger than 150. For the same reason as in the univariate case, we only plot results in the the Proportional Hazards model up to time 150.

First we look at the plots of the pointwise quartiles and average of the 500 estimated log hazards ratios in Figure 5.22. Comparing them with the true log ratio, we find that the estimates from the Cox method have the smallest bias, as we expected. The plots show that at time values within the range of knots, the biases of the estimates from the Same Knots method and the Different Knots method are almost the same except that the estimates from the Different Knots method have bigger bias at time zero. It looks like the correlation level does not affect the Same Knots and Different Knots estimates too much. But it is interesting to see that the estimates from the Cox method have the smallest bias when $\theta = 1.5$.

Then we look at the plots of the pointwise standard deviations of the 500 estimated log hazards ratios in Figure 5.23. We see that the estimates of the log hazards ratio from the Cox method have the smallest variability. There is no big difference between the variabilities of the estimated log hazards ratio from the other two methods. We compare Figure 5.23 with Figure 5.16, which shows the variability of the marginal log hazards. It seems that near time 0 the variabilities of the estimated log hazards ratio with the Same Knots and Different Knots methods are similar to the variabilities of the estimated marginal log hazards.

When we compare the estimated standard errors with the pointwise standard deviations in Figure 5.23, we find that the estimates of the standard error from the Cox methods have the smallest bias when θ is 1 and 1.5, that is when the paired data are independent or only slightly correlated. But when θ is 2.5, the estimated standard errors from the other two methods have smaller bias than the Cox method. The reason for the increasing bias of the estimated standard errors from the Cox method might be the violation of the assumption of independence of the two failure times. Over all three correlation levels, the biases of the estimated standard errors from the Same Knots method and the Different Knots method are very close to each other.

When we look at the plots of the quantiles of the pointwise z values of the estimated log hazards ratios, Figure 5.24, we note that the estimated log hazards ratios from the Cox method are more reliable than the estimates from the other two methods, and the best case for the Cox method is when $\theta = 1.5$. For the Same Knots and Different Knots methods, the reliability of the estimates of the log hazards ratio is fairly good when time t is between 5 and 70, which is about the range of the knots. Once more the reliability of the estimates from the Same Knots and Different Knots methods does not depend on the correlation levels, which does not surprise us because our estimated standard errors take the correlation into account.

Finally, we test the hypothesis that the failure times have proportional hazards by first estimating the two log hazards using the Same Knots method. We then test $\beta_{1j} - \beta_{2j} = 0$, $j = 1, 2, 3, 4, 6$, as indicated in Section 3.5.2. Using the χ^2 test statistics discussed in Section 3.5.2, we can calculate p-values for the test.

We look at Figure 5.25, the plots of the empirical distributions of the p-value for testing that the failure times have proportional hazards. The null hypothesis is rejected far more than we expected. We expected each plot would be a straight line since the null hypothesis is true and the p-value should be uniformly distributed. An interesting result is that the distribution curve is closest to the line $y = x$ when the data are highly correlated ($\theta = 2.5$).

Through the simulation results, we would like to say that for the data as in the proportional hazards model, the Cox method works well. The performance of the Same Knots and Different Knots methods is reasonable. The estimated standard errors from all three methods are good but those from the Cox model deteriorate as θ increases.

5.3.3 Non-proportional Hazards Model

Figure 5.26 shows the histograms of the two marginals and the non-censored failure times for the data from the non-proportional hazards model. As with the proportional hazard model, we only plot results in the the non-proportional hazards model up to time 150.

First we look at Figure 5.27, the plots of the true log ratio and the pointwise quantiles and average of the 500 estimated log hazards ratios. As we expected, the Cox method does not work. The estimates from the other two methods are very close. These estimates have smallest bias when the data are independent. But it is hard to see any difference between $\theta = 1.5$ and $\theta = 2.5$.

Next we look at Figure 5.28, the plot of the pointwise standard deviations of the 500 estimated log hazards ratios and the pointwise quartiles and mean of the 500 estimated standard errors. Since the Cox method does not work for estimating log hazards ratios for the non-proportional hazards data, we only discuss the results from the other two methods here. We find that the pointwise standard deviations from the Same Knots method are a little bit bigger than those from the Different Knots method. But the standard deviation from the Different Knots method has a bigger jump at $t = 0$. When we look at the variability of the estimated standard errors, we find that the bias difference between the Same Knots method and the Different Knots method is very small within the range of knots. The standard errors from the Same Knots and Different Knots methods are too small for $t \geq 70$. The Different Knots method performs slightly better in this range. The bias and the variability of the estimated standard errors do not depend on the correlation level of the paired data.

We also check the plots of the quantiles of the pointwise z values of the estimated log hazards ratios from the three methods, Figure 5.29. Those plots show that the quantile curves for the Same Knots and Different Knots methods mainly lie in the range $(-2, 2)$ with the best performance when the data are independent. It is not surprising that the Cox method performs poorly since the failure times do not have proportional hazards.

Now we consider testing the hypothesis that the failure times have proportional hazards. The test procedures are the same as those we used for the data sets from the proportional model in Section 5.3.2 When we look at the plot of the empirical distributions of the p-values, Figure 5.30, we find that the test power is not high but the lowest power is when $\theta = 1$.

5.3.4 Effect of the Number of Knots

In this section we study how the number of knots used in each simulation affects the estimates of the log hazards ratios and the standard errors. The data sets are generated as in the proportional hazards model and the non-proportional hazards model. For the data set from proportional hazards model, we use three, six, and nine knots to estimate the marginal hazards of T_1 and T_2. For the data generated from the non-proportional hazards model, we use three, six, and nine knots for failure time T_1 but only three knots for T_2. For the data from proportional hazards model, the true hazard of T_2 has three knots, see (5.5). In contrast, for the data from non-proportional hazards model, T_2 is exponentially distributed. As we found in the univariate simulation (Section 5.2), *heft* will rarely choose more than three knots to fit this exponential. Therefore, we do not need to study fits with more than three knots.

The simulation results are as we expected, that is, the more knots, the smaller the bias of the estimates but the bigger the variability of the estimates. We show the results in Figures 5.31 to 5.37. From those graphs we find that the estimates of the log hazard ratio by using three knots have the smallest variability but the biggest bias. There is no big difference in bias between using six knots and using nine knots but the estimates using nine knots have the biggest variability.

When we look at the estimated standard errors, we find that the estimates using six knots have the smallest bias and variability.

Figures 5.39 to 5.42 show the plots of the quantiles of the pointwise z values of the estimated log hazards ratios from the three methods. It can be seen that the performance of the estimates using six knots is similar to those with nine knots and superior to those using three knots.

Figures 5.43 and 5.44 present the p-values to test that the two failure times have proportional hazards for the data as in the proportional model and non-proportional model, respectively. We find that, for both data sets, the number of knots does not effect the test too much but the test performs better when T_1 and T_2 are dependent than when T_1 and T_2 are independent.

Overall, for the data generated as in the proportional hazards and non-proportional hazards models, using six knots seems better than using three knots or using nine knots. Of course, the value of six may also depend on the sample size and the true data distribution.

5.4 Summary of Simulations

Through this simulation study, we find the following.

In the univariate case:

1. With all the estimation methods, the estimates of the log hazard perform well within the range of knots used. However, if the range of knots is large, the Flexible Knots method might cause some numerical calculation problems in estimating the log hazards and in calculating standard errors.

2. With the Quantile Knots method, the estimates of standard errors perform well except that they slightly underestimate the variability of the estimated log hazards when the censoring rate is high.

3. The censoring rate affects the bias and variability of the estimated log hazards. The smaller the censoring rate, the better the estimates.

4. For all marginal models, the True Model method gives the best estimates for the log hazards.

5. The estimates using three knots look a little bit better than those using six knots for the data generated as in the exponential distribution (5.2) and the Weibull

distribution (5.3). But for the data generated as in the B-spline model (5.4), the estimates with six knots look more reliable.

6. The Flexible Knots method gives better estimated log hazards than other methods when the ratio of the largest knot to the smallest knot is not too big. Otherwise calculation of the estimated standard errors and the estimated log hazards causes numerical problems.

7. The test procedures for testing that the failure times follow an exponential and Weibull distribution perform well.

In the bivariate case:

1. With the Same Knots and Different Knots estimation methods, the estimates of the log hazards ratio perform well within the range of knots used. There is not a big difference between the estimates from the two estimation methods.

2. For the data with proportional hazards, the Cox method gives the least-biased estimates of the log hazards ratios when the failure times are independent. For the data from the non-proportional hazards model, the Cox proportional model does not work at all.

3. The test that the failure times have proportional hazards does not perform well. We expect a bigger sample size might improve the test procedure.

4. The correlation levels do not affect the estimated log hazards ratios too much. But they affect the test that the failure times have proportional hazards. The

test has the lowest power when the two failure times are independent. Further simulations are required to better understand the effects of correlation.

In this simulation we also check the test procedures for the hypothesis that the two failure times have the same distribution. But there are some numerical problems when we calculate p-values. We do not present the result here.

One more thing we would like to point out is that our estimated formulae for the standard errors assume for non-random knots. The procedures used in our simulations have random knots except for the True Model method in the univariate case. So this might be a reason that the estimated standard errors are a bit small.

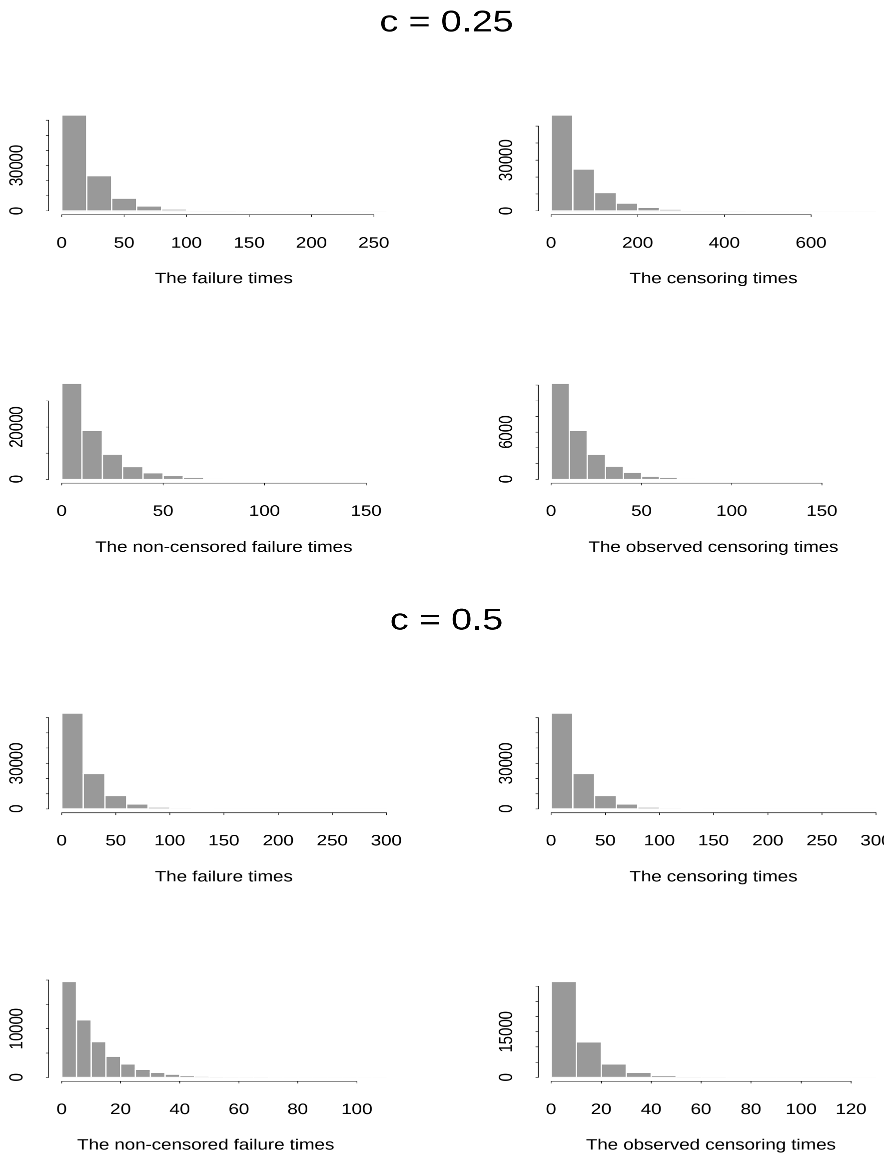

Figure 5.1: The histograms of the simulated failure times, censoring times, non-censored failure times, and observed censoring times for exponential data as in model (5.2).

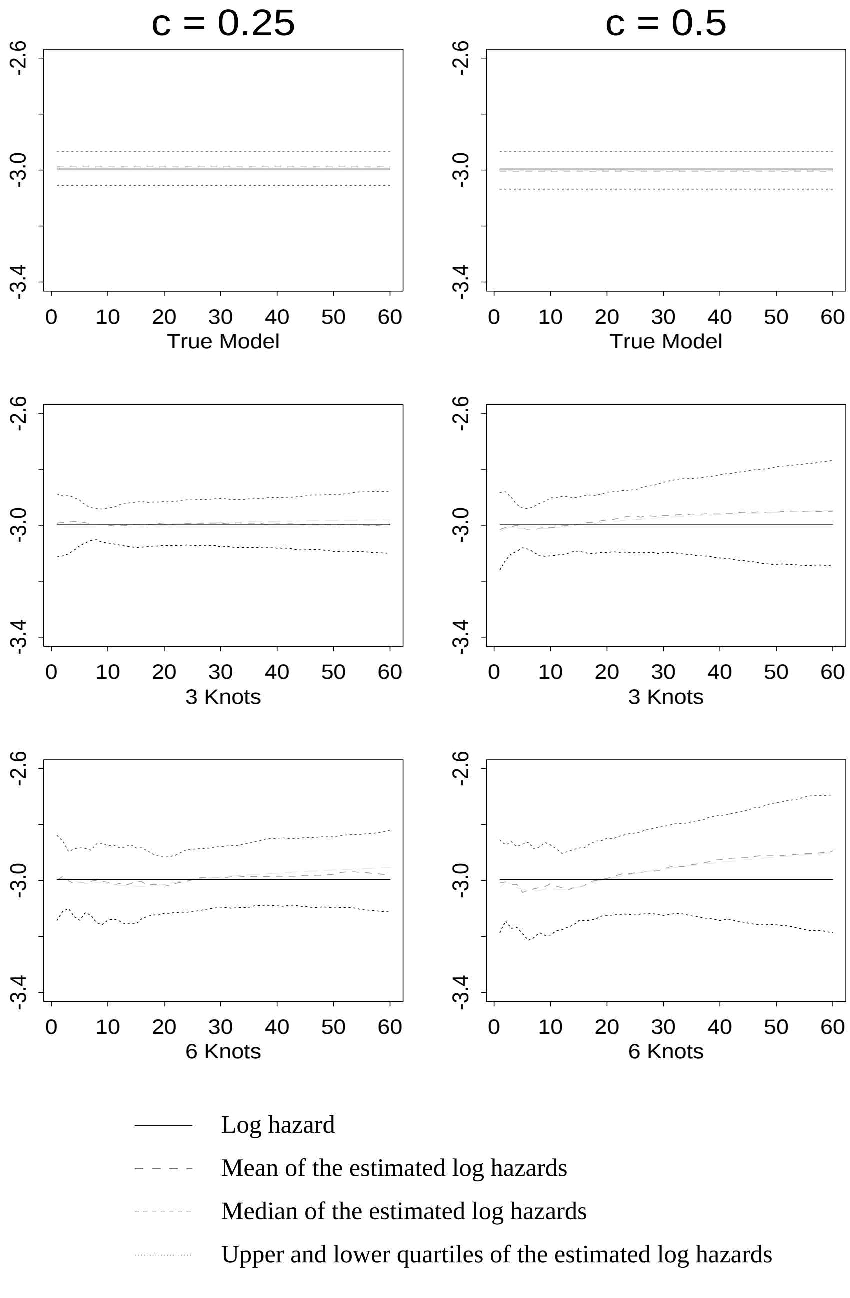

Figure 5.2: Log hazard of the exponential distribution (5.2) and the pointwise quartiles and empirical mean of the 500 estimated log hazards.

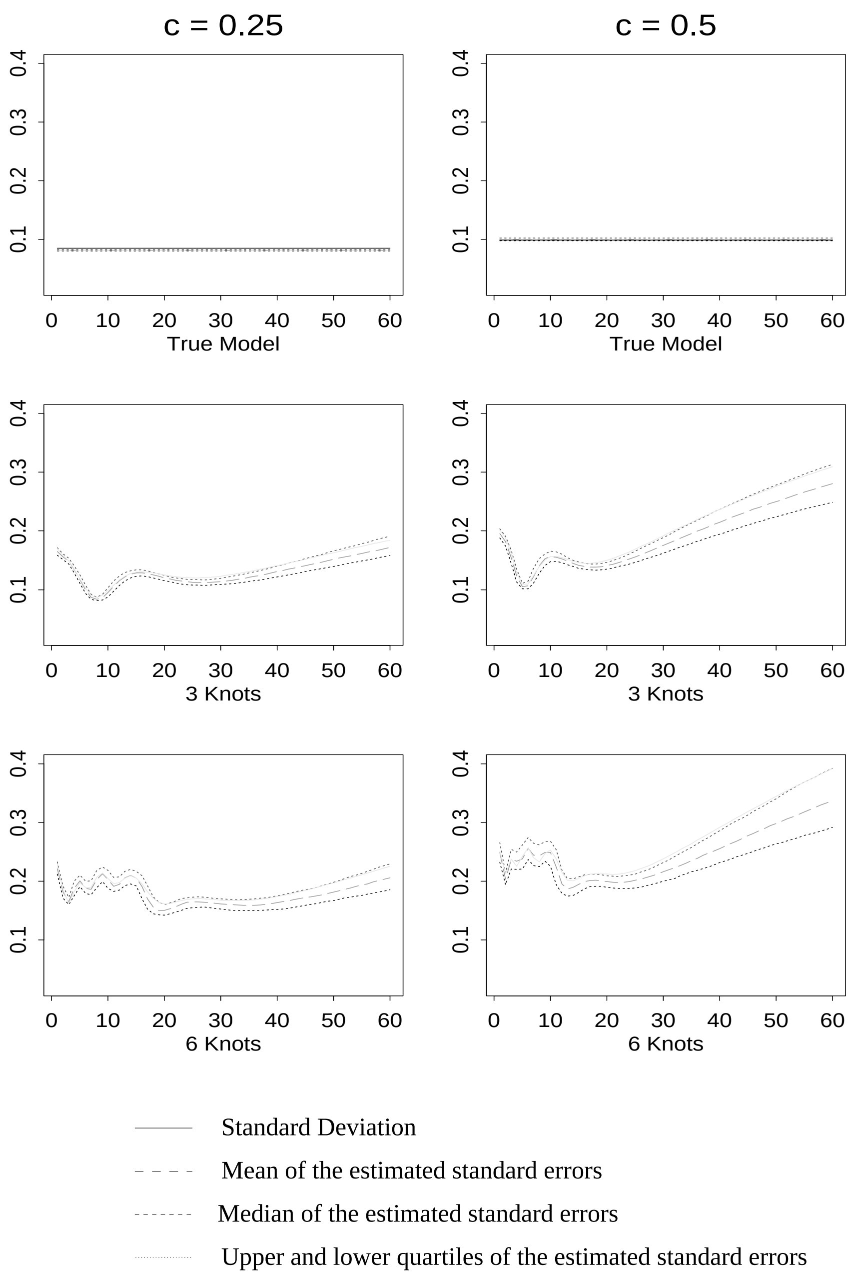

Figure 5.3: The pointwise standard deviations of the estimated log hazard for the exponential model (5.2) and the pointwise quartiles and empirical mean of the 500 estimated standard errors.

