Delft University of Technology

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft Institute of Applied Mathematics

Master of Science Thesis

Model Adequacy Test for Cox Proportional Hazard Model

by

Bowen Zhang

Delft, the Netherlands

July 2008

Members of the Committee

Chairperson of Graduate Committee:

Prof. dr. R. M. Cooke
Delft University of Technology, the Netherlands
Resources for Future, United States

Committee:

Prof. dr. R. M. Cooke
Delft University of Technology, the Netherlands
Resources for Future, United States

Dr. D. Kurowicka
Delft University of Technology, the Netherlands

Dr. R. Lopuhaa
Delft University of Technology, the Netherlands

Abstract

Title: Model Adequacy Test for Cox Proportional Hazard Model

By

Bowen Zhang

Cox proportional hazard model has been widely used in survival statistics since it is proposed in 1972. It is often used to regress hazard rate onto covariates that influence the survival time to some hazardous event, such as death. While much effort has been put on development of the regression model, a very important question should be answered as well before the model can be used for prediction or adopted in other fields. The question is how well the model has explained the hazard data? Here comes the issue of model adequacy test. A model adequacy test consists of the examination of the model's fit as well as adherence to model assumptions to check whether the fitted model provides an adequate summary of the data. This procedure is as important as model development.

Several effective methods to test model adequacy have been proposed to examine the model's fit (see [1]). One way is to estimate the coefficients for both the true model and incomplete model where some covariates are excluded, using the data generated from the true model. Estimates are compared to see if the incomplete model estimates well or in other words, to see if the missing covariate is influential to the estimation. Another way is to compare the population cumulative hazard function with estimated baseline cumulative hazard function to see if covariates are influential. These two methods work well when applied to independent covariate in [1]. However, when these covariates are dependent, or when interaction and quadratic terms are considered, results may be different. Moreover, we also have to apply some effective methods to test the adherence to model assumptions.

In this thesis we first extend the results of [1] through application of the methods mentioned above to dependent data and models with interaction and quadratic terms. Then we put forward another method to test model's fit: the likelihood ratio test. After that, we study the test of adherence to model assumption. We check whether the proportional assumption is satisfied for the true model as well as model of missing covariates by including time-dependent covariate into the model. Furthermore, since we assume that the covariates appear in their linear form in the true model we also use martingale residuals and Lowess smooth to test if the linear assumption holds for incomplete models. We apply all these methods to independence covariates, weakly dependent covariates and strongly dependent covariates to see if these methods work well in each situation.

∕∕ ⍝⋲⍝⌇⋾□⋭⋲ ⋭⩲⌇⋪ ⋭⩲⋲⋪⌇⋪ ⋭& ○&⦇ɸ ;□⍝ □⦈⍝ ⨯□⋭□⟧⌇⦈
⨯⌇&⋲□⦈⫫

Contents

Acknowledgement.....13

Chapter 1 Introduction.....21

1.1 Research Objectives.....21

1.2 Previous Works.....22

1.3 Thesis Outline.....23

Chapter 2 Testing Adequacy for Cox Proportional Hazard Model with Simulations.....25

2.1 Cox proportional hazard model.....25

2.1.1 Estimation of Coefficients *A, B* and *C*.....26

2.1.2 Estimation of Cumulative Baseline Hazard Function.....27

2.1.3 Estimation of Population Cumulative Hazard Function.....28

2.2 Model Adequacy Test.....28

2.3 Data Simulation.....29

2.3.1 Generating Samples of Data for Independent Covariates.....29

2.3.2 Generating Samples of Data for dependent Covariates.....30

Chapter 3 Futher Study of Previous Methods.....31

3.1 Comparison between Ordered Estimated Coefficients.....31

3.1.1 Review of the Method.....31

3.1.2 Study of Dependence.....33

3.1.3 Study of Interaction and Quadratic terms.....39

3.2 Comparison between Estimated Cumulative Baseline Hazard Function and Population Cumulative Hazard Function.....43

3.2.1 Review of the Method 44

3.2.2 Study of Dependence 45

3.2.3 Study of Interaction 49

Chapter 4 Other Methods for Model Adequacy Test 53

4.1 Likelihood Ratio Test 53

4.1.1 Introduction 53

4.1.2 Independent Covariates 54

4.1.3 Covariates Strongly Dependent 55

4.1.4 Covariates Weakly Dependent 55

4.2 Inclusion of Time Dependent Variable 56

4.2.1 Introduction 56

4.2.2 True Model *h(x,y,z)* 56

4.2.3 Models with two covariates missing 57

4.2.4 Models with one covariate missing 59

4.2.5 Wald Test 60

4.3 Martingale Residual 64

4.3.1 Martingale Residual and Lowess Smooth 64

4.3.2 An Improvement of This Method 66

4.3.3 Independent Covariates 66

4.3.5 Dependent Covariates 69

Chapter 5 Conclusion 75

Bibliography 77

Appendix 1 Proof of Pearson Theorem 79

Appendix 2 Additional Plots for Section 3.1 82

Appendix 3 Additional Plots for Section 3.2 89

Appendix 4 Tables for Section 4.1 90

Appendix 5 Additional Plots for Section 4.2 93

Appendix 6 Addition Plots for Section 4.3 95

Acknowledgement

I would like to express my thanks to Roger M. Cooke and Dorota Kurowicka for giving me the chance to study in the Risk and Environmental modeling group for two years. During these two years, they not only helped me to schedule my study, always accessible whenever I had questions, but also constructed our faculty into a family via lots of activities such as the picnic.

Many thanks to Patrycja L. Gradowska for helping me with every aspect of the thesis; she is always very patient with every question I asked and she put a lot of effort in helping me modify the thesis until it is completely ok. The most important thing I learned from her is how to present the result in a proper way, or in other words, how to "*sell your work*".

I wish to give my appreciation to my good friends Kasia Jozwiak, Kasia Krugla and Lukasz Bednara. I did almost all of the projects necessary for the master program together with Kasia Jozwiak. She is a good partner to work with and we really did some nice projects. Kasia Krugla gave me nice advice on a lot of things and encouraged me whenever I was in a bad mood. Lukasz Bednara helped a lot with my studies especially at the beginning of the master program. He also helped me to fix my laptop more than once so that all important documents I stored are ok.

Thanks very much for my family. They always encourage me to work hard and enjoy life when abroad. Especially, endless support from my parents is my source of strength to move on.

In particular, I wish to thank Catalin Ciobanu. He is always the person who moves closer to me when the rest of the world moves away from me. His caring made me feel that I have a family here and my loneliness vanished thanks to that.

List of Figures

Figure 3.1 100 ordered estimates of A for hxyz, hxy, hx, independent, C=1 32

Figure 3.2 100 ordered estimates of A for hxyz, hxy, hx, independent, C=5 32

Figure 3.3 100 ordered estimates, X strongly dependent on Y,Z, C=1, rank correlation of Y and Z is zero 33

Figure 3.4 100 ordered estimates, X and Z, X and Y strongly dependent, Y and Z weakly correlated, C=1...34

Figure 3.5 100 ordered estimates, X strongly dependent on Y and Z, Y and Z strongly correlated, C=1..35

Figure 3.6 100 ordered estimates all covariates weakly dependent, C=1....................................36

Figure 3.7 100 ordered estimates of A for hxyz, hxy, hx, strong dependence, C=5 37

Figure 3.8 100 ordered estimates of A for hxyz, hxy, hx, weak dependence, C=5.......................37

Figure 3.9 Negative and weak dependence...38

Figure3.10 Negative and strong dependence 39

Figure 3.11 100 ordered estimates of A for the true model and models with extra interaction or quadratic terms...40

Figure 3.12 100 ordered estimates of A for h(x,y,z,xy,yz,xz),h(x,y,xy),h(x), independent, C=5 ..41

Figure 3.13 100 ordered estimates of A for h(x,y,z,xy,yz,xz),h(x,y,xy),h(x), strong dependence, C=5..42

Figure 3.14 100 ordered estimates for model with both interaction and quadratic terms, strong dependence, C=5 42

Figure 3.15 100 ordered estimates for model with both interaction and quadratic terms, weak dependence, C=5 43

Figure 3.16 Cumulative population and baseline hazard, independent, C=1...............................44

Figure 3.17 Cumulative population and baseline hazard, independent, C=5...............................45

Figure 3.18 Cumulative population and baseline hazard, strong dependence, C=1......................46

Figure 3.19 Cumulative population and baseline hazard, strong dependence, C=5......................47

Figure 3.20 Cumulative population and baseline hazard, weak dependence, C=1 48

Figure 3.21 Cumulative population and baseline hazard, weak independence, C=5 48

Figure 3.22 Cumulative population and baseline hazard for model with interaction, independent, C=1 49

Figure 3.23 Cumulative population and baseline hazard for model with interaction, independent, C=5 50

Figure 3.24 Cumulative population and baseline hazard for model with interaction, strong dependence, C=5 51

Figure 3.25 Cumulative population and baseline hazard for model with interaction, weak dependence, C=5 51

Figure 4.1 100 ordered estimates of model coefficients for model with and without time dependent variable, true model, (A,B,C)=(1,1,5)...57

Figure 4.2 Estimates of A, inclusion of time dependent variable, for the model with two missing covariates, independent, (A,B,C)=(5,5,5)..58

Figure 4.3 Estimates of A, inclusion of time dependent variable, for the model with two missing covariates, strong dependence, (A,B,C)=(5,5,5)...59

Figure 4.4 Estimates of A, inclusion of time dependent variable, for the model with one missing covariates, independent, (A,B,C)=(5,5,5)..59

Figure 4.5 Estimates of A, inclusion of time dependent variable, for the model with one missing covariates, strong dependence, (A,B,C)=(5,5,5)...60

Figure 4.6 Martingale residual and Lowess smooth for h(x,y,z), independent.............................67

Figure 4.7 *fi* versus *xi* for h(x,y,z), independent 67

Figure 4.8 Martingale residual and Lowess smooth for h(x,y), independent...............................68

Figure 4.9 *fi* versus *xi* for h(x,y), independent 68

Figure 4. 10 *fi* versus *xi* for h(x), independent 69

Figure 4. 11 Martingale residual and Lowess smooth for h(x,y), strong dependence...................70

Figure 4.12 *fi* versus *xi* for h(x,y), strong dependence 70

Figure 4.13 *fi* versus *xi* for h(x),strong dependence 71

Figure 4. 14 Martingale residual and Lowess smooth for h(x,y), weak dependence 71

Figure 4.15 *fi* versus *xi* for h(x,y), weak dependence..72

Figure 4.16 *fi* versus *xi* for h(x), weak dependence..72

List of Tables

Table 4.1 Significance Test of Time Dependent Variable for model h(x,y,z) , X,Y,Z independent 61

Table 4.2 Significance Test of Time Dependent Variable for model h(x,y), X,Y,Z independent .62

Table 4.3 Significance Test of Time Dependent Variable for model h(x), X,Y,Z independent 62

Table 4.4 Significance Test of Time Dependent Variable for model h(x), X,Y,Z independent 63

Table 4.5 Significance Test of Time Dependent Variable for model h(x,y), X,Y,Z strongly dependent...63

Table 4.6 Significance Test of Time Dependent Variable for model h(x,y), X,Y,Z strongly dependent...63

Chapter 1

Introduction

1.1 Research Objectives

This thesis is a mathematical support for the European Union project entitled BENERIS. BENERIS project focuses on the analysis of health benefits and risks associated with food consumption. It aims to develop a comprehensive method that combines both the dose-response modeling and a user-friendly graphical model interface. Results of this thesis may be used in the BENERIS project in the future.

This thesis focuses on the Cox hazard proportional model. It was proposed by D.R. Cox in 1972 ([5]). The Cox model is a well-recognized statistical technique to explore the relationship between the survival of a subject to some hazardous event of research interest such as death and several explanatory variables. In this model the hazard rate (failure rate) is regressed onto covariates which have influence on the survival time to some hazardous event, such as death. Moreover, the Cox model is built on the proportional assumption that the cumulative hazard function over time can be factorized into time dependent part which describes how hazard (risk) changes over time and time independent part which describes how hazard relates to other factors. The time dependent part is expressed by the baseline hazard function. No particular shape is assumed for the baseline hazard and it is estimated by nonparametric estimation. Time independent part consists of explanatory covariates, and the coefficients of these covariates are estimated by maximizing the partial likelihood.

The Cox proportional hazard model has been widely used in survival statistics and software such as *SAS, R, STATA* and *Splus* make it easier to develop the Cox model, in other words, to find the relationship between hazard or survival of a subject and explanatory covariates. However, a very important question should be answered well before the model can be used for prediction or adopted in any other fields. The question is how well the model has explained the hazard data? Here comes the issue of model adequacy test. A well-fitted model should provide an adequate summary of the data it is based on and a model adequacy test consists of two vital parts. One part is the examination of the model's fit and the other is to check the adherence to model assumptions.

This procedure is as important as model development and the aim of this thesis is to study on different methods used for model adequacy test.

1.2 Previous Works

An extensive survey of the research in this area identifies several contributions. For instance, a theoretical explanation that missing covariates tend to underestimate the model coefficient was given by Bretagnolle and Huber-Carol (1988) ([13]) as well as Keiding et al. (1997) ([15]). Hougaard (2000) pointed out that when covariates are excluded from the true model, proportional hazard assumption may not be satisfied any more ([11]). Moreover, several effective methods have been developed to examine the model's fit ([1]). They assumed that there is a true model and instead of using real life data, they generate sample data from this true model. To verify that the true model fits better to the data than others, one way they proposed is to estimate the coefficient of the same model covariate for both the true model and some incomplete models where covariates are excluded, base on the data generated from the true model. This is done independently 100 times. Then the 100 estimates were put in order and the ordered estimates for both the true model and model of missing covariates are plotted in the same graph for comparison. If these estimates are significantly different, it implies that the missing covariates are influential and should be included in the model. This method has already been applied to independent covariates and the results imply that missing covariates lead to under-estimation. Moreover, the larger the coefficient of the missing covariate is, the more the under-estimation it leads to.

Another method proposed in [1] is through comparison between population cumulative hazard function and the estimated baseline cumulative hazard function. A null hypothesis that these two hazard functions are equal is introduced. If this null hypothesis cannot be rejected, then it means that the baseline hazard and population hazard are not significantly different and covariates are not influential. In this case Cox model is not indicated. This method works well when applied to independent covariates, where this null hypothesis fails to be rejected for the incomplete model in the examples in [1].

However, when these covariates are dependent, or when interaction and quadratic terms are considered, results for these methods may be different. Moreover, the adherence to model assumptions should also be tested in some effective way. These issues are what we studied in this thesis.

1.3 Thesis Outline

As a start we give an explanation of the Cox model, what is a model adequacy test and how we make such test in chapter 2. We also illustrate the procedure to generate data for both independent and dependent covariates in the same chapter.

In chapter 3 we continue with previous work through applying the two methods mentioned in section 1.2 to dependent covariates and the case where interaction terms and quadratic terms are considered. We also make a distinction between strong dependence and weak dependence and study each of them.

In chapter 4 first we study on the likelihood ratio test. This is another method to test model's fit. In this test two times of the difference between logarithms of partial likelihood of true model and some incomplete model is calculated. Then we decide by chi-square test to see if this difference is significantly large and if true model is better than the incomplete model with covariates missing. Then we focus on test for model assumptions. First we include time dependent variables into the model. If the estimates of the coefficient are different for the model with and without the time dependent variable, then the time dependent variable is influential and should be included in the model. This is in conflict with proportional assumption that all covariates are time independent. We also perform Wald test to check if those time dependent variables are significant or not. Furthermore we use martingale residuals and Lowess smooth to check if the assumption that covariates appear in Cox model in the linear form is satisfied or not. Martingale is the difference between the censoring variable and estimated cumulative hazard. We assume there is no censoring so censoring variable is equal to one for all subject. In this test we first exclude a covariate and fit the model. Then results are used to calculate martingale residual which is plotted against the excluded covariate. The shape in this plot indicates the form of this covariate in the Cox model. Hence if it is not linear model assumption is not obeyed. Lowess is a smoothing method for scattered plot via local regression. An improvement of this method is also discussed in this section. As before, in this chapter we sample data from the assumed true model instead of using real data. Moreover, we apply all these methods to independence covariates, weakly dependent covariates and strongly dependent covariates to see if these methods work well in each situation.

Chapter 2

Testing adequacy of Cox proportional hazard model with simulations

As mentioned in the introduction chapter, the main object of this thesis is to test model adequacy for Cox hazard model. Moreover, in this study, we generate data instead of using real-life data. As a start, in this chapter we give a discussion about what is Cox proportional hazard model, the idea of model adequacy test, as well we the procedure of simulation for both independent covariates and dependent covariates.

2.1 Cox proportional hazard model

Cox proportional hazard model was introduced by D.R.Cox in 1972. In this model, failure rate, or hazard rate is regressed onto explanatory variables. The Cox model is built on the proportional hazard assumption that the cumulative hazard function over time can be factorized into two parts: a time dependent part and a time independent part. The time dependent part describe how hazard (risk) changes over time, which is independent of all model covariates. Time independent part which depends on model covariates describes how hazard relates to other factors. The Cox proportional hazard model we mainly used in our study consists of three covariates *X, Y* and *Z* and is of the form below.

$$h(t, X, Y, Z) = h_0(t)e^{AX+BY+CZ}$$

h (X,Y,Z) is the cumulative hazard function and $h_0(t)$ is the cumulative baseline hazard function of the form

$$h_0(t) = \int_0^t \lambda_0(u)du$$

where λ_0 is baseline hazard rate.

Main assumptions for this project are listed as follows.

1. Proportional assumption. All model covariates are assumed to be time independent.

2. Linear assumption. We assume all covariates appear in the Cox model in the linear form, or in other words, we assume that log-hazard function is a linear combination of covariates.

3. No ties. We assume that there is only one possible death at any time point or in other words we exclude the case where more than one subjects could die at the same time.

4. No censoring. It is assumed that the event of interest occurs for every subject during the study period, or in other words, we are aware of the survival time for all subjects.

The cumulative baseline hazard function as well as coefficients *A, B* and *C* can be estimated when fitting the Cox model to our data. Moreover, the population cumulative hazard can be estimated directly from data. We will give a detailed explanation how these items are estimated in the following sections.

2.1.1 Estimation of Coefficients *A, B* and *C*

Cox (1972) proposed an expression he called a "partial likelihood function" which only depends on parameters of interest. In our case parameters of interest are *A, B* and *C* and the partial likelihood function is time independent. In the present setting, the partial likelihood function is given by

$$\prod_{i=1}^{N} \frac{e^{x_i A + y_i B + z_i C}}{\sum_{t_j \ge t_i} e^{x_j A + y_j B + z_j C}}.$$

The denominator is the sum over all subjects at risk at t_i, and coefficients *A, B* and *C* are estimated by maximizing the partial likelihood function or the log partial likelihood function which is given as

$$\log(\prod_{i=1}^{N} \frac{e^{x_i A + y_i B + z_i C}}{\sum_{t_j \ge t_i} e^{x_j A + y_j B + z_j C}}) = \sum_{i=1}^{N} (x_i A + y_i B + z_i C - \log(\sum_{t_j \ge t_i} e^{x_j A + y_j B + z_j C}))$$

The maximum partial likelihood estimator for the coefficients can be obtained using the Matlab command *coxphfit*, which solves the problem through calculations of the first and second derivatives of log partial likelihood functions with respect to parameters *A, B* and *C.*

2.1.2 Estimation of Cumulative Baseline Hazard Function

As mentioned in [1], if cumulative hazard function $h(X,Y,Z)=h_0(t)e^{AX+BY+CZ}$ holds then the survival function for an individual with covariate values (x,y,z) is

$$S(t)=e^{-h(x,y,z)}=e^{-h_0(t)\exp(Ax+By+Cz)}=(e^{-h_0(t)})^{\exp(Ax+By+Cz)}$$

Denote $S_0(t)=e^{-h_0(t)}$ then

$$S(t)=(S_0(t))^{\exp(Ax+By+Cz)}$$

Therefore,

$$\frac{S(t_i,x)}{S(t_{i-1},x)}=\frac{(S_0(t_i))^{\exp(Ax+By+Cz)}}{(S_0(t_{i-1}))^{\exp(Ax+By+Cz)}}=(\frac{S_0(t_i)}{S_0(t_{i-1})})^{\exp(Ax+By+Cz)}$$

Define the conditional baseline survival probability as

$$\alpha_i=\frac{S_0(t_i)}{S_0(t_{i-1})}$$

Then

$$\frac{S(t_i,x)}{S(t_{i-1},x)}=\alpha_i^{\exp(Ax+By+Cz)}$$

Denote

$$\hat{\theta}_i=\exp(x_i\, y_i z_i)(\hat{A}\,\hat{B}\,\hat{C})'$$

Note that $\frac{\hat{\theta}_i}{\sum_{t_j>t_i}\hat{\theta}_j}$ is the probability of hazard for subject x_i at time t_i and $1-\frac{\hat{\theta}_i}{\sum_{t_j>t_i}\hat{\theta}_j}$ is the probability of survival for subject x_i at time t_i. Since

$$S(t_i,x_i)=\prod_{j\le i}\Pr(x_i \quad survive \quad at \quad t_j)$$

we have

$$1-\frac{\hat{\theta}_i}{\sum\limits_{t_j>t_i}\hat{\theta}_j}=\frac{S(t_i,x_i)}{S(t_{i-1},x_i)}=\alpha_i^{\exp(Ax_i+By_i+Cz_i)}$$

The estimated conditional baseline survival probability $\hat{\alpha}_i$ can be solved from the expression above or through its transformation

$$\frac{\hat{\theta}_i}{1-\alpha_i^{\exp(Ax_i+By_i+Cz_i)}}=\sum_{t_j>t_i}\hat{\theta}_j$$

Then the estimated baseline hazard rate is 1-$\hat{\alpha}_i$ and the estimated cumulative baseline hazard function at t_i is the sum of the estimated baseline hazard rate until t_i.

2.1.3 Estimation of Population Cumulative Hazard Function

The Nelson Aalen estimator is used for the population cumulative hazard function. We denote the rank-ordered survival times as $t_{(1)}<t_{(2)}<\ldots<t_{(N)}$. Let the number at risk of dying at $t_{(i)}$ be denoted as n_i and the observed number of deaths be denoted as d_i . Then the population cumulative hazard rate at time $t_{(i)}$ is $\frac{d_i}{n_i}$. Since we do not consider ties, d_i is always equal to one, so in our case the population cumulative hazard rate is just $\frac{1}{n_i}$, and the estimated population cumulative hazard function at t_i is the sum of the population cumulative hazard rate until t_i.

2.2 Model Adequacy Test

The motivation to perform model adequacy test comes from the requirement that the fitted model should provide an adequate summary of the data upon which it is based. In this sense, a complete and thorough examination of a model's fit and adherence to model assumptions is as important as model development.

In our study, we assume the true model to be *h(x,y,z)* and generate sample data from this true model. Furthermore, two other models *h(x,y)* and *h(x)* are also considered, from which some covariates are excluded. We fit these two models with missing covariates to the data generated by the true model. In order to verify that model *h(x,y,z)* is the true model or the best model, we should prove that the two incomplete models where covariates are missing are not proper for the data, either that these models do not fit well to the data or model assumptions are violated. To examine the model's fit, we could compare the estimates of coefficients for true model and the incomplete model, or compare the population cumulative hazard function with the estimated baseline hazard function for all these models. We can also use the likelihood ratio test to see if the true model is significantly better than the others. Martingale residual and Lowess smooth can be used to test adherence to the linear assumption for covariates, while we can include a time dependent variable to test if covariates are time independent. All these methods mentioned above are further studied in detail in chapter 3 and chapter 4 of this thesis.

In this thesis, we consider three forms of the true model. The first one is without interaction terms or quadratic terms; the second one is with interaction terms but without quadratic terms; the third is with both interaction and quadratic terms.

2.3 Data Simulation

2.3.1 Generating Samples of Data for Independent Covariates

Instead of using real-life data we generate sample data for both model covariates and survival time from the true model used in the thesis. Data for all covariates can be sampled from specified joint distribution. Then data for survival times could be generated using sample data for model covariates and relation $T \sim -\ln(U)/e^{AX+BY+CZ}$. Now we give a detailed explanation of this expression of relation between survival time and model covariates.

Cox model used in this thesis is of the following form as mentioned in section 2.1:

$$h(X,Y,Z) = h_0(t)e^{AX+BY+CZ}$$

where $h_0(t) = \int_0^t \lambda_0(u)du$.

When the baseline hazard rate λ_0 is constant and scaled to one we have that $h_0(t)=t$. Moreover, when (x,y,z) fixed and t random the survival function

$$S(t) = \exp(-h_0(t)e^{Ax+By+Cz}) = \exp(-te^{Ax+By+Cz})$$

, as a function of t, is uniformly distributed on [0,1] . Furthermore, when $t=0$, $S(t)=1$, when t tends to infinity, $S(t)$ tends to zero.

Therefore we have

$$\begin{aligned}
&\exp(-te^{Ax+By+Cz}) \sim U[0,1] \Leftrightarrow \\
&te^{Ax+By+Cz} \sim -\ln(U) \Leftrightarrow \\
&t \sim -\ln(U)/e^{Ax+By+Cz}
\end{aligned}$$

This holds for each subject (sample) and we can rewrite it as

$$T \sim -\ln(U) / e^{AX+BY+CZ}.$$

Now we could summarize the steps of generating data for independent covariates which is on the basis of discussion above:

1. Specify values of *A B* and *C*, scale baseline hazard rate to one and choose a distribution for *(X,Y,Z)*.

2. Sample 100 values for *(X,Y,Z)* since we consider 100 subjects in our study and 100 values from uniform distribution on [0,1]. Then use $T \sim -\ln(U) / e^{AX+BY+CZ}$ to calculate corresponding value of survival time for each subject.

After data is generated coefficients and baseline hazard can be estimated using the methods discussed in section 2.1.

2.3.2 Generating Samples of Data for dependent Covariates

With the help of normal transformation, we can sample data from arbitrary marginal distributions with specified rank correlation matrix. Procedure of sampling dependent data in this way is illustrated below.

