EVALUATION OF METHODS FOR SURVIVAL ANALYSIS IN THE PRESENCE OF EXTREMELY FEW EVENTS PER VARIABLE

By

Austin Marcus Lanser

Thesis
Submitted to the Faculty of the
Vanderbilt University Department of Biostatistics
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE
in

BIOSTATISTICS

May, 2015

Nashville, TN

Approved:
Qingxia Chen, Ph.D.
Dandan Liu, Ph.D.

ACKNOWLEDGMENTS

I would like to thank the people that made this work possible. First and foremost I need to thank my advisor, Dr. Cindy Chen, for her patience and guidance. Likewise, I must thank Dr. Dandan Liu for being part of my committee. Furthermore, I thank the crew of ACCRE for education about and use of their computing cluster. I could not have done this without the sense of community and support provided by my cohort: Christina Tripp, David Schlueter, Derek Smith, Mark Giganti, and Minchun Zhou. Finally, I want to acknowledge the ones who have been there from the very beginning, my parents, Melissa and Peter Lanser.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

CI Confidence Interval

CoxPH Cox Proportional Hazards

EPV Events Per Variable

ESE Empirical Standard Error

ICD International Classification of Diseases

IPTW Inverse Probability of Treatment Weighting

LASSO Least Absolute Shrinkage and Selection Operator

LASSONoPen Least Absolute Shrinkage and Selection Operator with No Penalty on Exposure

MAD Median Absolute Deviation

MAE Median Absolute Error

MSE Mean Square Error

PMLE Penalized Maximum Likelihood Estimation

PS Propenisty Score

PSA Propensity Score Adjustment

PSH Propensity Score Adjustment with additional Heterogeneity adjustment

RCS Restricted Cubic Spline

RidgeNoPen Ridge regression with No Penalty on Exposure

CHAPTER I

Introduction

I.1 Basic Review of Survival Analysis

The idea of studying and modeling the time until some event occurs is not a new concept. The earliest known example of survival analysis is John Graunts 1662 life table, or mortality table [19]. Graunt's life table, though novel, was very basic. Major developments in survival analysis (and statistics in general) did not occur until the early 20th century when World War II drove investigations of human mortality and military equipment failure. The term 'survival analysis' is usually used in a biomedical context and is referred to by other names in different fields including 'event-history analysis' in sociology, 'reliability analysis' in engineering, 'duration/duration-time analysis' in economics, and 'time-to-event analysis' more generally. Some common examples of events studied in biostatistics are time until death, time until the development of infection or disease, time until appearance of cancer, or time until remission after treatment. Some additional examples in other fields are time until equipment failure, time until an earthquake, time until dropping out of school, or time until policy adoption. The rest of this section will be dedicated to a brief review of survival analysis, the problem to be investigated, and the motivating example for this paper.

A defining feature of survival analysis is the potential to have censored event times, that is, for some events the time is known exactly, but for others the time is only known to be within some interval. One might wonder why not simply exclude censored data and approximate using some standard regression technique. The problem with that approach is that there is information to be gained by the fact that a subject has gone some amount of time without experiencing the event. Among other problems, analysis using only subjects with observed events could produce very biased inferences. This problem is easily demonstrated when considering the study of a relatively rare event. For example, suppose for a 10 year study of a randomly sampled cohort of 1000 Americans, ages 18-25 years, death is the outcome of interest and 50 die. The analysis using only the 50 deceased would produce a result indicating the expected lifetime of Americans from that age group is less than 35 years (25+10), which is far less than reality. By investigating the deceased, one is essentially doing a subgroup analysis of those with a predisposition to die young. In order to research the broader population, analysis must be conducted including censored information.

The three main categories of censoring are left, interval, and right censoring. For the purposes of this paper, censoring may be best explained through another example. Consider a study for which subjects have appointments every three months for a period of two years, and the outcome of interest is time until infection by some disease. If a subject is diagnosed with the disease of interest at the first appointment then they are known to have been infected sometime before then, but not exactly when, and the outcome is considered left censored. Interval censoring occurs when a subject is diagnosed on or after the second appointment because we can only say the subject was infected sometime since the previous appointment. Finally, if the study ends and after two years a subject has not been diagnosed with the disease, the subject is considered right censored, because it is possible the subject becomes infected in the unobserved future. Another example of right censoring is when a subject is lost to followup or otherwise drops out of a study [13].

Examples of methods to model survival data include parametric, nonparametric and semiparametric approaches. Parametric models assume the distribution of survival times follow some known distribution chosen by the investigator. A few common choices of distribution are exponential, Weibull and log-normal distributions. Using the assumed distribution, parameters are then estimated using maximum likelihood estimation [13, p. 395]. In contrast to parametric models, nonparametric methods do not assume a specific distribution. A popular example of a nonparametric method is the Kaplan-Meier method (i.e. Product Limit Estimation). The foundation of the Kaplan-Meier method is an estimator for survival at time t defined as $\hat{S}(t) = \prod_{t_i \leq t}[1 - \frac{d_i}{R_i}]$ where d_i is the number of events at time t_i, and R_i is the number of subjects at risk prior to t_i [13, p. 92]. In general, when the correct distribution is specified, parametric models tend to produce more precise estimation than nonparametric methods; however, when the specified distribution does not fit the data, one essentially has a consistent estimate of the wrong quantity, revealing the advantageous flexibility of nonparametric modeling [13, p. 393]. A major downside to the Kaplan-Meier approach is the inability to directly adjust for confounding. Semiparamtric modeling serves as a sort of compromise between the other two approaches. The most common example of a semiparametric survival method is Cox proportional hazards regression (CoxPH) and is the primary method of interest for this paper [5]. Unlike the Kaplan-Meier method, one can adjust for confounding using CoxPH. Also, CoxPH models require fewer assumptions about the underlying distribution than parametric models do and the risks of mispecification are mitigated. CoxPH modelling has two key assumptions. The first, is proportional hazards. The hazard funtions for two stratified groups of the data must be proportional. The second assumption for CoxPH modelling is non-informative censoring. The mechanism causing the censoring of individuals must not be related to the chance of an event occuring.

I.2 The Problem

Since the outcome of interest is the time it takes for events to occur, it is generally agreed that the more of those events in the sample being studied, the stronger the inference that can be drawn. Rules of thumb vary in the literature for regression modelling, but generally at least 10 to 15 events per variable (EPV) minimum are required to produce fairly reliable inference in terms of accuracy and precision of estimates [9][10][15]. Sometimes, however, even in well-designed studies, events occur very rarely, and despite the presence of a large total sample, a low EPV ratio damages the accuracy and precision of a regression model's coeffiecient estimates [15]. Some scenarios in which we expect a low EPV ratio are when the number of confounders is unavoidably large, the subject of study is a rare disease, or subgroup analysis is of great interest. In these cases investigators must make due with few EPV.

I.3 Motivating Scenario

The objective of a study by Xu et al. [23] was to use electronic health record data to confirm the association of a recently repurposed diabetes drug, metformin, with reduced cancer mortality. Their study sample consisted of 32,415 records of cancer patients extracted using the Vanderbilt tumor registry, and a separate study sample of 79,258 records from the Mayo Clinic. Xu et al. used stratified CoxPH models to estimate hazard ratios for all-cause mortality adjusting for age at diagnosis, sex, race, body mass index, tobacco use, insulin use, cancer type, tumor stage (4 levels for tumor stage: 0,1, 2 or 3 combined, and 4), and non-cancer Charlson Comorbidity Index (CCI). Age, BMI, and CCI were modeled as restricted cubic spline functions with four knots. The CCI was developed by Charlson et al. [3], and is an index that categorizes comorbidities of subjects based on International Classification of Diseases (ICD) codes commonly used in medical center administrative data. Each diagnosis category within the CCI has assigned weights determined by the adjusted risk of mortality or treatment resource use. The sum of those weights is the CCI. The higher the CCI, the higher likelihood the predicted outcome will result in mortality or higher treatment resource use. The advantage of using the CCI is the reduction of thousands of ICD codes to one index that has been validated for use in regression [6]. The samples for this study were large enough to observe statistically significant results with regard to the four most common forms of cancer (colorectal, breast, lung and prostate). Xu et al., however, state that they were unable to stratify by histologic subtype within each cancer type due to small sample sizes within each cancer. Among other suggestions for future research, Xu et al. called for studies with greater statistical power in order to evaluate these less frequently observed cancers. The obvious answer is to get more data, but less obvious is how to obtain that data. Xu et al. use multi-cite electronic health records and applied advanced informatics methods to get as much out of their data as possible. The usual limitations of time and money are also factors that prevent more accumulation of some types of data. Sometimes, even in the age of big data, a researcher must make due with what little there is. This concession is the motivation for evaluating methods on survival data with few events.

The potential lack of validity of a standard CoxPH model means exploring the implementation of additional methods to control for confounding. Some broad categories of analysis to be considered are propensity score methods (PS) and penalized maximum likelihood methods (i.e., shrinkage methods). This paper considers a couple variations of each set of methods: propensity score adjustment (PSA) [1], propensity score adjustment with additional heterogeneity adjustment (PSH) [1], ridge regression [22], ridge regression with no penalty on the exposure variable (RidgeNoPen) [4], least absolute shrinkage and selection operator (LASSO) [21], LASSO with no penalty on the exposure variable (LASSONoPen)[4]. In addition, the CoxPH model unadjusted for any covariates will be considered for the purpose of comparison with the other models. Further discussion of these methods can be found later in the Methods section. The goal of this simulation study is to use Monte Carlo simulations to evaluate and compare the performance of the traditional CoxPH, propensity score adjustment and shrinkage regression methods when there are few events compared to the number of variables.

CHAPTER II

Methods

II.1 Regression Methods

II.1.1 Regular Cox Proportional Hazards Model (CoxPH)

Survival data for the j-th subject consists of three parts: the possibly censored time on study (T_j), whether or not an event occured during study (δ_j), and the vector of measured covariates including the exposure of interest (Z_j for the exposure of interest and $\mathbf{X_j}$ for the remaining covariates). For this paper, the exposure is a binary variable repressenting treatment status ($Z_j = 1$ for treatment group, $Z_j = 0$ for control group). The time on study is either an event time or right censored ($\delta_j = 1$ for event or $\delta_j = 0$ for censored). Let $h(t|Z_j, \mathbf{X_j})$ denote the hazard rate at time t with treatment Z_j and covariate vector $\mathbf{X_j}$. The hazard funtion specified by the Cox PH model is

$$h(t|Z_j, \mathbf{X}) = h_0(t_j) exp(\beta_Z Z_j + \boldsymbol{\beta}'\mathbf{X_j}),$$

where $h_0(t_j)$ is the baseline hazard, and $\boldsymbol{\beta}$ is a vector of regression coefficients corresponding to the covariates in $\mathbf{X_j}$. Note that $\boldsymbol{\beta}'$ is the transpose of the vector $\boldsymbol{\beta}$. The term containing a linear combination of the covariates ($\beta_Z Z_j + \boldsymbol{\beta}'\mathbf{X_j}$) (equivalently written as $\beta_Z Z_j + \beta_1 X_{j,1} + \beta_2 X_{j,2} + ... + \beta_p X_{j,p}$ for p covariates) will occasionally be referred to as the 'linear predictor' of the regression model.

Cox PH is a semiparametric method and assumes a parametric form for the covariate effects, but models the baseline hazard nonparametrically. The assumed form for the baseline hazard is multiplicative. The reason for the label 'proportional hazards' can be briefly demonstrated by considering the following hazard ratios of two individuals, one with exposure Z and covariate vector $\mathbf{X}$ and the other with Z^* and $\mathbf{X}^*$:

$$\frac{h(t|Z, \mathbf{X})}{h(t|Z^*, \mathbf{X}^*)} = \frac{h_0(t) exp[\beta_Z Z + \sum_{k=1}^{p} \beta_k X_k]}{h_0(t) exp[\beta_Z Z^* + \sum_{k=1}^{p} \beta_k X_k^*]} = exp[\beta_Z (Z - Z^*) + \sum_{k=1}^{p} \beta_k (X_k - X_k^*)]$$

The hazard ratio is a constant with respect to t, so the hazard rates are proportional. The main benefit of the proportional hazards assumption is that regression coefficients can be estimated without specifying the baseline hazard. Coefficient estimation for the Cox model is done by maximizing the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{d \in D} \frac{exp(\beta_Z Z_d + \boldsymbol{\beta}'\mathbf{X_d})}{\sum_{r \in R} exp(\beta_Z Z_r + \boldsymbol{\beta}'\mathbf{X_r})} \tag{II.1}$$

where D is the subset of subjects for which events were observed, and R is the subset of subjects still at risk.

II.1.2 Unadjusted CoxPH

The unadjusted CoxPH model is simply the Regular CoxPH with no attempt made to adjust for covariates. To be precise, that means the only regression coefficient is β_Z making the Unadjusted CoxPH model

$$h(t|\mathbf{X}) = h_0(t_j)exp(\beta_Z Z_j)$$

II.1.3 Propensity Score Adjustment (PSA)

A propensity score (PS) is the conditional probability of a subject receiving the binary exposure of interest given the set of confounders for that subject [17]. Whenever treatment groups are not randomly assigned, it is possible that there is some amount of confounding between treatment assignment and patient characteristics. That confounding is one motivation for using PS methods as researchers can utilize them to control for systematic differences between treatment groups that are predictable from observed characteristics and data [1]. As outlined by P. Austin [1], there are four common options for PS methods to implement in regression modeling: stratification, matching, inverse probability weighting, and adjustment.

To use the first option, ‘stratification,’ we partition the entire data set according to estimated PS. There are several variations one can use to stratify, but one common scheme is to divide the data into five even groups based on the quintiles of the PS [1]. Once stratified, one can then adjust for the stratified group by adjusting for the strata with a categorical variable or by modeling each stratum separately.

The second option, ‘matching,’ is used by forming matched sets of treatment groups based on estimated PS. For example, consider a subject from the treatment group in a study that has two groups, treatment and control. Now, for a one-to-one matching scheme, this subject is paired with a subject from the control group that is within a predetermined range of the first subject’s propensity score. Once the groups are made, one can model the outcome similar to the stratification method by either adjusting for the matched group using a categorical variable, or modeling the groups separately.

The third approach is to use PS for inverse probability of treatment weighting (IPTW). The weight used in IPTW for a subject is a function of exposure status and the estimated PS [14]:

$$(Z/PS) + [(1-Z)/(1-PS)]$$

Here, subjects who are less likely to be in the treatment group receive more weight so they are not under represented in the model. Using the result to weight the sample creates a new sample that no longer is confounded with treatment assignment [1]. This sample is then used to model the outcome as usual.

The final option is PSA where one directly adjusts for the PS in the outcome model. This last approach is the one considered for the simulation study. To be precise, for this paper the PS is adjusted using a restricted cubic spline (RCS) with four knots and no other covariates other than the exposure of interest. With the RCS, the linear predictor for PS adjustment is

$$\beta_Z Z + \beta_1 X_{PS} + \beta_2 (X_{PS} - x_1)^3 + \beta_3 (X_{PS} - x_2)^3 ... \beta_5 (X_{PS} - x_4)^3,$$

where x_1 to x_4 are the knots, which are selected using quantiles of the PS [9, p. 27].

The model used to estimate the PS for subject j is a logistic regression model

$$logit(P(Z_j = 1|\mathbf{X_j})) = \alpha_0 + \alpha_1 X_{j,1} + \alpha_2 X_{j,2} + ... + \alpha_p X_{j,p},$$

where the $\boldsymbol{\alpha}$s are regression coefficients corresponding to $\mathbf{X_j}$ and p is the number of measured covariates.

II.1.4 Propensity Score Adjustment with Additional Heterogeneity Adjustment (PSH)

Two kinds of effects are commonly seen in regression modelling; marginal effects and conditional effects [8]. A marginal effect is the effect of a covariate on an outcome given all other covariates are held constant. Another way to think about marginal effects is as population level effects on the outcome after changing one covariate value to another for the whole population. A conditional effect is the average effect of a covariate on an individual subject's outcome given that subjects covariates. Collapsibility is a property of some regression models that marginal and conditional treatment effects will be the same. Some models produce treatment effect estimates that are collapsible, such as linear treatment effects [1]. Unfortunately, survival outcomes (i.e. hazard ratio) are not collapsible [7]. P. Austin [1] points out that no variation of Cox model using the PS that was used estimated marginal treatment effects, but instead produced conditional effect estimates. One result of non-collapsibility is that effect estimates will be biased [16], and this has been confirmed in simulations for Cox models using PS [1]. Included in this paper is a method that should avoid these problems of non-collapsibility; PSH.

For PSH [4], we take PSA and include additional confounding variables to the model. Specifically, for this study, three confounders with relatively large effect sizes were chosen to be included. By including confounders in addition to the propensity score adjustment, heterogeneity in the population is accounted for, which relieves the issue of non-collapsibility of the hazard ratio [4]. The linear predictor used in the model for PSH is

$$\beta_Z Z + \beta_1 X_{PS} + \beta_2 (X_{PS} - t_1)^3 + ... + \beta_5 (X_{PS} - t_4)^3 + \eta_1 X_{H1} + \eta_2 X_{H2} + \eta_3 X_{H3},$$

where η_1,η_2, and η_3 are regression coefficients for the three confounders chosen for the heterogeneity adjustment.

II.1.5 Ridge

The remaining four methods fall under the category of penalized maximum likelihood estimation (PMLE) [22]. PMLE methods are also referred to as shrinkage methods because their defining trait is the reduction in magnitude of coefficients based on some criterion. The effect of this shrinkage is regression coefficients that are biased towards the null, but are also lower in mean squared error and useful in avoiding overfitting [9]. PMLE can be a compromise when choosing between a parsimonious model and a more complex model that might fit the data better. PMLE solves the dilemma by providing a well-defined, often unique, and reproducible process to shrink coefficient estimates. Many different approaches for PMLE exist and are often defined based on variations in penalization criteria.

Ridge regression (i.e. quadratic PMLE, or Tikhonov regularization) was first introduced for linear models by Hoerl and Kennard [12]. Ridge regression coefficients are subject to the constraint of the L2-norm:

$$\beta_Z^2 + \sum_{j=1}^{p} \beta_j^2 \leq s$$

where p is the number of model parameters, and s is a constant selected using cross-validation to minimize a user defined loss function [21][20]. Given $\boldsymbol{\beta}$ that maximizes the Cox model's partial likelihood (Eq: II.1), if $s \geq \beta_Z^2 + \sum_{j=1}^{p} \beta_j^2$ then the Ridge coefficient estimates are the usual partial likelihood estimates produced by CoxPH, but if $s < \beta_Z^2 + \sum_{j=1}^{p} \beta_j^2$ then the coefficient estimates are scaled towards zero. Though coefficient estimates are scaled toward zero, they will never be exactly zero (as they might be in LASSO).

II.1.6 Ridge Without Penalizing Exposure of Interest (RidgeNoPen)

RidgeNoPen is the same as Ridge except the coefficient for the exposure of interest is not penalized. So, if β_Z is the regression coefficient for the exposure of interest, the constraint for RidgeNoPen is revised to be:

$$\sum_{j=1}^{p} \beta_j^2 \leq s$$

The logic behind not penalizing the exposure of interest is while we do not want to overfit the model, we also do not want to bias the estimated effect for the exposure of interest [4]. Penalizing other covariates should still help avoid overfitting and reduce the MSE for the model's coefficient estimates.