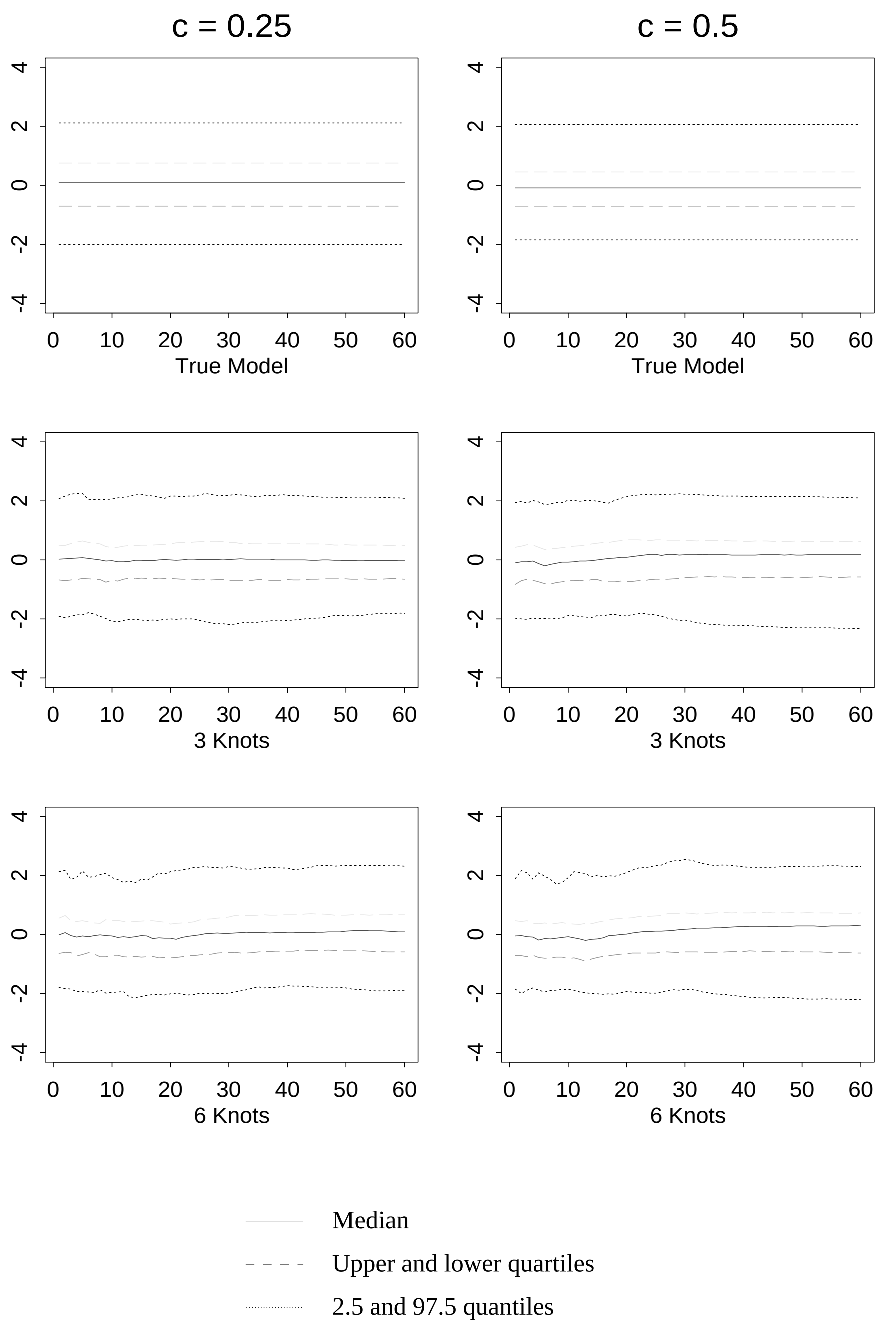

Figure 5.4: The quantiles of the pointwise z values of the estimated log hazards for the exponential distribution (5.2).

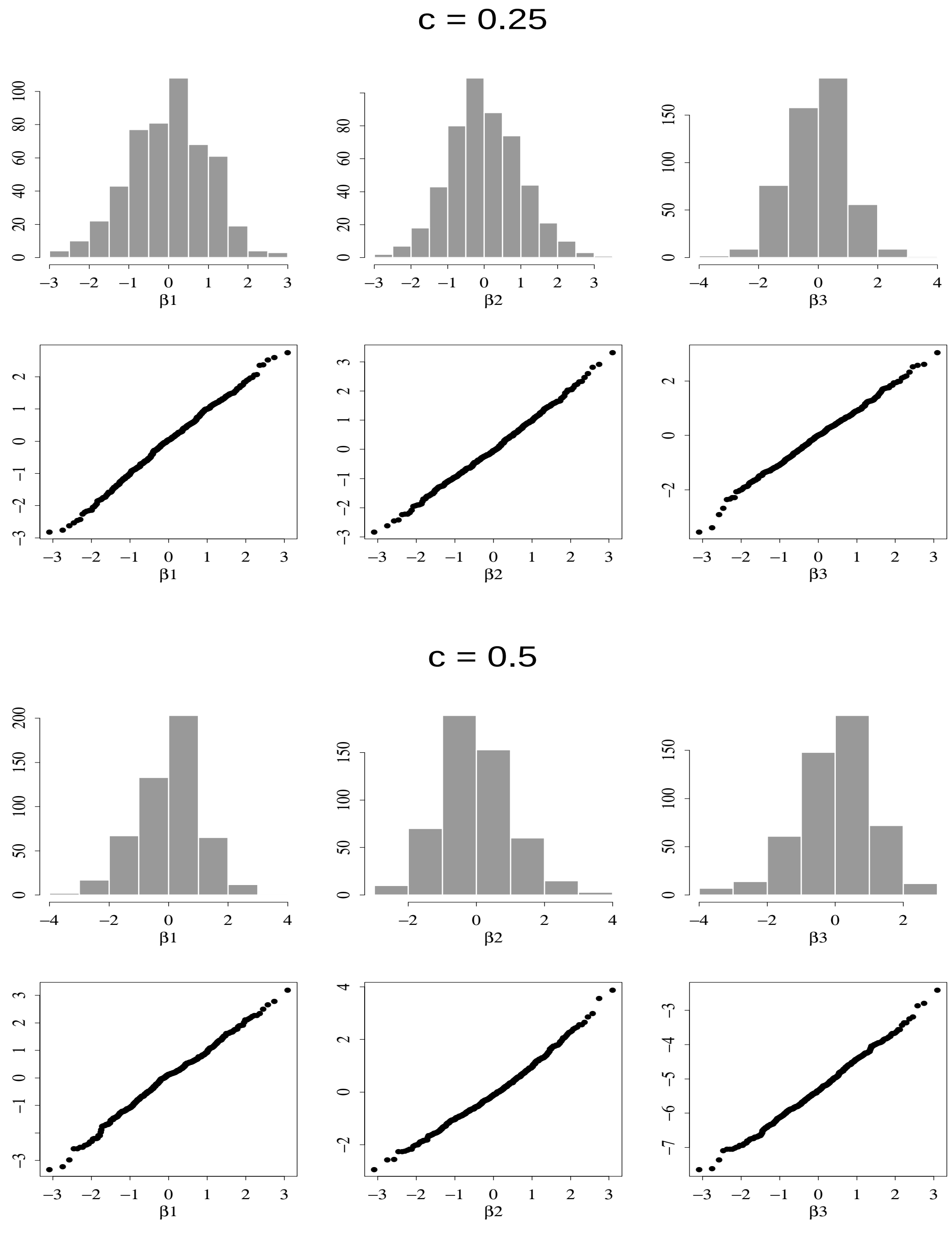

Figure 5.5: Histograms and qq-plots of the normalized estimate of $\boldsymbol{\beta}$ from the Quantile Knots methods with three knots for the exponential model (5.2).

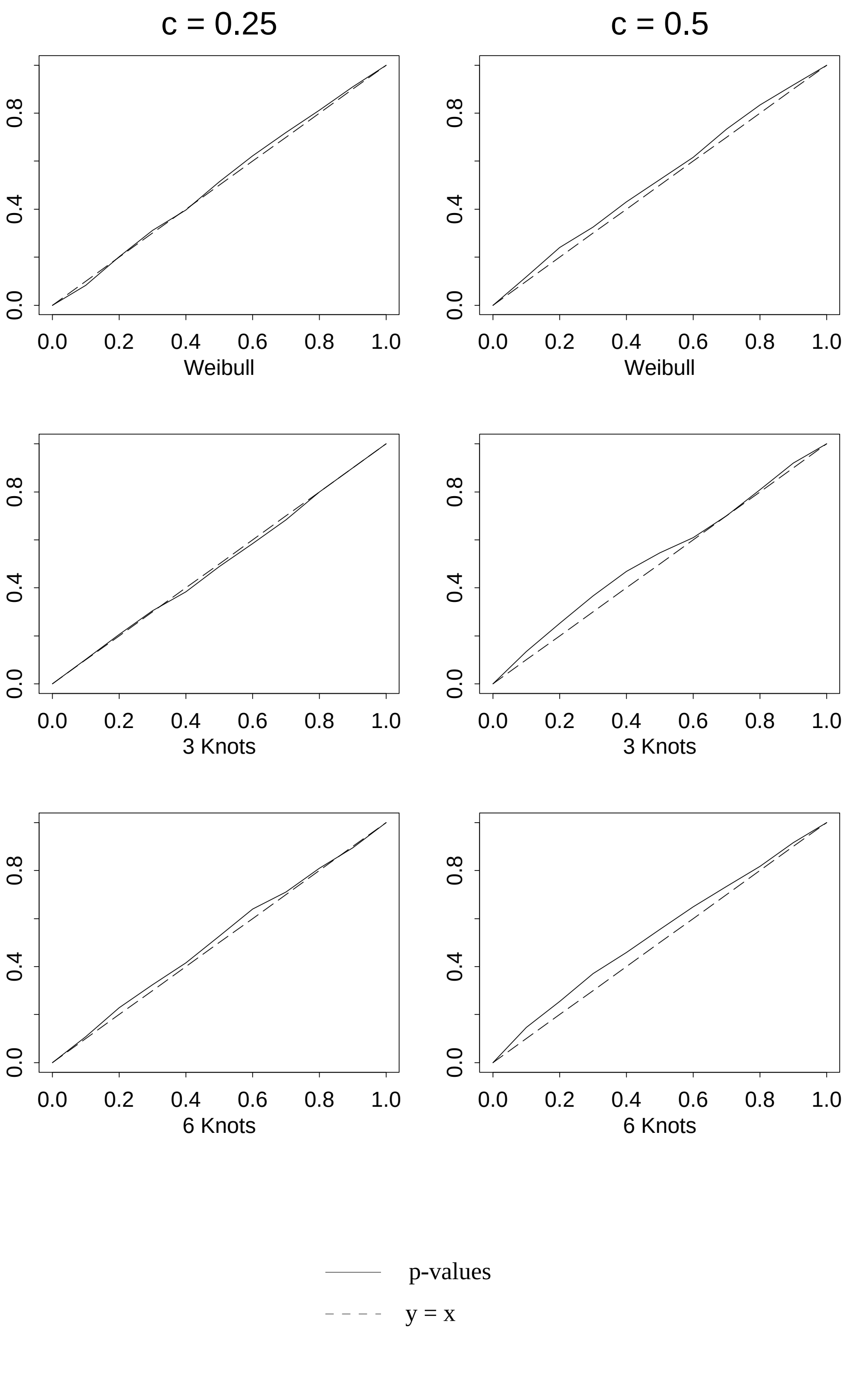

Figure 5.6: Empirical distribution functions of the p-values for testing that the failure times are exponentially distributed.

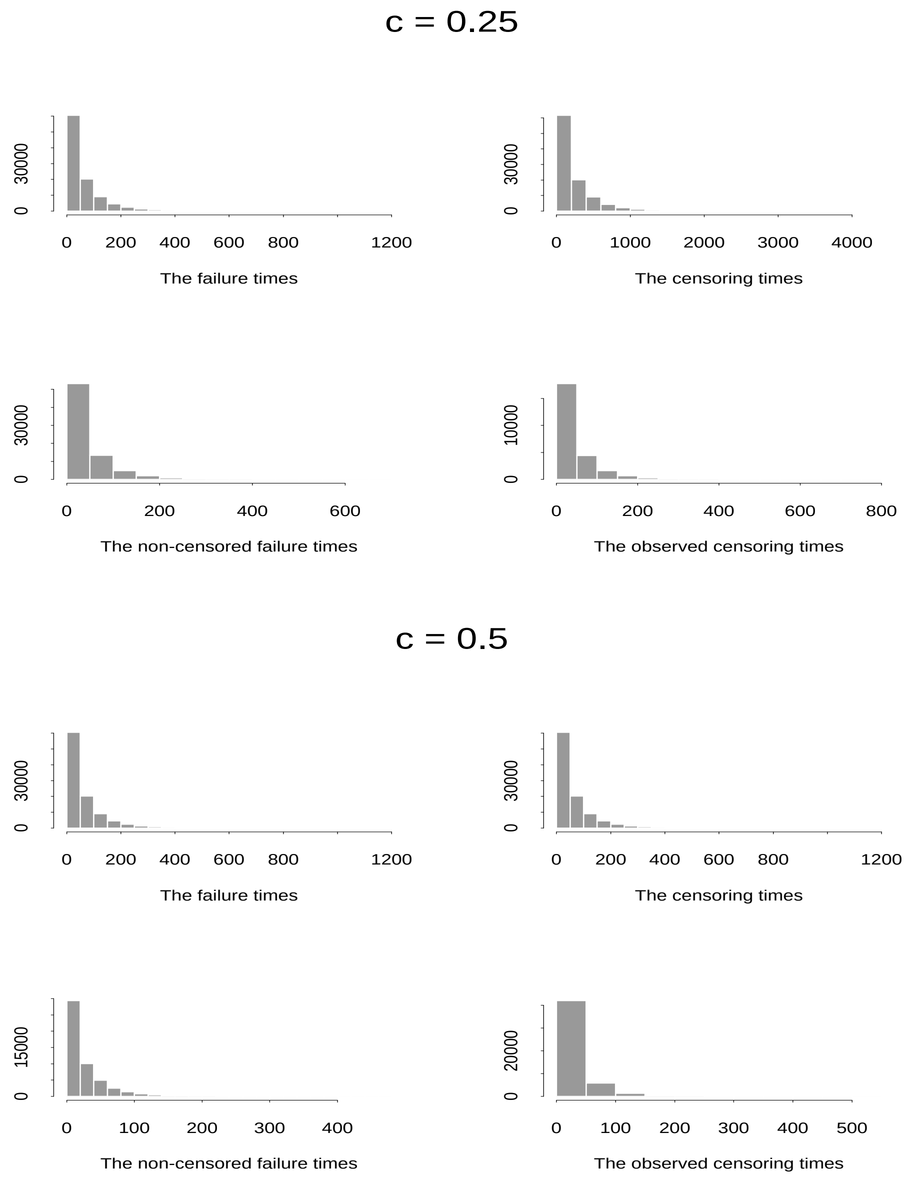

Figure 5.7: The histograms of the simulated failure times, censoring times, non-censored failure times, and observed censoring times for Weibull data as in model (5.3).

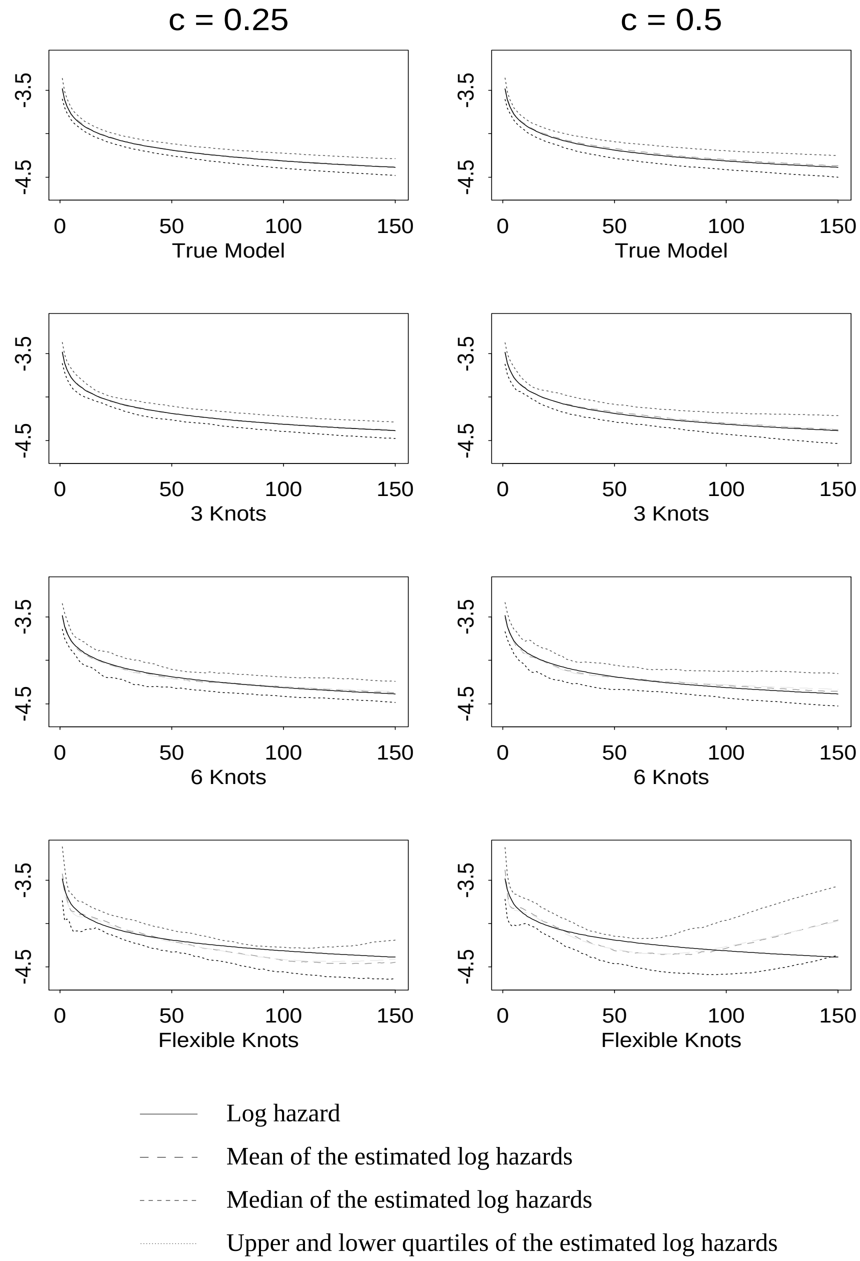

Figure 5.8: Log hazard of the Weibull distribution (5.3) and the pointwise quartiles and empirical mean of the estimated log hazards.

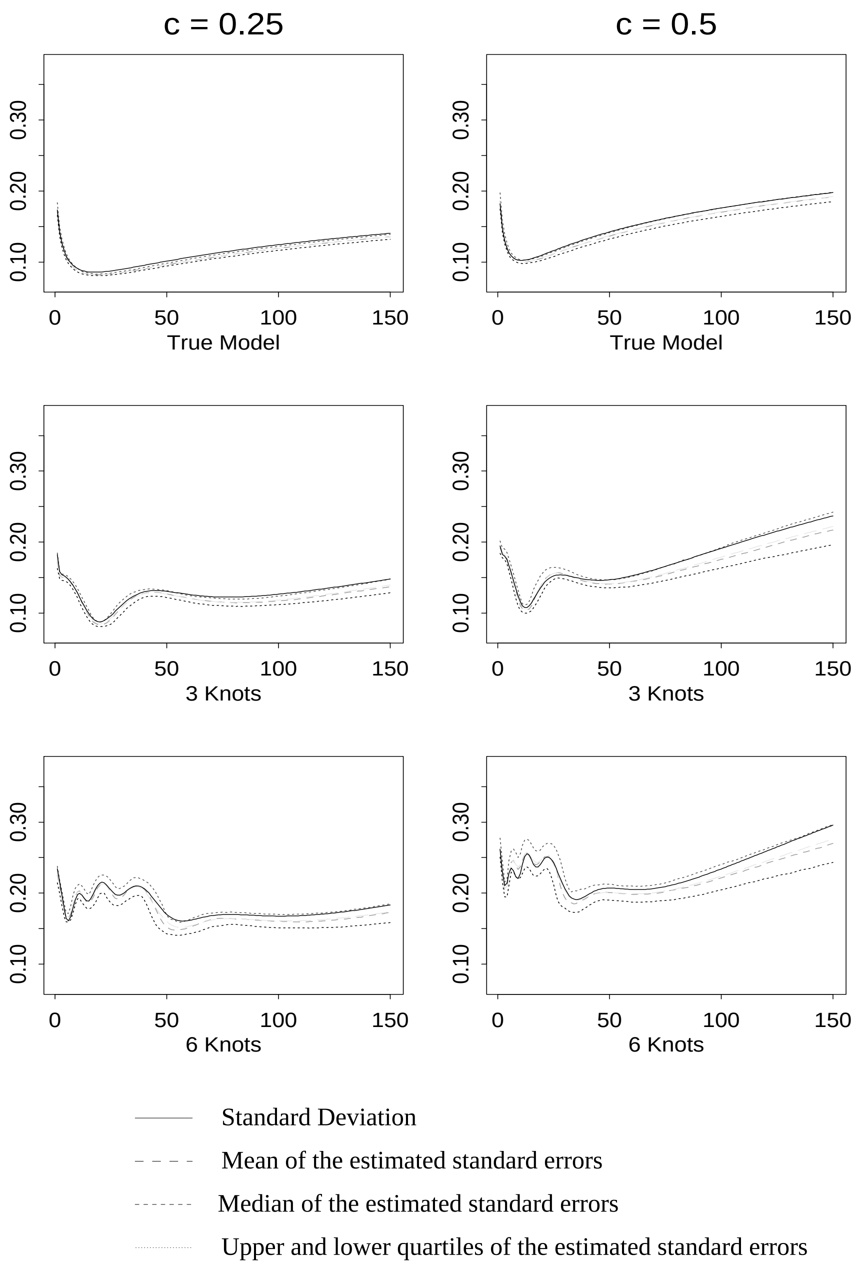

Figure 5.9: The pointwise standard deviations of the estimated log hazards for the Weibull distribution (5.3) and the pointwise quartiles and empirical mean of the 500 estimated standard errors.

