A STUDY OF THE ROBUSTNESS OF COX'S PROPORTIONAL HAZARDS MODEL USED IN TESTING FOR COVARIATE EFFECTS

by

MINGWEI FEI

M.S., Shanghai Jiaotong University, China, 2006

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2012

Approved by:

Major Professor
Paul Nelson

Copyright

Abstract

There are two important statistical models for multivariate survival analysis, proportional hazards(PH) models and accelerated failure time(AFT) model. PH analysis is most commonly used multivariate approach for analysing survival time data. For example, in clinical investigations where several (known) quantities or covariates, potentially affect patient prognosis, it is often desirable to investigate one factor effect adjust for the impact of others. This report offered a solution to choose appropriate model in testing covariate effects under different situations.

In real life, we are very likely to just have limited sample size and censoring rates(people dropping off), which cause difficulty in statistical analysis. In this report, each dataset is randomly repeated 1000 times from three different distributions (Weibull, Lognormal and Loglogistc) with combination of sample sizes and censoring rates. Then both models are evaluated by hypothesis testing of covariate effect using the simulated data using the derived statistics, power, type I error rate and covergence rate for each situation.

We would recommend PH method when sample size is small($n \leq 20$) and censoring rate is high($p \geq 0.8$). In this case, both PH and AFT analyses may not be suitable for hypothesis testing, but PH analysis is more robust and consistent than AFT analysis. And when sample size is 20 or above and censoring rate is 0.8 or below, AFT analysis will have slight higher convergence rate and power than PH, but not much improvement in Type I error rates when sample size is big($n \geq 50$) and censoring rate is low($p \leq 0.3$). Considering the privilege of not requiring knowledge of distribution for PH analysis, we concluded that PH analysis is robust in hypothesis testing for covariate effects using data generated from an AFT model.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my appreciation to the following persons for their continuous guidance and support all through my M.S. study:

- Dr. Paul Nelson for his great help, advice, and patience during the research and writing of this report.
- Dr. James Neill, and Dr. Nora Bello for serving on my committee.
- Professors and Staff in Statistic department for great teaching, service and help during my entire study.
- My family, Kebing Yu and Julianne Yu, for their continuous love, support, and bearing with me all these years.
- Finally I thank my parents, Jinyu Fei, Jianlan Zhang, and all of my friends for their encouragement and help.

Dedication

Dedicate to my parents, Jinyu Fei, Jianlan Zhang, my husband Kebing Yu, and my little girl, Julianne Yu.

Chapter 1

Introduction

Survival analysis examines and models the times it takes for events recorded on experimental units to occur. The term 'survival' arose from early applications where the event was death or component failure. Nowadays survival analysis has been applied to a variety of areas, such as economic, public health, industry, etc. The survival analysis is competitive for prediction in comparison with usual or logistic regression.[2]Early work in survival analysis ignored heterogeneity among the units on which event times were recorded and analyzed data as being a random sample from a family of continuous distributions specified up to an unknown parameter. This early approach did not adjust the event times for measurable differences in the units, which can, in the modern era, be accounted for by incorporating covariates in the model. For example, covariates such as weight, age and smoking status of individuals could have important effects on their lifetimes. In a clinical trial, covariates are used to represent different treatments and/or treatment doses. In reliability, covariates such as the turning speed of a machine tool or the stress applied to a component can affect the lifetime of a component.

Survival analysis provides a framework for the inclusion as time-varying covariates, such as macroeconomic variables, interest rate and unemployment index. But in this report, we only study the fixed covariate effect. Suppose that based on a random sample $\{(T_i, z_i); i = 1, 2, \ldots, n\}$, lifetimes $\{T_i\}$, covariates $\{z_i\}$ represent characteristics of the unit on which lifetime is recorded and do not change over time. It is desired to test if the components of a

fixed covariate vector z jointly effect the distribution of lifetime X. For example, the useful life of a battery life may be heavily influenced by environment such as heat and moisture but not by the gender of the driver of the car.

The effect of covariates on survival often complicates the analysis of a set of lifetime data. Two methods that are often used to incorporate the effect of covariates on lifetimes are based on the *accelerated failure time (AFT)* and the *proportional hazards (PH) models*, the latter developed by Cox (1972). Accelerated life models are parametric and PH models are referred to as being semi-parametric. Both models make strong assumptions. The purpose of this report is to study via simulation the robustness of the PH model in testing for covariate effects when the data come from an accelerated life model whose hazards may not be proportional. As described below, my study will allow right random censoring of the event times. I will not investigate diagnostic methods for assessing the validity of each model.

The survival function and the hazard function are important descriptions of lifetime distributions. For a positive, continuous random variable having probability density function f(x), the probability of an individual surviving beyond time x (experiencing the event after time x) is given by the survivor function defined as:

$$S(x) = Pr(X > x) = \int_x^\infty f(t)dt \tag{1.1}$$

The hazard function is the instantaneous failure rate, also known as the conditional failure rate is defined by

$$h(x) = \lim_{\triangle x \to 0} \frac{P[x \leq X < x + \triangle x \mid X \geq x]}{\triangle x} = \frac{f(x)}{S(x)} = -d \ln \frac{[S(x)]}{dx} \tag{1.2}$$

There are some typical hazard functions[3] used in real application shown in 1.1:

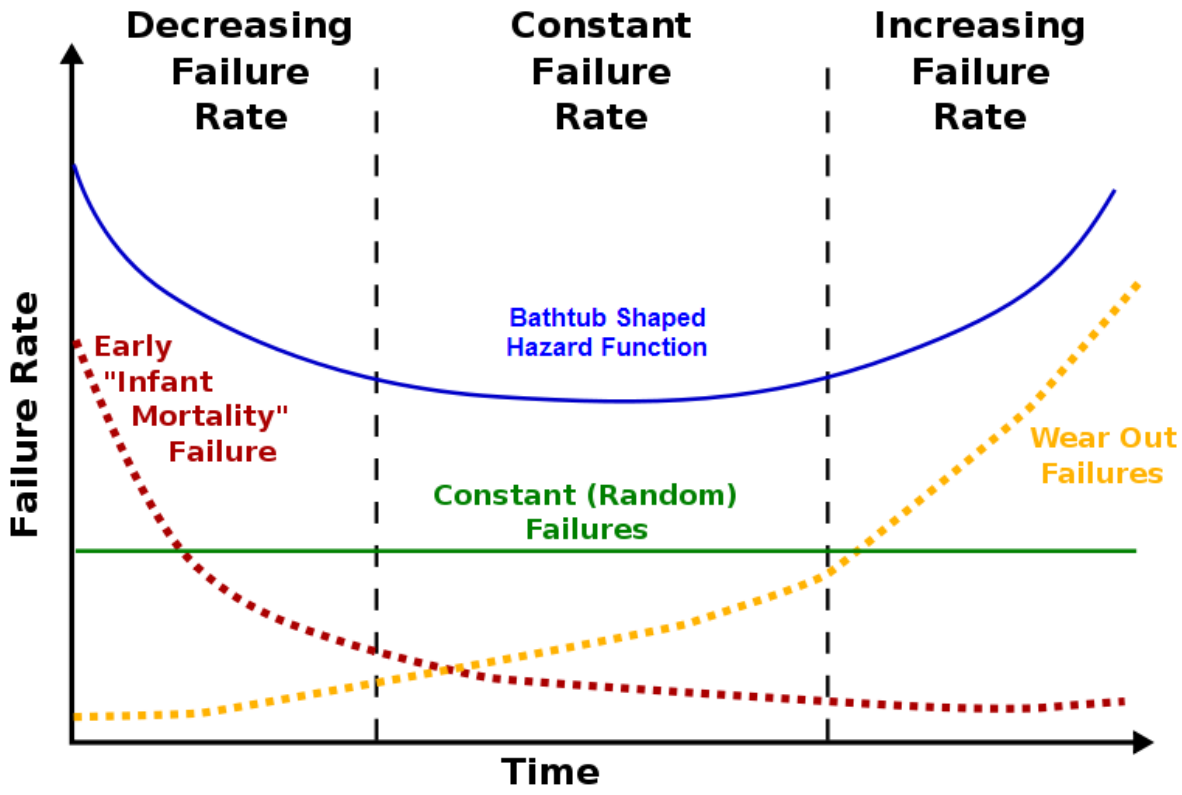


Figure 1.1: Hazard functions

Censoring is a form of missing data problem which is common in survival analysis. If it is known only that the date of death is after some date, this is called right censoring. Right censoring will occur for those subjects whose birth date is known but who are still alive when they are lost to follow-up or when the study ends. If a subject's lifetime is known to be less than a certain duration, the lifetime is said to be left-censored. $\{X_i\}$ denotes the lifetime, and $\{C_i\}$ is the censoring time for $i = 0, 1, \ldots, n$. The actual observations consists of $\{T_i, \delta_i\}$, where $T_i = \min\{X_i, C_i\}$ and $\delta_i = I(X_i \leq C_i)$ is an indicator of the censoring status of $\{T_i\}$. The right censored data is actually observed by the vectors $\{(T_i, \delta_i = 0, z_i)\}$, where, for unobserved censoring variables $\{(T_i, \delta_i = 1, z_i)\}$.

Examples of right censoring are shown in 1.2:

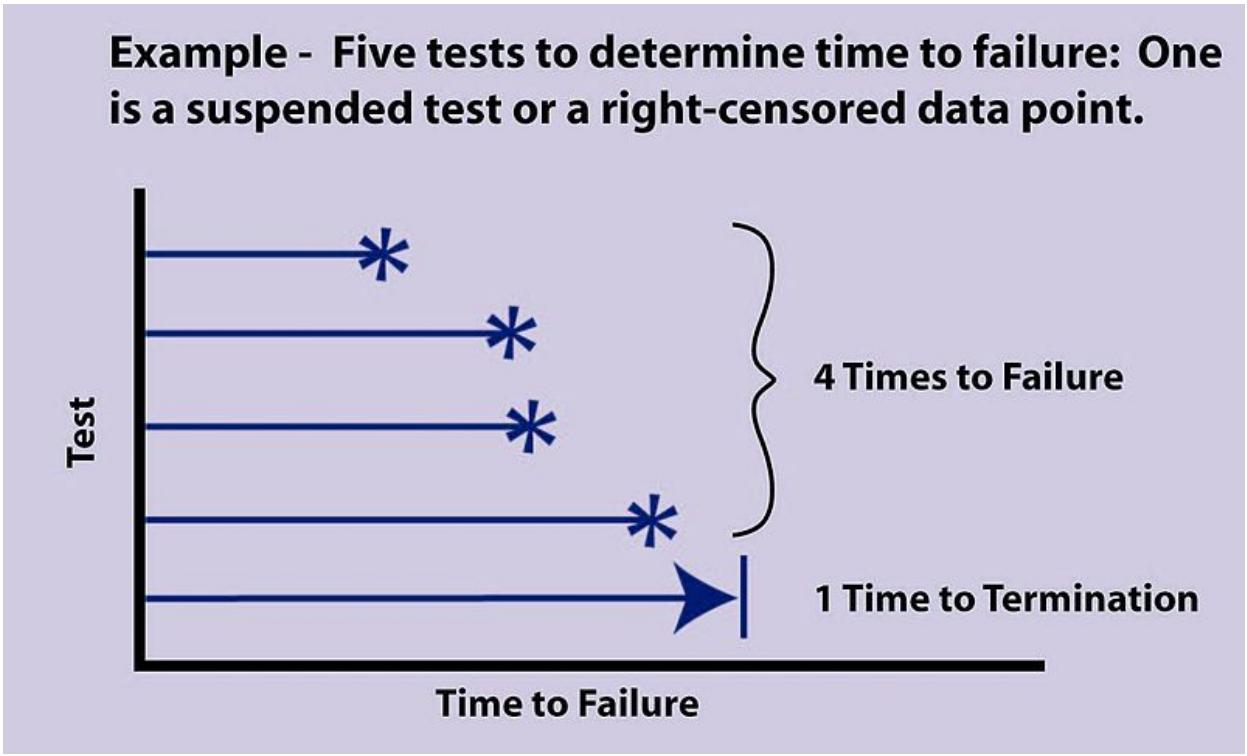


Figure 1.2: Right censoring and uncensoring

1.1 Proportional Hazards Models

Cox's (1972) semi-parametric proportional hazards (PH) model relates the distribution of continuous lifetime X to covariate z through a function of the form $g(z, \beta)$,usually calibrated so that the parameter vector β is zero only if the covariates acting jointly play no role in the distribution of event times. Cox's PH model is used widely in model selection.[2] The PH model assumes that the hazard function of a component having covariate vector z, denoted $h(\cdot|z)$ usually has the from

$$h(t|z) = \exp(z'\beta)h_0(t) \tag{1.3}$$

where $h_0(\cdot)$ is an unknown, continuous, baseline hazard function and $g(z, \beta) = \exp(z'\beta)$. The PH model implies that $h(t|z_1)/h(t|z_2)$ is free of time t for all pairs of covariate vectors z_1 and z_2, hence the term *proportional hazards*, which is a very restrictive assumption. Since the baseline hazard portion of the model is unspecified and the influences of the explanatory variables are described in a parametric linear-regression type model, the Cox model is said to be *semi-parametric*. Inference about β can be carried out using Cox's (1972) partial likelihood, denoted $L_p(\beta)$. I will focus on testing for a covariate effect by testing

$$H_0 : \beta = 0 \quad \text{v.s.} \quad H_0 : \beta \neq 0 \tag{1.4}$$

1.2 Partial likelihoods for PH model

As indicated earlier, our data are based on a sample of size n consisting of the triple $(T_j, \delta_j, z_j), j = 1, 2, \ldots, n$. Recall that due to censoring, we observe $T_i = \min(Y_i, C_i), \delta_i = I(Y_i \leq C_i)$. We assume that censoring is non-informative in that, given z_j , the event and censoring time for the jth observation are independent, $j = 1, 2, \ldots, n$ and that the distribution of censoring time is free of unknown parameters. For simplicity, suppose there are no ties between the event times. Let $t_1 < t_2 < \cdots < t_D$ denote the ordered, distinct event

times and $Z_{(i)k}$ be the kth covariate associated with the individual whose failure time is t_i . The partial likelihood based on the hazard function as specified by Eq.(1.3), is expressed by reference[6],[4],[7] and[5]

$$L_p(\beta) = \prod_{i=1}^{D} \frac{exp[\sum_{k=1}^{p} \beta_k z_{(i)k}]}{\sum_{j \in R(t_i)} exp[\sum_{k=1}^{p} \beta_k z_{jk}]} \tag{1.5}$$

where $R(T_i) = \{j : T_j > T_i\}$ is the 'risk set' at time T_i. This partial likelihood can be treated as a standard likelihood and inference carried out by usual means. Let $LL(\beta) = ln(L_p(\beta))$. We can write $LL(\beta)$ as

$$LL(\beta) = \sum_{i=1}^{D} \sum_{k=1}^{p} \beta_k z_{(i)k} - \sum_{i=1}^{D} ln[\sum_{j \in R_(t_i)} exp[\sum_{k=1}^{p} \beta_k z_{jk}]]. \tag{1.6}$$

The partial maximum likelihood estimates for β are found by solving the equations obtained by setting the partial derivatives of $LL(\beta)$ with respect to β equal to zero. The partial information matrix is the negative of the matrix of second derivatives of the log likelihood and is given by $I(\beta) = [I_{gh}(\beta)]_{p \times p}$ with the (g, h) element given by

$$\begin{aligned} I_{gh}(\beta) = & \sum_{i=1}^{D} \frac{\sum_{j \in R_(t_i)} z_{jg} z_{jb} exp[\sum_{k=1}^{p} \beta_k z_{jk}]}{\sum_{j \in R_(t_i)} exp[\sum_{k=1}^{p} \beta_k z_{jk}]} \\ & - \sum_{i=1}^{D} \frac{[\sum_{j \in R_(t_i)} z_{jg} exp[\sum_{k=1}^{p} \beta_k z_{jk}]][\sum_{j \in R_(t_i)} z_{jh} exp[\sum_{k=1}^{p} \beta_k z_{jk}]]}{[\sum_{j \in R_(t_i)} exp[\sum_{k=1}^{p} \beta_k z_{jk}]][\sum_{j \in R_(t_i)} exp[\sum_{k=1}^{p} \beta_k z_{jk}]]} \end{aligned} \tag{1.7}$$

There are three main tests for hypothesis about regression parameters β . Let $\hat{\beta}_p = (\hat{\beta}_{p1}, \hat{\beta}_{p2}, \cdots, \hat{\beta}_{pk})$ denote the partial maximum likelihood estimates of β obtained as discussed above and let $I(\hat{\beta}_p)$ be the $k \times k$ information matrix evaluated at $\hat{\beta}_p$ and defined by Eq. (1.7). The Wald test of the global hypothesis of $H_0 : \beta = 0, \quad$ v.s.$H_a : \beta \neq 0$ uses the test statistic:

$$X_w^2 = \hat{\beta}_p' I(0) \hat{\beta}_p \tag{1.8}$$

which for large samples has approximately a chi-squared distribution with p degrees of freedom if H_0 is true.

The *Likelihood ratio test* of the global hypothesis uses the test statistic:

$$X^2_{LR} = 2[LL(\hat{\beta}_p) - LL(0)] \tag{1.9}$$

which for large samples has approximately a chi-squared distribution with p degrees of freedom if H_0 is true.

The Score test is based on the efficient scores, $U(\beta_p) = (U_1(\beta_{p1}), U_2(\beta_{p2}), \cdots, U_k(\beta_{pk}))$. The efficient score equation are found by taking partial derivatives of Eq.(1.6) with respect to β_p as follows. Let

$$U(\beta_p) = \partial LL(\beta_p)/\partial\beta_h, \quad h = 1, 2, \cdots, k \tag{1.10}$$

$$U_h(\beta_p) = \sum_{i=1}^{D} Z_{(i)h} - \sum_{i=1}^{D} \frac{\sum_{j\in R_(t_i)} Z_{jh} exp[\sum_{l=1}^{k}]\beta_l Z_{jl}}{\sum_{j\in R_(t_i)} exp[\sum_{l=1}^{k}]\beta_l Z_{jl}} \tag{1.11}$$

For large samples, $U(\beta_p)$ is asymptotically distributed as k-variates normal distribution with mean 0 and covariance $I(\beta_p)$ when H_0 is true. Hence, a test statistic with an asymptotic chi-square distribution is given by

$$X^2_{SC} = U(\hat{\beta}_p)' I^{-1}(0) U(\hat{\beta}_p) \tag{1.12}$$

1.3 Accelerated Failure Time model

In the accelerated failure time model (AFT), covariates act multiplicatively on lifetime, as given by

$$X = \exp(\mu + \gamma' z + \sigma W) = \exp(\gamma' z)\exp(\mu + \sigma W) = \exp(\gamma' z)\exp(\mu) T^{\sigma}, \sigma > 0 \tag{1.13}$$

where μ is a location parameter, σ a positive scale parameter and $T = exp(W)$ is a baseline lifetime whose distribution is fully specified. Here, a test for joint covariate effects becomes

$$H_0 : \gamma = \underline{0}, \quad \text{v.s.} \quad H_a : \gamma \neq \underline{0} \tag{1.14}$$

The accelerated failure model is best understood on a log transformed scale, $Y = \ln(X)$. Then, we obtain a linear model,

$$Y = lnX = \mu + \gamma' z + \sigma W \tag{1.15}$$

where $\gamma' = (\gamma_1, \cdots, \gamma_p)$ is a vector of regression coefficients and W may be viewed as an error term even though it does not necessarily have mean zero. Interpretations of the regression coefficients γ differ among the distributions of W. A variety of models can be used for W or, equivalently, for baseline lifetime distribution $T = exp(W)$. I will investigate three families of lifetimes, the *lognormal*, *log-logistic* and *Weibull*. These can be modeled by Eq.(1.13). Only the *Weibull* is also a *PH* model. A full likelihood incorporating right censoring can easily be constructed for the accelerated failure time model in Eq.(1.13). It is given by[6]

$$L(\gamma, \mu, \delta) = \Pi_{j=1}^{n} [f_Y(y_j|\gamma, \mu, \delta)]^{\delta_j} [S_Y(y_j|\gamma, , \mu, \delta)]^{1-\delta_j} \tag{1.16}$$

where $f_Y(y_j|\gamma, , \mu, \delta), S_Y(y_j|\gamma, , \mu, \delta)$ are respectively the density function and survival function of log-lifetime Y. And we assume that the distributions of the covariates $\{z_i\}$ do not contribute to the likelihood of (μ, σ, γ). Numerical methods must be used to maximize these complicated likelihoods.

Likelihood based inference for AFT models can be used to test Eq.(1.14), which like Eq.(1.4) addresses the question: Do the covariates do jointly effect the distribution of lifetime X? There are three likelihood based tests, likelihood ratio, score and Wald, similar to their partial counterparts described above.

Since sure knowledge as to whether a *PH* model, an *AFT* model or some other model holds is rarely available, robustness of these tests with respect to model misspecification is an important issue. To explore this issue, my report used simulation to assess and compare the performance of Cox's *PH* partial likelihood ratio test for joint covariate effects in terms

of size and power to the 'correct' AFT likelihood ratio test when data are generated by one of the three AFT models described above. Due to time limitations, I will not compare the performance of the AFT tests among themselves.

Chapter 2

Distributions

2.1 Weibull distribution

If we specify a standard extreme value distribution for W, $T = exp(W)$ has a standard exponential distribution[?] . Then, lifetime X then has a *Weibull* distribution, denoted $X \sim We(\alpha, \lambda)$, $\alpha > 0, \lambda > 0$, whose density, survivor function and hazard function have respectively, for $x > 0$, the forms are:

$$\begin{aligned} f(x|\alpha,\lambda) &= \lambda\alpha x^{\alpha-1} exp(-\lambda x^{\alpha}) \\ S(x|\alpha,\lambda) &= exp(-\lambda x^{\alpha}) \\ h(x|\alpha,\lambda) &= \lambda\alpha x^{\alpha-1} \end{aligned} \tag{2.1}$$

with

$$\alpha = 1/\sigma, \quad \lambda = exp(-(\mu + \gamma' z)/\sigma), \tag{2.2}$$

where γ are coefficients of covariate z given in Eq.(1.13), μ is location parameter on log scale given in Eq.(1.15).

This is also a *PH* model since, letting R.R denote relative risk, with

$$\beta' = -\frac{\gamma}{\sigma}$$
$$R \cdot R(z_1, z_2) = \frac{h(t|Z_1)}{h(t|Z_2)} = exp^{\beta(z_1 - z_2)} \tag{2.3}$$

is free of lifetime t. It is the only AFT model that also has a proportional hazards representation.

Below are plots of probability density functions of some *Weibull* distributions:

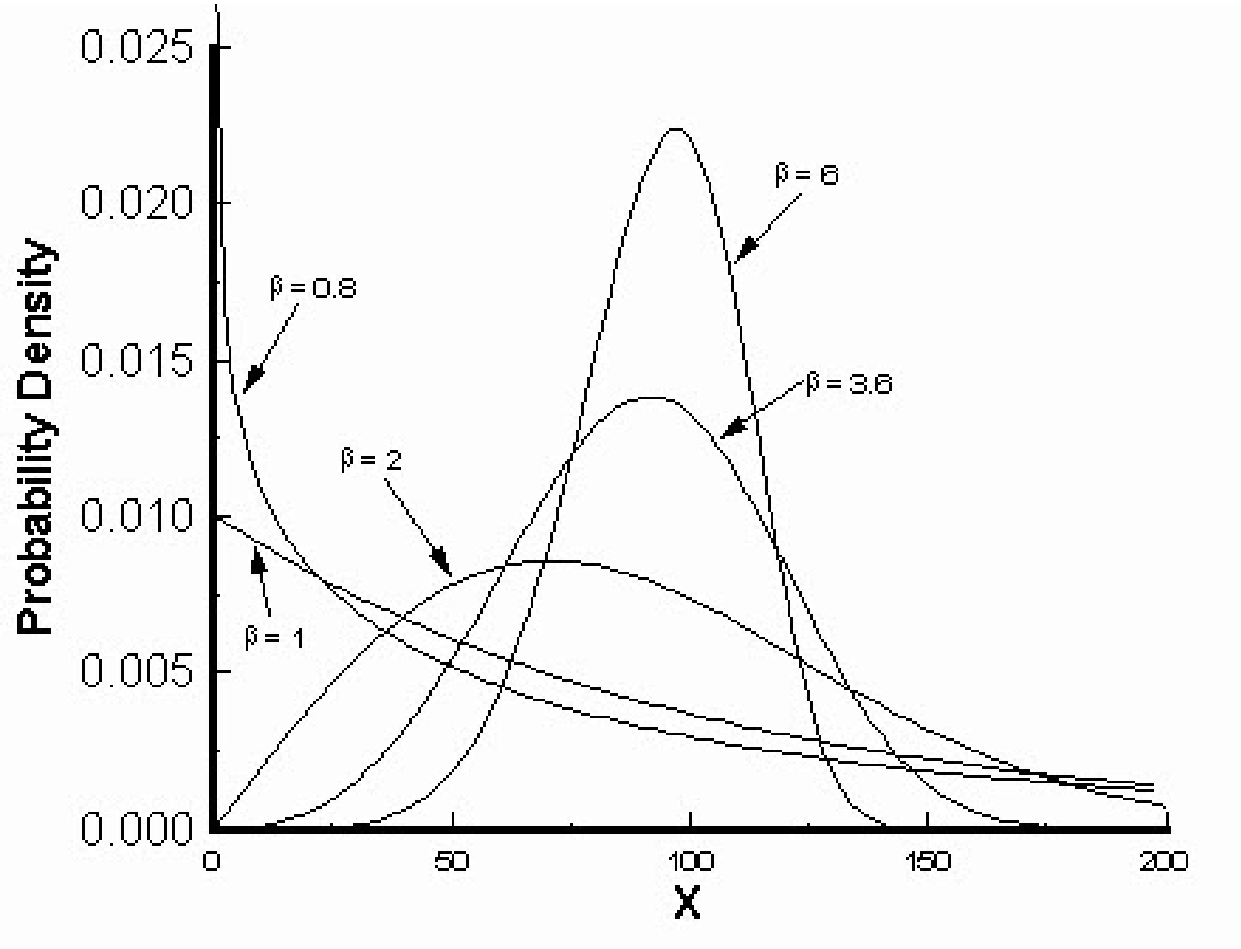


Figure 2.1: Probability density functions of *Weibull* distribution

2.2 Lognormal Distribution

The *lognormal* distribution is popular because of its relationship to the normal distribution. Specifically, if X is lognormal, $\ln(X)$ is normal. Further, the Lognormal hazard function has non-monotone behavior. It increases initially, then decreases and eventually approaches zero. This means that lifetimes with a Lognormal distribution have an increasing rate of failure as they age for some period of time[?] . But, after survival to a specific age, the rate of failure decreases as time increases.