II.1.7 LASSO (Least Absolute Shrinkage and Selection Operator)

LASSO is another shrinkage method. The main difference between Ridge and LASSO is the penalty. The penalty for LASSO is called the L1-norm penalty and is expressed as

$$|\beta_Z| + \sum_{j=1}^{p} |\beta_j| \leq s$$

where p, β_j and s are the same as defined for Ridge. Similar to use of the Ridge penalty, given $\boldsymbol{\beta}$ that maximizes the partial likelihood (Eq.: II.1), if $s \geq |\beta_Z| + \sum_{j=1}^{p} |\beta_j|$ then the LASSO coefficient estimates are the usual partial likelihood estimates, but if $s < |\beta_Z| + \sum_{j=1}^{p} |\beta_j|$ then the coefficient estimates are scaled to zero [21]. LASSO is one method used for variable selection because under the L1-norm penalty many coefficients will be penalized exactly to zero. One problem with the LASSO is if two covariates are strongly correlated, then the LASSO will pick one to be non-zero and penalize the other to zero [20]. In light of this, Ridge better handles highly correlated covariates, whereas LASSO produces a more parsimonious model. LASSO was suggested for use with the Cox model by Tibshirani [21] for situations where the number of covariates is greater than the number of study subjects, which causes instability in the Cox model.

II.1.8 LASSO Without Penalizing Exposure of Interest (LASSONoPen)

LASSONoPen is the same as LASSO except the coefficient for the exposure of interest is not penalized. So, similar to the RidgeNoPen, if β_1 is the regression coefficient for the exposure of interest, the constraint for LASSONoPen is revised to be

$$\sum_{j=1}^{p} |\beta_j| \leq s$$

This way we will not penalize the coefficient for the exposure of interest to zero when the effect is close to null, but still avoid overfitting by shrinking the coefficients for the remaining covariates [4].

II.2 Simulation Methods

II.2.1 Simulation Procedures

This section will describe the procedures used to simulate the data used for this paper. Every simulation was run under the constraint of a low EPV ratio, but several other factors were varied to evaluate performance under different circumstances. Factors considered were sample size, percentage that recieved treatment/exposure of interest, the true effect of treatment on event time and number of confounders. We considered five different sample sizes: 80, 120, 160, 200, and 240. Having always included 18 covariates with a percentage of events in the sample was approximately 20% for all setups, so smaller sample size meant very few EPV and larger sample size meant more EPV. The resulting EPV ratios (rounded to the nearest tenth) were 0.4, 0.9, 1.3, 1.8, and 2.2 respectively. Remember Harrell's [**?**] recommendation of 15 EPV and note that the largest EPV considered here is about seven times smaller [9]. Two different percentages of subjects recieving exposure were considered: 30% and 50%. Five different hazard ratios for the effects of treatment were considered: 1/3 (very strong protective), 1/2 (strong protective), 1 (no effect), 2 (strong causitive), 3 (very strong causitive). Last, two different confounder settings were examinged. The first confounder setup was such that all covariates considered had a non-zero effect on the outcome, and had various levels of correlation with the exposure variable. In contrast, for the second setup all covariates except two had 0 effect, and only those two with non-zero effects were correlated with the exposure variable. The first and second confounder setups will hereafter be referred to as the 'many confounder setup' and the 'two confounder setup', respectively. Five sample sizes $\times$ two exposure rates $\times$ five treatment effects $\times$ 2 confounder setups makes a total of 100 scenarios considered. It should be noted that initially, a sample size option of 40 was not included in the simulation and 240 was instead included. Due to the severely low event rate with that small of a sample, regression models suffered from severe monotone likelihood problems among possibly other problems and simulation results were riddled with errors and warnings. For that reason, we excluded the sample size option of 40 and added the option of 240.

For each scenario, m=1,200 Monte Carlo simulations were attempted, but in several scenarios a few simulations would result in errors, and more would produce results, but accompanied by warnings. Fortunately, the impact of the model errors was minimal, with most scenarios completing all simulations without issue. Considering individual scenarios (a scenario being a set of 1,200 simulations under a particular set of factors) at most 9 simulations (9/1,200 = 0.75%) included results of NA for atleast one method considered. Since so few models encountered errors, we analyzed only results of simulations that had model output for each method, which eliminated a total of 18 simulations from the many confounder setup and 11 simulations from the two confounder setup. Of greater concern for validity is the number of warnings, in particular for the regular CoxPH method. See Figures 25, 26, 27, and 28 to see the prevelance of warnings. These issues are further discussed in the Discussion section under the heading 'NA Coefficient Estimates and Monotone Likelihoods'.

Attempting to simulate realistic correlation structures for the baseline covariates, data was used from the drug repurposing study by Xu et al. mentioned as the motivating example in the Introduction [23]. The finalized set of covariates used for the simulation included 34,904 observations of 18 variables including dummy variables for tumor stage and 5 simulated continuous covariates (BMI and age were the only original continuous covariates).

With the covariates set, the next step was to randomize treatment selection and simulate survival outcomes. An overview of what we did was 1) generate treatment statuses, 2) generate true event times, 3) generate censoring times, and 4) compare each event time to the corresponding censoring time as well as an end of study cutoff to determine whether the event was observed. The final result for each subject is a time, and an event indicator. Parameters selected for the models used for event generation and censoring scheme were jointly selected to create our desired scenario of few events. The subjects had their binary treatment status determined by subject specific Bernoulli distributions with the probability of treatment (p_j) chosen by using a function of the linear predictor of subject covariates, $\mathbf{X_j}$:

$$p_j = \frac{1}{1+exp(\phi_1 X_{j,1} + \phi_2 X_{j,2} + ... + \phi_{18} X_{j,18})}$$

where ϕ_i is the effect of the ith covariate on treatment selection detemined to result in the desired exposure rate. Each subject has a treatment status Z_j and a set of covariates $\mathbf{X_j}$. From there, survival outcomes were simulated using an approach similar to Bender et al. [2]. Event time T_i was generated parametrically using subject specific Weibull distributions with shape parameter 10 and scale parameter a function of the linear predictors:

$$scale = 2.23exp[(-1/10)(\beta_Z Z_j + \beta_1 X_{j,1} + ... + \beta_{18} X_{j,18})]$$

The selection of the Weibull distribution was heavily influenced by its flexibility in modelling different kinds of hazard fuctions, other models proving difficult to work with. A censoring time was generated for each subject using a uniform distribution. Finally, the survival time actually recorded for each subject was the smallest number out of the event time, the censoring time, and the end of study time. If the number recorded was the event time, then the event indicator was recorded as 1, otherwise the subject was censored and the event indicator was recorded as 0. For each simulation, the survival outcomes were generated, a sample of the desired size drawn and the eight methods performed on the sample.

Simulations and statistical analysis of results were performed using R version 3.1.1. Several R packages were used including ‘survival’ for Cox models, ‘Hmisc’ for imputation, and ‘glmnet’ for Ridge and LASSO methods. Graphics were produced using the R package ‘lattice’.

II.2.2 Evaluation Criteria

The primary focus of simulation evaluation was placed on the bias, standard error, mean square error (MSE), median absolute error (MAE), and coverage probability. Descriptions and simulation results for some additional evaluation tools including the empirical standard error (ESE), lower coverage, upper coverage, and median confidence interval (CI) length, are included in the Appendix. For this section, let $\hat{\theta}_i$ stand for the estimated log-hazard of treatment effect produced by simulation i, let $\bar{\hat{\theta}}_i$ be the mean of the estimated log-hazards, and let θ be the true log-hazard used in data simulation.

II.2.2.1 Bias

The bias is calculated using the formula Bias($\hat{\theta},\theta$)=$\frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i-\theta)$, where m is the number of simulations. The bias can be thought of as the distance between the estimated log-hazard and the true log-hazard. For positive θ, a positive bias means overestimation of θ and a negative bias means underestimation of θ. For negative θ, a positive bias means underestimation of θ and a negative bias means overestimation of θ.

II.2.2.2 Mean and Median Standard Error

Standard Errors (SE) of the estimated coefficient for the exposure of interest were reported from software output for CoxPH, Unadjusted CoxPH, PSA, and PSH. For LASSO, LASSONoPen, Ridge, and RidgeNoPen, bootstrap estimates of the SE were calculated based on 500 bootstrap samples. The median SE was reported in addition to the mean as a measure that is more robust to outliers.

II.2.2.3 Mean Square Error (MSE)

MSE is calculated by the formula MSE($\hat{\theta},\theta$)= $\frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i-\theta)^2$. It can be shown that the MSE($\hat{\theta},\theta$)= Var($\hat{\theta}$)+$(Bias(\hat{\theta},\theta))^2$. Being a combination of variance and bias makes the MSE a decent metric to summarize the quality of an estimate using one number. Lower MSE means the estimates have low variability and are not very biased, meaning lower is considered better. A questionable aspect of the MSE (and means in general) is that outliers are highly influential, meaning the mean is not always the best way to characterize a sample. Squaring the values further exacerbates the high influence, making the MSE very sensitive to outliers.

II.2.2.4 Median Absolute Error (MAE)

Due to the MSE's sensitivity to outliers a common metric used in conjunction with the MSE is the MAE. The MAE is calculated using the median (opposed to the mean) making it more robust to the effects of outliers. The formula used to calculate it is MAE($\hat{\theta},\theta$)=median$|\hat{\theta}-\theta|$.

II.2.2.5 Empirical Coverage Probability

For each method we estimated the empirical coverage probability of 95% confidence intervals (CI) by finding the percent of simulation estimates that are within the limits of their associated CI. For CoxPH, Unadjusted CoxPH, PSA, and PSH, we used the CIs produced by the software. For LASSO, LASSONoPen, Ridge and RidgeNoPen, the CIs are calculated as CI=($\hat{\theta}-1.96*1.4826*MAD$,$\hat{\theta}+1.96*1.4826*MAD$) where MAD stands for 'median absolute deviation'. The MAD is calculated using the formula MAD=$median_i(|\hat{\theta}_i-$

$median_j(\hat{\theta}_j)|)$. Our purpose for calculating the MAD is specifically to use with bootstrap estimates from LASSO and Ridge methods. We used the bootstrap with 500 iterations on the LASSO, LASSONoPen, Ridge and RidgeNoPen. We then calculated the MAD of the bootsrap estimates in order to robustly estimate standard deviations for use in normal approximation confidence intervals. When multiplied by a factor of 1.4826, the MAD becomes a consistent estimator of the standard deviation for Normal distributions [18]. In general we like to see the coverage be around 95%. If the coverage is greater than 95% the method is too conservative, and if the coverage is less than 95% then the method is too liberal.

CHAPTER III

Results

Each type of result is accompanied by four figures. The first and second figures contain results under the many confounder setup with exposure rates of 30% and 50% respectively. The third and fourth figures contain results under the two confounder setup with exposure rates of 30% and 50% respectively. Each figure contains five panels, each panel presenting the results for simulations under a different true value for β_Z. Each of the eight methods discussed in the Methods section has a different line type and color combination.

III.1 Bias (Figs. 1, 2, 3, 4)

Under the many confounder setup, when the exposure is balanced (50%) and there is no association (log haz=0) between exposure of interest and outcome, bias for all methods is very close to 0 (Fig. 2). For the same exposure and association under the two confounder setup (Fig. 4), the methods CoxPH, PSA, PSH and LASSO still have biases very close to 0, though the other methods show some biases with Unadjusted CoxPH having the largest. The log-hazard=0 frames for imbalanced exposure (30%) (Figs. 1, 3) look similar regardless of confounder setup, again with near 0 bias for CoxPH, PSA, PSH and LASSO, and some bias for the rest. For other log-hazards, CoxPH starts extremely biased for number of EPV less than about 1.5, but quickly attains the smallest bias for more moderate numbers of EPV. For 1.5 EPV or fewer, PSA and PSH seem to consistently have low bias, and are the closest to 0 for all scenarios except when exposure rate=.5 with the many confounder setup in which the PS methods are met or beat by RidgeNoPen, LASSONoPen and the Unadjusted CoxPH over the range of EPV. Unadjusted CoxPH has surprisingly low bias when the log hazard is positive in the many confounder setup (Figs. 1, 2). Notice that the CoxPH is the only method that has substantially decreasing bias as number of EPV increases, whereas the other methods stay relatively flat.

Figure 1: Bias of Estimates (Many Confounder Setup and Exposure Rate=.3)

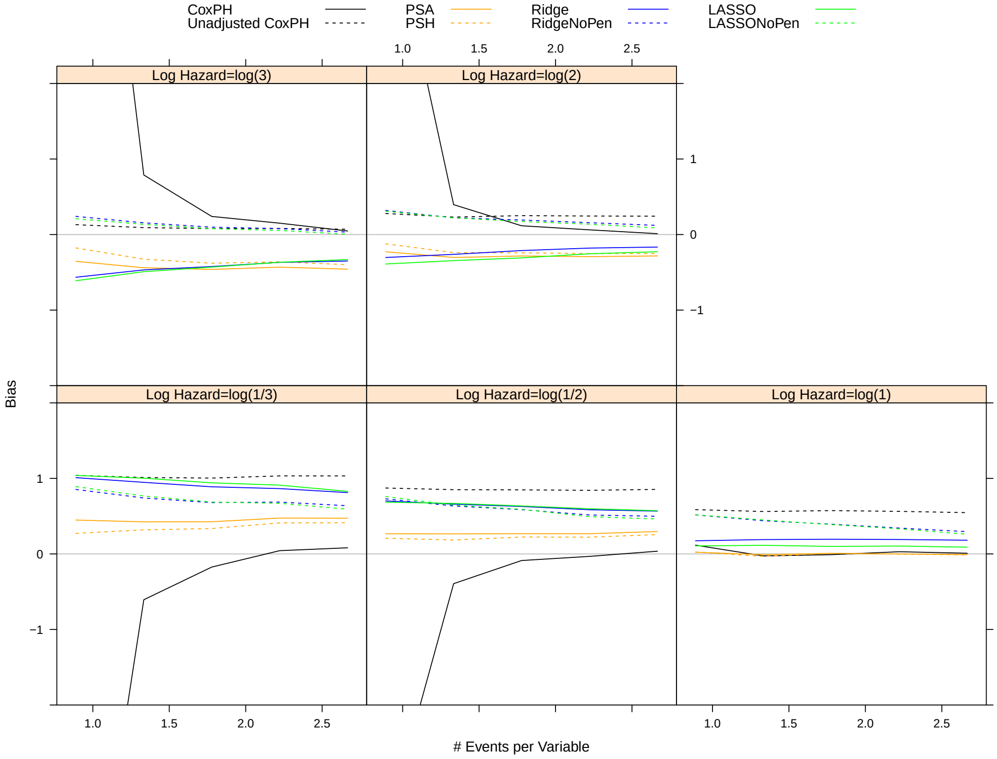


Figure 2: Bias of Estimates (Many Confounder Setup and Exposure Rate=.5)

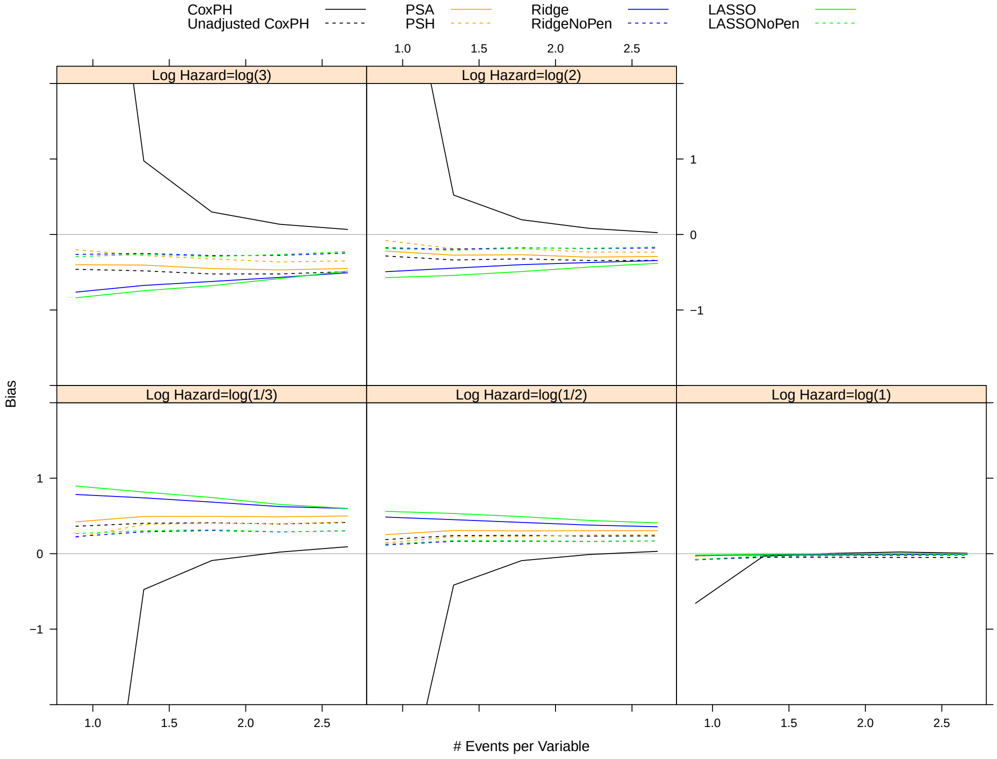


Figure 3: Bias of Estimates (Two Confounder Setup and Exposure Rate=.3)

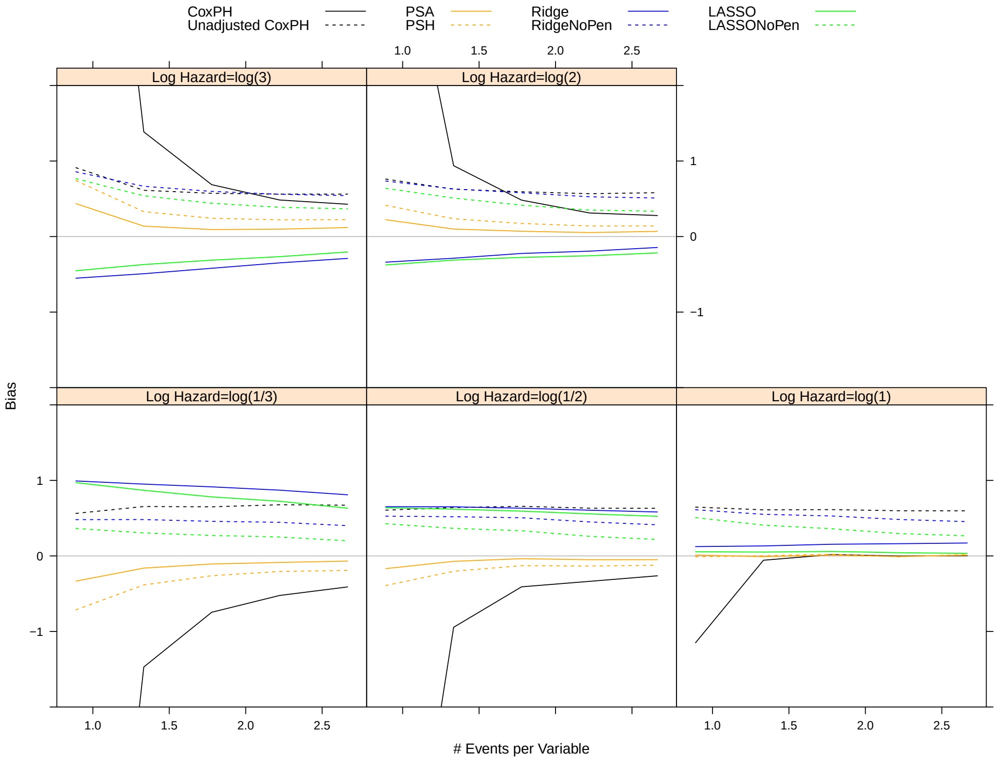
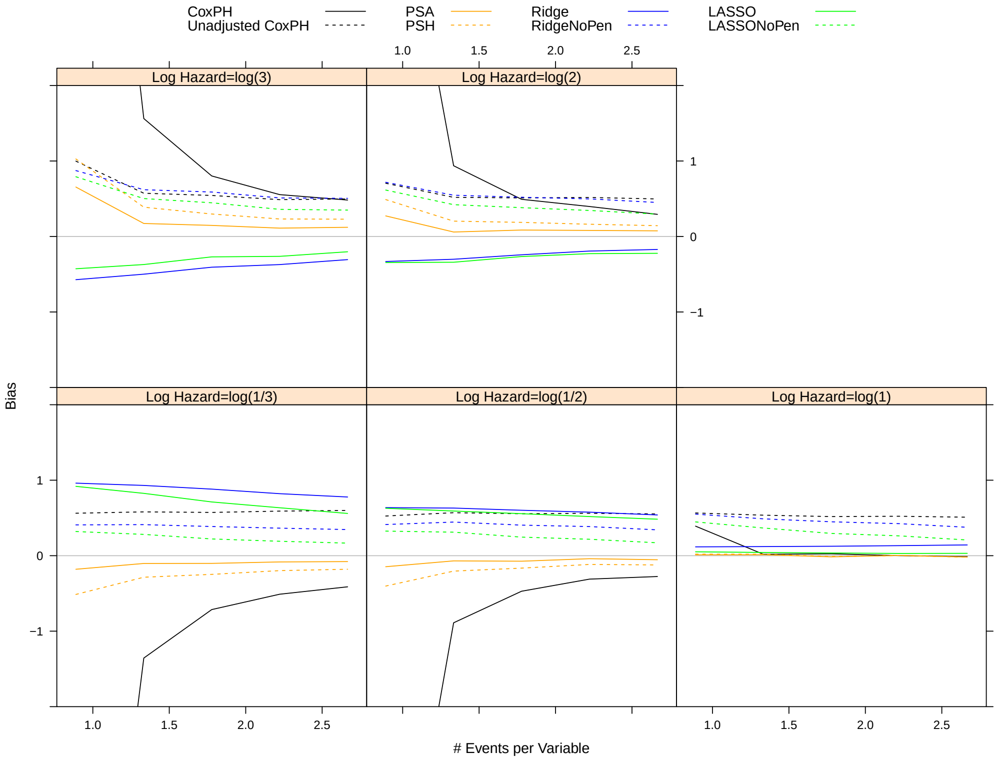


Figure 4: Bias of Estimates (Two Confounder Setup and Exposure Rate=.5)



III.2 Mean and Median Standard Error (Figs. 5, 6, 7, 8, 9, 10, 11, 12)

The picture of the mean and median standard errors are very similar accross different scenarios. The standard errors start high and drop considerably as the number of EPV increases. LASSO almost always shows smallest standard errors, met or closely followed by Ridge. The next lowest cluster of methods includes LASSONoPen, RidgeNoPen and Unadjusted CoxPH. After those, are PSA and PSH performing fairly similarly, and finally CoxPH always has the largest standard errors.