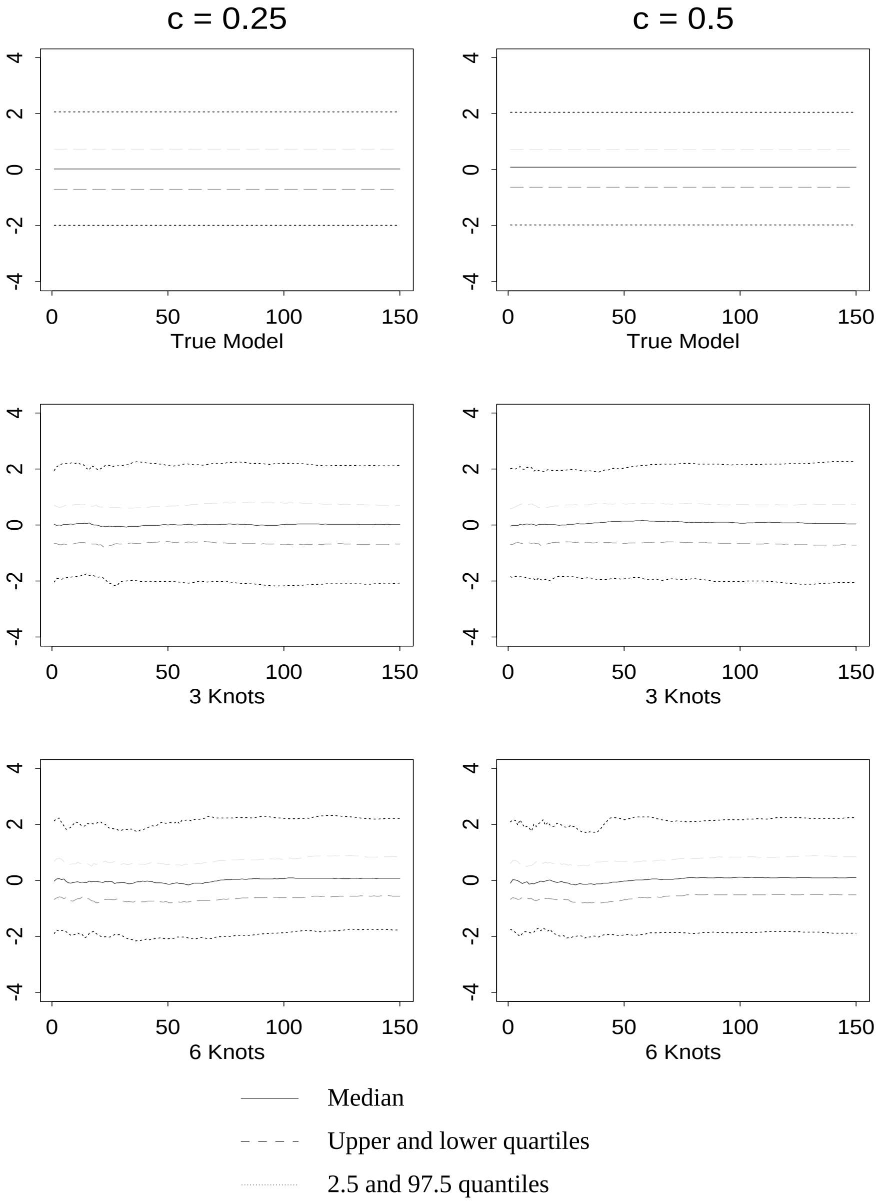

Figure 5.10: The quantiles of the pointwise z values of the estimated log hazards for the Weibull model (5.3).

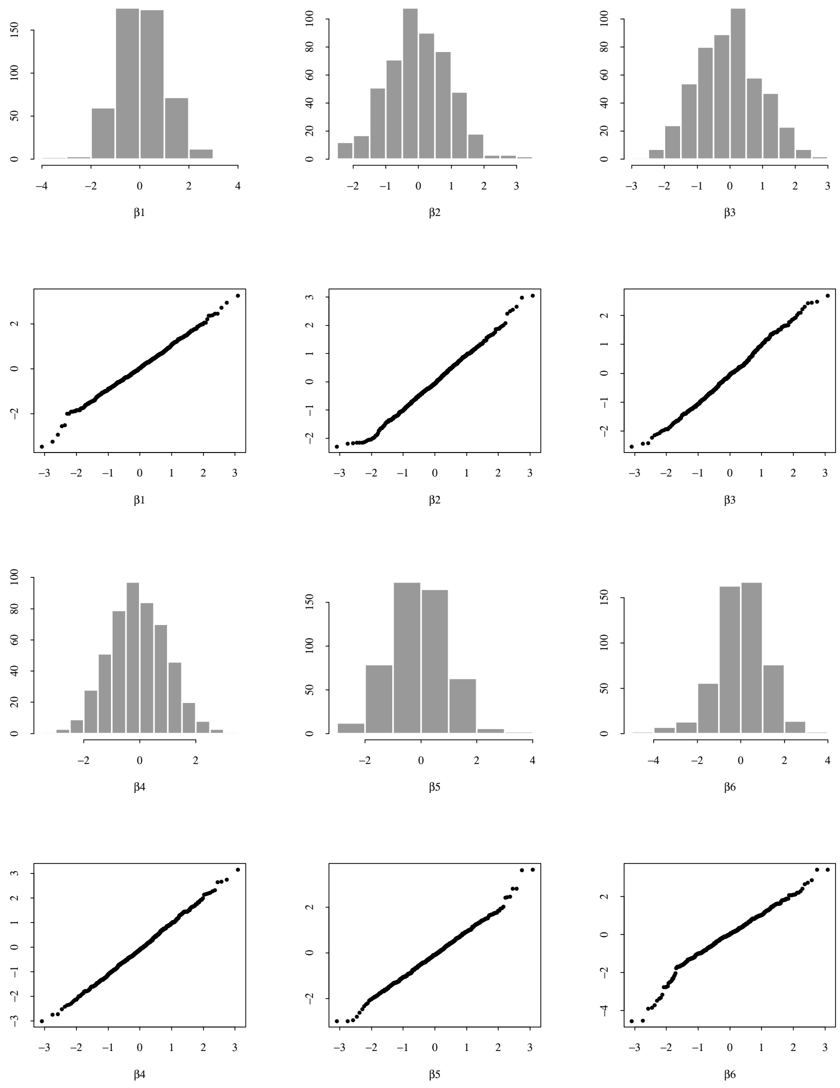

Figure 5.11: The histograms and the qq-plots of the standardized estimates of the parameter $\boldsymbol{\beta}$ from the method 5.1 with six knots for the Weibull model (5.3).

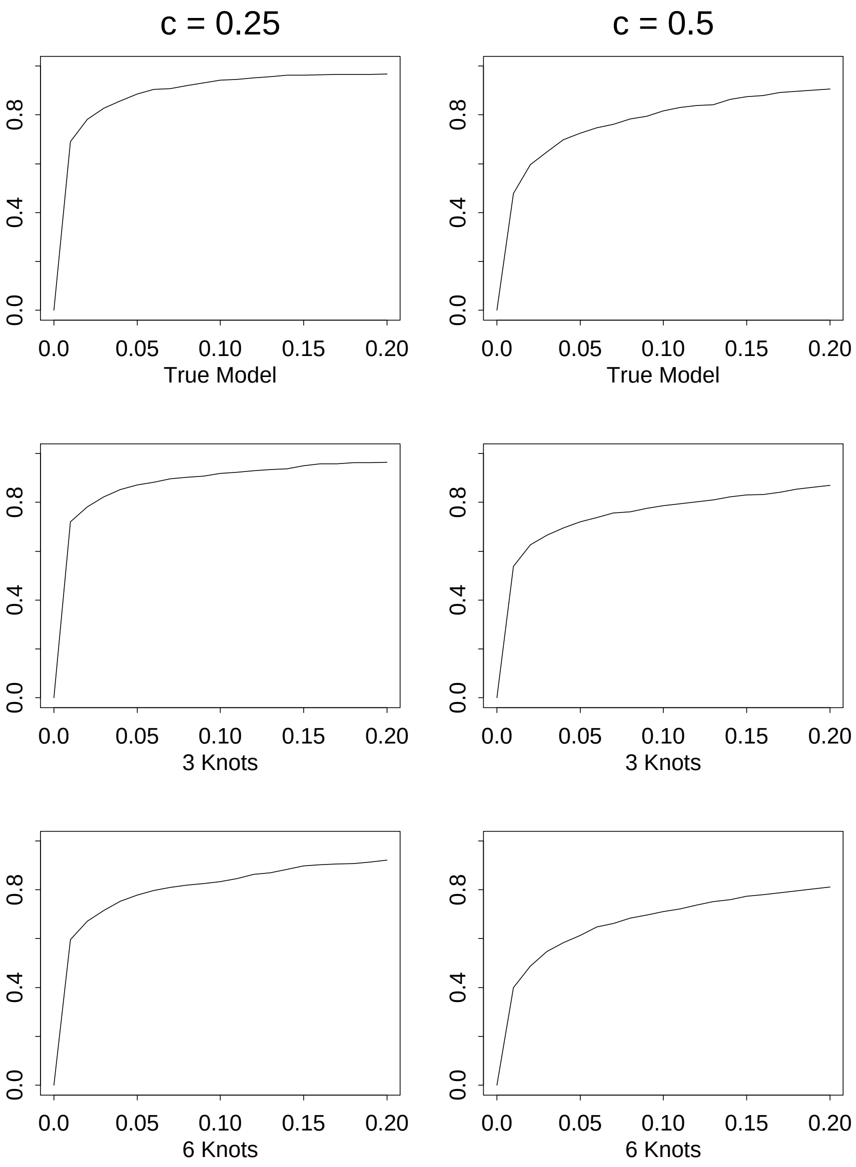

Figure 5.12: Empirical distribution functions of the p-values for testing the hypothesis that the failure time distribution is exponential.

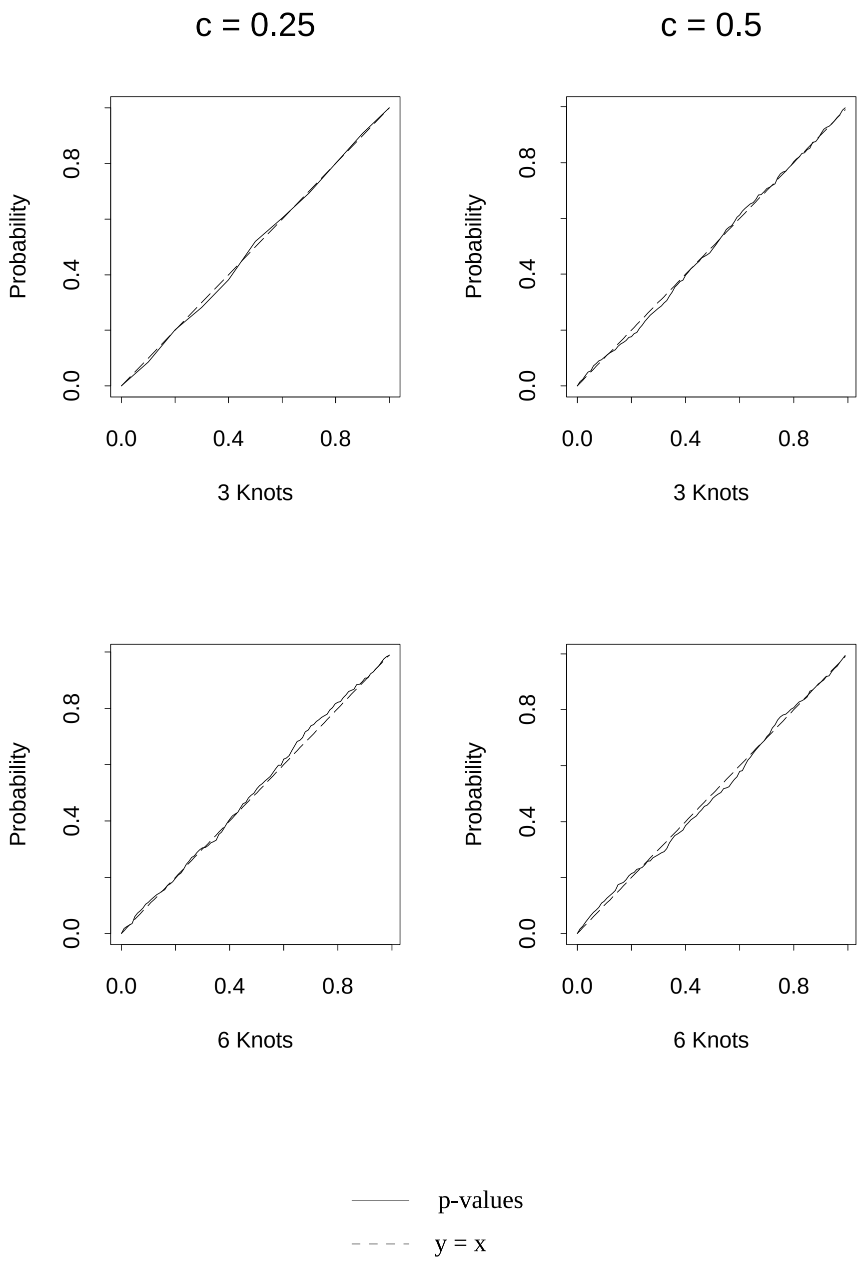

Figure 5.13: Empirical distribution functions of the p-values for testing that the failure times have a Weibull distribution.

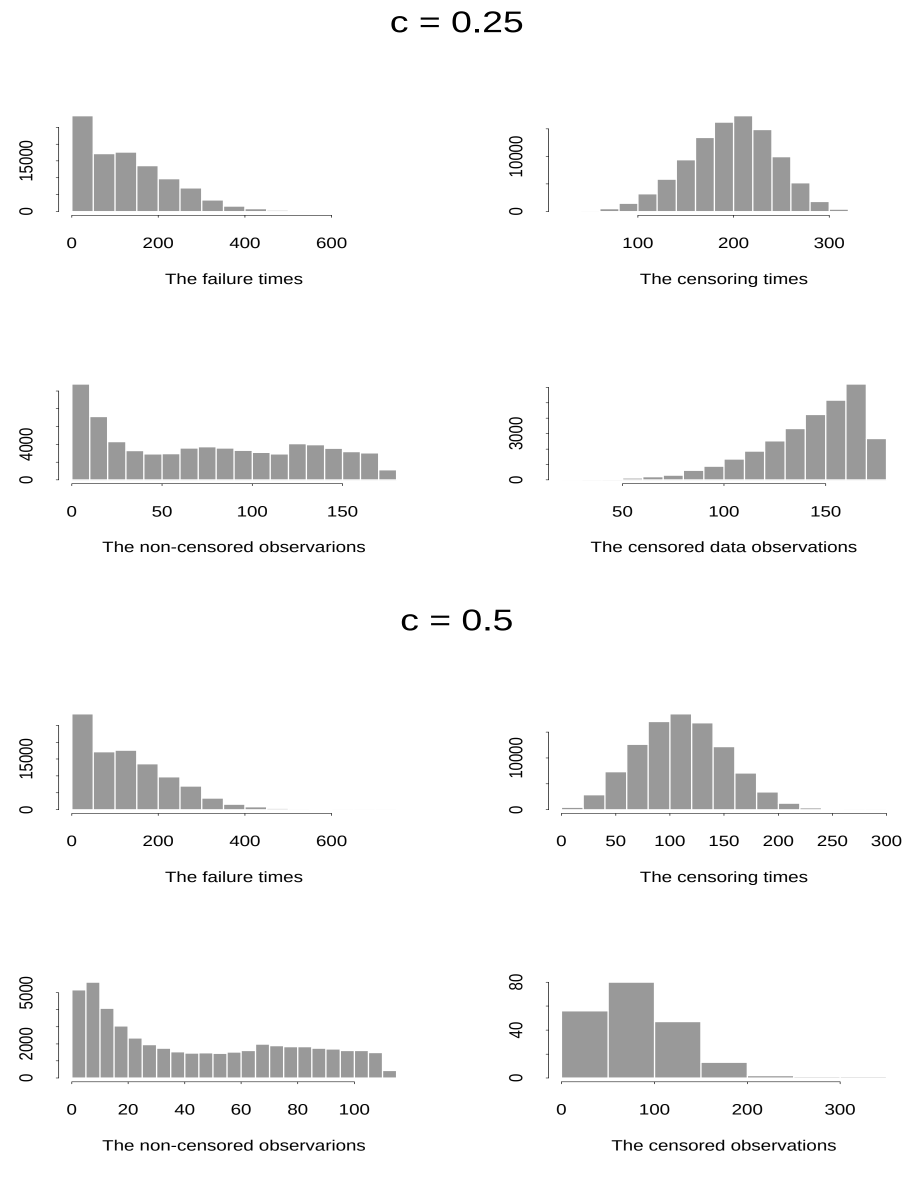

Figure 5.14: The histograms of the simulated failure times, censoring times, non-censored failure times, and observed censoring times for the data as in the B-spline model (5.4).

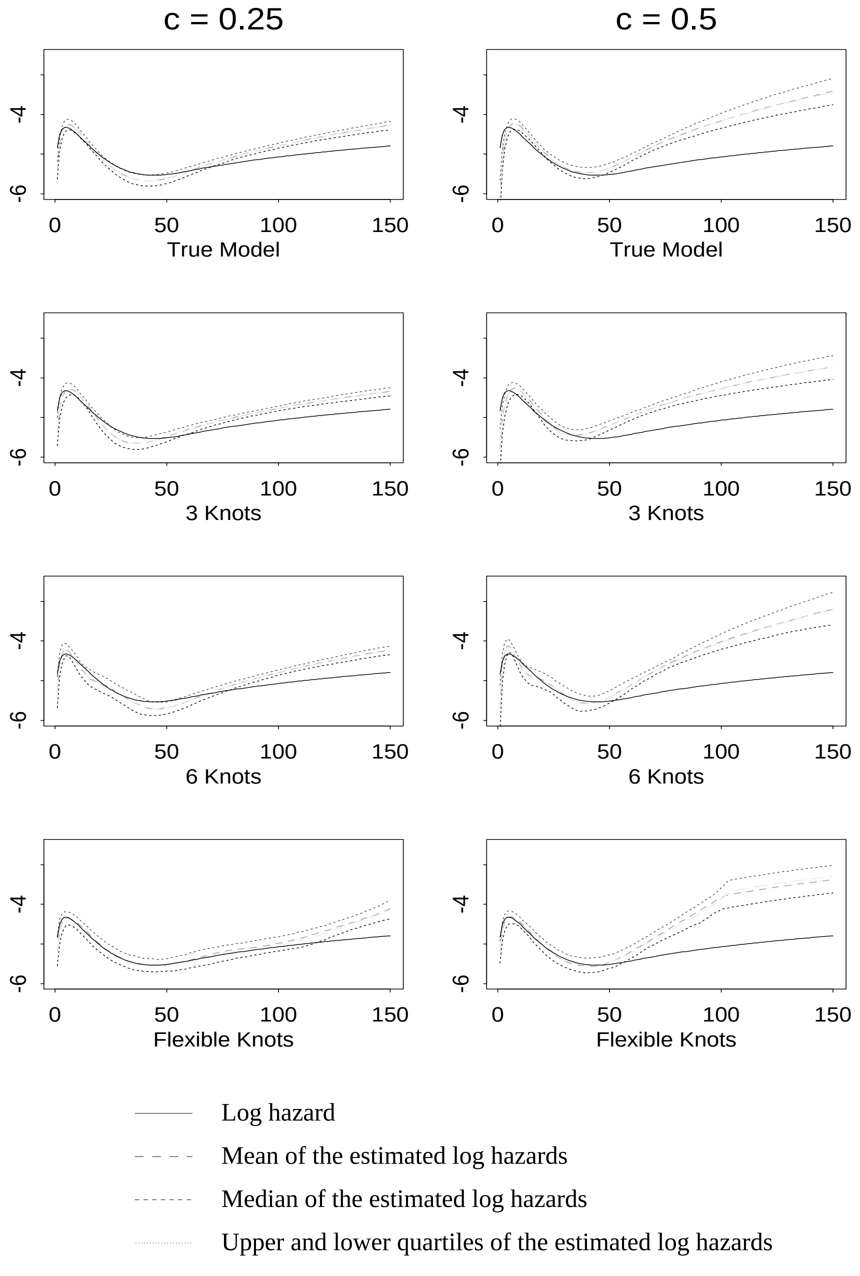

Figure 5.15: Log hazard of the the B-spline data as in model (5.4) and the pointwise quartiles and empirical mean of the estimated log hazards.

