EFFECTS OF MODEL MISSPECIFICATION IN ESTIMATING COVARIATE EFFECTS IN SURVIVAL ANALYSIS FOR SMALL SAMPLE SIZES

Yi-Hwei Li
Department of Biostatistics
University of Pittsburgh

John P. Klein
Division of Biostatistics
The Medical College of Wisconsin

and

M. L. Moeschberger
Department of Preventive Medicine and Statistics
The Ohio State University

ABSTRACT

The results of a Monte Carlo study of the size and power of parametric and semi-parametric approaches to inference on covariate effects in survival (time-to-events) models in the presence of model misspecification and an independent censoring mechanism are reported. Basic models employed are a parametric model, where both a baseline distribution and the dependence structure of covariates on the failure times are fully specified (exponential, Weibull, log logistic, log normal, and normal regression models are studied), and a semi-parametric approach (due to Cox) in which the baseline distribution is unspecified. The Cox model performs very well compared to the parametric models for distributions with proportional hazard rates and appears to be robust with regard to the proportional hazard assumption. Appropriate parametric models have the potential of improving the size and power of the tests, although overall they are not appreciably better than the Cox model. In comparing the small sample performance of three statistical tests, the likelihood ratio and Wald tests closely agree with each other (likelihood ratio having a slight advantage), while the score test tends to perform much poorer since it inflates the size and power more than the other two.

KEY WORDS: Cox proportional hazards regression, parametric regression, censoring, small sample size power, robustness, model misspecification

1. INTRODUCTION

Of primary interest in scientific research is the investigation of the relationship between covariates, such as treatment or subject characteristics (possibly risk factors), and the time to occurrence of an event such as death, disease recurrence or cure. Through a regression model, in which the lifetimes have a distribution depending upon the covariates, we can examine the association between selected concomitant variables and the lifetimes. Two major approaches to the regression problem have been suggested. These are the parametric approach, where both a baseline distribution and the dependence structure of covariates on the failure time are fully specified, and the semi-parametric approach in which we construct a regression model without the specification of a baseline distribution. Both approaches are available in many standard statistical packages such as SAS. The objective of this study is to compare the performance of these two approaches for a variety of underlying distributions, censoring proportions, and sample sizes.

Although the large sample properties of the tests for covariate effects for the models discussed in this paper have been studied, their performance for small samples, as typically encountered in practice, has not been thoroughly investigated. In addition, the behavior of these statistics when the model is misspecified is not clear. Johnson et al. (1982) examined the small sample performances of the maximum partial likelihood estimators derived using the Cox model when there is Type II censoring and the covariate model is misspecified. Lagakos and Schoenfeld (1984), and Lagakos(1988) investigated the properties of proportional hazards score tests under misspecification of the functional form of the regression portion of the Cox model. Solomon (1984) and Struthers and Kalbfleisch (1986) studied the properties of an estimator based on a proportional hazards model when the true model is an accelerated failure time model. However, the size (or significance level) and the power of the likelihood ratio, score and Wald test statistics involving independent censoring under a variety of hazard distributions have not been evaluated for either the Cox or parametric models. Nor have the small sample size power of the statistical tests for the parametric regression models been examined when the model is incorrect. Lee, et al. (1983) evaluated the small sample performance of the three tests discussed in this paper, along with two other adjusted score statistics, under the Cox model in the presence of multiple covariates and censoring based on 1000 simulated samples of size n = 50 from the exponential distribution. Peace and Flora (1978) assessed size and

power of the likelihood ratio and Wald tests for the Cox, exponential, Weibull, and Gompertz models given that the model is correctly specified for samples of size n = 25, 50 and 100. Lininger, et al. (1979) studied the size and power of the Mantel-Haenszel test (with and without continuity correction), the generalized Gehan test, and the Cox likelihood ratio test under a variety of sample sizes and censoring schemes. In this report, a comparison of the small sample properties of the likelihood ratio, the score and the Wald statistics, as well as a comparison of the parametric and semi-parametric procedures as applied to hypothesis testing for covariates in survival models, will be made.

Section 2 introduces the estimation procedures employed in this study. The assumed models and underlying distributions for survival are described in Section 3. The simulation procedure is described in detail in Section 4. In Sections 5 and 6, we report the results of a Monte Carlo study of the size and power of the parametric and semi-parametric approaches to inference on covariate effects in the presence of model misspecification and an independent censoring mechanism.