Figure 5: Mean Standard Error of Estimates (Many Confounder Setup and Exposure Rate=.3)

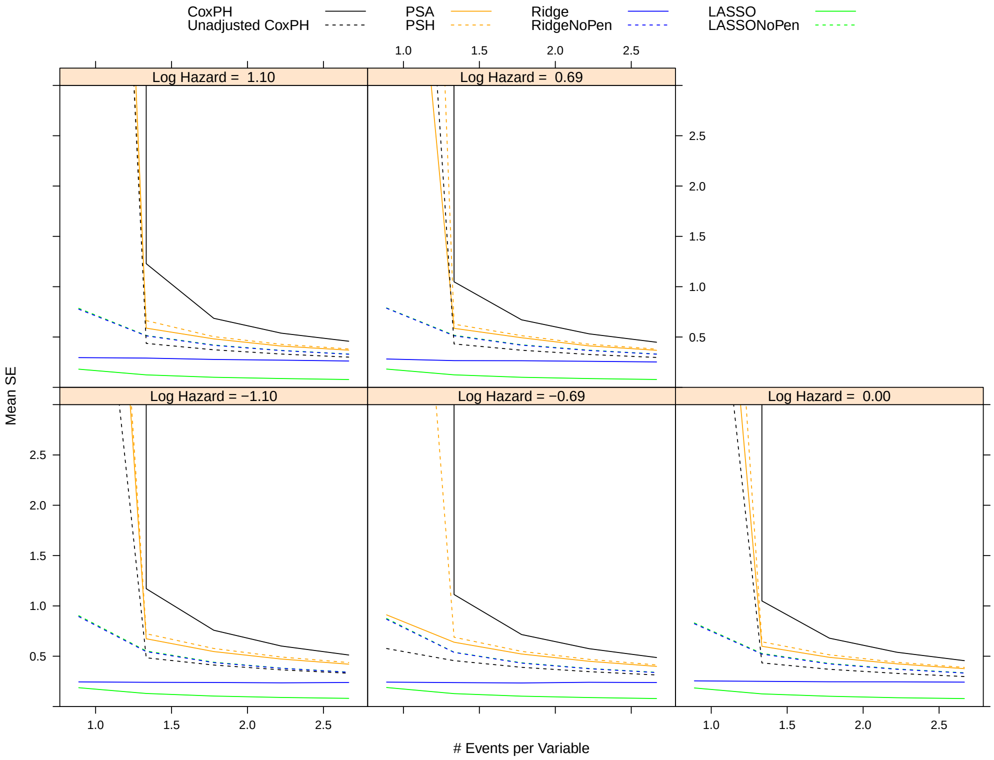


Figure 6: Mean Standard Error of Estimates (Many Confounder Setup and Exposure Rate=.5)

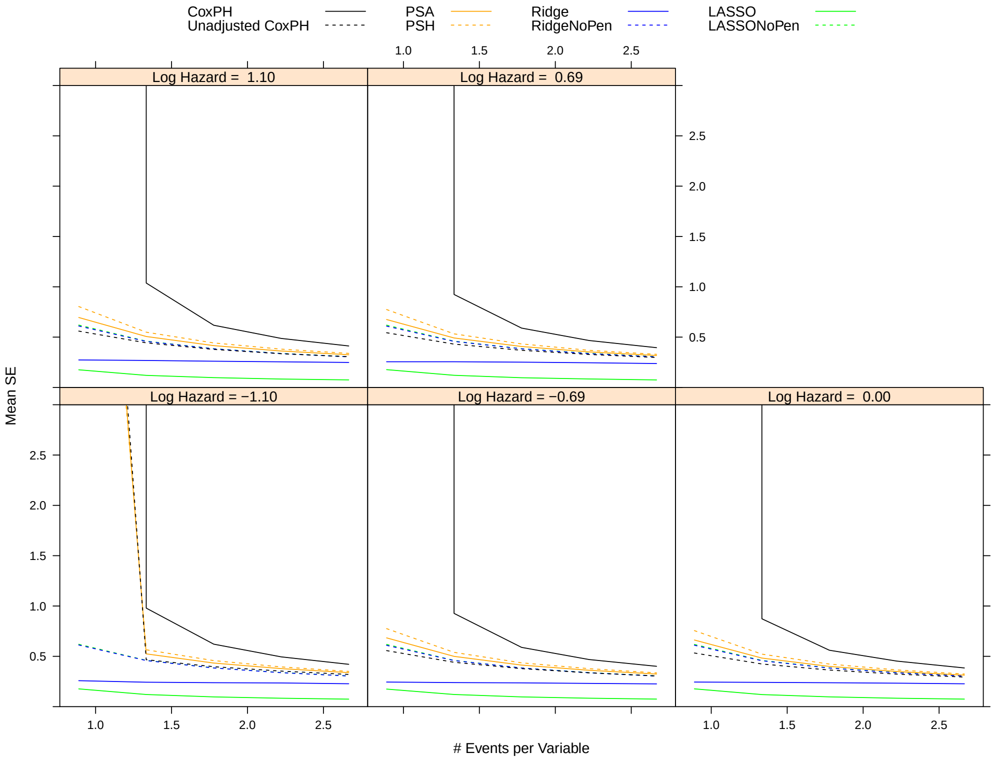


Figure 7: Mean Standard Error of Estimates (Two Confounder Setup and Exposure Rate=.3)

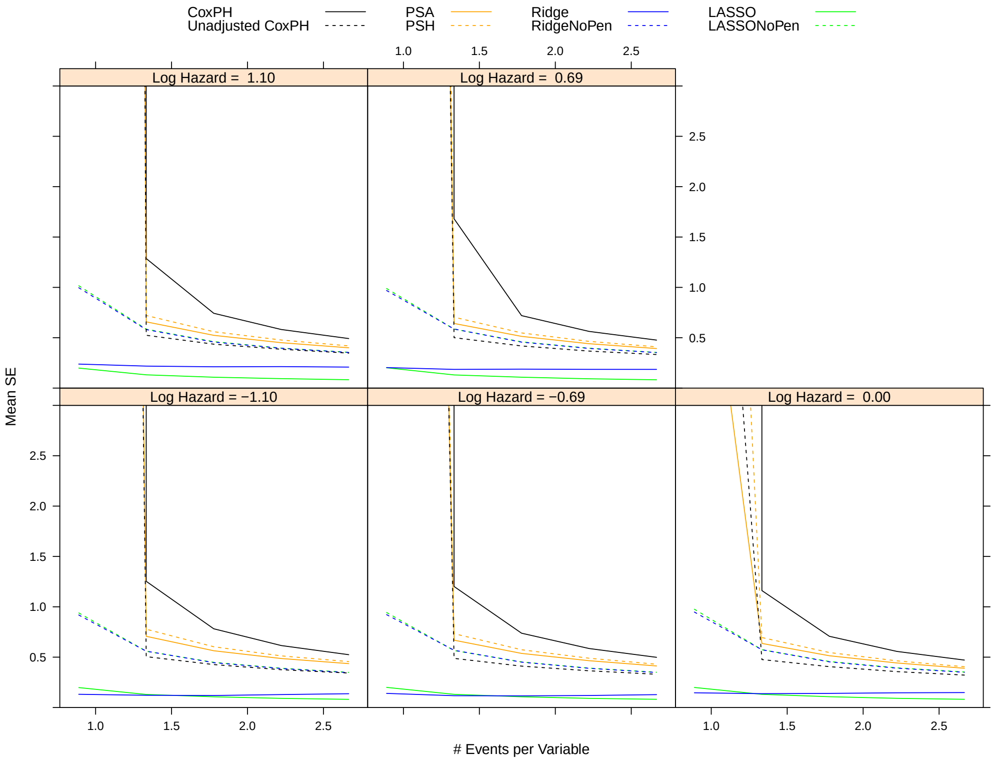


Figure 8: Mean Standard Error of Estimates (Two Confounder Setup and Exposure Rate=.5)

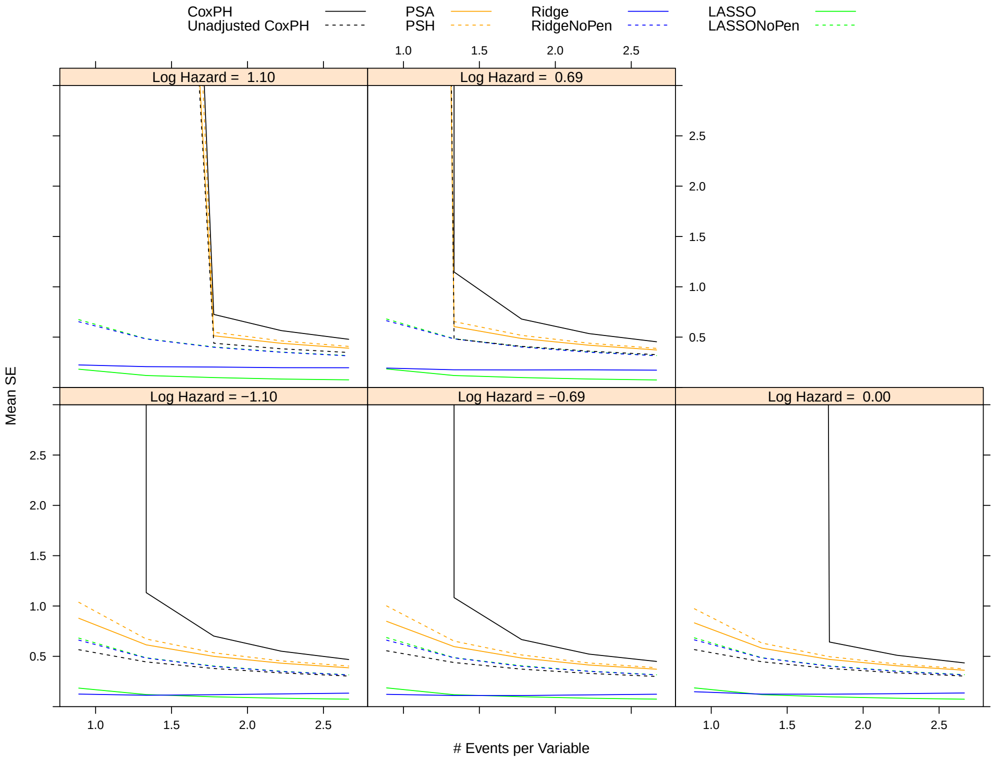


Figure 9: Median Standard Error of Estimates (Many Confounder Setup and Exposure Rate=.3)

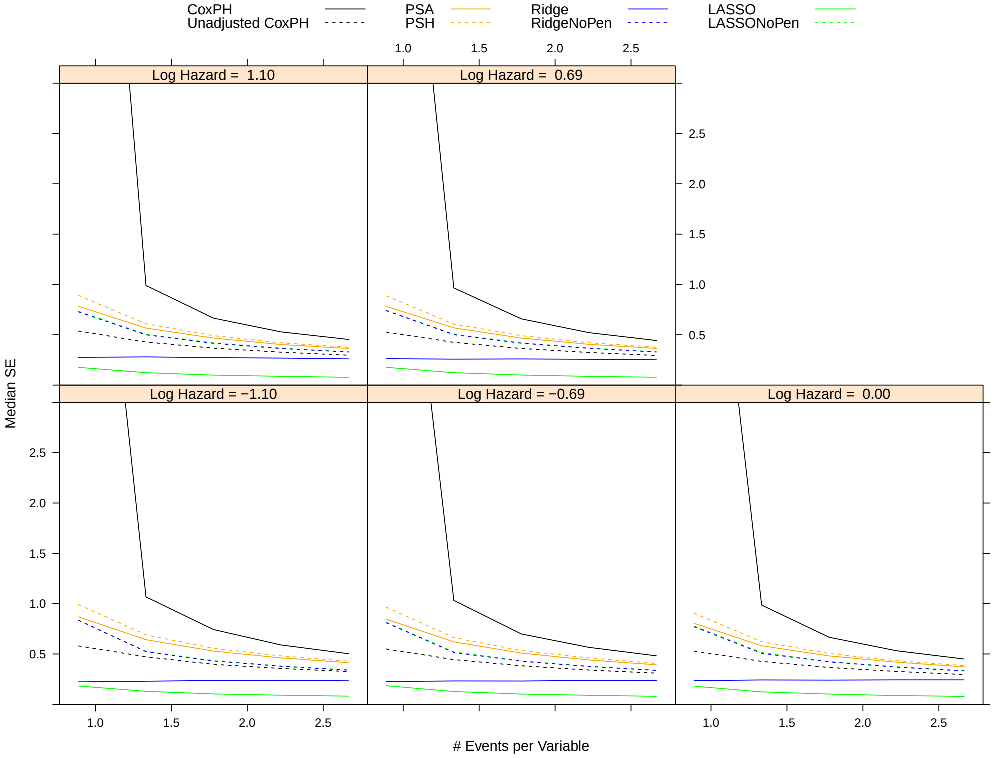


Figure 10: Median Standard Error of Estimates (Many Confounder Setup and Exposure Rate=.5)

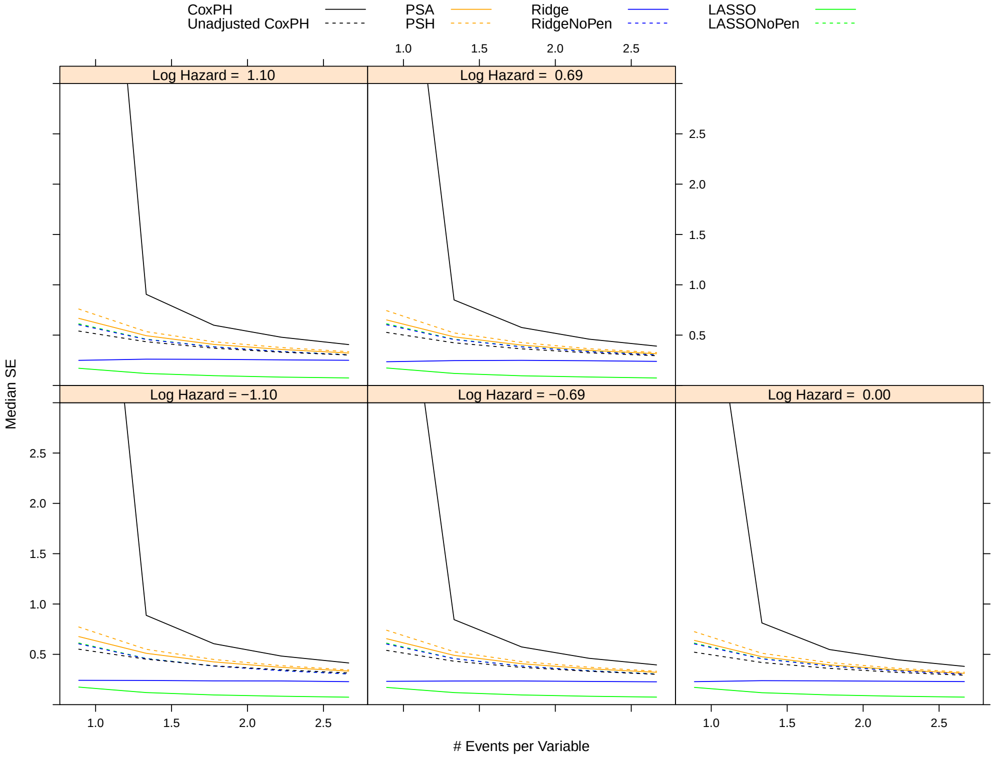


Figure 11: Median Standard Error of Estimates (Two Confounder Setup and Exposure Rate=.3)