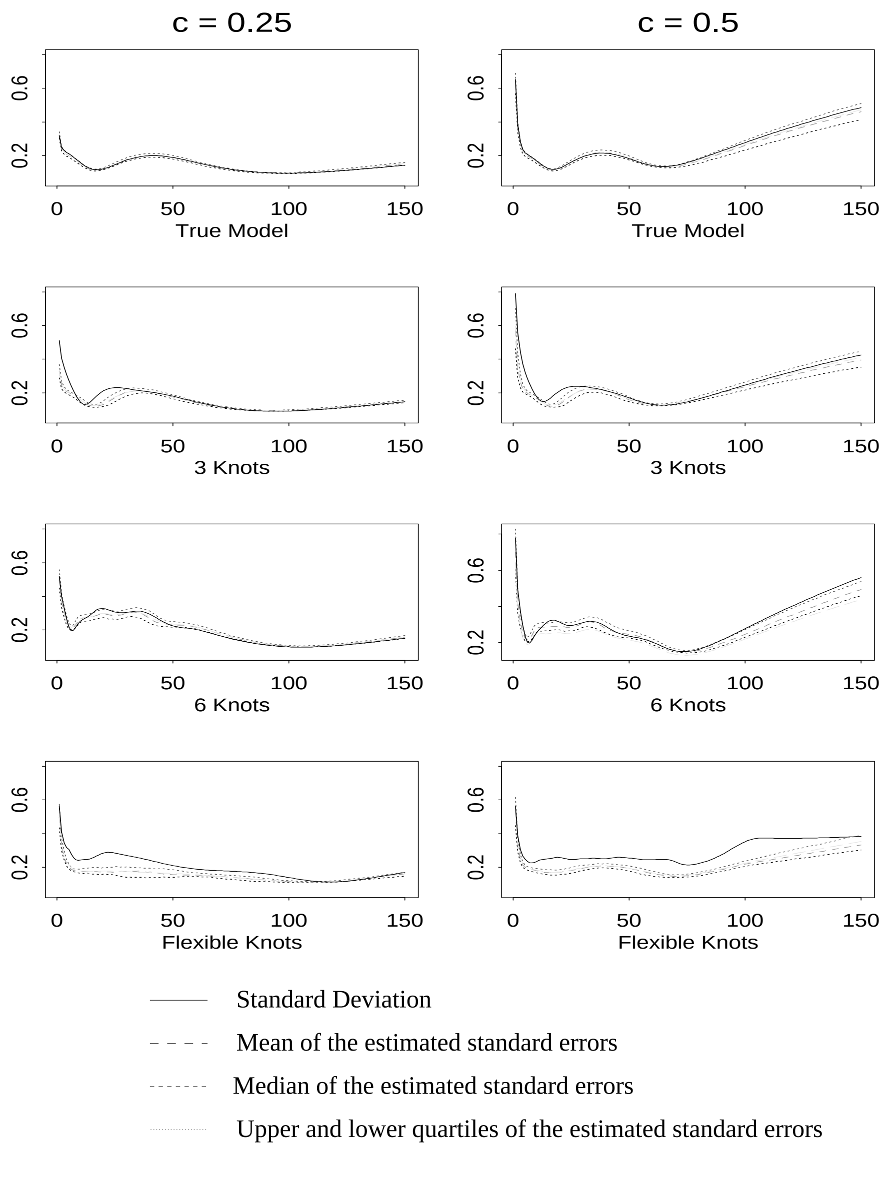

Figure 5.16: The pointwise standard deviations of the estimated log hazards for the B-spline model (5.4) and the pointwise quartiles and empirical mean of the estimated standard errors.

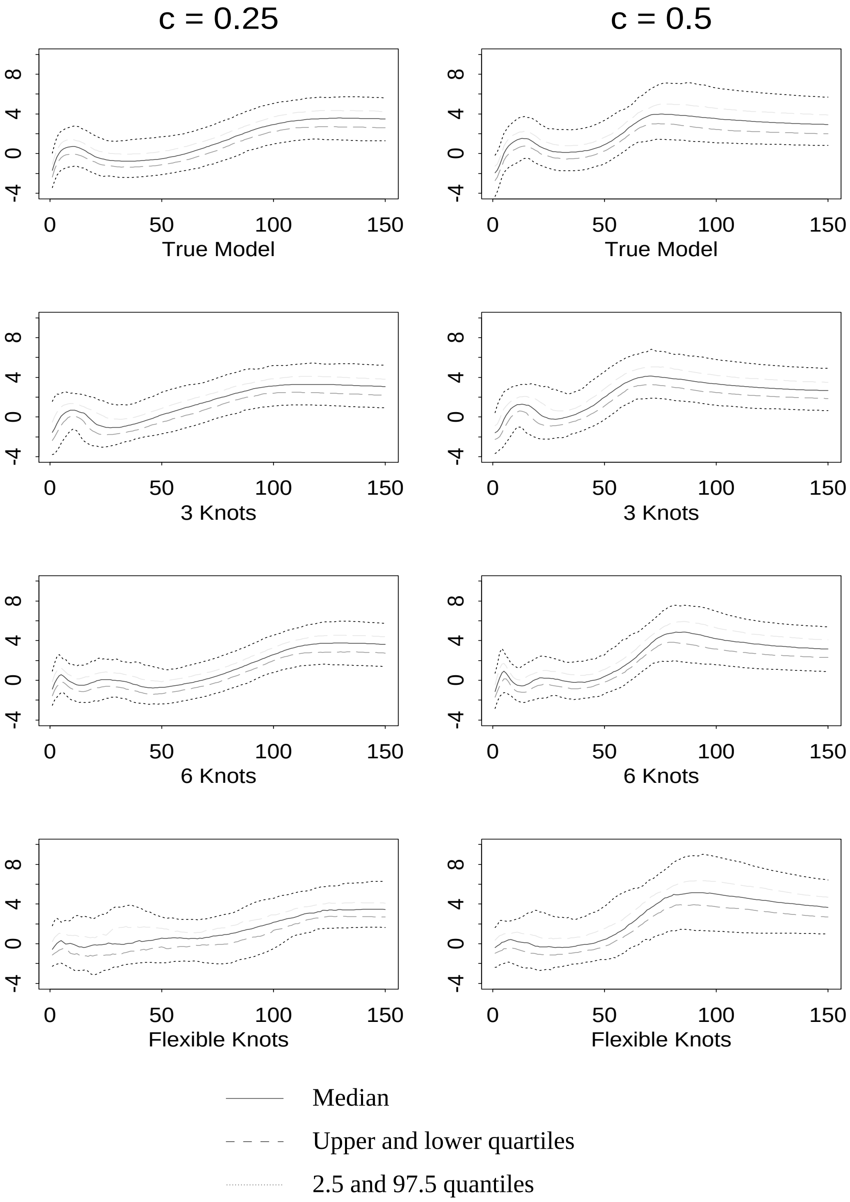

Figure 5.17: The quantiles of the pointwise z values of the estimated log hazards for the B-spline model (5.4).

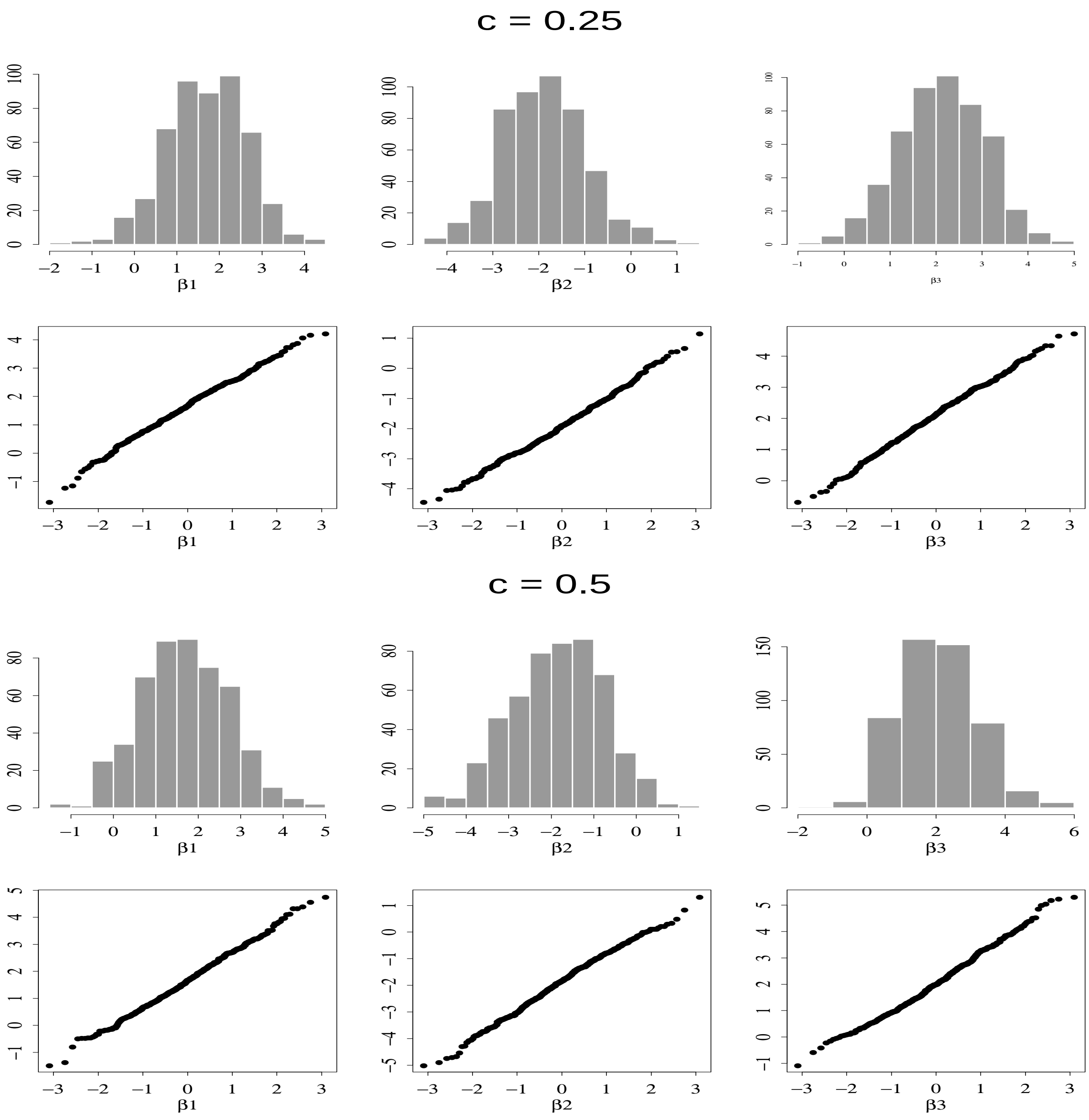

Figure 5.18: The histograms and the qq-plots of the standardized estimates of the parameter $\boldsymbol{\beta}$ from the True Knots method for the B-spline model (5.4).

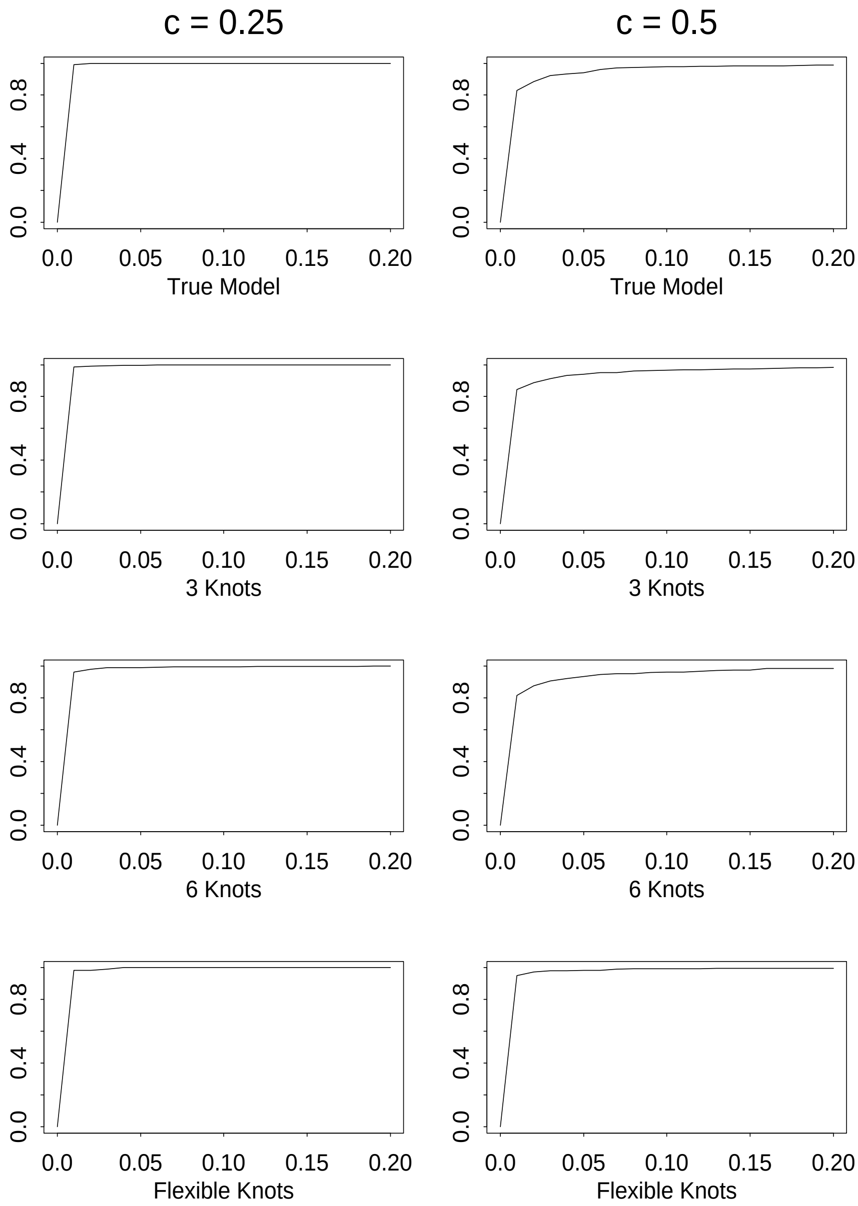

Figure 5.19: Empirical distribution functions of the p-values for testing the hypothesis that the failure times have a Weibull distribution.

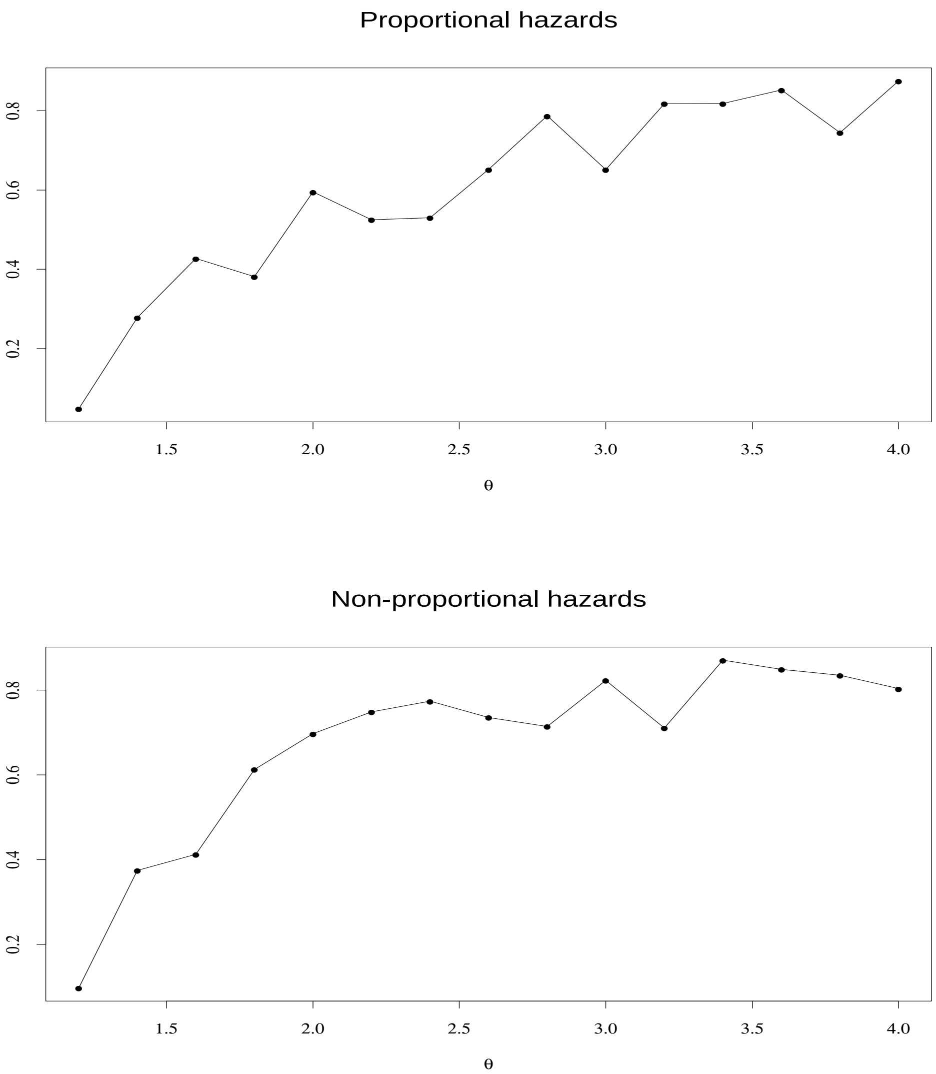

Figure 5.20: The plots of the correlation coefficients versus θ for the data as in the proportional hazards and non-proportional hazards models

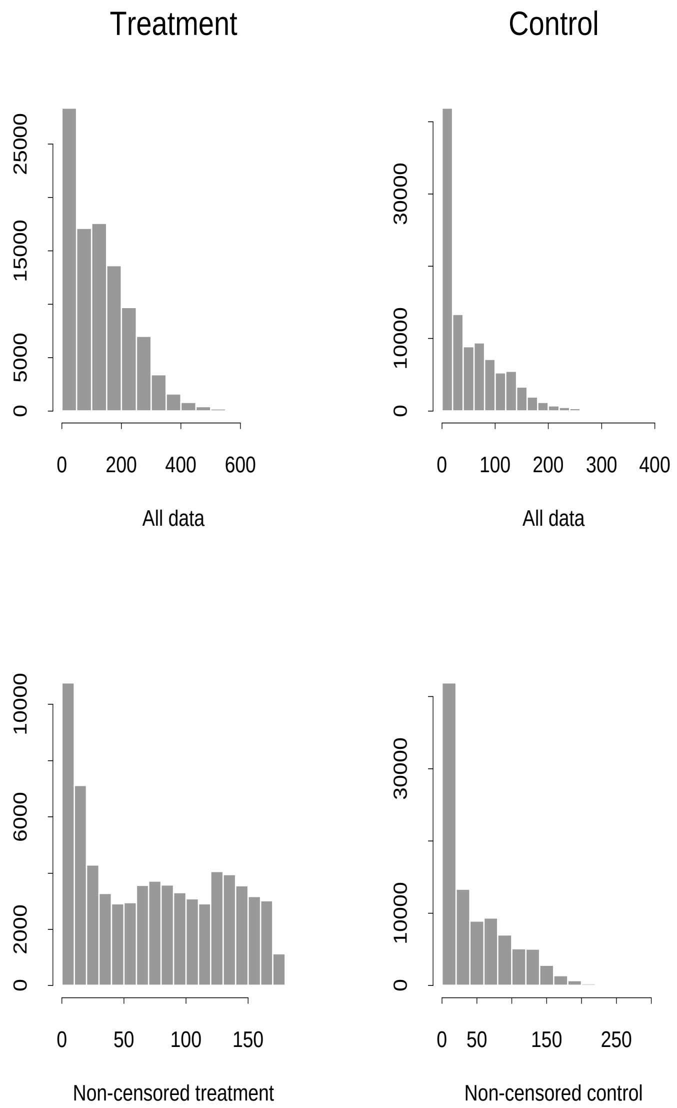

Figure 5.21: The histograms of the marginal failure times, and non-censored marginal failure times for the data generated according to the proportional hazards model in Section 5.1. The censoring rate for the "treatment" is 25%.

