ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis [Econometrics and Operational Research]

Confidence Intervals in Sparse High-Dimensional Models

Student name: Quinten Blankenborg

Student ID: 509059

Supervisor: A.J. Koning

Second assessor: Max Welz

Date final version: 02-07-2021

Abstract

Many data-types like genomes, pictures and sounds can be characterised by their large number of descriptive attributes. Analysis on these kinds of data generally requires the use of high-dimensional statistical models. For these models, much of the classic statistical theory is inadequate, making this a difficult problem to resolve. Ning and Liu (2017) attempt to tackle this issue and present theoretical derivations for tests and confidence intervals in several kinds of sparse high-dimensional models. In this paper I evaluate their theory on confidence intervals for linear regression and additive hazard models through application and simulation. I find that the theory for linear regressions does not deal well with increasing dimensionality, whereas the theory for additive hazards models provides slightly more consistent results regardless of dimensionality.

Contents

1 Introduction

As time passes and information technology improves, so do our options for analyzing more complex data structures. Due to these developments, many of the relevant current data analysis problems are high-dimensional (Donoho, 2000). Examples of such research include research within evolutionary biology as presented by Collyer et al. (2015) and a multitude of other subjects where objects with a large number of attributes, like images, texts or sounds, are analysed (Francois, 2007).

However, performing proper statistical analysis on high-dimensional data poses multiple computational and theoretical challenges. In certain algorithms for optimization, approximation or numeric integration, high-dimensionality can make the procedures practically infeasible (Donoho, 2000). According to Fan and Li (2000), the field of computational biology occasionally struggles statistically with the nature of their data. They often deal with data regarding a wide array of different genes, of which only a select few are statistically relevant. However, since the number of observations is often much smaller than the number of genes that are investigated, much of the classical theory on testing the significance of parameters is not adequate. In general, for sparse high-dimensional models, the classical statistical methods and tests simply do not provide sufficient functionality. This particular issue is addressed by Ning and Liu (2017), who provide a theoretical framework for hypothesis testing and constructing confidence intervals for parameters of interest, for sparse high-dimensional models.

In this article, I will evaluate the findings and derivations of Ning and Liu (2017), specifically their findings on confidence interval in linear regression and additive hazards models. First of all, their findings regarding confidence intervals for the two aforementioned models is applied on actual datasets. For the linear regression model the dataset concerns the fraction of votes incumbent parties receive in elections. This variable is then regressed on several macro-economic and/or societal indicators for how well any given country performs in respect to others. For the additive hazards model, the chosen dataset contains information on the survival time of seedlings under different circumstances. Secondly, the derived theory is evaluated through a simulation study. Artificial datasets are constructed to evaluate whether the proposed methods provide correct coverage for confidence intervals for the two models in question. This subsequently leads to the following research question: To what extent is the theory provided by Ning and Liu (2017), on confidence intervals for linear regression and additive hazards models, valid and thus useful in practice?

The paper proceeds as follows: In section two, the methodology will be discussed in-depth, in section three the results are presented, the fourth section concerns the conclusion, containing a reflection on the methods and results and the fifth and last section presents a discussion on shortcomings, future improvements and/or additions.

2 Methodology

The methodology can be divided into two distinct parts. The first part, consisting of sections 2.1, 2.2 and 2.3, gives an overview of the relevant theory and methods presented by Ning and Liu (2017) and thus all methods presented in this part should be accredited to them accordingly. This part specifically describes the construction of a decorrelated score function and the subsequent construction of a confidence interval for a univariate variable in a high-dimensional model, using the said score function. The second part, sections 2.4 and 2.5, provide the steps taken to apply and validate the theory discussed in the first part. The validation is done through a simulation study, in which the coverage of the derived confidence interval is examined. All the programming done for these purposes, is written and executed in MATLAB.