1. Specify a rank correlation matrix *ρr*, and then apply Pearson Theorem to get the corresponding product moment correlation coefficient matrix *ρ*. Pearson Theorem is described as follows:

Let *(X, Y)* be random vectors with joint normal distribution then

$$\rho(i,j) = 2\sin(\frac{\pi}{6}\rho r(i,j))$$

Where $\rho(i,j)$ and $\rho r(i,j)$ are the *(i,j)th* cell of matrices ρ and ρr respectively.

Proof of Pearson Theorem is given in Appendix 1.

2. Let W_1, W_2, W_3 be independent standard normal variables and sample 100 times independently for (W_1, W_2, W_3). Apply Cholesky decomposition to *ρ*, denote $\rho=LL^T$ and then put *V=LW*, where $V=(V_1\ V_2\ V_3)'$and $W=(W_1\ W_2\ W_3)'$. Here $(V_1\ V_2\ V_3)$ is joint normally distributed with standard normal margins and rank correlation *ρr.*

3. Specify invertible marginal distribution F_1 F_2 and F_3. Moreover denote Φ as standard normal distribution function. For each of the 100 subjects, put $x_i=F_1^{-1}(\Phi(v_{1i}))$, $y_i=F_2^{-1}(\Phi(v_{2i}))$, $z_i=F_3^{-1}(\Phi(v_{3i}))$, $i=1,\ldots,100$. Then *(X,Y,Z)* has marginal distribution F_1 F_2 and F_3 and rank correlation matrix *ρr.*

Chapter 3

Further Study of Previous Methods

In this part, we give a review and further study on two methods put forward in [1] to test model adequacy based on data simulation discussed in section 2.3.

3.1 Comparison between Ordered Estimated Coefficients

3.1.1 Review of the Method

The first method to test model adequacy we study in this thesis is to fit both the true model and the model with missing covariates to the data generated by the true complete model 100 times ([1]). Then the 100 ordered estimates of model coefficient are plotted for comparison. If the estimates of a coefficient for the incomplete model (with missing covariates) is significantly different from the estimates of the same coefficient for the true complete model, it implies that the missing covariates are significantly influential and should not be excluded from the true model, or in other words, the incomplete model for which some covariates are excluded is wrong. Here three models are compared: true model, denoted as *h(x,y,z)*, the model for which covariate *Z* is excluded, denoted as *h(x,y)*, as well as the model where both *Y* and *Z* are excluded, denoted as *h(x).* All these three models were fit 100 times to the data generated by the true model and the resulting 100 ordered estimates of the coefficient of *X*, i.e. *A,* for all models were plotted into the same graph for comparison.

This method has been proved effective when applied to independent covariates without interaction or quadratic terms ([1]). Figure 3.1 is the plot of 100 ordered estimates of *A* for model *h(x,y,z) h(x,y)* and *h(x)* where *(A,B,C)=(1,1,1)* while in figure 3.2 coefficient *C,* which is the coefficient of missing covariate *Z,* increases from 1 to 5. From these two graphs we see that missing covariates result in under-estimation. Moreover when the coefficient of missing covariate is larger, the under-estimation is more pronounced.

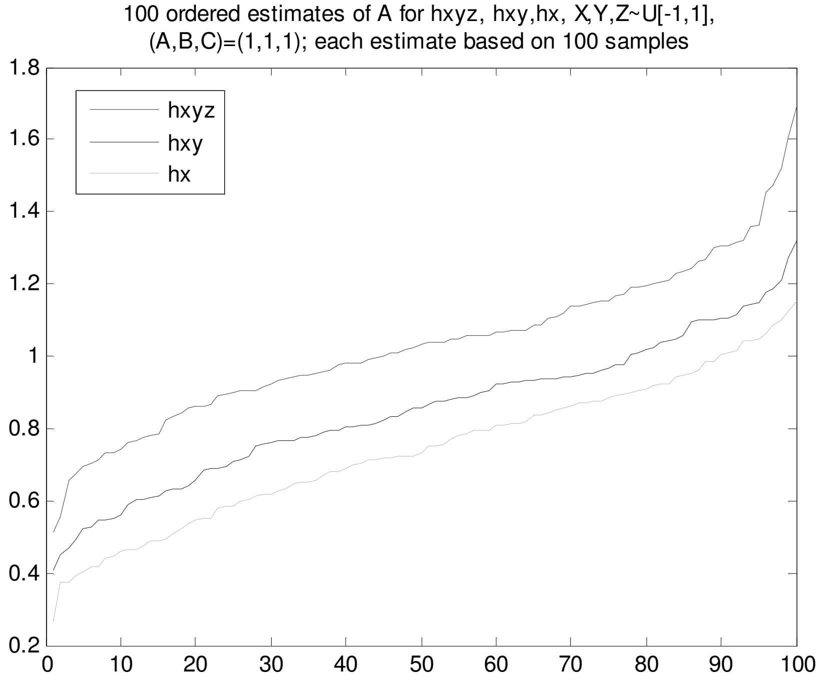


Figure 3.1 100 ordered estimates of A for hxyz, hxy, hx, independent, C=1

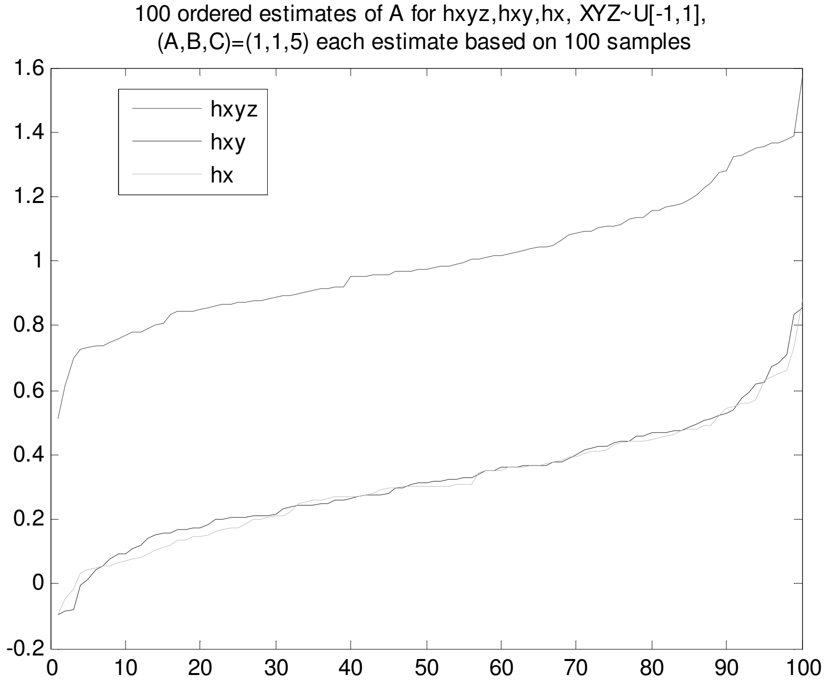


Figure 3.2 100 ordered estimates of A for hxyz, hxy, hx, independent, C=5

Now we continue with a further study of this method. First we apply this method to dependent covariates. Then interaction terms and quadratic terms are discussed.

3.1.2 Study of Dependence

As a start, let us look at the situation where covariate *X* is dependent on *Y* and *Z* to check whether the estimated coefficient of *X* will be influenced by such dependence.

Figure 3.3 is the graph of 100 ordered estimates of *A* for *h(x,y,z)*, *h(x,y)*, and *h(x)*, where the rank correlations between *X* and *Z* is 0.9 and the rank correlation between *X* and *Y* is 0.4. It is suggested not to set the rank correlation between each pair of covariates all very high, in order to avoid the situation that after Pearson Transformation the product moment correlation matrix is not positive definite and Cholesky decomposition cannot be performed. In this example, we put zero to the rank correlation between *Y* and *Z.*

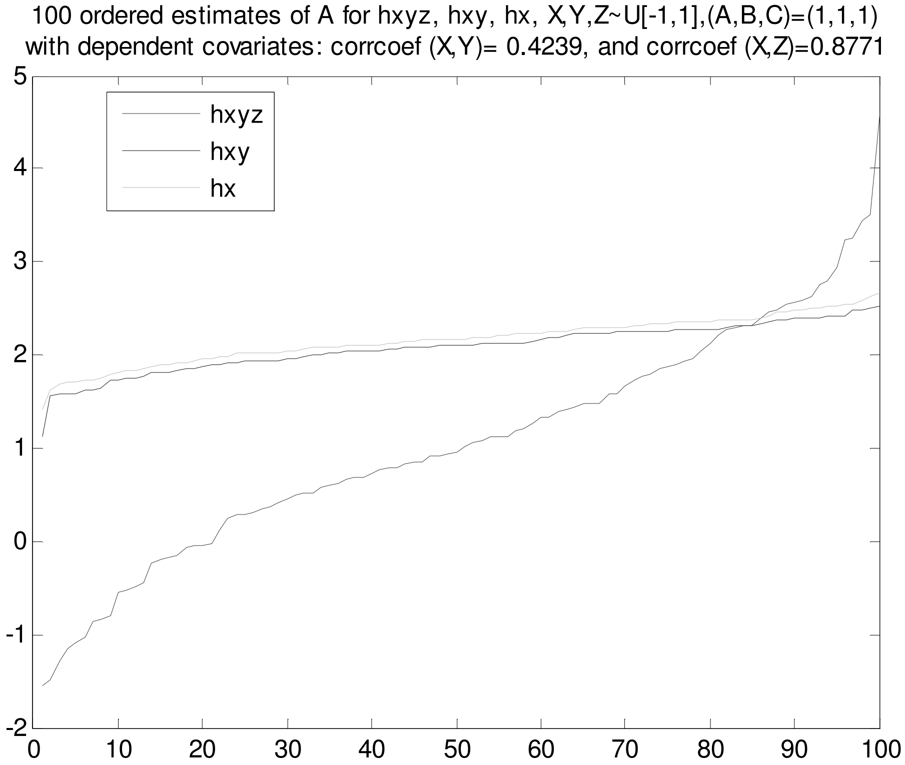


Figure 3.3 100 ordered estimates, X strongly dependent on Y, Z, C=1, rank correlation of Y and Z is zero

As shown in the title of figure 3.3, *corrcoef* is short for product moment correlation coefficient, which depends on both the specified rank correlation matrix and sampling. From this plot, we see that unlike the case for independent covariates where incomplete model tends to under-estimate the coefficient, as shown in figure 3.1 and 3.2, in this case both incomplete models over-estimate the value of coefficient, which implies the significant influence of missing covariate and incorrectness of the incomplete models.

Now we give two examples where rank correlation between *Y* and *Z* are not zero. Results are demonstrated in figure 3.4 and figure 3.5. Figure 3.4 is for the situation where *Y* and *Z* are weakly correlated while figure 3.5 is a result of the case that *Y* and *Z* are stronger correlated.

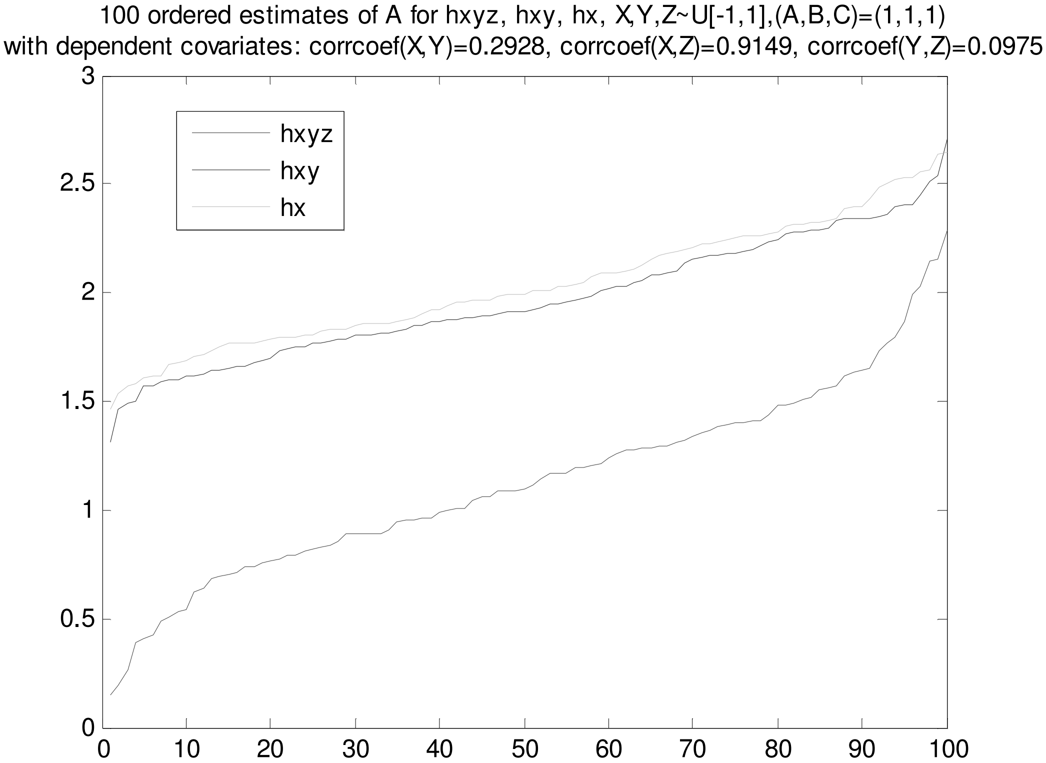


Figure 3.4 100 ordered estimates, X and Z, X and Y strongly dependent, Y and Z weakly correlated, C=1

From both figure 3.4 and figure 3.5 we see that when *Y* and *Z* are also rank correlated with each other, the ordered estimates for incomplete models deviate more from the ordered estimates for the true model. In other words, the coefficient is over-estimated by incomplete models to a greater extent. However, the similarity of the two graphs also tells us that how strongly or weakly *Y* and *Z* are correlated has little influence on the result.

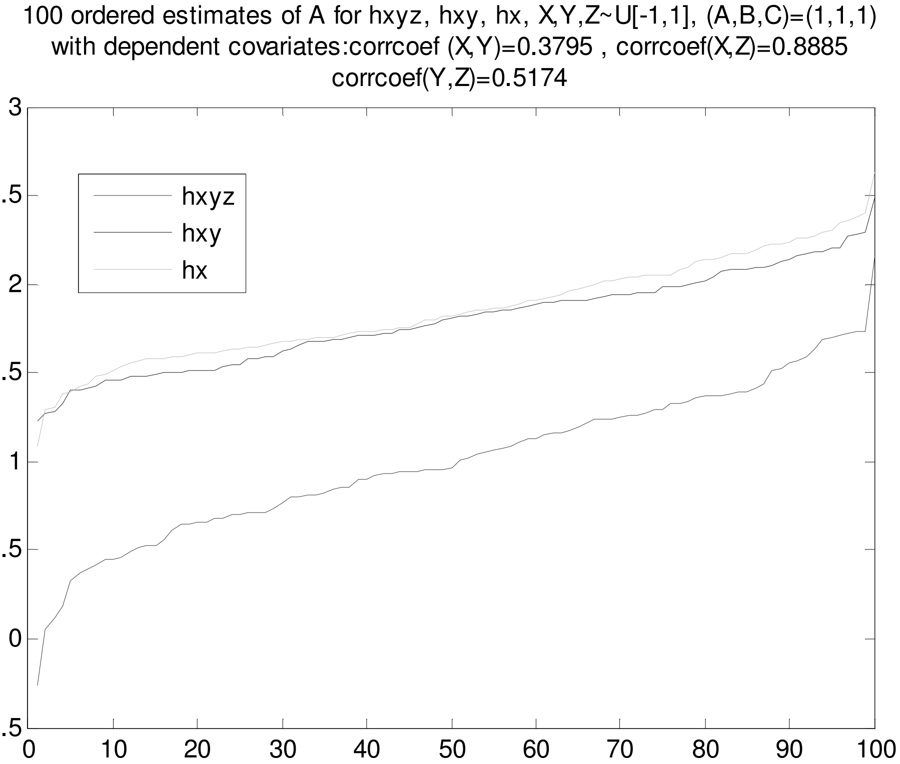


Figure 3.5 100 ordered estimates, X strongly dependent on Y and Z, Y and Z strongly correlated, C=1

However, while studying on the situations where all covariates are weakly correlated (here we put rank correlation for each pair of covariates 0.1), we notice from figure 3.6 that the incomplete models tend to under-estimate the coefficient, as in the cases of independent covariates, instead of over-estimation, as in the cases of strong dependence between *X* and the missing covariates.

Moreover, when compared figure 3.3 with figure 3.6, we notice that strong dependence leads to larger variance of estimates for the true model. Moreover, when covariates are strongly dependent, variance of estimates for the true model is larger than that of the incomplete models.

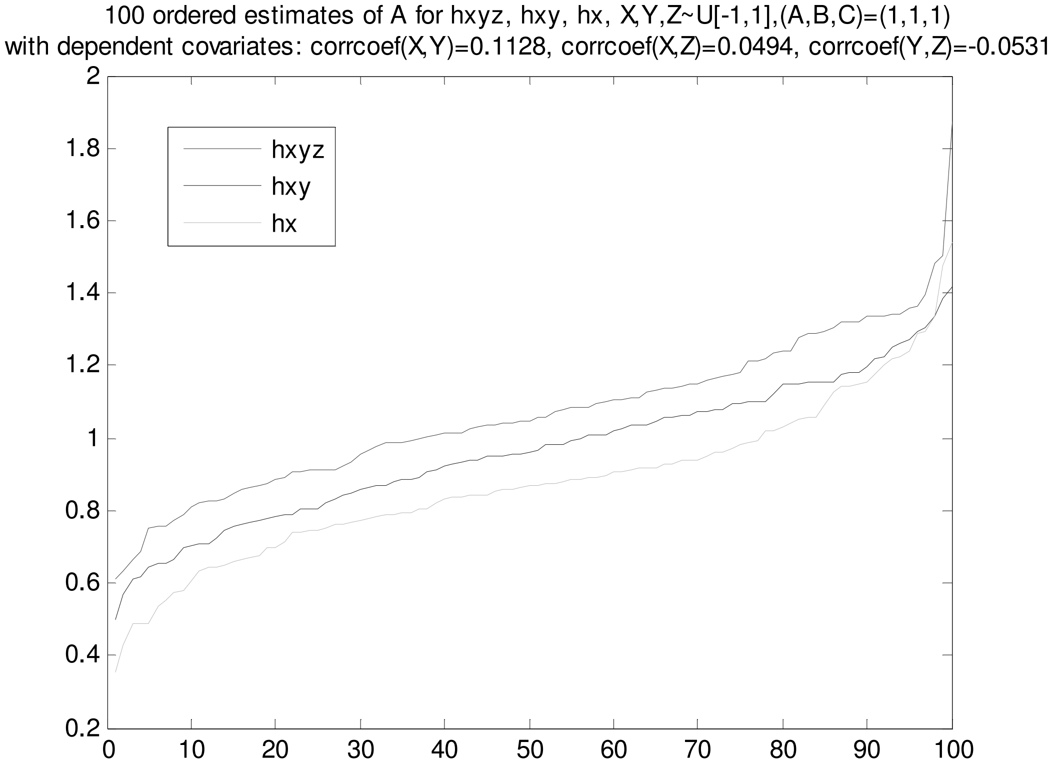


Figure 3.6 100 ordered estimates all covariates weakly dependent, C=1

As shown in figure 3.7 and figure 3.8, when we set the coefficient of missing covariate *Z* as *C=5* at the beginning of simulation, then in the case where *X* is strongly corrected with *Z* (here in figure 3.7 we still set rank correlation of *X* and *Z* as 0.9 and that of *X* and *Y* as 0.4), the over-estimation for incomplete models are more obvious. However, when all covariates are weakly dependent on each other, for instance when all rank correlation are 0.1, as shown in figure 3.8, then incomplete models tend to under-estimate the coefficient just like in the case of independent covariates, and here under-estimation is more obvious than in the case when *C=1* and all covariates of the same rank correlation (which is the case illustrated in the figure 3.6).

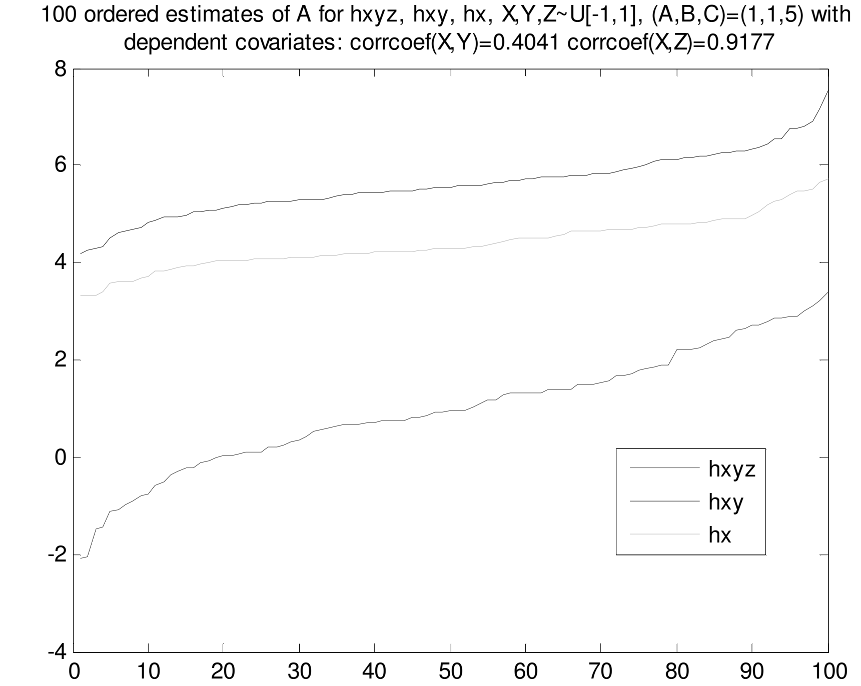


Figure 3.7 100 ordered estimates of A for hxyz, hxy, hx, strong dependence, C=5

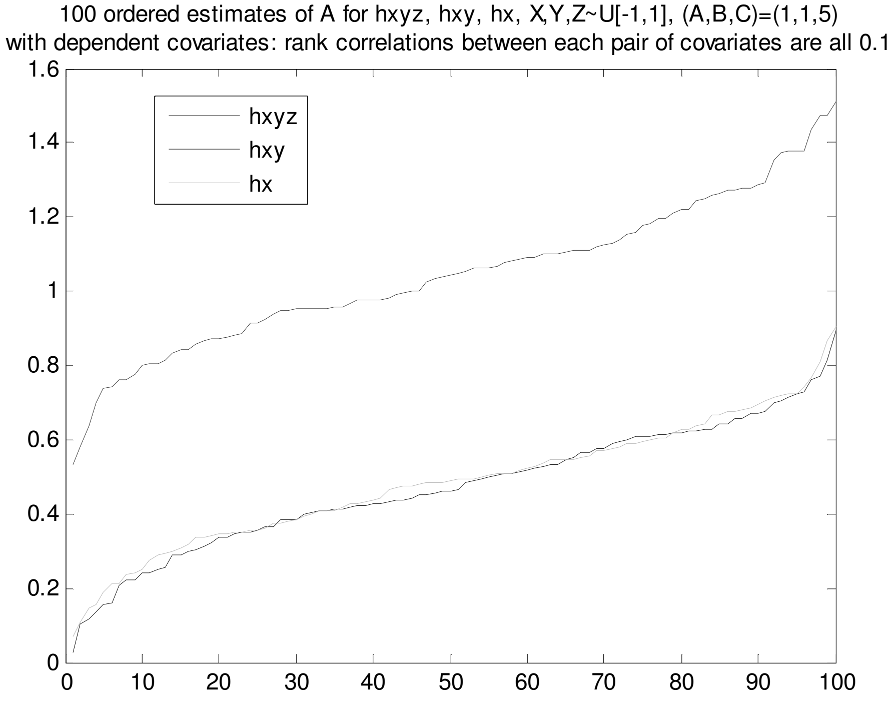


Figure 3.8 100 ordered estimates of A for hxyz, hxy, hx, weak dependence, C=5

Now we have a look at the influence of negative dependence. From figure 3.9 we see that when *X* and missing covariate *Z* are negatively and strongly dependent, missing covariates results in under-estimation of coefficient of *X*. This is different from the situation when they are strongly but positively dependent as shown in figure 3.7 (rank correlation matrices we set in figure 3.7 and 3.9 are only different in the sign). Hence in the case of strong dependence, the sign of correlation between covariates, or in other words, whether covariates are positively or negatively dependent is also influential to the estimates.

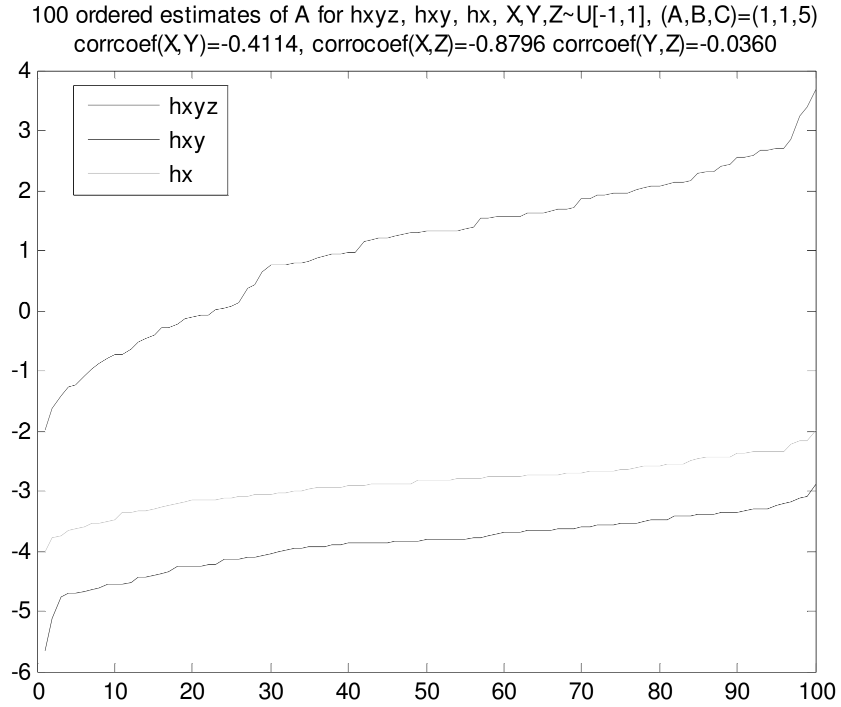


Figure 3.9 Negative and strong dependence

As a contrast, let us look at figure 3.10. Here covariates are negatively and weakly dependent. In this case rank correlation between covariates only different from figure 3.8 in the sign. From this plot we see that missing covariates results in under-estimation. Note that this is the same as in figure 3.8 where covariates are positively dependent. Therefore for weakly dependent covariates whether they are positively or negatively dependent is not influential to the estimates.