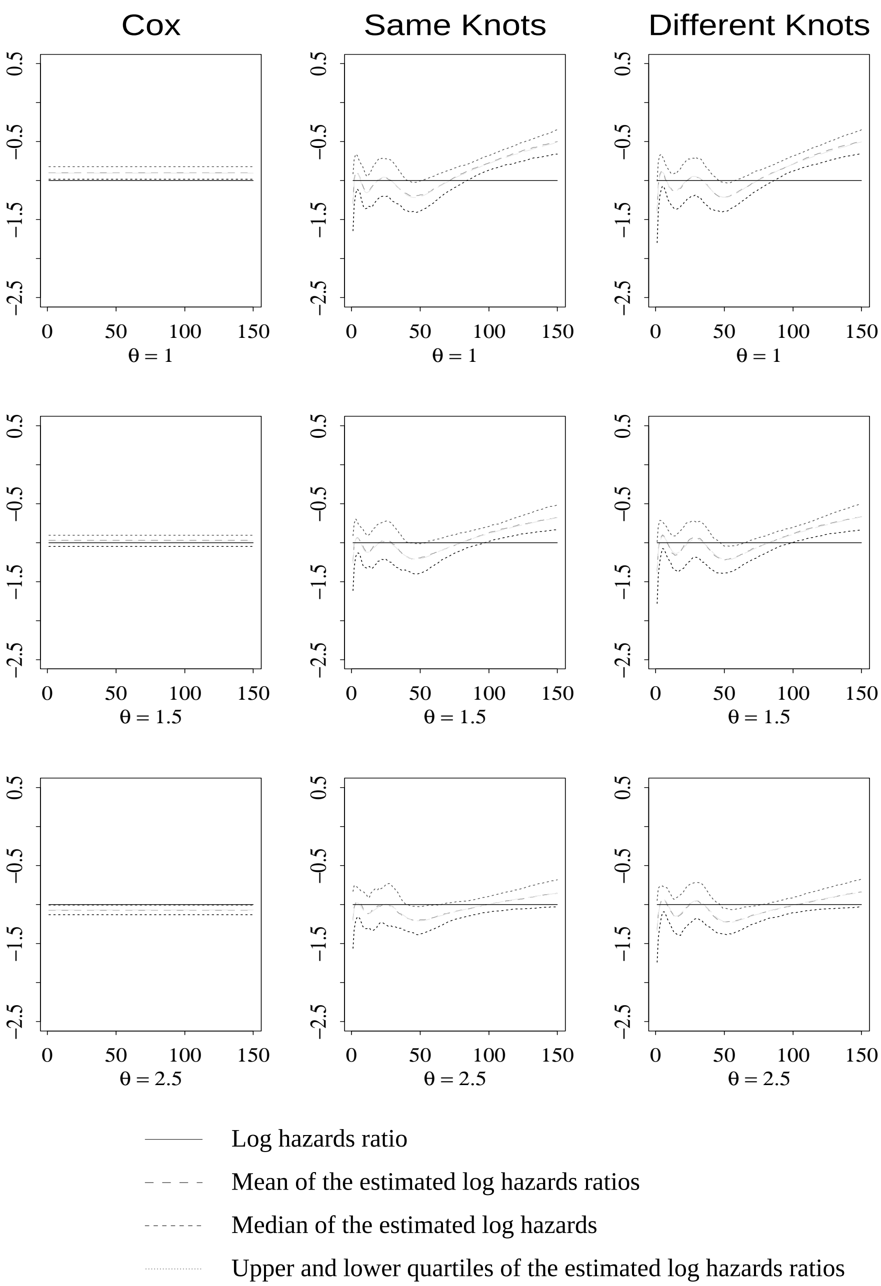

Figure 5.22: The true log hazards ratio of the data from the proportional hazards model defined in Section 5.1 and the pointwise quartiles and empirical mean of the 500 estimated log hazards ratios.

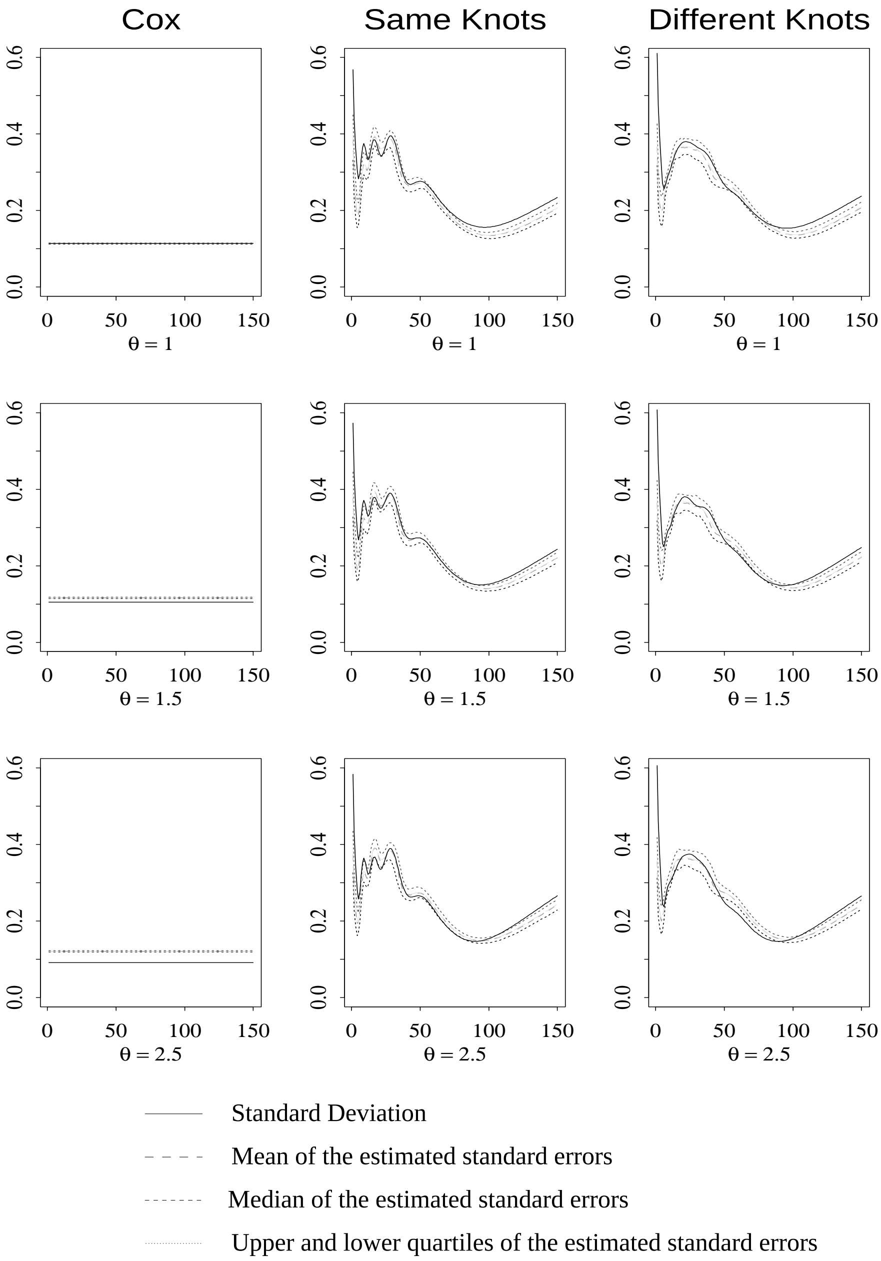

Figure 5.23: The standard deviations of the estimated log hazards ratio for data generated according to the proportional hazards model. The censoring rate is 25% for the "treatment".

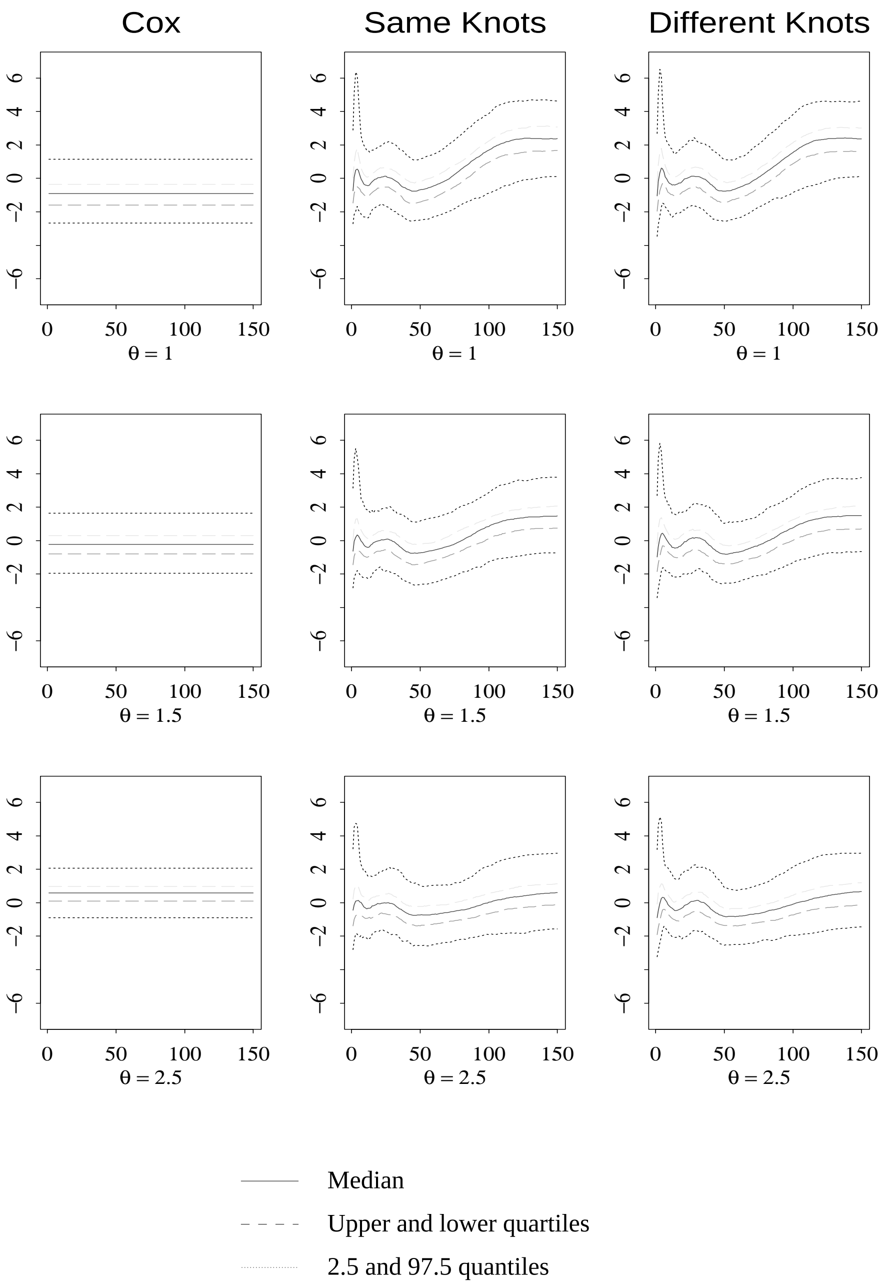

Figure 5.24: The quantiles of the pointwise z values of the estimated log hazards ratio for the data generated according to the proportional hazards model. The censoring rate is 25% for the "treatment".

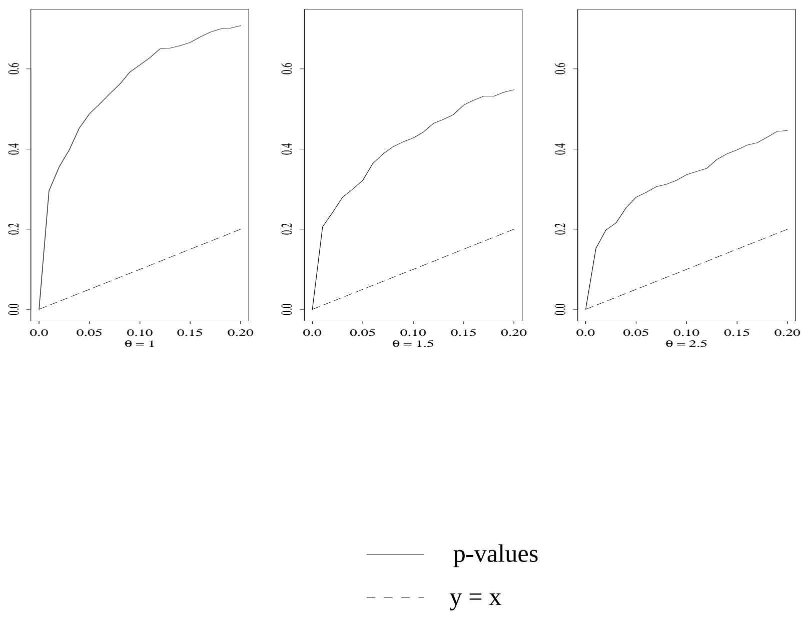

Figure 5.25: Empirical distribution functions of the p-values for testing the hypothesis that the two failure times of the data generated as in the proportional hazards model in Section 5.1 have proportional hazards.

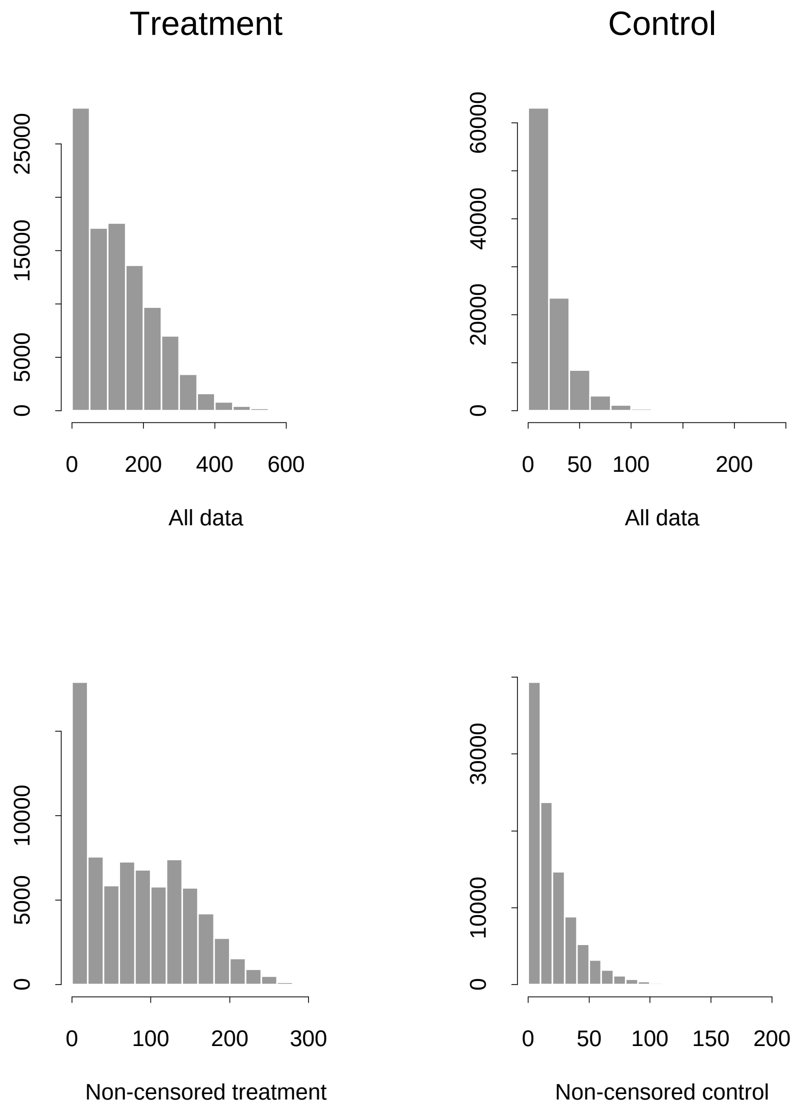

Figure 5.26: The histograms of the marginal failure times, and the non-censored marginal failure times for the data from the non-proportional hazards model in Section 5.1. The censoring rate for the "treatment" is 25%.

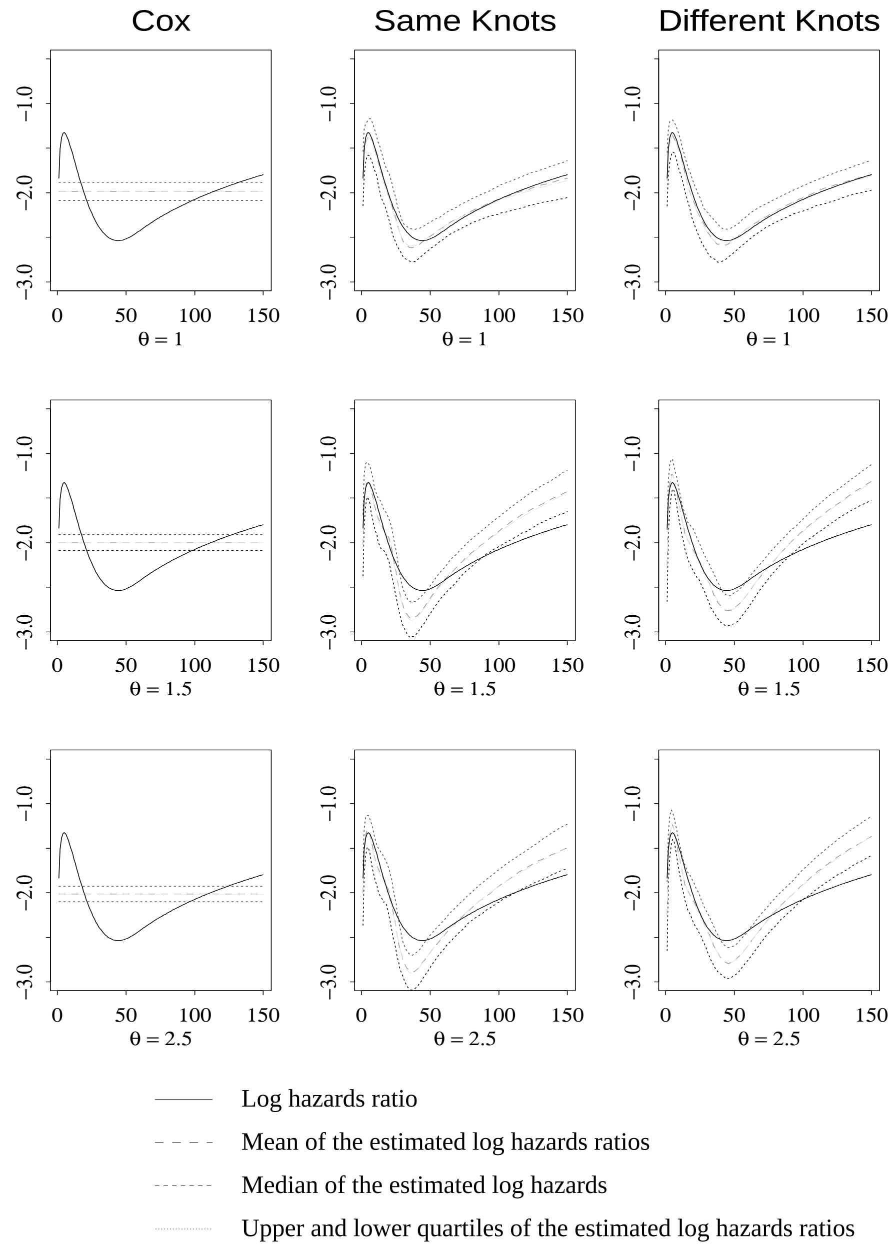

Figure 5.27: The true log hazards ratio of the data from the non-proportional hazards model defined in Section 5.1 and the pointwise quartiles and empirical mean of the estimated log hazards ratios.