If we specify that $T = exp(W)$ in Eq.(1.13) with $W \sim N(0,1)$, life time X has a Lognormal distribution,

$$X \sim lognormal(\mu + \gamma' z, \quad \sigma^2) \tag{2.4}$$

and hence,

$$Y \sim normal(\mu + \gamma' z, \quad \sigma^2) \tag{2.5}$$

Equivalently, the density of X is given by

$$f_X(x, \mu + \gamma' z, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} exp^{-\frac{(lnx-\mu-\gamma' z)^2}{2\sigma^2}} \tag{2.6}$$

The survivor function of X is then expressed as

$$S(x) = 1 - \Phi[\frac{ln(x) - \mu - \gamma' z}{\sigma}], \tag{2.7}$$

where Φ is the distribution function of a standard normal distribution. Hence,

$$h(x) = \frac{f_X(x)}{S(x)} = \frac{\frac{1}{x\sigma\sqrt{2\pi}} exp^{-\frac{(lnx-\mu-\gamma' z)^2}{2\sigma^2}}}{1 - \Phi[\frac{ln(x)-\mu-\gamma' z}{\sigma}]} \tag{2.8}$$

Then, the ratio of the hazard functions for two different covariates is given by

$$R \cdot R(z_1, z_2) = \frac{h(t|Z_1)}{h(t|Z_2)} = \frac{\frac{f(t|Z_1)}{S(t|Z_1)}}{\frac{f(t|Z_2)}{S(t|Z_2)}}$$
$$= \frac{\frac{\frac{1}{t\sigma\sqrt{2\pi}} exp^{-\frac{(lnt-\mu-\gamma' z_1)^2}{2\sigma^2}}}{1-\Phi[\frac{ln(t)-\mu-\gamma' z_1}{\sigma}]}}{\frac{\frac{1}{t\sigma\sqrt{2\pi}} exp^{-\frac{(lnt-\mu-\gamma' z_2)^2}{2\sigma^2}}}{1-\Phi[\frac{ln(t)-\mu-\gamma' z_2}{\sigma}]}} \tag{2.9}$$

which is not a free of t and hence not a PH model.

Below are plots of the probability density functions of some *Lognormal* distributions:

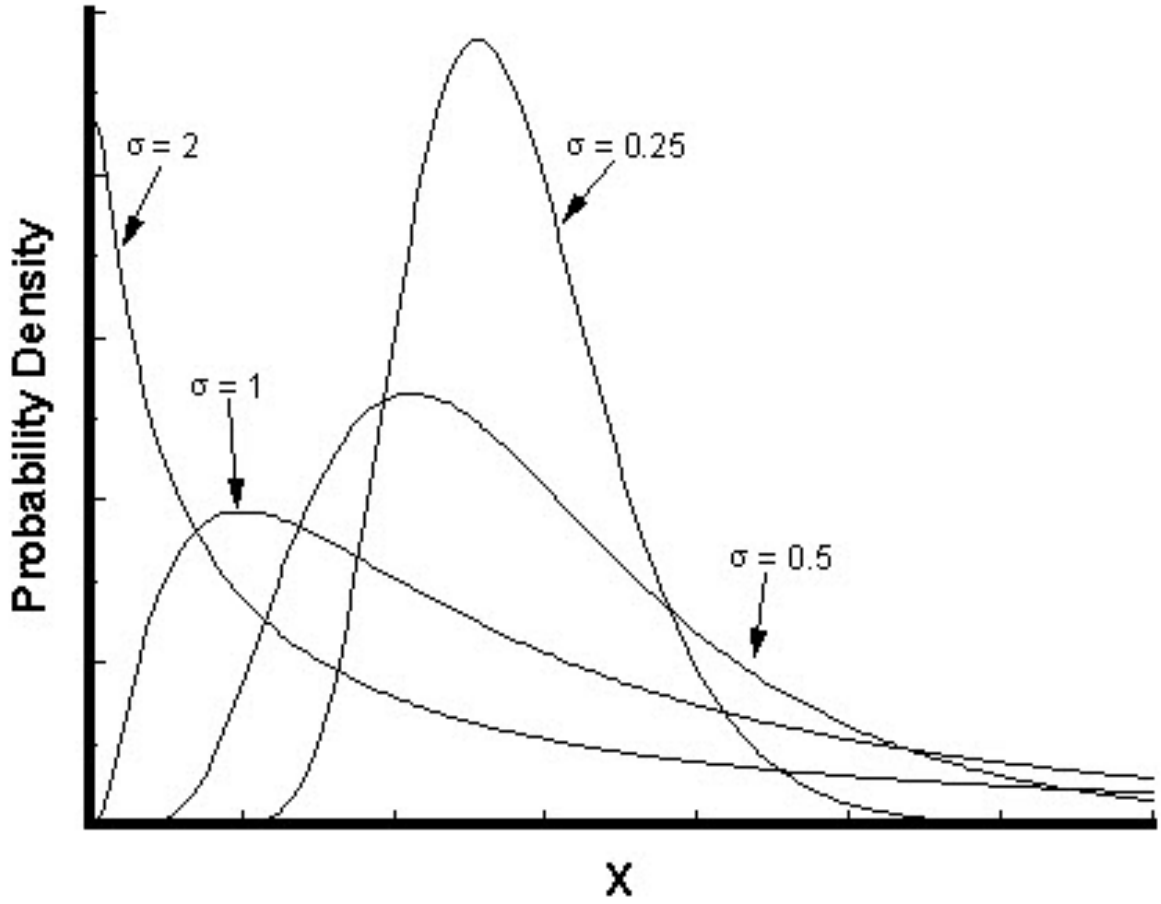


Figure 2.2: Probability density functions of *lognormal* distribution

2.3 Loglogistic Distribution

The log-logistic distribution has the following density function[?] :

$$f_X(x,\gamma) = \frac{\gamma x^{\gamma-1}}{(1+x^{\gamma})^2} \quad x>0, \gamma>0. \tag{2.10}$$

If we specify that W has a logistic distribution so that $T = exp(W)$ in Eq.(1.13) has a *Loglogistic* distribution with density denoted by

$$f_X(x|\mu,\gamma,\sigma) \sim loglogistic(\frac{1}{\sigma}, \exp((\mu+\gamma' z))), \tag{2.11}$$

which is

$$f_X(x|\mu,\gamma,\sigma) = \frac{1}{\sigma}(\frac{x}{\exp(\mu+\gamma' z)})^{\frac{1}{\sigma}-1}\exp(\mu+\gamma' z) \tag{2.12}$$

Then the survivor function and hazard function of lifetime X are given respectively by

$$\begin{aligned} S(x) &= [1+(\frac{x}{\exp(\mu+\gamma' z)})^{\frac{1}{\sigma}}]^{-1} \\ h(x) &= f(x)/S(x) = \frac{\frac{1}{\sigma}(\frac{x}{\exp(\mu+\gamma' z)})^{\frac{1}{\sigma}-1}\exp(\mu+\gamma' z)}{1+(\frac{x}{\exp(\mu+\gamma' z)})^{\frac{1}{\sigma}}} \end{aligned} \tag{2.13}$$

which is hump-shaped. The ratio of hazard functions for two different covariates is given by

$$\begin{aligned} R\cdot R(z_1,z_2) = \frac{h(t|Z_1)}{h(t|Z_2)} &= \frac{\frac{f(t|Z_1)}{S(t|Z_1)}}{\frac{f(t|Z_2)}{S(t|Z_2)}} \\ &= \frac{\frac{(\frac{t}{\exp(\mu+\gamma' z_1)})^{\frac{1}{\sigma}-1}}{1+(\frac{t}{\exp(\mu+\gamma' z_1)})^{\frac{1}{\sigma}-1}}}{\frac{(\frac{t}{\exp(\mu+\gamma' z_2)})^{\frac{1}{\sigma}-1}}{1+(\frac{t}{\exp(\mu+\gamma' z_2)})^{\frac{1}{\sigma}-1}}} \end{aligned} \tag{2.14}$$

which is not a free of t and hence not a PH model.

This model is the only AFT model that also has a representation as a proportional odds model. Specifically, for the loglogistic AFT model, the odds of survival beyond time t are given by

$$\frac{S(x|Z)}{1-S(x|Z)} = \exp(\beta' Z)\frac{S_0(x)}{1-S_0(x)} \tag{2.15}$$

where

$$\beta' = -\frac{\gamma}{\sigma} \tag{2.16}$$

Below are plots of the probability density functions of some *Loglogistic* distributions:

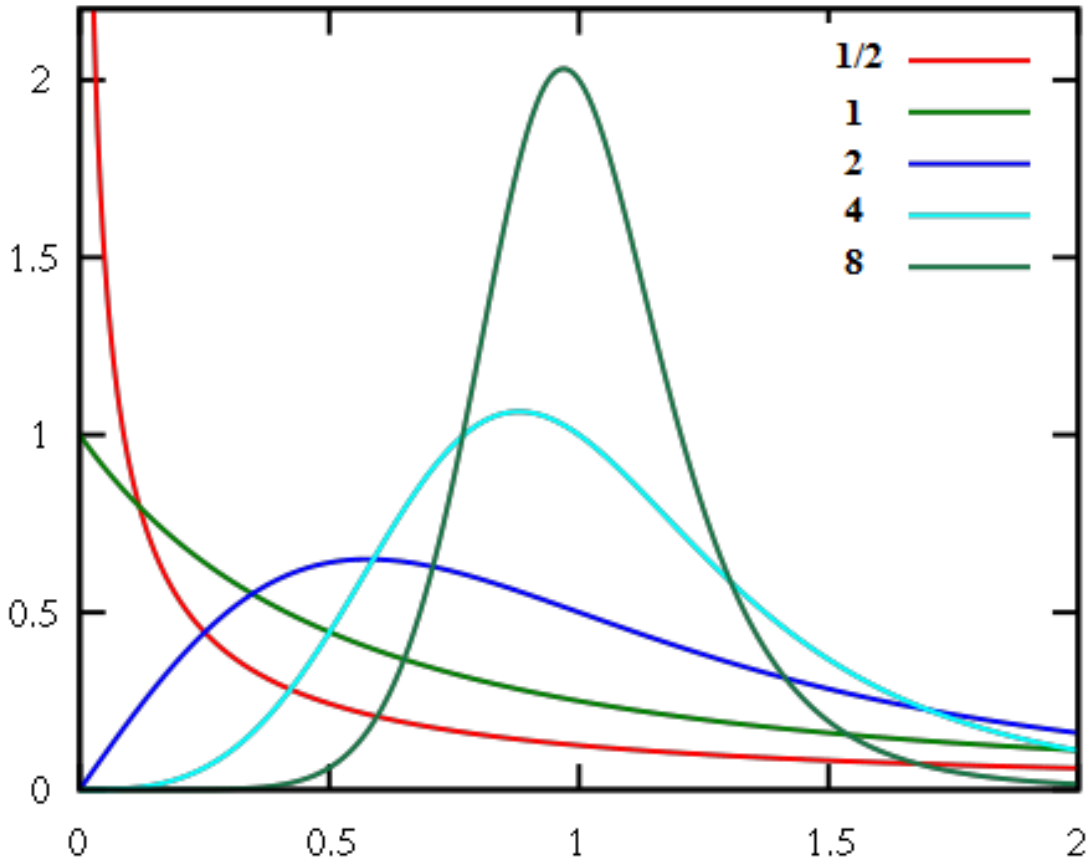


Figure 2.3: Probability density functions of *loglogistic* distribution

Chapter 3

Simulation

3.1 Introduction

Simulation is a numerical technique for conducting experiments on a computer. In statistics simulation experiments are most often used to study properties of statistical methods which cannot otherwise be easily evaluated. Monte Carlo simulations in statistics are computer experiments involving random sampling from probability distributions to study properties of statistical methods[?] . Here we will use software R to generate random numbers using Monte Carlo methods.

I generated data from the three families of parametric AFT models described above, the *Weibull, Lognormal* and *Loglogistic*. Two tests for no covariate effect were carried out on each data set: the likelihood ratio test for the AFT model used to generate the data and the partial likelihood ratio test for Cox's PH model. Both tests are 'correct' for the *Weibull* model, but only the parametric likelihood ratio tests for the other two. Comparisons of estimated type I error rates and power allow me to assess the robustness and performance of Cox's test relative to the *Lognormal* and *Loglogistic* likelihood ratio tests and compare Cox's semiparametric procedure to the fully parametric procedure in the case of the *Weibull* distribution.

In this study, typically a sample of n random numbers $X_1, \ldots, X_n$ were generated as observations from the three accelerated failure time models described above. Specifically,

my simulation study entails random sampling from the *Weibull, Lognormal,* and *Loglogistic* distributions.

3.2 Simulation Settings

My report designed and implemented a simulation study of size and power of the tests described above. The independent variables are: sample size, distribution of the simulated data, parameter values, censoring rate and number of covariates.

3.2.1 Distributions

For simplicity, I limited my study to the widely used Weibull, Lognormal and Loglogistic distributions. For the Lognormal distribution, we chose parameter values $\mu = 0, \sigma = 0.5$, and the resulting lognormal density is shown in Fig.3.1b. For the Weibull distribution, the shape parameter $\beta = 2.5$ was used. For the Loglogistic, the shape parameter $\beta = 0.25$ was used. The corresponding densities are shown in Fig.3.1a- 3.1c. Note that since I just studied the three specific distributions, the conclusions in this report need not apply to other members of these families or other distributions.

3.2.2 Right Censored Data and Sample Size

In this simulation, we only generated right-censored data, in which the study units are lost to follow-up (or the study ends) and might have experienced a recurrence of the event at some time in the future. But, the researcher wouldn't know if or when this happened. Generating censored data requires specifying both the lifetime and censoring distributions. One of the simulation methods discussed as the random censorship model[9] and[8]. is that, we assume that we have n independent, identically distributed lifetimes (that is, nonnegative random variables), $\{X_i\}$, with continuous distribution function F, and n independent, identically distributed censoring times, $\{C_i\}$, with continuous distribution function G. We also assume that $\{X_i\}$ and $\{C_i\}$ are independent for $i = 1, 2, \ldots n$. The distributions for F and G in

reference [3] were set as normal, exponential, weibull distributions.

Censoring index δ_i is determined by comparing $\{X_i\}$ and $\{C_i\}$, and in the reference article, it simulated distributions of $\{X_i\}$ and $\{C_i\}$ under assumptions that they follow the same and specific distribution. But there are various distributions for F and G, and they don't necessarily follow the same distribution. It is not necessary for us to study all kinds of lifetime and censoring time distributions in this paper. We will simplify things and represent censoring as follows.

Recall that lifetimes $\{T_i\}$ are assumed to be independent, a lifetime is censored only if the censoring index $\delta = 0$ and that the censoring rate is denoted by p. I assume the censoring index δ is independent from lifetime $\{T_i\}$. I set $\delta = 0$ for the first $m =$ censoring rate $*$ sample size observations. This procedure, although it does not exactly conform to the usual right censoring model, provides an easy way to approximate $100p\%$ non-informative right censoring.

What constitutes a reasonable sample size depends to some extent on the number of covariates and the censoring rate. I used one and three covariates, censoring rates p in the range of$(0.2 \sim 0.9)$ and sample sizes n in the range of $(10 \sim 50)$ to represent typical medical or engineering type problems. Medical experiments often deal with many covariates and high censoring rates.

3.2.3 Covariate vectors and coefficient

Also for simplicity, I generated covariates by sampling from a uniform distribution on the unit interval. For the one covariate model, its coefficient γ was set to range over the interval [0,5) by increments 0.1 to study power. So, there are 50 γs in total. Recall that

$$\text{lifetime} = exp(\mu + \gamma \times Z)T^{\sigma}. \tag{3.1}$$

In case of the three covariate lifetime model, I generated independent uniforms $\{U_i\}_{i=1}^3$ for each data set and let covariates $\{Z_i\}_{i=1}^3$ be the resulting order statistics, which are correlated, as often happens in the real world. The covariate coefficients γ's are random

numbers in the range of (0,5). But, I made sure $\gamma_1 = \gamma_2 = \gamma_3 = 0$ occurred so that I study the type I error rate. Recall that the three covariate lifetime model is then given by

$$lifetime = exp(\mu + \gamma_1 \times Z_1 + \gamma_2 \times Z_2 + \gamma_3 \times Z_3)T^{\sigma}. \tag{3.2}$$

In sum, I considered cases where the coefficients γ of the covariates equal zero to check type I error rates and some values of γ not equal to the zero vector to study power. I carried out some preliminary tests to find reasonable values for γ so that at least some of the tests have estimated powers close to one.

3.2.4 Monte Carlo Replicates N

Since the experimentation is done on a computer, we can easily replicate the experiments. The larger the number of Monte Carlo replicates, N, the better the approximation will be. However, computing time and computer or software (e.g. SAS, R) memory may be limited, making it necessary to run the N Monte Carlo replicates in smaller batches.

The empirical standard deviation for the estimated true level $\hat{\alpha}$ for testing no treatment effects is

$$sd(\hat{\alpha}) = \sqrt{\frac{(\hat{\alpha})(1-\hat{\alpha})}{N}} \tag{3.3}$$

If $\hat{\alpha} = 0.05$, and we want $sd(\hat{\alpha}) \leq 0.01$,

$$sd(\hat{\alpha}) = \sqrt{\frac{(0.05)(0.95)}{N}} \leq 0.01 \tag{3.4}$$

then we must choose $N \geq 475$. If $\hat{\alpha} = 0.05$ and $N = 1000$, $sd(\hat{\alpha}) = 0.0069$ and the corresponding 95% confidence interval for α is $0.05 \pm (1.96)(0.0069) = [0.0365, 0.0635]$.

In this experiment, we will use N=1000. I simulated combinations of 100 covariate coefficients, 4 sample sizes, 7 censoring rates and 3 distributions. For each combination, I generated 1000 datasets in order to average out the effects of randomness. Time is always an issue when it comes to large scale simulations. In my report I used the software package R to generate a $1000 * (100 * 4 * 7 * 3)$ data matrix and carried out PH and AFT analyses

on each data set, taking around 10 hours of CPU time. I used Java to do further analyses of the output from R in Chapter 4.

3.3 Goodness of fit test to Simulation data

There are several statistical software packages such as R and SAS that can be used for simulation in reference[1]. R is very convenient for programming and making graphs. It is free and widely used. I used functions "rweibull","rlnorm" and "rlogis" in R to generate data. But, the accuracy of its functions and macros is not validated by any one reliable source. We may therefore have some underlying risk in simulation and analysis when using R. In order to have some confidence in using R, I carried out χ^2 goodness of fit tests to test if the data it generated came from the specified distribution, denoted F. Recall that Fig. 3.1a- 3.1c plot the density functions I used in my simulation study. The three density distributions have similar shapes.

The Pearson's chi-square test is used to test the following hypothesis:

$$\begin{aligned} H_0 &: \text{The data are a random sample from distribution F} \\ H_a &: \text{The data are not a random sample from F} \end{aligned} \tag{3.5}$$

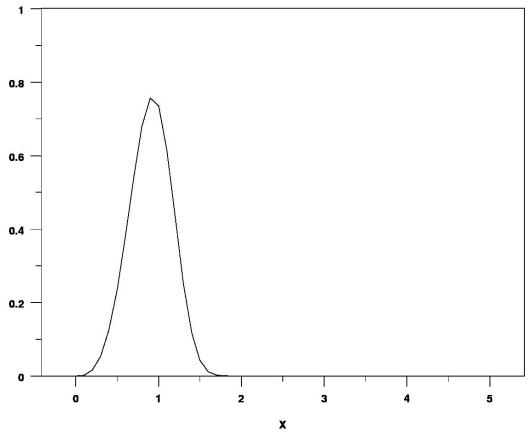
Pearson's chi-square test uses a measure of goodness of fit which is the sum of differences between observed and expected outcome frequencies (that is, counts of observations). The data generated by R were divided into k bins (defined below) and the test statistic defined as:

$$\chi^2 = \Sigma_{i=1}^{k}(O_i - E_i)^2/E_i, \tag{3.6}$$

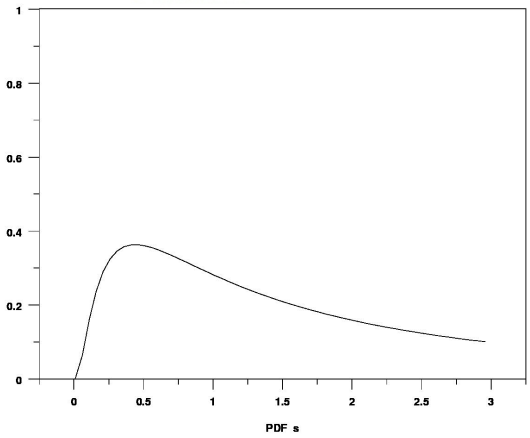
where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i under H_0. The expected frequency is calculated by

Figure 3.1: Densities of Studied Distribution

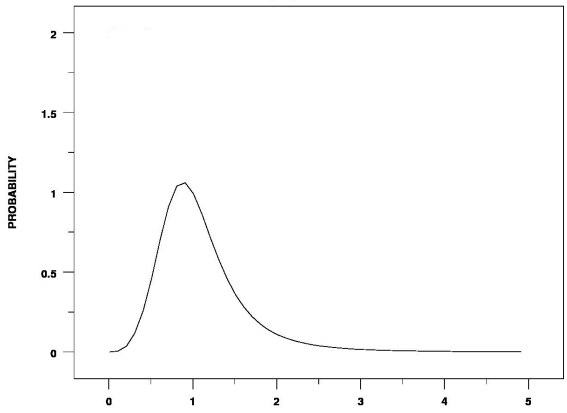
(a) *Weibull* density function



(b) *Lognormal* density function



(c) *Loglogistic* density function



$$E_i = n(F(Y_u) - F(Y_l)), \tag{3.7}$$

where Y_u is the upper limit for class i, Y_l is the lower limit for class i, and n is the sample size. This test is sensitive to the choice of bins. There is no optimal choice for the bin width since the optimal bin width depends on the distribution. Most reasonable choices should produce similar, but not identical results. I used $0.3 * s$, where s is the sample standard deviation, for the class width. The lower and upper bins are at the sample mean plus and minus $6.0 * s$, respectively, resulting in a target of 20 bins. In order to determine the degrees of freedom of the chi-squared distribution, one takes the total number of bins and subtracts one. For example, since there are 20 bins, I compared to Eq. (3.6) a chi-squared distribution with 19 degrees of freedom.

For large sample size, under H_0, the test statistic follows, approximately, a chi-square distribution with $(k-1)$ degrees of freedom where k is the number of non-empty bins Therefore, the hypothesis that the data are from a population with the specified distribution is rejected at nominal type I error rate α if

$$\chi^2 > \chi^2_{(\alpha,k-1)} \tag{3.8}$$

where $\chi^2_{(\alpha,k-1)}$ is the $100(1-\alpha)$ chi-square percentage point of a χ^2 distribution with $k-1$ degrees of freedom.

For example, I performed the chi-square test with $n = 50$ observations generated using R from the Weibull, Lognormal and Loglogistic distributions. The chi-square test was applied to each specific distribution respectively, as shown in Table 3.1. The test statistics are all small and we don't reject the null hypothesis at any reasonable type I error rates. So I concluded that R could be used to generate data for my simulation study.

Now, I use sample data generated from Weibull distribution to illustrate survival analysis carried out by R. There is only one covariate Z in this example, which is sampled from a uniform distribution, $Z \sim U(0,1)$. For sample size $n = 50$, and censoring rate $p = 0.5$, there

Table 3.1: Goodness of fit, 50 observations

	chi-square goodness of fit		
	weibull	**lognormal**	**log-logistic**
Test Statistic	12.2566	9.2372	10.36
P-Value	0.129	0.600	0.416

are $p \times n = 25$ right censored data points. The accelerated failure lifetime model is given by:

$$\text{lifetime} = \exp(5Z)T, \tag{3.9}$$

where T has a standard exponential distribution so that T has the *Weibull* distribution we specified above.