When comparing figure 3.9 and figure 3.10 we also notice that for negatively dependent covariates, strong dependence leads to more profound under-estimation and larger variance of estimates for the true model.

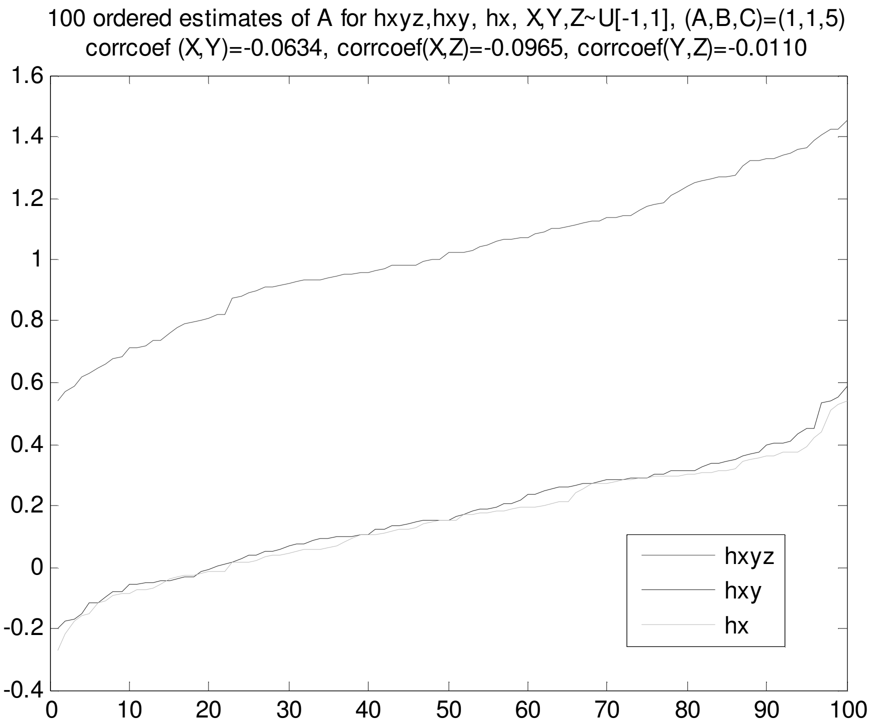


Figure3.10 Negative and weak dependence

3.1.3 Study of Interaction and Quadratic terms

Let us move on to study the interaction and quadratic terms. First we assume that there are no such terms in the true model as in the previous cases and check if such terms are influential in the model or not via the same method as before. Then we assume that the true model contains such terms and give a similar discussion for model adequacy.

3.1.3.1 No Interaction or Quadratic Terms in the True Model

In this case we compare the ordered estimates of coefficient of *X* for the true model with those estimated for the models with extra term x^2, *xy*, *xz, xyz* respectively. We study both the cases where *C=1* and *C=5*. Moreover, various dependence situations are also studied. When all covariates are independent and *C=1*, result is illustrated in figure 3.11. From this figure we notice that all the extra interaction or quadratic terms are not significant for the Cox proportional hazard model since the estimates of *A* are almost the same for the models with and without those terms. Therefore these terms are not necessary to be included in the true model *h(x,y,z)*. In other situations results are similar and all the extra terms are not influential either. Plots for dependent covariates and for the case of *C=5* are shown in Appendix 2.

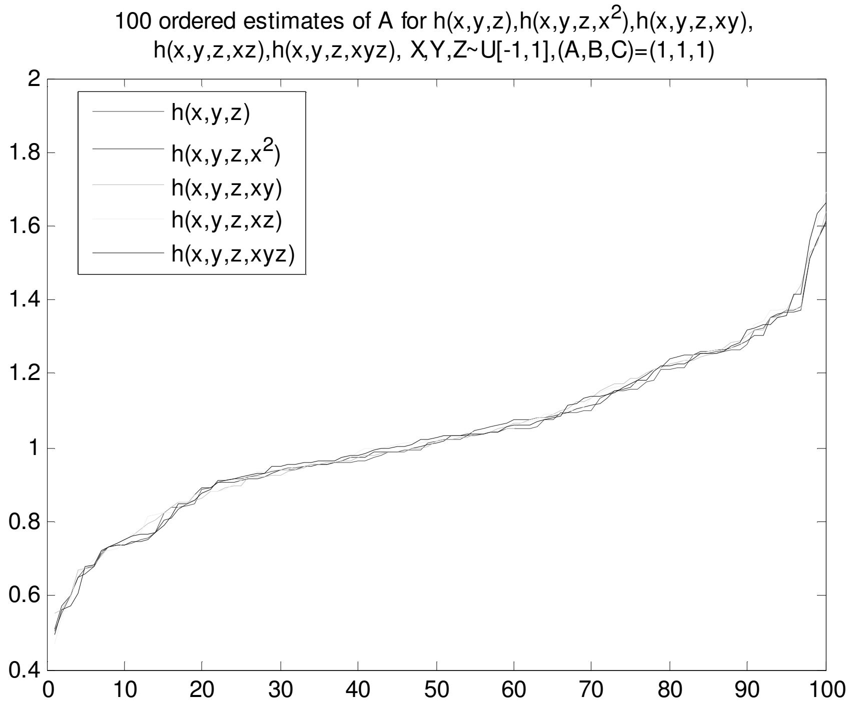


Figure 3.11 100 ordered estimates of A for the true model and models with extra interaction or quadratic terms

3.1.3.2 True Model with Interaction Terms

Now let us discuss about the case that the true model is assumed to have interaction terms. We follow the same way to test model adequacy, the only different is that here the three models for comparison are the true Cox proportional hazard model *h(x,y,z,xy,yz,xz)* and two incomplete models *h(x,y,xy)* and *h(x)*.

Figure 3.12 results from independent covariates. We see that although here interaction is considered, missing covariates still lead to under-estimation of the coefficient, like in the case where there is no interaction in the model. Figure 3.13 also shows similar results to the case of no interaction. In this plot, over-estimation can be seen due to strongly dependence between *X* and missing covariates. In both figures *C=5*.

Plots for the cases of weak dependence and *C=1* are demonstrated in Appendix 2. From these figures, as well as figure 3.12 and figure 3.13, we see that including interaction in the model results in similar performance of the estimates. That is, when covariates are independent or weakly dependent, missing covariates result in under-estimation of the coefficient, while for strong dependence, over-estimation will be the consequence. Moreover, over-estimation or under-estimation will be more significant if coefficient of missing covariate *Z* is larger.

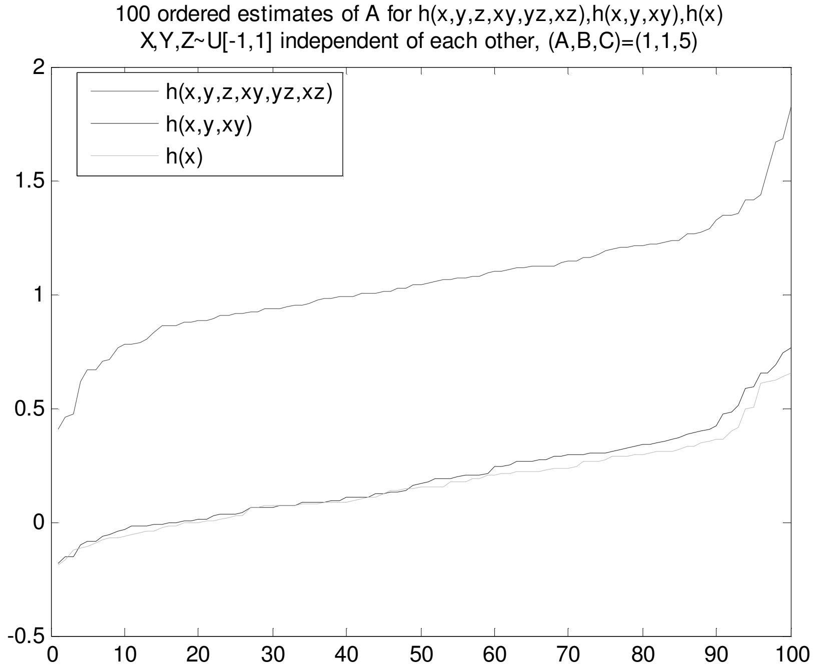


Figure 3.12 100 ordered estimates of A for h(x,y,z,xy,yz,xz),h(x,y,xy),h(x), independent, C=5

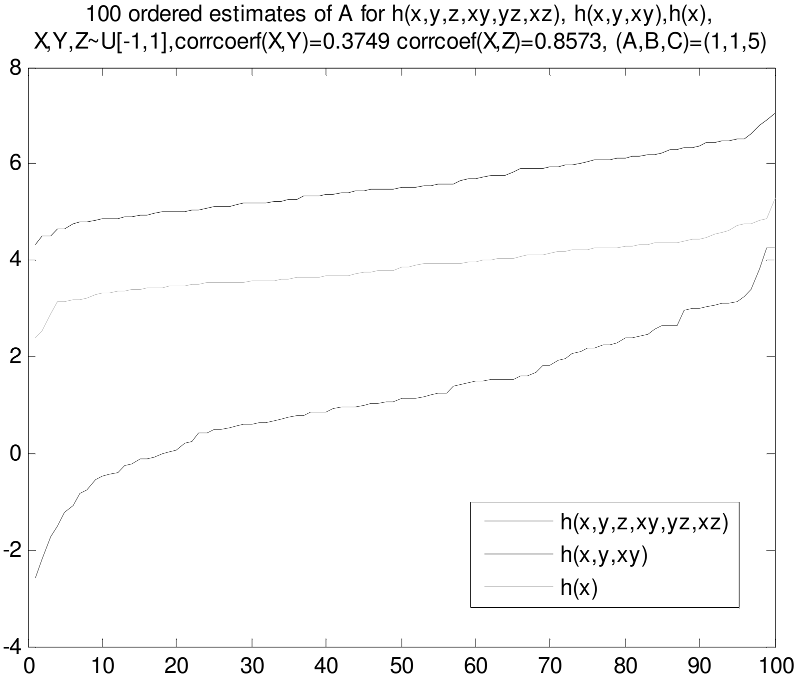


Figure 3.13 100 ordered estimates of A for h(x,y,z,xy,yz,xz),h(x,y,xy),h(x), strong dependence, C=5

3.1.3.3 True Model with Both Interaction Terms and Quadratic Terms

In this section, we consider both interaction and quadratic terms. In this case, the true model is *h(x,y,z,x^2,y^2,z^2,xy,yz,xz)*. Model adequacy test is performed through comparison between ordered estimates of coefficient *A* for true model and two incomplete models with missing covariates, *h(x,y,x^2,y^2,xy)* and *h(x,x^2)*. As before, different situations such as independence, strong dependence and weak dependence between covariates are studied and compared. We also vary the value of *C* which is the coefficient of missing covariate *Z*. Results for all these situations are similar to those obtained from the model with no interaction or quadratic terms and the model with only interaction terms, where missing covariates results in under-estimation if independent or weakly dependent, and over-estimation otherwise.

For instance let us look at figure 3.14. This plot results from strongly dependent covariates where product moment correlation coefficient between *X* and missing covariate *Z* is 0.9140. Over-estimation is obvious as a consequence. Note that none of the estimates of *A* for the two incomplete models is close to 1which is the value of *A*. Actually, as shown in the plot, estimates for model *h(x,x^2)*are all above 2, while estimates for model *h(x,y,x^2,y^2,xy)* are all above 4.

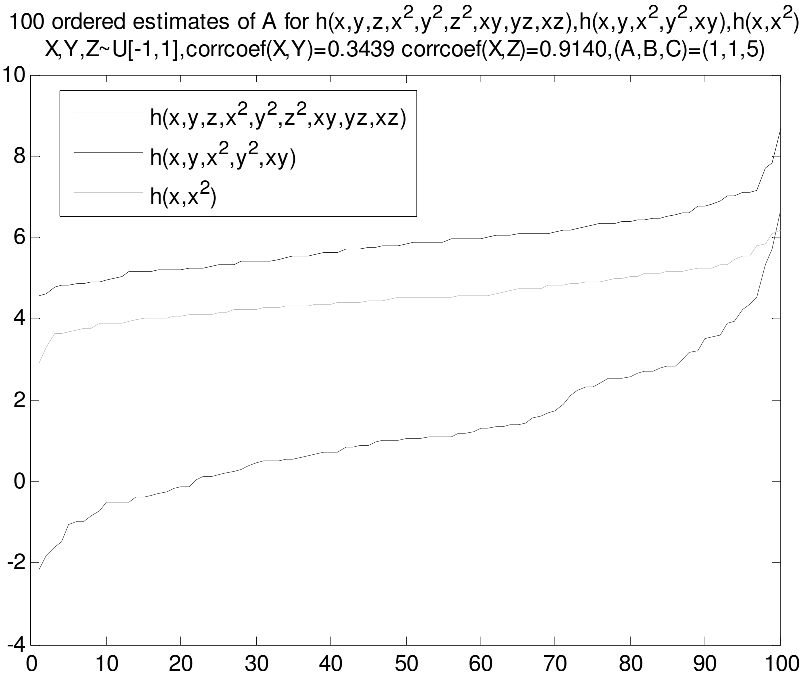


Figure 3.14 100 ordered estimates for model with both interaction and quadratic terms, strong dependence, C=5

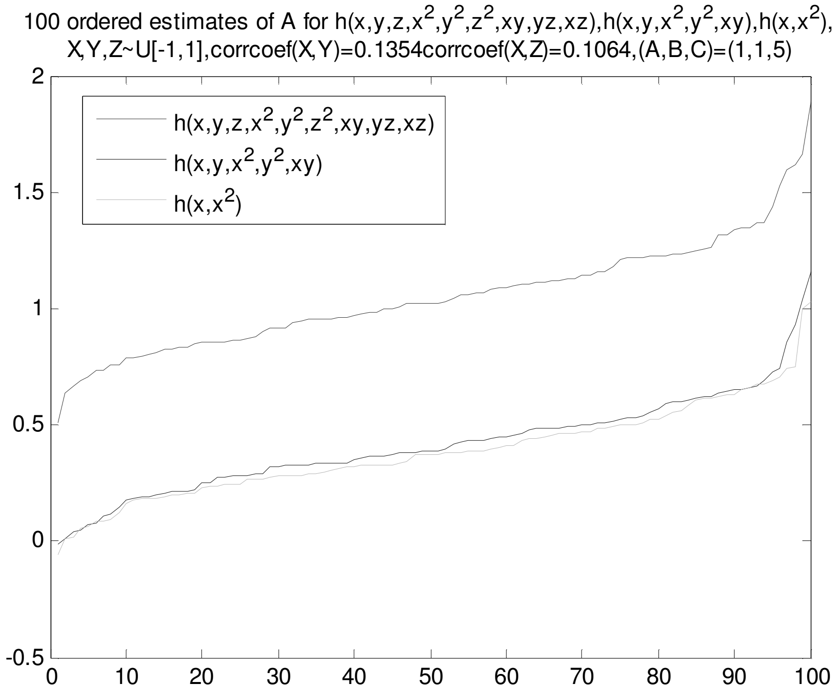


Figure 3.15 100 ordered estimates for model with both interaction and quadratic terms, weak dependence, C=5

Figure 3.15 implies under-estimation for weak dependence. Here, most of the estimates for the two models with missing covariates are below 1. In both figure 3.14 and figure 3.15 we have *C=5*. Plots for independent covariates and for the case of *C=1* are shown in Appendix 2.

Studies in this section imply that the method to test model adequacy through comparison between ordered estimates of coefficient works well for dependence covariates, for models with interaction terms and models with both interaction and quadratic terms.

3.2 Comparison between Estimated Cumulative Baseline Hazard Function and Population Cumulative Hazard Function

Another method used to test model adequacy is through comparison between estimated cumulative baseline hazard function and population cumulative hazard function ([1]). This is the main focus of this section.

3.2.1 Review of the Method

From the form of Cox proportional hazard model $h(X,Y,Z) = h_0(t)e^{AX+BY+CZ}$, we see that if all covariates are excluded, or all coefficients are equal to zero, then the cumulative baseline hazard function is equal to the population cumulative hazard function, which can be put as a null hypothesis to test model adequacy. If this hypothesis can be rejected, in other words, if the estimated cumulative baseline hazard function significantly deviates from the population cumulative hazard function, it implies that effect of covariates is significant and covariates should be included in the Cox model. Otherwise, the Cox model is not appropriate if this null hypothesis cannot be rejected. Two-sigma (variance) bands of population cumulative hazard function are applied to show whether the estimated cumulative baseline hazard function is significant different from the population cumulative hazard function. The variance estimator for the Nelson-Aalen estimator for the population cumulative hazard function used here is ([4], *P25*)

$$\hat{V}(t) = \sum_{t_j \le t} \frac{d_j(n_j - d_j)}{n_j^3}$$

where n_j and d_j are the same as defined in section 2.1.3. Since we assume no ties, $d_j \equiv 1$.

Applications of this method to independent covariates were already discussed ([1]). Results are shown in figure 3.16 and figure 3.17.

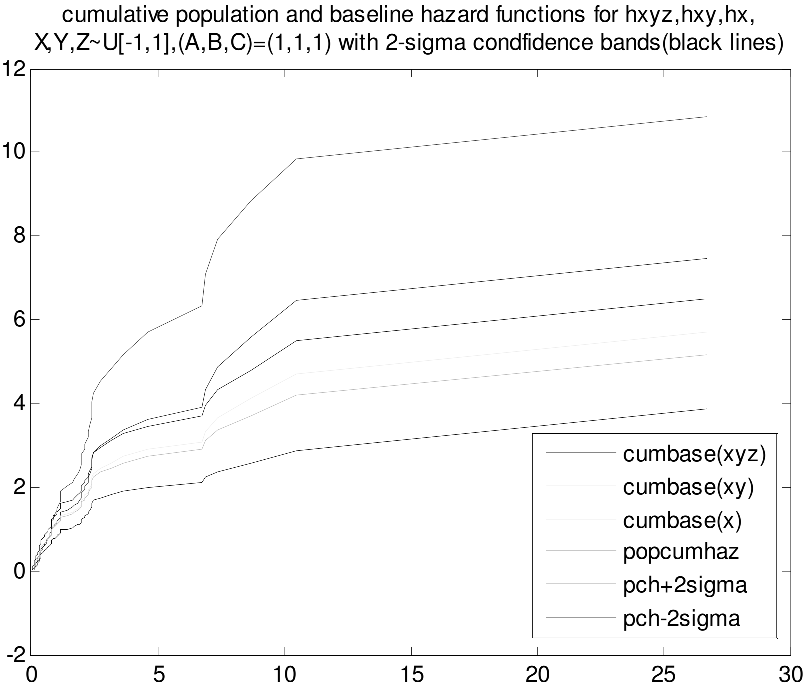


Figure 3.16 Cumulative population and baseline hazard, independent, C=1

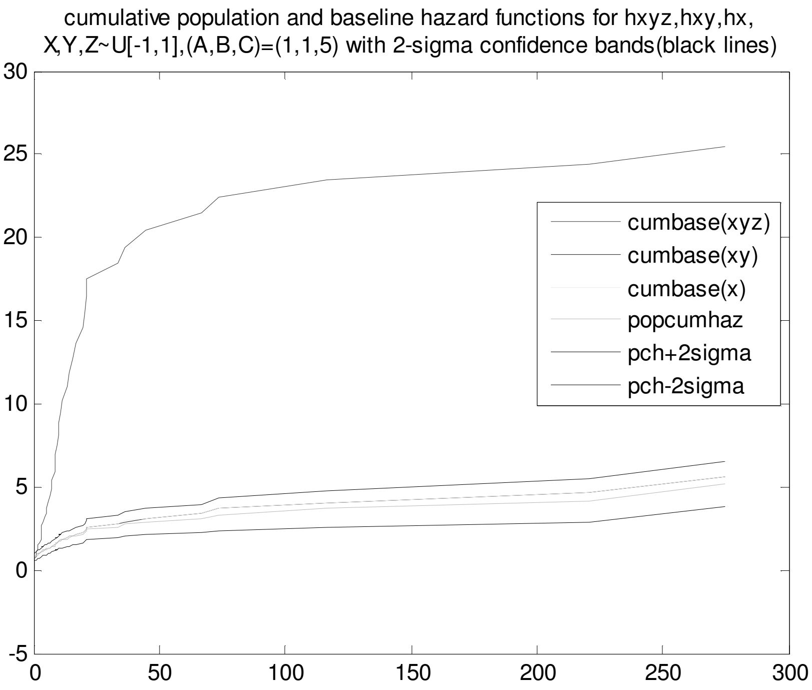


Figure 3.17 Cumulative population and baseline hazard, independent, C=5

From figure 3.16 and figure 3.17 we see that when *C=1*, the null hypothesis fails to be reject for model *h(x),* since the estimated cumulative baseline hazard for this model is within the 2-sigma bands of the population cumulative hazard; when *C=5,* the null hypothesis cannot be rejected for both *h(x,y)* and *h(x)* and the estimated cumulative baseline hazard for these two model is almost the same as the population cumulative hazard. As shown in the legend, *cumbase (xyz), cumbase (xy)* and *cumbase (x)* stand for the estimated cumulative baseline hazard for model *h(x,y,z), h(x,y)* and *h(x)* respectively. Moreover, *popcumhaz* stands for population cumulative hazard and pch+2sigma, pch-2sigma are 2-sigma bands for the population cumulative hazard.

In our thesis, we first extend this method to dependent covariate. Then we present a further study on the models with interaction terms. Here we use the same method as before when sampling data from dependent covariates.

3.2.2 Study of Dependence

3.2.2.1 Strong Dependence

Our discussion begins with the case that *X* is dependent on *Y* and *Z*. First, strong dependence between *X* and missing covariate *Z* is studied. We consider both the situations where *C=1* and *C=5* .Results are shown in figure 3.18 and figure 3.19 respectively.

Figure 3.18 results from the situation of *C=1*. Unlike figure 3.16, in this case the estimated baseline cumulative hazard functions for both incomplete models are close to that of the true model, instead of the population cumulative hazard function. Since the estimated cumulative baseline hazard for all models are different from the population cumulative hazard, the null hypothesis that the cumulative baseline hazard function is equal to the population cumulative hazard function can be rejected for all models.

Let us move on to the situation of *C=5*. As *C* increases from 1 to 5, when all covariates are independent, as shown in figure 3.17, the estimated cumulative baseline hazard functions for models *h(x,y)* and *h(x)* move closer to and actually is not significantly different from the population cumulative hazard function. However, from figure 3.19, we see that when *X* is highly correlated with the missing covariates, although the estimated cumulative baseline hazard functions for the two incomplete models are not so close to that function of the true model as in figure 3.18, they are still significantly different from the population cumulative hazard function. Therefore the null hypothesis can be rejected for both incomplete models.

Moreover, when *C=5* we notice from figure 3.19 the difference in values of survival time (the horizontal axis). This is due to the heavier loading of covariate *Z*.

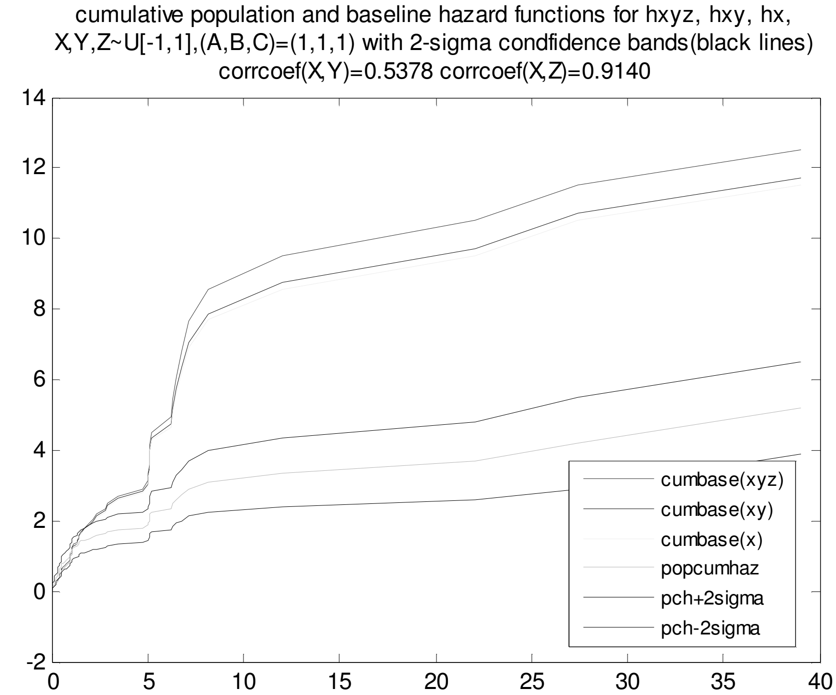


Figure 3.18 Cumulative population and baseline hazard, strong dependence, C=1

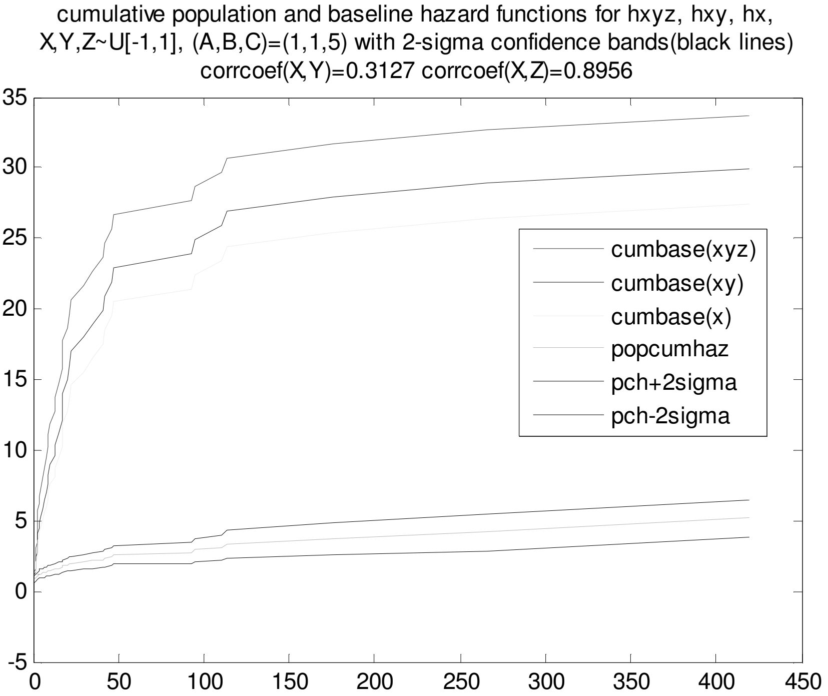


Figure 3.19 Cumulative population and baseline hazard, strong dependence, C=5

3.2.2.2 Weak Dependence

Furthermore, from figure 3.20 and figure 3.21 we see that when covariates are dependent, but weakly dependent, the resulting performances of the estimated cumulative baseline hazard functions for all models are similar to the case of independence. Figure 3.20 is the result of *C=1*, where the estimated cumulative baseline hazard functions for *h(x,y)* and *h(x)* are close to the population cumulative hazard function, unlike the case of strong dependence.