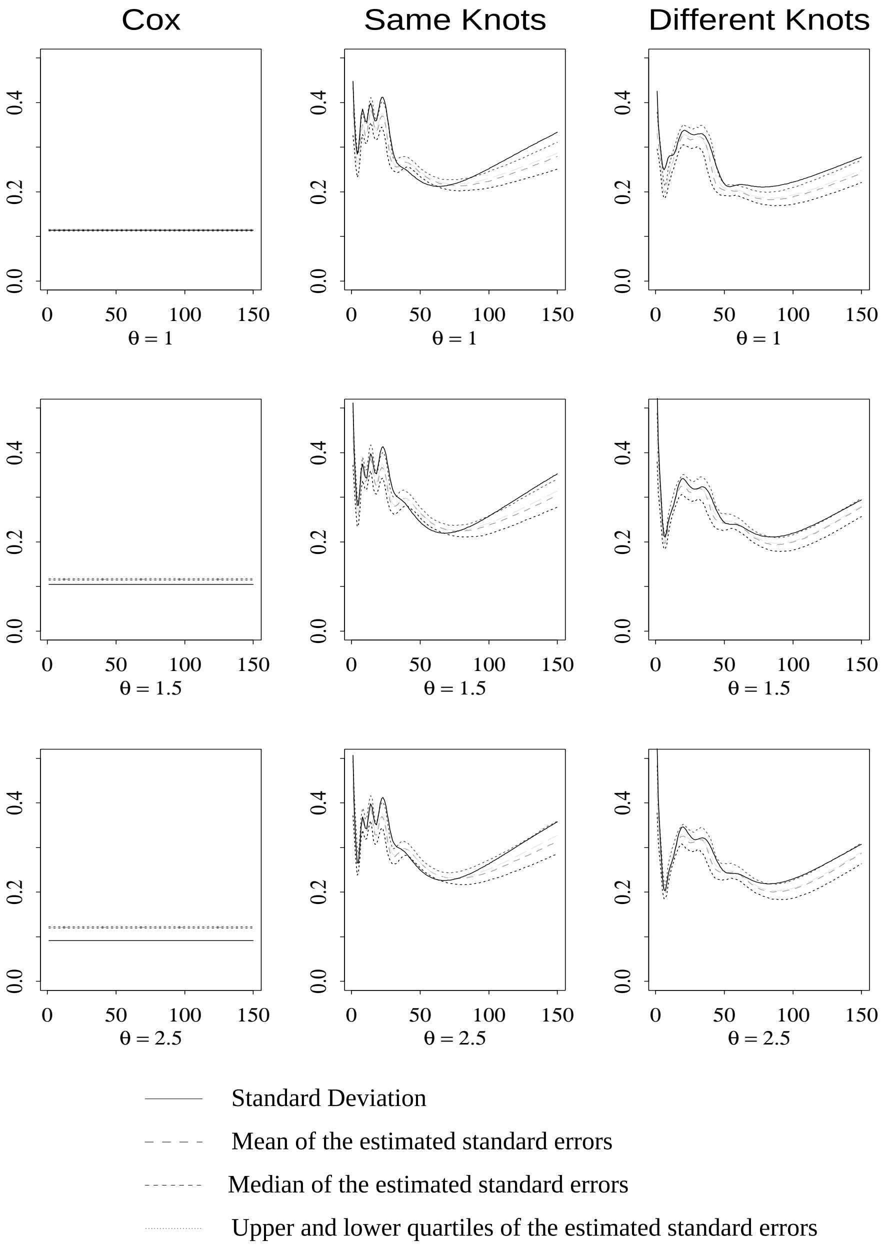

Figure 5.28: The standard deviations of the estimated log hazards ratio for data generated according to the non-proportional hazards model. The censoring rate is 25% for the "treatment".

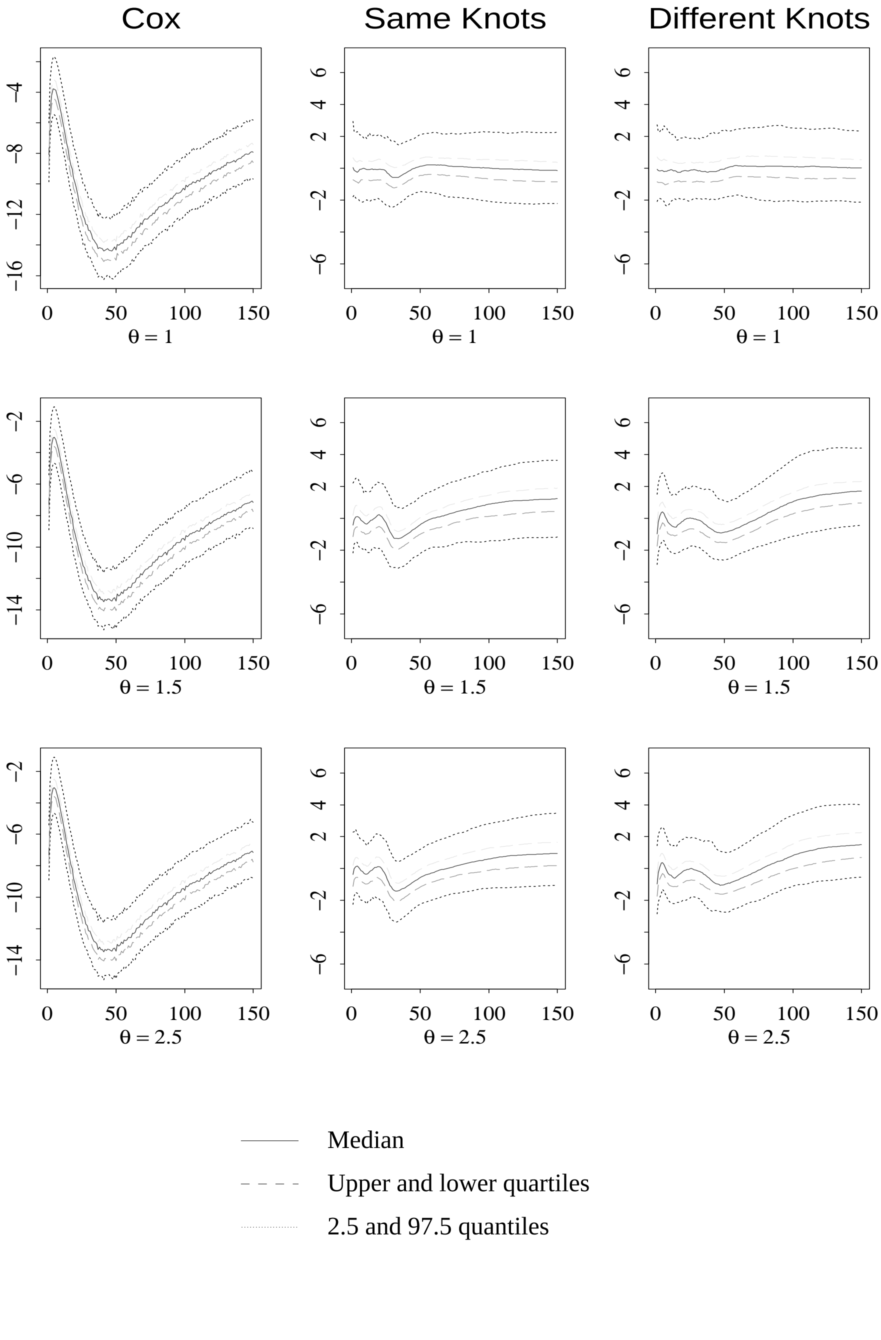

Figure 5.29: The quantiles of the pointwise z values of the estimated log hazards ratio for the data generated according to the non-proportional hazards model. The censoring rate is 25% for the “treatment”.

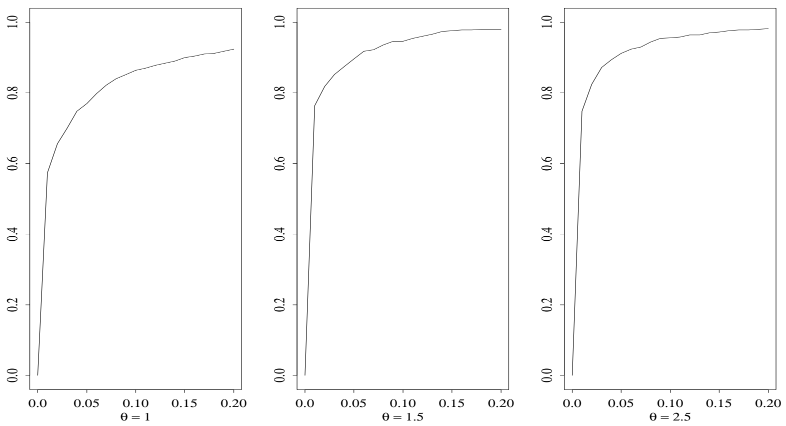

Figure 5.30: Empirical distribution functions of the p-values for testing the hypothesis that the two failure times of the data generated as in the non-proportional hazards model in Section 5.1 have proportional hazards.

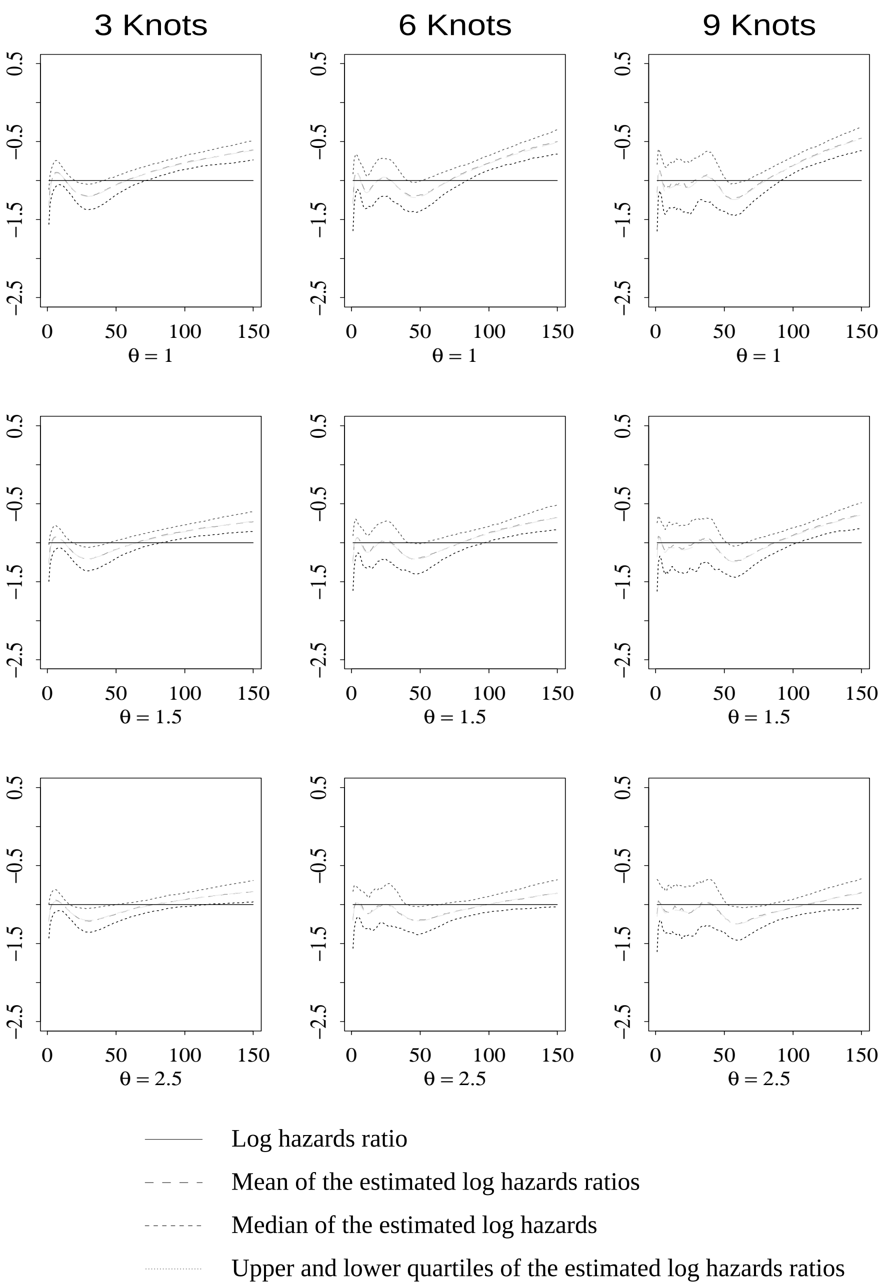

Figure 5.31: The true log hazards ratio of the data from the proportional hazards model defined in Section 5.1 and the pointwise quartiles and empirical mean of the estimated log hazards ratios by the Same Knots method.

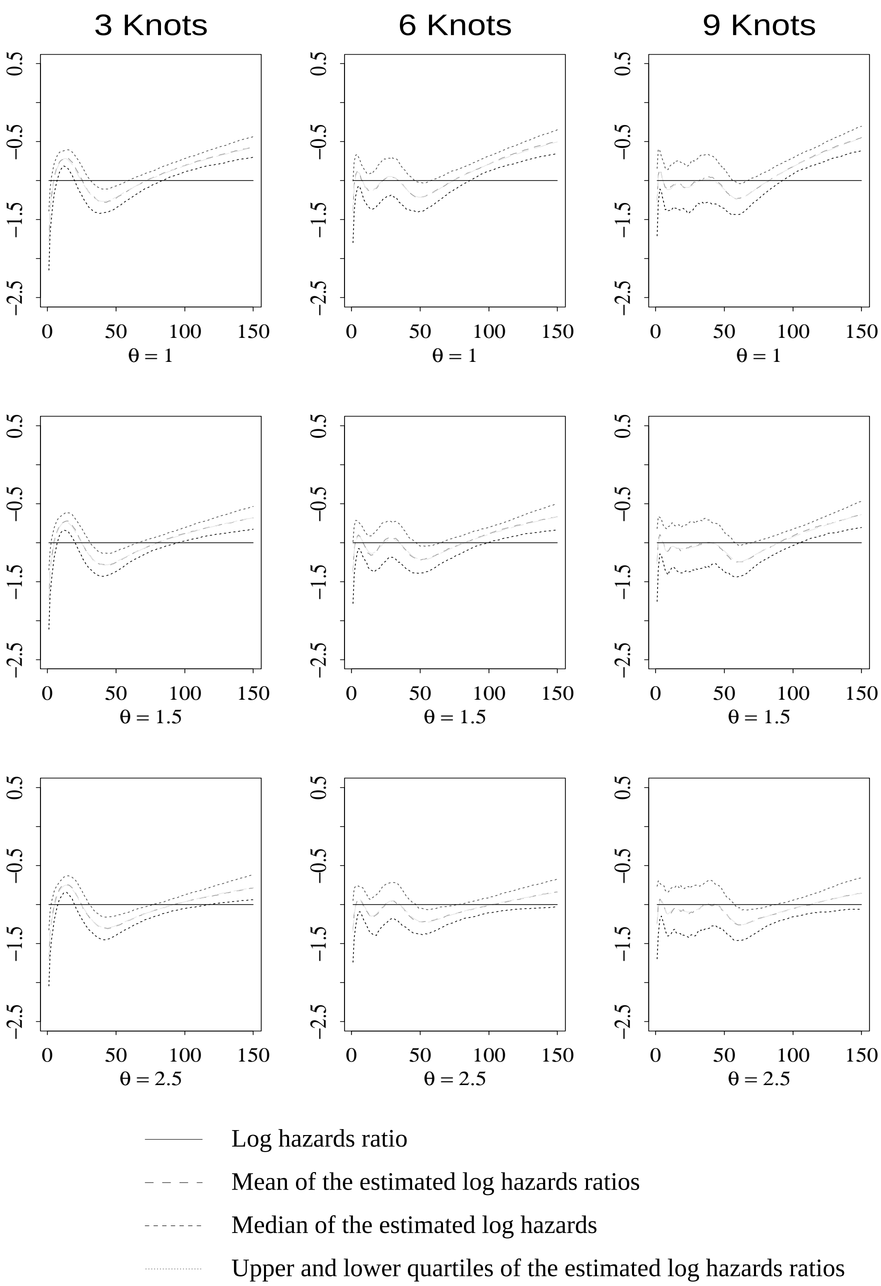

Figure 5.32: The true log hazards ratio of the data from the proportional hazards model defined in Section 5.1 and the pointwise quartiles and empirical mean of the estimated log hazards ratios by the Different Knots method.

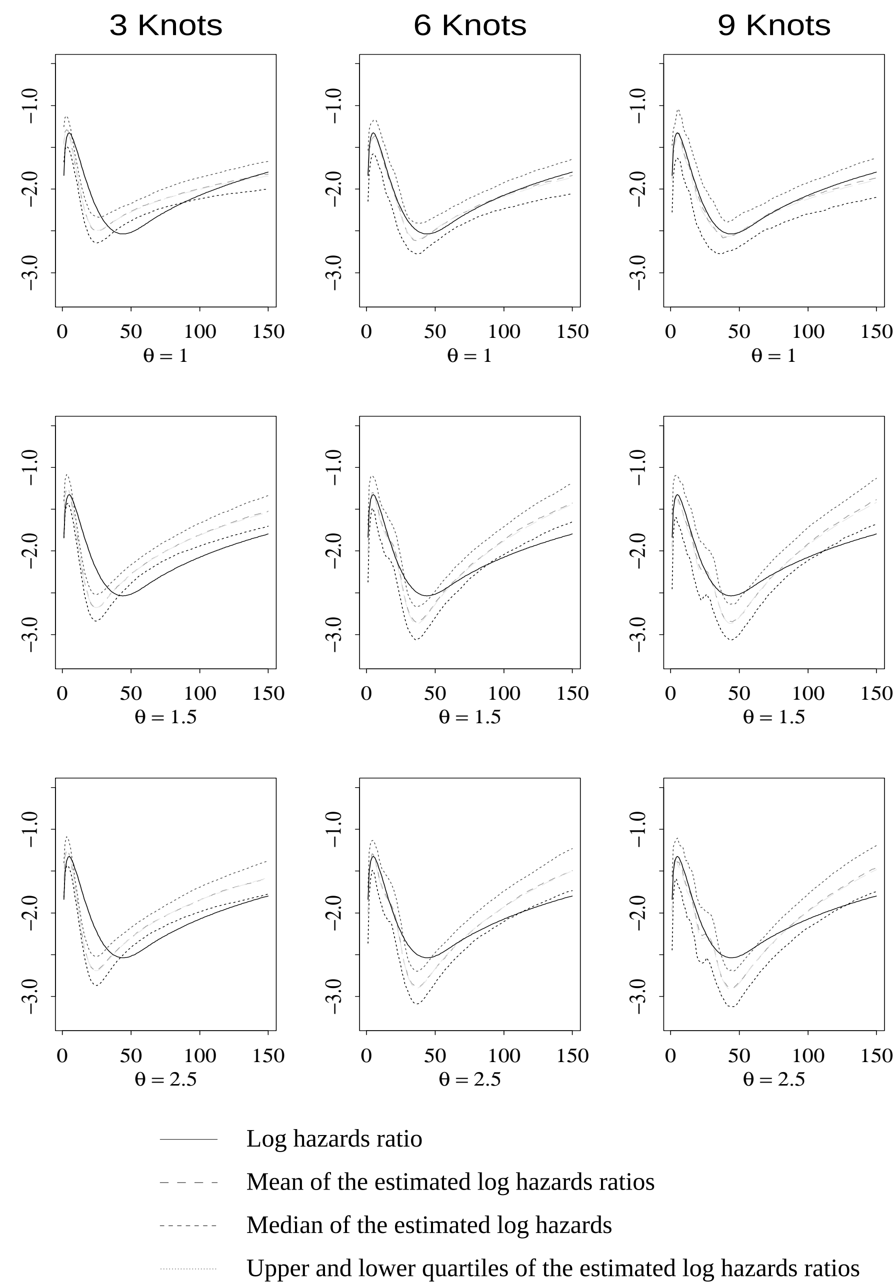

Figure 5.33: The true log hazards ratio of the data from the non-proportional hazards model defined in Section 5.1 and the pointwise quartiles and empirical mean of the estimated log hazards ratios by the Same Knots method.