All methods concern statistical analysis on models with the d-dimensional parameter vector (θ, γ^T), in which θ is a univariate parameter and γ is a (d-1)-dimensional parameter vector. Let β be defined as $\beta = (\theta, \gamma^T)^T$ For any of the models used, β is estimated by:

$$\hat{\beta} = argmin_{\beta} l(\beta) + P_{\lambda}(\beta) \tag{1}$$

where hereafter any parameter that is adorned with a hat-symbol, is an estimator of that parameter. Furthermore $l(\beta)$ is a loss function and $P_{\lambda}(\beta)$ is a penalty function with tuning parameter λ. For the application and simulation, no penalty function is actually implemented as to not influence the results through the choice of a particular penalty function.

2.1 Confidence Intervals

For both linear regression and additive hazards models a decorrelated score function $S(\theta, \gamma)$ is estimated. The exact procedure for this estimation is described separately per model in the subsequent sections. Using the decorrelated score function, theory can be derived for the construction of confidence intervals for the univariate parameter θ. For this purpose we first define the one step estimator $\tilde{\theta}$ as follows:

$$\tilde{\theta} = \hat{\theta} - \hat{S}(\hat{\beta})\hat{\mathbf{I}}_{\theta|\gamma} \quad with \quad \hat{\mathbf{I}}_{\theta|\gamma} = \nabla^2_{\theta\theta} l(\hat{\beta}) - \hat{\mathbf{w}}^T \nabla^2_{\gamma\theta} l(\hat{\beta}) \tag{2}$$

Let $\eta_i(n)$ be a sequence converging to zero as n approaches infinity. If $(\eta_1(n) + \eta_2(n)\sqrt{log(d)} = o(1)$, and $\hat{\mathbf{I}}_{\theta|\gamma}$ is a consistent estimator of the true parameter $\mathbf{I}^*_{\theta|\gamma}$ and $\mathbf{I}^*_{\theta|\gamma} \geq C$ for some constant $C > 0$, then under certain assumptions, which hold for both linear regression and additive hazards models, $\tilde{\theta}$ is asymptotically normal with mean θ^*, the true value of θ as hereafter indicated by the addition of the * symbol. Thus, this quantity can be used to construct confidence intervals for θ. The mentioned assumptions and the proof that they hold in the discussed models, can be found in the paper by Ning and Liu (2017).

2.2 Linear Regression

Consider a linear regression of the following form:

$$Y_i = \theta^* Z_i + \gamma^{*T} \mathbf{X}_i + \epsilon_i \tag{3}$$

where Z_i is a univariate variable and $\mathbf{X_i}$ is a (d-1)-dimensional variable-vector. Furthermore it is assumed to be true that $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = \sigma^2$. The loss function for this model that is employed in equation (1), is the negative log-likelihood.

2.2.1 Decorrelated Score Function

To start, the decorrelated score function is constructed, such that it is applicable for statistical derivations in high-dimensional linear regression models. For linear regression models, this score function takes the following form:

$$S(\theta, \gamma) = \nabla_\theta l(\theta, \gamma) - \mathbf{w}^T \nabla_\gamma l(\theta, \gamma) \tag{4}$$

Lastly, $\mathbf{w}^T$ is defined as follows:

$$\mathbf{w}^T = \mathbf{I}_{\gamma\gamma}^{-1} \mathbf{I}_{\gamma\theta} \tag{5}$$

where $\mathbf{I}_{\gamma\gamma}^{-1}$ and $\mathbf{I}_{\gamma\theta}$ are partitions of the Fisher information matrix $\mathbf{I} = E_\beta(\nabla^2 l(\beta))$. The decorrelated score function is decorrelated in the sense that the correlation between itself and $\nabla_\gamma l(\beta)$ is equal to zero.

Given the tuning parameter λ', the score function can be estimated through a simple algorithm.
Step 1: Calculate $\hat{\beta}$ through equation (1).
Step 2: Estimate $\mathbf{w}$ through the following optimization problem:

$$\hat{\mathbf{w}} = argmin_{\mathbf{w}} \sum_{i=1}^{d} |\mathbf{w}_i| \quad s.t. \;\; max_{1 \le i \le d} \{|\alpha_i| : \alpha_i \in (\nabla^2_{\theta\gamma} l(\hat{\beta}) - \mathbf{w}^T \nabla^2_{\gamma\gamma} l(\hat{\beta}))\} \le \lambda' \tag{6}$$