As *C* increases to 5, from figure 3.21 we see that the estimated cumulative baseline hazard functions for *h(x,y)* and *h(x)* are almost the same as the population cumulative hazard function, and the null hypothesis cannot be rejected for neither *h(x,y)* nor *h(x)*. This implies that using Cox model would not be indicated for these two models with missing covariates.

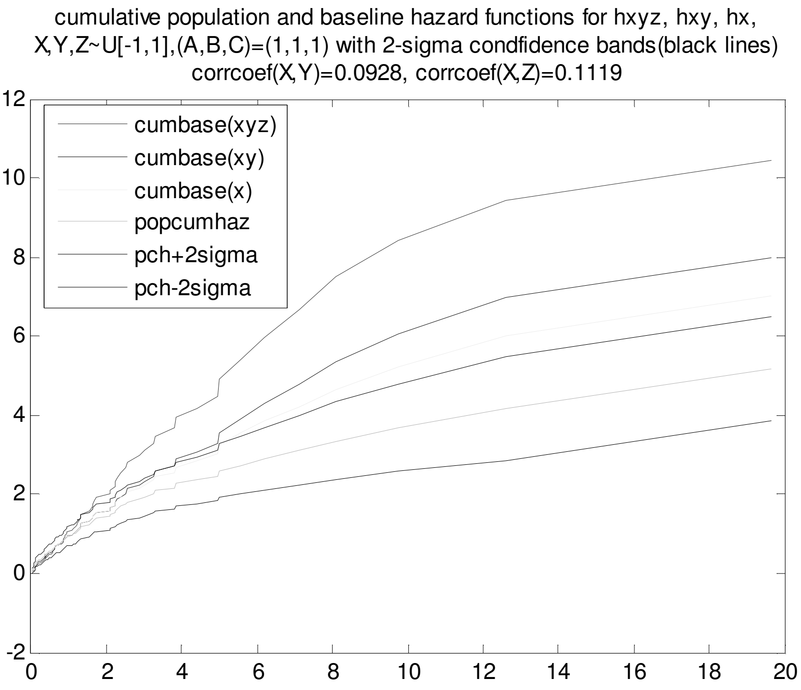


Figure 3.20 Cumulative population and baseline hazard, weak dependence, C=1

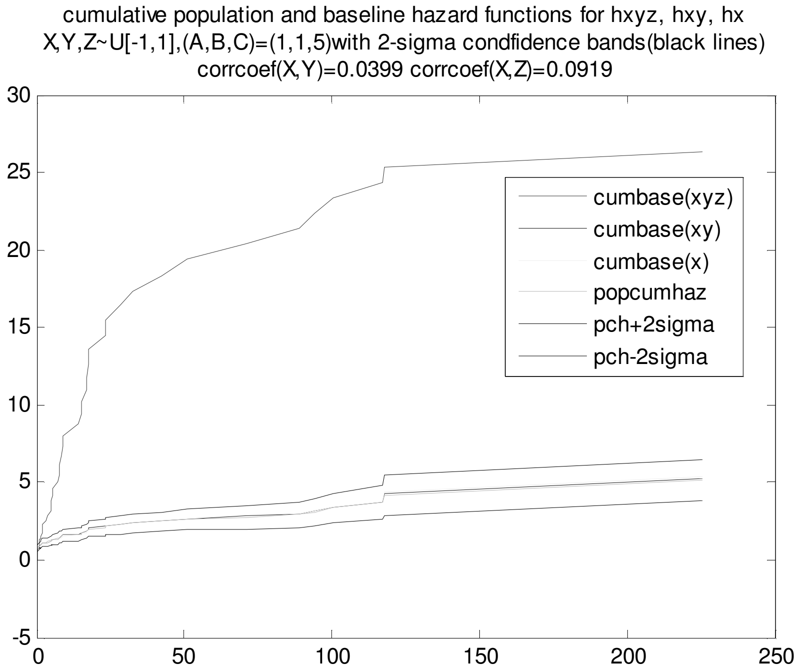


Figure 3.21 Cumulative population and baseline hazard, weak independence, C=5

3.2.3 Study of Interaction

In this section interaction terms are included in the model and the true model will be *h(x,y,z,xy,yz,xz)*. Furthermore, the two incomplete models with missing covariates for comparison are *h(x,y,xy)* and *h(x)*.

3.2.3.1 Covariates Independent of Each Other

We start with the assumption that all covariates are independent of each other. When *C=1,* as shown in figure 3.22, the estimated cumulative baseline hazard functions for *h(x)* is not different from the population cumulative hazard function and the null hypothesis fails to be rejected for model *h(x)*. Hence using the Cox model for this model with two missing covariates is not appropriate.

When *C=5*, as shown in figure 3.23, the null hypothesis fails to be rejected for both *h(x,y,xy)* and *h(x)* which means that Cox model is not indicated for neither of them.

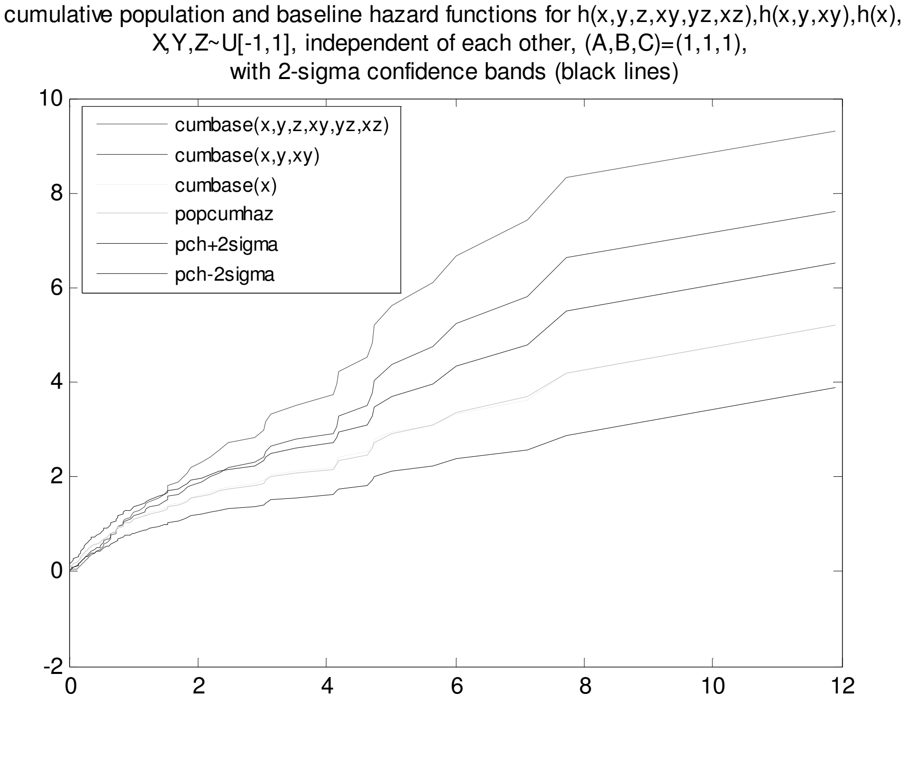


Figure 3.22 Cumulative population and baseline hazard for model with interaction, independent, C=1

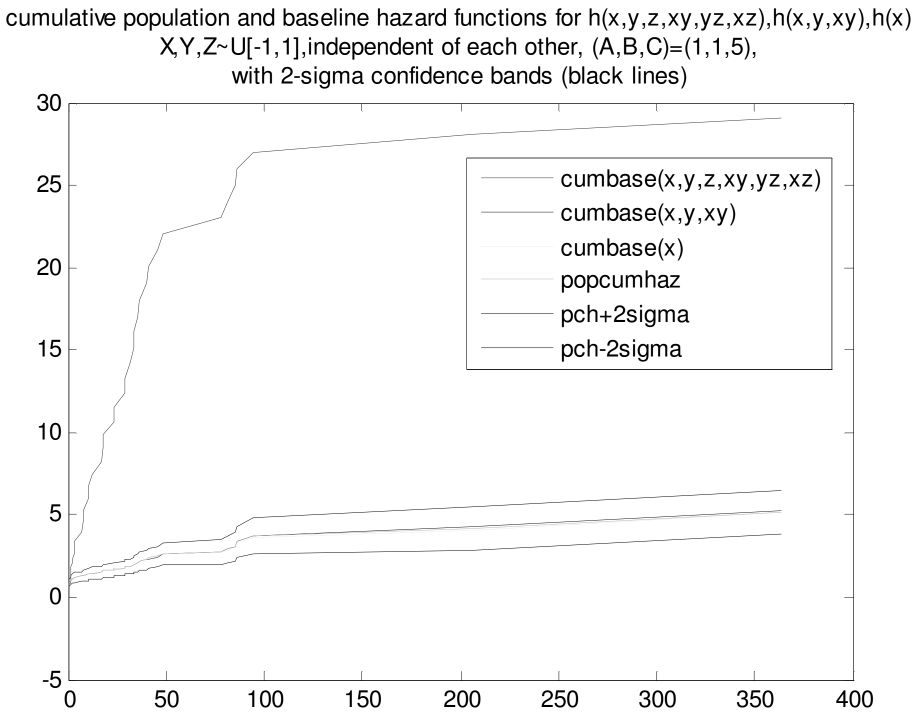


Figure 3.23 Cumulative population and baseline hazard for model with interaction, independent, C=5

3.2.3.2 Dependence Covariates

As shown in figure 3.24, when *X* is highly correlated with missing covariate *Z* the null hypothesis can be rejected for all models, since all of the estimated baseline cumulative hazard functions are significantly different from the population cumulative hazard function. However, when *X* is weakly dependent on the missing covariates which is the case of figure 3.25, the estimated cumulative baseline hazard functions for the incomplete models are almost the same as the population cumulative hazard function. As a consequence, the null hypothesis fails to be rejected for both *h(x,y)* and *h(x)*. Figure 3.24 and figure 3.25 demonstrate the results for *C=5*. When C=1, null hypothesis can be rejected for both incomplete models in the case of strong dependence while for the situation where covariates are weakly dependent null hypothesis fails to be rejected for model *h(x)*. Relevant plots are shown in Appendix 3.

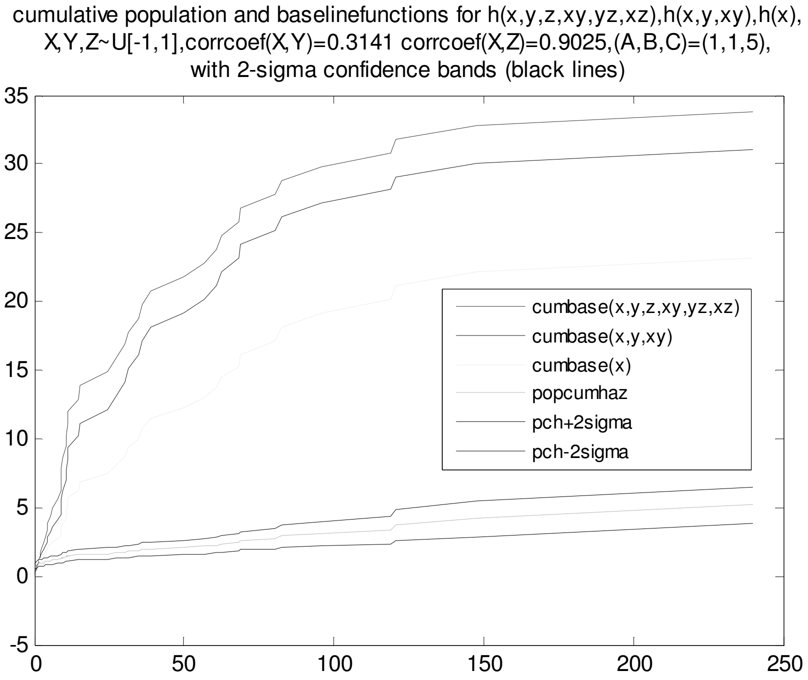


Figure 3.24 Cumulative population and baseline hazard for models with interaction, strong dependence, C=5

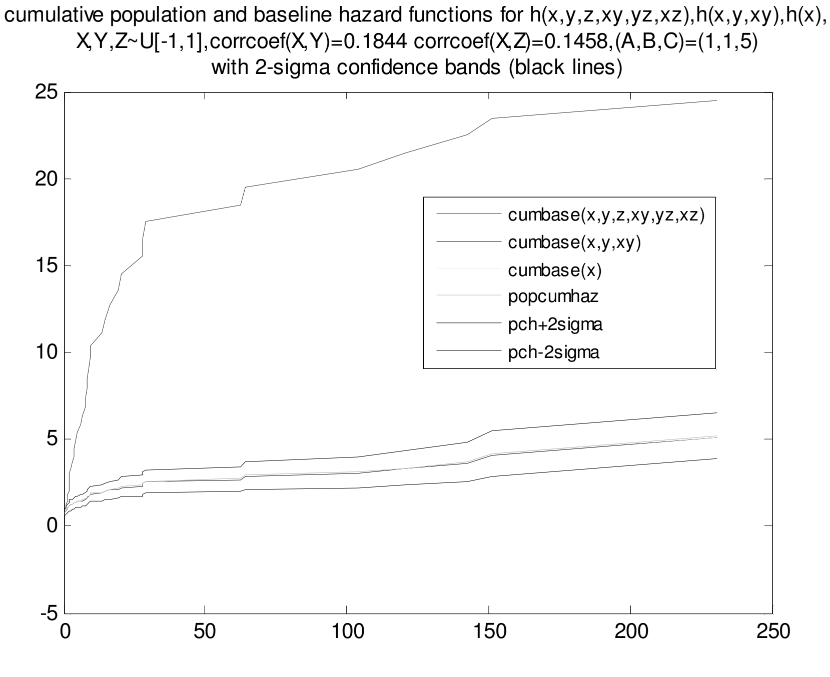


Figure 3.25 Cumulative population and baseline hazard for model with interaction, weak dependence, C=5

Chapter 4

Other Methods for Model Adequacy Test

In this chapter, we investigate three other methods used to test model adequacy for the Cox proportional hazard model. Our discussion starts with likelihood ratio test to see if the true complete model fit the data significantly better than the model with missing covariates. Moreover, by including time dependent variables we can check if the proportional hazard assumption that covariates are independent of time is satisfied. In the end, we use the Martingale residual to test if the model with missing covariates still satisfies the assumption that the form of covariates in the Cox model is linear.

4.1 Likelihood Ratio Test

In this part, the main focus is likelihood ratio test which will tell us if one model is better than another model through comparison of the log-likelihood of them. In this thesis likelihood ratio test is applied to check if the true model is better than the model with missing covariates. This is done through the following way. We calculate the log-likelihood of both the true model and the model with missing covariates and check if the difference between them is significantly large, or in other words, if the log-likelihood of the true model is significantly larger than the incomplete model. If so then we can say that true model is better than the incomplete model. In this section we discuss about the influence of independence, weak dependence and strong dependence between covariates. Moreover, for each situation we consider three kinds of true model: model without interaction terms or quadratic terms, model with interaction terms but not quadratic terms together with the model with both interaction and quadratic terms.

4.1.1 Introduction

Idea of likelihood ratio test is as follows. As a start the likelihood ratio test statistics, denoted as G, is calculated as twice the difference between the log partial likelihood of two models. This difference will be asymptotically χ^2 distributed with degrees of freedom equal to the difference in dimensionality between these two models. Then $Pr\ (\chi^2\ (n) \geq G)$ is calculated where n is the degrees of freedom of χ^2 distribution. If this probability is smaller than some threshold value, say

0.05, then it implies that *G* or in other words, the difference in log-likelihood between two models is significantly large and one model is significantly better than the other. In this section, we use the likelihood ratio test to check if the true model fits the data better than the models with missing covariates. First we calculate the log-likelihood of the true model, of all the models with one missing covariates, and of all the models with two missing covariates. Then we pick up the best 2-covariate model (where one covariate is excluded) which has the largest log-likelihood among all 2-covariate models and also the best 1-covariate model (where two covariates are excluded). First, *G* is calculated as twice the difference between the log partial likelihood of the true model and the best 2-covariate model. If $Pr\ (\chi^2\ (1) \geq G)$ is significantly small then it implies that the true model is better than the best 2-covariate model and no covariate should be excluded. The same procedure will be applied to check if the true model is better than the best 1-covariate model. However, this time degrees of freedom of χ^2 distribution will be 2. As a start let us have a look at the situation of independent covariates.

4.1.2 Independent Covariates

In this section six tests are performed as follows:

G1 is the likelihood ratio test for the true model and the best 2-covariate model while G2 is the test for the true model and the best 1-covariate model. For these two tests interaction between covariates and quadratic terms are not considered.

Test G3 and test G4 are similar to G1 and G2 respectively. The only difference lies in the inclusion of interaction terms in the true model which has the form *h(x,y,z,xy,yz,xz)*. Moreover, in this case 2-covariate models are *h(x,y,xy), h(y,z,yz)* and *h(x,z,xz)* while 1-covariate models are *h(x), h(y)* and *h(z)* .

Test G5 and test G6 are for the situation when both interaction terms and quadratic terms are considered. In this case, true model is of the form $h(x,y,z,x^2,y^2,z^2,xy,yz,xz)$, 2-covariate models are $h(x,y,x^2,y^2,xy)$, $h(y,z,y^2,z^2,yz)$ and $h(x,z,x^2,z^2,xz)$, and 1-covarite models are of the form $h(x,x^2)$, $h(y,y^2)$ and $h(z,z^2)$. G5 is the likelihood ratio test for the true model and the best 2-covariate model while G6 is the likelihood ratio test for the true model and the best 1-covariate model.

For all the six tests *X, Y, Z* are sampled from centered uniform distribution on [-1, 1] and *(A, B, C)* takes the value *(1, 1, 5)*. 10 simulations are performed for each test and in each simulation probability $Pr\ (\chi^2\ (n)\ \geq G)$ for all the tests is calculated. From the results which are shown in Appendix 4, we see that in each simulation for all tests, probability of $\chi^2(n)\geq G$ are all much smaller than 0.05. The largest value of that probability appears in the fourth simulation when we compare true model with the best 2-covariate model without consideration of interaction or quadratic terms and the value is 2.2206e-003. This implies that true model is better than all other models, whether interaction terms and quadratic terms are in the model or not.

4.1.3 Covariates Strongly Dependent

Here we consider strong dependence between some covariates. We put rank correlation between *X* and *Z* as 0.9 while rank correlation between *X* and *Y* is set as 0.4. Similarly to section 4.1.2, we perform six tests as below:

G7: test between full model and the best 2-covariate model, true model without interaction or quadratic terms.

G8: test between full model and the best 1-covariate model, true model without interaction or quadratic terms.

G9: test between full model and the best 2-covariate model, true model with interaction terms.

G10: test between full model and the best 1-covariate model, true model with interaction terms.

G11: test between full model and the best 2-covariate model, true model with interaction and quadratic terms.

G12: test between full model and the best 1-covariate model, true model with interaction and quadratic terms.

Here marginal distribution of *X, Y, Z* are still *U* [-1, 1] and we set *(A, B, C)* as *(1, 1, 1).* Results obtained from 10 simulations are shown in Appendix 4.

From our results we notice that when strong dependence exists between some covariates, log-partial likelihood of the true model is not always significantly larger than that of the best 2-covariate model. This is especially obvious for the true model without interaction or quadratic terms, where the values of $Pr(\chi^2(n)\geq G)$ are larger than 0.1 in all simulations. For simulation 7 and 8, this probability is even larger than 0.5. However, in the case of strong dependence, true model is significantly better than the best 1-covariate model, since the resulting $Pr\ (\chi^2(n)\geq G)$ is very small in each simulation. Actually when we compare true model with the best 1-covariate model the largest value of $Pr\ (\chi^2(n)\geq G)$ appear in simulation 8 for the model without interaction or quadratic terms and the value is only *6.4418e-006*, which is much smaller than the frequently used significance level *0.05*.

4.1.4 Covariates Weakly Dependent

When we apply the tests in section 4.1.3, denoted in this section as test G13-G18 respectively, to weakly dependent covariates (here rank correlations between *X* and *Y* as well as *X* and *Z* are both set as 0.1), results are similar to that obtained in section 4.1.2. In all simulations and for all tests true model is better than all the other models since the difference between log partial likelihood of them are significantly large in each case. As shown in the end of Appendix 4 the largest value of *Pr* $(\chi^2(n) \geq G)$ is only*1.9149e-003* .

4.2 Inclusion of Time Dependent Variable

4.2.1 Introduction

In this part time dependent variable *ln(t)*covariate* is introduced. We compared the estimates of coefficient for the models with and without this time dependent variable. As in section 3.1, we simulate 100 times and compare the ordered estimates in plots. If the 100 ordered estimates for models with and without this time dependent variable are different, or if this time dependent variable can help better fit the data, then this time dependent variable is significant to the Cox model thus should be included. If so, then it is in conflict with the proportional assumption that all covariates are time independent. In addition to the method above, at the end of this section we perform Wald test on the coefficients of the time dependent variables to check if time dependent variables are significant. True model used in this section is *h(x,y,z).*

4.2.2 True Model *h(x,y,z)*

At first let us have a look at the true model. Here we compare the 100 ordered estimates of *A* for models *h(x,y,z)* and *h(x,y,z,xlnt)*, the 100 ordered estimates of *B* for models *h(x,y,z)* and *h(x,y,z,ylnt)* and the 100 ordered estimates of *C* for models *h(x,y,z)* and *h(x,y,z,zlnt)*. In this way we can test if all these three covariates are time independent or not.

Figure 4.1 is obtained from independent covariates and in this example *C=5*. From this figure we see that for the true model the ordered estimates are almost the same for the model with and without the time dependent variable. This implies that for the true model the proportional hazard assumption is satisfied. When some covariates are strongly dependent or in the case of weak dependence results are similar which we put in Appendix 5 and proportional hazard assumption is satisfied in those cases.

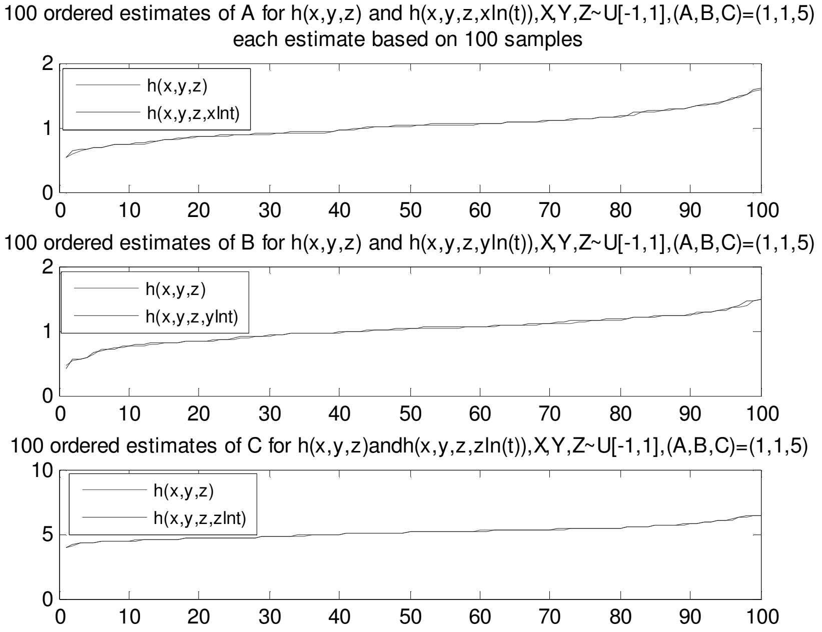


Figure 4.1 100 ordered estimates of model coefficients for model with and without time dependent variable, true model, (A,B,C)=(1,1,5)

4.2.3 Models with two covariates missing

As a contrast we see from figure 4.2 that for the incomplete model *h(x)* where two covariates are excluded, the ordered estimates for the model with time dependent variable are different from those for the model without the time dependent variable. Moreover, we notice from figure 4.2 that inclusion of the time dependent variable actually help make smaller the under-estimation due to missing covariates. Although estimates for both models are smaller than 5, estimates which include the time dependent variable are generally larger. Therefore this time dependent variable is significant to the model and should be included for instance as a covariate and this is in conflict with the assumption that all covariates are time independent. Similar results are obtained for the model where *X, Y* and *Z* are weakly dependent (see Appendix 5). However, when data is sampled from strongly dependent covariates, as shown in figure 4.3, the difference between ordered estimates for the two models is not as obvious as in figure 4.2 where the same values of coefficients *(5,5,5)* are set for simulations. It is hard to say from figure 4.3 whether the time dependent variable is influential or not.