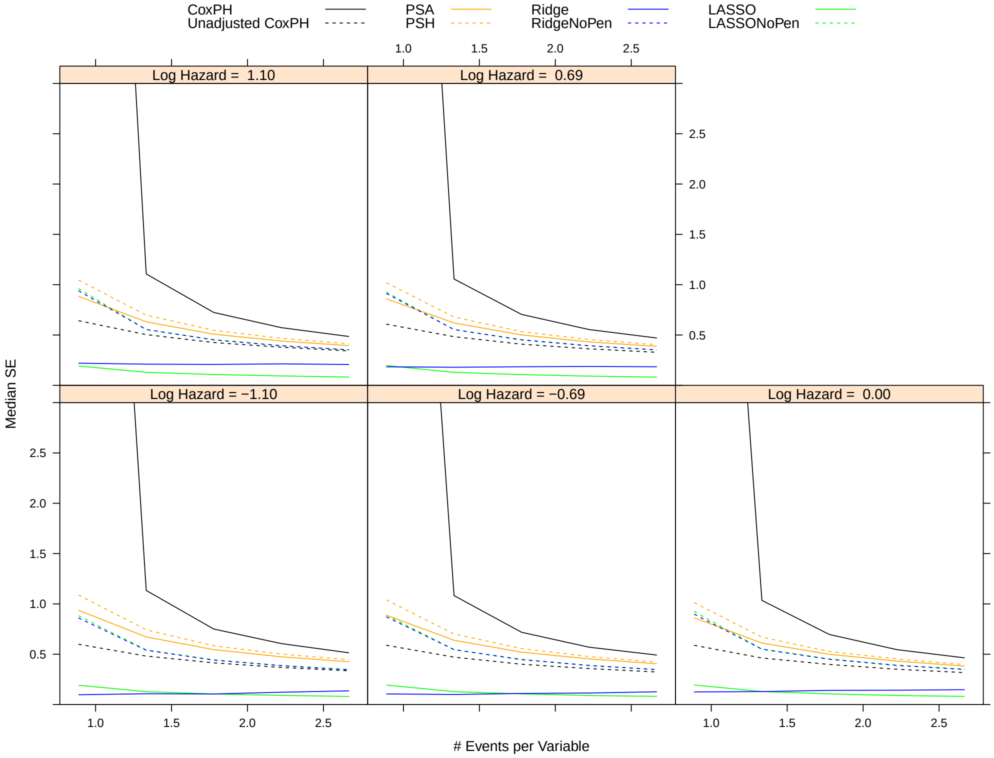
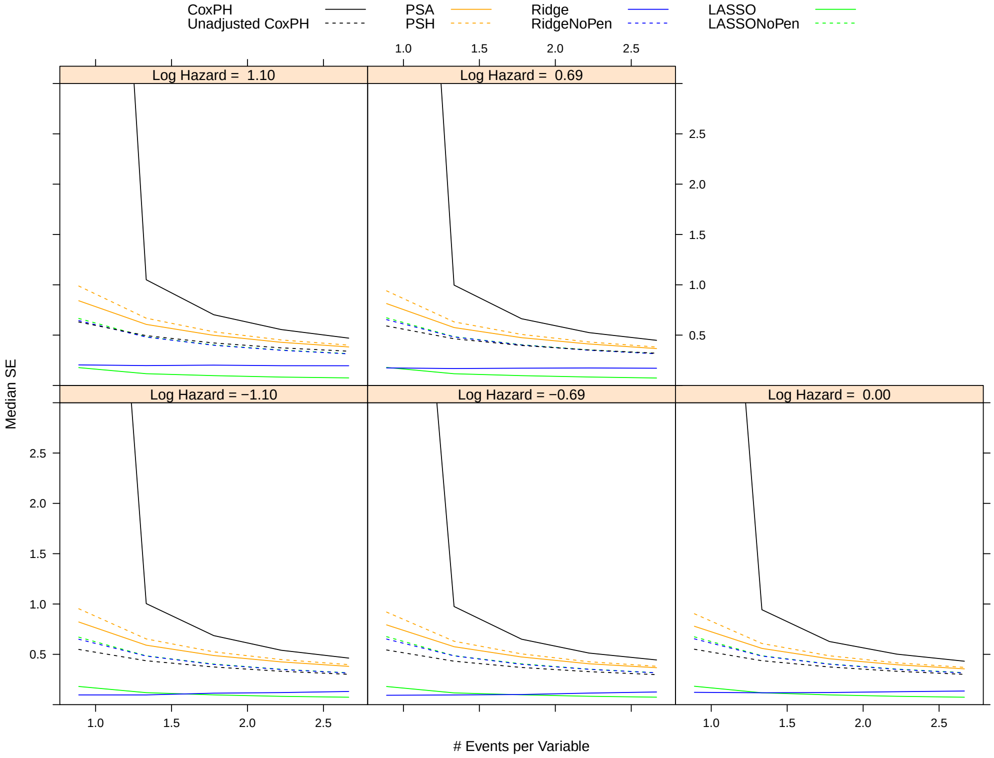


Figure 12: Median Standard Error of Estimates (Two Confounder Setup and Exposure Rate=.5)



III.3 Mean Square Error (MSE) (Figs. 13, 14, 15, 16)

CoxPH starts extremely high, but drops quickly around 2 EPV to become more competetive with the other methods. Unadjusted CoxPH has surprisingly low MSE when the log-hazard is positive in the many confounder setup, which is probably due to the very low bias. LASSO and Ridge seem to do better as the log-hazard approaches 0. Other methods do not appear to vary much accross log-hazard.

Figure 13: Mean Square Error of Estimates (Many Confounder Setup and Exposure Rate=.3)

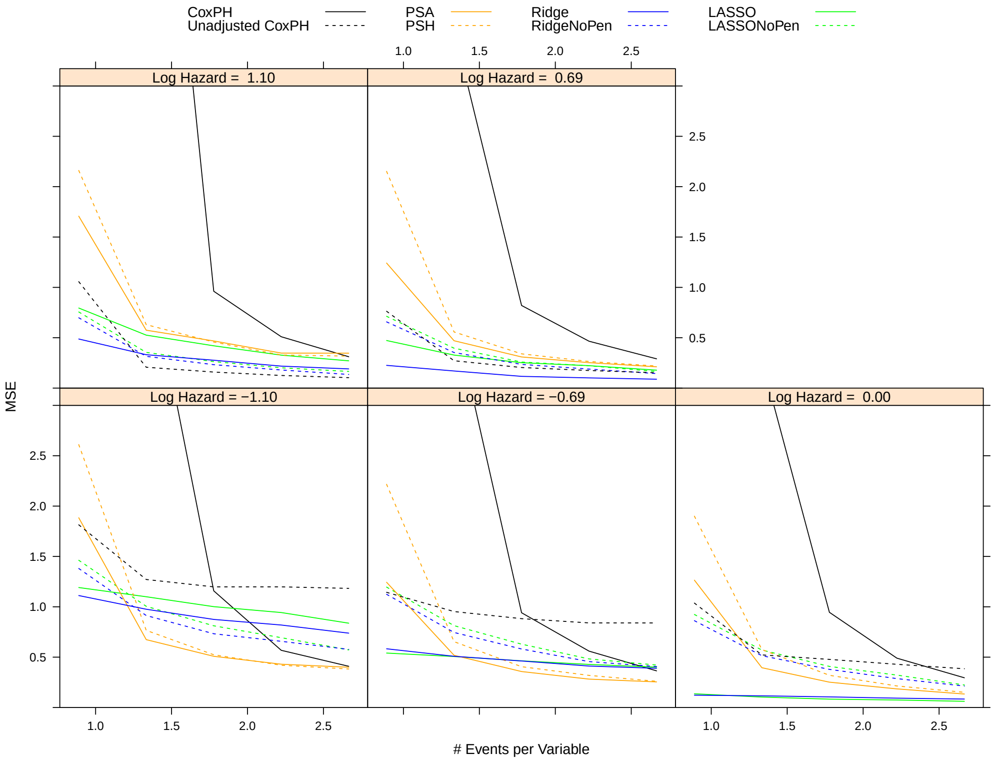


Figure 14: Mean Square Error of Estimates (Many Confounder Setup and Exposure Rate=.5)

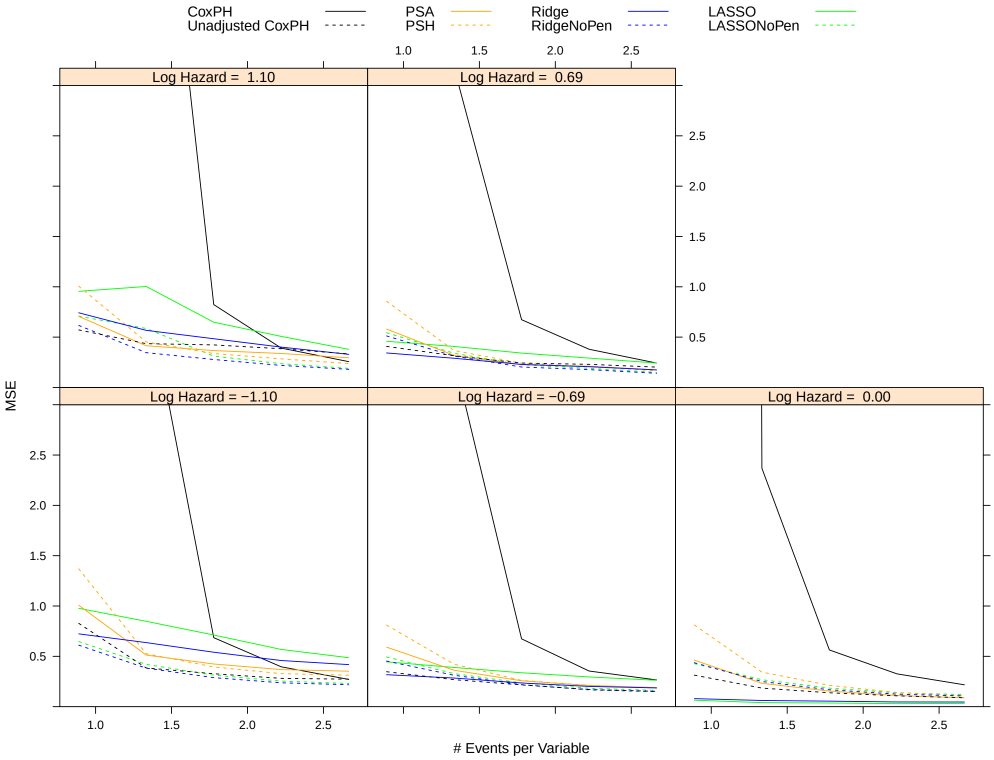


Figure 15: Mean Square Error of Estimates (Two Confounder Setup and Exposure Rate=.3)

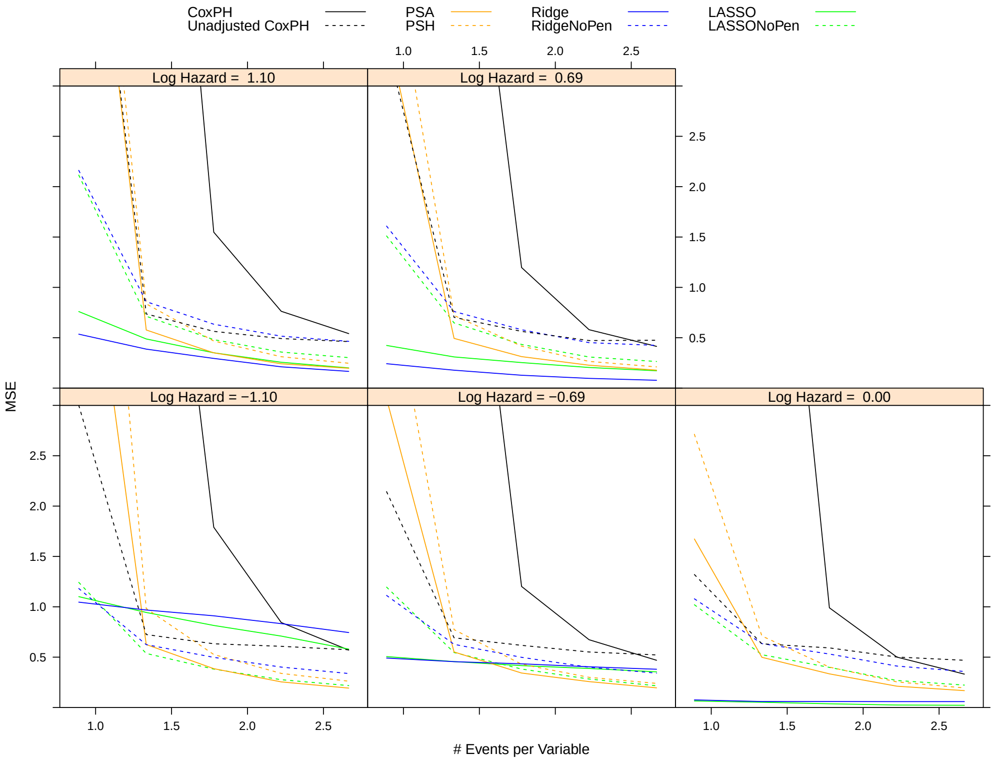
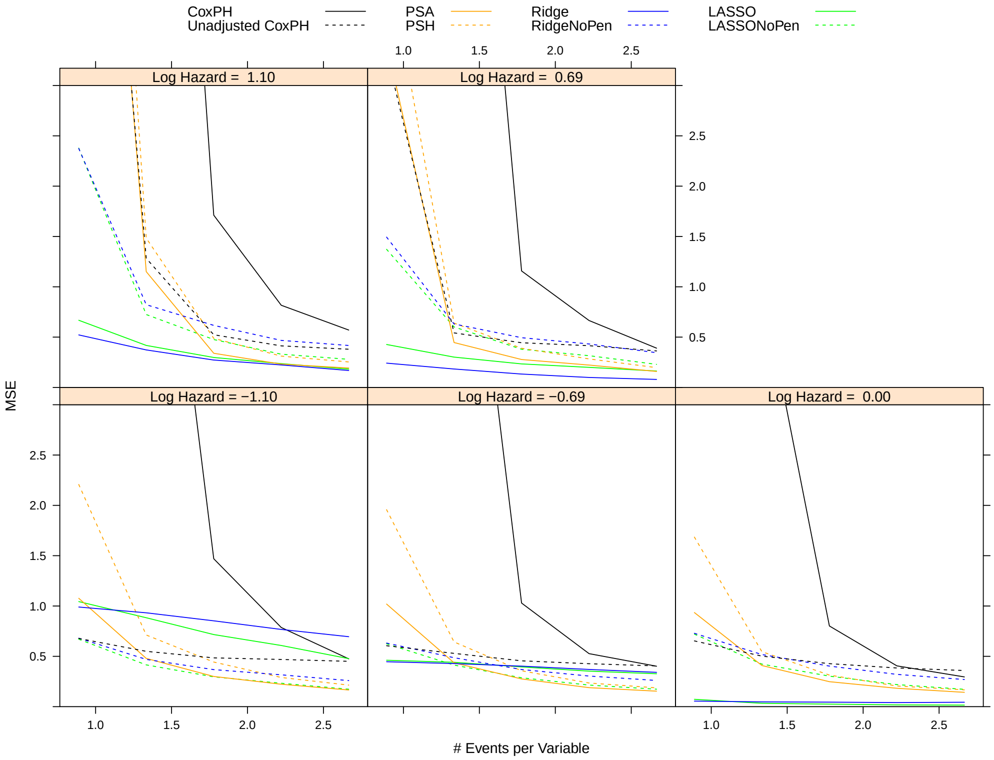


Figure 16: Mean Square Error of Estimates (Two Confounder Setup and Exposure Rate=.5)



III.4 Median Absolute Error (MAE) (Figs. 17, 18, 19, 20)

Methods that had very high MSEs in the low EPV range, have lower MAE. CoxPH, for instance, is still quite large around 1.5, but compared to the MSE which was off the chart at that level, it shows the influence of outliers in this case. There are some similar patterns here to those seen for the MSE, with the CoxPH performing quite poorly overall. PSA and PSH frequently have the lowest MAE in the two confounder case, but is consistently beat by the LASSONoPen and RidgeNoPen in the many confounder setup. Like with MSE, LASSO and Ridge seem to improve most dramatically as the log-hazard gets closer to 0, with those methods being lowest in the log-hazard=0 frame.

Figure 17: Median Absolute Error of Estimates (Many Confounder Setup and Exposure Rate=.3)

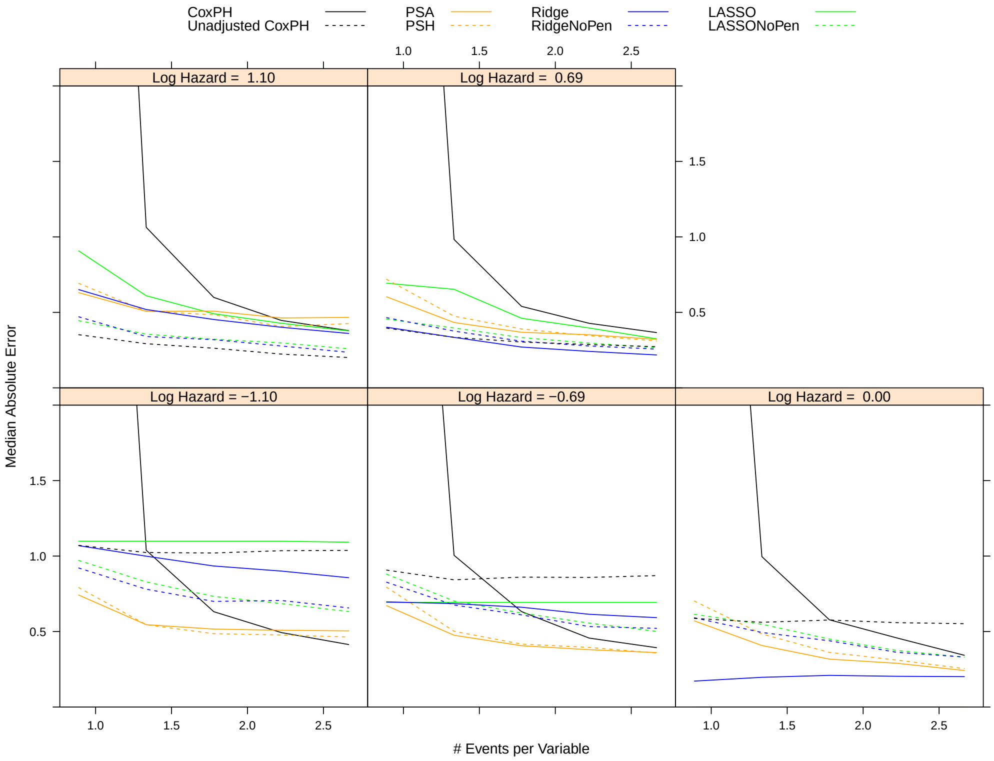


Figure 18: Median Absolute Error of Estimates (Many Confounder Setup and Exposure Rate=.5)

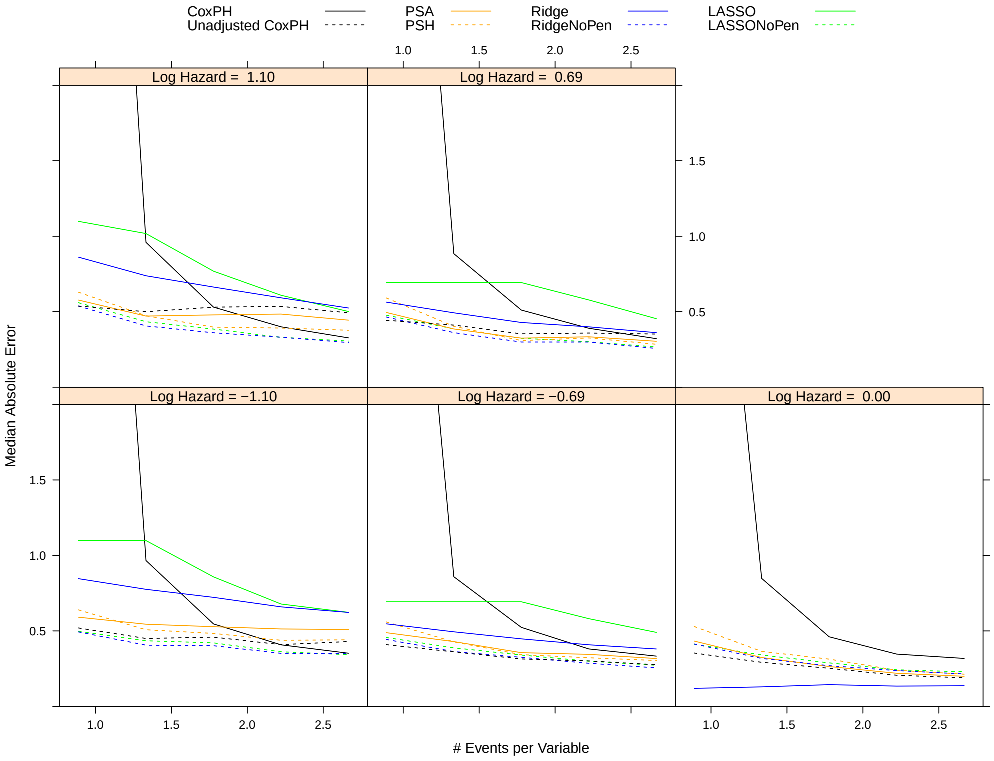


Figure 19: Median Absolute Error of Estimates (Two Confounder Setup and Exposure Rate=.3)