Step 3: Estimate the decorrelated score function:

$$\hat{S}(\hat{\theta}, \hat{\gamma}) = \nabla_\theta l(\hat{\theta}, \hat{\gamma}) - \hat{\mathbf{w}}^T \nabla_\gamma l(\hat{\theta}, \hat{\gamma}) \tag{7}$$

For a linear regression model of the discussed form, if $n^{-1/2}(s' \vee s^*) log(d) = o(1)$, where s' and s^* are $|supp(\mathbf{w}^*)|$ and $|supp(\beta^*)|$ respectively and if $C \le \lambda/\lambda' \le C'$ and $C \le \lambda'/\sqrt{log(d)/n} \le C'$ for constants $C, C' > 0$, then $n^{-1/2}(\tilde{\theta} - \theta^*)\hat{\mathbf{I}^{\mathbf{1/2}}_{\theta|\gamma}}$ converges weakly to standard normal. Considering this result a $(1-\alpha) \times 100\%$ confidence interval for θ^* can be derived as: $[\tilde{\theta} - n^{-1/2}\Phi^{-1}(1-\alpha/2)\hat{\mathbf{I}}^{-1/2}_{\theta|\gamma}, \tilde{\theta} + n^{-1/2}\Phi^{-1}(1-\alpha/2)\hat{\mathbf{I}}^{-1/2}_{\theta|\gamma}]$.

2.3 Additive Hazards Model

Let us first establish the general framework of an additive hazards model. Let T and R be the time to the studied event and the right side censoring time respectively. Next, let $\mathbf{Q}(t) = (Z(t), \mathbf{X}^T(t))^T$ be a d-dimensional time-dependent covariate vector at time t. For the application and simulations in this paper, data is used or simulated such that the data is actually constant over time for any given observation. This choice is made as doing so notably eases the simulation procedure and improves computational feasibility. Lastly, $W = min\{T, R\}$ and $\Delta = I(T \leq R)$. With this notation, the hazard function of the model can be defined as:

$$\lambda(t|\mathbf{Q}(t)) = \lambda_0(t) + (\theta^*, \gamma^{*T})\mathbf{Q}(t) \tag{8}$$

where $\lambda_0(t)$ is an unknown baseline hazard function, which for the application and simulations will simply be set as a constant or zero.

Next, define some additional notations and identities: $\mathbf{v}^{\otimes 0} = 1$, $\mathbf{v}^{\otimes 1} = \mathbf{v}$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$. Furthermore, $N_i(t) = I(W_i \leq t, \Delta_i = 1)$, $Y_i(t) = I(W_i \geq t)$ and $\bar{\mathbf{Q}}(t) = \sum_{i=1}^n Y_i(t)\mathbf{Q}_i(t) / \sum_{i=1}^n Y_i(t)$. Using this notation, the following identities can be defined:

$$\begin{aligned} \mathbf{b} &= n^{-1} \sum_{i=1}^{n} \int_0^\tau (\mathbf{Q}_i(t) - \bar{\mathbf{Q}}(t)) dN_i(t) \\ \mathbf{V} &= n^{-1} \sum_{i=1}^{n} \int_0^\tau Y_i(t)(\mathbf{Q}_i(t) - \bar{\mathbf{Q}}(t))^{\otimes 2} dt \end{aligned} \tag{9}$$

where τ is the last time t included in the sample. Given these definitions, the loss function used for this model is formulated as:

$$l(\beta) = \frac{1}{2}\beta^T \mathbf{V}\beta - \mathbf{b}^T\beta \tag{10}$$

Next define $s^{(k)}(t) = E(Y_i(t)Q_i(t)^{\otimes k}), \quad for \;\; k = 0, 1, 2.$ Using this definition, construct the following equations:

$$\begin{aligned} \mathbf{V}^* &= E(\int_0^\tau Y_i(t)(\mathbf{Q}_i(t) - \frac{s^1(t)}{s^0(t)})^{\otimes 2} dt) \\ \mathbf{W}^* &= E(\int_0^\tau (\mathbf{Q}_i(t) - \frac{s^1(t)}{s^0(t)}^{\otimes 2}) dN_i(t)) \end{aligned} \tag{11}$$