Therefore for the model with two covariates excluded proportional assumption does not hold for weakly dependent or independent covariates. However, when it comes to strong dependence this method does not work well and it is advisable to use other ways for model adequacy test.

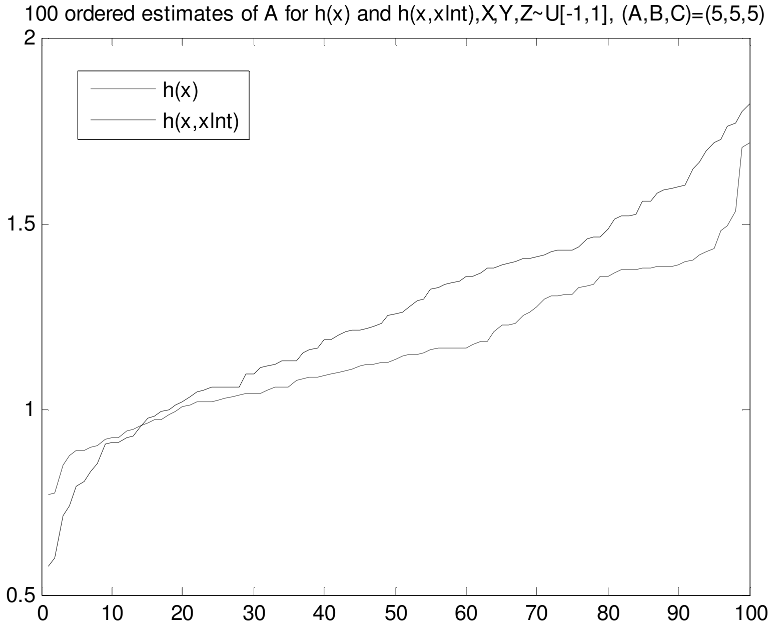


Figure 4.2 Estimates of A, inclusion of time dependent variable, for the model with two missing covariates, independent, (A,B,C)=(5,5,5)

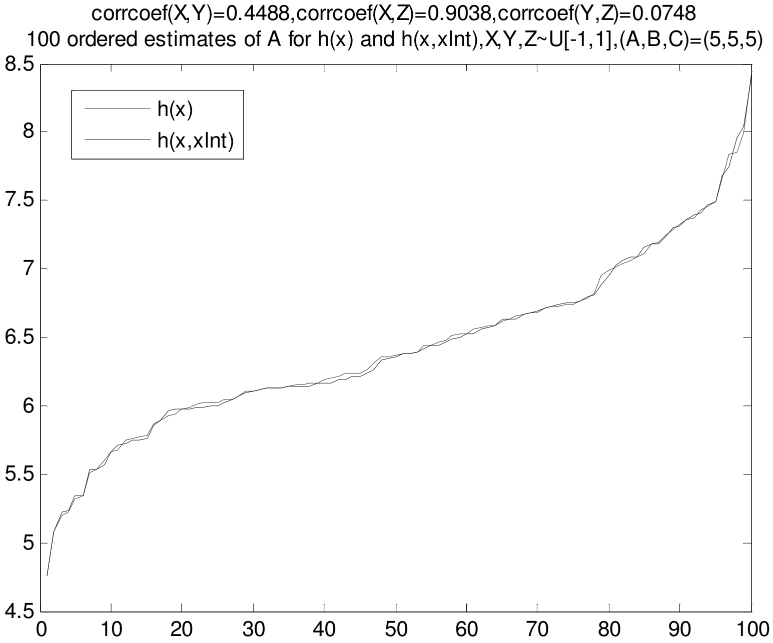


Figure 4.3 Estimates of A, inclusion of time dependent variable, for the model with two missing covariates, strong dependence, (A,B,C)=(5,5,5)

4.2.4 Models with one covariate missing

Figure 4.4 is obtained from independent covariates. It shows that when taking the same coefficient values *(5,5,5)* as in figure 4.2 where two covariates are excluded from the model, there is difference between the ordered estimates for models with and without the time dependent variables but the different is not as large as in figure 4.2. As shown in figure 4.5, when covariate X is strongly dependent on missing covariate Z (here product moment correlation coefficient between these two covariates is larger than 0.9), difference between estimates from the model with and without the time dependent variables is small and it is hard to judge from the graph if the difference is significant or not. In other words, apart from the graphic method, it is important for us to use some statistical test to check if time dependent variables are significant or not. This is what we will do in the next section.

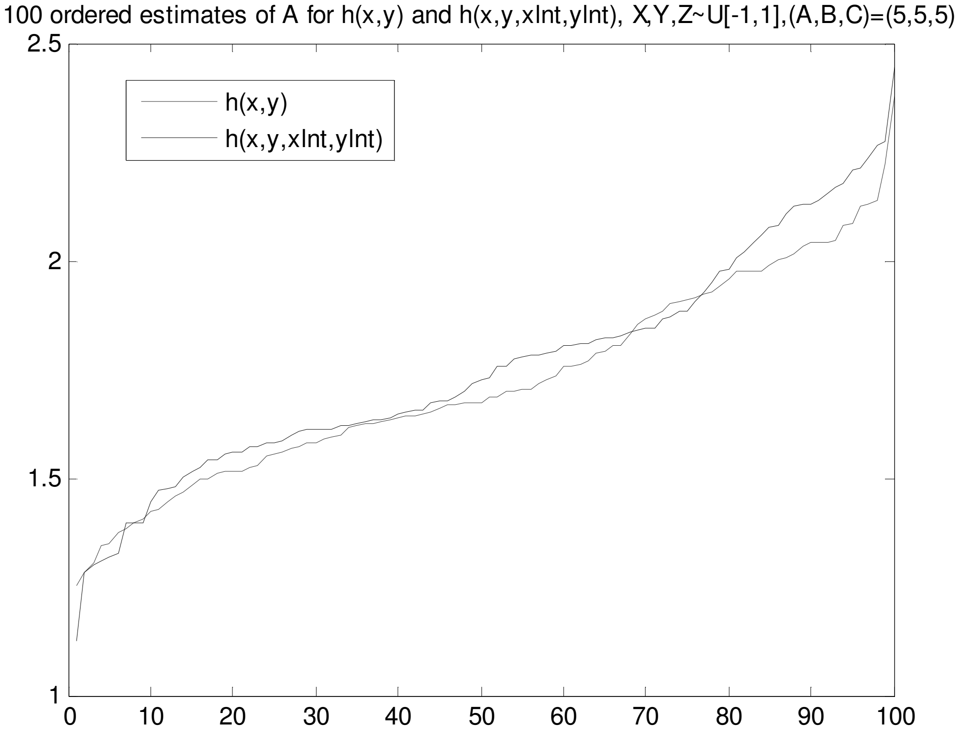


Figure 4.4 Estimates of A, inclusion of time dependent variable, for the model with one missing covariates, independent, (A,B,C)=(5,5,5)

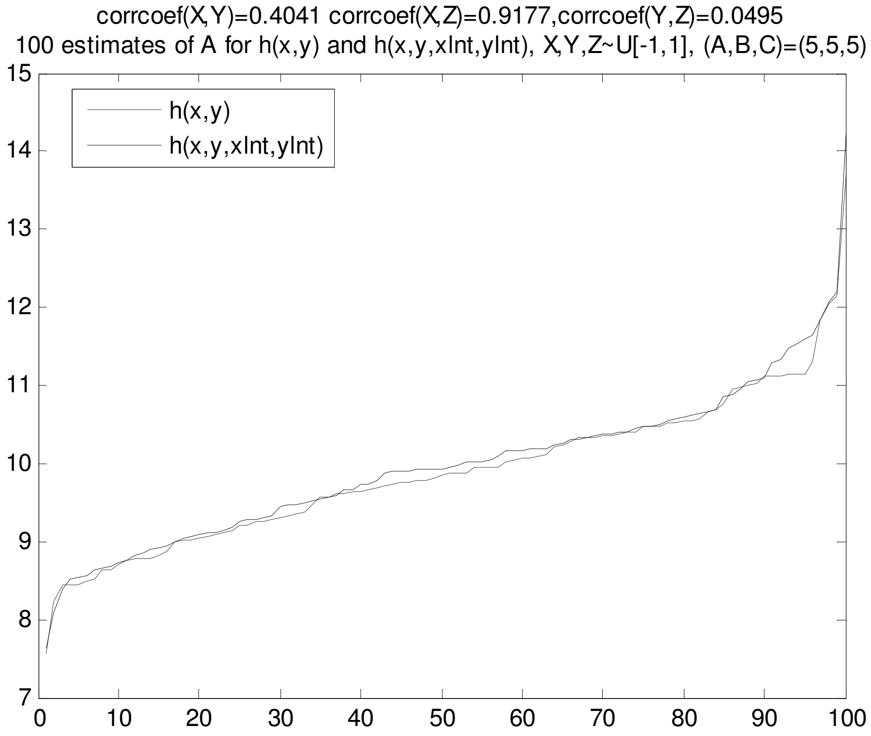


Figure 4.5 Estimates of A, inclusion of time dependent variable, for the model with one missing covariates, strong dependence, (A,B,C)=(5,5,5)

4.2.5 Wald Test

Wald statistic is used to test significance of coefficient. It is the ratio of the estimated coefficient to its estimated standard error. The estimator of the standard error is the positive square root of the variance estimator which is calculated in the following way.

Denote the coefficient we need to test as β and the corresponding covariate as x. Support there are m subjects $x_1,...x_m$ with survival times $t_{(1)} \le t_{(2)} \le ... \le t_{(m)}$. First we calculate the observed information as follows

$$I(\beta) = -\frac{\partial^2 L_p(\beta)}{\partial \beta^2} = -\sum_{i=1}^{m} \sum_{t_j \ge t_{(i)}} w_{ij} (x_j - \bar{x}_{w_i})^2$$

$L_p(\beta)$ is the log partial likelihood.

$$w_{ij}(\beta) = \frac{e^{x_j\beta}}{\sum\limits_{t_l \geq t_{(i)}} e^{x_j\beta}}, \quad \bar{x}_{w_i} = \sum_{t_j \geq t_{(i)}} w_{ij}(\beta) x_j$$

Then the estimator of variance can be calculated by

$$\hat{Var}(\hat{\beta}) = I(\hat{\beta})^{-1}$$

Under the null hypothesis that the coefficient is equal to zero, along with other mathematical conditions, the Wald statistics will follow a standard normal distribution and the equation for the Wald statistics is

$$z = \frac{\hat{\beta}}{\hat{SE}(\hat{\beta})}$$

$\hat{SE}(\hat{\beta})$ is the estimated standard error and two-tailed test is applied here.

First assume that covariates are independent. We regress the hazard onto both the covariates and time dependent variables and then apply Wald test to check if the time dependent variables are significant or not. Tables 4.1, 4.2 and 4.3 are summaries of Wald test for the true model and models with missing covariates. Here *Beta* is the estimated coefficient of the time dependent variable, *se* is its estimated standard error, *z* is the Wald statistics and *p* is the two-tailed *p-value*. Significance level applied here is 0.05. If the *p-value* is smaller than 0.05 then the null hypothesis that the coefficient is zero is rejected. Moreover, the time dependent variable is significant and the proportional hazard assumption is not satisfied.

From table 4.1 we see that for the true model, *p-values* for all time dependent variables are all larger than 0.05. Hence in the true model all the time dependent variables are not significant and proportional hazard assumption is satisfied.

	Beta	*se*	*z*	*p*
Xlnt	*0.0273*	*0.0844*	*0.3229*	*0.7468*
Ylnt	*0.0906*	*0.0742*	*1.2218*	*0.2218*
Zlnt	*0.0482*	*0.0762*	*0.6320*	*0.5274*

Table 4.1 Wald Test of Time Dependent Variable for model h(x,y,z) X,Y,Z independent

As we include time dependent variables *Xlnt* and *Ylnt* into the model *h(x,y)*, it is shown in table 4.2 that the *p-value* of *Xlnt* is smaller than the significance level 0.05, which implies that this time dependent variable is significant and proportional assumption is not satisfied.

	Beta	*se*	*z*	*p*
Xlnt	*-0.1489*	*0.0527*	*-2.8241*	*0.0047*
Ylnt	*0.0670*	*0.0501*	*1.3381*	*0.1809*

Table 4.2 Wald Test of Time Dependent Variable for model h(x,y), X,Y,Z independent

When it comes to the model with two covariates missing, we see from table 4.3 that *p-value* of *Xlnt* is much smaller than 0.05 hence it is very influential to the model and proportional hazard assumption is not obeyed.

	Beta	*se*	*z*	*p*
Xlnt	*-0.1630*	*0.0452*	*-3.6084*	*0.0003*

Table 4.3 Wald Test of Time Dependent Variable for model h(x), X,Y,Z independent

Now, let us have a look at strong dependence between covariates. In this case,

$$\rho(X,Y) = 0.4742, \rho(X,Z) = 0.8970, \rho(Y,Z) = 0.5589,$$

	Beta	*se*	*z*	*p*
Xlnt	*-0.0182*	*0.1003*	*-0.1819*	*0.8557*
Ylnt	*0.0371*	*0.0574*	*0.6461*	*0.5182*
Zlnt	*-0.0129*	*0.0995*	*-0.1294*	*0.8971*

Table 4.4 Wald Test of Time Dependent Variable for model h(x), X,Y,Z independent

From table 4.4 we see that for the true model, in the case of strong dependence, none of the time dependent variables is significant. However this is also the case for model *h(x,y)*, as shown in table 4.5, which is identical to the results shown in figure 4.4 and figure 4.5.

	Beta	*se*	*z*	*p*
Xlnt	*-0.0328*	*0.0663*	*-0.4955*	*0.6202*
Ylnt	*0.0295*	*0.0524*	*0.5637*	*0.5730*

Table 4.5 Wald Test of Time Dependent Variable for model h(x,y), X,Y,Z strongly dependent

	Beta	*se*	*z*	*p*
Xlnt	*-0.0591*	*0.0392*	*-1.5079*	*0.1316*

Table 4.6 Wald Test of Time Dependent Variable for model h(x,y), X,Y,Z strongly dependent

In table 4.6 for model *h(x), p-value* of *Xlnt* is also larger than the significance level 0.05 but much smaller than that in model *h(x,y)* and *h(x,y,z)*. Therefore time dependent variable *Xlnt* is much more significant for the model *h(x),* than the other two models, but still not significant enough that we can reject the null hypothesis and claim that proportional hazard assumption is violated.

4.3 Martingale Residual

In this section we test model adequacy by checking if the linearity assumption still holds for the incomplete model where covariates are missing. True model used here is *h(x,y,z).* Linearity assumption tells us that covariates appear in the Cox model in the linear form, or in other words, log-hazard function is linear to all covariates. First let us review definition of Martingale residual.

4.3.1 Martingale Residual and Lowess Smooth

As a start of this section we review the definition of martingale residual and Lowess smooth then we show how to apply this method to test model adequacy.

The martingale residual for the *i*th subject at the end of follow-up is of the form

$$\hat{M}_i = c_i - \hat{h}_i = c_i - \hat{h}(t_i, x, \overset{\wedge}{\beta})$$

$\hat{M}_i$ is the estimated martingale residual for the *i*th subject and c_i is the censoring variables for the *i*th subject. Since we assume no censoring all c_i are equal to 1. Moreover, $\hat{h}_i$ is the estimated cumulative hazard and in this section it is estimated in the following way.

The expression for the cumulative baseline hazard is

$$h_0(t_i) = \sum_{t_j \le t_i} \lambda_0(t_j)$$

The value of the baseline hazard at t_i is expressed as (for the full model)

$$\lambda_0(t_i) = \frac{c_i}{\sum_{j \in R(t_i)} \exp(\hat{A} x_i + \hat{B} y_i + \hat{C} z_i)} = \frac{1}{\sum_{j \in R(t_i)} \exp(\hat{A} x_i + \hat{B} y_i + \hat{C} z_i)}$$

Note that in our paper we assume that there are no ties.

Then the cumulative hazard $\hat{h}_i$ is estimated as

$$\hat{h}_i = h_0(t_i)\exp(\hat{A}\,x_i + \hat{B}\,y_i + \hat{C}\,z_i)$$

t_i is the survival time, x represents the covariates which in our case are *X,Y,* and *Z*, and $\hat{\beta}$ is the estimated coefficient. In our model the estimated coefficients are $\hat{A}, \hat{B}$ and $\hat{C}$.

The residual

$$\hat{M}_i = c_i - \hat{h}_i = c_i - \hat{h}(t_i, x, \hat{\beta})$$

has also been called the Cox-Snell or modified Cox-Snell residual ([3]), also see Cox and Snell (1968) and Collett (1994). This terminology is due to the work of Cox and Snell.

Lowess smooth is used here in addition to the martingale residual. It is a weighted scatter plot smoothing method which fits simple models to localized subsets of the data to build up a function that describes the deterministic part of the variation in the data, point by point. The smoothing parameter, denoted as α, is the proportion of data used in each fit and the subset of data used in each weighted least squares fit hence comprises the *nα* (*n* is the number of points in the plot) points whose values are closest to the point at which the response is being estimated. Large values of α produce the smoothest functions while for smaller α the regression function will be closer conform to the data. However, when using a too small α the regression function will eventually start to capture the random error in the data. When we take α=0.5 the Lowess regression function is smooth enough for our study and we keep this setting all through this section. We use the Matlab command '*smooth*' to make Lowess smooth.

To use martingale residual and Lowess smooth to test model adequacy, Therneau, Grambsch and Fleming (1990) suggest fitting a model that excludes the covariate of interest. The results are used to calculate the martingale residual and to generate Lowess smoothed values. These are then plotted against the values of the excluded covariate and the shape of the plot, especially the smooth, provides an estimate of the functional form of the covariate in the model. Hence, if the shape of martingale residual and Lowess smooth turns out to be linear then it implies that the model satisfies the linearity assumption and vice versa.

In this part we apply this method to model *h(x,y z)* and *h(x,y),* the covariate of interest or in other words, the excluded covariate will be *X*. We exclude *X* and fit both models and then plot the calculated martingale residual and Lowess smooth values versus the excluded covariate *X*. Then we can check if the linearity assumption holds in each case, in other words, to see if the assumption that log-hazard function is linear to covariate *X* holds for both true model *h(x,y z)*and the incomplete model *h(x,y).*

4.3.2 An Improvement of This Method

Grambsch, Therneau and Felming (1995) expand on their earlier work and suggest that one begin with a fit of the model containing all covariates and then plot the log of ratio of a smoothed c to a smoothed $\hat{H}$ versus the covariate of interest. They found out that in this way it has greater diagnostic power than their earlier proposed method which has been described in section 4.3.1.

In our paper this method is applied in the following way. We fit models *h(x,y,z), h(x,y)* and *h(x)*, all including the covariate off interest, *X*. Then the cumulative hazard $\hat{H}_i$ is estimated using the way introduced in section 4.3.1. The values of $\hat{H}_i$ are plotted versus *X* and a Lowess smooth was calculated and saved, denoted as $\hat{H}_{smooth}$. Moreover, smoothed c is always one since in our case we do not consider censoring hence all c_i are one. The smoothed values were then used to calculate

$$f_i = \ln(\frac{\hat{c}_{smooth}}{\hat{H}_{smooth}}) + \hat{A}x_i = \ln(\frac{1}{\hat{H}_{smooth}}) + \hat{A}x_i$$

For models *h(x,y,z), h(x,y)* and *h(x)*we plot f_i against x_i, if the resulting plot is not linear then it implies that *X* is not of the linear form in the model, which is in conflict with the linearity assumption .

4.3.3 Independent Covariates

Let us first have a look at the situation where all covariates are independent. Figure 4.6 is the martingale residual and Lowess smooth plot for model *h(x,y,z)* versus *X*. Figure 4.7 is the plot of f_i for model *h(x,y,z)*against *X*. From these two graphs we see that for the true model, both of them show a linear shape which implies that the true model satisfies the linear assumption.

Figures 4.8 and 4.9 are the corresponding plots for model *h(x,y)*. Both plots have a non-linear shape which implies that covariate *X* appears as non-linear in the model. This is a violation of the linear assumption. Similar result is shown in Figure 4.10 which is the plot of f_i for model *h(x).* From the non-linear shape of this figure we see that *h(x)* do not obey the linearity assumption of the model either.