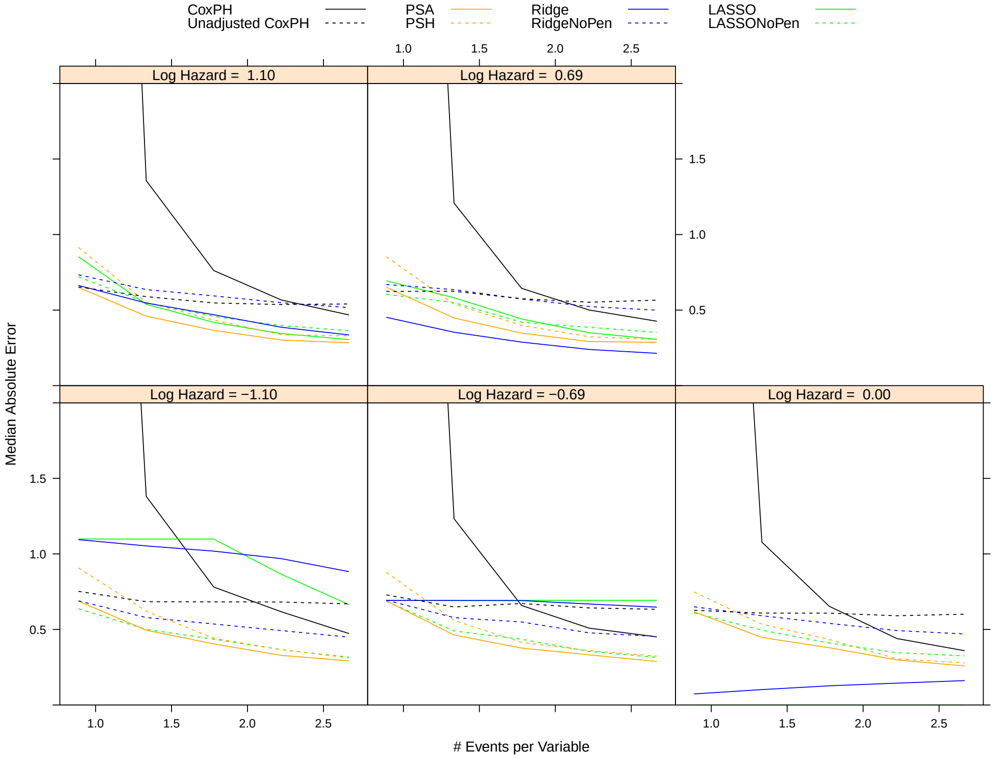
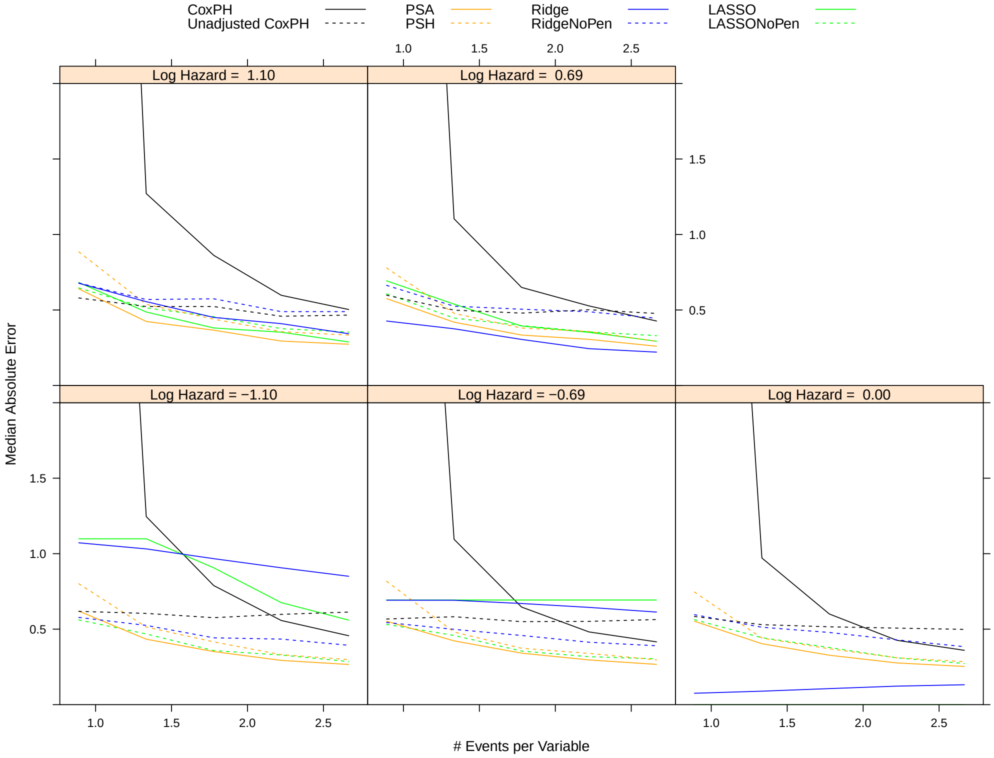


Figure 20: Median Absolute Error of Estimates (Two Confounder Setup and Exposure Rate=.5)



III.5 Empirical Coverage Probability (Figs. 21, 33, 37, 22, 34, 38, 23, 35, 39, 24, 36, 40)

The coverage of PSA and PSH seems to be quite close to 95%, especially in the two confounder setup. LASSONoPen and RidgeNoPen are more competetive in the many confounder setup. In general, methods seem to have coverage well below the 95% level, LASSO and Ridge being particularly severe when the true log-hazard is not 0.

Figure 21: Coverage Probability of 95% Confidence Intervals (Many Confounder Setup and Exposure Rate=.3)

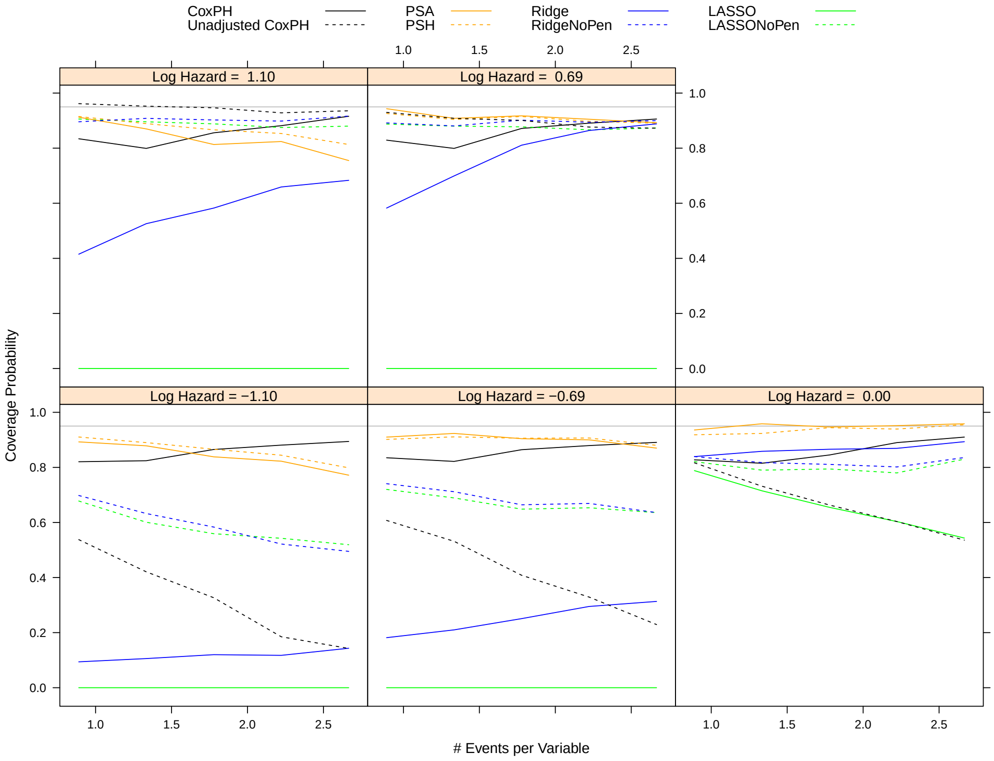


Figure 22: Coverage Probability of 95% Confidence Intervals (Many Confounder Setup and Exposure Rate=.5)

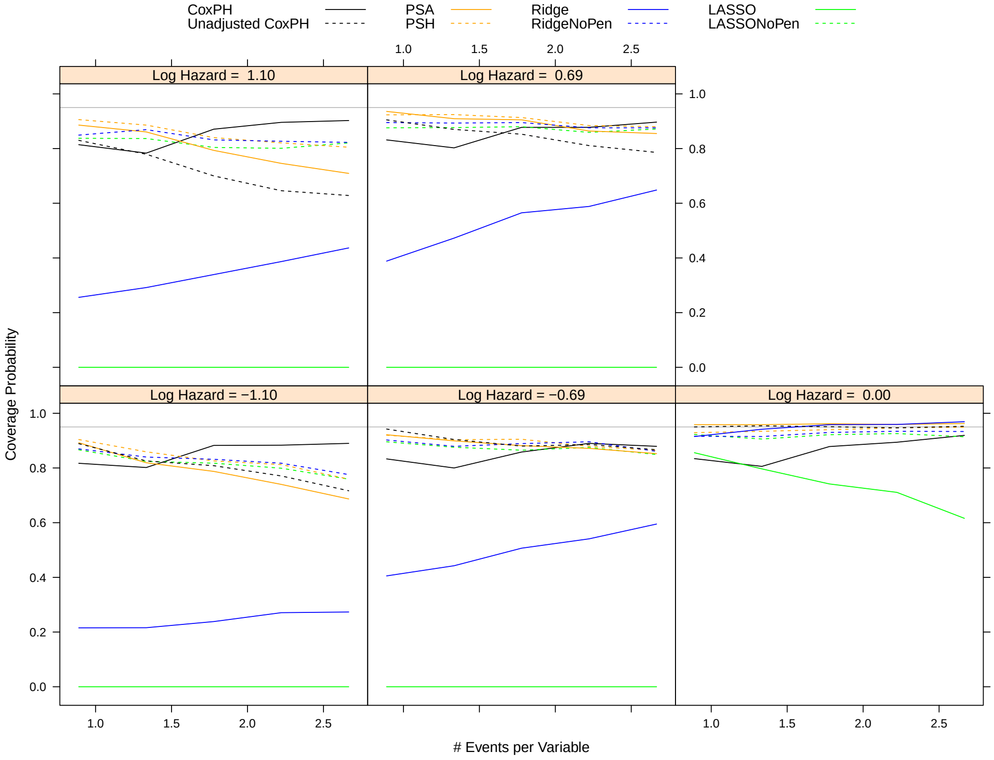


Figure 23: Coverage Probability of 95% Confidence Intervals (Two Confounder Setup and Exposure Rate=.3)

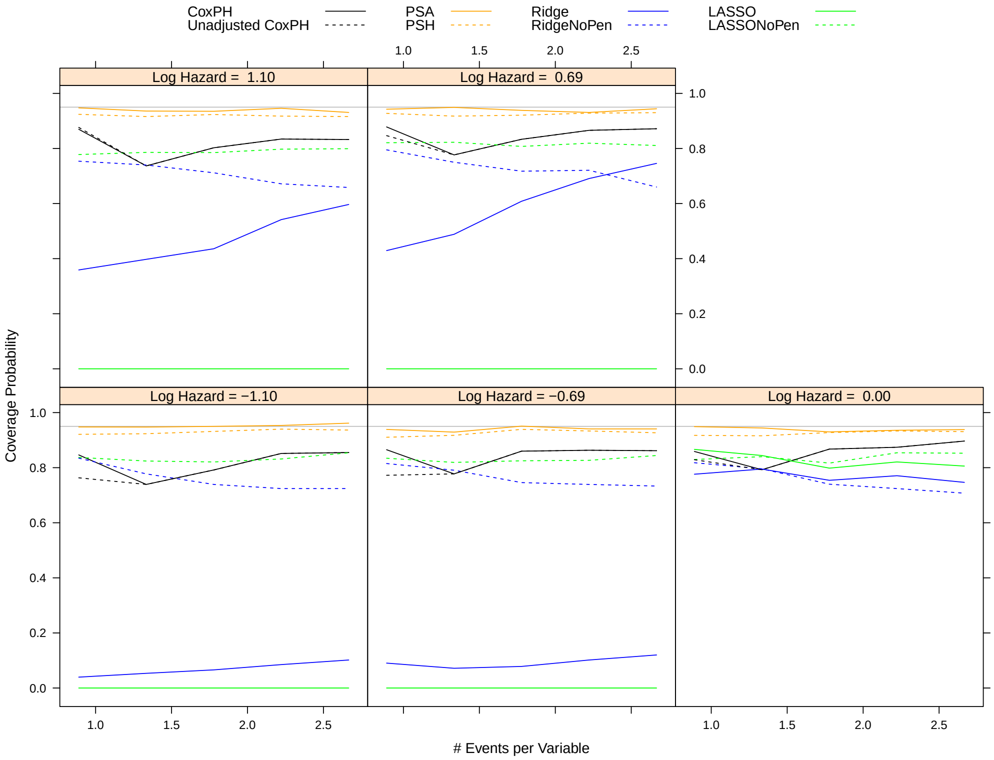
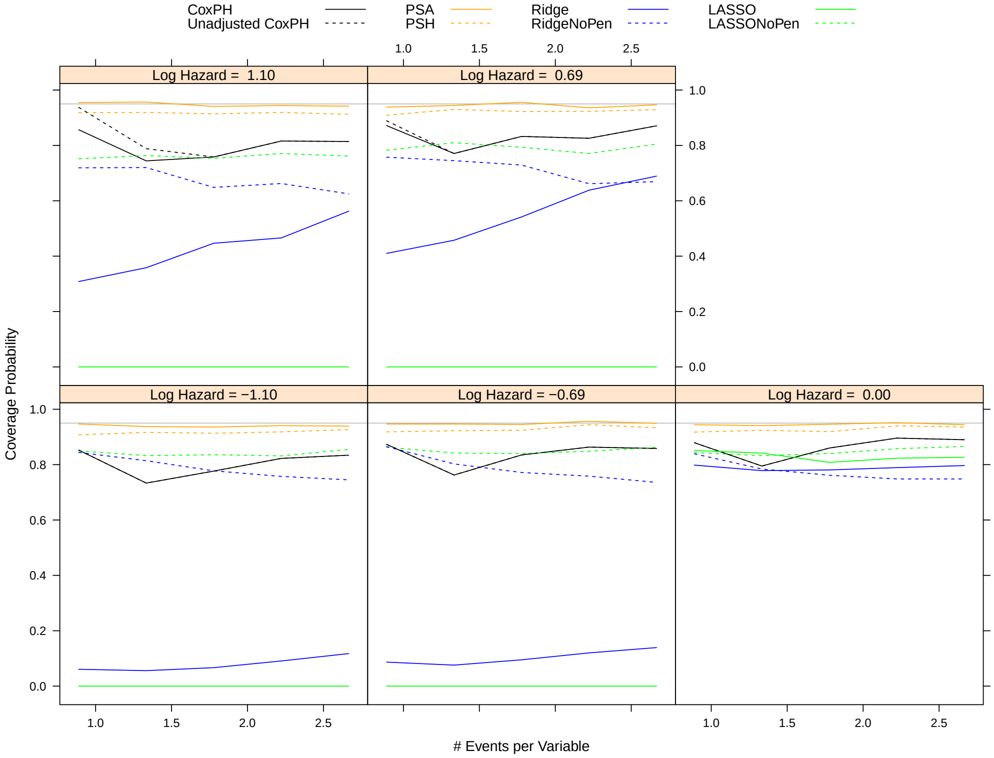


Figure 24: Coverage Probability of 95% Confidence Intervals (Two Confounder Setup and Exposure Rate=.5)



CHAPTER IV

Discussion

Interestingly, the Unadjusted CoxPH method is frequently less biased than CoxPH when the number of EPV is less than 1.5. This difference in performance is likely explained by the Unadjusted CoxPH avoiding monotone likelihood problems by not adjusting for any covariates other than the exposure of interest. Still, it is surprising how low the bias is for Unadjusted CoxPH when the log-hazard is positive in the many confounder setup. Intuition says that with more confounders present, bias will be improved by adjusting for those confounders, but apparently other underlying factors are causing other methods to give more biased results than the unadjusted model. Potential explanation for high bias in methods other than PSH is the non-collapsibility of survival outcomes, particularly if needed covariates are omitted or penalized [7].

The low standard errors of the LASSO method are in accordance with previous literature [21][9]. In general, methods avoiding monotone likelihood issues have a significant advantage by most measures. Some mean standard errors at low number of EPV are pulled to extreme values due to a few models providing estimated coefficient values that did not converge. The median standard error is less influenced by those extreme values, which again is one reason why the median standard errors are also reported.

There is some concern about the validity of coverage probabilty figures. Concerning the LASSO and Ridge, it makes some sense that these methods would occasionally penalize the exposure, however the LASSO appears to do so almost exclusively, and the estimated standard errors are very low due to how consistently the exposure is being penalized. Consequences of the penalty are clearly present in the bias. When the log-hazard is negative, the exposure is penalized up towards 0 resulting in positive bias and when the log-hazard is positive, the exposure is penalized down towards 0 resulting in negatively biased estimates. Looking at any of the bias figures and comparing the LASSO and Ridge with their unpenalized exposure counterparts, LASSONoPen and RidgeNoPen, one can see that the choice of whether to penalize the exposure is a gamble. By allowing the exposure to be penalized, the bias might be reduced (see Figure 3, panel with log-hazard=log(2)), or the bias might be increased (see Figure 3, panel with log-hazard=log(1/2).

A general trend in the simulation results is that RidgeNoPen and LASSONoPen methods appear more sensitive to exposure imbalance. It is possible the effects of exposure imbalance are inflated by the choice to not penalize it. For instance the MSE (Figs. 13, 14, 15, 16) is frequently lower for RidgeNoPen and LASSONoPen when the exposure rate is .5 compared to when it is .3, whereas Ridge and LASSO are more similar across exposure rates.

IV.1 NA Coefficient Estimates and Monotone Likelihood

As Heinze [11] explains, it is possible for the iterative fitting process of a regression model to not converge for one or more of the parameters being estimated, and this problem is referred to as the 'monotone likelihood problem'. In the simple case of a univariate regression, monotone likelihood happens when the covariate's values for subjects experiencing events are always the largest or always the smallest values in the sample [11]. Monotone likelihood is simply demonstrated by considering a contingency table of a dichotomous variable (X) and event status (δ). In this example there are four types of subject: ($X = 1,\delta = 1$), ($X = 1,\delta = 0$), ($X = 0,\delta = 1$), and ($X = 0,\delta = 0$). Monotone likelihood occurs when either the (X=1,δ=1) cell is empty (X is always 0 for subjects experiencing events), or when the (X=0,δ=1) cell is empty (X is always 1 for subjects experiencing events). When considering multivariate regression, monotone likelihood occurs when the same happens for a linear combination of covariates [11]. Sample size, the number of events, and the magnitudes of association with the outcome, degree of balance, and number of binary covariates all effect the prevalence of monotone likelihood. Lower sample size means fewer events, and that means higher probability of empty contingency cells. If a dichotomous covariate is very strongly associated with the outcome, it is more likely to have uniform values. Similarly, if the covariate is very imbalanced (mostly 1 or mostly 0) then the under-represented value has a high chance to induce monotone likelihood. Furthermore, a larger number of dichotomous variables means more opportunity for there to be a problematic covariate as described present within the set of covariates.

We see two different types of monotone likelihood problems. The first, is specifically when the coefficient for the exposure of interest does not converge and reports 'NA' for the coefficient estimate. In the case of monotone likelihood due to the exposure of interest, careful error catching had to be implemented to make sure batches of simulations did not prematurely end due to errors caused by the problem. We kept tally of the iterations producing NA results for model coefficient estimates. Fortunately, out of 60,000 (1,200 simulations $\times$ 50 scenarios = 60,000) total simulations under the many confounder setup, only 18 simulation results contained an NA result. For the 60,000 simulations under the two confounder setup only 11 simulation results contained an NA result. These NAs were a symptom of the fairly extreme conditions that were being modelled. NA results occured exclusively in scenarios which used the smallest sample size considered (80). In addition, assigning a very strong protective treatment effect and a lower rate of exposure (30%) results in more NAs being produced. Remembering that the number of events for the simulated data is low to begin with, it should be no surprise that monotone likelihood problems arrose in several samples.