In summary, our interest in this research has been focused on the following specific questions:

(i) How sensitive is the test of the null hypothesis of no covariate effect, based on parametric analyses, to model misspecification?

(ii) How robust is this test when made using Cox's proportional hazards model when the underlying distribution does not have proportional hazards?

(iii) Which test statistic, the likelihood ratio (LR), score (SC), or Wald (WA) statistic is more efficient?

(iv) What are the effects of censoring?

Finally, Section 7 includes a discussion and summary of our results.

2. PARAMETRIC AND SEMI-PARAMETRIC ESTIMATION PROCEDURES

The proportional hazards and accelerated failure time classes of regression models are commonly used in the analysis of survival data. For the proportional hazards model, the hazard rates of individuals with different explanatory variables are proportional to each other. Here there is a baseline hazard, $h_0(x)$, corresponding to a standard condition and the explanatory variables, **z**, act multiplicatively on the baseline hazard. That is, the effect of

the covariates is to increase or decrease the hazard. Thus, the proportional hazards model is defined by $h(x; z) = h_0(x) \exp(\mathbf{z})$. For the accelerated failure time model, the explanatory variables act multiplicatively on a baseline time X_0 so that their effect in this case is to accelerate or decelerate the failure time X, namely, $X = X_0\exp(\beta\mathbf{z})$, so that the log lifetime follows a linear model: $\log X = \beta \mathbf{z} + E$, where $E = \log X_0$ is an error random variable.

Both the proportional hazards and accelerated life models require two model assumptions, namely,

(1) a baseline distribution where the standard condition holds; and

(2) a functional form for the dependency of the lifetime on the covariates, often in terms of some parametric model.

In this paper, we examine two approaches to regression models: a fully parametric approach where those two parts are explicitly specified and a semi-parametric approach where only (2) is identified.

The standard approach to inference for parametric regression models is the maximum likelihood method. Here, if we observe a subject who failed at time t, then the contribution to the likelihood is $f(t ; \theta , \mathbf{z})$, the density function at t. The contribution from a subject censored at t is $S(t ; \theta , \mathbf{z})$, the probability of survival beyond t. Thus the full likelihood based on the data $(t_i , \delta_i , \mathbf{z}_i)$, i = 1,....,n, is (c.f. Lawless, 1982)

$$L(\theta) = \prod_{i=1}^{n} f(t_i ; \theta, \mathbf{z}_i)^{\delta_i} S(t_i ; \theta, \mathbf{z}_i)^{1-\delta_i}, \qquad (2.1)$$

where δ_i's are event indicator variables: $\delta_i = 1$ if the i^{th}

where $R(t_i)$ is the risk set at time t_i. The semi-parametric estimation techniques of this paper will be based on Cox's partial likelihood.

3. ASSUMED MODELS AND UNDERLYING DISTRIBUTIONS OF SURVIVAL.

In the simulation study, described in Section 4, we consider a single-covariate regression problem. We assume that the dichotomous covariate, z, equals 1 for treatment and -1 for control. We investigated five fully parametric regression models (the exponential, Weibull, log logistic, log normal, and normal) and Cox's semiparametric regression model because these models are widely applied in practice (partly because of their availability in SAS), often without checking the model assumptions. These models are the "assumed" models in the Monte Carlo experiments.

We briefly review these models.

(1) Cox proportional hazards model: $h(x; z) = h_0(x)\, e^{\beta z}$, $x \geq 0$,where $h_0(.)$ is any arbitrary nonnegative function.

(2) Exponential regression model: $h(x; z) = \lambda\, e^{\beta z}$, $\lambda > 0$, $x \geq 0$,where the baseline hazard function remains constant.

(3) Weibull regression model: $h(x; z) = \lambda \rho\, x^{\rho-1} e^{\beta z}$, $\lambda, \rho > 0$, $x \geq 0$.

The baseline hazard rate increases with time when $\rho > 1$ and decreases when $\rho < 1$.

Note that both the exponential and Weibull regression models can be written as accelerated failure time models with extreme value errors.

(4) Log logistic regression model: $Y = \log X = \mu + \beta z + \sigma E$, where E follows a standard logistic distribution. Here the baseline hazard function is monotone decreasing, for $\sigma \geq 1$, and, for $\sigma < 1$, increases initially to a maximum, then decreases to zero as time approaches infinity .

(5) Log normal regression model: $Y = \log X = \mu + \beta z + \sigma E$, where E has a standard normal distribution. For all log normal distributions, the hazard function is hump-shaped. That is, h(x) increases from 0 to a maximum, then declines to 0 monotonically as $x \to \infty$. However, for $\sigma \leq 0.2$, h(x) increases over most of the distribution. For $\sigma \geq 0.8$, h(x) decreases over most of the distribution.