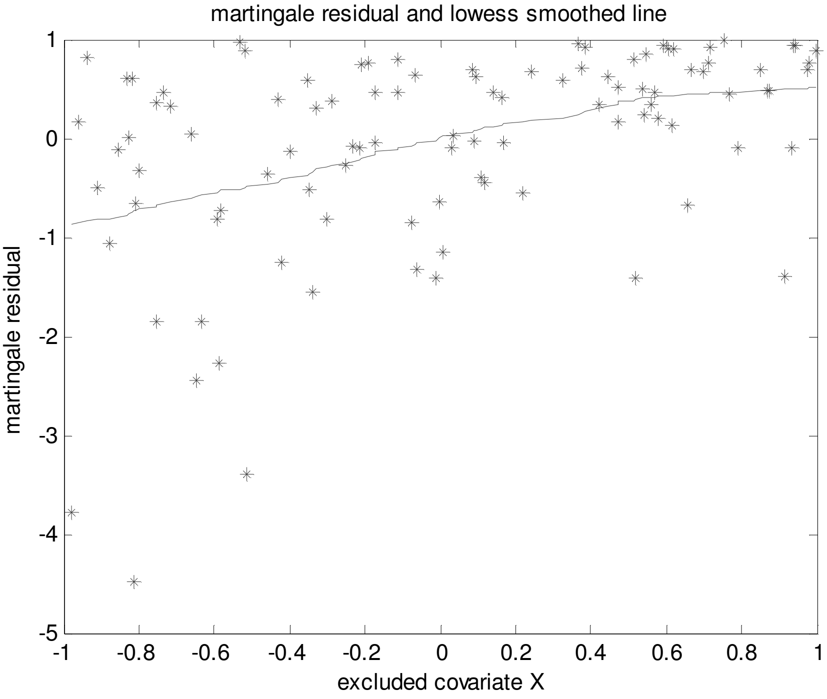


Figure 4.6 Martingale residual and Lowess smooth for h(x,y,z), independent

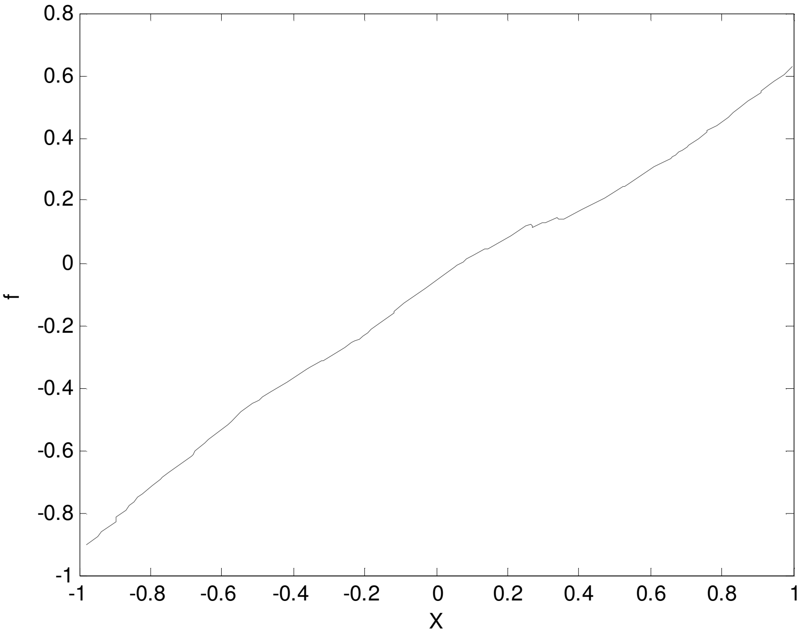


Figure 4.7 *fi* versus *xi* for h(x,y,z), independent

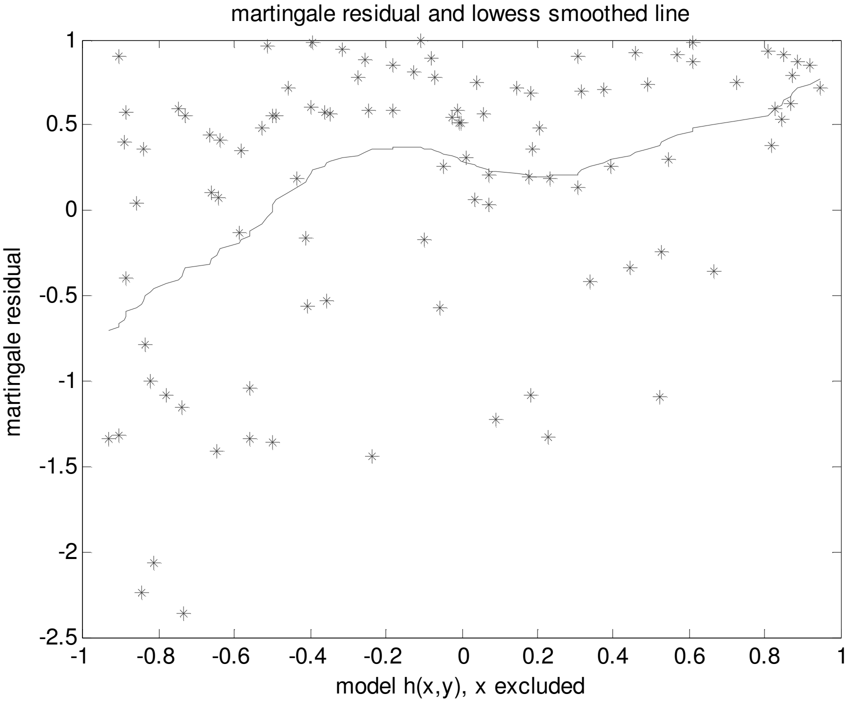


Figure 4.8 Martingale residual and Lowess smooth for h(x,y), independent

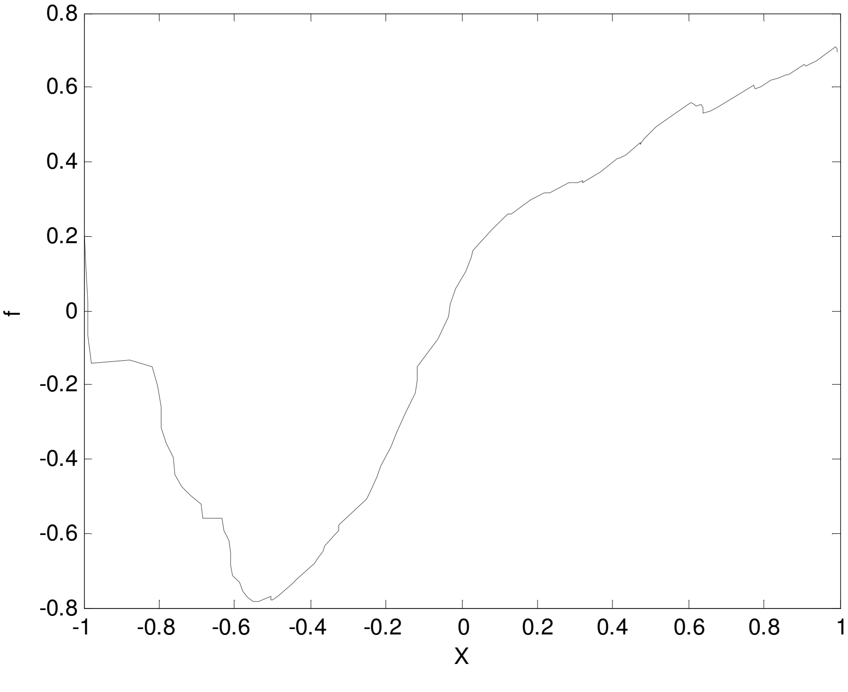


Figure 4.9 *fi* versus *xi* for h(x,y), independent

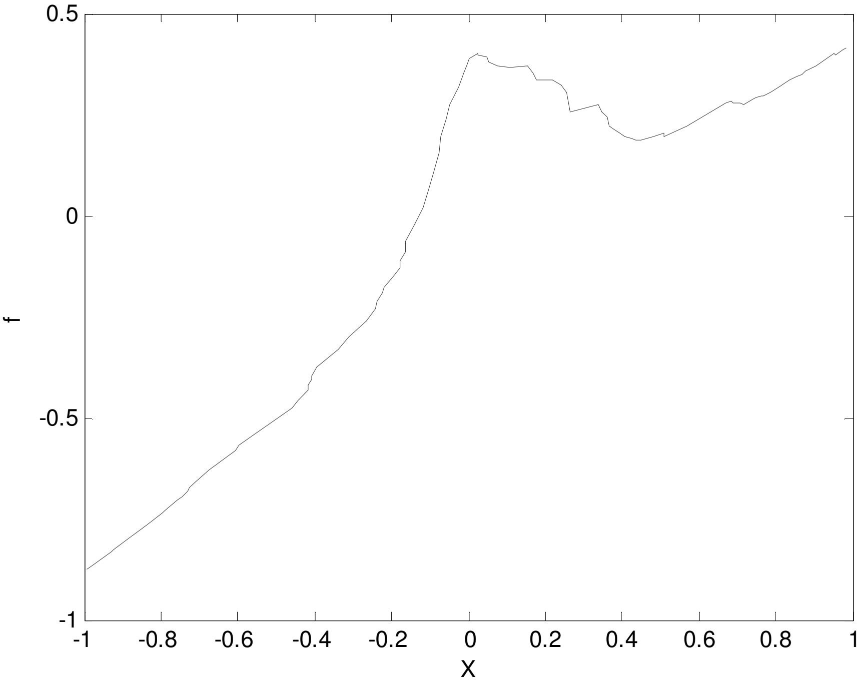


Figure 4. 10 *fi* versus *xi* for h(x), independent

4.3.5 Dependent Covariates

First let us talk about the influence of strong dependence. Figure 4.11 and figure 4.12 are plots of the martingale residuals and f_i for model *h(x,y)*. Although non-linear shape can still be seen from both these two graphs which implies violation of linear assumption for covariate *X*, deviation from linearity is not as much as in the case of independence as shown in figure 4.8 and figure 4.9. This is also the case for model *h(x)* as shown in figure 4.13. There is deviation from linear at the beginning, but it seems when *X* takes values from -0.6 to 1, a linear shape can be seen within this interval.

Similar to the situation of independence, when we look at figures for weakly dependent covariates (figures 4.14, 4.15, 4.16), non-linear shapes are obvious in the martingale residuals for *h(x,y)* and in the plots of f_i for both *h(x,y)*and *h(x)*. This implies covariate *X* does not appear in its linear form in the model hence linear assumption is not satisfied. Furthermore, plots of martingale residuals and f_i deviate more from linearity in the case of weak dependence than in the case of strong dependence.

In all situations of dependence, model *h (x,y,z)* satisfies the linear assumption since all plots demonstrate a linear shape. Relevant plots for *h(x,y,z)* are put in Appendix 6.

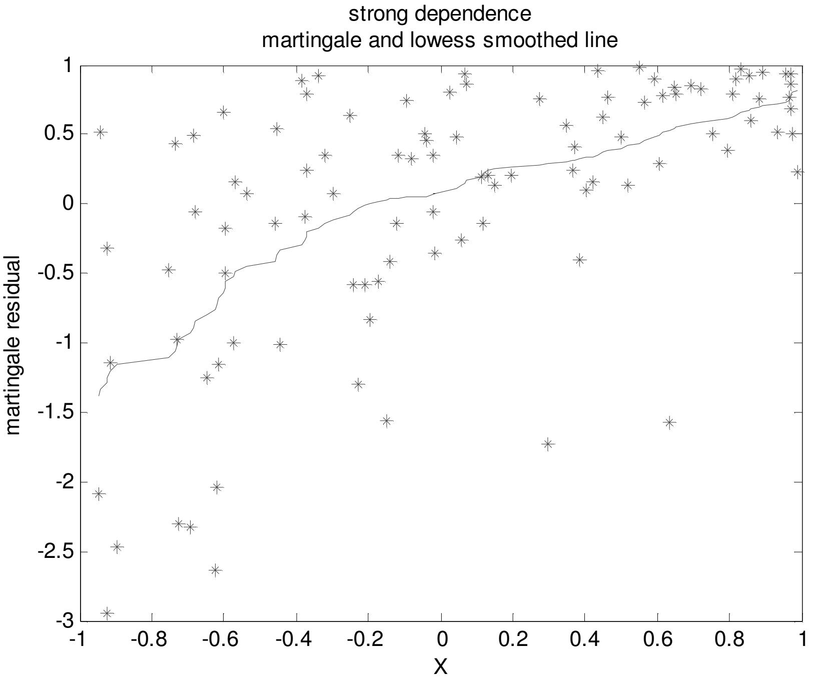


Figure 4. 11 Martingale residual and Lowess smooth for h(x,y), strong dependence

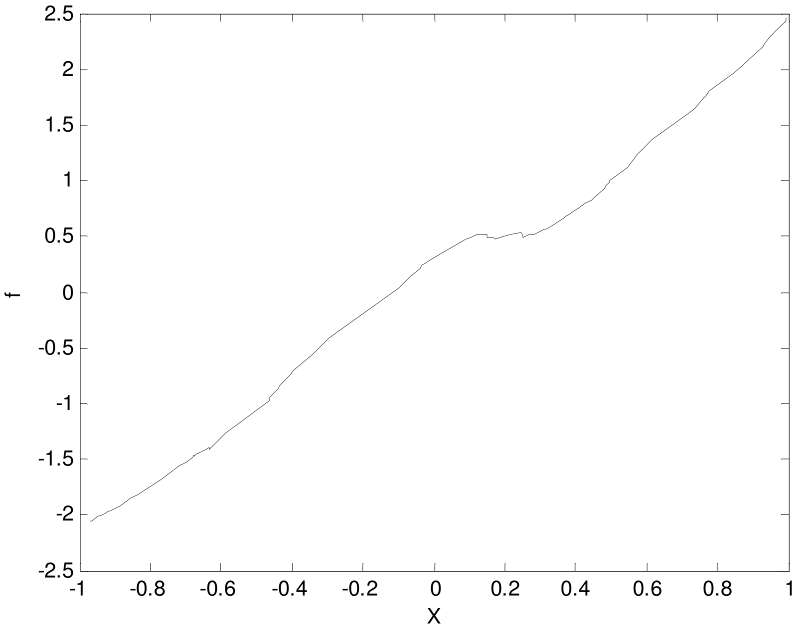


Figure 4.12 *fi* versus *xi* for h(x,y), strong dependence

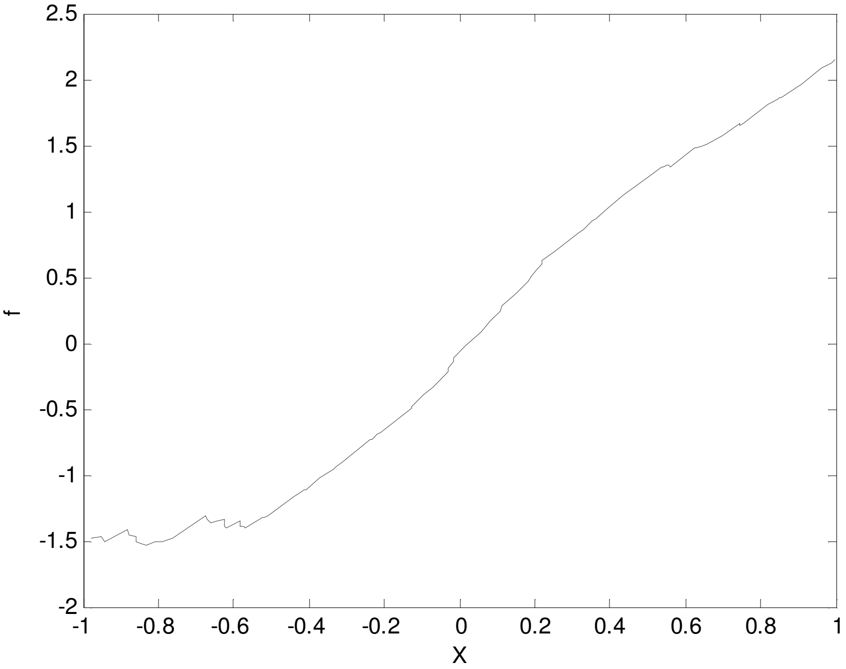


Figure 4.13 f_i versus *xi* for h(x),strong dependence

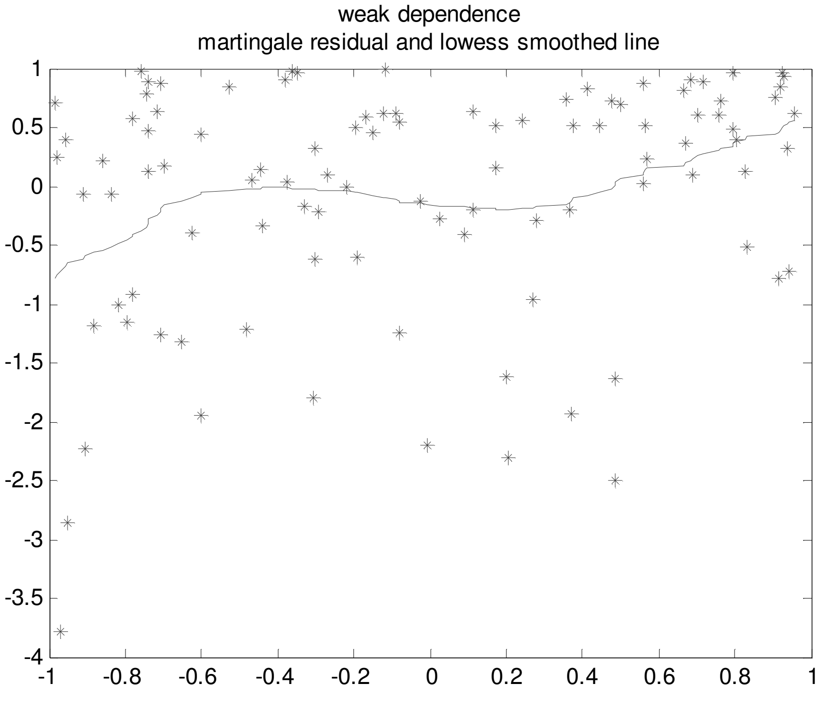


Figure 4. 14 Martingale residual and Lowess smooth for h(x,y), weak dependence

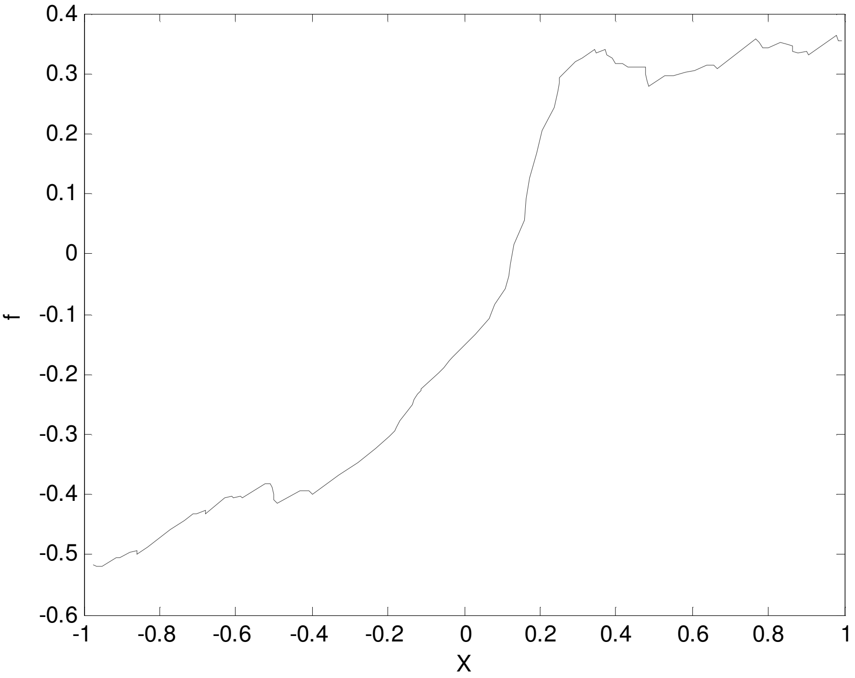


Figure 4.15 ***fi*** **versus** ***xi*** **for h(x,y), weak dependence**

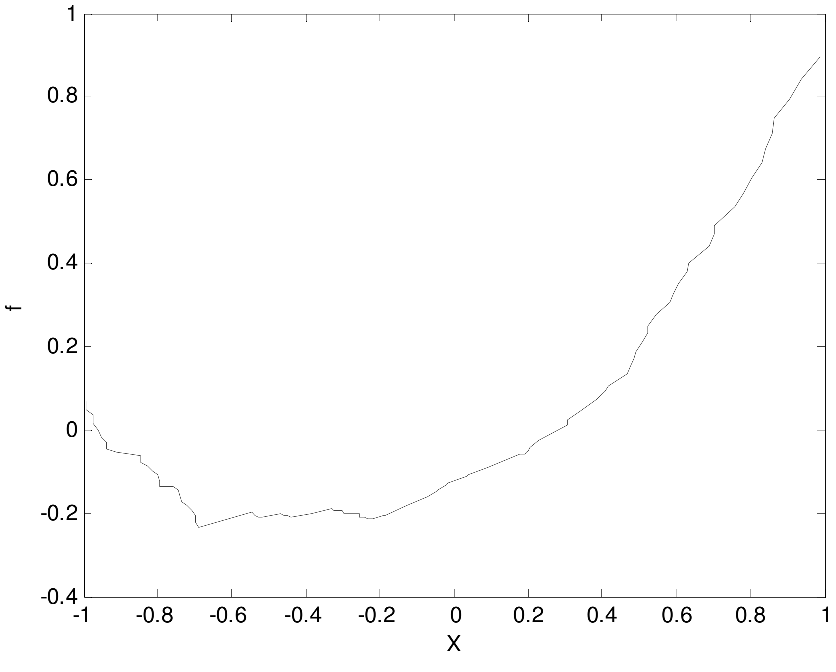


Figure 4.16 ***fi*** **versus** ***xi*** **for h(x), weak dependence**

At the end of this chapter we give a talk about the Schoenfeld residual which is commonly used as a way to test model adequacy. More precisely, it is used to test adherence to the proportional hazard assumption. Schoenfeld (1982) proposed the first set of residuals for use with a fitted proportional hazard model and these residuals are referred to as Schoenfeld residuals. These residuals are based on the individual contribution to the derivative of log partial likelihood. Suppose there are *p* covariates and *n* subjects. The derivative of log partial likelihood with respect to the *kth* covariate is give below. Note that we assume no censoring, all c_i equal to one.

$$\frac{\partial L_p(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} c_i (x_{ik} - \frac{\sum_{t_j \geq t_i} x_{jk} e^{x_j'\beta}}{\sum_{t_j \geq t_i} e^{x_j'\beta}}) = \sum_{i=1}^{n} c_i (x_{ik} - \bar{x}_{w_i,k}) = \sum_{i=1}^{n} (x_{ik} - \bar{x}_{w_i,k})$$

where

$$\bar{x}_{w_i,k} = \frac{\sum_{t_j \geq t_i} x_{jk} e^{x_j'\beta}}{\sum_{t_j \geq t_i} e^{x_j'\beta}}$$

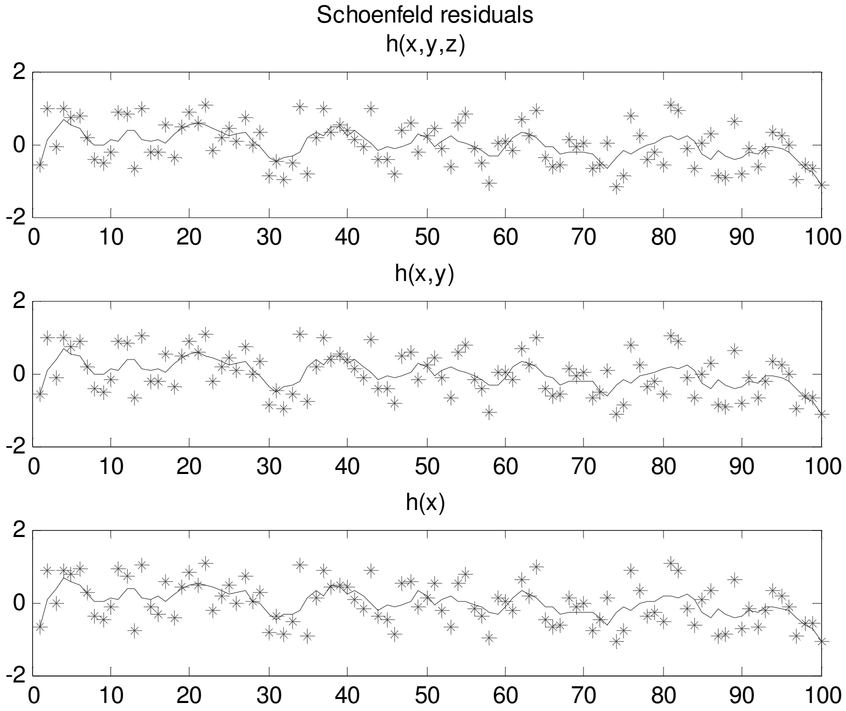
The estimator of the Schoenfeld residual for the *ith* subject on the *kth* covariate is obtained by substituting the partial likelihood estimator of the coefficient, denoted as $\hat{\beta}$:

$$\hat{r}_{ik} = \sum_{i=1}^{n} c_i (x_{ik} - \bar{x}_{w_i,k}) = \sum_{i=1}^{n} (x_{ik} - \hat{\bar{x}}_{w_i,k})$$

where

$$\hat{\bar{x}}_{w_i,k} = \frac{\sum_{t_j \geq t_i} x_{jk} e^{x_j'\hat{\beta}}}{\sum_{t_j \geq t_i} e^{x_j'\hat{\beta}}}$$

If these residuals have a trend over time, or there is some slope on the plot of these residuals, then it implies that time has influence on the residuals and the assumption that all covariates are time dependent is not satisfied. The figure below is the Schoenfeld residuals for covariate *X* calculated from model *h(x,y,z), h(x,y)* and *h(x).* None of them have a slope or some trend over time, but on the other hand there is no significant difference between results of these three models. In this sense this method is not as powerful as inclusion of time dependence variables to test adherence of the proportional hazard assumption. Therefore we did not use this method in our study.



Chapter 5

Conclusion

The aim of this thesis is to test modal adequacy, which includes the test of how well the model fits the data and the test of whether model assumptions are satisfied. As in [1], we generate data from the assumed true model, and fit both the true model and models with covariates excluded to the generated data. Results of regression for each model are used in different methods designed to test model adequacy. In this thesis plenty of further study of [1] has been done by applying two methods proposed in [1] to test model's fit to dependent covariates, to interaction and to quadratic terms. We also applied another method, likelihood ratio test to check model's fit. Moreover, methods to test adherence to model assumptions are also studied. We include time dependent variables to test adherence to proportional assumption. Furthermore, martingale residual and Lowess smooth plots are used to test if the linearity assumption holds for true model and models with missing covariates. Results of simulations are summarized as follows.

When covariates are independent or weakly dependent, missing covariates result in under-estimation of coefficient. However, when covariates are strongly dependent, missing covariates result in over-estimation of coefficient. Larger value set for coefficient of missing covariate when generating data results in greater under-estimation or over-estimation. Models with interaction terms and models with both interaction and quadratic terms have similar results.

As for the null hypothesis that the estimated baseline cumulative hazard function is equal to the population cumulative hazard function, when covariates are strongly dependent on, this null hypothesis is rejected for *h(x,y)* and *h(x)* since the estimated baseline hazard functions for these two models of missing covariates are close to that of *h(x,y,z),* instead but not the population cumulative hazard. When covariates are weakly dependent, results are similar to independent case, that is, when $C=1$, this null hypothesis fails to be rejected for *h(x),* and when $C=5$, it is failed to be rejected for both *h(x,y)* and *h(x).*When interaction is considered results are similar in each situation.

As for the likelihood ratio test, when covariates are independent or weakly dependent, true model is significantly better than models with one or two missing covariates. This is also the case when interaction and quadratic terms are considered. However, when covariates are strongly dependent, true model is not so significantly better than the models with one missing covariate, and this is especially the case for the model with no interaction or quadratic terms. But the true model is still significantly better than model with two missing covariates and this is also the case when interaction and quadratic terms are considered.

As for the test of proportional assumption, for the true model, no matter whether covariates are independent of each other, strongly dependent or weakly dependent, no matter what value *C* takes, proportional assumption is always satisfied. For models with two missing covariates, proportional assumption does not hold any more in each case. When only one covariate is excluded, the estimates for models with and without time dependent variable are different but not so much as when two covariates are excluded, especially when covariates are strongly dependent.

When covariates are independent or weakly dependent, plots of martingale residual and the smoothed censoring over smoothed hazard both imply linear form of covariates for model *h(x,y,z),* and non-linear form for incomplete models. When covariates are strongly dependent, violation of the model with missing covariates from linearity is not as much as for the case of independence and weak dependence, but both models with covariates excluded do not satisfy the linear assumption any more.

In summary, the method that compares the ordered estimates for the true model and models of missing covariates works well for both independent and dependent covariates. It can also be applied to models with interaction terms and quadratic terms. However, the comparison between the estimated cumulative baseline hazard function and the population cumulative hazard function does not work well for strongly dependent covariates. Moreover, in the case of strong dependence, likelihood ratio test does not work well when comparing the true model with the best 2-covariate model, especially when there is no interaction or quadratic terms in the model. Hence when covariates are strongly dependent, it is advisable to test the model's fit through comparison of ordered estimates of coefficients. As for the tests of adherence to model assumptions, inclusion of time dependent variables, which is used to test if the proportional assumption is satisfied, works better for independent covariates and weakly dependent covariates than for strongly dependent covariates. Furthermore, the method we used to test adherence to the linear assumption through martingale residuals and Lowess smooth works well in each situation.

Bibliography

[1] Roger M. Cooke, *the Cox Proportional Hazard Model.*

[2] Dorota Kurowicka, Roger M. Cooke, *Uncertainty Analysis with High Dimensional Dependence Modeling*, Wilely, 2006.

[3] Hosmer, D.W. and Lemeshow, S, *Applied Survival Analysis*, Wiley, New York, 1999.

[4] John D. Kalbfleisch and Ross L. Prentice, *the Statistics Analysis of Failure Time Data*, Second Edition, Wiley, 2002.

[5] D. R. Cox, Regression Models and life-tables, Royal Statistics Society, Series B, Vol. 34, No. 2, 1972, pp. 187-220.

[6] Patricia M. Grambsch, Terry M. Therneau, Thomas R. Fleming, *Diadnostic Plots to Reveal Functional Form for Covariates in Multiplicative Intensity Models*, Biometric, Vol. 51, No. 4, (Dec., 1995), pp. 1469-1482.

[7] Terry M. Therneau, Patricia M. Grambsch, Modeling Survival Data, Springer, New York, 2000.

[8] William S. Cleveland and Susan J. Devlin, *Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting*, Journal of the American Statistics Association, Vol. 83, No. 403, (Sep., 1988), pp. 596-610.

[9] Cox, D.R., and D. Oakes, *Analysis of Survival Data*, Chapman & Hall, Boca Raton, 1984.

[10] Lawless, J.F., *Statistical Models and Methods for Lifetime Data*, Wiley, New York, 2003.

[11] Hougaard, P, *Analysis of Multivariate Survival Data*, Springer-Verlag, New York, 2000.

[12] Roger M. Cooke, Oswald Morales-Napoles, *Competing risk and the Cox proportional hazard model*, Journal of Statistical Planning and Inference vol. 136, no 5, pp. 1621-1637, 2006.

[13] Bretagnolle, J. and Huber-Carol, C. *Effects of omitting covariates in Cox's Model for survival data,* Scandinavian Journal of Statistics, 15, 125-138, 1988.