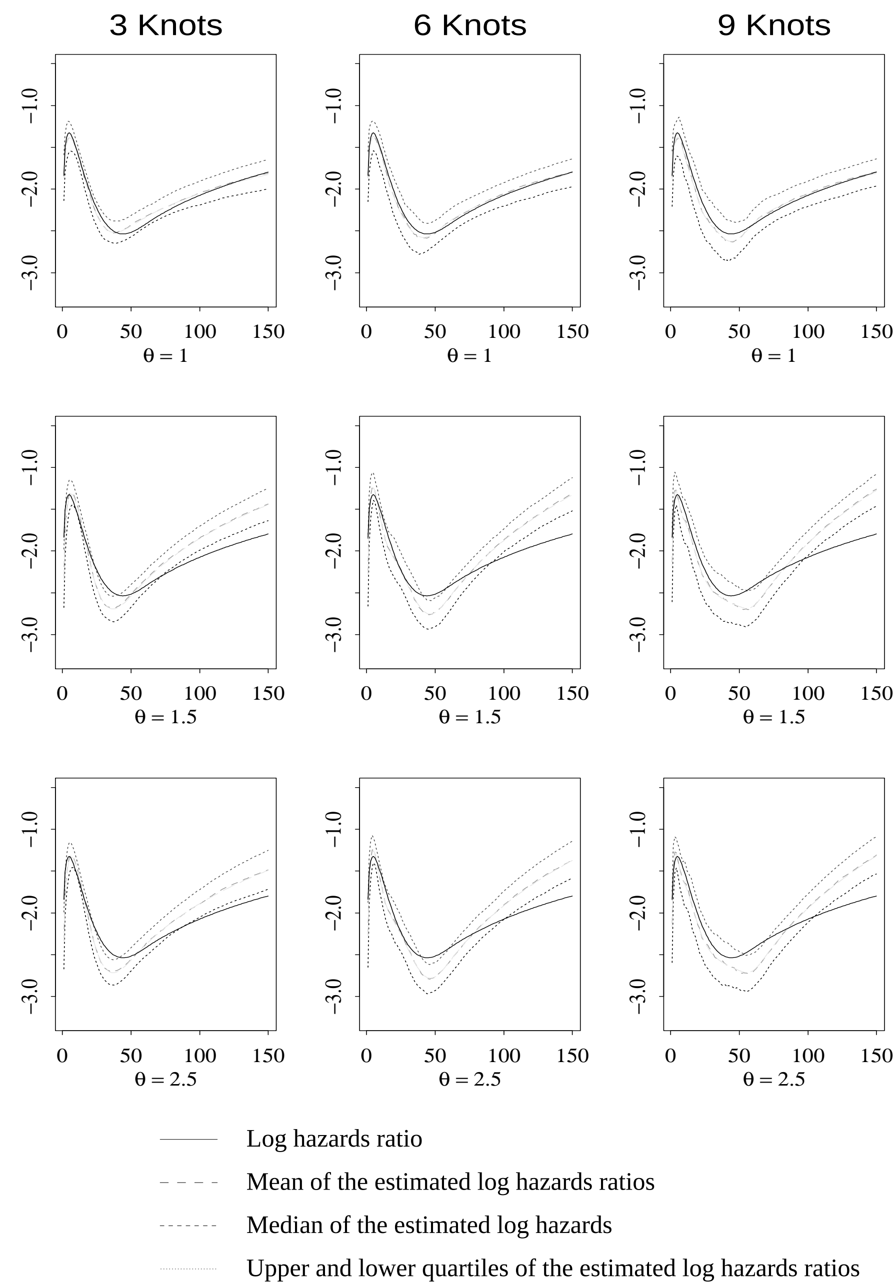

Figure 5.34: The true log hazards ratio of the data from the non-proportional hazards model defined in Section 5.1 and the pointwise quartiles and empirical mean of the estimated log hazards ratios by the Different Knots method.

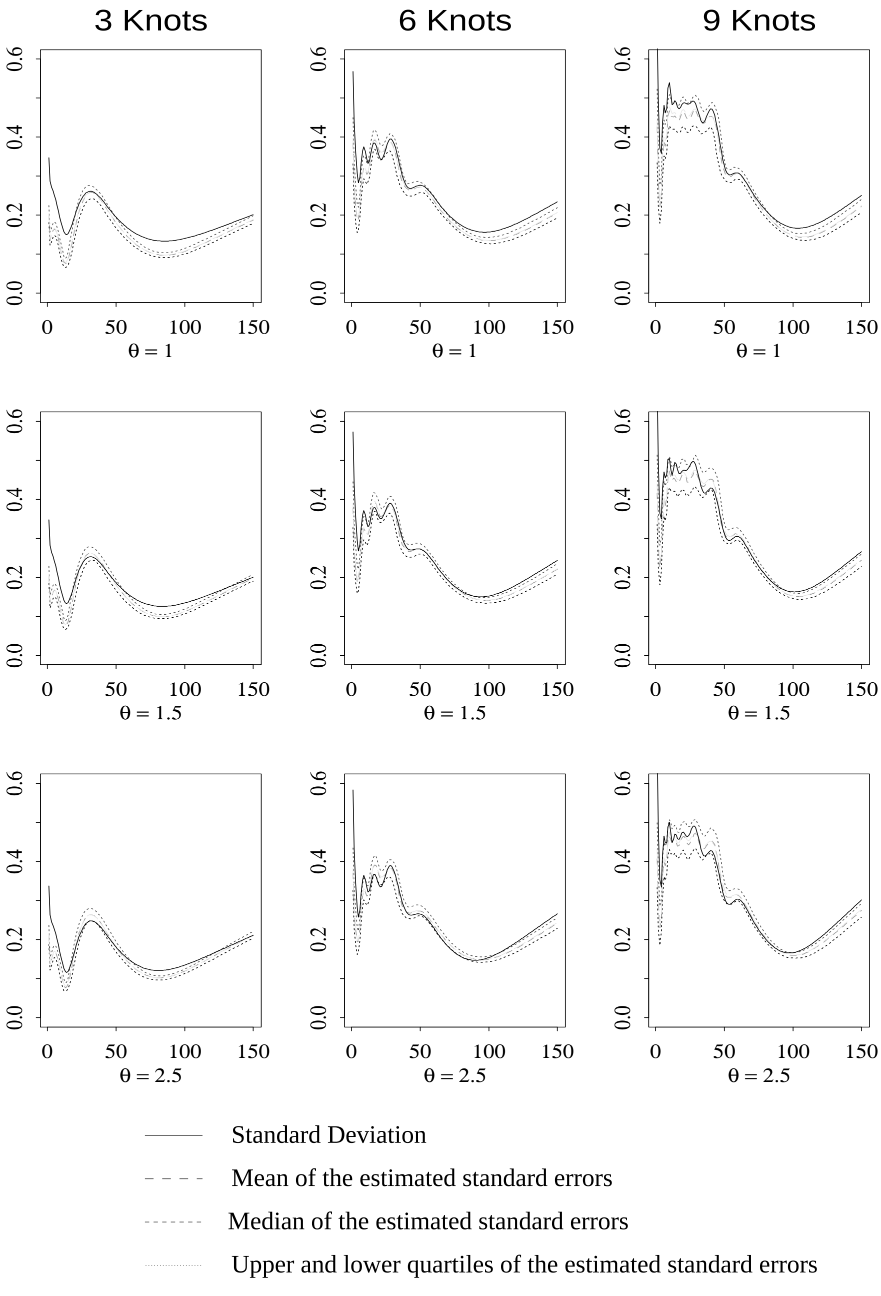

Figure 5.35: The standard deviations of the estimated log hazards ratio by the Same Knots method for data generated according to the proportional hazards model. The censoring rate is 25% for the “treatment”.

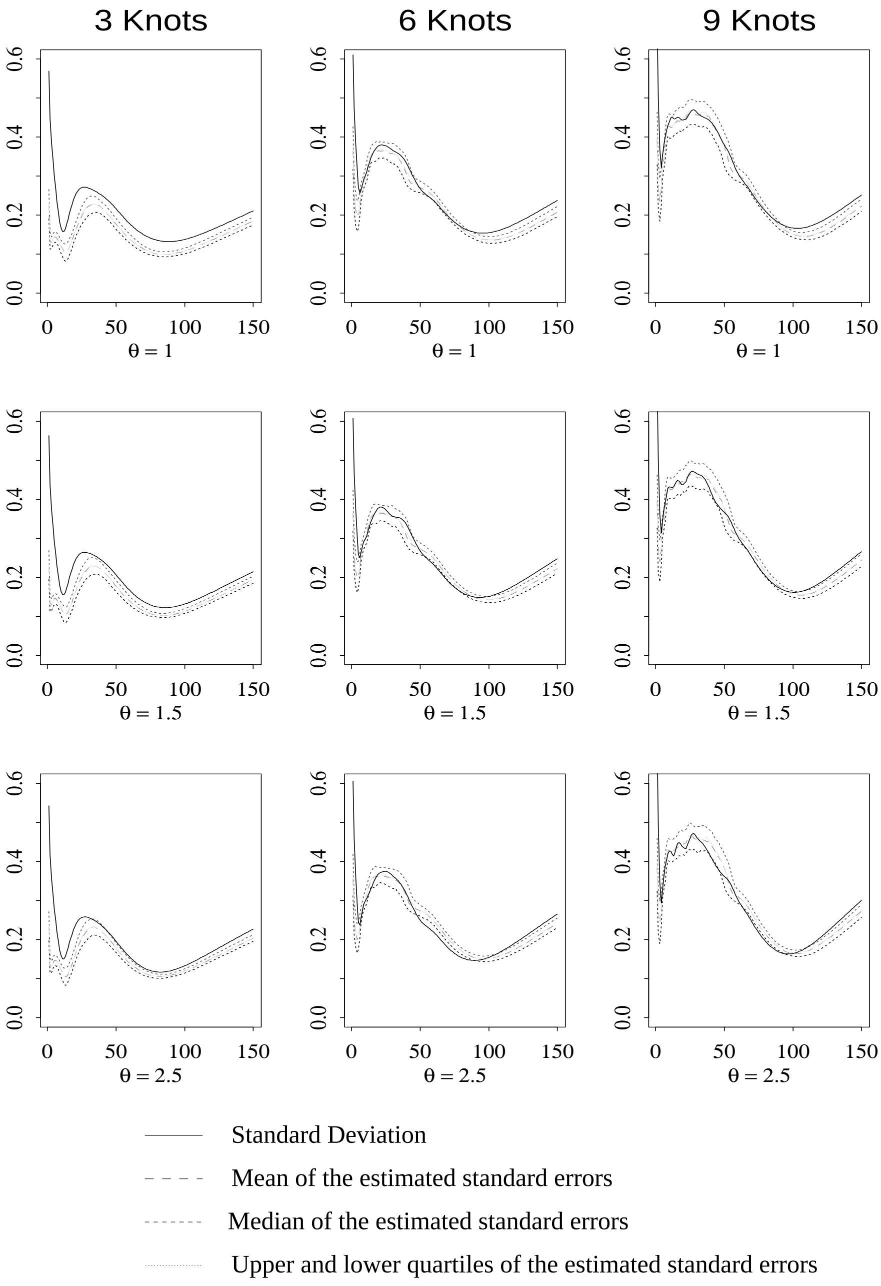

Figure 5.36: The standard deviations of the estimated log hazards ratio by the Different Knots method for data generated according to the proportional hazards model. The censoring rate is 25% for the "treatment".

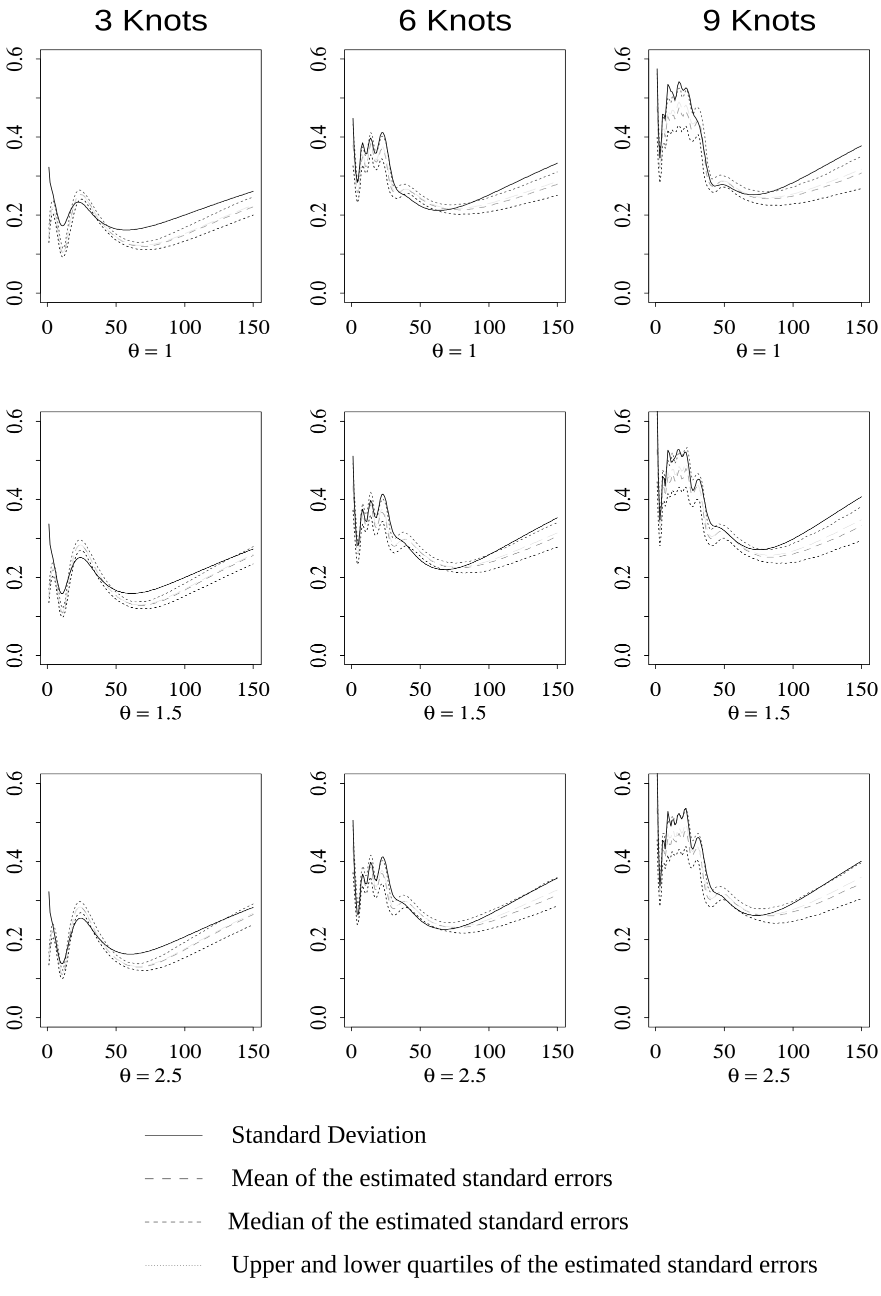

Figure 5.37: The standard deviations of the estimated log hazards ratio by the Same Knots method for data generated according to the non-proportional hazards model. The censoring rate is 25% for the "treatment".

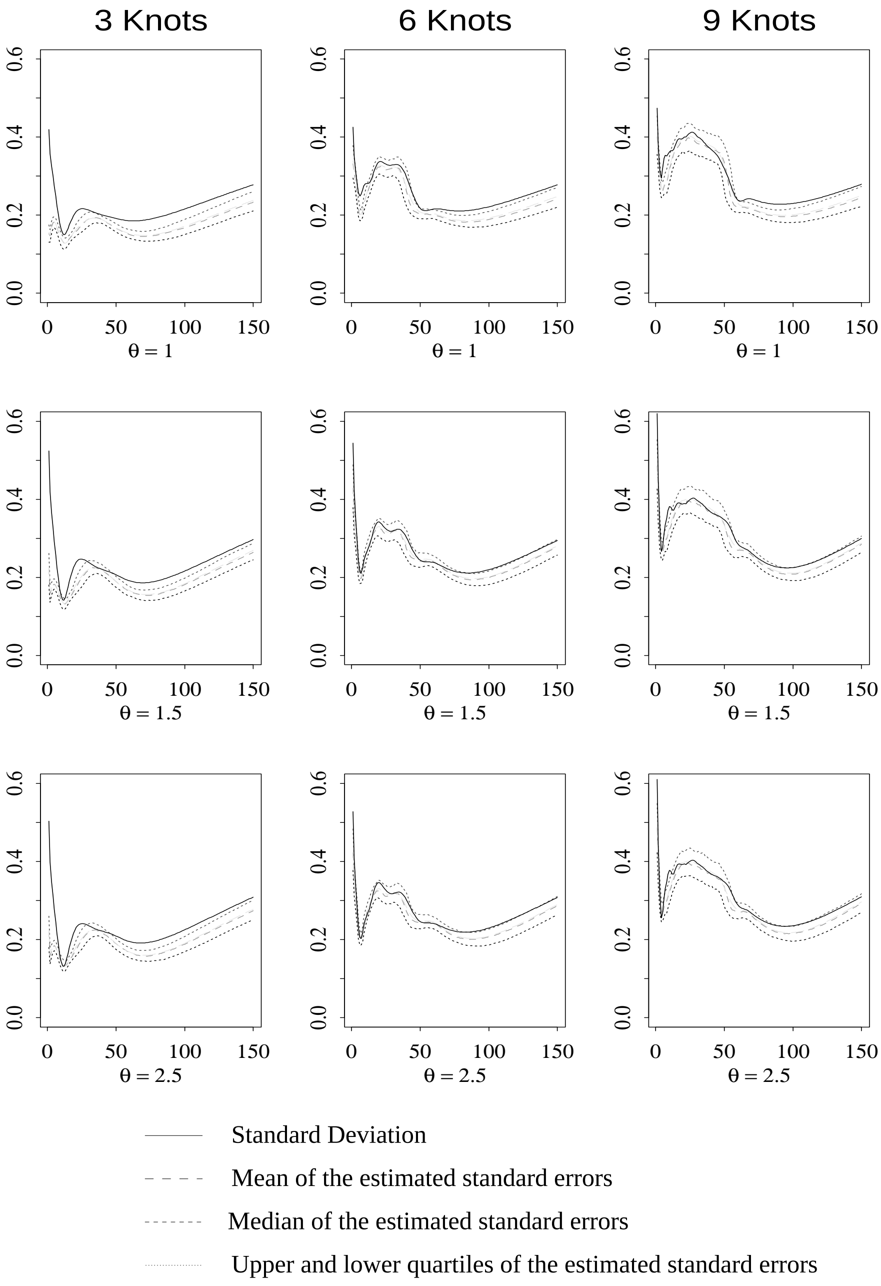

Figure 5.38: The standard deviations of the estimated log hazards ratio by the Different Knots method for data generated according to the non-proportional hazards model. The censoring rate is 25% for the "treatment".