(6) Normal regression model: $X = \mu + \beta z + \sigma E$, where E has a standard normal distribution, and the hazard function is strictly increasing. For this reason, the normal distribution may be appropriate for the life of products with wear-out types of failure.

Our primary concern is inference on the regression coefficient β. We consider the hypothesis tests for H_0: $\beta = 0$ based on the likelihood function (2.1) (or based on the partial likelihood (2.2) for the Cox model.) Three large sample test statistics widely used in applications are the score statistic (defined as the quadratic form based on the efficient score vector and the observed information matrix computed using the maximum likelihood estimates (mle) under the restricted model with $\beta = 0$), Wald's statistic (defined as the quadratic form based on the mle and the covariance matrix computed under the full model), and the likelihood ratio statistic (defined as twice the log of the ratio of the likelihood under the full model to that under the restricted model). (See Lawless (1982) for more complete and precise definitions of the test statistics.) Each of these statistics has an asymptotically chi squared distribution under H_o.

There is no absolute criterion to determine which of these procedures gains superiority over the others. However, many authors [e.g., Kalbfleisch and Prentice (1980) and Lawless (1982)] have pointed that the distribution of the likelihood ratio statistic often appears to reach its limiting distribution more rapidly than the other two statistics. In addition, the likelihood ratio statistic has the invariance property that the test result is independent of the choice of parameterization.

Generally speaking, the likelihood ratio statistic is more efficient than the others if the model is correctly specified. Nevertheless, the properties of these statistics are still not very clear if the model is misspecified or if the sample size is small.

For the Cox model, we use the bisection method to solve the score equation for β and perform the likelihood ratio (LR), score (SC), and Wald (WA) tests of H_0: $\beta = 0$ based on the partial likelihood (2.2). For the exponential model where there exists a closed form solution for the nuisance parameter, we use the bisection method to search for the root of β. We applied the Marquart's method (David and Moeschberger, 1978) to solve the score equations for the Weibull and log logistic models. We use the EM algorithm procedure (Lawless, 1982) to search for the maximum likelihood estimates for the log normal and normal models.

power, we choose alternatives β = 0.20, 0.35, 0.55 and 0.70. Each power estimate is based on 1000 simulated samples.

We suppose each individual has a lifetime X and a censoring time L, where X and L are independent continuous random variables. We assume that X follows one of the underlying distributions, and L has an exponential distribution with parameter λ_p. For a given censoring percentage P and failure time distribution with density, $f_x(x)$, we solved iteratively for λ_p in $P = 100\,[\,Pr(\,L < x)\,] = 100 \int_0^\infty [\,1 - e^{-\lambda x}\,]\, f_x(x)\, dx$. (See Li (1991) for more detail on the numerical value of λ based on different failure time distributions and 20 and 40 percent censoring.)

The failure and the censoring times (except for the normal, log normal and gamma variates) were generated from the uniform 0-1 congruential generator function RANF available on the Cray computers. For each subject we generated two independent variates X and L_p. The min (X, L_p) represents the observed time of that subject corresponding to an expected percentage censorship, P.

In the Monte Carlo studies, we use RNNOA routine in IMSL, where an acceptance and rejection technique (Kinderman and Ramage, 1976) is used, to generate pseudo random numbers from a standard normal distribution. RNGAM routine is called to generate variates from Gamma (κ) in which squared and halved normal deviates are used for κ = 0.5 and a ten-region rejection procedure developed by Schmeiser and Lal (1980) is used for κ = 2. We apply the general inverse-transform method to simulate the Weibull, log logistic, log Laplace, Gompertz and Smith-Bain random variables from the RANF U(0,1) random number generator on the Cray Y-MP computer. A more complete description of the algorithms used to generate the underlying distributions may be found in Li (1991).

5. SIZE CONSIDERATIONS.

We first investigate the hypothesis-testing properties by means of a 'size' comparison. With the assumption of asymptotic normality and a nominal 5% significance level, the 'size' in the Monte Carlo simulation is defined as the fraction of replications for which the chi-squared test statistic (LR, SC or WA) is greater than 3.841. We compare the precision of estimates for the six models and the three types of tests by counting the number of times

that the observed sizes fall above or below the symmetric 99.88% probability interval (0.040, 0.060) .