Fig.3.2 below is a plot of simulated lifetimes vs their corresponding covariate Z. The red circle points are censored lifetime and the blue plus points are true lifetimes. The lifetimes, covariates and censoring variables used to generate the plot are given in Table 3.2. For this sample data, both the PH and AFT based tests report small p-values, p-value = 0.00001. Hence, both tests yields the same conclusion, reject H_0 in Eq.1.4, which is the correct decision here since $\gamma = 5$ is not zero.

In sum, I carried out a simulation study using one or three predictors and different combinations of sample size and censoring rates. I generated 1000 data sets for each value of γ from 0 to 5 in increments of 0.1 from each of the three AFT models described above. Tests with nominal type I error rate 0.05 were carried out by rejecting the null hypothesis, that $\gamma = 0$ if the reported p-value was at most 0.05. The estimated power of the test for each value of γ was obtained by tallying the proportion of the 1000 data sets that led to rejection of H_0 . The PH partial likelihood ratio tests and the correct AFT likelihood ratio test were carried out for each model. The power plots and data sheets were presented in Chapter 4.

Figure 3.2: Simulated data from Weibull

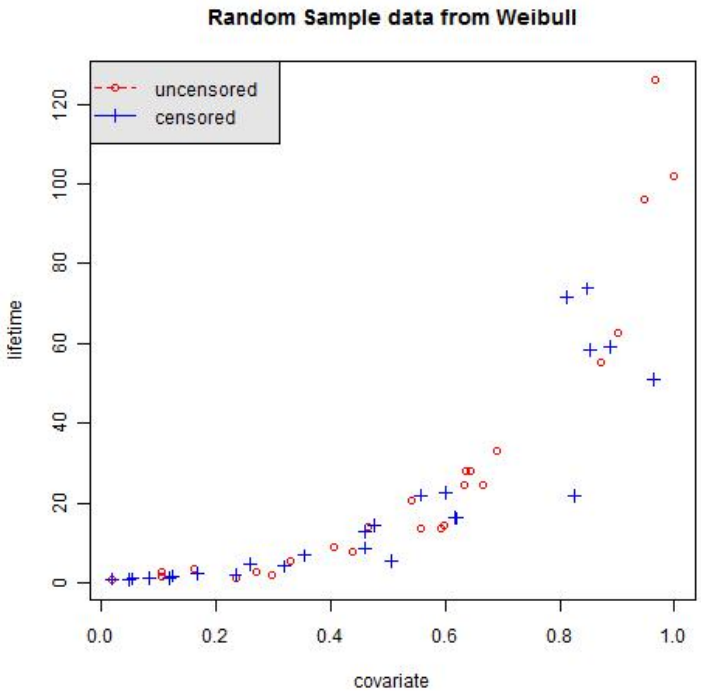


Table 3.2: Random Sample data from Weibull

Lifetime	Covariate	Censor
2.540639	0.445524	0
17.50793	0.644185	0
3.939988	0.22012	0
185.55	0.930428	0
4.252931	0.225385	0
2.095554	0.170033	0
99.41442	0.863456	0
70.00906	0.878021	0
93.47316	0.944982	0
23.1394	0.542224	0
0.602702	0.24987	0
6.165455	0.484835	0
98.16364	0.92288	0
2.152622	0.207742	0
4.78132	0.295521	0
50.02159	0.861418	0
12.1357	0.54584	0

Table 3.2: Random Sample data for Weibull

Lifetime	Covariate	Censor
46.78178	0.789511	0
3.190324	0.295456	0
143.2624	0.970317	0
0.684381	0.230309	0
15.40339	0.415117	0
2.044376	0.208141	0
1.743995	0.131936	0
4.884992	0.263578	0
1.050948	0.08123	1
22.05879	0.688035	1
50.31091	0.757446	1
16.50599	0.528918	1
19.6669	0.700643	1
8.548341	0.523447	1
27.72673	0.547905	1
30.39002	0.785065	1
7.97606	0.342266	1
50.23128	0.995804	1
34.87643	0.715028	1
40.63037	0.762608	1
1.260967	0.13311	1
12.92038	0.493604	1
1.065531	0.204003	1
3.955906	0.249272	1
52.48039	0.762551	1
16.96415	0.517036	1
0.694777	0.000712	1
6.133684	0.907436	1
73.22687	0.866193	1
18.81652	0.63778	1
14.922	0.600493	1
6.667701	0.334432	1
6.67046	0.441778	1

Chapter 4

Survival analysis methods Assessment and Application

In this chapter, we assess the Cox's PH and AFT model in terms of the convergence rate, type I error rate and power on the simulated data. And we will apply both PH and AFT methods in a real medical example and evaluate the analysis performance.

4.1 Non-convergence of Maximization Algorithm

The *AFT* analysis is based on maximizing the complicated likelihood given in Eq.(1.16). The *PH* analysis is based on solving the system of equations given in Eq.(1.5). It is possible that either or both of the algorithms used to carry out these processes fail to converge for a given data set. The boxplots in Fig.4.1 show the non-convergence rates (NR) for sample size less than 20, when sample size is larger than 20, NR decreases rapidly to less than $< 1\%$. The comparison for PH and AFT model is shown in Table **??**.

Table 4.1: Model comparison in non-convergence rate

Model	$n <= 20$	$n > 20$
PH	2.5%	$< 1\%$
AFT	5%	$< 1\%$

In Fig.4.2, we found out that AFT analysis only has higher NR's for small sample size(10) or censoring rate(0.8,0.9). PH analysis is stable and robust under extreme situation, for

example, small sample size(10) and high censoring rate(0.9), while AFT method does not hold. The comparison between models in censoring rate is shown in Table 4.2.

Table 4.2: Model comparison in non-convergence rate

Model	$p <= 0.8$	$p > 0.8$
PH	$< 2\%$	2%
AFT	$<= 1.5\%$	6%

Overall, We would recommend choosing between PH and AFT analyses according to sample sizes and censoring rates for one-covariate model. Observed NR's for both one or three predictor cases are presented in Tables A.1- A.3in the Appendix. It can be seen that, NR gets worse as censoring increases and is a bigger problem for PH than for AFT analysis. Going forward, data sets where an algorithm failed to converge for a method were deleted from the tally of results for that method.

Figure 4.1: Boxplots of NR: PH V.S. AFT, 1 covariate

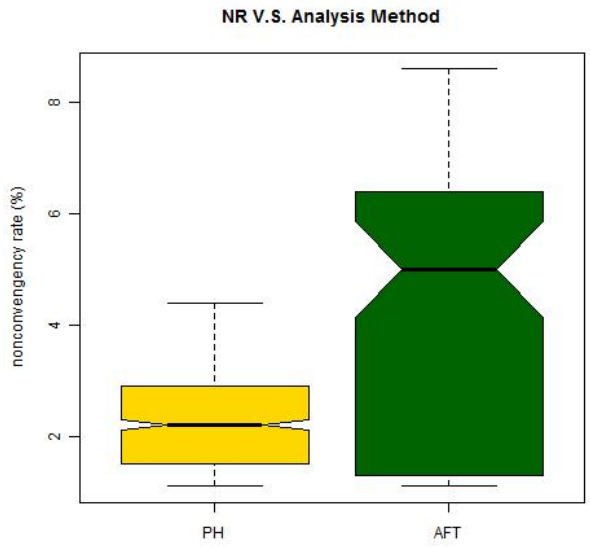
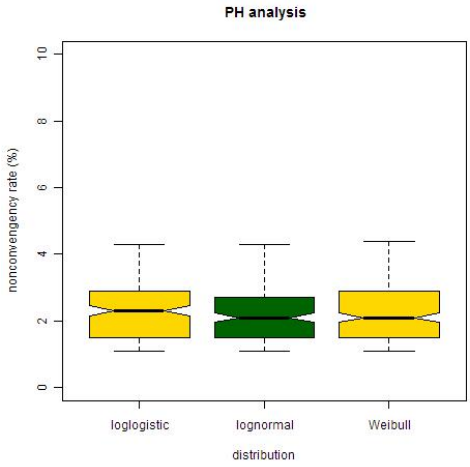
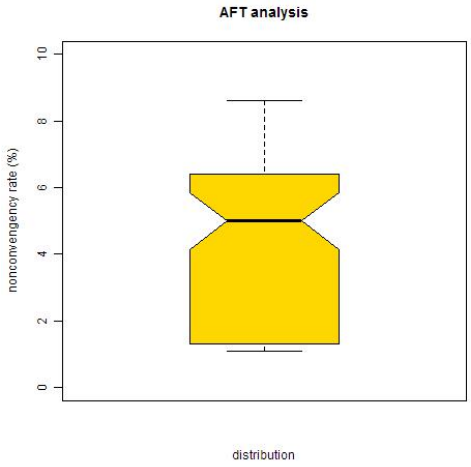


Figure 4.2: NR Comparisons for PH and AFT Analyses, 1 covariate

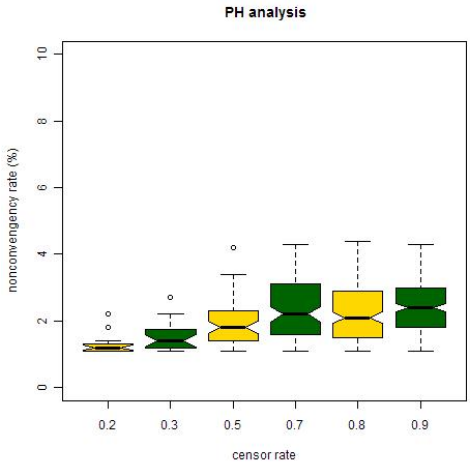
(a) NR V.S. Distribution



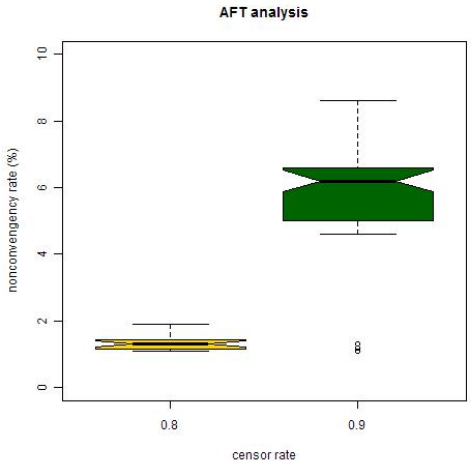
(b) NR V.S. Distribution



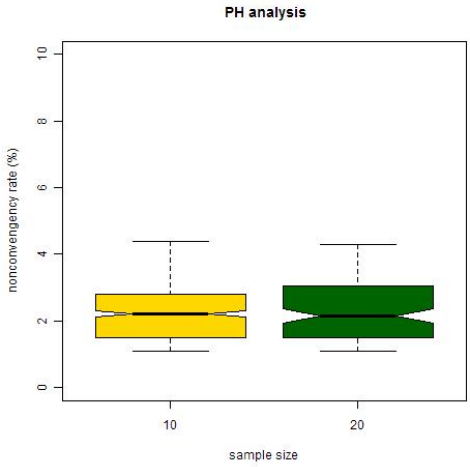
(c) NR V.S. Censor rate



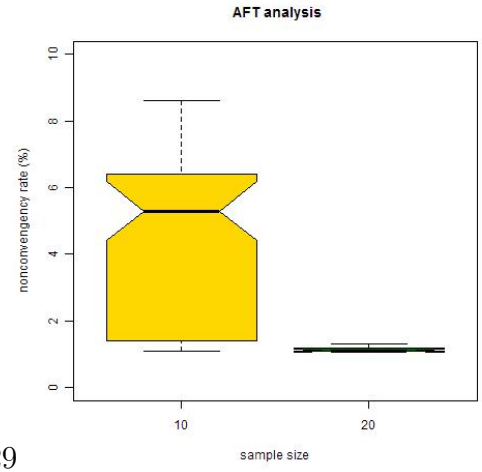
(d) NR V.S. Censor rate



(e) NR V.S. Sample Size



(f) NR V.S. Sample Size



4.2 Type I error rate study

Type I error, also known as false positive, occurs when a statistical test rejects a true null hypothesis. For example, in the one covariate study, the null hypothesis states that $\gamma = 0$. A type I error occurs if $\gamma = 0$ and the test rejects the hypothesis, falsely suggesting that $\gamma \neq 0$. I set the nominal type I error rate at 0.05 in my simulation study. But when sample size is small ($n \leq 20$), and the censoring rate is high ($p \geq 0.8$), it is expected that actual type I error rates may differ from 0.05. Recall that, I simulated data sets having high censoring rates and small sample sizes with one covariate and three covariates, which are common situations in real life. For example, in some medical studies of cancer, we are often interested in more than one factor, have small samples and many subjects may not complete the study due to a variety of reasons. Type I error rate performance is an important criterion for choosing between PH analysis or AFT analysis.

In my simulation for the one-covariate model, sample sizes were set as 10,20,30,50, and censoring rates are set as 0.2,0.3,0.5,0.7,0.8,0.9. The estimated type I error rates are given in Table 4.5. In Fig.4.3, the median of type I error rate of PH analysis is about 0.05, which is lower than the median of AFT analysis (0.08).

Fig.4.4 presents the comparisons of estimated Type I error rates (α) for the one-covariate model according to censoring rates, sample sizes, distributions adjusting for PH and AFT analyses. We could see a significant type I error rate at sample size(10) or censoring rate(0.9) for both of the analyses, up to 0.4, which are not acceptable. When sample size is over 20, type I error rates do not change very much. The comparisons are shown in Table 4.3-4.4.

Table 4.3: Model comparison in type I error rate

Model	$n <= 10$	$n >= 20$
PH	0.09	0.05
AFT	0.15	0.05

Furthermore, both PH and AFT analyses are not suitable for small sample sizes and

Figure 4.3: Boxplot of Type I Error Rates V.S. Analysis Methods, 1 covariate

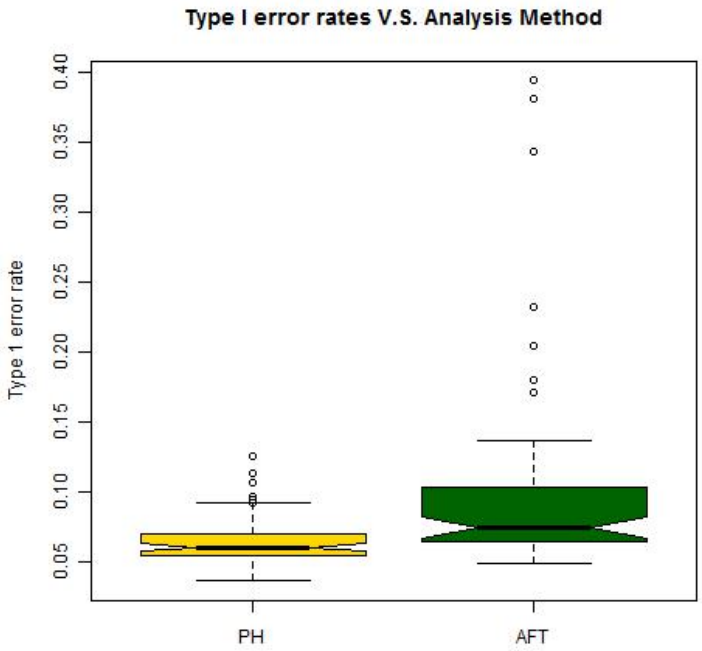


Table 4.4: Model comparison in type I error rate

Model	$p <= 0.8$	$p >= 0.9$
PH	0.06	0.12
AFT	0.05	0.4

high censoring rates(n=10,p=0.9), since they have large type I error rates, up to 0.2, as shown in Table 4.5.Overall, we don't recommend using PH and AFT analysis under this situation. A good performance occurs when an actual type I error rate is close to 0.05, this occurs, for example, when sample size is above 30, and censoring rate is not higher than 0.5.

For the 3 covariate model, estimated type I error rates are given in Table A.4 in the Appendix. As shown in the tables, the type I error rates for the 3 covariate model are much higher than for the 1 covariate model. For small sample sizes and high censoring rates, the estimated type I error rates are as high as 0.4, which means both PH or AFT analysis are

Table 4.5: Type I error rates for 1 covariate model

PH	AFT	censor rate	sample size	distribution
0.084	0.103	0.2	10	Weibull
0.074	0.102	0.2	10	lognormal
0.082	0.099	0.2	10	loglogistic
0.065	0.096	0.3	10	Weibull
0.065	0.102	0.3	10	lognormal
0.092	0.113	0.3	10	loglogistic
0.08	0.123	0.5	10	Weibull
0.062	0.112	0.5	10	lognormal
0.086	0.133	0.5	10	loglogistic
0.098	0.171	0.7	10	Weibull
0.092	0.136	0.7	10	lognormal
0.097	0.137	0.7	10	loglogistic
0.113	0.234	0.8	10	Weibull
0.099	0.192	0.8	10	lognormal
0.100	0.198	0.8	10	loglogistic
0.097	0.390	0.9	10	Weibull
0.101	0.358	0.9	10	lognormal
0.092	0.355	0.9	10	loglogistic
0.056	0.056	0.2	20	Weibull
0.077	0.077	0.2	20	lognormal
0.055	0.055	0.2	20	loglogistic
0.048	0.061	0.3	20	Weibull
0.066	0.074	0.3	20	lognormal
0.062	0.077	0.3	20	loglogistic
0.050	0.065	0.5	20	Weibull
0.058	0.068	0.5	20	lognormal
0.049	0.071	0.5	20	loglogistic
0.060	0.091	0.7	20	Weibull
0.060	0.081	0.7	20	lognormal
0.063	0.087	0.7	20	loglogistic
0.059	0.108	0.8	20	Weibull
0.063	0.104	0.8	20	lognormal
0.066	0.098	0.8	20	loglogistic
0.101	0.209	0.9	20	Weibull
0.086	0.127	0.9	20	lognormal
0.085	0.141	0.9	20	loglogistic

Table 4.5: Type I error rates for 1 covariate model

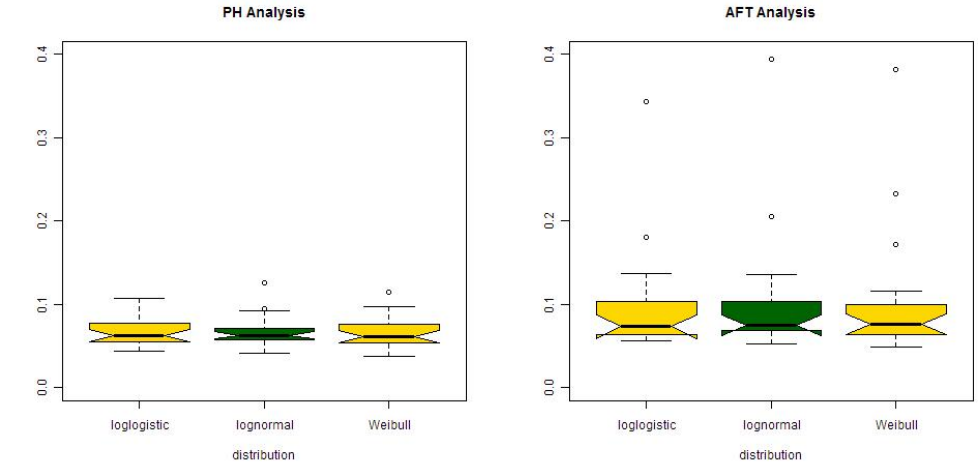
PH	AFT	censor rate	sample size	distribution
0.056	0.064	0.2	30	Weibull
0.062	0.071	0.2	30	lognormal
0.06	0.059	0.2	30	loglogistic
0.055	0.067	0.3	30	Weibull
0.066	0.073	0.3	30	lognormal
0.059	0.056	0.3	30	loglogistic
0.05	0.065	0.5	30	Weibull
0.057	0.066	0.5	30	lognormal
0.057	0.073	0.5	30	loglogistic
0.062	0.075	0.7	30	Weibull
0.054	0.071	0.7	30	lognormal
0.047	0.064	0.7	30	loglogistic
0.073	0.093	0.8	30	Weibull
0.055	0.065	0.8	30	lognormal
0.054	0.07	0.8	30	loglogistic
0.079	0.130	0.9	30	Weibull
0.065	0.078	0.9	30	lognormal
0.076	0.105	0.9	30	loglogistic
0.062	0.064	0.2	50	Weibull
0.059	0.057	0.2	50	lognormal
0.063	0.069	0.2	50	loglogistic
0.057	0.06	0.3	50	Weibull
0.049	0.063	0.3	50	lognormal
0.063	0.064	0.3	50	loglogistic
0.047	0.055	0.5	50	Weibull
0.041	0.052	0.5	50	lognormal
0.044	0.062	0.5	50	loglogistic
0.037	0.049	0.7	50	Weibull
0.050	0.057	0.7	50	lognormal
0.051	0.064	0.7	50	loglogistic
0.053	0.059	0.8	50	Weibull
0.057	0.078	0.8	50	lognormal
0.055	0.06	0.8	50	loglogistic
0.061	0.08	0.9	50	Weibull
0.063	0.075	0.9	50	lognormal
0.057	0.071	0.9	50	loglogistic

not suitable. We don't recommend using PH and AFT analysis when sample size is low, say 10, and censoring rate is at least 0.9.

Figure 4.4: Type I Error Rate Comparisons for PH and AFT Analyses, 1 covariate

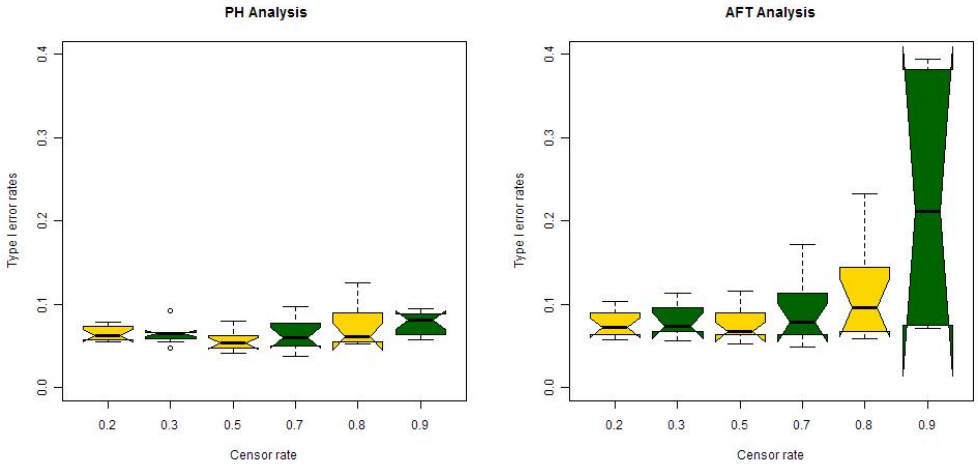
(a) Distribution

(b) Sample Size



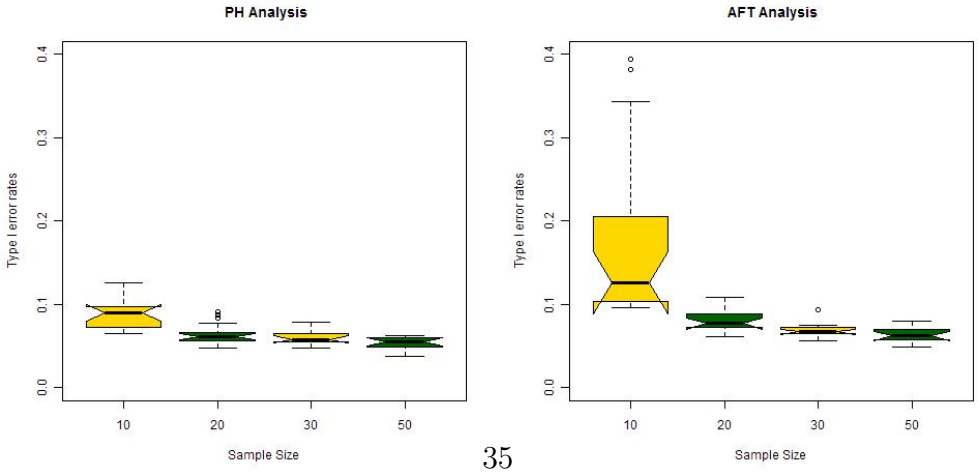
(c) Censor rate

(d) Censor rate



(e) Sample Size

(f) Sample Size



4.2.1 Estimated Type I Error Rates Test

When sample size is 10 and censoring rate is 0.9, the type I error rates are quite high which indicate that both PH and AFT analyses are not suitable in this case. In this section, I only considered the type I error rates for $n \geq 20, p \leq 0.8$. I carried out tests of hypotheses to check if actual type I error rates α differ from 0.05. Specifically, I tested:

$$H_0 : \alpha = 0.05, \quad \text{v.s.} \quad H_a : \alpha \neq 0.05 \tag{4.1}$$

Estimated type I error rates based on N datasets tend to approximately follow a standard normal distribution, $\hat{\alpha} \sim N(\alpha, s^2)$, $s = \sqrt{\alpha(1-\alpha)/N}$. Accordingly, I used the test statistic

$$\begin{aligned} Z_{obs} &= \frac{\hat{\alpha} - 0.05}{s} \\ s &= \sqrt{\frac{0.05(0.95)}{N}} = \sqrt{\frac{0.05(0.95)}{1000}} = 0.0069 \end{aligned} \tag{4.2}$$

We reject H_0 in Eq.(4.1) if $|Z_{obs}| \geq 1.96$, and conclude that the actual type I error rate is not equal to 0.05 at the 5% significance level. The 95% fail to reject H_0 range for α is (0.037,0.063). For the one covariate model, most of the estimated type I error rates of PH and AFT analyses fall into this range, as shown in Table 4.5. Although about 10% of those type I error rates are out of this range, they are almost all smaller than 0.11, we could conclude that actual type I error rates are mostly acceptably close to 0.05 for practical use.