[14] C. A. Struthers and J. D. Kalbfleisch, *Misspecified Proportional Hazard Models*, *Biometrika*, Vol. 73, No. 2 (Aug., 1986), pp. 363-369.

[15] Keiding, N., Andersen, P.K., Klein, J.P. *The role of frailty time models in describing heterogeneity due to 35 omitted covariates.* Statist. Medicine 16, 215–224, 1997

[16] Cox, D. R. and Snell, E. J., *A general definition of residuals with discussion*, Journal of Royal Statistical Society: Series A, 30: 248-275, 1968

Appendix 1

Pearson Theorem

Let (X, Y) be random vectors with joint normal distribution then

$$\rho(X,Y) = 2\sin(\frac{\pi}{6}\rho r(X,Y))$$

Proof of Pearson Theorem

Density function of the standard normal vector *(X, Y)* with product moment correlation ρ is the following

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp[-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}]$$

Ranks measured around their mean 0.5:

$$\zeta = \int_{-\infty}^{\infty}\int_{-\infty}^{x} fdvdy = \frac{1}{\sqrt{2\pi}}\int_{0}^{x} e^{-\frac{v^2}{2}}dv$$

$$\eta = \int_{-\infty}^{\infty}\int_{-\infty}^{y} fdwdx = \frac{1}{\sqrt{2\pi}}\int_{0}^{y} e^{-\frac{w^2}{2}}dw$$

Then we have that the rank correlation is

$$\rho r = 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\zeta\eta fdxdy$$

and

$$\frac{\partial\rho r}{\partial\rho} = 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\zeta\eta\frac{\partial f}{\partial\rho}dxdy$$

Moreover we can show that

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial f}{\partial \rho}$$

Therefore,

$$\frac{\partial \rho r}{\partial \rho} = 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \zeta \eta \frac{\partial^2 f}{\partial x \partial y} dxdy$$

By a partial integration with respect to x we obtain

$$\frac{\partial \rho r}{\partial \rho} = -12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \eta \frac{\partial \zeta}{\partial x} \frac{\partial f}{\partial y} dxdy$$

Again by a partial integration with respect to y we get

$$\frac{\partial \rho r}{\partial \rho} = 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\partial \zeta}{\partial x} \frac{\partial \eta}{\partial y} fdxdy$$

Since

$$\frac{\partial \eta}{\partial y} = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$

$$\frac{\partial \zeta}{\partial x} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

we obtain

$$\frac{\partial \rho r}{\partial \rho} = \frac{12}{4\pi^2 \sqrt{1-\rho^2}} \int_{R^2} e^{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}} e^{-\frac{x^2 + y^2}{2}} dxdy = \frac{12}{4\pi^2 \sqrt{1-\rho^2}} \int_{R^2} e^{-\frac{(2-\rho^2)x^2 - 2\rho xy + (2-\rho^2)y^2}{2(1-\rho^2)}} dxdy$$

which can be written in the following form

$$\frac{\partial \rho r}{\partial \rho} = \frac{3}{\pi^2 \sqrt{1-\rho^2}} \int_{R^2} e^{-\frac{1}{2}([\sqrt{\frac{4-\rho^2}{2-\rho^2}} x]^2 + [\frac{\sqrt{2-\rho^2}\, y - \frac{\rho}{\sqrt{2-\rho^2}} x}{\sqrt{2-\rho^2}}]^2)} dxdy$$

Applying substitution

$$s = \sqrt{\frac{4-\rho^2}{2-\rho^2}}x$$

$$t = \frac{\sqrt{2-\rho^2}\,y - \frac{\rho}{\sqrt{2-\rho^2}}x}{\sqrt{2-\rho^2}}$$

With Jacbian

$$\sqrt{\frac{1-\rho^2}{4-\rho^2}}$$

It can be obtained

$$\frac{\partial \rho r}{\partial \rho} = \frac{6}{\pi\sqrt{4-\rho^2}}$$

From abve we get

$$\rho r = \frac{6}{\pi}\arcsin(\frac{\rho}{2})$$

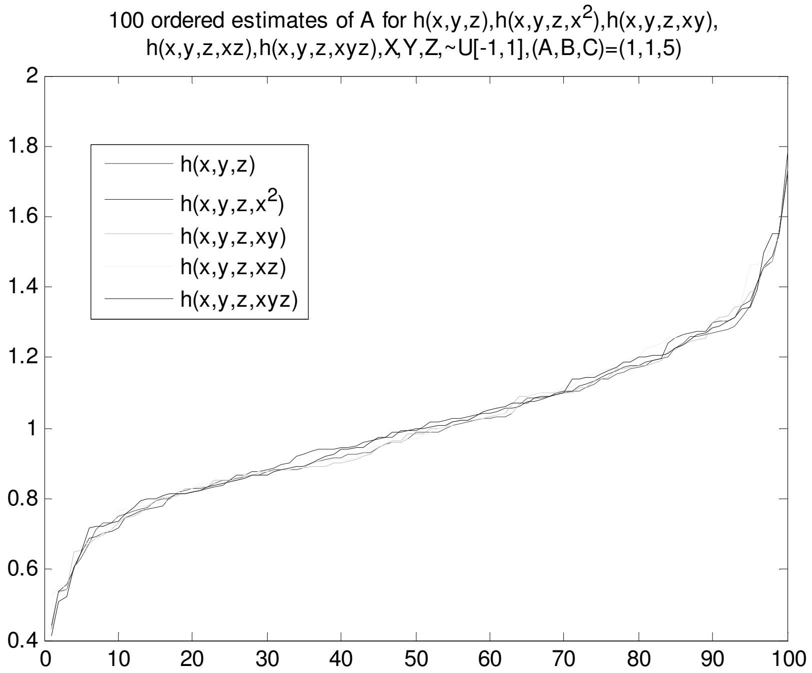
or written as

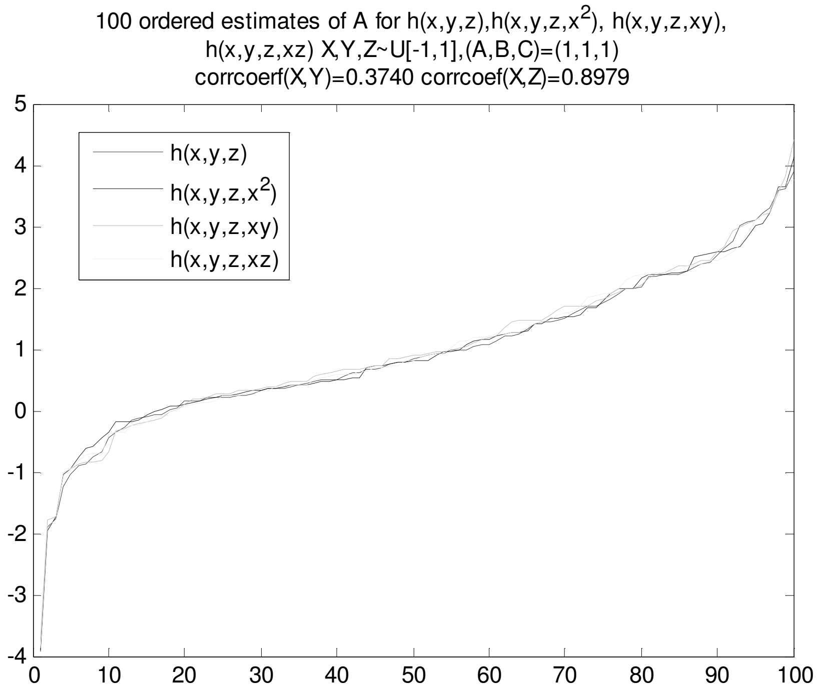
$$\rho = 2\sin(\frac{6}{\pi}\rho r)$$

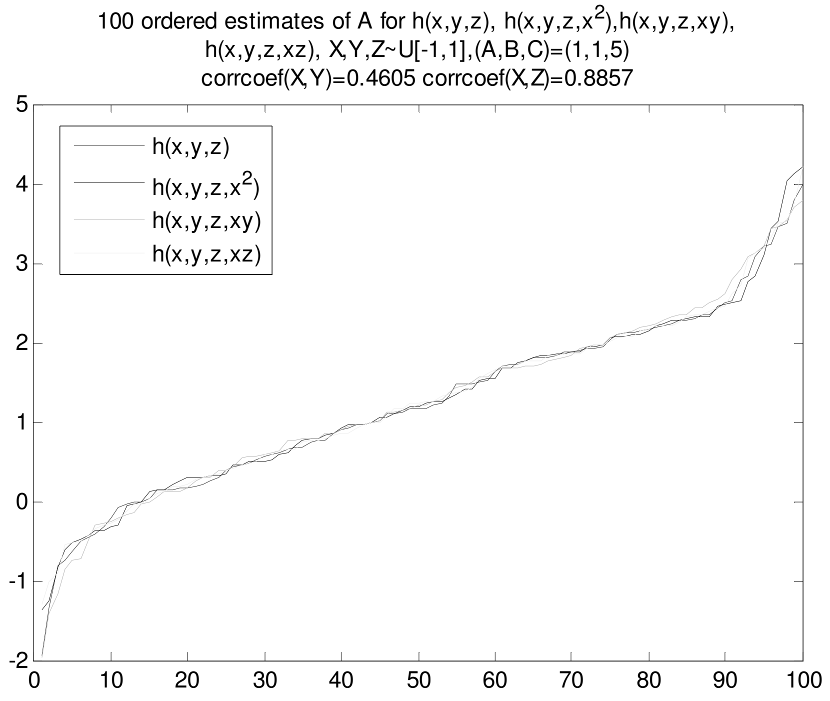
This finishes the proof.

Appendix 2

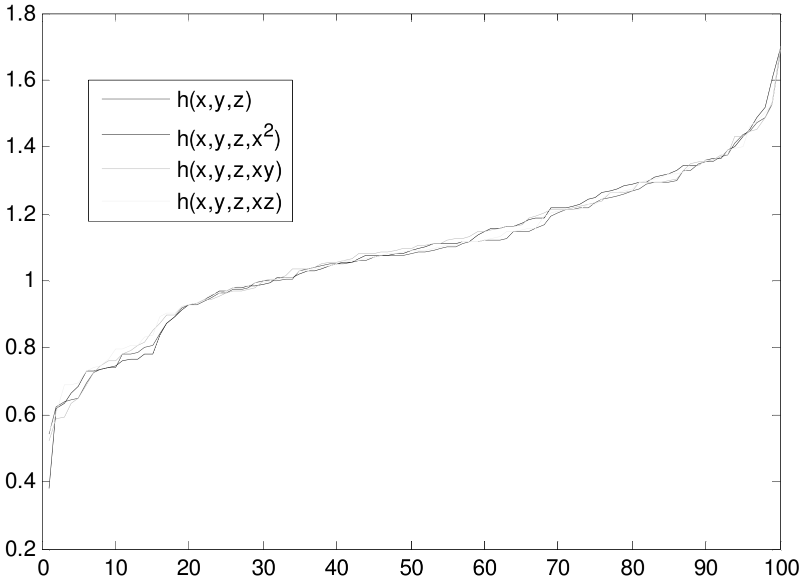
Additional plots for Section 3.1



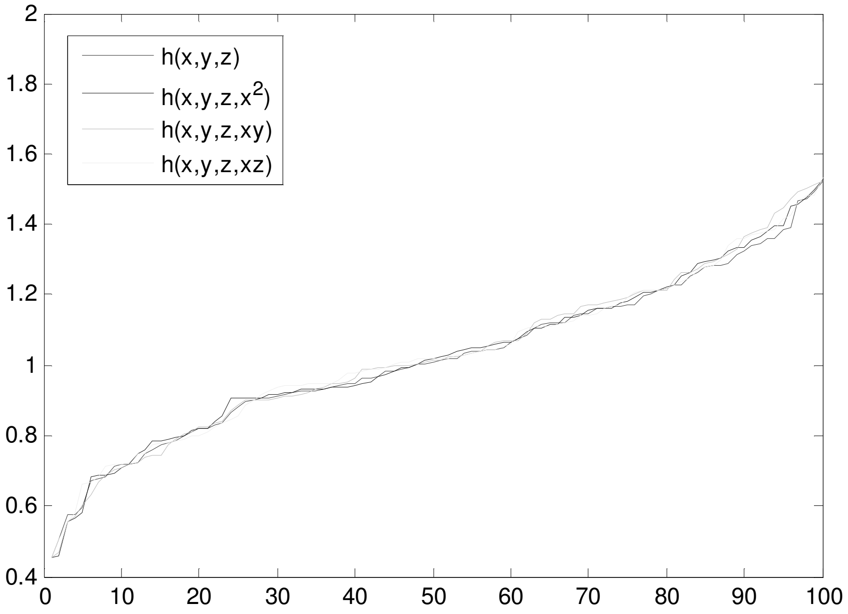

100 ordered estimates of A for h(x,y,z),h(x,y,z,x^2), h(x,y,z,xy),
h(x,y,z,xz) X,Y,Z~U[-1,1],(A,B,C)=(1,1,1)
corrcoerf(X,Y)=0.3740 corrcoef(X,Z)=0.8979
h(x,y,z)
h(x,y,z,x^2)
h(x,y,z,xy)
h(x,y,z,xz)
5
4
3
2
1
0
-1
-2
-3
-4
0
10
20
30
40
50
60
70
80
90
100

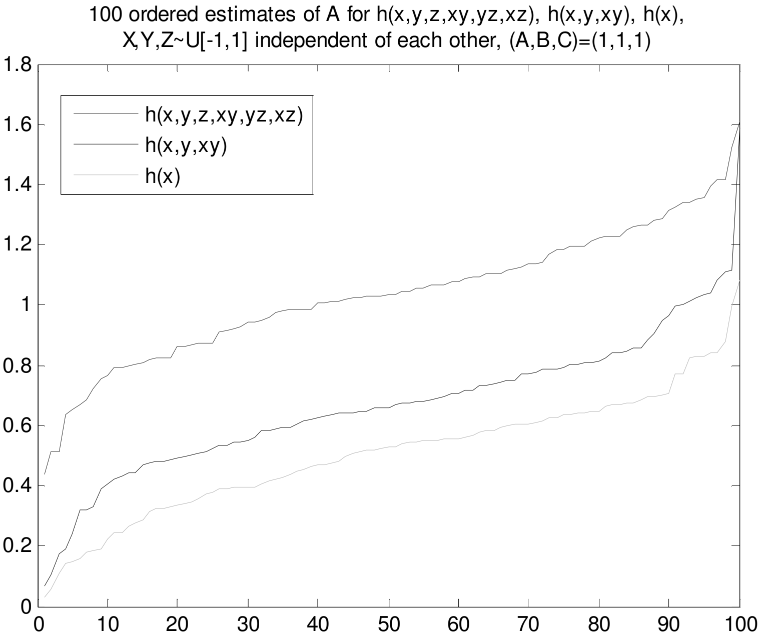

100 ordered estimates of A for h(x,y,z), h(x,y,z,x^2),h(x,y,z,xy),
h(x,y,z,xz), X,Y,Z~U[-1,1],(A,B,C)=(1,1,5)
corrcoef(X,Y)=0.4605 corrcoef(X,Z)=0.8857
h(x,y,z)
h(x,y,z,x^2)
h(x,y,z,xy)
h(x,y,z,xz)
5
4
3
2
1
0
-1
-2
0
10
20
30
40
50
60
70
80
90
100

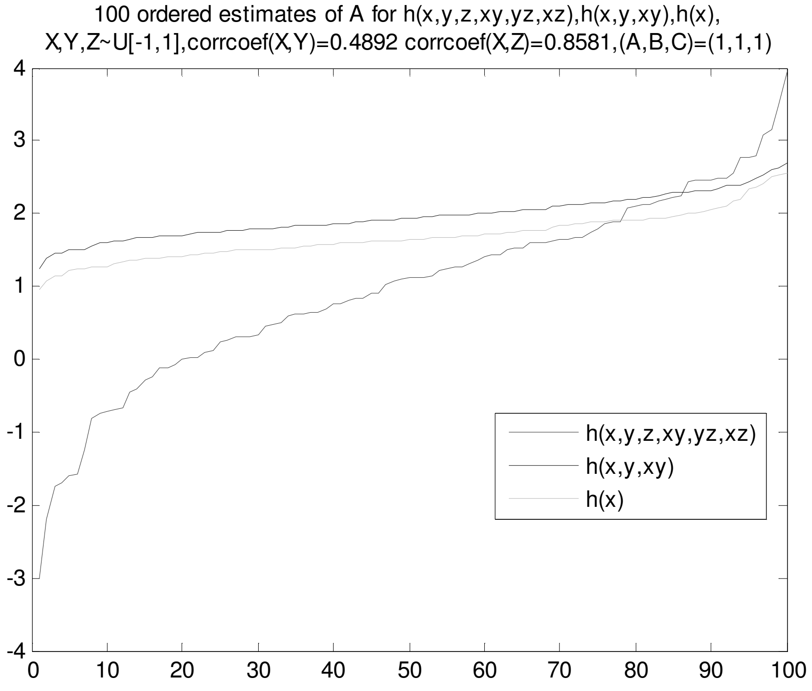
100 ordered estimates of A for h(x,y,z), h(x,y,z,x^2),h(x,y,z,xy),
h(x,y,z,xz) X,Y,Z,~U[-1,1],(A,B,C)=(1,1,1)
corrcoef(X,Y)= 0.0423 corrcoef(X,Z)=0.2007

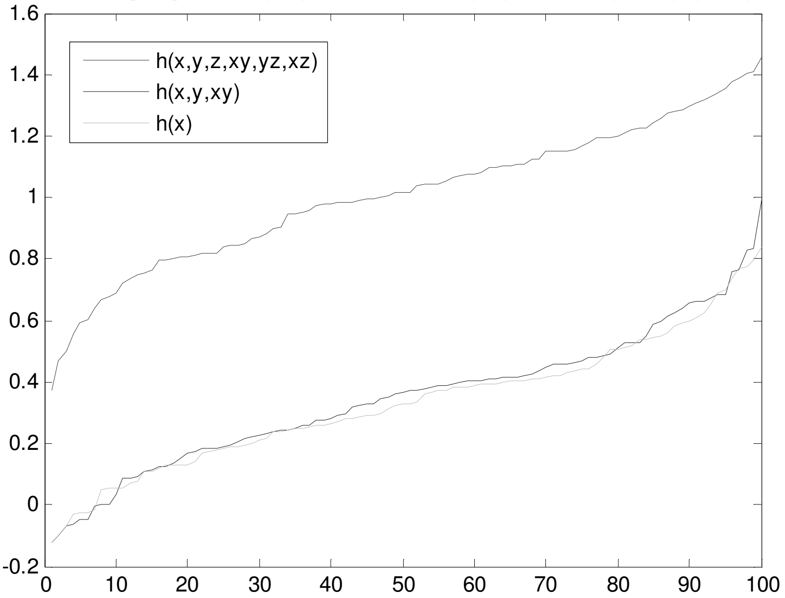


100 ordered estimates of A for h(x,y,z), h(x,y,z,x^2),h(x,y,z,xy),
h(x,y,z,xz),X,Y,Z~U[-1,1],(A,B,C)=(1,1,5)
corrcoef(X,Y)=0.09 corrcoef(X,Z)=-0.0729

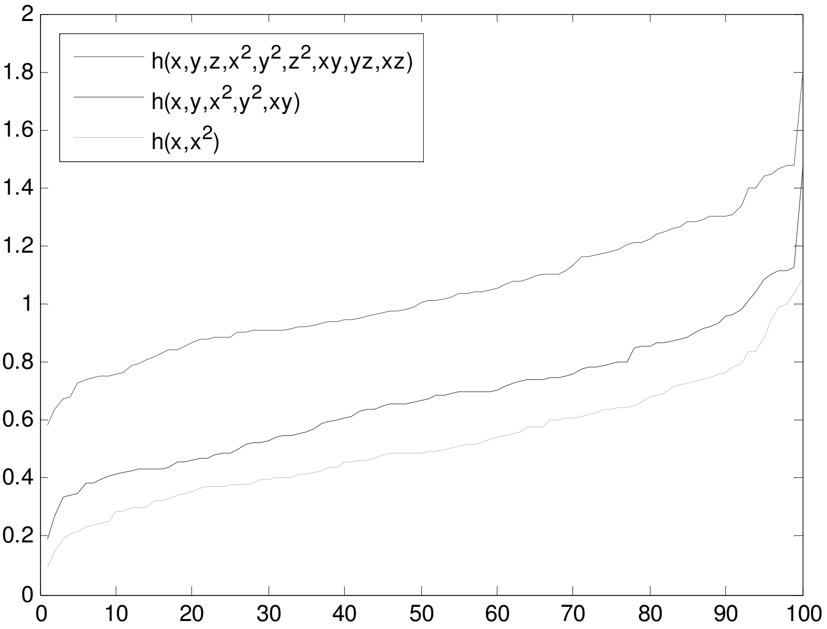



100 ordered estimates of A for h(x,y,z,xy,yz,xz), h(x,y,xy), h(x),
X,Y,Z~U[-1,1] independent of each other, (A,B,C)=(1,1,1)
h(x,y,z,xy,yz,xz)
h(x,y,xy)
h(x)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
0
10
20
30
40
50
60
70
80
90
100

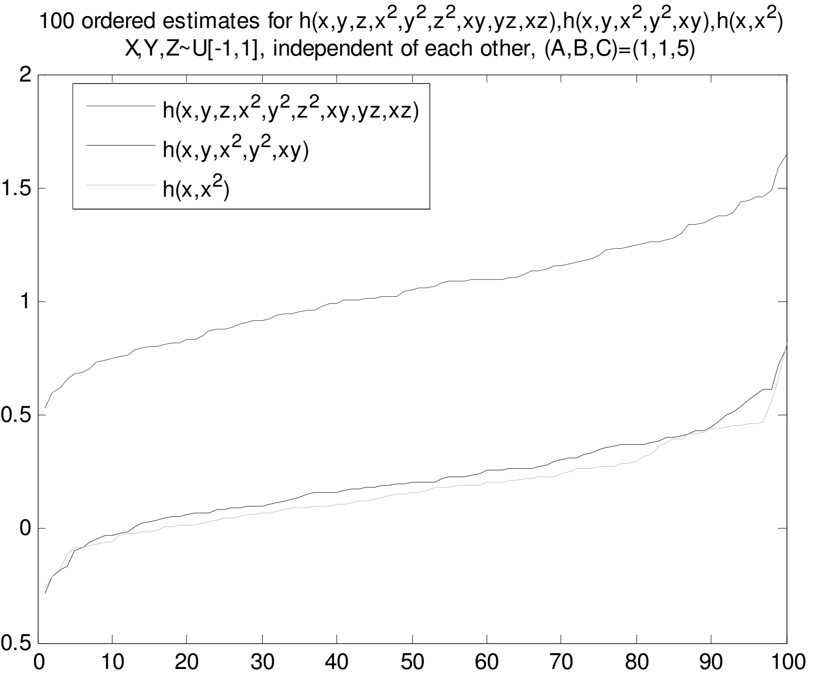

100 ordered estimates of A for h(x,y,z,xy,yz,xz),h(x,y,xy),h(x),
X,Y,Z~U[-1,1],corrcoef(X,Y)=0.4892 corrcoef(X,Z)=0.8581,(A,B,C)=(1,1,1)
h(x,y,z,xy,yz,xz)
h(x,y,xy)
h(x)
-4
-3
-2
-1
0
1
2
3
4
0
10
20
30
40
50
60
70
80
90
100

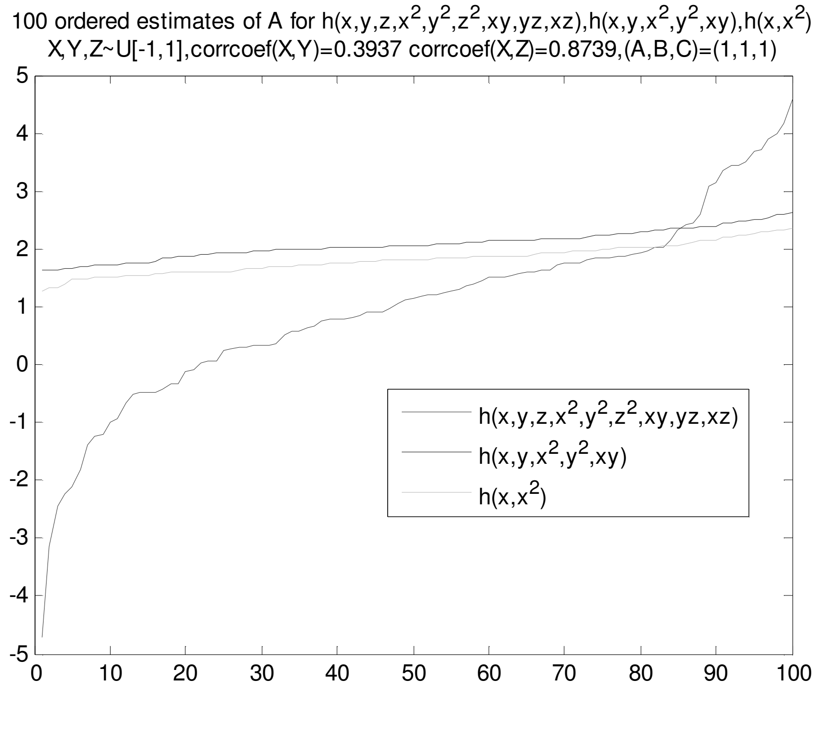


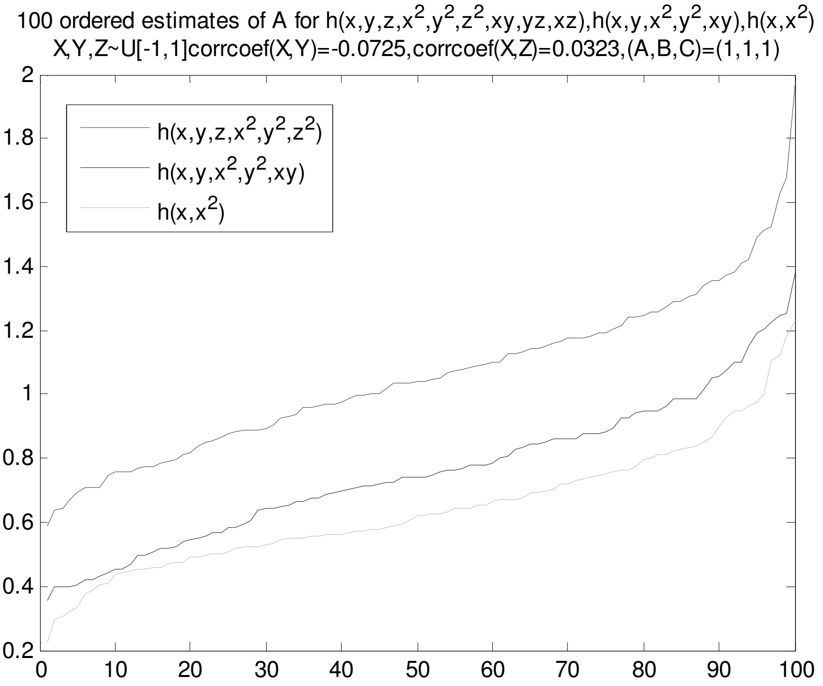
100 ordered estimates of A for h(x,y,z,xy,yz,xz), h(x,y,xy),h(x),
X,Y,Z~U[-1,1],corrcoef(X,Y)=0.1006,corrcoef(X,Z)=-0.0329, (A,B,C)=(1,1,5)