The Cox model appears to maintain the nominal significance level well. In only 2 out of 114 independent experiments utilizing the partial likelihood ratio test and the Wald test, the observed sizes fall above the 99.88% probability interval, and no observation falls below the interval in our experiments. The Scores test does poorer, relative to the partial likelihood ratio test and the Wald test, in that in 12 out of 114 independent experiments the observed sizes fall above the probability interval. For other models, excluding the exponential, all three tests tend to overestimate size with the scores test overestimating more severely than the likelihood ratio or the Wald tests.

Here, due to space consideration, we present only partial simulation results (see TABLE 5.1) for selected underlying distributions representative of a variety of hazard rates with n = 40. More complete results may be found in Li (1991) and Li, et al (1993).

[Table 5.1 here]

More specifically, the results for selected underlying distributions categorized by type of hazard rate, sample size = 40, and censoring fractions of 0%, 20%, and 40% are presented for the likelihood ratio, scores, and Wald tests for the six assumed models in Table 5.1.

Overall, one can easily see that the exponential model performs poorly for all tests and all underlying distributions except for the case of no misspecification, namely, in the case of constant hazard rate. Further examination of the specific models reveals that the exponential model substantially overestimates the nominal 5% level when the hazard rate is decreasing or bathtub shaped and substantially underestimates the 5% level when the hazard rate is increasing. If the hazard rate is hump-shaped, then the exponential both

the symmetric 95% probability interval (0.044, 0.056) and then performing a logistic analysis to assess whether the differences we observe may be explained by chance.

The three types of test perform differently ($p < 0.0001$) with the likelihood ratio performing best, the Wald next best, and the scores test performing much poorer than either of the other two. In fact, when the score test was excluded, the remaining two types of tests interacted with the models ($p=0.015$). The Wald and likelihood ratio tests are similar for the Cox and Weibull models but the Wald test performs poorer, on the whole, than the likelihood ratio test for the log logistic, log normal, and normal models. Sample size appears to be a large factor in the relative performance of the three tests with the Wald test being closer to the likelihood ratio test for samples of size 100. This property agrees with Peace and Flora (1978) who investigated the effect of sample size on size of the test when the underlying distributions were correctly specified as exponential, Weibull, or Gompertz.

The five models perform differently ($p < 0.001$) with the Cox model outperforming the other four. The Weibull model performs below expectations in that it overestimates the size badly for the log logistic, log normal, and log Laplace underlying distributions. The log logistic model does slightly better than the Weibull model, though not significantly better. Overall, both the log logistic and Weibull models do significantly better than the log normal and normal models. Of course, in the case of no misspecification of the model, each correctly specified model performs well.

An interesting observation, previously made by Lee, et al (1983) and confirmed in our study, is that, for the Cox model, the censoring fraction does not seem to be very strongly related to the performance of the various types of tests or to the performance of the different models, although, for fixed sample size, the observed sizes for 40% censoring are less than those for 0% censoring and the sizes at 20% censoring compared to those at 0% (no censoring) are not different. With the score test excluded, there is no significant difference in censoring fractions ($p=0.19$) when adjustment is made for type of test.

Examining, in a descriptive way, the performance of the various tests and models more closely for the various types of hazard functions is a bit more illuminating. For constant or decreasing hazard rates, the likelihood ratio statistic mirrors the above general statements and outperforms the other two. The Wald statistic does a bit poorer for $n = 40$ but is as good as the likelihood ratio test for $n = 100$, and the score statistic performs by far the

poorest. As noted earlier, the exponential only does well for a constant hazard rate and the log logistic and log normal generally outperform the normal and Weibull models, although not always very convincingly. All parametric models are inferior, on the whole, to the Cox model with regard to model misspecification.

For increasing hazard rates, the general statement about the test statistics is as before. Here the parametric models do substantially poorer than Cox for n = 40. The Weibull, normal, and log normal do improve for n = 100 using the likelihood ratio or Wald tests.

For bathtub-shaped hazard rates, Cox is again the best with the log normal, log logistic, and Weibull showing some promise, in that order, provided one uses the likelihood ratio test.

For hump-shaped hazard rates, the parametric models again do substantially poorer than Cox for both n = 40 and n = 100 . Again, the exponential and Weibull perform very poorly with the log logistic, log normal, and normal showing varying degrees of promise.

The Weibull model tends to overestimate the nominal level when the hazard rate is decreasing or hump-shaped. It tends to underestimate when the hazard rate is bathtub shaped. Not surprisingly, it has the most success when the hazard rate is increasing.