For the 3 covariate model, the type I error rates are given in Table A.4. By carrying out the equivalence test of the estimated type I error rates to 0.05, most of the type I error rates are around 0.1, higher than for the 1 covariate model.

4.3 Power study

In this section, I study "power", which is the probability of rejecting H_0 when it is not true. We see from the definition that power is related to the type I error rate. Generally, lowering

the probability of type I errors, raises the probability of type II errors and lowers power. In some cases we compare the powers of tests that have different type I error rates. How to balance Type I error and power depends on the researcher's judgement and objective condition.

The estimated power plots in Fig.4.5, show the power trends versus the single covariate case as the coefficient γ varies from 0 to 5. For the 1 covariate model, the horizontal axis represents γ^2, and the vertical axis represents estimated power. Note that estimated powers increase as the horizontal scale γ^2 increases from 0-25. In the power plots for 3 covariate model in Fig.4.8, the horizontal axis is $\frac{\gamma_1^2+\gamma_2^2+\gamma_3^2}{3}$, and the vertical axis represents power. Note that estimated powers increases as the horizontal scale $\frac{\gamma_1^2+\gamma_2^2+\gamma_3^2}{3}$ increases from 0-25.

We superimposed the two power plots of PH and AFT analyses adjusting for sample size, censoring rate and distribution. Estimated PH powers are denoted by blue, blank circles, and AFT powers by red plus sign. In general, we see that for the one covariate case, *AFT* analysis has higher power than PH analysis. The power plots of PH and AFT analyses are almost identical as powers converge to 1, and the biggest differences between the power of two analyses occur before power converges to 1. The comparisons of power convergence rates are given in Table 4.6- 4.7. Fig.4.5,4.6,4.7 show the power plots for the one-covariate model with different sample sizes and censoring rates. The power plots for other sample sizes and censoring rates are given in Fig.B.1 in the Appendix.

Table 4.6: Model comparison in power convergence to 1

Model	$n = 20$	$n = 50$
PH	$\gamma^2 = 5$	$\gamma^2 = 2.5$
AFT	$\gamma^2 = 5$	$\gamma^2 = 2.5$

The power plots for the three covariate model are given in Fig.4.8, Fig.4.9 as well as in Fig.B.2, we see that there is a great difference in the power trends between the one covariate and three covariate models. Because the different ways to simulate data in Section 3.2.3, the power plots for three covariate model are not as smooth and continuous as one

Table 4.7: Model comparison in power convergence to 1

Model	$p = 0.2$	$p = 0.5$
PH	$\gamma^2 = 2.5$	$\gamma^2 = 5$
AFT	$\gamma^2 = 2.5$	$\gamma^2 = 5$

covariate model. We could see that there is a sharp increase of power reaching to 1 for three covariate model. AFT analysis has higher power and type I error rates than PH analysis and it converges slightly less rapidly to 1 than PH analysis. As sample sizes increase or censoring rates decrease, the two power trends converge. The power plots don't differ much across the distributions either. I discuss how the sample sizes, censoring rates, distributions and covariate numbers affect power difference between PH and AFT analyses later in this chapter.

Figure 4.5: Power Plots for 1 Covariate Model,n=20,p=0.2

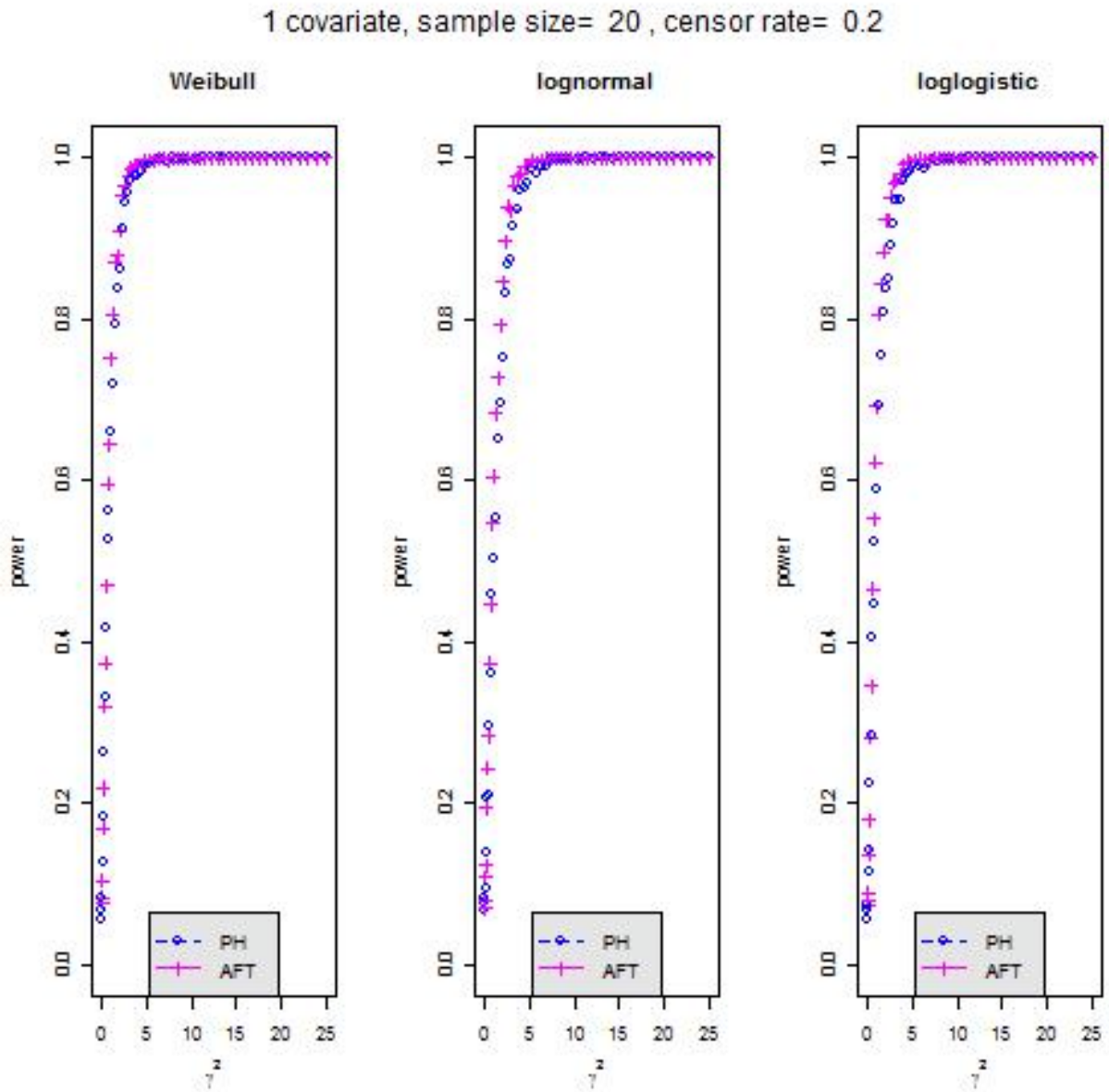


Figure 4.6: Power Plots for 1 Covariate Model,n=20,p=0.5

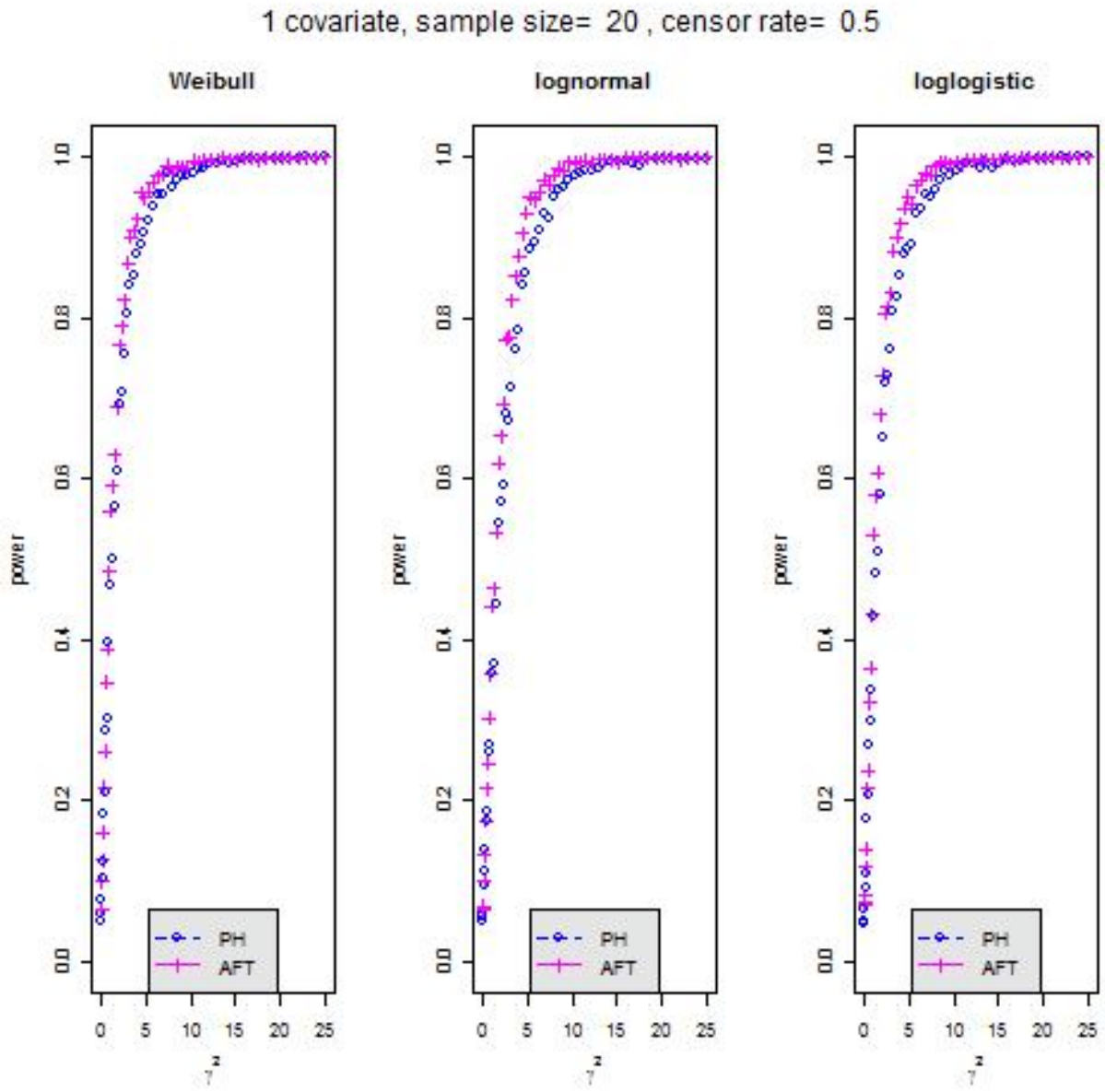


Figure 4.7: Power Plots for 1 Covariate Model,n=50,p=0.5

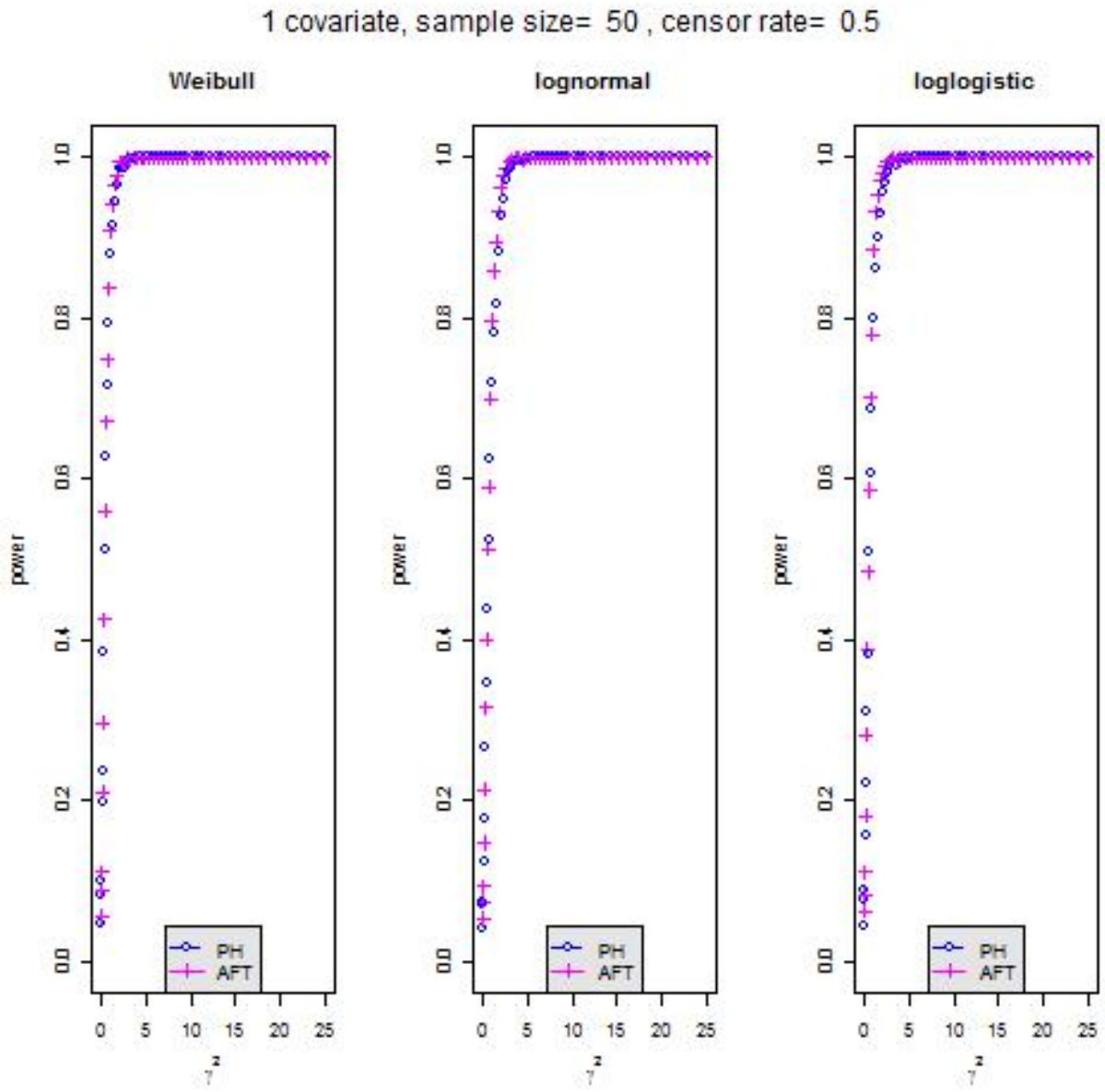


Figure 4.8: Power Plots for 3 Covariate Model,n=20,p=0.5

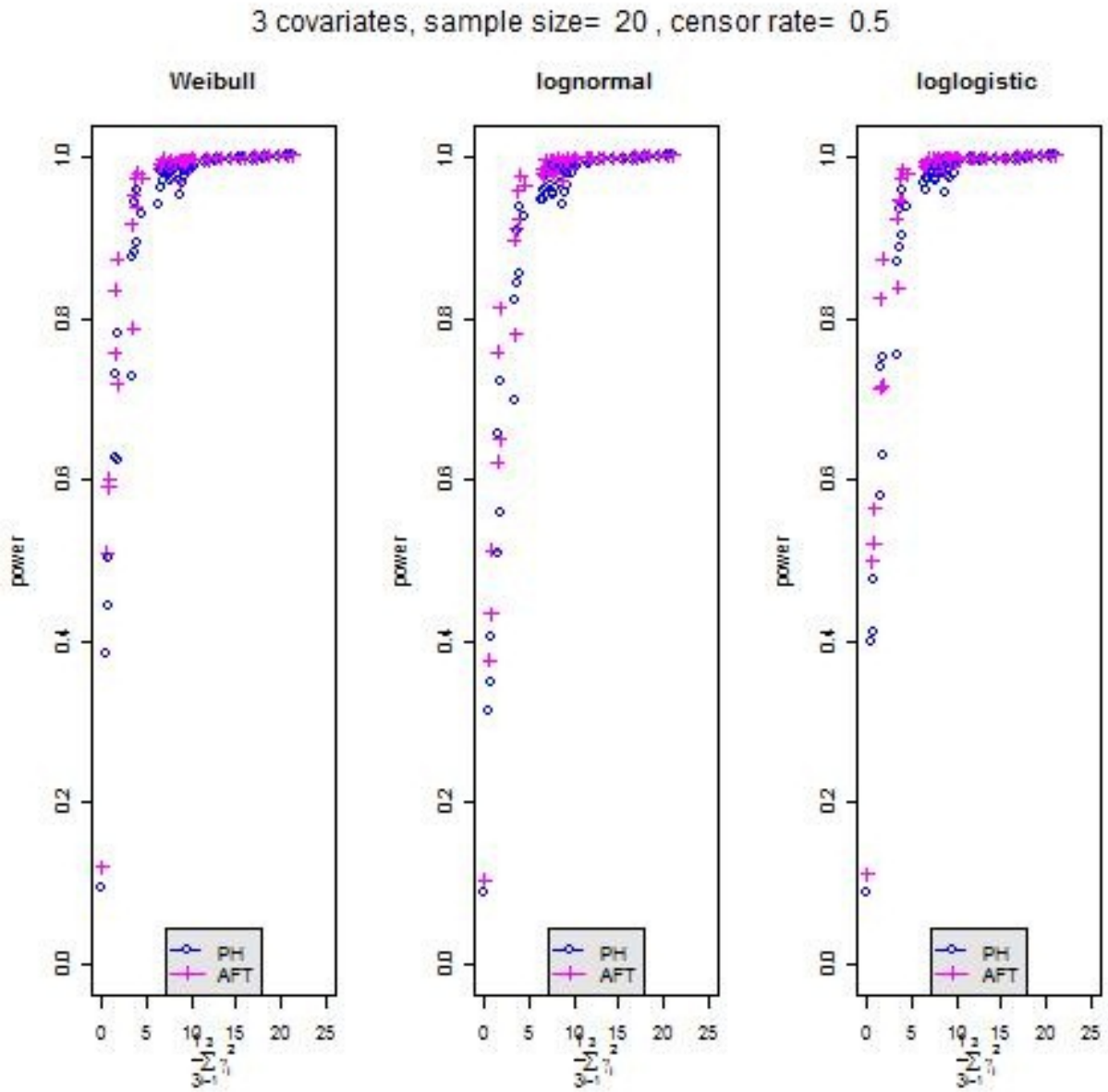
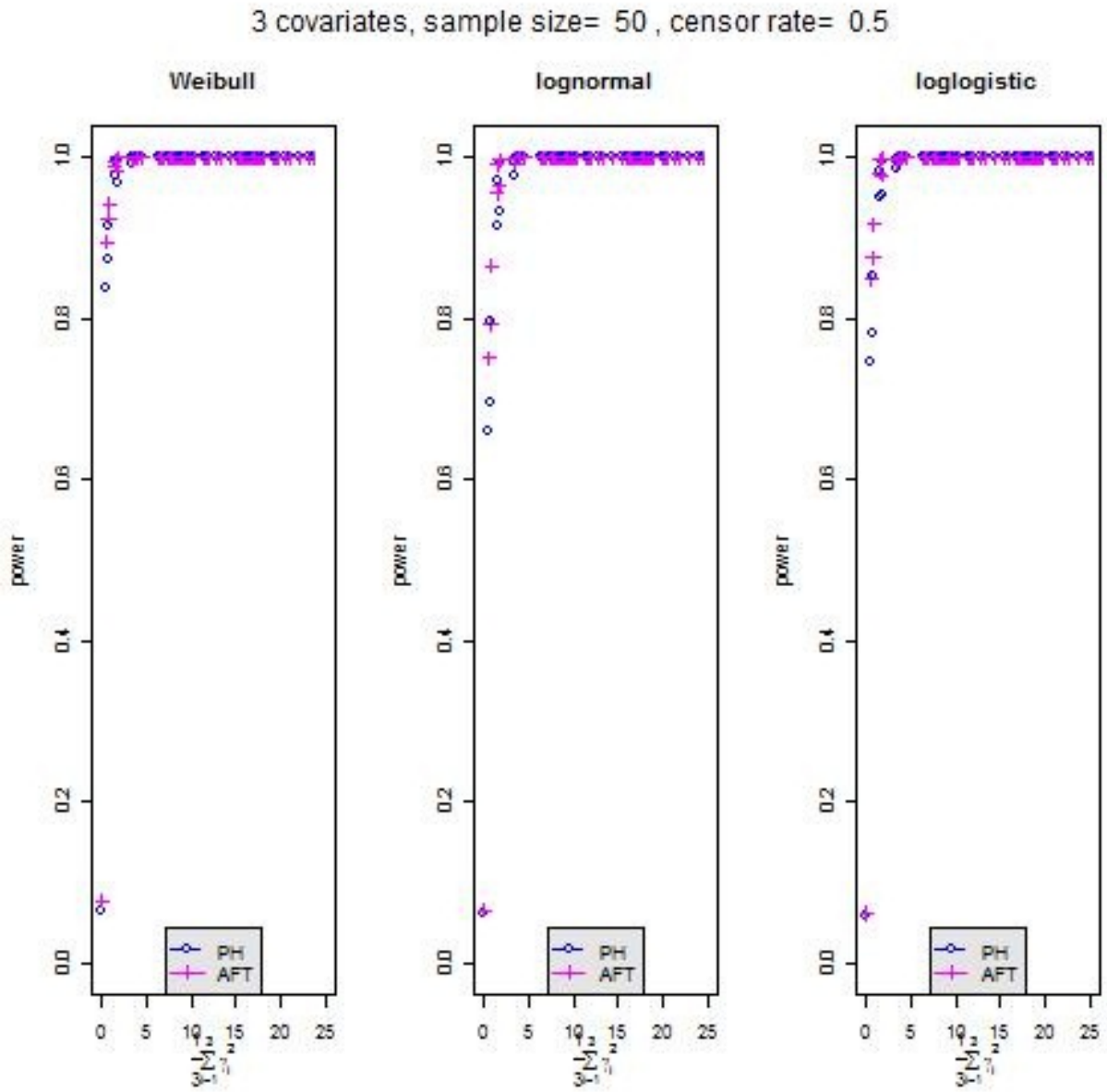


Figure 4.9: Power Plots for 3 Covariate Model,n=50,p=0.5



4.4 McNemar's Test for PH and AFT Analysis

Referring to Section 4.3, when γ is small from (0-1.5), estimated PH and AFT power trend lines appear to be different. With large sample size and small censoring rate, for example, n=50, p=0.2, the two power trend lines appear to be identical in this case. In order to test if there is difference between the two analyses, I applied two criteria: type I error rate and maximum power difference. McNemar's test is used to evaluate these criteria in the following sections.

McNemar's test is used to test if there is difference in population proportions based upon experiments where both responses are recorded on each experimental unit. It is named after Quinn McNemar, who introduced it in 1947[?] . McNemar's test is most easily carried out by summarizing the data in the "four fold" table given in Table 4.8, where n_{ij} is the observed count in row i, column j and p_{ij} is the corresponding population proportion under the circumstance of a specific γ, sample size, censoring rate of one or three covariate model. The estimates of $\{p_{ij}\}$ are given by

$$\begin{aligned}
\hat{p}_{11} &= \frac{n_{11}}{n_{11}+n_{12}}, \\
\hat{p}_{12} &= \frac{n_{12}}{n_{11}+n_{12}}, \\
\hat{p}_{21} &= \frac{n_{21}}{n_{21}+n_{22}}, \\
\hat{p}_{22} &= \frac{n_{22}}{n_{21}+n_{22}}.
\end{aligned} \tag{4.3}$$

The null hypothesis of marginal homogeneity states that the marginal probabilities, i.e. $p_{11}+p_{12} = p_{11}+p_{21}$ and $p_{21}+p_{22} = p_{12}+p_{22}$ are equal. Thus, the hypothesis of McNemar's Test is given by

$$H_0 : p_{12} = p_{21} \quad \text{v.s.} \quad H_a : p_{12} \neq p_{21} \tag{4.4}$$

McNemar's test statistic with Yates' correction for continuity is given by:

Table 4.8: Four Fold Table

	Test 2 rejects	**Test 2 not reject**
Test 1 rejects	$n_{11}(p_{11})$	$n_{12}(p_{12})$
Test 1 not reject	$n_{21}(p_{21})$	$n_{22}(p_{22})$

$$\chi^2 = \frac{(\mid n_{12} - n_{21} \mid -0.5)^2}{n_{12} + n_{21}}. \tag{4.5}$$

Under the null hypothesis, with a sufficiently large number of discordant (counts n_{12} and n_{21}), χ^2 has a approximately chi-squared distribution with 1 degree of freedom. If either n_{12} or n_{21} is small ($n_{12}+n_{21} < 25$) then χ^2 may not be well-approximated by the chi-square distribution. The binomial distribution can be used to obtain the "exact" distribution for obtaining p-values. In this formulation, under H_0, n_{12} has conditional on $n_{12}+n_{21}$, a binomial distribution with size parameter equal to $n_{12}+n_{21}$ and "probability of success" $= 0.5$, and is essentially the sign test. For $n_{12} + n_{21} < 25$, the binomial calculations should be performed and result in an what's called an exact test. If the statistic provides sufficient evidence to reject the null hypothesis, in favor of the alternative hypothesis that $H_a : p_{12} \neq p_{21}$, conclude in the setting discussed here that the power functions for the PH and AFT model analysis are not identical. This procedure is carried out for $\gamma = 0$ and $\gamma \neq 0$ in the following sections.