Lastly a couple of assumptions need to be made for the model: $\int_0^\tau \lambda_0(t)dt < \infty$, $P(Y_i(t) = 1) \geq 0$, $max_{1 \leq i \leq d} |\mathbf{Q}_i(t)| \leq A$, $|\mathbf{w}^{*T}\mathbf{X}_i(t)| \leq K$, $\lambda_{min}(\mathbf{V}^*) \geq \kappa^2$ and $\lambda_{min}(\mathbf{W}^*) \geq \kappa^2$, where for some constants $A, K, \kappa > 0$, and where $\lambda_{min}(D)$ is the smallest eigenvalue of matrix D.

Using this framework the score function takes the form:

$$\hat{S}(\hat{\beta}) = \hat{\mathbf{v}}^T(\mathbf{V}\hat{\beta} - \mathbf{b}) \tag{12}$$

where $\mathbf{v} = (1, -\hat{\mathbf{w}}^T)^T$ with $\hat{\mathbf{w}}$ being computed through the following optimization:

$$\begin{aligned} &\hat{\mathbf{w}} = argmin_{\mathbf{w}} \sum_{i=1}^{d} |\mathbf{w}| \\ &s.t. \;\; max_{1 \leq i \leq d}\{|\alpha_i| : \alpha_i \in (\mathbf{V}_{\theta\gamma} - \mathbf{w}^T\mathbf{V}_{\gamma\gamma})\} \leq \lambda' \end{aligned} \tag{13}$$

where $\mathbf{V}_{\theta\gamma}$ and $\mathbf{V}_{\gamma\gamma}$ are partitions of $\mathbf{V}$.

Now, let $\hat{\mathbf{W}} = n^{-1}\sum_{i=0}^{n}\int_0^\tau (\mathbf{Q}_i(t) - \bar{\mathbf{Q}}(t))^{\otimes 2} dN_i(t)$ be a consistent estimator for $\mathbf{W}^*$. If $n^{-1/2}(s' \vee s^*)log(d) = o(1)$, where s' and s^* are $|supp(\mathbf{w}^*)|$ and $|supp(\beta^*)|$ respectively and if $C \leq \lambda/\lambda' \leq C'$ and $C \leq \lambda'/\sqrt{log(d)/n} \leq C'$ for constants $C, C' > 0$, then $n^{1/2}(\tilde{\theta} - \theta^*)\hat{\mathbf{I}}_{\theta|\gamma}\hat{\sigma}_s^{-1/2}$ converges weakly to standard normal, where $\hat{\mathbf{I}}_{\theta|\gamma} = \mathbf{V}_{\theta\theta} - \hat{\mathbf{w}}^T\mathbf{V}_{\gamma\theta}$ and where $\hat{\sigma}_s = \hat{\mathbf{v}}^T\hat{\mathbf{W}}\hat{\mathbf{v}}$ is an estimator for the asymptotic variance σ_s^*. Following this result, a $(1-\alpha) \times 100\%$ confidence interval for θ for an additive hazards model can be constructed as: $[\tilde{\theta} - n^{-1/2}\Phi^{-1}(1-\alpha/2)\hat{\mathbf{I}}_{\theta|\gamma}^{-1}\hat{\sigma}_s^{1/2}, \tilde{\theta} + n^{-1/2}\Phi^{-1}(1-\alpha/2)\hat{\mathbf{I}}_{\theta|\gamma}^{-1}\hat{\sigma}_s^{1/2}]$.

2.4 Application

The first step in testing the validity of the derived statistical theory is done through application on actual datasets. Due to the different nature of the two considered models, both models will be applied on a separate dataset, both with a ninety-five percent confidence interval. In terms of the tuning parameter λ', it will take on a value of 0.5 and one for the linear regression and additive hazards models respectively.