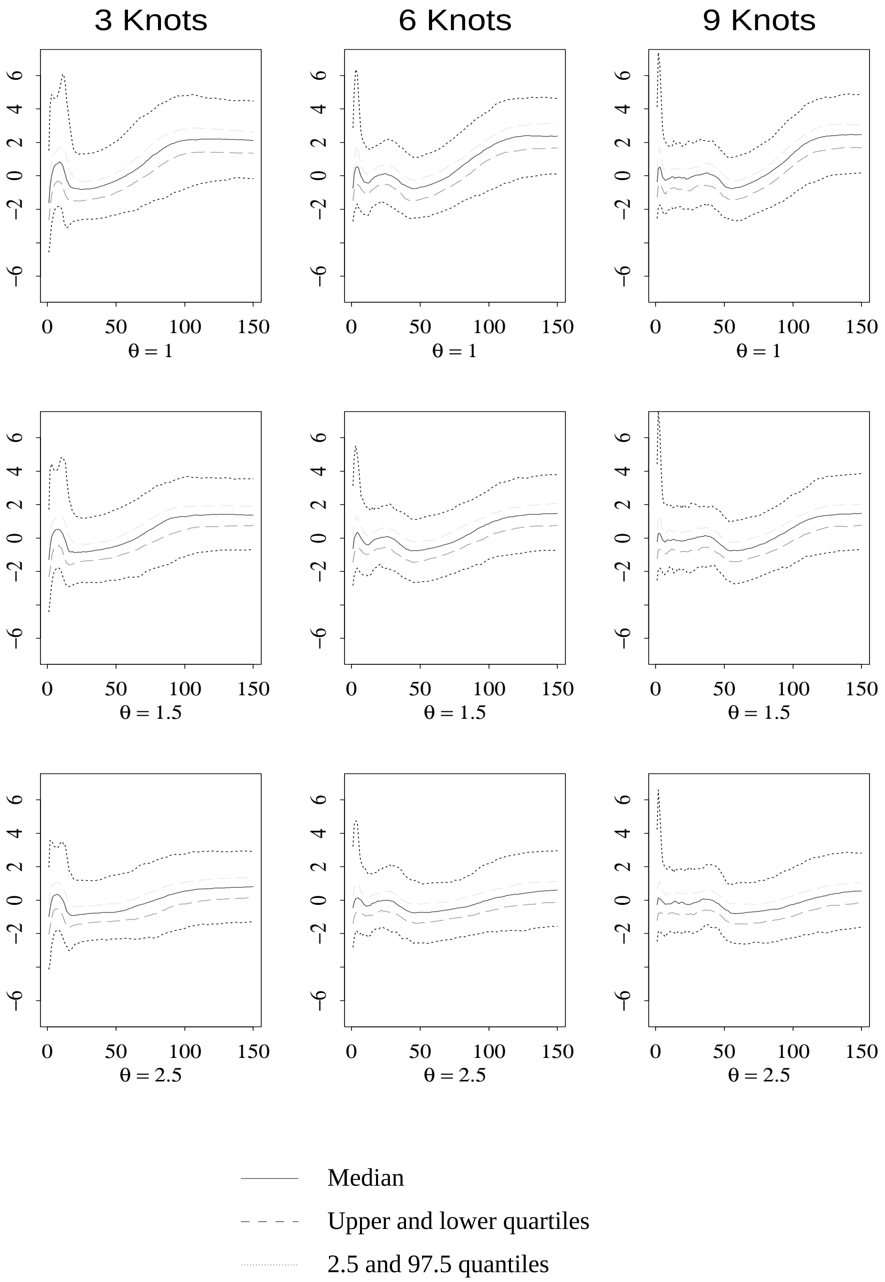

Figure 5.39: The quantiles of the pointwise z values of the estimated log hazards ratio by the Same Knots method for the data generated according to the proportional hazards model. The censoring rate is 25% for the "treatment".

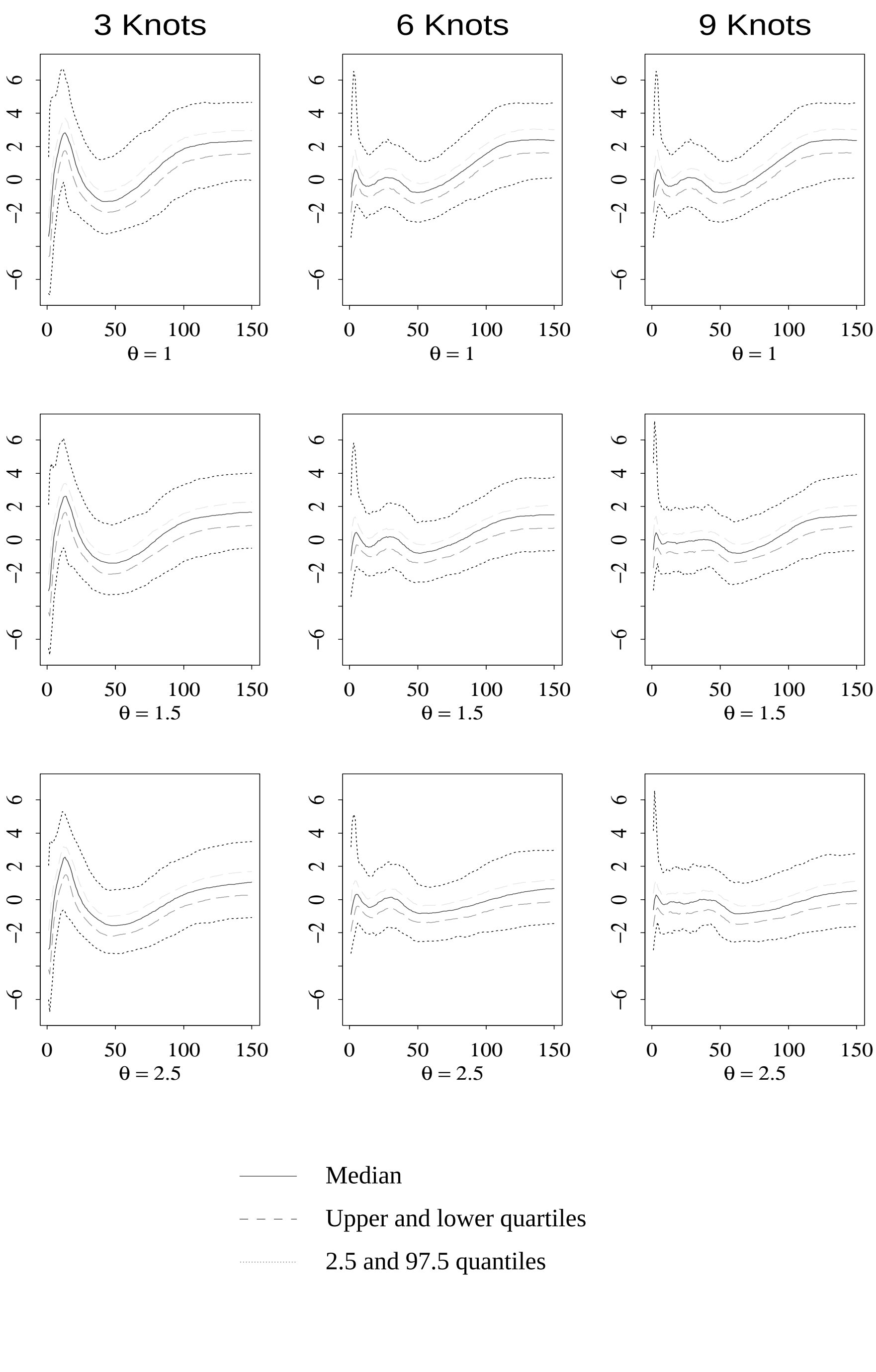

Figure 5.40: The quantiles of the pointwise z values of the estimated log hazards ratio by the Different Knots method for the data generated according to the proportional hazards model. The censoring rate is 25% for the "treatment".

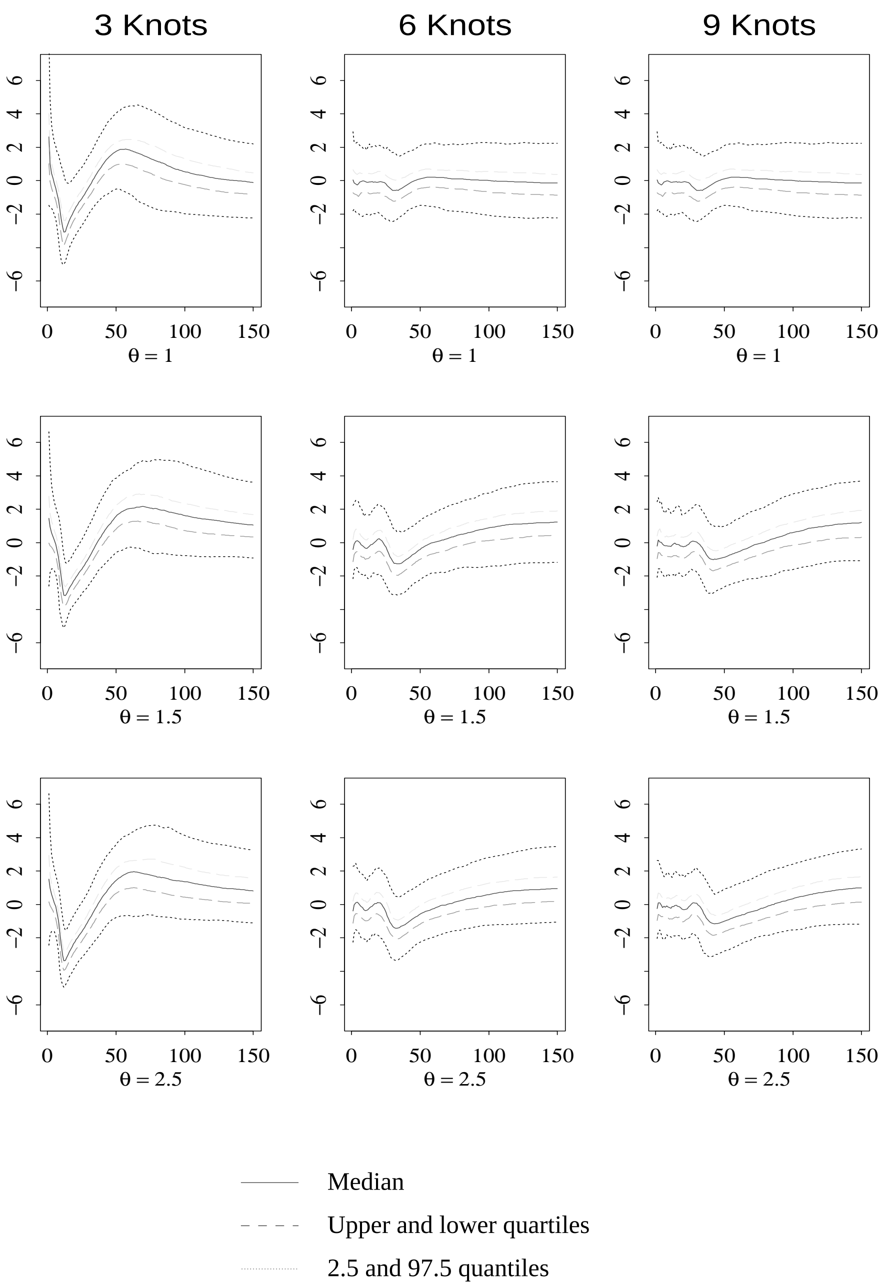

Figure 5.41: The quantiles of the pointwise z values of the estimated log hazards ratio by the Same Knots method for the data generated according to the non-proportional hazards model. The censoring rate is 25% for the “treatment”.

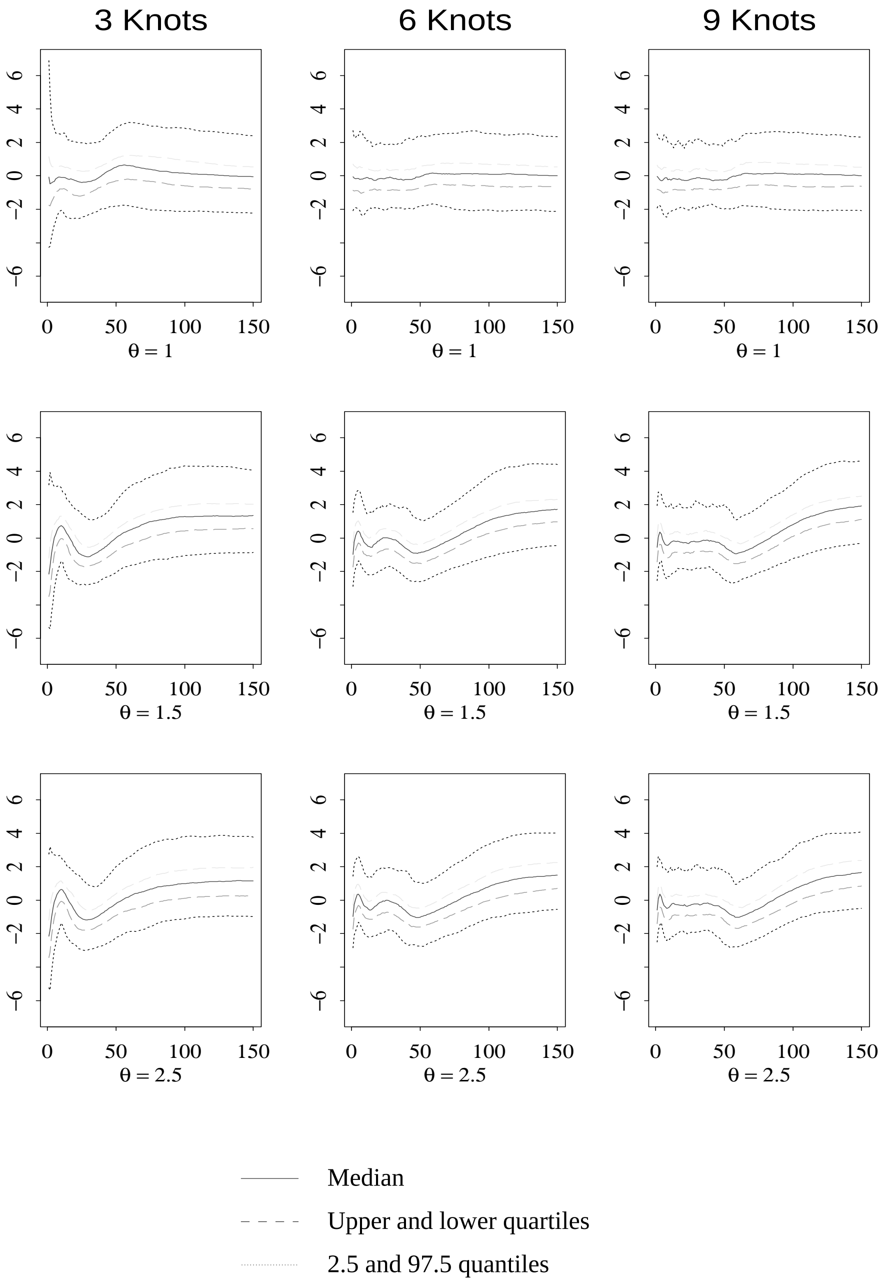

Figure 5.42: The quantiles of the pointwise z values of the estimated log hazards ratio by the Different Knots method for the data generated according to the non-proportional hazards model. The censoring rate is 25% for the "treatment".

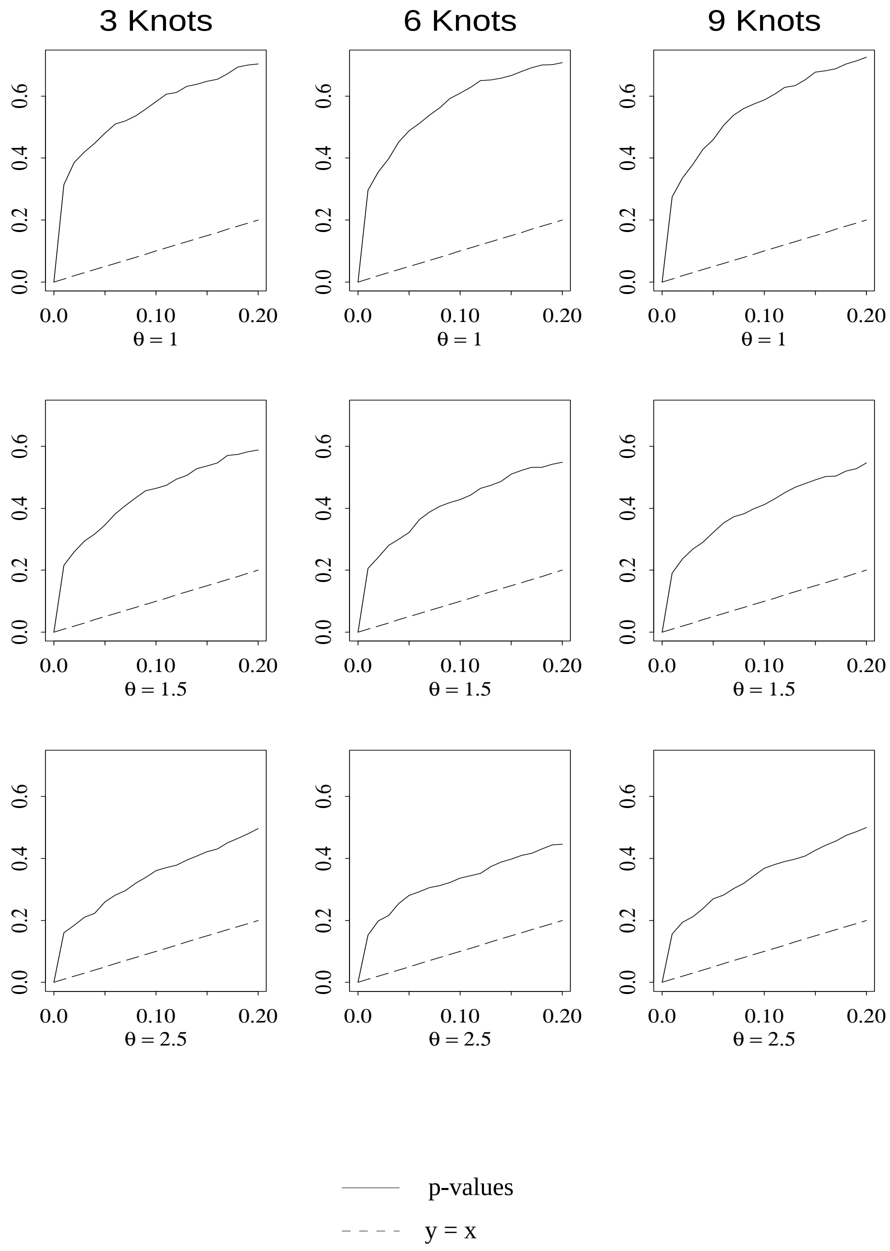

Figure 5.43: By the Same Knots method empirical distribution functions of the p-values for testing the hypothesis that the two failure times of the data generated as in the proportional hazards model in Section 5.1 have proportional hazards.

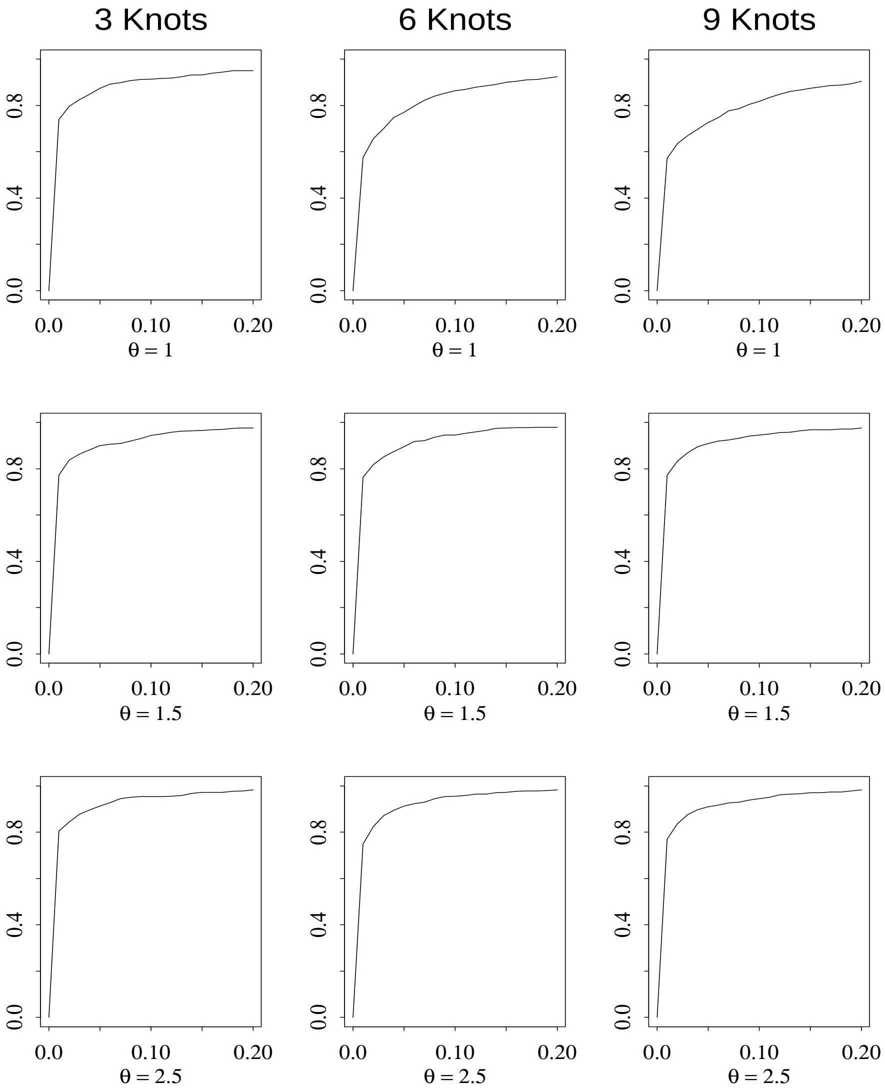

Figure 5.44: By the Same Knots method empirical distribution functions of the p-values for testing the hypothesis that the two failure times of the data generated as in the non-proportional hazards model in Section 5.1 have proportional hazards.