The log logistic and log normal tend to overestimate the 5% level for n = 40, but improve substantially as sample size increases to n = 100.

The normal model fluctuates considerably in estimating the size of the test for decreasing and hump-shaped hazard rates and tends to overestimate size for increasing, constant, and bath-tub shaped hazard rates.

6. POWER CONSIDERATIONS

We examined the powers of the tests for the assumed models against the alternative values of beta = .20, .35, .55 and .70 by computing the fraction of replications for which the test statistic is larger than 3.841, i.e., the nominal level of significance or size will be nearly .05 when the model is correct. If the underlying distributions are correctly specified, then the parametric models perform slightly better than the Cox model (a point already noted for the likelihood ratio test and Wald test by Peace and Flora (1978) for the exponential, Weibull, and Gompertz models). For fixed models and underlying distributions, the powers of the tests with increased censoring are consistently smaller than those without censoring but the relative order of the power of the assumed models does not

ratio, the score and the Wald tests based on these models. In addition, we have examined the effects of an independent censoring mechanism on these inference procedures.

The Cox proportional hazards regression model in a hypothesis-testing framework appears to be robust with respect to the proportional hazards assumption. Although, the Cox model will reflect the relative importance of the covariate effect when the underlying distribution does not possess proportional hazards, power can be improved by using the appropriate underlying parametric distribution as can be seen in Figure 3. Indeed, the appropriate parametric models have the potential of improving the precision and power of

In summary, the results of this study suggest that, if interest centers upon making an inference on a covariate effect, the Cox model may be used with confidence. The likelihood ratio test is slightly better than the Wald test and the score test should rarely be used. If parametric models are used, the prior knowledge of the hazard rate would suggest the model to be used (most likely, the log logistic or log normal model). In rare instances the Weibull might be used and the exponential would never be used.

ACKNOWLEDGMENT

This research was supported by Grant 1 R01 CA54706-01 from the National Cancer Institute and by a Grant from the Ohio Super Computer Center.

REFERENCES

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* **B34**, 187-220.

David, H. A. and Moeschberger, M. L. (1978). *The theory of competing risks*. griffin no. 39, London.

Elandt-Johnson, R. C. and Johnson, N. L. (1980). *Survival models and data analysis*. Wiley, New York.

Johnson, M. E., Tolley, H. D., Bryson, M. C. and Goldman, A. S. (1982). Covariate analysis of survival data: A small-sample study of Cox's model. *Biometrics* **38**, 685-698.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Kinderman, A. J. and Ramage, J. G. (1976). Computer generation of normal random variables. *Journal of the American Statistical Association* **71**, 893-896.

Klein, J. P. and Moeschberger, M.L. (1989). The robustness of several estimators of the survivorship function with randomly censored data. *Communications in Statistics* **18(3)**, 1087-1112.

Lagakos, S. and Schoenfeld, D. (1984). Properties of proportional hazards score tests under misspecified regression models. *Biometrics* **40**, 1037-1048.

Lagakos, S. (1988). The loss in efficiency from misspecifying covariates in proportional hazards regression models. *Biometrika* **75**, 156-160.

Lawless, J. F. (1982). *Statistical models and methods for lifetime data*. Wiley, New York.

Lee, L.L., Harrell, F.E.,Jr., Tolley, H.D., and Rosati, R.A. (1983). A comparison of test statistics for assessing the effects of concomitant variables in survival analysis. *Biometrics* **39**, 341-350.

Li, Y.-H. (1991). *Regression analysis of failure time data*. Ph.D. Dissertation at The Ohio State University, Columbus, Ohio.

Li, Y.-H., Klein, J.P., and Moeschberger, M.L. (1993). Effects of model misspecification in estimating covariate effects in survival analysis for small sample sizes. Technical Report Ns0E511, Department of Statistics, The Ohio State University, Columbus, OH.

7ezberger, M.L. (1993).,fe6 0ce02 Tcl an4 TD 71001 Tc 0.006 Tw*Biometrics 1Tj -3.4735 -24iation*

Lininger, L., Gail, M.H., Green, S.B., and Byar, D.P. (1979). Comparison of four tests for equality of survival curves in the presence of stratification and censoring. *Biometrika* **66**, 419-428.