2.4.1 Linear Regression

The dataset that is used for the linear regression part of the application, is a dataset concerning voting behaviour under different economic circumstances (Harvard Dataverse, 2018). The dataset originally is presented as replication data for the results found in a paper by Arel-Bundock et al. (2019), an article that expands on earlier results given by Kayser and Peress (2012) and Aytaç (2017), but in this paper the dataset is used to demonstrate and possibly give an indication of the validity of the previously discussed statistical methods. The following list presents and elaborates on the variables from the dataset that are used for this purpose:

- *inc_vote* is the dependent variable and represents the fraction of votes the party that previously provided the previous president/prime-minister received in the subsequent elections.
- *vote_prev* is the fraction of votes the incumbent party received in the previous elections.
- *growth_year* is the growth of the country's real GDP during the election year.

- *enp*, or 'effective number of parties', is a measure first introduced by Laakso and Taagepera (1979) to indicate the extent of fragmentation within a political system.
- *edu* is the average number of years of schooling the adult population in a country has received at the time of the elections.
- *log(inc)* is the logarithm of the GDP per capita for any given country during their election year.
- *trade* is a measure of how reliant a country's economy is on trade at the time of the elections.

Using these variables, the statistical methods are demonstrated assuming the following model:

$$inc\_vote_i = \theta growth\_year_i + \gamma_1 vote\_prev_i + \gamma_2 enp_i + \gamma_3 edu_i + \gamma_4 log(inc_i) + \gamma_5 trade_i + \epsilon_i \tag{14}$$

As is usually done, the error terms ϵ_i are assumed to be normally distributed with mean zero and variance σ^2. Moreover, the error terms are assumed not to be serially correlated.

2.4.2 Additive Hazards Model

To demonstrate the proposed theory on additive hazards models, the methodology is applied on a dataset regarding seedling survival data (Pillay et al., 2018). The event T is the event in which a seedling dies. The study starts on the twelfth of august in 2014, ends on the fourth of November of the same year and contains a total of 2069 observations. W is subsequently defined as the minimum of either the time that the seedling survived or the right censoring time for that seedling. The following is a list of the included covariates, the first of which is the parameter of interest for which θ is to be estimated.

- *canopy_cov* is the proportion of canopy cover available for the seedling. In other words, the fraction of a seedling's vertical projection that is covered by other vegetation.
- *total_dens_seedtrap* is a measure of the average seed density in any given seed's vicinity.
- *consp* denotes the number of conspecific seeds located on the same plot as the seed in question.

With the inclusion of these variables the hazard function is assumed to take the following form:

$$\lambda(t|\mathbf{Q}(t)) = \theta canopy\_cov + \gamma_1 total\_dens\_seedtrap + \gamma_2 consp \tag{15}$$

where the hazard function is assumed to be continuous. Furthermore, for this dataset, the covariate matrix Q is constant over time per observation. This means that several equations can be simplified in the procedure, thus easing computational load. First of all, $\bar{\mathbf{Q}}(t)$ can be redefined as follows:

$$\bar{\mathbf{Q}}(t) = \sum_{i=1}^{n} Y_i(t)\mathbf{Q}_i / \sum_{i=1}^{n} Y_i(t) \tag{16}$$

Recall that $Y_i(t)$ is an indicator for whether an observation has survived up and till time t. Now, any observation is present in the sample for any given time A_i. In what part of the studied time period this A_i takes place is completely arbitrary. In other words, moving the observation forward or backwards should not change analysis outcomes as long as it does not change A_i through censoring. Thus, a more computationally efficient proxy for $\bar{\mathbf{Q}}(t)$ would be a time independent weighted average, weighted on A_i:

$$\bar{\mathbf{Q}} = \sum_{i=1}^{n} A_i \mathbf{Q}_i / \sum_{i=1}^{n} A_i \tag{17}$$

With this definition, the identities in (9) simplify into the following:

$$\begin{aligned} \mathbf{b} &= n^{-1} \sum_{i=1}^{n} \Delta_i (\mathbf{Q}_i - \bar{\mathbf{Q}}) \\ \mathbf{V} &= n^{-1} \sum_{i=1}^{n} A_i (\mathbf{Q}_i - \bar{\mathbf{Q}})^{\otimes 2} \end{aligned} \tag{18}$$

The old identities are substituted for its newly derived counterparts and the methodology proceeds the same way as it would have otherwise.