The second type of monotone likelihood problem were due to monotone likelihood resulting from covariates other than the exposure of interest. When these issues occur with other covariates, a coefficient estimate for the exposure of interest is still produced, but the output includes a warning message concerning lack of convergence for model coefficients. The method most significantly effected by the monotone likelihood problem is the regular CoxPH method (see Figures 25, 26, 27, and 28). In his paper, Heinze includes simulation results for prevalences of monotone likelihood [11]. Estimated probabilities were based on 1000 simulations per senario and sample sizes of 50, 100, and 200. Heinze reports prevalences of monotone likelihood as extreme as 100% for scenarios across all sample sizes considerd with 10% censoring rate, and 5 dichotomous covariates, each with a relative risk of 16, and 1:4 ratio of 1s to 0s. For the simulations in this paper, the pattern of monotone likelihood prevalence was quite similar accross confounder setup. Since cases of monotone likelihood due to the exposure of interest all resulted in NAs, the pattern of warnings was naturally similar across exposure rate and across true log-hazards of association between the outcome and exposure. For sample sizes of 80, the prevalence of monotone likelihood for CoxPh is always between 60% and 80%, then sharply falls to near 0% for sample sizes of 100 and up. Other methods never exceded 10% even at sample size of 80, and likewise fall to near 0% for sample size 100 and up. Noticing monotone likelihood problems helps explain why there are huge spikes in quantities like bias and standard error when the sample size is low, which is consistent with other literature [11][21]. Heinze reports average biases to be very large (from 70 to 400 times true relative risk) when more than 50% of samples produced monotone likelihood.

Figure 25: Ratio of Warnings to Number of Simulations (Many Confounder Setup and Exposure Rate=.3)

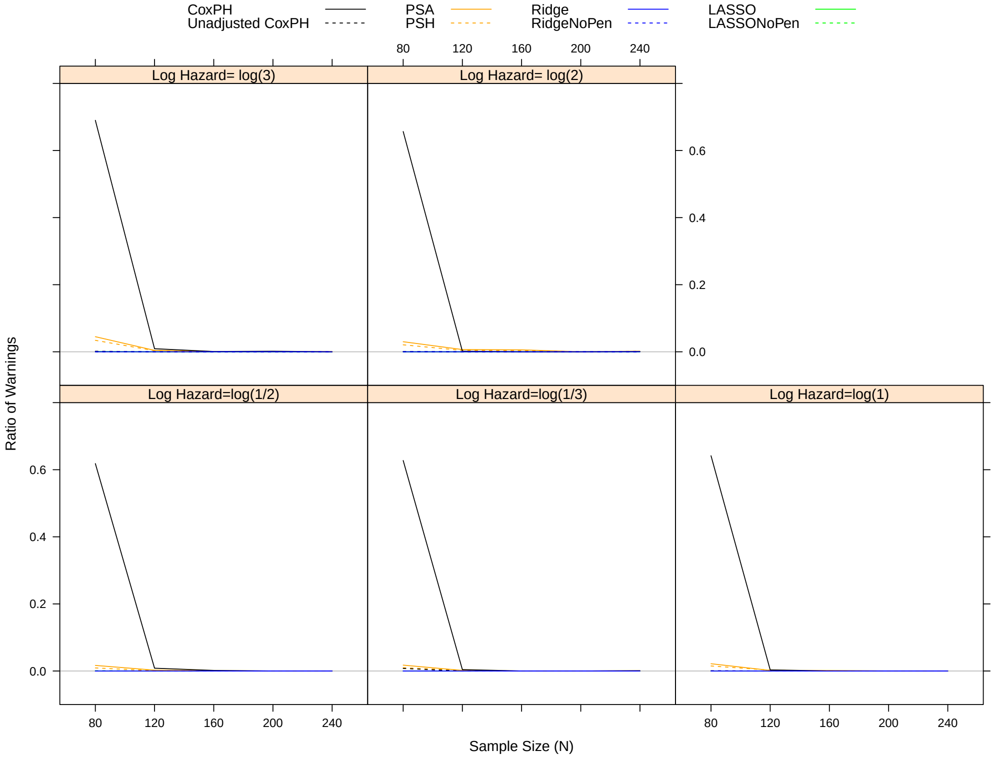


Figure 26: Ratio of Warnings to Number of Simulations (Many Confounder Setup and Exposure Rate=.5)

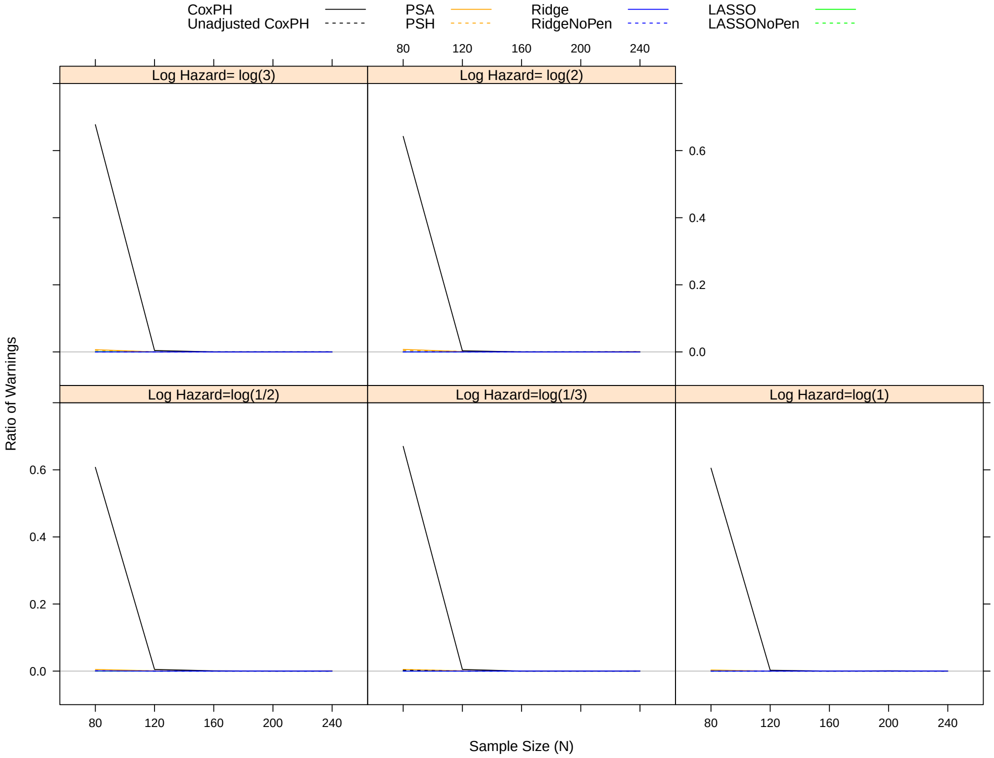


Figure 27: Ratio of Warnings to Number of Simulations (Two Confounder Setup and Exposure Rate=.3)

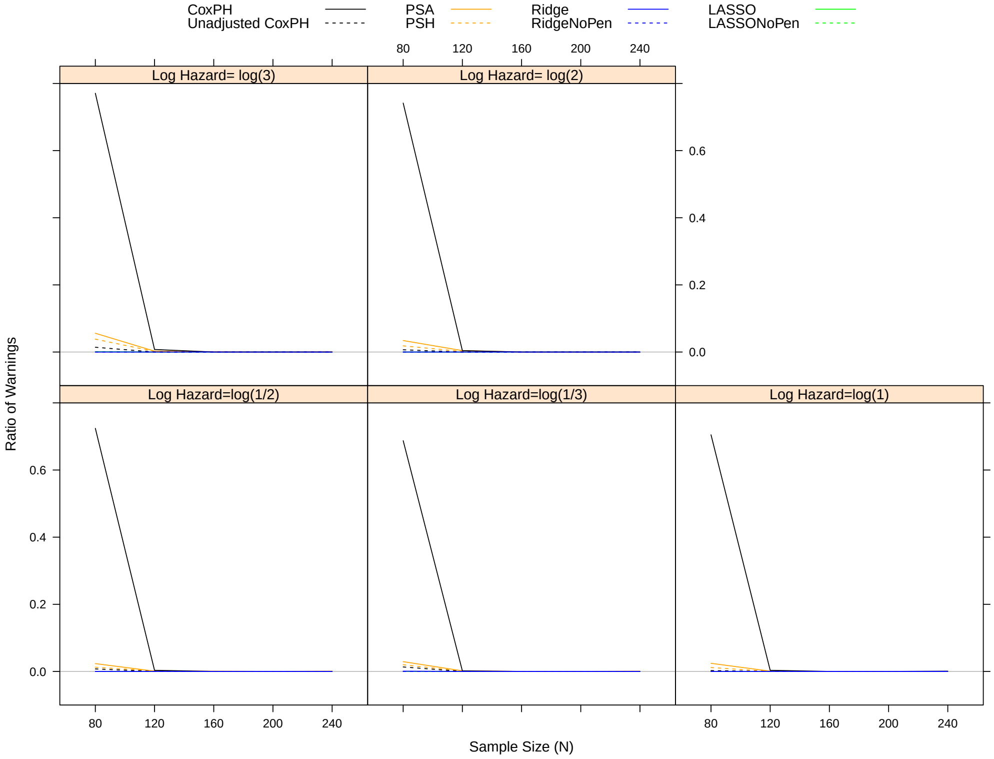
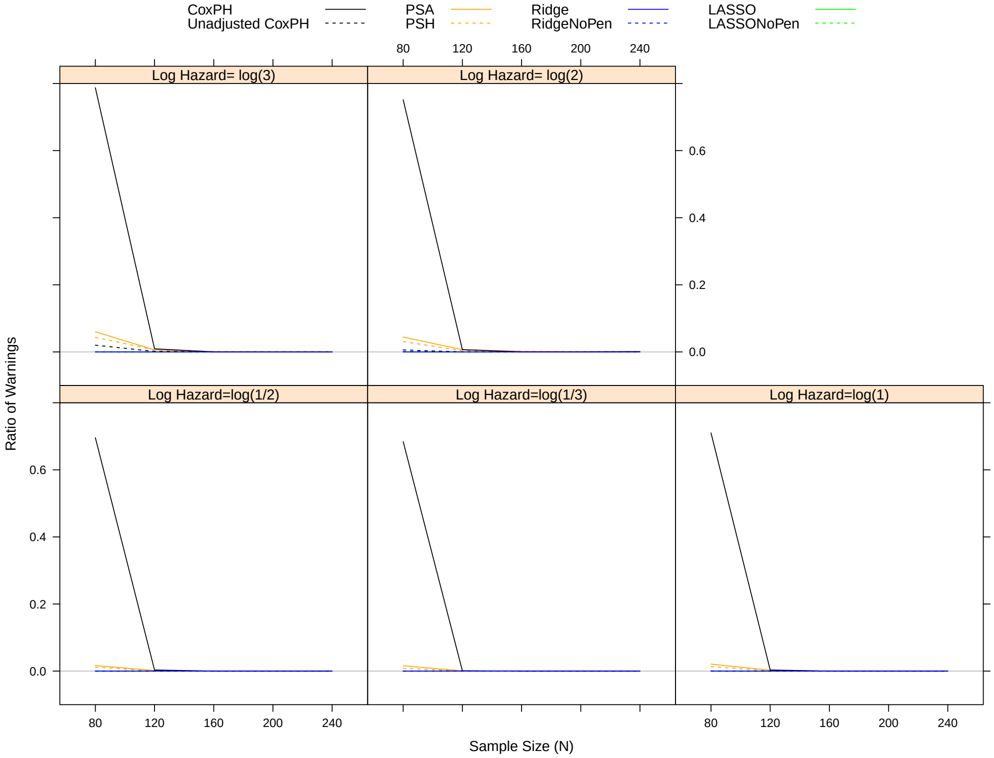


Figure 28: Ratio of Warnings to Number of Simulations (Two Confounder Setup and Exposure Rate=.5)



IV.2 Conclusions

One clear cut conclusion is that for very low EPV, CoxPH has a fairly high chance of producing severely biased and non-significant results, and should not be considered for use when there are fewer than two EPV. RidgeNoPen, LASSONoPen, PSH and PSA all have consistently low MSE and MAE compared to other methods, are less suseptible to monotone likelihood problems and better CI coverage compared to Ridge and LASSO, making them reasonable options for use under the constraint of few EPV. Out of those four methods, when it is believed that there are few large confounding effects, perhaps the safest choices are PSH or PSA which consistently (though not always) have lower MSE and MAE in those circumstances, and have coverage closest to 95%. When there are believed to be many confounders, LASSONoPen and RidgeNoPen seem to be good choices based on MSE and MAE, though the coverage still is not as close to 95% as PSH or PSA

IV.3 Potential Future Work

Many variations exist for the methods selected and evaluated in this paper. For instance, P. Austin tests 8 different variations of just PS methods [1]. Different researchers likely have different prefered methods they would be interested in testing. In particular, one might want to try implementing Heinze's [11], or others's proposed solutions to the monotone likelihood problem. In addition to trying different models and methods, different features of the data can be varied, for example including missing values and the effects of imputation in a low EPV setting. Furthermore, it is necessary to validate these simulation results using real data.

CHAPTER V

Appendix

This appendix contains descriptions and simulation results for secondary evaluation criteria that were not of primary interest in this paper. Note that MAD plots and plots of IQR of bootstrap estimates only contain penalized methods, which are the only methods considered that use the bootstrap.

V.1 Empirical Standard Error (ESE) (Figs. 29, 30, 31, 32)

ESE is calculated by taking the standard deviation of the estimated log-hazards, that is $ESE(\hat{\theta})=\sqrt{\frac{\sum_{i=1}^{m}(\hat{\theta}_i-\bar{\hat{\theta}}_i)^2}{m}}$. The ESE is one standard by which to quantify the variability of estimates. Smaller ESE indicates more precision.

The ESE results are very similar to the model reported standard errors from the section above. Like the standard errors, the ESE has a very similar picture accross scenarios and has the same form in that it starts high and drops towards 0 as the number of EPV increases. The main difference from the model reported standard errors is that, when they are not tied, Ridge seems to frequently have smaller ESEs than those of LASSO (the opposite was true in the previous section).

Figure 29: Empirical Standard Error of Estimates (Many Confounder Setup and Expsoure Rate=.3)

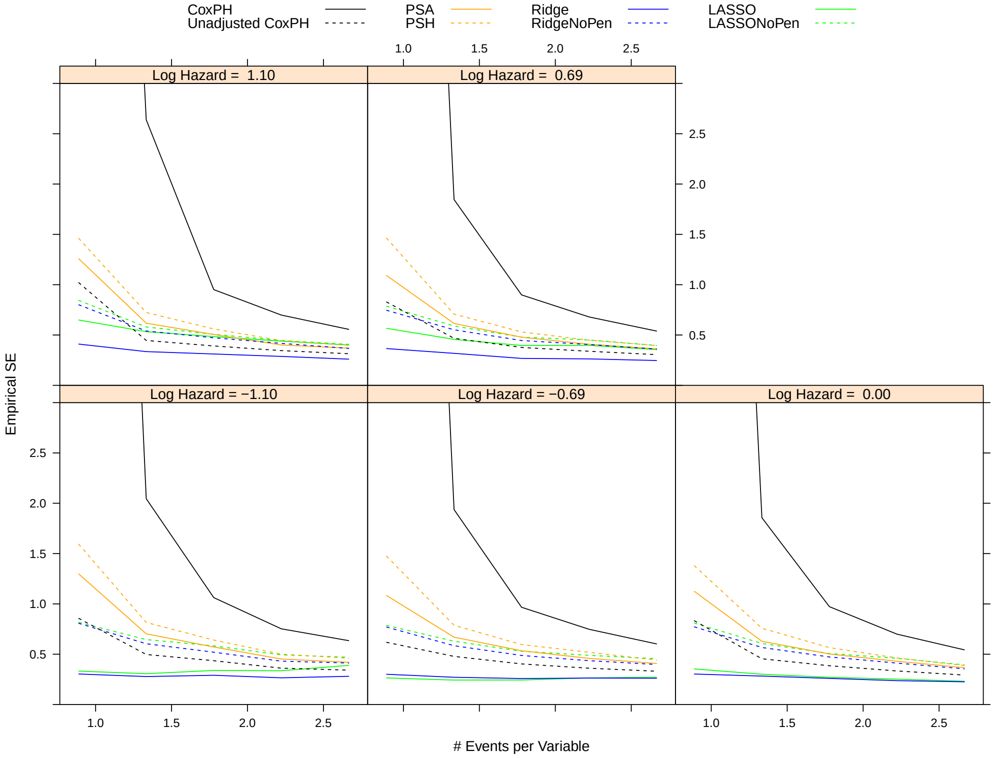


Figure 30: Empirical Standard Error of Estimates (Many Confounder Setup and Expsoure Rate=.3)

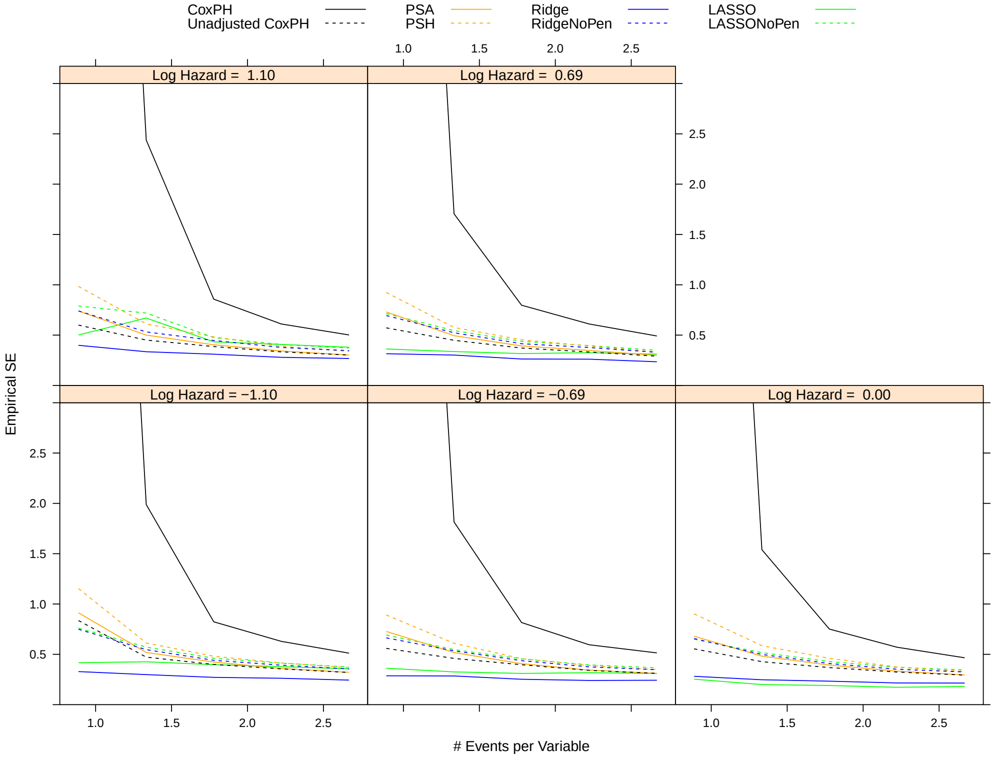


Figure 31: Empirical Standard Error of Estimates (Two Confounder Setup and Expsoure Rate=.3)