100 ordered estimates of A for $h(x,y,z,x^2,y^2,z^2,xy,yz,xz)$,$h(x,y,x^2,y^2,xy)$,$h(x,x^2)$,
X,Y,Z~U[-1,1], independent of each other, (A,B,C)=(1,1,1)


100 ordered estimates for h(x,y,z,x^2,y^2,z^2,xy,yz,xz),h(x,y,x^2,y^2,xy),h(x,x^2)
X,Y,Z~U[-1,1], independent of each other, (A,B,C)=(1,1,5)
h(x,y,z,x^2,y^2,z^2,xy,yz,xz)
h(x,y,x^2,y^2,xy)
h(x,x^2)
2
1.5
1
0.5
0
-0.5
0
10
20
30
40
50
60
70
80
90
100

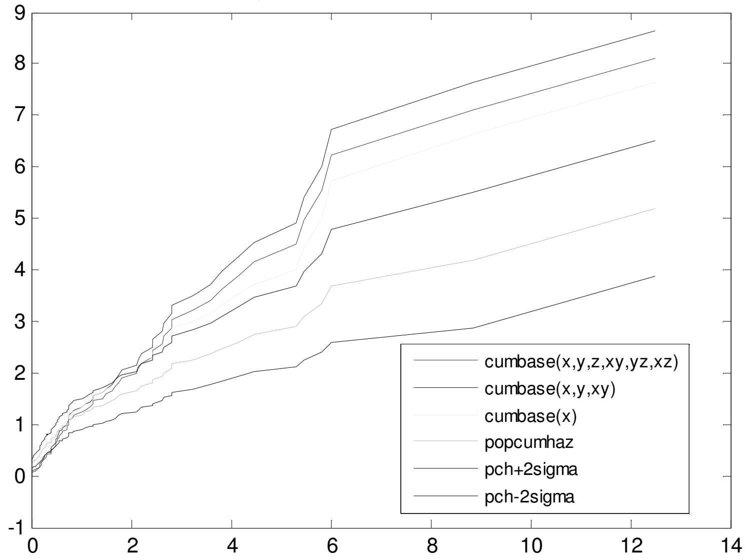

100 ordered estimates of A for h(x,y,z,x^2,y^2,z^2,xy,yz,xz),h(x,y,x^2,y^2,xy),h(x,x^2)
X,Y,Z~U[-1,1],corrcoef(X,Y)=0.3937 corrcoef(X,Z)=0.8739,(A,B,C)=(1,1,1)
h(x,y,z,x^2,y^2,z^2,xy,yz,xz)
h(x,y,x^2,y^2,xy)
h(x,x^2)
5
4
3
2
1
0
-1
-2
-3
-4
-5
0
10
20
30
40
50
60
70
80
90
100


100 ordered estimates of A for h(x,y,z,x2,y2,z2,xy,yz,xz),h(x,y,x2,y2,xy),h(x,x2)
X,Y,Z~U[-1,1]corrcoef(X,Y)=-0.0725,corrcoef(X,Z)=0.0323,(A,B,C)=(1,1,1)
h(x,y,z,x2,y2,z2)
h(x,y,x2,y2,xy)
h(x,x2)
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
90
100

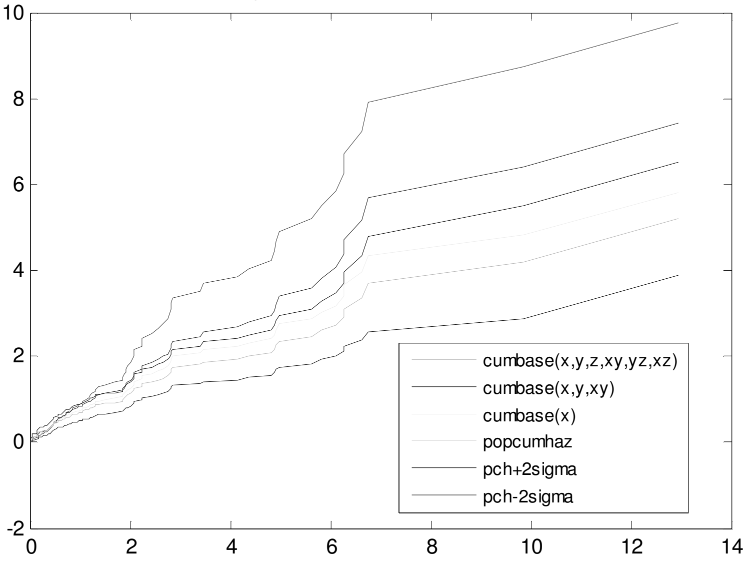
Appendix 3

Additional plots for Section 3.2

cumulative population and baseline hazard functions for h(x,y,z,xy,yz,xz),h(x,y,xy),h(x),
X,Y,Z~U[-1,1],corrcoef(X,Y)=0.4443 corrcoef(X,Z)=0.8899, (A,B,C)=(1,1,1),
with 2-sigma confidence bands (black lines)



cumulative population and baseline hazard functions for h(x,y,z,xy,yz,xz),h(x,y,xy),h(x),
X,Y,Z~U[-1,1],corrcoef(X,Y)=-0.0011corrcoef(X,Z)=0.0513,(A,B,C)=(1,1,1)
with 2-sigma confidence bands (black lines)



Appendix 4

Tables for Section 4.1

$Pr(\chi^2(n)\geq G)$	G1	G2	G3	G4	G5	G6
Simulation1	1.3415e-006	2.3548e-013	1.7529e-008	5.9960e-012	9.1381e-008	1.8059e-012
Simulation2	7.7231e-008	4.0251e-012	4.9890e-010	5.2125e-013	1.8379e-011	1.7764e-015
Simulation3	2.3599e-006	1.8258e-011	2.9640e-012	6.6613e-016	7.2525e-011	1.1102e-015
Simulation4	2.2206e-003	2.9128e-007	1.7361e-005	1.8773e-010	8.7909e-007	6.3607e-012
Simulation5	1.4966e-007	1.1102e-015	3.2687e-008	6.4393e-015	3.4579e-008	2.2315e-014
Simulation6	1.6544e-005	8.2868e-009	8.9964e-008	3.1288e-010	3.6310e-008	5.6234e-012
Simulation7	2.1681e-004	3.8520e-009	1.1252e-006	5.7676e-013	4.7510e-008	6.5048e-013
Simulation8	2.7987e-005	2.0100e-008	5.5872e-010	1.1335e-012	1.0898e-010	3.4717e-013
Simulation9	8.1895e-006	1.9696e-011	2.4526e-005	6.3114e-011	1.6577e-007	2.4860e-011
Simulation10	4.6840e-006	9.2930e-011	3.7270e-006	1.4997e-008	2.3297e-004	8.1172e-006

Likelihood Ratio Test for Independent covariates

$Pr(\chi^2(n)\geq G)$	G7	G8	G9	G10	G11	G12
Simulation1	2.5901e-001	1.3053e-007	3.0026e-003	1.0031e-012	2.0035e-002	1.3940e-009
Simulation2	1.5331e-001	2.0546e-010	1.5878e-002	1.9207e-014	5.3600e-003	8.6553e-013
Simulation3	2.0945e-001	4.6584e-009	7.1778e-002	2.1538e-014	1.8493e-002	9.7700e-015
Simulation4	2.1293e-001	5.5865e-010	2.8448e-003	4.2633e-014	4.2701e-002	4.5286e-013
Simulation5	2.9425e-001	4.2245e-009	1.0234e-004	3.6926e-013	1.3295e-002	3.9541e-011
Simulation6	2.8013e-001	6.3442e-007	2.6363e-002	3.7818e-011	1.7017e-002	9.2381e-010
Simulation7	5.3966e-001	4.4694e-011	1.6541e-002	1.4988e-014	1.6124e-001	3.5527e-015
Simulation8	5.2999e-001	6.4418e-006	5.7771e-002	1.1940e-012	2.4034e-002	4.8591e-010
Simulation9	4.6604e-001	6.4418e-008	9.8655e-002	3.4084e-014	1.4403e-001	2.1118e-012
Simulation10	3.5509e-001	4.3317e-007	1.9452e-004	6.4837e-014	8.3254e-005	3.2032e-012

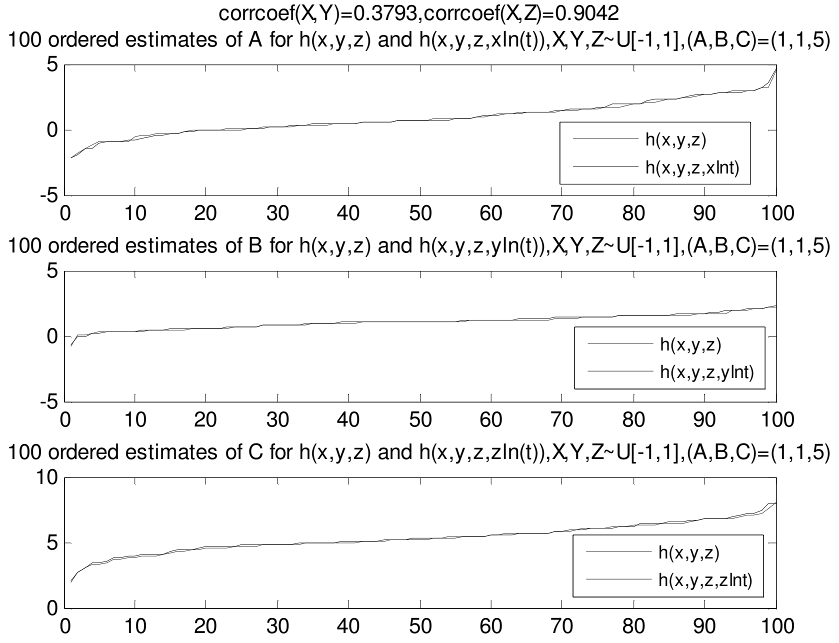
Likelihood Ratio Test with Strong Dependence between Covariates

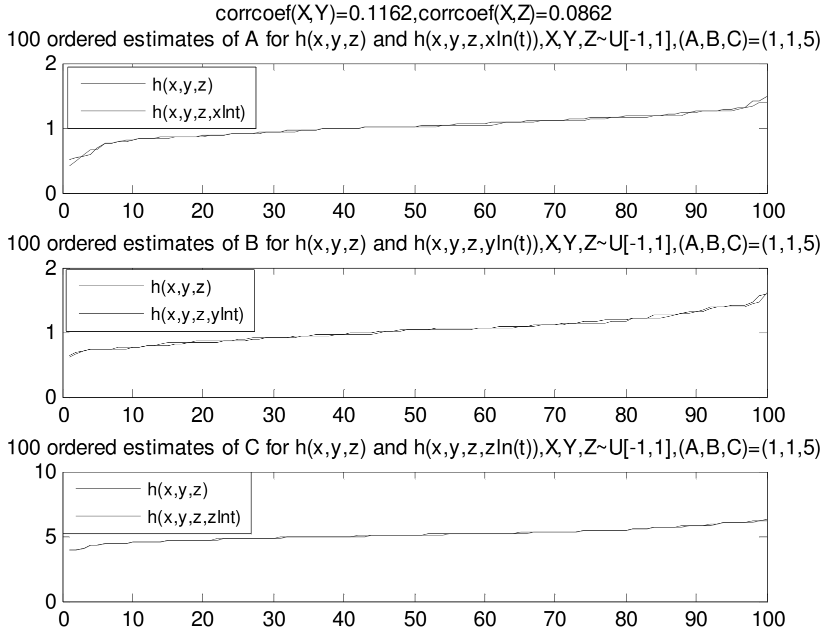
$Pr(\chi^2(n)\geq G)$	G13	G14	G15	G16	G17	G18
Simulation1	4.5819e-006	5.3557e-013	4.2481e-008	5.5511e-016	1.1840e-007	1.2546e-014
Simulation2	8.8171e-007	2.4114e-013	1.4465e-009	2.6201e-014	1.0287e-010	8.5487e-015
Simulation3	1.8026e-006	1.8562e-010	1.0288e-009	7.8137e-013	6.0133e-011	1.8319e-014
Simulation4	5.3727e-005	1.6965e-008	4.2787e-008	4.9574e-012	4.3013e-010	3.6637e-015
Simulation5	1.0055e-004	1.3514e-011	6.7195e-009	2.5435e-013	4.2951e-010	4.4409e-016
Simulation6	5.4273e-005	1.3717e-011	2.0629e-005	9.2226e-013	6.0984e-006	4.3322e-012
Simulation7	1.9149e-003	1.2750e-006	4.6106e-005	6.4430e-008	1.5335e-007	8.0461e-012
Simulation8	8.2769e-004	1.1795e-009	3.2055e-008	2.1965e-011	1.5458e-008	1.0466e-012
Simulation9	3.2331e-004	4.7785e-009	1.0890e-006	2.8092e-009	2.3362e-007	7.4496e-014
Simulation10	1.7482e-005	7.8251e-012	4.0866e-009	3.0642e-014	5.7860e-009	1.5654e-014

Likelihood Ratio Test with Weak Dependence between Covariates

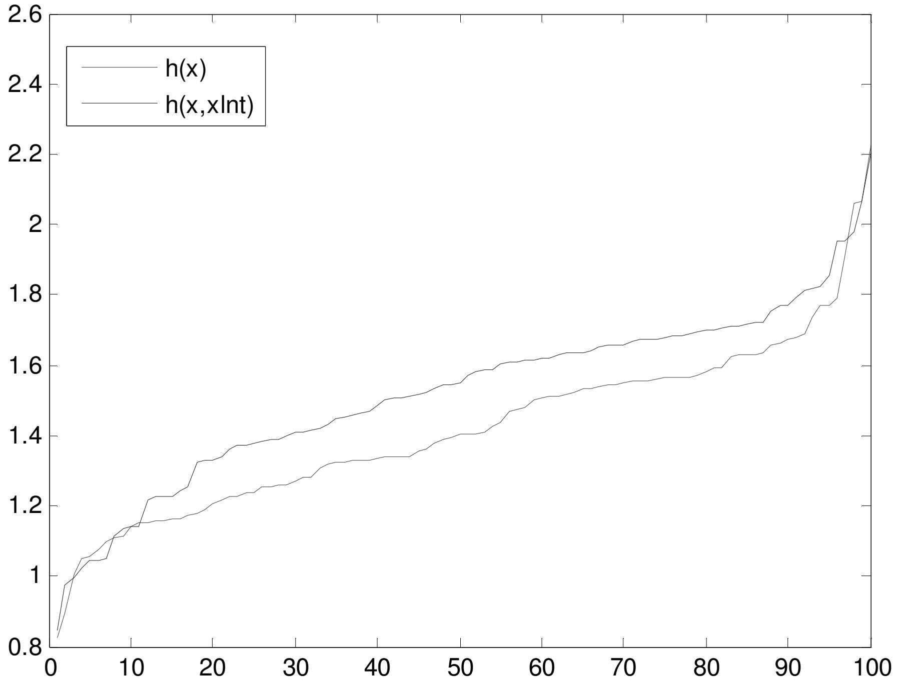
Appendix 5

Additional plots for Section 4.2

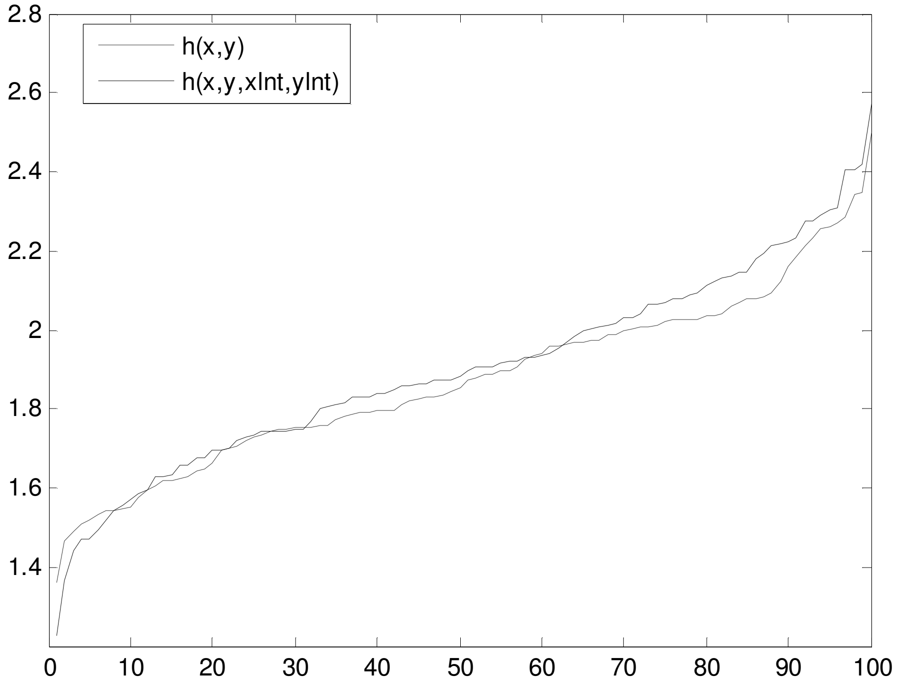




100 ordered estimates of A for h(x) and h(x,xInt), X,Y,Z~U[-1,1],(A,B,C)=(5,5,5)
corrcoef(X,Y)=0.1576,corrcoef(X,Z)=0.2124,corrcoef(Y,Z)=0.0795

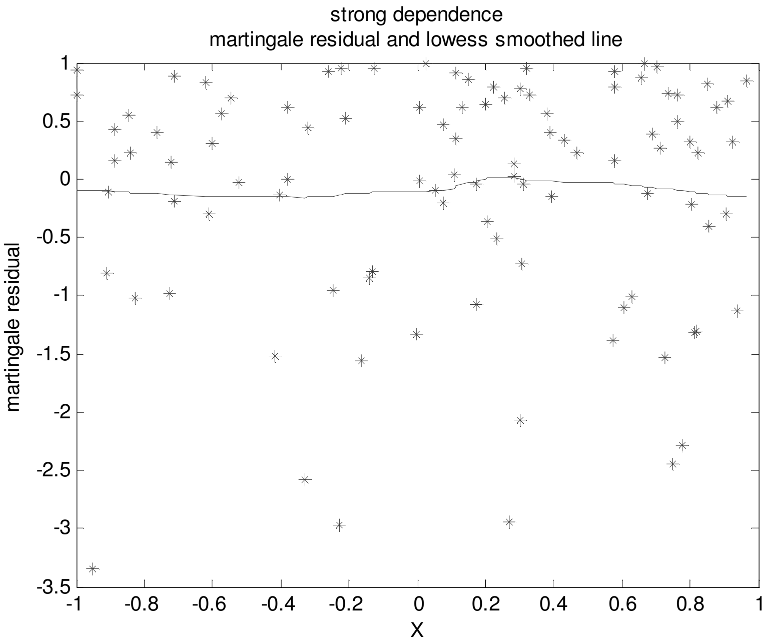


100 estimates of A for h(x,y) and h(x,y,xInt,yInt),X,Y,Z~U[-1,1],(A,B,C)=(5,5,5)
corrcoef(X,Y)=-0.1192, corrcoef(X,Z)=-0.0700,corrcoef(Y,Z)=0.1324

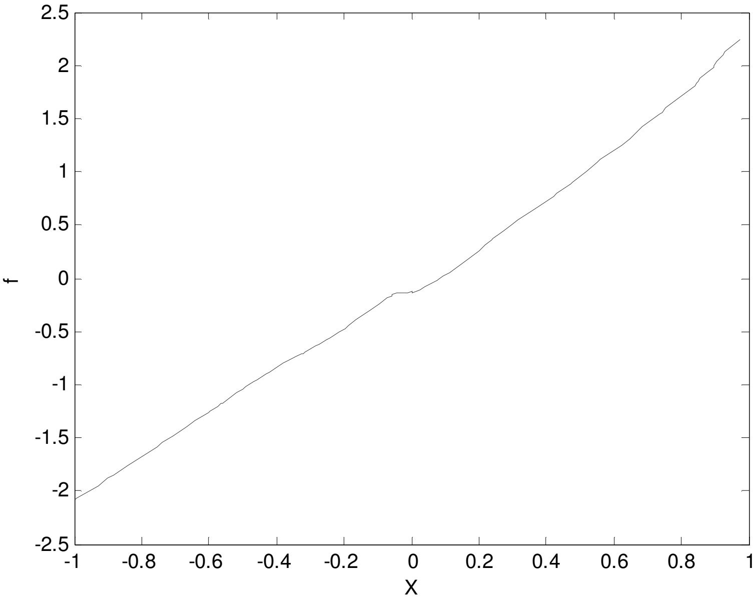


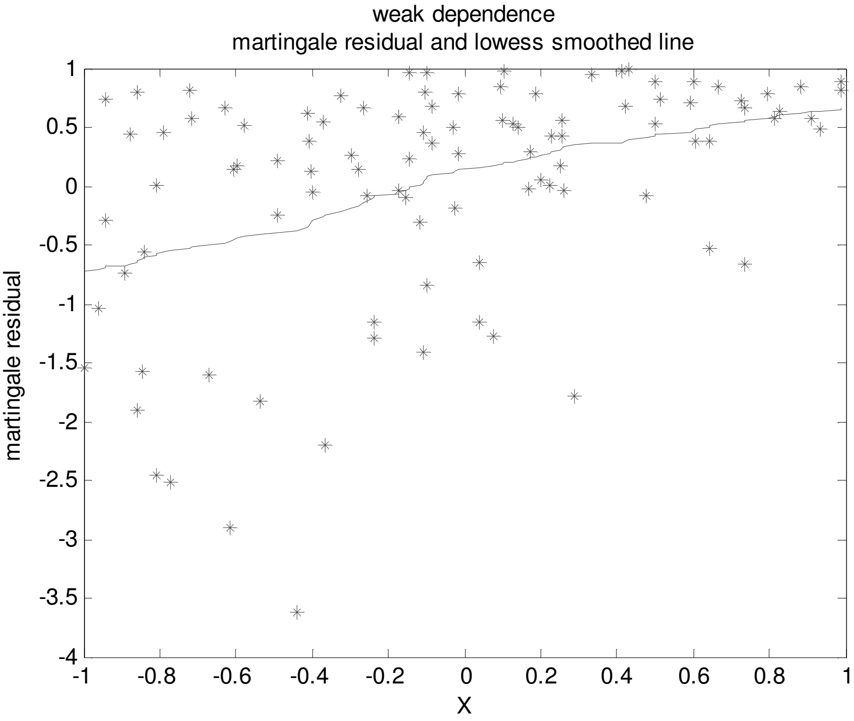
Appendix 6

Additional plots for Section 4.3

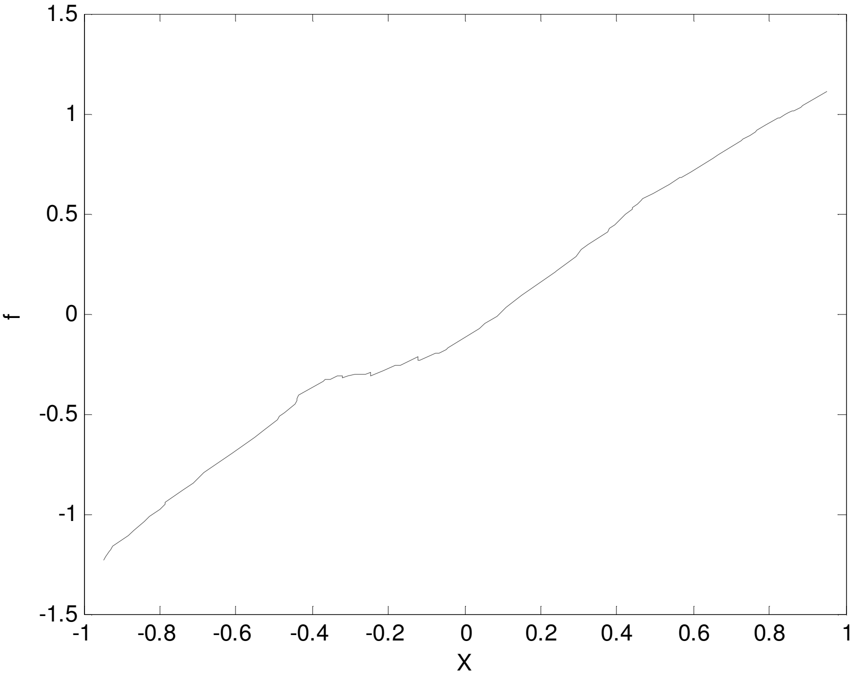


f_i versus x_i for $h(x,y,z)$, strong dependence





f_i versus x_i for $h(x,y,z)$, weak dependence


f
X
1.5
1
0.5
0
-0.5
-1
-1.5
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1