2.5 Simulation

The second step in evaluating the validity of the presented statistical theory, is done through a simulation study. For both of the procedures a thousand models are simulated each time. Each model contains one hundred observations with ten, twenty-five, fifty, seventy-five, ninety or a hundred parameters to give an indication of the methodology's performance as the dimensionality increases. Each model is simulated such that it adheres to all the necessary assumptions. After a model is simulated, the first parameter is assigned to be the θ parameter. The methodology is then applied to the model after which the program validates whether or not the true parameter value falls within the derived confidence bounds. After a thousand iterations this will produce a hit-rate of some number out of a thousand. This procedure is repeated for ninety, ninety-five and ninety-nine percent confidence intervals. Afterwards the theory can be evaluated by comparing the hit-rate, the rate at which the true parameter θ^* falls within the derived confidence interval, with the supposed coverage of the computed confidence interval. If these values are close to one another, it would suggest the theory to be (partially) valid. If it does not, the theory might not hold in practice, or at least not completely. Lastly this procedure is then repeated for different values for the tuning parameter λ' to study the influence the choice of this parameter has on the results. The theory's performance will as such be evaluated for three cases, where λ' takes on the value of either 0.5, one or two.

2.5.1 Linear Regression

To simulate a linear regression model, first the independent variables and error terms are drawn from a normal distribution both with means equal to zero and standard deviations of five and one respectively.

Subsequently the corresponding parameters are drawn from a normal distribution with mean zero and a standard deviation equal to five. Using these simulated quantities the dependent variable is computed for each simulated observation through the following formula:

$$Y_i = \theta^* Z_i + \gamma^{*T} X_i + \epsilon_i \tag{19}$$

Now, the general methodology to analyse a singular model requires two minimisations and although Matlab has several methods for this purpose, it is still rather taxing computationally, which is not desired when simulating a thousand models at a time. However, because of the choice of excluding a penalty function, equation (1) boils down to minimising just the negative log likelihood, for which the estimated beta can be computed analytically as follows (Heij et al., 2004):

$$\hat{\beta}_{MLE} = ((Z, X)^T (Z, X))^{-1} (Z, X)^T Y \tag{20}$$

If possible, doing things analytically is thus preferred over a less efficient numerical approach. So additionally to doing this minimization, the computing of gradients and Hessians is also done analytically to further streamline the process.

2.5.2 Additive Hazards Model

To properly simulate data for an additive hazards model, we will first need a derivation of the inverse cumulative density function $F^{-1}(t)$ from the hazard function. For this derivation we start with some mathematical relations that hold true for continuous hazard models (Kleinbaum and Klein, 2012):

$$\begin{aligned} \Lambda(t) &= \int_0^t \lambda(s) ds \\ S(t) &= exp(-\Lambda(t)) \\ F(t) &= 1 - S(t) \end{aligned} \tag{21}$$

The models are simulated using $\lambda(t|\mathbf{Q}(t)) = (\theta^*, \gamma^{*T})\mathbf{Q}(t)$ as the hazard function. Seeing as the simulated independent variables are going to be constant over time, $\lambda(t|\mathbf{Q}(t)) = \lambda_c$, is constant over time and the following holds:

$$\Lambda(t) = t\lambda_c \tag{22}$$

Given these mathematical relations and results, we can subsequently continue the derivations:

$$\begin{aligned} S(t) &= exp(-t\lambda_c) \\ F(t) &= 1 - exp(-t\lambda_c) \end{aligned} \tag{23}$$

Given this cumulative density function $F(t)$, the inverse cumulative density function takes the following form:

$$F^{-1}(t) = \frac{ln(1-t)}{-\lambda_c} \tag{24}$$

This inverse can be easily proven to be correct through some simple computations:

$$\begin{aligned} F(F^{-1}(t)) &= 1 - exp(-(\frac{ln(1-t)}{-\lambda_c})\lambda_c) \\ F(F^{-1}(t)) &= 1 - exp(ln(1-t)) \\ F(F^{-1}(t)) &= 1 - (1-t) \\ F(F^{-1}(t)) &= t \end{aligned} \tag{25}$$