4.4.1 Type I Error Rate Criterion

Type I error rate is the first key issue in comparisons of PH and AFT analyses. I used the type I error rate difference(ERD) to measure the difference of type I error rate between two analyses, denoted by

$$ERD = \mid \alpha_{AFT} - \alpha_{PH} \mid \tag{4.6}$$

where α_{AFT} and α_{PH} are type I error rates for AFT and PH analyses at $\gamma = 0$. Let $\hat{\alpha}_{AFT}$and $\hat{\alpha}_{PH}$ be estimated type I error rates of the AFT and PH analysis respectively. Then ERD can be estimated by

$$\widehat{ERD} = \mid \hat{\alpha}_{AFT} - \hat{\alpha}_{PH} \mid \tag{4.7}$$

Since both methods were applied to the same simulated data set, I used McNemar's test to test the difference between the two analyses in type I error rate. Thus, the hypothesis of McNemar's test is given by

$$H_0 : ERD = 0, \quad \text{v.s.} \quad H_a : ERD \neq 0. \tag{4.8}$$

For example, the type I error rates given in Table 4.5 is 0.062 for PH analysis, and 0.064 for AFT analysis. McNemar's test for the equality of correlated type I error rates yields:

$$\chi^2 = \frac{(\mid 8 - 6 \mid -0.5)^2}{8 + 6} = 0.285 \tag{4.9}$$

with $df = 1, p - value = 0.593$ and we fail to reject $H_0 : ERD = 0$ and we don't have evidence to conclude that there is a statistically significant difference between the type I error rates of the PH and AFT analysis at the 5% significance level when sample size is 50 and censoring rate is 0.2 for the 1 covariate model.

In order to test whether PH and AFT analyses are different in terms of type I error rates, I carried out McNemar's tests in all cases and found out that PH and AFT don't differ much in type I error rates in most cases for Lognormal and Loglogistic distributions. Some of the tables are given in Tables A.5. But, for the Weibull distribution, there is a significant difference between the two analyses. For example, when sample size is 30 and censoring rate 0.5 for 1 covariate model, the four fold tables and McNemar's tests for the three distributions are shown in Table 4.9a- 4.9c. With $p - value = 0.016$ for Weibull, we reject $H_0 : ERD = 0$ and conclude that there is a significant different in type I error rates for Weibull distribution when sample size is 30 and censoring rate is 0.5 for the 1 covariate model. With $p - value = 0.069, 0.3$ for Lognormal and Loglogistic distributions, we fail to reject $H_0 : ERD = 0$ and we don't have evidence to conclude that there is a statistically significant difference between the type I error rates of the PH and AFT analysis at the 5%

significance level for Lognormal and Loglogistic distributions when sample size is 30 and censoring rate is 0.5 for the 1 covariate model. I performed McNemar's tests in all cases and found out that only for sample size 50 and censoring rate less than or equal to 0.5, there are no significant differences in type I error rates among methods. When sample size is small < 50, PH analysis has smaller type I error rate than AFT analysis. The four fold tables are given in Tables A.5 in the Appendix.

Table 4.9: Four Fold Table for Type I error rate, 1 covariate model

(a) n=30, p=0.5, Weibull

ERD	**PH**		
		reject	not reject
AFT	**reject**	56	18
	not reject	6	918
χ^2	5.51		
p-value	0.016		

(b) n=30, p=0.5, Lognormal

ERD	**PH**		
		reject	**not reject**
AFT	**reject**	61	26
	not reject	14	899
χ^2	3.31		
p-value	0.069		

(c) n=30, p=0.5, Loglogistic

ERD	PH		
		reject	not reject
AFT	reject	52	12
	not reject	7	929
χ^2	1.066		
p-value	0.3		

(d) n=50, p=0.2, Weibull

ERD	PH		
		reject	not reject
AFT	reject	38	8
	not reject	6	948
χ^2	0.285		
p-value	0.593		

(e) n=50, p=0.2, Lognormal

ERD	PH		
		reject	not reject
AFT	reject	32	27
	not reject	24	917
χ^2	0.176		
p-value	0.67		

(f) n=50, p=0.2, Loglogistic

ERD	PH		
		reject	not reject
AFT	reject	33	21
	not reject	13	933
χ^2	1.88		
p-value	0.17		

4.4.2 Maximum Power Difference Criterion

In order to test whether PH and AFT analyses are different in terms of power, I tested statistical signficiance of the difference between the two estimated power plots. Whether the power differences are statistically significant or not is a key question for helping us to choose between the two analyses. I used the maximum power difference(MPD) to measure the difference of power between two analyses, defined by

$$MPD = \max\{\mid K_{AFT}^{(i)} - K_{PH}^{(i)} \mid \quad , i = 1, 2, \cdots, m\} \tag{4.10}$$

where $K_{AFT}^{(i)}$ and $K_{PH}^{(i)}$ are powers for AFT and PH analyses at alternative $\gamma_i \neq 0$. Let $\hat{K}_{AFT}^{(i)}$and $\hat{K}_{PH}^{(i)}$ be estimated powers of the AFT and PH analysis respectively. Then MPD can be estimated by

$$\widehat{MPD} = \max\{\mid \hat{K}_{AFT}^{(i)} - \hat{K}_{PH}^{(i)} \mid \quad , i = 1, 2, \cdots, m\}, \tag{4.11}$$

Since both methods were applied to each simulated data set, I used McNemar's test to test

$$H_0 : \text{MPD} = 0, \quad \text{v.s.} \quad H_a : \text{MPD} \neq 0. \tag{4.12}$$

In my case, the units are data sets and the responses are rejection of the hypothesis of the covariates being zero. I could use McNemar's test at each alternative value of the covariate(s). Instead, I carried out McNema's test at the alternative where the estimated power curves are furthest apart, which is an adaptive way to test $H_0 : MPD = 0$ in Eq. (4.12). A small p-value could then be interpreted as evidence that the power curves are not identical. Since for Weibull distribution, both tests are appropriate as discussed in Section 2.1, it is not surprising that the two power trend plots appear be similar. For example, the smallest MPD given in Table 4.11 is 0.047 with sample size 50, censoring rate 0.2 and Weibull distribution. The four fold table in Table 4.10 of rejections of H_0 is shown below.

McNemar's test for the equality of correlated power rates yields:

Table 4.10: Four fold table for MPD with n=50, p=0.2, Weibull 1 covariate Model

MPD		PH	
		reject	not reject
AFT	reject	702	55
	not reject	16	227

$$\chi^2 = \frac{(\mid 55 - 16 \mid -0.5)^2}{55 + 16} = 21.422 \tag{4.13}$$

with $df = 1, p - value = 4.115E - 5$ and we reject $H_0 : MPD = 0$ and conclude that there is a statistically significant difference between the powers of the PH and AFT analysis at the 5% significance level when sample size is 50 and censoring rate is 0.2 for the 1 covariate model. We concluded that the power of AFT analysis differs from PH analysis in this case. But, as we look at the power plots in Fig.4.10, the two power trend lines for PH and AFT analysis are almost the same after the powers reach 1, while they are quite different when γs are close to 0. Statistical significance is probably obtained here because of the large number (1000) of data sets.

Since when $n \leq 10, \quad p \geq 0.9$, it is not suitable to use PH and AFT analyses because of high NR and type I error rates. So I just did McNemar's tests to cases $n \geq 20$ or $p \leq 0.8$, and found out that AFT analysis has different power in testing for covariates than PH analysis. The four fold tables are given in Tables A.6. So far, by carrying McNemar's test for the equality of correlated powers and type I error rates, we conclude that there are differences in powers between PH and AFT analysis, while no difference in type I error rates for one covariate effect with sample size 50 and censoring rate 0.2 at 5% significance level. We also learned that when sample sizes are 20 or above and censoring rates are less than 0.9, the non-convergence rate is low and type I error rates are around 0.05. By carrying out McNemar's test for these cases, $\widehat{MPD} > 0$ and there was an indication of a superior performance for AFT analysis in those cases where it's estimated type I error rate is close to nominal. But whether using PH or AFT depends mainly on a researcher's judgement.

Figure 4.10: Power Plots for 1 Covariate Model,n=50,p=0.2

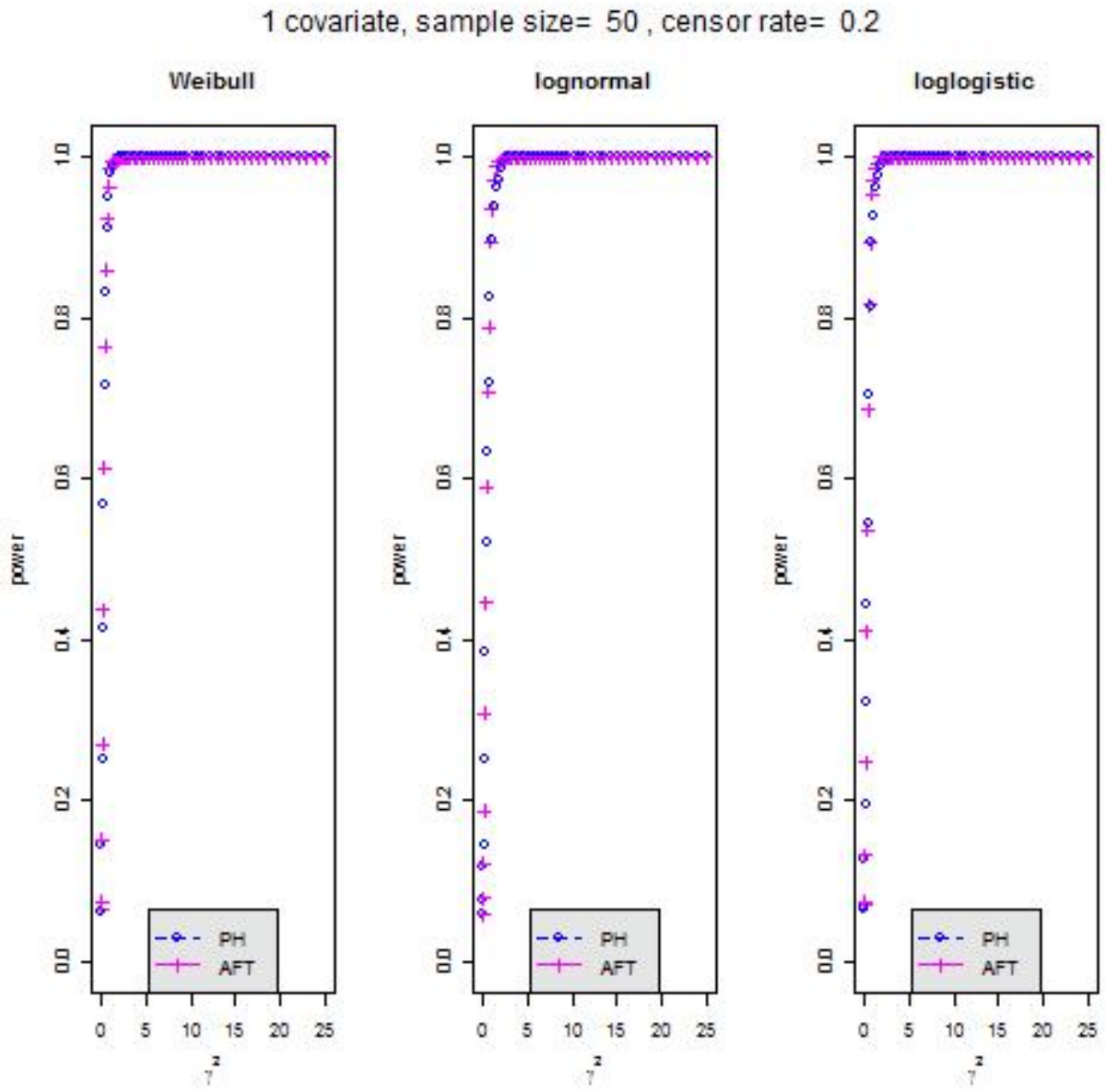


Table 4.11: Maximum Power Difference for 1 Covariate Model

$\max\{\lvert\widehat{PH} - \widehat{AFT}\rvert\}$	censor	size	distribution
0.091	0.2	20	Weibull
0.131	0.2	20	lognormal
0.113	0.2	20	loglogistic
0.060	0.2	30	Weibull
0.101	0.2	30	lognormal
0.130	0.2	30	loglogistic
0.097	0.3	20	Weibull
0.108	0.3	20	lognormal
0.107	0.3	20	loglogistic
0.094	0.5	20	Weibull
0.110	0.5	20	lognormal
0.101	0.5	20	loglogistic
0.108	0.7	20	Weibull
0.117	0.7	20	lognormal
0.128	0.7	20	loglogistic
0.131	0.8	20	Weibull
0.128	0.8	20	lognormal
0.112	0.8	20	loglogistic
0.069	0.3	30	Weibull
0.097	0.3	30	lognormal
0.101	0.3	30	loglogistic
0.071	0.5	30	Weibull
0.095	0.5	30	lognormal
0.110	0.5	30	loglogistic

Table 4.11: Maximum Power Difference for 1 Covariate Model

$\max\{\lvert\widehat{PH} - \widehat{AFT}\rvert\}$	censor	size	distribution
0.086	0.7	30	Weibull
0.106	0.7	30	lognormal
0.095	0.7	30	loglogistic
0.090	0.8	30	Weibull
0.111	0.8	30	lognormal
0.093	0.8	30	loglogistic
0.059	0.5	50	Weibull
0.079	0.5	50	lognormal
0.102	0.5	50	loglogistic
0.052	0.7	50	Weibull
0.098	0.7	50	lognormal
0.087	0.7	50	loglogistic
0.078	0.8	50	Weibull
0.091	0.8	50	lognormal
0.079	0.8	50	loglogistic
0.047	0.2	50	Weibull
0.075	0.2	50	lognormal
0.142	0.2	50	loglogistic

4.4.3 Modeling MPD

As we learned from the last two sections, there is a statistically significant difference between the powers of the PH and AFT analyses, and the MPD changes as sample sizes and censoring rates change. Here, I studied how MPD is effected by sample size, censoring rate, distribution, method of analysis and number of covariates.

As noted above, PH and AFT models are equivalent for the Weibull distribution. We therefore used the Weibull distribution as a reference category, and compared power performance of Lognormal v.s. Weibull, Loglogistic V.S. Weibull. We built a first order linear regression model for MPD, considering the independent variables sample size, censoring rate, covariate number, distribution and method of analysis. We used least squares to fit the linear regression model

$$\text{MPD} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 W_1 + \beta_4 W_2 + \epsilon, \tag{4.14}$$

where we denote X_1 for sample size with levels(20,30,50),X_2 for censoring rate with levels(0.2,0.3,0.5,0.7,0.8). W_1 is a dummy variable for the Lognormal distribution, W_2 is a dummy variable for the Loglogistic distribution and $\epsilon \sim N(0, \sigma^2)$. Specifically, $W_1 = 1, W_2 = 0$ denotes Lognormal distribution, $W_1 = 0, W_2 = 1$ denotes Loglogistic distribution, and $W_1 = 0, W_2 = 0$ denotes Weibull distribution.

The fitted model obtained from R for MPD is shown as Eq.(4.15):

$$\widehat{MPD} = 0.107 - 0.001 X_1 + 0.008 X_2 + 0.022 W_1 + 0.026 W_2. \tag{4.15}$$

with $R^2 = 0.7$, all factors are significant with p-values < 0.05, except censoring rate with p-value=0.424.

From the MPD fitted model Eq.(4.15), the coefficient of sample size is -0.001, which means if sample size increases by 10, we estimate that MPD decreases by 0.01, adjusted for the factors censoring rate, distribution. The estimated coefficient of censoring rate is 0.008, which means if censoring rate increases 0.1, the MPD increases by 0.0008 which is quite

small, adjusted for the factors sample size and distribution. Since the p-value for censoring rate is greater than 0.05, we concluded that the censoring rate effect is not statistically significant, the other factors being fixed.

In Eq.(4.15), the Weibull distribution is the reference category, and the coefficient is 0.022 for Lognormal distribution dummy variable and 0.026 for Loglogistic distribution dummy variable, adjusting for other factors. So, as expected the MPDs between *PH* and *AFT* analysis increase if the distribution is Lognormal or Loglogistic rather than Weibull, if the other factors are fixed. From Table 4.11, the MPD is smallest for Weibull distribution, and relatively smaller for Lognormal distribution than Loglogistic distribution.

4.5 PH and AFT Survival Analysis Application

A study was conducted on the effects of ploidy on the prognosis of patients with cancers of the mouth[6]. Patients were selected who had a paraffin-embedded sample of the cancerous tissue taken at the time of surgery. Follow-up survival data was obtained on each patient. The tissue samples were examined using a flow cytometer to determine if the tumor had an aneuploid (abnormal) or diploid (normal) DNA profile using a technique discussed in SickleCSantanello et al. (1988). The data in Table 4.12 is on patients with cancer of the tongue. Times are in weeks.

Table 4.12: Death times (in weeks) of patients with cancer of the tongue

Aneuploid Tumors:	
Death Times:	1, 3, 3, 4, 10, 13, 13, 16, 16, 24, 26, 27, 28, 30, 30, 32, 41, 51, 65, 67, 70, 72, 73, 77, 91, 93, 96, 100, 104, 157, 167
Censored Observations:	61, 74, 79, 80, 81, 87, 87, 88, 89, 93, 97, 101, 104, 108, 109, 120, 131, 150, 231, 240, 400
Diploid Tumors:	
Death Times:	1, 3, 4, 5, 5, 8, 12, 13, 18, 23, 26, 27, 30, 42, 56, 62, 69, 104, 104, 112, 129,181
Censored Observations:	8, 67, 76, 104, 176, 231

Table 4.13: Survival Analysis Application

	PH analysis	**AFT analysis**		
		Weibull	**Lognormal**	**Loglogisitc**
likelihood ratio test	0.3378	0.88	23.79	26.47
p-value	0.5611	0.3490	< 0.0001	< 0.0001

In the study patients were classified as having either an aneuploid or diploid DNA profile. As Weibull regression model is widely used in medical study, we applied weibull distribution to this data with a single covariate, Z, that is equal to 1 if the patient had an aneuploid DNA profile and 0 otherwise. In order to test the hypothesis of effect of ploidy on survival using the likelihood ratio test, I used maximum likelihood function to estimate Weibull distribution for Aneuploid tumor, in which weibull scale parameter $\lambda = 0.016$ and weibull shape parameter $\alpha = 0.832$. And Weibull distribution for Diploid tumor, in which weibull scale parameter $\lambda = 0.775$ and weibull shape parameter $\alpha = 0.036$.

I applied the PHREG procedure for PH analysis and LIFEREG procedure for AFT analysis with distributions Weibull, Lognormal and Loglogistic in SAS to test covariate effect of aneuploid DNA profile. The likelihood ratio test statistics are given in Table 4.13. Both PH analysis and AFT test using Weibull distribution give p-value greater than 0.05, which we failed to reject the null hypothesis, while the other AFT tests using lognormal and loglogistic distributions yield opposite conclusion. We could see that selection of distribution greatly effect AFT test.

An appealing feature of the Cox model is that the baseline hazard function is estimated nonparametrically, and so unlike most other statistical models, the survival times are not assumed to follow a particular statistical distribution. According to PH analysis output in Table 4.13, the estimated coefficient of variable tumor is $\hat{\beta} = 0.16929$, the $Se\{\hat{\beta}\} = 0.28955$, and the estimated hazard ratio is $\exp(\hat{\beta}) = 1.184$, which means the hazard risk of death would be 1.184 times larger if the tumor is Diploid than Aneuploid. While the confidence interval for β is (-0.398,0.737), and the confidence interval of hazard

ratio is $(\exp(-0.398), \exp(0.737)) = (0.671, 2.089)$. The hazard ratio confidence interval contain 1 indicates that the covariate(tumor type) effect is probably not associated with the event(death) probability , and thus not associated with the length of survival.

Chapter 5

Conclusion and Further Study

5.1 Conclusion

The Proportional hazard model is a class of survival model often used in Medical, Biological, and Engineering fields, etc. In a proportional hazards model, the effect of a unit increase in a covariate is multiplicative with respect to the hazard rate. Another type of survival model, the accelerated failure time model, does not require such proportional hazards. Both models are used to assess the effects of covariates on lifetimes. In order to study and compare the performance of these models, we applied PH model analysis to data simulated using AFT models. We simulated accelerated lifetime model data with one covariate and three covariates in Chapter 3. The comparison of estimated powers from the two models presented in Chapter 4 leads to the following conclusions.

1. Under extreme circumstance, such as low sample sizes (n = 10), or high censoring rates (p = 0.8), the PH model has higher percent of non-convergence estimates than the AFT model. Overall, both PH model and AFT analyses have low power in testing for covariate effects. But, the PH analysis has lower Type I error than AFT analysis, indicating its robustness.

2. When sample size increases from 20-50, the power plots of PH and AFT model become increasingly similar to each other, although there is still a statistically significant difference in power. The biggest difference in power exists when γ is small. Overall, AFT analysis has

a slightly higher power than PH analysis.

3. When sample size is large $n \geq 50$ and censoring rate is small $p \leq 0.3$, there is not a statistically significant difference in type I error rates for the two analyses. Otherwise, PH analysis has a slightly lower type I error rate than AFT analysis.

Since using an AFT analysis requires specifying the distribution of lifetime data, we would prefer applying a PH analysis rather than AFT analysis in practice. Generally, we conclude that PH model is quite robust in terms of type I error rate and type II error rate with respect to AFT model in applied statistics.

5.2 Further Study

There are some questions in my simulation and analysis need to be considered in future work.

(1) Censoring index

In my study, I simplified the generation of censoring index as an independent variable with lifetime as shown in Section 3.2.2. In the future work, we should simulate the censoring distribution as well as the lifetime distribution.

(2) Distribution

In the report, I only compared the three distributions(Weibull, Lognormal and Loglogistic) which are used widely in biology and engineering. And I selected specific distributions with similar shape from the three families as shown in Section 3.2.1. We don't know anything about the performance of PH and AFT tests on other distributions in these families or other families of distribution, which limits the applicability of the conclusions we obtained in this report.

(3) Other types of censoring data

In my report, I only simulated the right censoring data with censoring rate $p = 0.2 \sim 0.9$. There are other types of censoring data like left censoring, interval censoring, and truncated data.

(4) Assumption of applying correct distribution in AFT analysis

A limitation of using AFT analysis is that we need to specify the distribution of the data. In my simulation, I always applied the correct distribution for AFT analysis. Consequently, we don't know anything about the performance of AFT analysis in case of using the wrong distributions. Future studies should compare the power performance of PH analysis and AFT analysis with the wrong distributions.

(5) How fast power reaches 1

As we look at the power plots for the two analyses, it seems that AFT analysis converges faster in power than PH analysis until power reaches 1. This behavior should be investigated in future work.

(6) Computing time

Because the simulated data are stored in a large dimension matrix, as shown in Section **??**, it consumed a lot of computer memory and required a long time to execute the analyses. I suggest that future studies use multi-thread programming to decrease the running time.

Bibliography

[1] P.D. Allison. *Survival analysis using SAS: A practical guide*. SAS publishing, 2010.

[2] T. Bellotti and J. Crook. Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, 60(12):1699–1707, 2008.

[3] D. M. Dabrowska, K. A. Doksum, and J. K. Song. *Graphical Comparison of Cumulative Hazards for Two Populations*, volume 76. Biometrika, 1989.

[4] Cox D.R. and Oakes D. *Analysis of survival data*, volume 2nd Ed. Chapman & Hall, 1984.

[5] T. A. Gooley, J. Leisenring, W.and Crowley, and B. Storer. Estimation of failure probabilities in the presence of competing risks: New representations of old estimators. *Statistics in Medicine*, 18:695–706, 1999.

[6] Melvin L. Moeschberger John P. Klein. *SURVIVAL ANALYSIS Techniques for Censored and Truncated Data*. Springer-Verlag New York Berlin Heidelberg, 2003.

[7] X. Luo, G.Z. Cui, F.L. Le, and S.J. Wang. Tests of the functional coefficients in the varying-coefficient cox proportional hazard model. *IEEE*, pages 5708–5713, 2011.

[8] G. Zhao. *Nonparametric and parametric survival analysis of censored data with possible violation of method assumptions*. The University of North Carolina at Greensboro, 2008.

[9] Ming Zhong and Kenneth R. Hess. Mean Survival Time from Right Censored Data. *The Berkeley Electronic Press*, 2009.