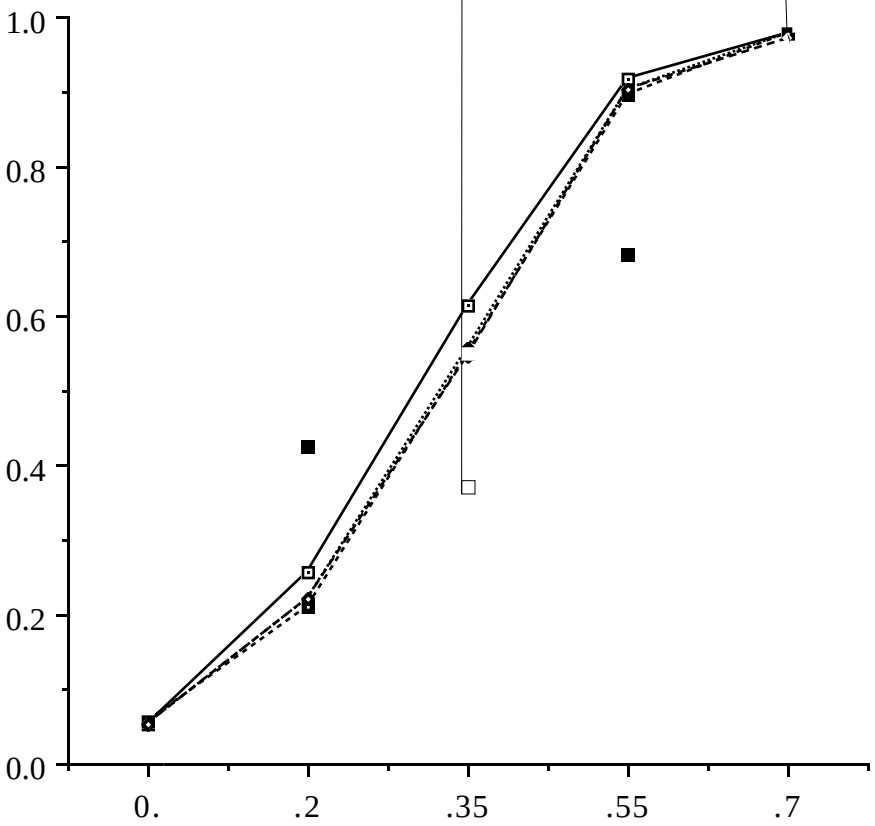
Peace, D.E. and Flora, R.E. (1978). Size and power assessments of tests of hypotheses on survival parameters. *Journal of the American Statistical Association* **73**, 129-132.

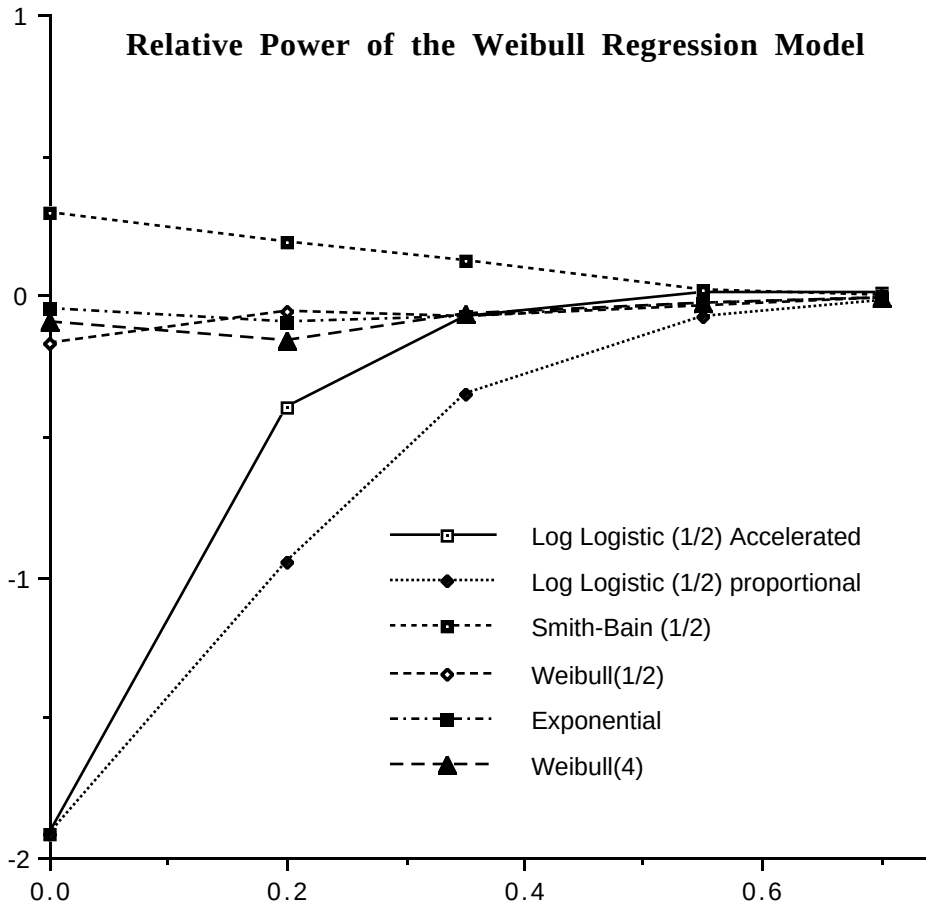
Schmeiser, B. W. and Lal, R. (1980). Squeeze methods for generating gamma variates, *Journal of the American Statistical Association* **75**, 679-682.