This derivation can now be used to actually generate usable data. The independent variables and parameters are randomly drawn from a normal distribution both with mean 0.5 and standard deviations of 0.5 and 0.25 respectively. These values alone generally assure that the hazard is positive at all times, as should be. Although, just in case, the code also contains aspects to completely guarantee this holds true. With these aspects simulated, the hazard can be computed for any given observation. Next the start time, the time when the observation enters the study, is drawn from a uniform distribution between zero and one. Lastly using another random drawn variable from the same uniform distribution, the time of death for the observation is computed by taking the start time and adding the value you get by plugging the second uniformly distributed variable into the derived inverse cumulative distribution function. If the time of death exceeds one, the end time for the simulated study, Δ_i takes the value of zero and W_i takes the value of one minus the observation's starting time, as this is the right censoring time for this observation. If the time of death does not exceed the end time of the simulated study, Δ_i takes on the value of one and W_i gets assigned the value of the death time minus the starting time, resulting in the survival time of the observation. With these values set, the result is a fully simulated survival dataset, fit for analysis.

As mentioned previously, the covariate matrix $\mathbf{Q}(t)$ is simulated to be time independent. Because of this, similar to the application part of this paper, the identities presented in equation (18) can be used to improve the computational feasibility of the procedures.

3 Results

3.1 Application

3.1.1 Linear Regression

After applying the methodology on the election dataset, the following estimations presented themselves:

Table 1: The results from applying the linear regression methodology on the voting dataset

$\hat{\theta}$	$\hat{\gamma_1}$	$\hat{\gamma_2}$	$\hat{\gamma_3}$	$\hat{\gamma_4}$	$\hat{\gamma_5}$	$\hat{S}(\hat{\beta})$	lower bound θ	upper bound θ
0.8692	0.7183	-1.2616	-0.0485	0.9932	-0.0216	1.6857×10^{-12}	-4.5864	6.3247

What stands out from these results in table 1, is first of all, that the value of the estimated score function is very small in magnitude relative to that of the parameters. This makes it highly unlikely that its implementation is actually going to have a significant impact on the resulting confidence interval. Secondly, the derived confidence bounds are quite far apart relative to the absolute values of the parameter estimations. This might indicate that the methodology constructs too large of a confidence interval that is much larger than the ninety-five percent coverage it should correspond to.

3.1.2 Additive Hazards Model

Applying the discussed methodology on the seedling survival data produced the following estimations:

Table 2: The results from applying the additive hazards model methodology on the seedling survival dataset

$\hat{\theta}$	$\hat{\gamma_1}$	$\hat{\gamma_2}$	$\hat{S}(\hat{\beta})$	lower bound θ	upper bound θ
0.0307	-1.3435	-1.3903×10^{-5}	2.2568×10^{-8}	-0.1134	0.1748

Similarly to the linear regression application, the estimated score function is rather small relative to the magnitude of the estimated parameters, making its impact and usefulness doubtful. Moreover, once again, the derived confidence interval is rather broad relative to the absolute value of the estimated parameters, again indicating the theory might not be valid in this regard as it possibly provides an interval with a coverage much larger than ninety-five percent.

3.2 Simulation

3.2.1 Linear Regression

For the linear regression model, the simulations provided the results as presented in tables 3, 4 and 5.

Table 3: Hit-rates for linear regression simulations for different numbers of parameters and theoretical coverages with $\lambda' = 0.5$

# parameters\(1-α)%	90%	95%	99%
10	100.0%	100.0%	100.0%
25	100.0%	100.0%	100.0%
50	100.0%	100.0%	100.0%
75	99.9%	100.0%	100.0%
90	90.8%	93.2%	98.3%
100	0.0%	0.0%	0.0%

Table 4: Hit-rates for linear regression simulations for different numbers of parameters and theoretical coverages with $\lambda' = 1$

# parameters\(1-α)%	90%	95%	99%
10	100.0%	100.0%	100.0%
25	100.0%	100.0%	100.0%
50	100.0%	100.0%	100.0%
75	100.0%	100.0%	100.0%
90	90.2%	93.4%	97.1%
100	0.0%	0.0%	0.0%