Appendix A

Tables

Table A.1: Non-convergency rates for Weibull, 1 covariate

PH (%)	AFT (%)	gamma	Censoring rate	Sample Size
1.2	1.6	0	0.8	10
0.9	1.3	0.5	0.8	10
1	1.4	1	0.8	10
1.3	1.1	1.5	0.8	10
1.2	1.8	2	0.8	10
1.1	1.4	2.5	0.8	10
1.8	1.6	3	0.8	10
1.2	1.3	3.5	0.8	10
1.1	1.8	4	0.8	10
2	1.7	4.5	0.8	10
1.7	1.7	5	0.8	10
0	0.2	0.5	0.8	20
0	0.3	1	0.8	20
0	0.1	1.5	0.8	20
0.3	0.4	2.5	0.8	20
0	0.2	3	0.8	20
0.1	0.2	3.5	0.8	20
0.3	0.2	4	0.8	20
0.4	0	4.5	0.8	20
0.2	0.2	5	0.8	20
0	0.1	0	0.5	10
0	0.1	0.5	0.5	10
0.2	0.1	1.5	0.5	10
0.4	0	2	0.5	10
0.1	0	2.5	0.5	10
0.4	0	3.5	0.5	10
0.6	0	4	0.5	10
0.5	0.1	4.5	0.5	10
2.9	8.2	0	0.9	10
0.9	9.1	0.5	0.9	10
1.6	7.7	1	0.9	10
1.7	8.3	1.5	0.9	10
2	7.8	2	0.9	10
2.7	6.7	2.5	0.9	10
2	6.7	3	0.9	10
1.8	5.6	3.5	0.9	10
1.6	7.4	4	0.9	10
2.3	5.7	4.5	0.9	10
2.3	5.9	5	0.9	10

Table A.2: Non-convergency rates for Lognormal, 1 covariate

PH(%)	AFT(%)	gamma	censor rate	Sample size
0.6	0	0	0.8	10
1	0	0.5	0.8	10
0.9	0	1	0.8	10
1.2	0	1.5	0.8	10
0.8	0	2	0.8	10
1.7	0	2.5	0.8	10
2.1	0	3	0.8	10
1.3	0	3.5	0.8	10
1.1	0	4	0.8	10
1.9	0	4.5	0.8	10
1.5	0	5	0.8	10
0.2	0	0	0.8	20
0.1	0	1	0.8	20
0.1	0	1.5	0.8	20
0.1	0	2	0.8	20
0.2	0	2.5	0.8	20
0.2	0	3	0.8	20
0.2	0	3.5	0.8	20
0.1	0	4	0.8	20
0.3	0	4.5	0.8	20
0.1	0	5	0.8	20
0.1	0	0.5	0.5	10
0.3	0	1.5	0.5	10
0.1	0	2.5	0.5	10
0.3	0	3	0.5	10
0.4	0	3.5	0.5	10
0.3	0	4	0.5	10
0.7	0	4.5	0.5	10
0.9	0	5	0.5	10
1.8	0.3	0	0.9	10
1.3	0.5	0.5	0.9	10
1.9	0.3	1	0.9	10
2.3	0.4	1.5	0.9	10
1.8	0.1	2	0.9	10
2.4	0.3	2.5	0.9	10
1.5	0.2	3	0.9	10
1.3	0.4	3.5	0.9	10
2	0.3	4	0.9	10
2.4	0.2	4.5	0.9	10
2	0.2	5	0.9	10

Table A.3: Non-convergency rates for Loglogistic, 1 covariate

PH(%)	AFT(%)	gamma	censor rate	Sample size
0.9	0	0	0.8	10
0.9	0	0.5	0.8	10
1.2	0	1	0.8	10
0.5	0	1.5	0.8	10
0.9	0	2	0.8	10
1.2	0	2.5	0.8	10
1.4	0	3	0.8	10
1.5	0	3.5	0.8	10
1.4	0	4	0.8	10
2	0	4.5	0.8	10
1.8	0	5	0.8	10
0.1	0	0.5	0.8	20
0.1	0	2.5	0.8	20
0.1	0	3	0.8	20
0.1	0	4.5	0.8	20
0.1	0	5	0.8	20
0.1	0	1	0.5	10
0.1	0	1.5	0.5	10
0.1	0	2.5	0.5	10
0.2	0	3.5	0.5	10
0.3	0	4	0.5	10
0.3	0	4.5	0.5	10
0.5	0	5	0.5	10
1.3	0.5	0	0.9	10
1.6	0.9	0.5	0.9	10
2.1	0.1	1	0.9	10
0.9	0.3	1.5	0.9	10
1.3	0.5	2	0.9	10
2	0.3	2.5	0.9	10
2	0.2	3	0.9	10
2.1	0.6	3.5	0.9	10
1.4	0.3	4	0.9	10
1.6	0.2	4.5	0.9	10
1.3	0.6	5	0.9	10

Table A.4: Type I error rates for 3 covariate model

PH	AFT	censor rate	sample size	distribution
0.119	0.216	0.2	10	Weibull
0.124	0.178	0.2	10	lognormal
0.134	0.207	0.2	10	loglogistic
0.154	0.223	0.3	10	Weibull
0.123	0.18	0.3	10	lognormal
0.155	0.228	0.3	10	loglogistic
0.177	0.267	0.5	10	Weibull
0.192	0.279	0.5	10	lognormal
0.196	0.281	0.5	10	loglogistic
0.263	0.401	0.7	10	Weibull
0.289	0.382	0.7	10	lognormal
0.256	0.347	0.7	10	loglogistic
0.092	0.492	0.8	10	Weibull
0.097	0.444	0.8	10	lognormal
0.097	0.463	0.8	10	loglogistic
0.082	0.101	0.2	20	Weibull
0.075	0.092	0.2	20	lognormal
0.094	0.114	0.2	20	loglogistic
0.087	0.111	0.3	20	Weibull
0.073	0.096	0.3	20	lognormal
0.095	0.113	0.3	20	loglogistic
0.093	0.12	0.5	20	Weibull
0.088	0.103	0.5	20	lognormal
0.088	0.111	0.5	20	loglogistic
0.113	0.163	0.7	20	Weibull
0.113	0.138	0.7	20	lognormal
0.093	0.137	0.7	20	loglogistic
0.106	0.171	0.8	20	Weibull
0.129	0.161	0.8	20	lognormal
0.123	0.178	0.8	20	loglogistic
0.09	0.099	0.2	30	Weibull
0.073	0.079	0.2	30	lognormal
0.066	0.073	0.2	30	loglogistic
0.085	0.101	0.3	30	Weibull
0.065	0.077	0.3	30	lognormal
0.064	0.079	0.3	30	loglogistic

Table A.4: Type I error rates for 3 covariate model

PH	AFT	censor rate	sample size	distribution
0.09	0.117	0.5	30	Weibull
0.072	0.089	0.5	30	lognormal
0.072	0.099	0.5	30	loglogistic
0.085	0.118	0.7	30	Weibull
0.085	0.104	0.7	30	lognormal
0.084	0.119	0.7	30	loglogistic
0.102	0.138	0.8	30	Weibull
0.085	0.101	0.8	30	lognormal
0.064	0.106	0.8	30	loglogistic
0.058	0.070	0.2	50	Weibull
0.066	0.080	0.2	50	lognormal
0.06	0.077	0.2	50	loglogistic
0.063	0.077	0.3	50	Weibull
0.058	0.068	0.3	50	lognormal
0.065	0.070	0.3	50	loglogistic
0.063	0.076	0.5	50	Weibull
0.06	0.065	0.5	50	lognormal
0.059	0.061	0.5	50	loglogistic
0.055	0.069	0.7	50	Weibull
0.080	0.082	0.7	50	lognormal
0.064	0.079	0.7	50	loglogistic
0.069	0.087	0.8	50	Weibull
0.069	0.083	0.8	50	lognormal
0.062	0.077	0.8	50	loglogistic
0.059	0.074	0.2	70	Weibull
0.067	0.067	0.2	70	lognormal
0.046	0.04	0.2	70	loglogistic
0.055	0.067	0.5	70	Weibull
0.07	0.061	0.5	70	lognormal
0.057	0.062	0.5	70	loglogistic
0.061	0.068	0.8	70	Weibull
0.067	0.084	0.8	70	lognormal
0.056	0.079	0.8	70	loglogistic

Table A.5: Four Fold Table for ERD, 1 covariate model

(a) n=30, p=0.3, Weibull

ERD		PH	
		reject	not reject
AFT	reject	41	25
	not reject	6	928

(b) n=30, p=0.3, Lognormal

ERD		PH	
		reject	not reject
AFT	reject	33	20
	not reject	22	925

(c) n=30, p=0.3, Loglogistic

ERD		PH	
		reject	not reject
AFT	reject	58	16
	not reject	10	916

(d) n=30, p=0.2, Weibull

ERD		PH	
		reject	not reject
AFT	reject	45	20
	not reject	7	927

(e) n=30, p=0.2, Lognormal

ERD		PH	
		reject	not reject
AFT	reject	35	31
	not reject	19	915

(f) n=30, p=0.2, Loglogistic

ERD		PH	
		reject	not reject
AFT	reject	44	22
	not reject	18	916

(g) n=50, p=0.5, Weibull

ERD		PH	
		reject	not reject
AFT	reject	46	13
	not reject	15	926

(h) n=30, p=0.3, Lognormal

ERD		PH	
		reject	not reject
AFT	reject	35	20
	not reject	19	926

(i) n=30, p=0.3, Loglogistic

ERD		PH	
		reject	not reject
AFT	reject	31	21
	not reject	12	936

(j) n=50, p=0.7, Weibull

ERD		PH	
		reject	not reject
AFT	reject	50	20
	not reject	7	922

(k) n=50, p=0.7, Lognormal

ERD		PH	
		reject	not reject
AFT	reject	38	18
	not reject	14	930

(l) n=50, p=0.7, Loglogistic

ERD		PH	
		reject	not reject
AFT	reject	44	23
	not reject	5	928

Table A.6: Maximum power differences for 3 covariate model

$\max\{\lvert PH - AFT\rvert\}$	censor	size	distribution
0.206	0.2	10	Weibull
0.171	0.2	10	lognormal
0.175	0.2	10	loglogistic
0.236	0.3	10	Weibull
0.182	0.3	10	lognormal
0.182	0.3	10	loglogistic
0.194	0.5	10	Weibull
0.171	0.5	10	lognormal
0.168	0.5	10	loglogistic
0.225	0.7	10	Weibull
0.220	0.7	10	lognormal
0.240	0.7	10	loglogistic
0.512	0.8	10	Weibull
0.535	0.8	10	lognormal
0.537	0.8	10	loglogistic
0.130	0.2	20	Weibull
0.108	0.2	20	lognormal
0.115	0.2	20	loglogistic
0.141	0.3	20	Weibull
0.114	0.3	20	lognormal
0.109	0.3	20	loglogistic
0.148	0.5	20	Weibull
0.112	0.5	20	lognormal
0.135	0.5	20	loglogistic
0.165	0.7	20	Weibull
0.140	0.7	20	lognormal
0.126	0.7	20	loglogistic
0.175	0.8	20	Weibull
0.159	0.8	20	lognormal
0.160	0.8	20	loglogistic

Table A.7: Maximum power differences for 3 covariate model

$\max\{\lvert PH - AFT\rvert\}$	censor	size	distribution
0.081	0.2	30	Weibull
0.094	0.2	30	lognormal
0.078	0.2	30	loglogistic
0.078	0.3	30	Weibull
0.110	0.3	30	lognormal
0.115	0.3	30	loglogistic
0.101	0.5	30	Weibull
0.120	0.5	30	lognormal
0.125	0.5	30	loglogistic
0.132	0.7	30	Weibull
0.122	0.7	30	lognormal
0.124	0.7	30	loglogistic
0.152	0.8	30	Weibull
0.135	0.8	30	lognormal
0.126	0.8	30	loglogistic
0.016	0.2	50	Weibull
0.057	0.2	50	lognormal
0.067	0.2	50	loglogistic
0.020	0.3	50	Weibull
0.079	0.3	50	lognormal
0.063	0.3	50	loglogistic
0.058	0.5	50	Weibull
0.097	0.5	50	lognormal
0.104	0.5	50	loglogistic
0.103	0.7	50	Weibull
0.122	0.7	50	lognormal
0.126	0.7	50	loglogistic
0.112	0.8	50	Weibull
0.100	0.8	50	lognormal
0.120	0.8	50	loglogistic

Table A.8: Four Fold Table for MPD, 1 covariate model

(a) n=20, p=0.2, Weibull

MPD		PH	
		reject	not reject
AFT	reject	642	89
	not reject	20	249

(b) n=20, p=0.2, Lognormal

MPD		PH	
		reject	not reject
AFT	reject	537	132
	not reject	12	319

(c) n=20, p=0.2, Loglogistic

MPD		PH	
		reject	not reject
AFT	reject	646	123
	not reject	15	216

(d) n=30, p=0.2, Weibull

MPD		PH	
		reject	not reject
AFT	reject	534	79
	not reject	20	367

(e) n=30, p=0.2, Lognormal

MPD		PH	
		reject	not reject
AFT	reject	450	123
	not reject	29	398

(f) n=30, p=0.2, Loglogistic

MPD		PH	
		reject	not reject
AFT	reject	681	122
	not reject	18	179

(g) n=30, p=0.5, Weibull

MPD		PH	
		reject	not reject
AFT	reject	459	84
	not reject	15	442

(h) n=30, p=0.5, Lognormal

MPD		PH	
		reject	not reject
AFT	reject	476	101
	not reject	21	402

(i) n=30, p=0.5, Loglogistic

MPD		PH	
		reject	not reject
AFT	reject	437	92
	not reject	19	434

(j) n=50, p=0.5, Weibull

MPD		PH	
		reject	not reject
AFT	reject	253	37
	not reject	20	690

(k) n=50, p=0.5, Lognormal

MPD		PH	
		reject	not reject
AFT	reject	834	72
	not reject	4	90

(l) n=50, p=0.5, Loglogistic

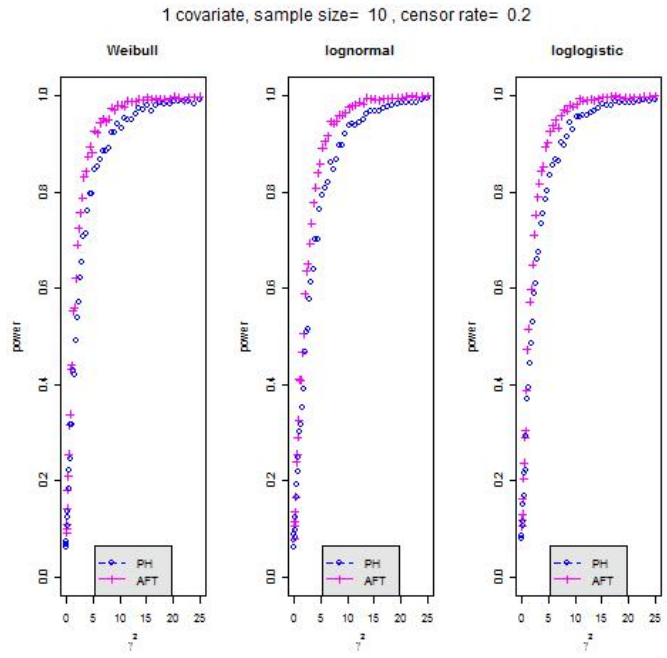
MPD		PH	
		reject	not reject
AFT	reject	414	110
	not reject	14	462

Appendix B

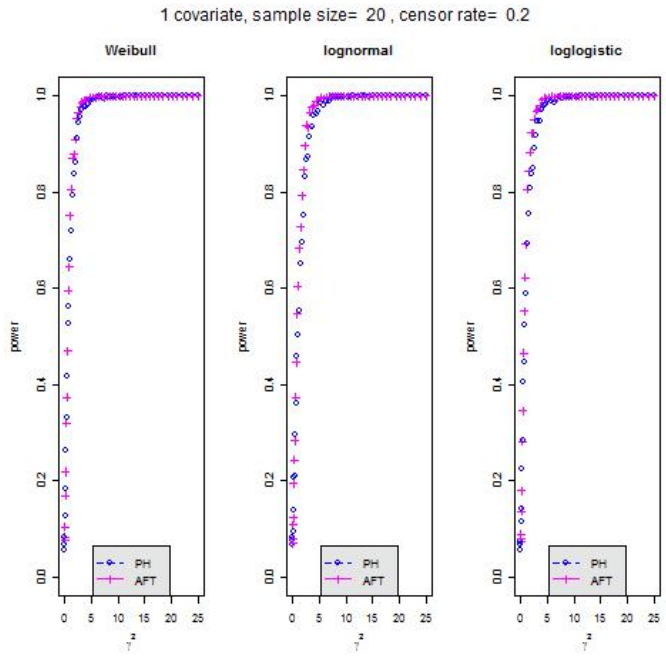
Graphs

Figure B.1: Power plots for 1 covariate model

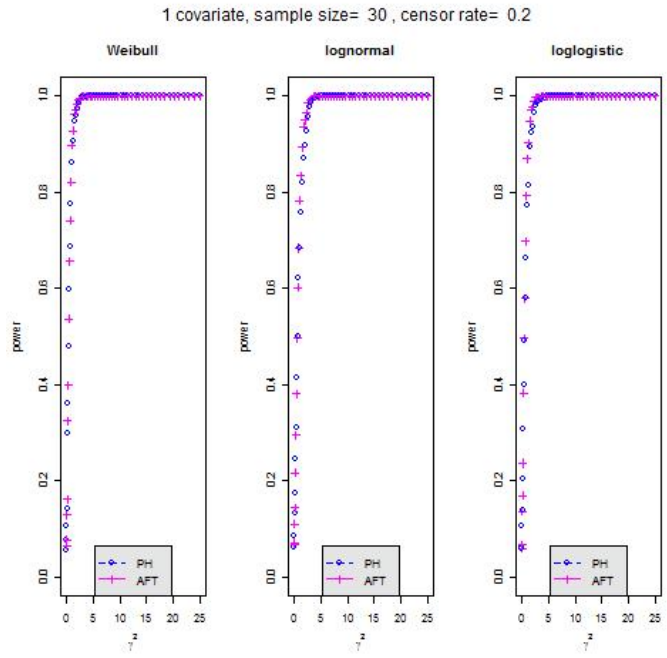
(a) n=10,p=0.2



(b) n=20,p=0.2



(c) n=30,p=0.2



(d) n=50,p=0.2

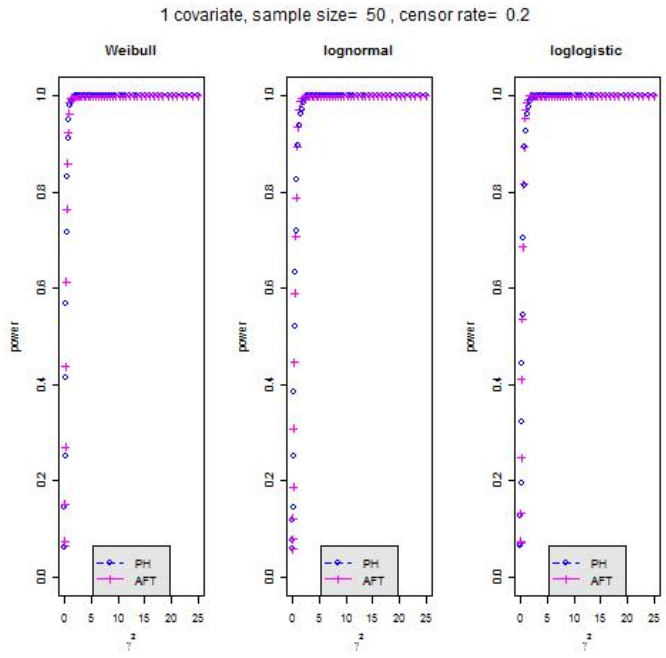
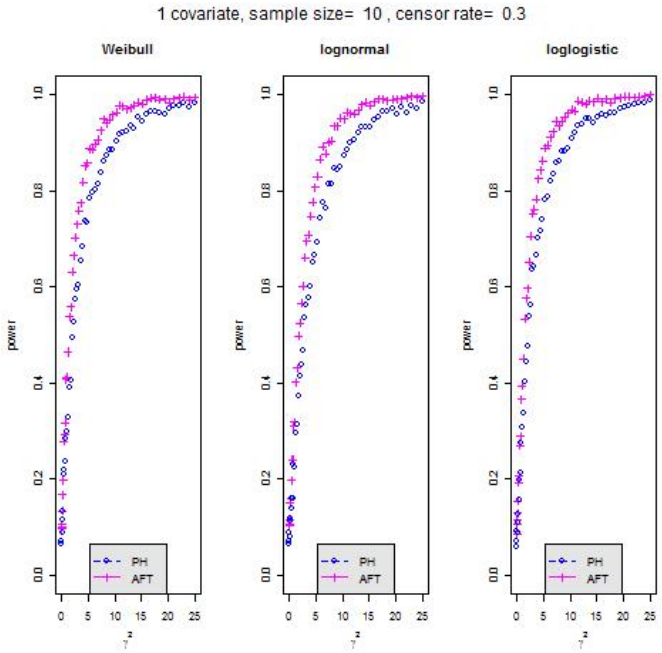
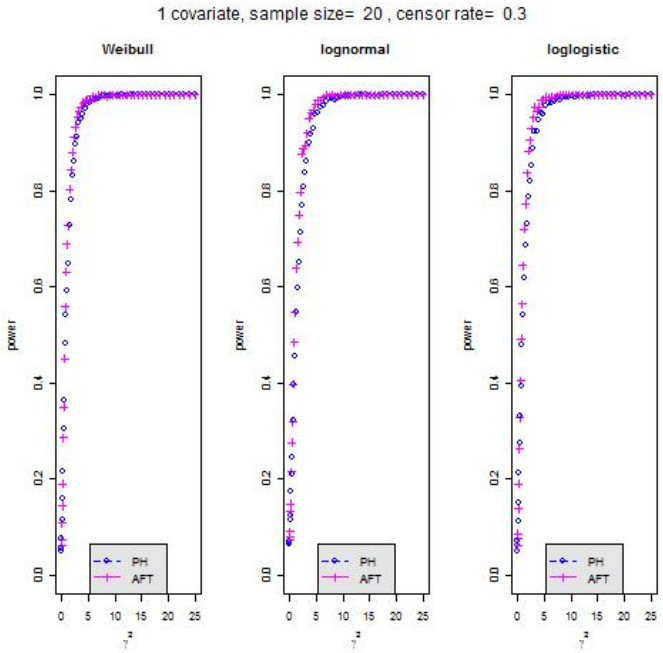


Figure B.1: Power plots for 1 covariate model

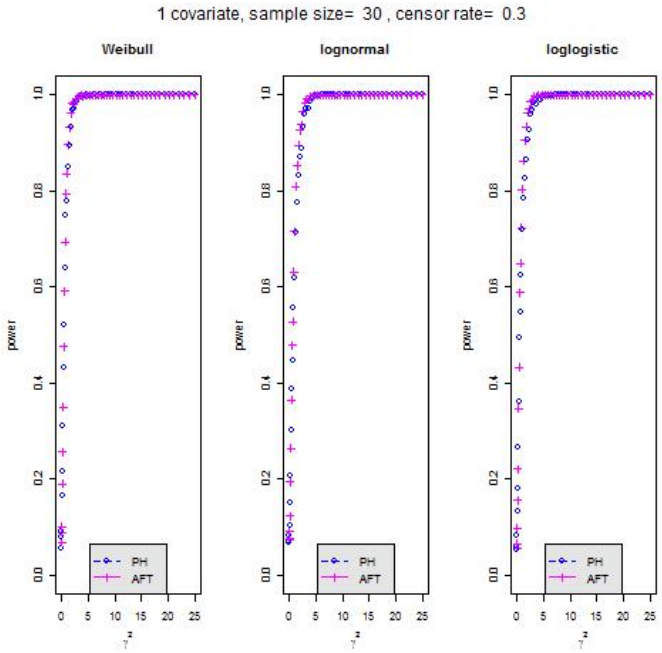
(e) n=10,p=0.3



(f) n=20,p=0.3



(g) n=30,p=0.3



(h) n=50,p=0.3

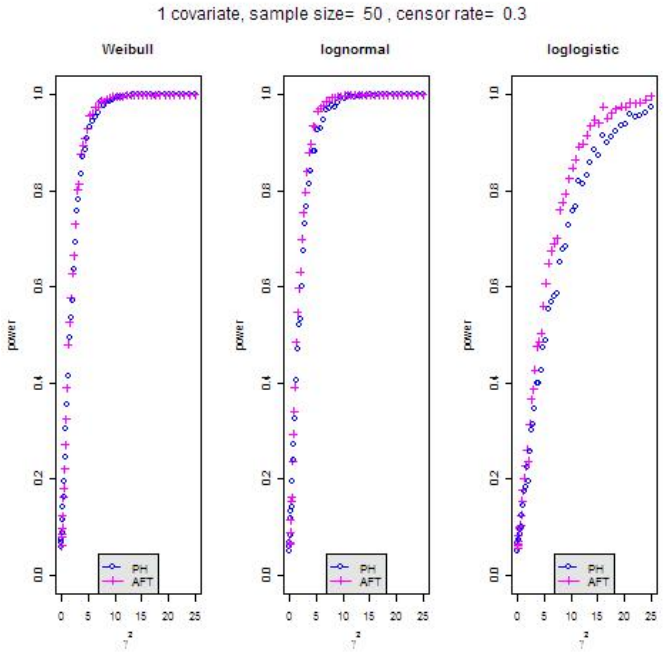
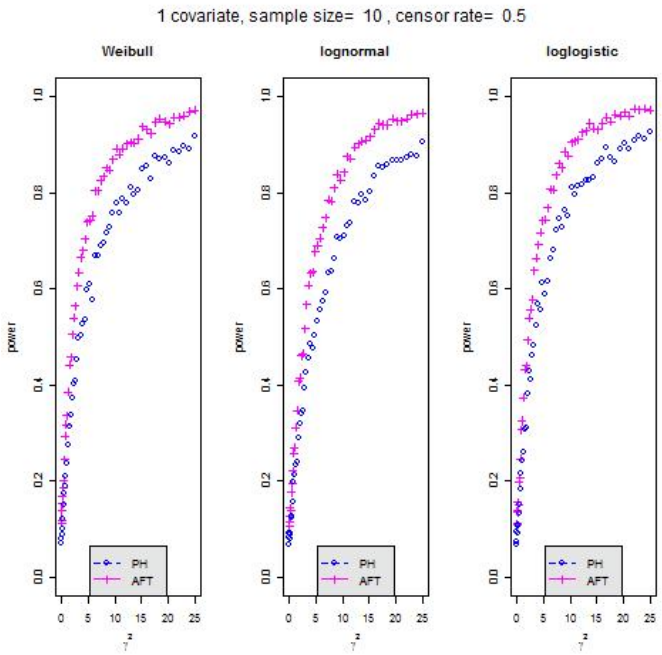
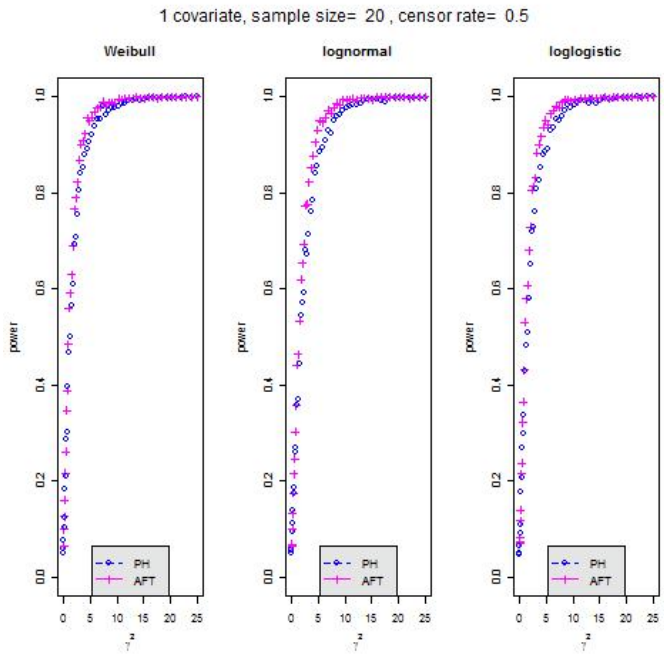