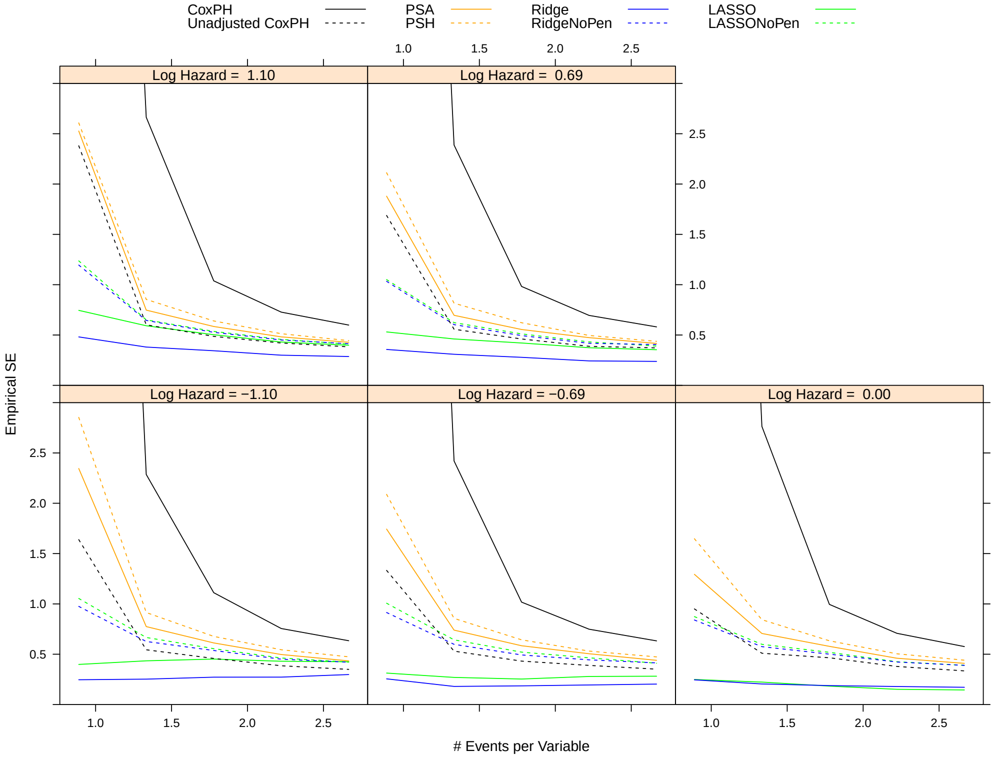
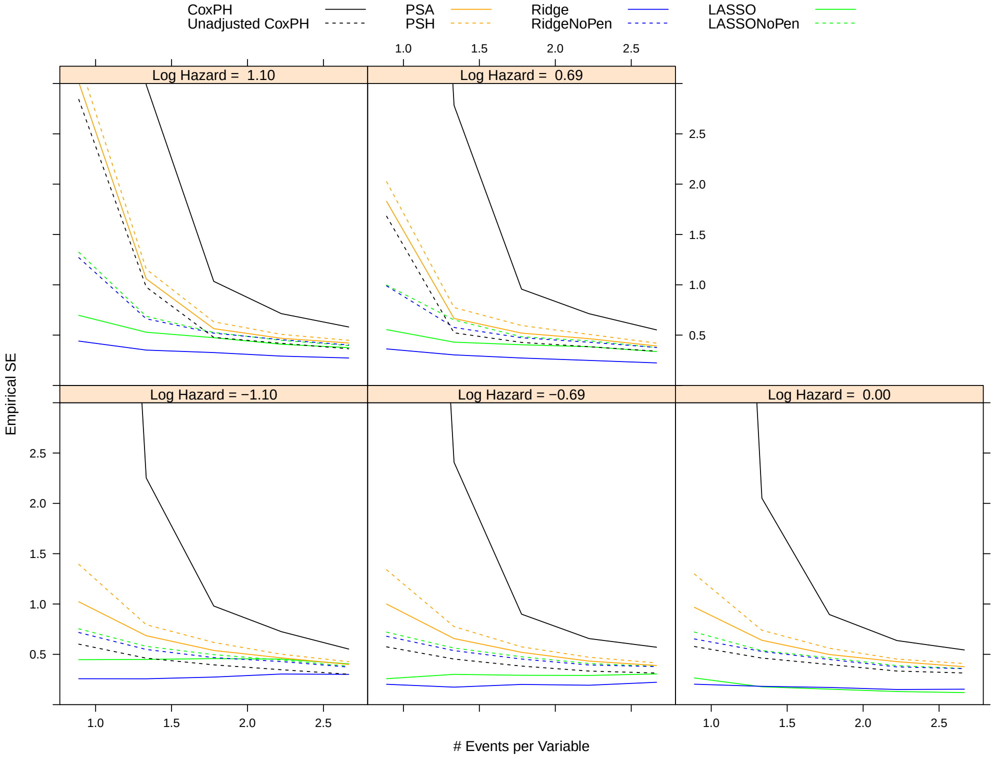


Figure 32: Empirical Standard Error of Estimates (Two Confounder Setup and Expsoure Rate=.5)



V.2 Additional Confidence Interval Figures

In addition to coverage of the CIs, we also looked at the individual bounds of the CIs separately. That is, we report the percent of simulations that have CI lower bounds less than the true log-hazard and seperately the percent of simulations that have CI upper bounds greater than the true log-hazard. Furthermore, we examined the median lengths of the CIs. The mean CI length was not used because of the instability of the mean in relation to outliers.

Figure 33: Probability of Lower Bound of CI Below True Hazard (Many Confounder Setup and Exposure Rate=.3)

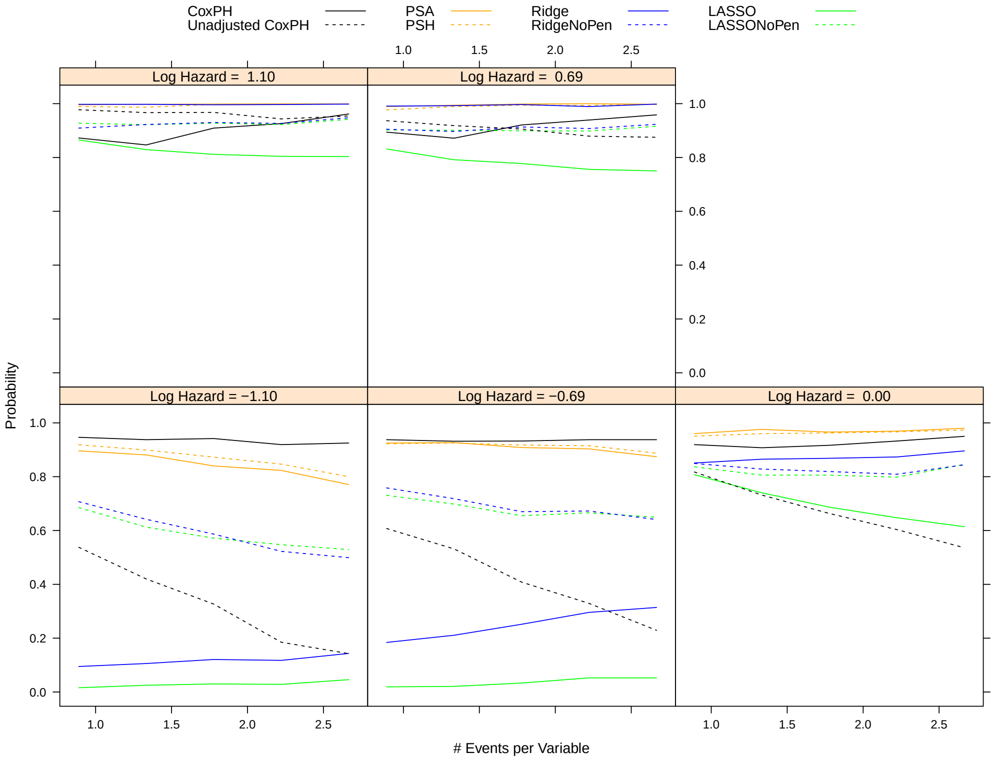


Figure 34: Probability of Lower Bound of CI Below True Hazard (Many Confounder Setup and Exposure Rate=.5)

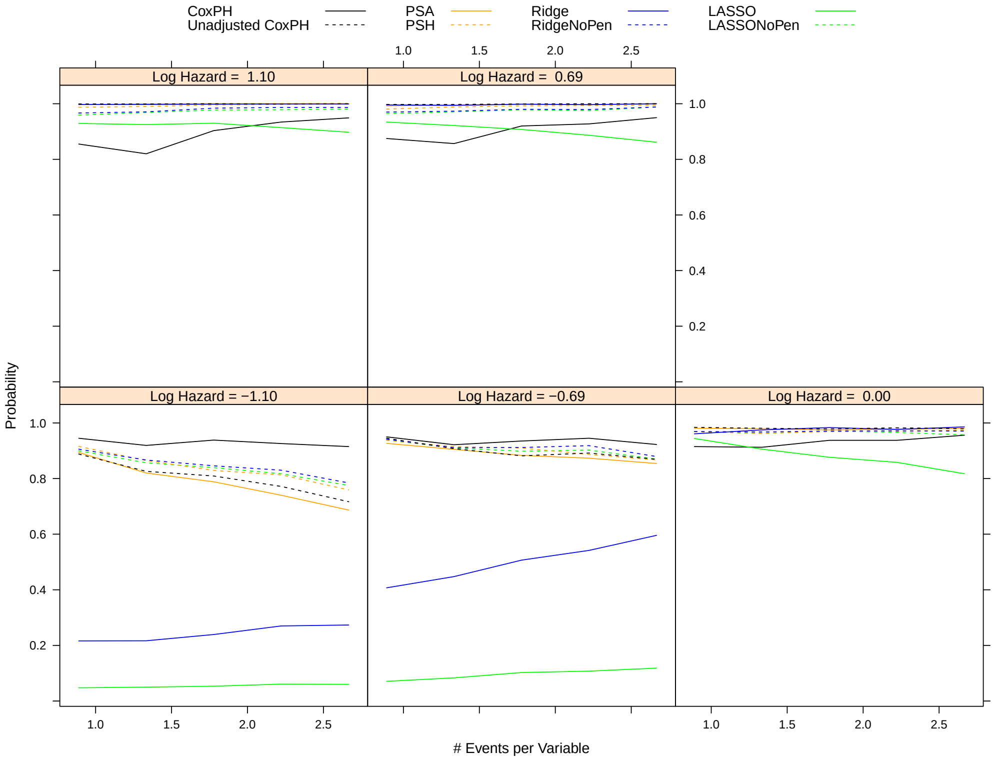


Figure 35: Probability of Lower Bound of CI Below True Hazard (Two Confounder Setup and Exposure Rate=.3)

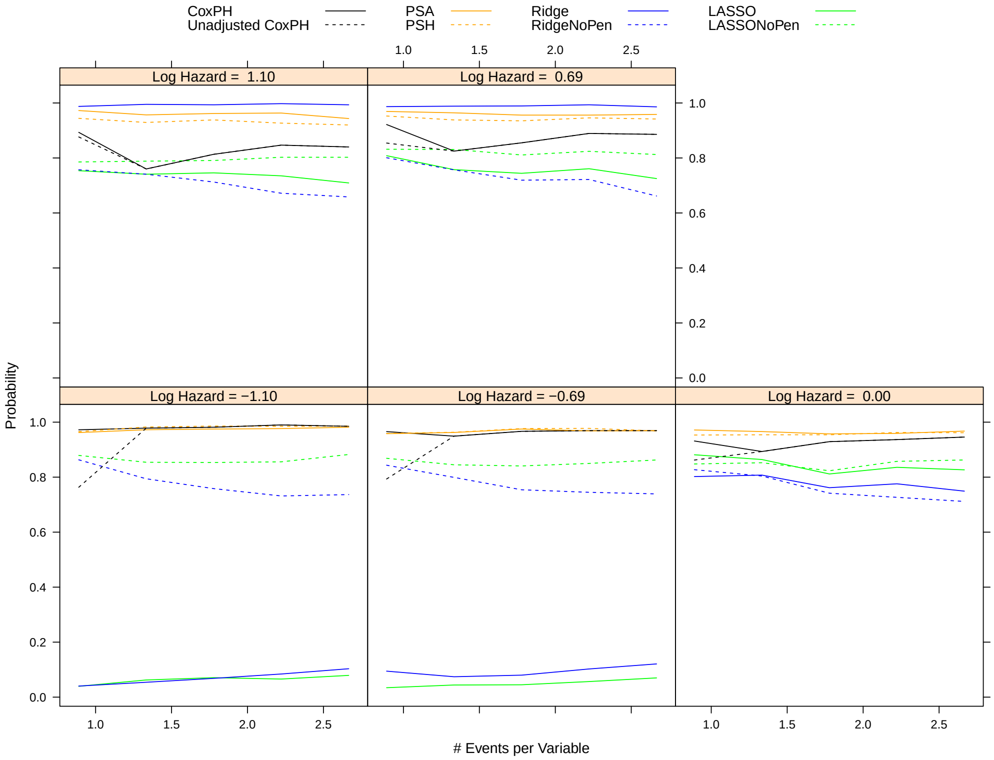


Figure 36: Probability of Lower Bound of CI Below True Hazard (Two Confounder Setup and Exposure Rate=.5)

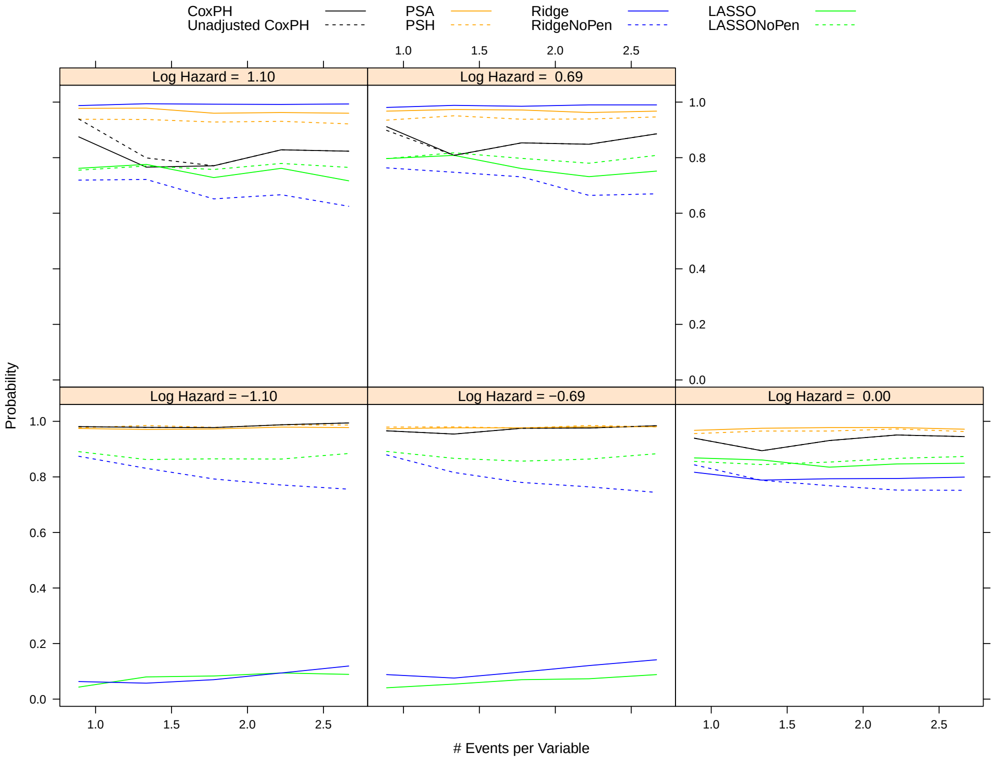


Figure 37: Probability of Upper Bound of CI Above True Hazard (Many Confounder Setup and Exposure Rate=.3)

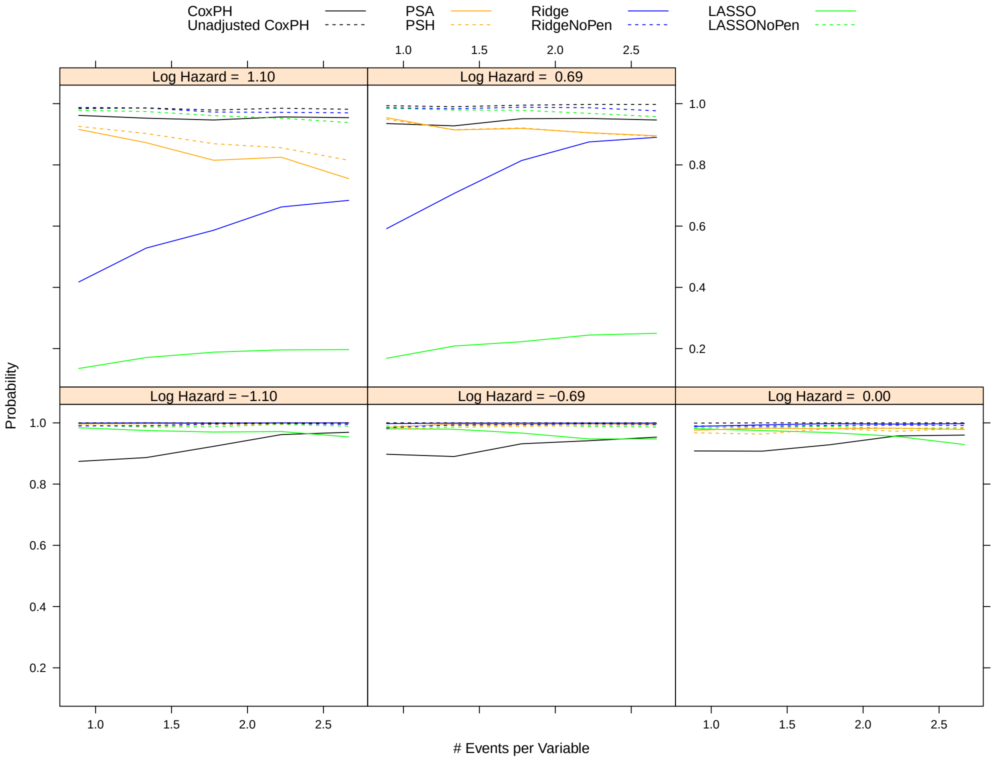


Figure 38: Probability of Upper Bound of CI Above True Hazard (Many Confounder Setup and Exposure Rate=.5)

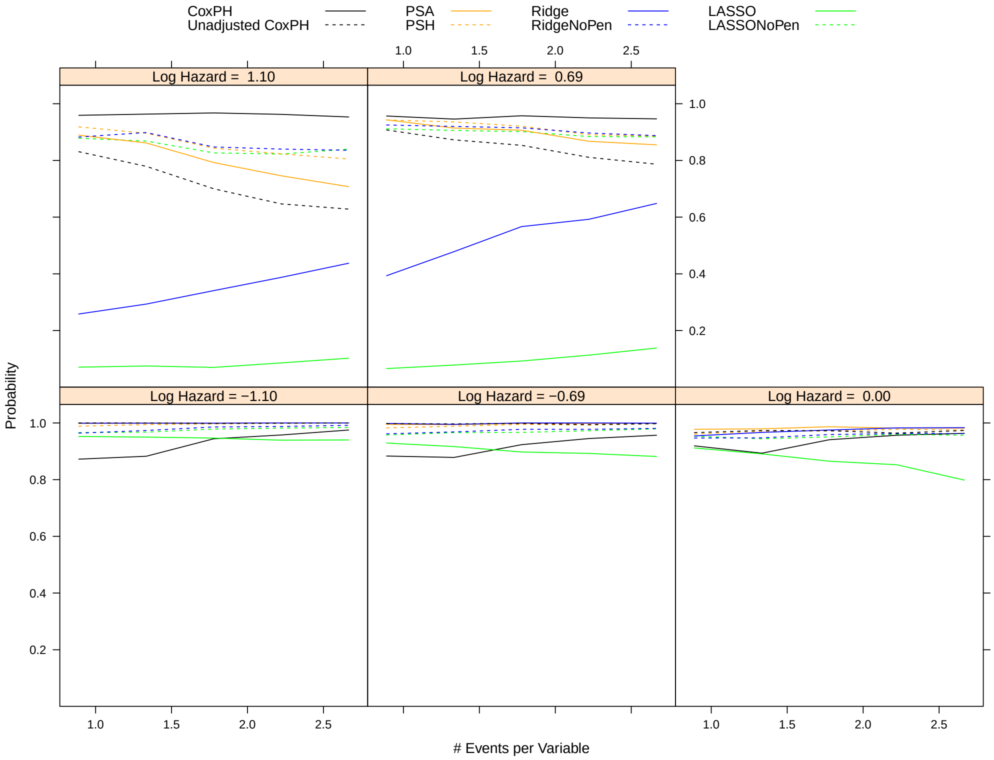


Figure 39: Probability of Upper Bound of CI Above True Hazard (Two Confounder Setup and Exposure Rate=.3)