Smith, R. M. and Bain, L. J. (1975). An exponential power life-testing distribution. *Communications in Statistics* **4(5)**, 469-481.

Solomon, P. J. (1984). Effects of misspecification of regression models in the analysis of survival data. *Biometrika* **71**, 291-298. Amendment (1986), **73**, 245.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73**, 363-369.


1.0
0.8
0.6
0.4
0.2
0.0
0.
.2
.35
.55
.7


Relative Power of the Weibull Regression Model
1
0
-1
-2
0.0
0.2
0.4
0.6
Log Logistic (1/2) Accelerated
Log Logistic (1/2) proportional
Smith-Bain (1/2)
Weibull(1/2)
Exponential
Weibull(4)

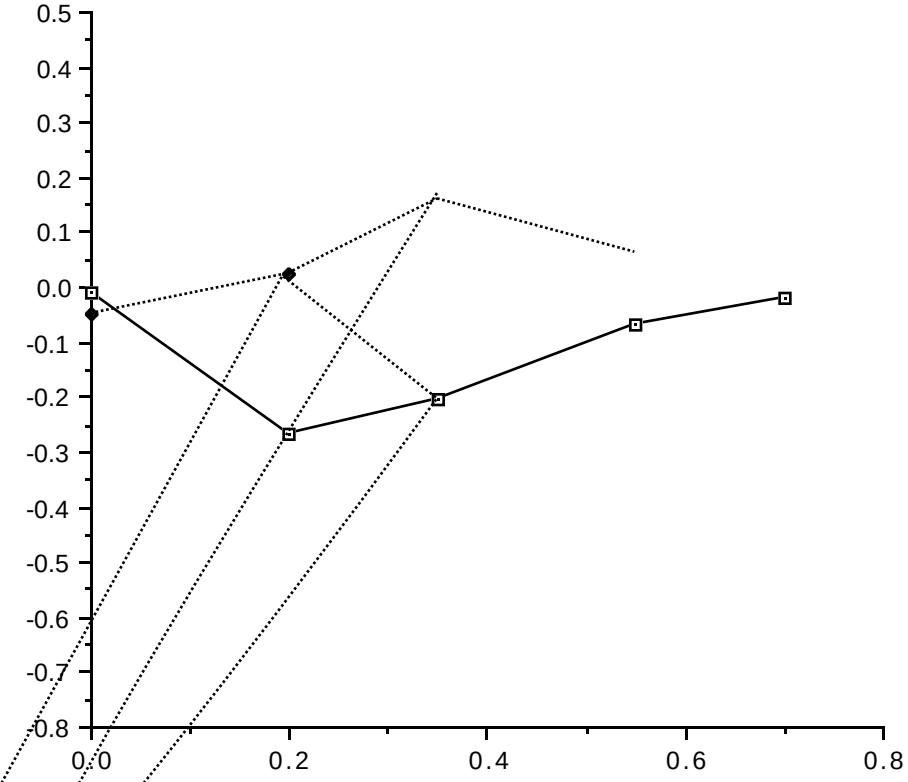
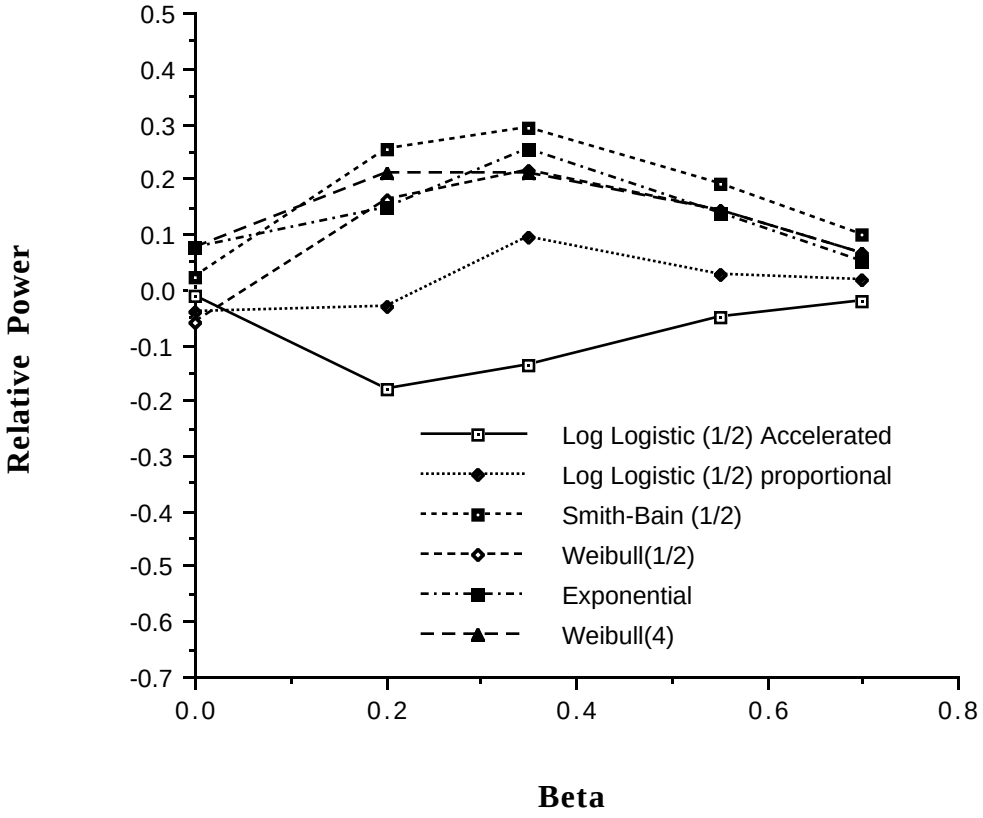
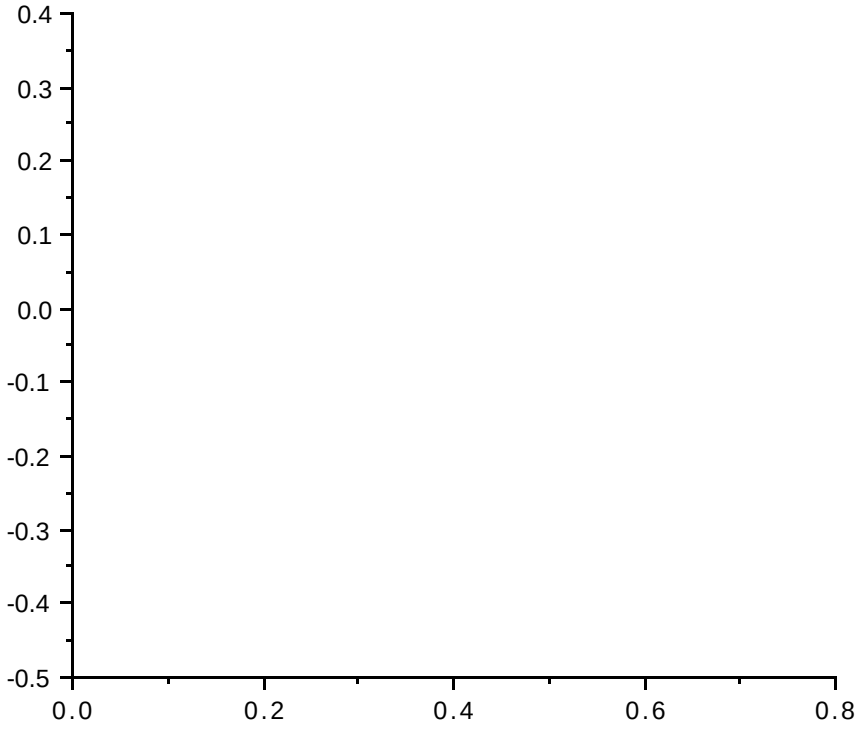

0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
0.0
0.2
0.4
0.6
0.8

Figure 4

Relative Power of the Log Normal Regression Model




0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
0.0
0.2
0.4
0.6
0.8