Figure B.1: Power plots for 1 covariate model

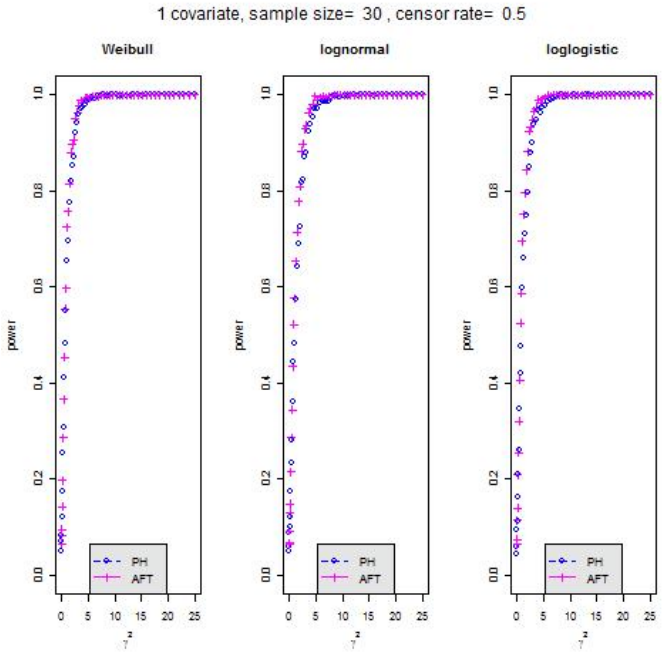
(i) n=10,p=0.5



(j) n=20,p=0.5



(k) n=30,p=0.5



(l) n=50,p=0.5

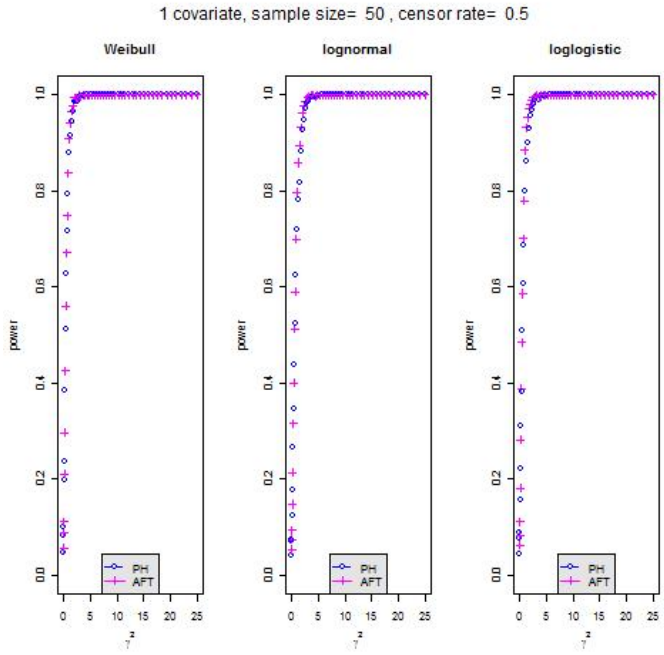
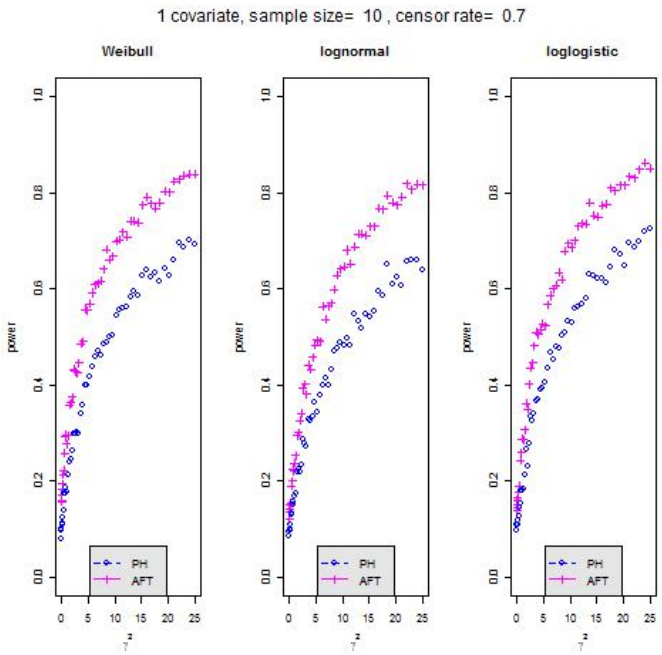
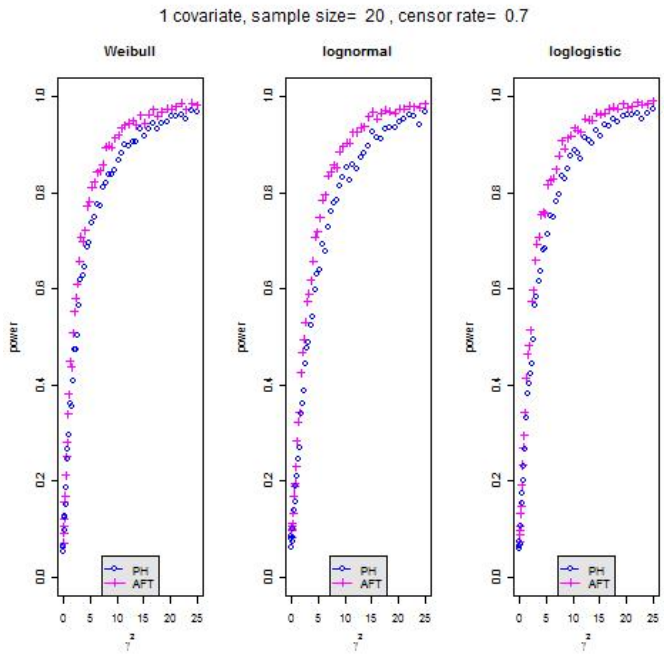


Figure B.1: Power plots for 1 covariate model

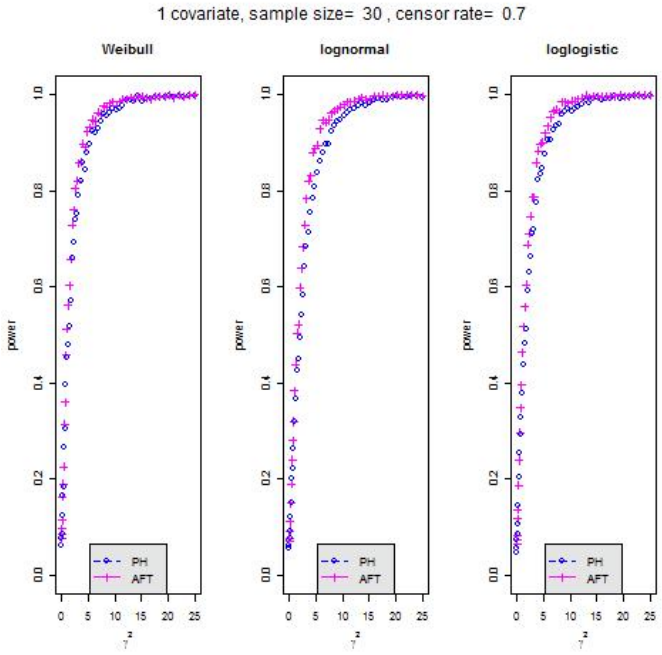
(m) n=10,p=0.7



(n) n=20,p=0.7



(o) n=30,p=0.7



(p) n=50,p=0.7

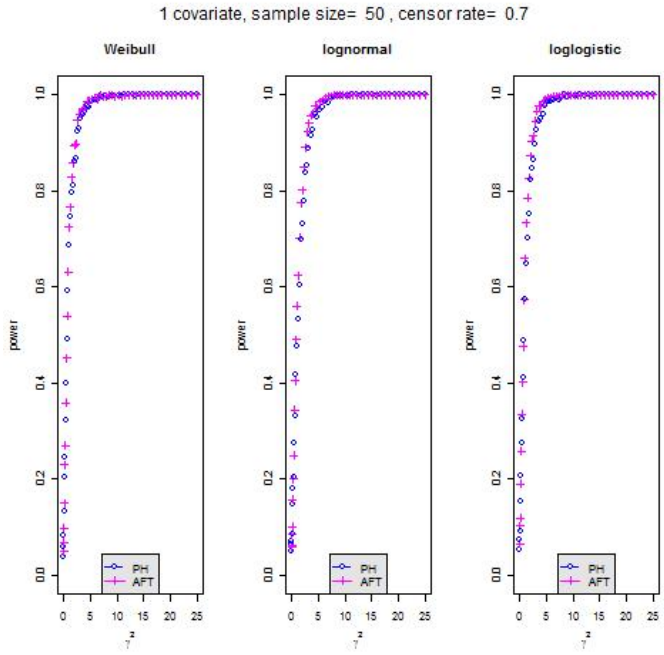
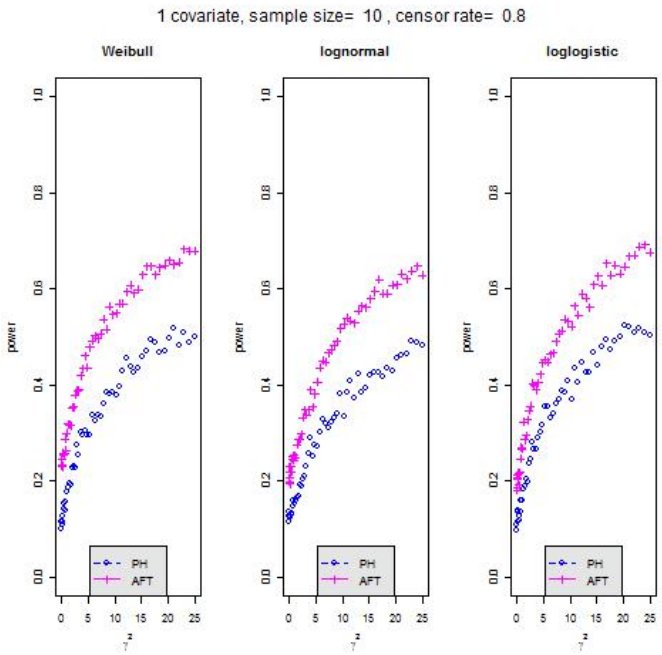
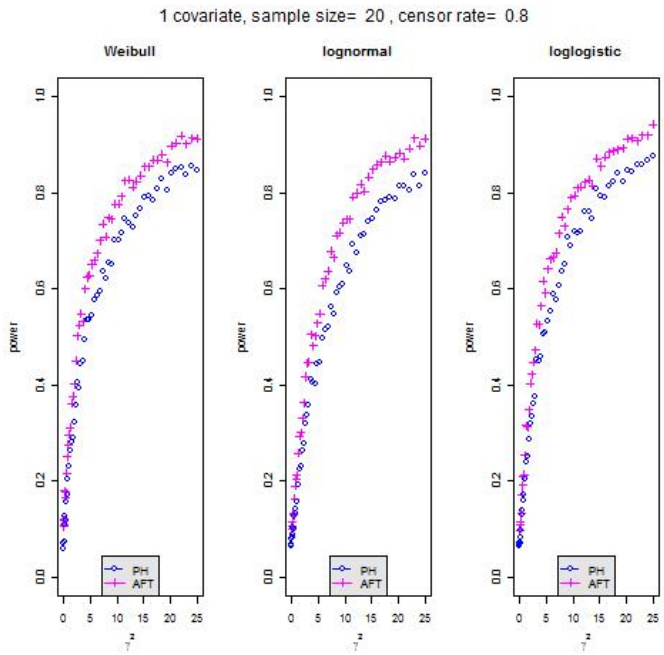


Figure B.1: Power plots for 1 covariate model

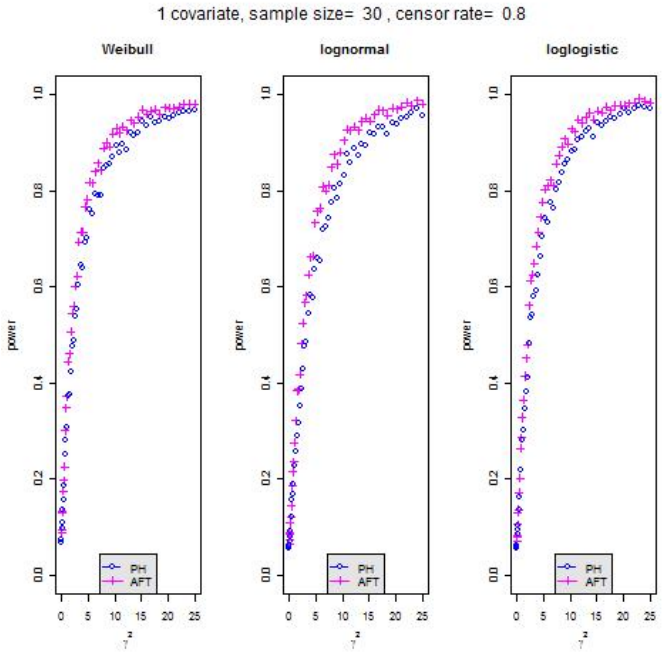
(q) n=10,p=0.8



(r) n=20,p=0.8



(s) n=30,p=0.8



(t) n=50,p=0.8

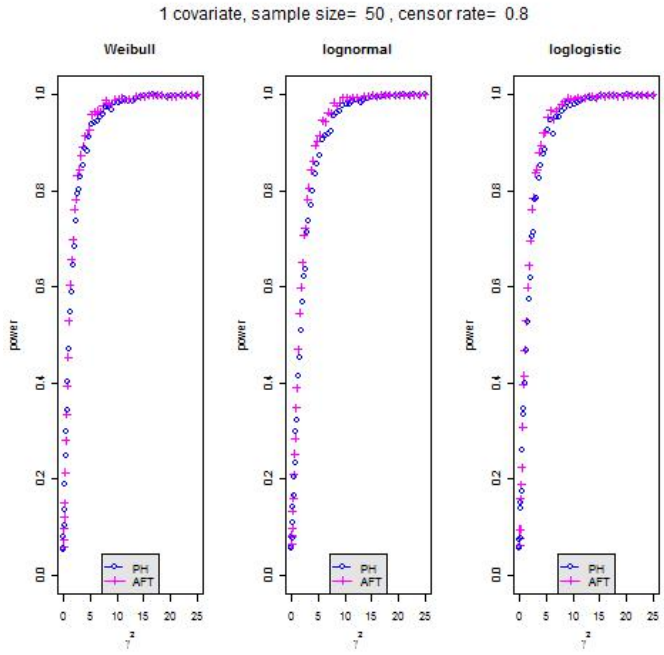
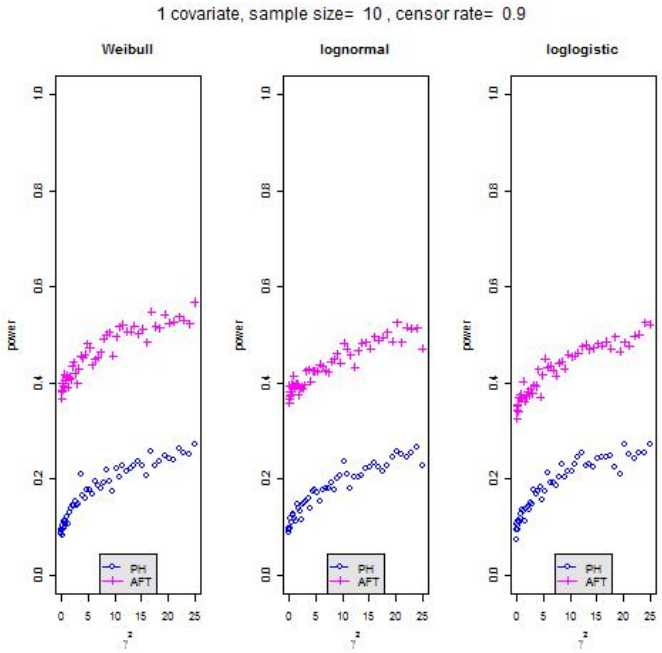
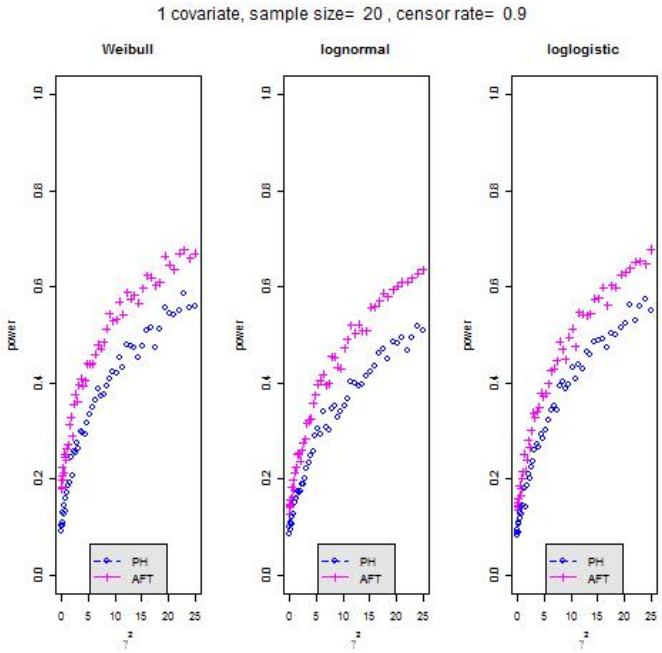


Figure B.1: Power plots for 1 covariate model

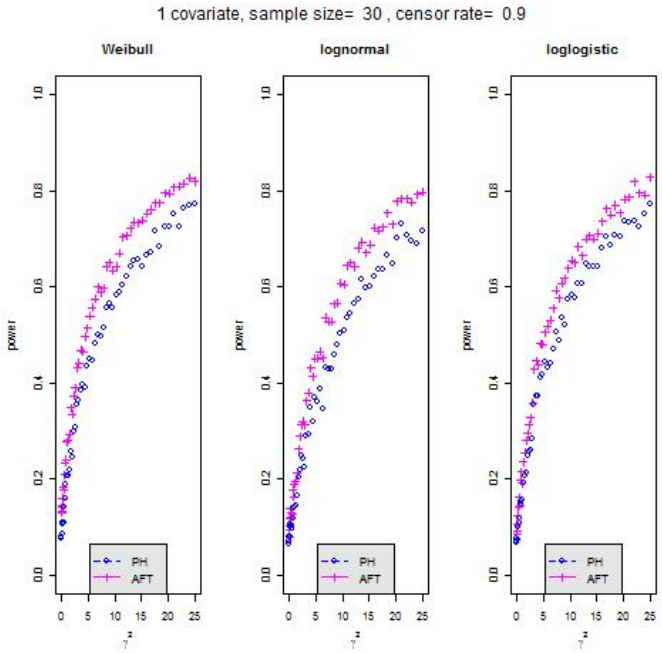
(u) n=10,p=0.9



(v) n=20,p=0.9



(w) n=30,p=0.9



(x) n=50,p=0.9

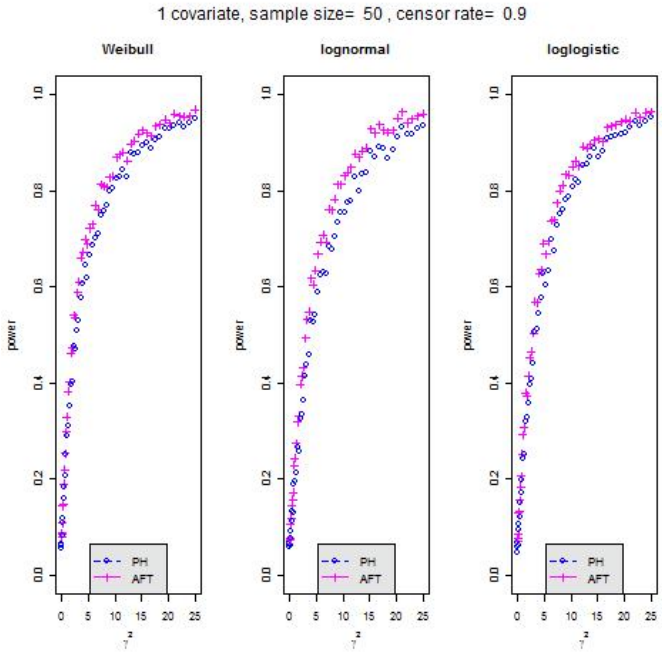
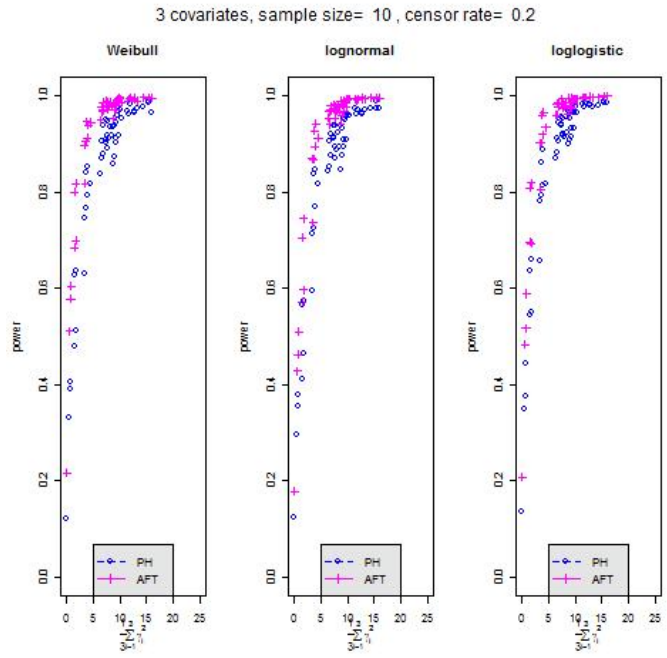
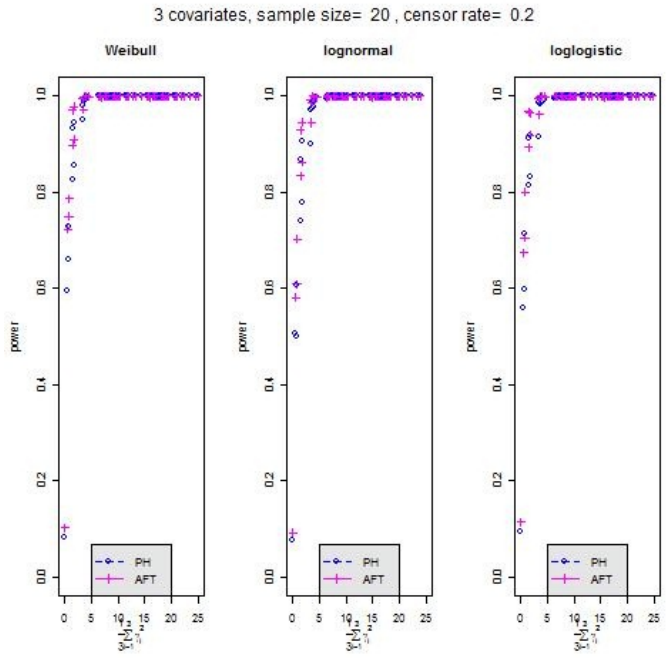