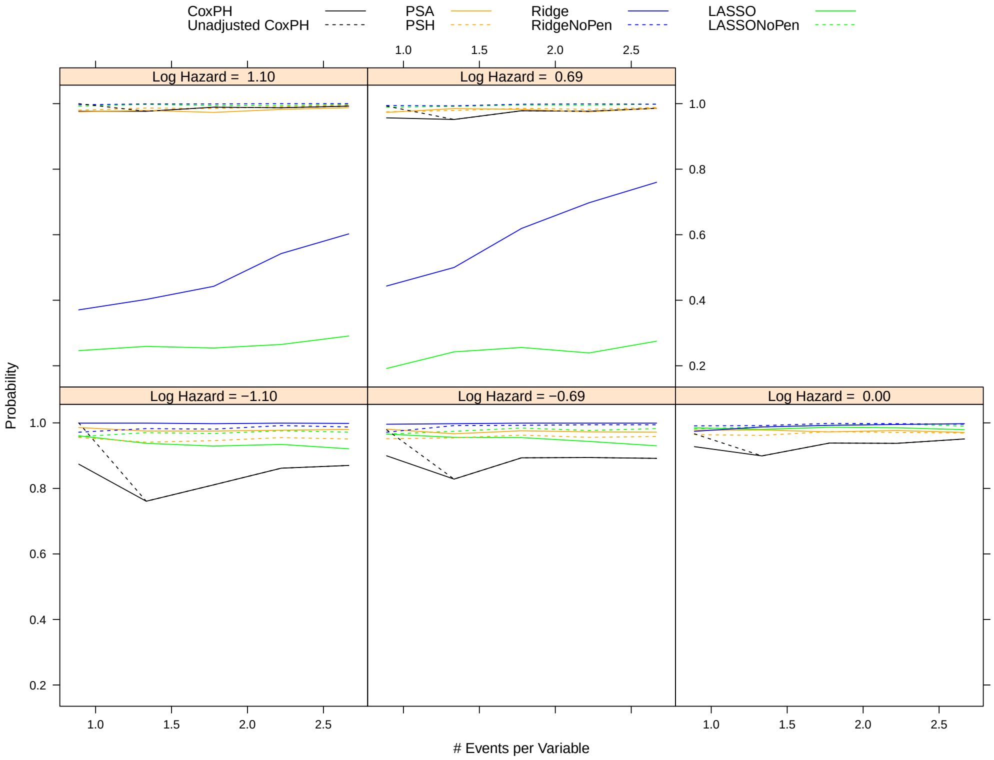


Figure 40: Probability of Upper Bound of CI Above True Hazard (Two Confounder Setup and Exposure Rate=.5)

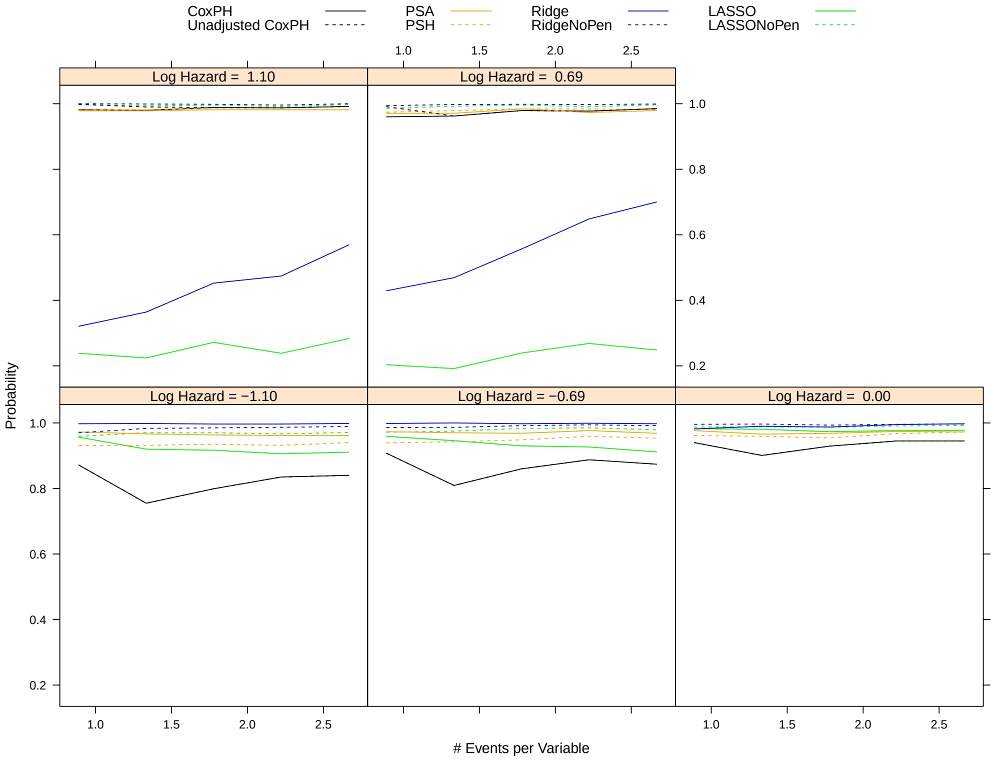


Figure 41: Median Confidence Interval Length (Many Confounder Setup and Exposure Rate=.3)

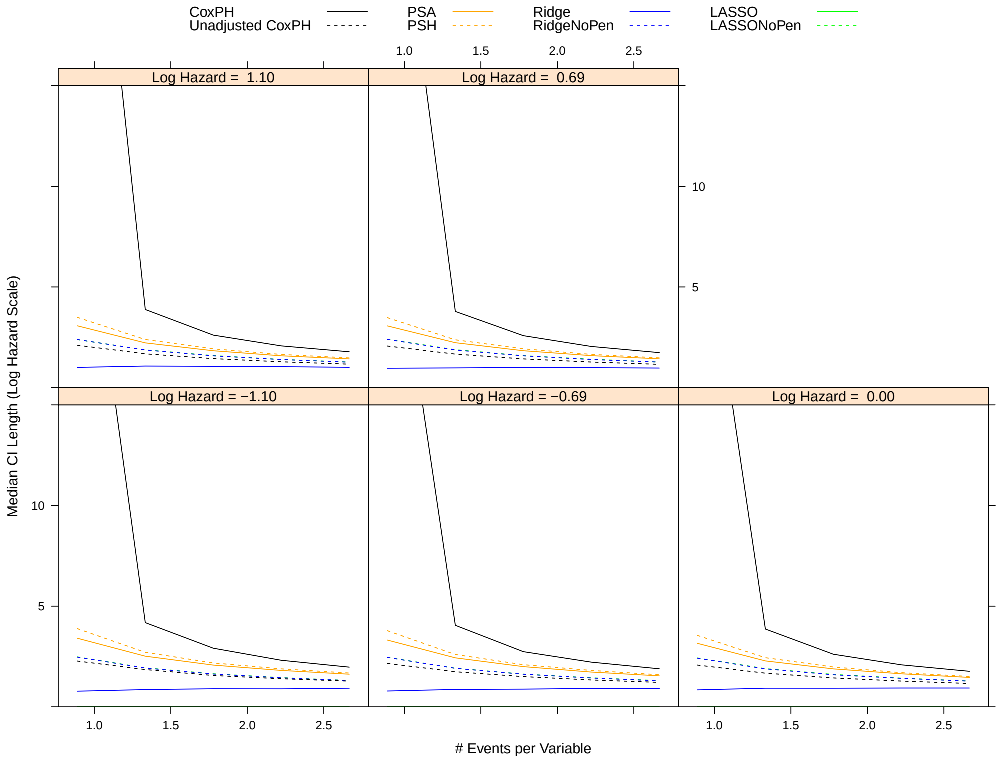


Figure 42: Median Confidence Interval Length (Many Confounder Setup and Exposure Rate=.5)

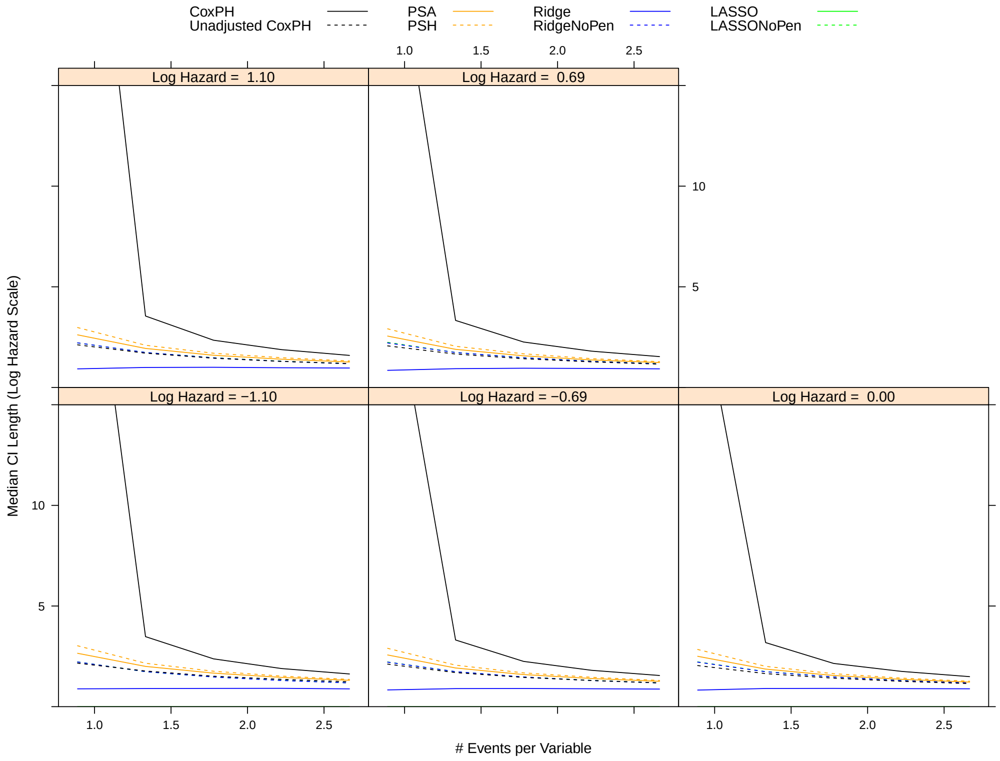


Figure 43: Median Confidence Interval Length (Two Confounder Setup and Exposure Rate=.3)

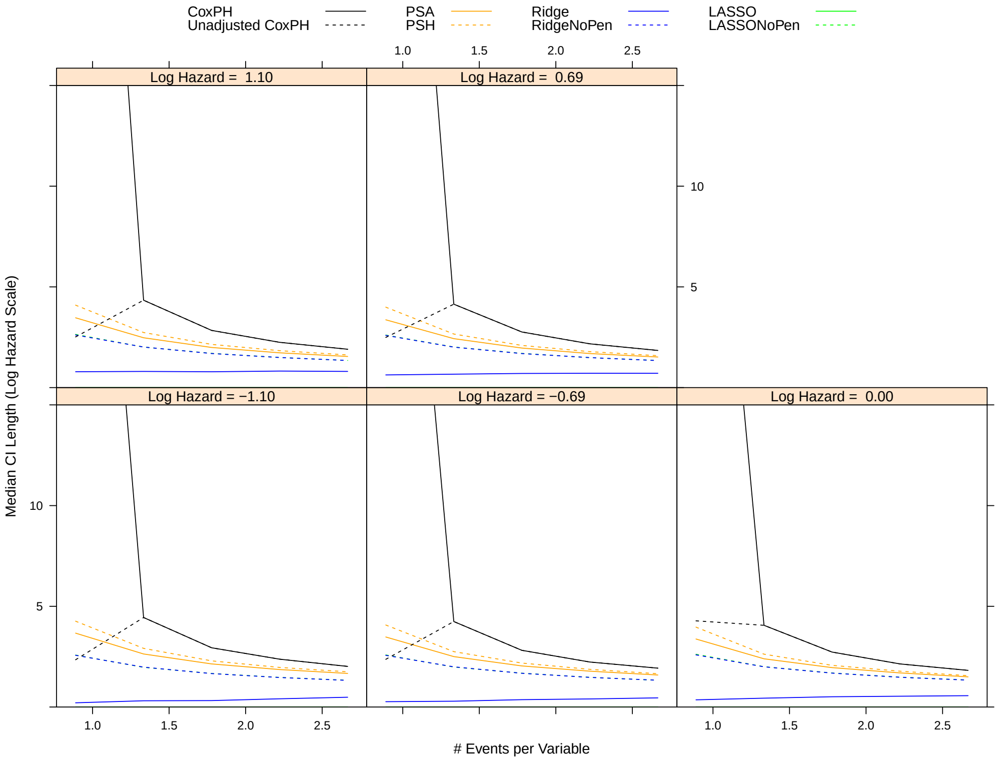
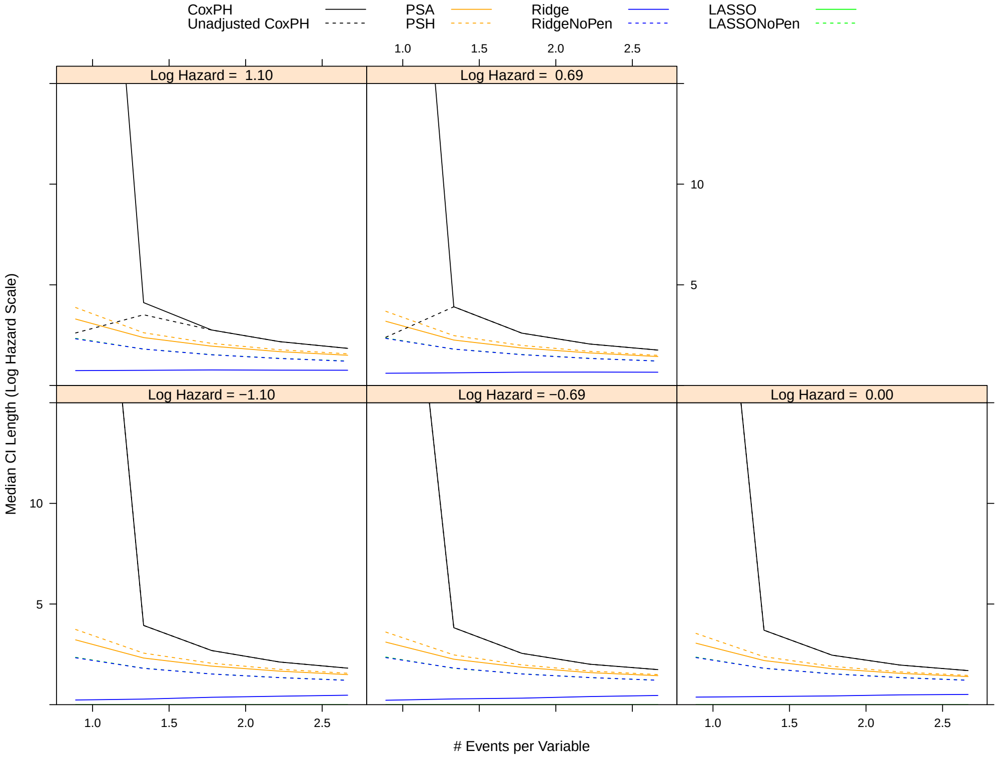


Figure 44: Median Confidence Interval Length (Two Confounder Setup and Exposure Rate=.5)



BIBLIOGRAPHY

[1] AUSTIN, P. C. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in medicine 32*, 16 (2013), 2837–2849.

[2] BENDER, R., AUGUSTIN, T., AND BLETTNER, M. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine 24*, 11 (2005), 1713–1723.

[3] CHARLSON, M. E., POMPEI, P., ALES, K. L., AND MACKENZIE, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases 40*, 5 (1987), 373–383.

[4] CHEN, Q., NIAN, H., ZHU, Y., TALBOT, H. K., AND GRIFFIN, MARIE R., H. F. E. Too many covariates and too few cases? – a comparative study. *Manuscript submitted for publication* (2015).

[5] COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*, 2 (1972), 187–220.

[6] D'HOORE, W., BOUCKAERT, A., AND TILQUIN, C. Practical considerations on the use of the charlson comorbidity index with administrative data bases. *Journal of clinical epidemiology 49*, 12 (1996), 1429–1433.

[7] GAIL, M. H., WIEAND, S., AND PIANTADOSI, S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika 71*, 3 (1984), 431–444.

[8] GREENLAND, S. Interpretation and choice of effect measures in epidemiologic analyses. *American journal of epidemiology 125*, 5 (1987), 761–768.

[9] HARRELL JR, F. E. *Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis*, 2 ed. Springer, 2014.

[10] HARRELL JR, F. E., LEE, K. L., MATCHAR, D. B., AND REICHERT, T. A. Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer treatment reports 69*, 10 (1985), 1071–1077.

[11] HEINZE, G., AND SCHEMPER, M. A solution to the problem of monotone likelihood in cox regression. *Biometrics 57*, 1 (2001), 114–119.

[12] HOERL, A. E., AND KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 1 (1970), 55–67.

[13] KLEIN, J. P., AND MOESCHBERGER, M. L. *Survival analysis: techniques for censored and truncated data*. Springer, 2003.

[14] LUNCEFORD, J. K., AND DAVIDIAN, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine 23*, 19 (2004), 2937–2960.

[15] PEDUZZI, P., CONCATO, J., FEINSTEIN, A. R., AND HOLFORD, T. R. Importance of events per independent variable in proportional hazards regression analysis ii. accuracy and precision of regression estimates. *Journal of clinical epidemiology 48*, 12 (1995), 1503–1510.

[16] POCOCK, S. J., ASSMANN, S. E., ENOS, L. E., AND KASTEN, L. E. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in medicine 21*, 19 (2002), 2917–2930.

[17] ROSENBAUM, P. R., AND RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 1 (1983), 41–55.

[18] ROUSSEEUW, P. J., AND CROUX, C. Alternatives to the median absolute deviation. *Journal of the American Statistical association 88*, 424 (1993), 1273–1283.

[19] SEAL, H. Early uses of graunt's life table. *Journal of the Institute of Actuaries 107*, 04 (1980), 507–511.

[20] SIMON, N., FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R., ET AL. regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software 39*, 5 (2011), 1–13.

[21] TIBSHIRANI, R., ET AL. The lasso method for variable selection in the cox model. *Statistics in medicine 16*, 4 (1997), 385–395.

[22] VERWEIJ, P. J., AND VAN HOUWELINGEN, H. C. Penalized likelihood in cox regression. *Statistics in medicine 13*, 23-24 (1994), 2427–2436.

[23] XU, H., ALDRICH, M. C., CHEN, Q., LIU, H., PETERSON, N. B., DAI, Q., LEVY, M., SHAH, A., HAN, X., RUAN, X., ET AL. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *Journal of the American Medical Informatics Association* (2014).