Figure B.2: Power plots for 3 covariate model

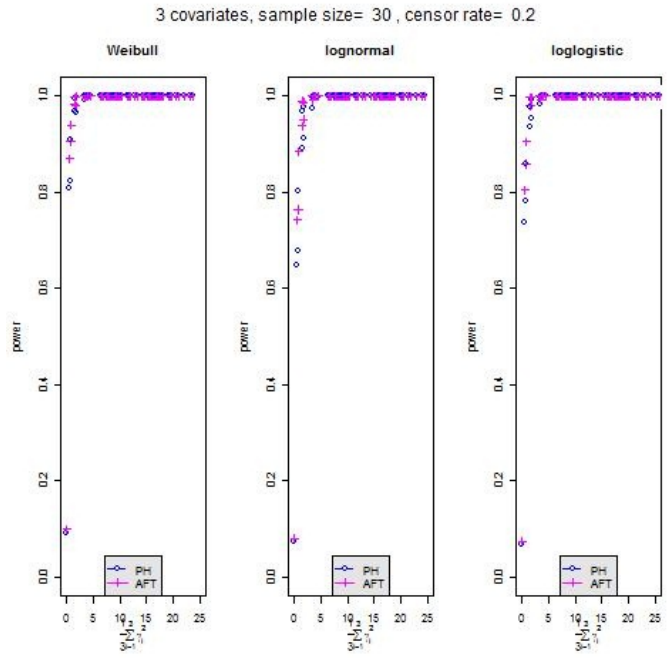
(a) n=10,p=0.2



(b) n=20,p=0.2



(c) n=30,p=0.2



(d) n=50,p=0.2

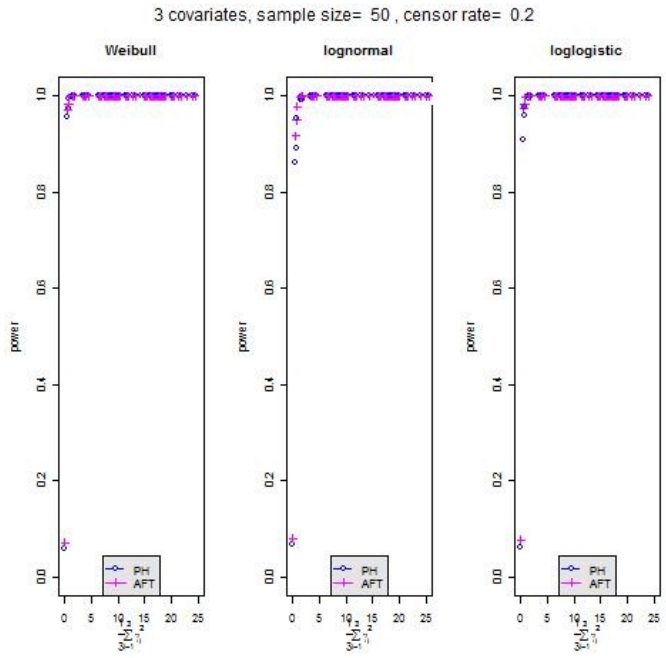
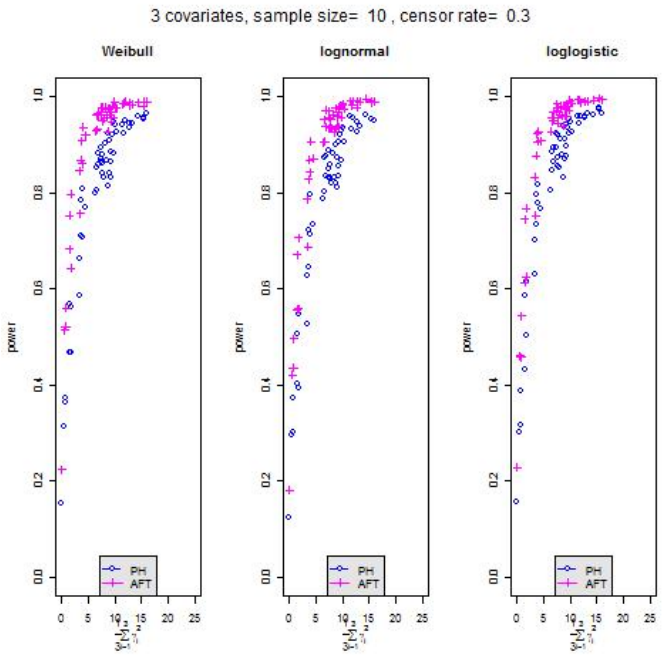
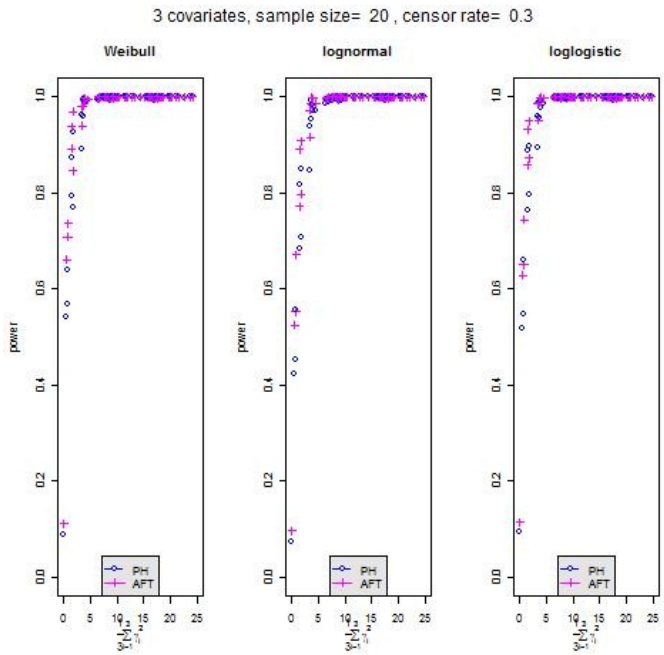


Figure B.2: Power plots for 3 covariate model

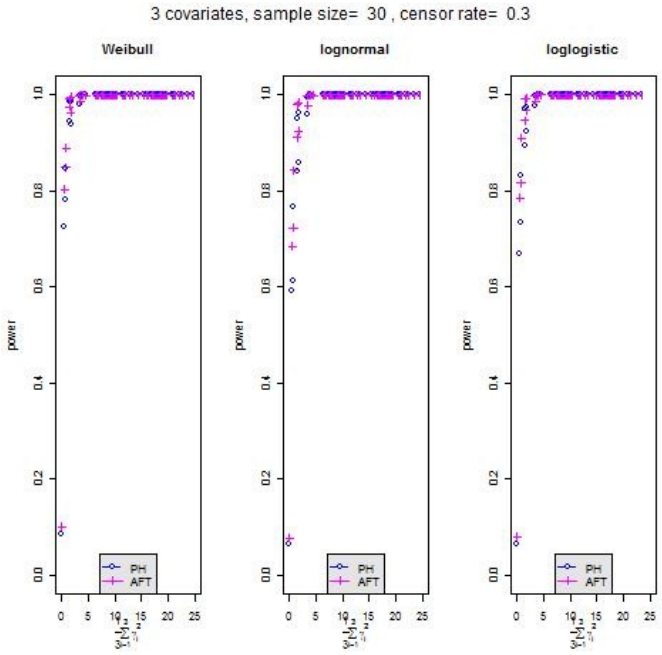
(e) n=10,p=0.3



(f) n=20,p=0.3



(g) n=30,p=0.3



(h) n=50,p=0.3

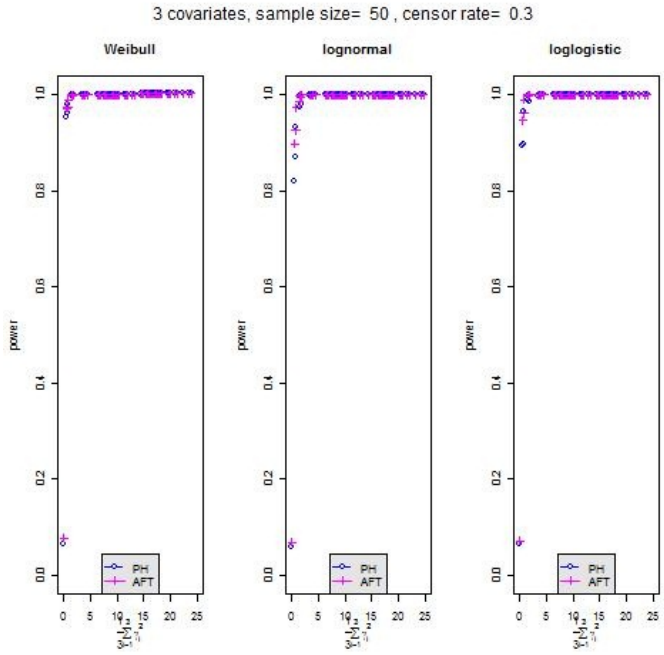
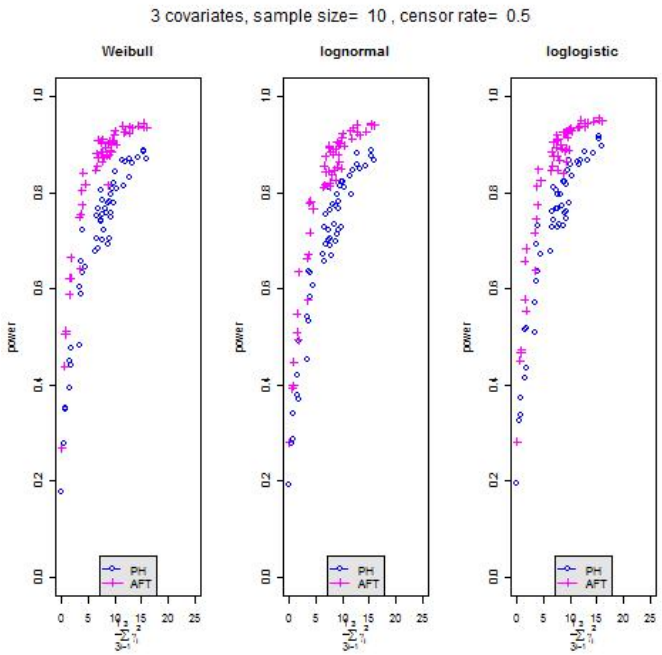
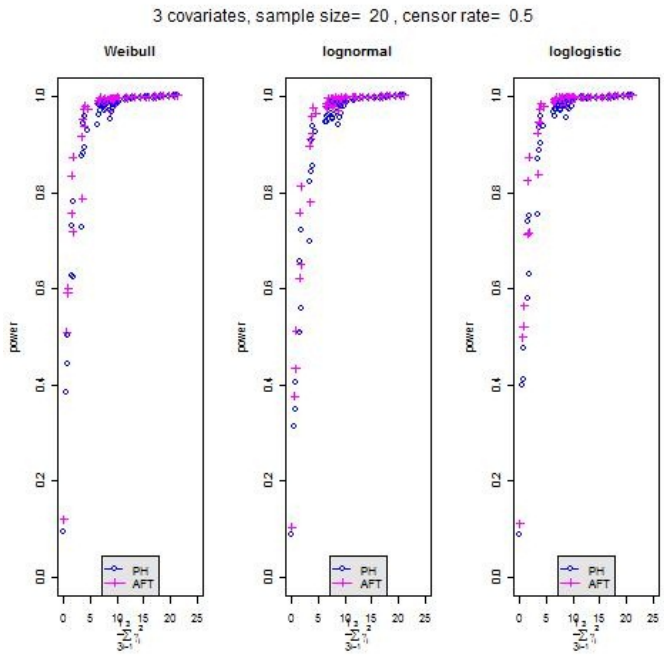


Figure B.2: Power plots for 3 covariate model

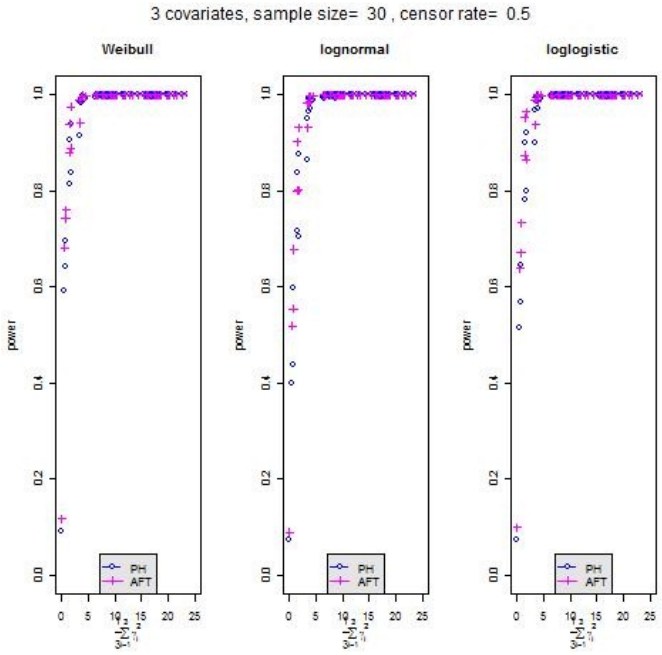
(i) n=10,p=0.5



(j) n=20,p=0.5



(k) n=30,p=0.5



(l) n=50,p=0.5

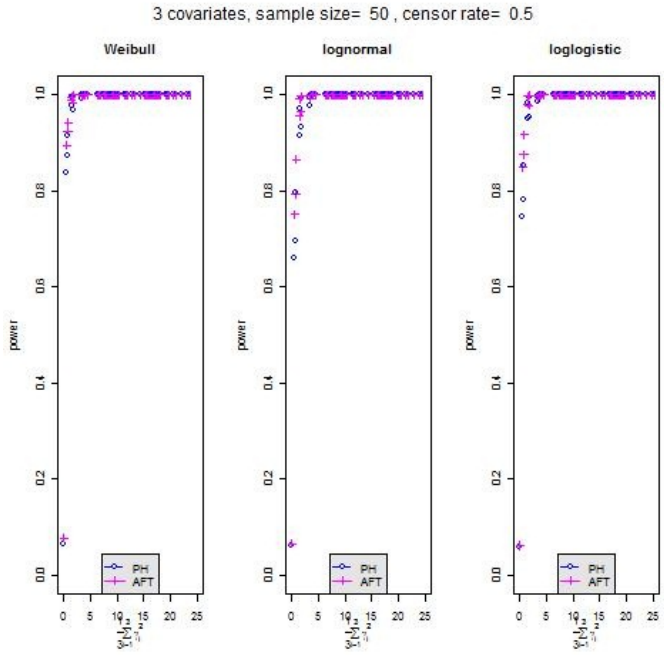
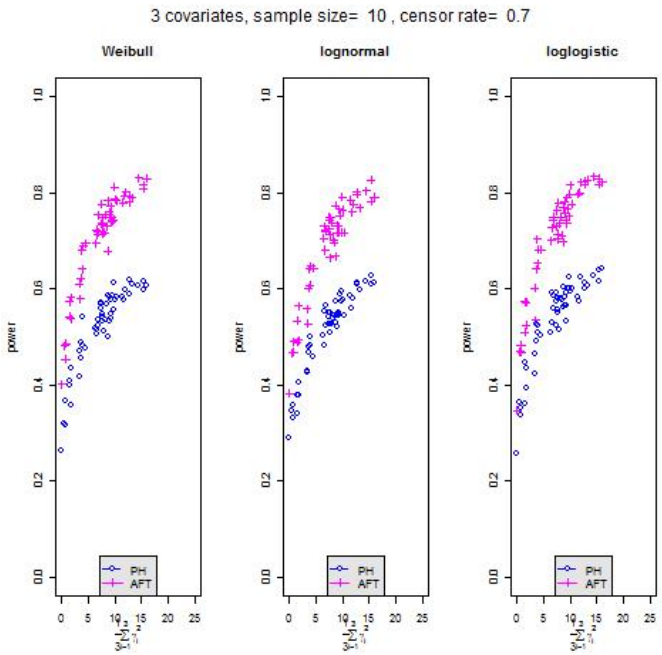
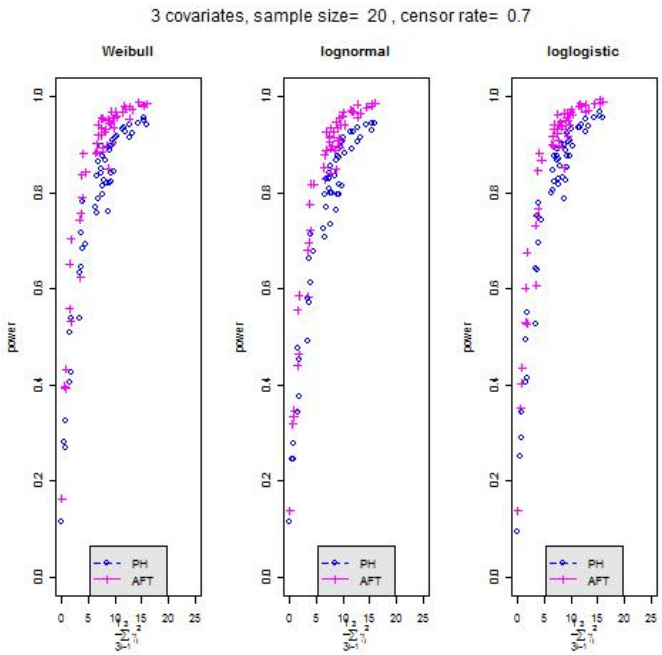


Figure B.2: Power plots for 3 covariate model

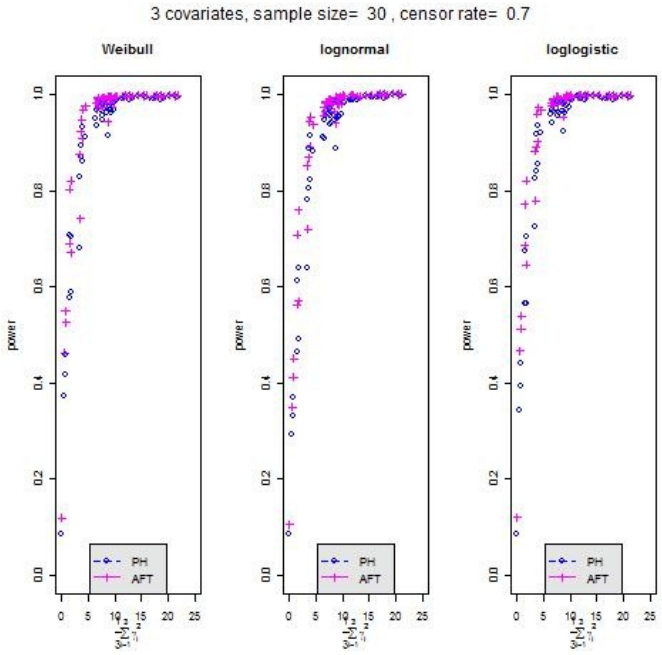
(m) n=10,p=0.7



(n) n=20,p=0.7



(o) n=30,p=0.7



(p) n=50,p=0.7

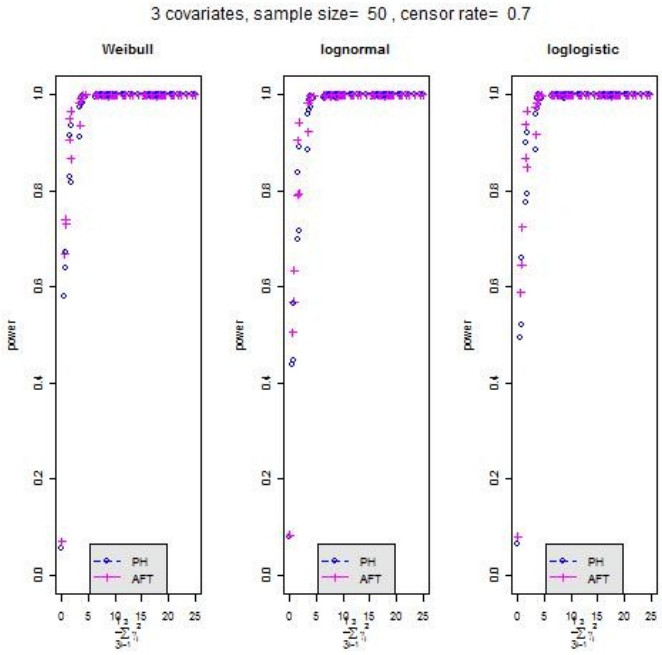
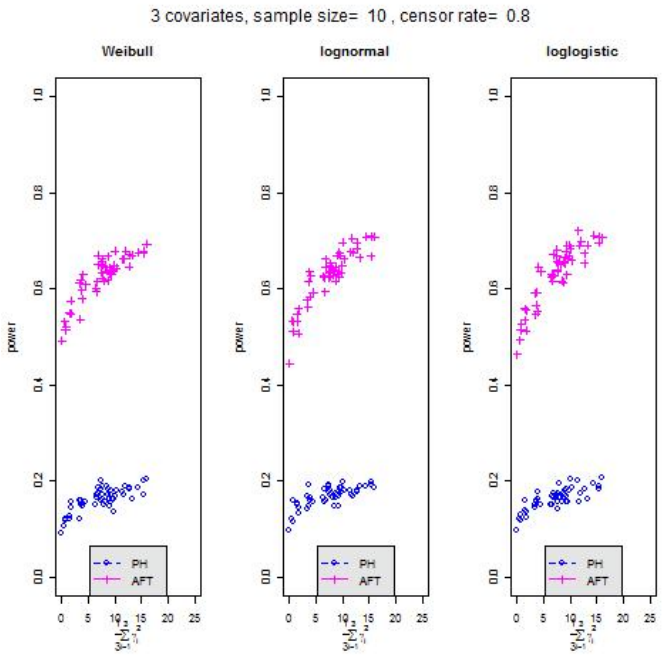
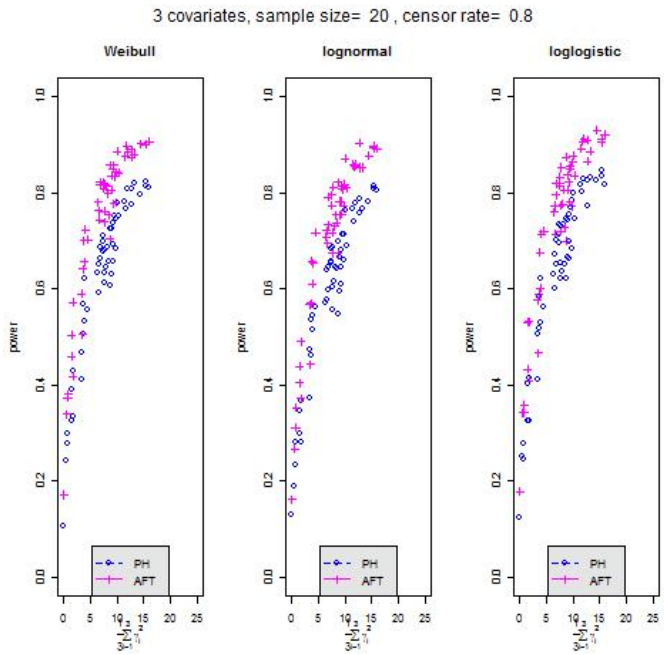


Figure B.2: Power plots for 3 covariate model

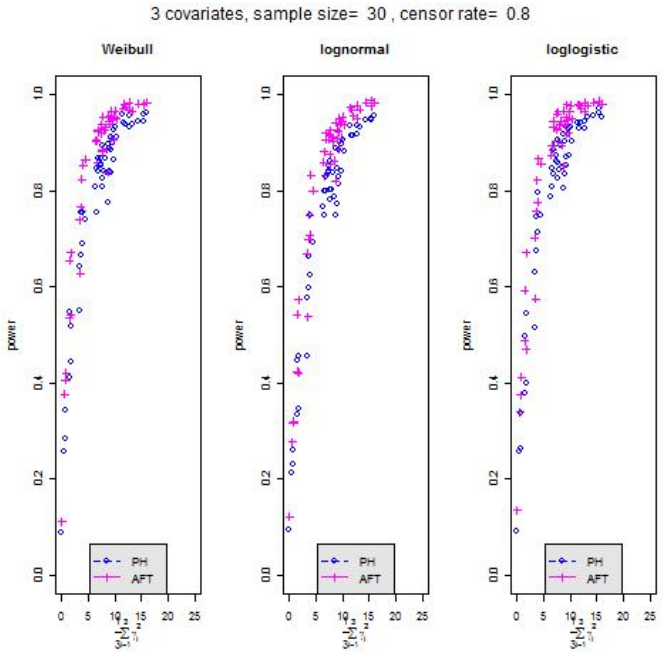
(q) n=10,p=0.8



(r) n=20,p=0.8



(s) n=30,p=0.8



(t) n=50,p=0.8