Table 5: Hit-rates for linear regression simulations for different numbers of parameters and theoretical coverages with $\lambda' = 2$

# parameters\(1-α)%	90%	95%	99%
10	100.0%	100.0%	100.0%
25	100.0%	100.0%	100.0%
50	100.0%	100.0%	100.0%
75	100.0%	100.0%	99.9%
90	90.0%	92.9%	96.9%
100	0.0%	0.0%	0.0%

The proposed procedures seem to provide too large of an interval for the models with a lower number of parameters. Yet as the number of parameters approaches the number of observations, the functionality of the methods rapidly collapses. Whereas the hit-rates for the models with 90 parameters still reach at least ninety percent for all theoretical coverages, the hit-rates for the models with a hundred parameters are equal to zero on every occasion. Thus the coverage is somewhat adequate for a ninety parameter model, which is promising, but beyond that, the effectiveness of the methods rapidly decreases. The effect of the slight differences in the choice of the tuning parameter λ' seems to be rather negligible. Any slight differences between the tables for the different values for λ' can most likely be interpreted as coincidental. Thus the simulations seem to indicate that the proposed theory is not completely valid in practice.

3.2.2 Additive Hazards Model

For the additive hazards model, the simulations provided the results as presented in tables 6,7 and 8.

Table 6: Hit-rates for additive hazards model simulations for different numbers of parameters and theoretical coverages with $\lambda' = 0.5$

# parameters\(1-α)%	90%	95%	99%
10	50.7%	57.0%	70.4%
25	52.6%	60.5%	72.9%
50	49.6%	58.1%	67.4%
75	56.2%	60.1%	72.2%
90	58.3%	63.1%	74.5%
100	58.3%	67.6%	77.2%

Table 7: Hit-rates for additive hazards model simulations for different numbers of parameters and theoretical coverages with $\lambda' = 1$

# parameters\(1-α)%	90%	95%	99%
10	48.5%	58.2%	72.7%
25	52.0%	61.5%	71.0%
50	52.7%	54.2%	68.3%
75	51.2%	63.2%	74.2%
90	56.0%	64.0%	73.6%
100	57.4%	64.2%	75.8%

Table 8: Hit-rates for additive hazards model simulations for different numbers of parameters and theoretical coverages with $\lambda' = 2$

# parameters\(1-α)%	90%	95%	99%
10	47.6%	54.3%	72.2%
25	54.1%	60.1%	71.1%
50	51.2%	56.2%	68.3%
75	53.4%	62.5%	74.3%
90	55.1%	62.3%	73.5%
100	57.1%	66.0%	77.6%

For none of the individual models, do the hit-rates closely match the corresponding theoretical coverage with the hit-rates generally being off by thirty to forty-five percent. However, it is promising that as dimensionality increases, the hit-rates remain relatively consistent. Looking at the results one might notice a slight upwards pattern in the hit-rates, but this pattern is broken for all three tables at the fifty parameter model. But as stated, the differences between the models is rather small. Subsequently the methodology does seem to deal relatively well with increasing dimensionality, although it possibly suffers from a slight scaling issue. If all the confidence intervals are scaled differently, specifically to cover a wider interval, the methodology could potentially provide correct coverages quite consistently, even when dealing with increasing dimensionality. Any differences between models for different values of tuning parameter λ', appear to be fairly arbitrary. When looking at the differences, there is no noticeable pattern or anything that hints to a significant impact due to the choice of the tuning parameter. This in part, displays an element of robustness for the methodology regarding the additive hazards models.

4 Conclusion

4.1 Linear Regression

The application of the methodology on the election dataset showed that the methodology potentially constructed confidence intervals that were too broad. The simulations partially confirmed this suspicion, showing that the confidence intervals constructed were in fact too extensive for models with a lower amount of parameters. Howbeit, as dimensionality increases, specifically as the number of parameters approaches the number of observations, the actual accomplished coverage starts to collapse towards zero. Thus, the procedure does not appear to function well with increasing dimensionality. There is no sign that these results are dependent on the choice of the tuning parameter λ', as the results showed only marginal, most likely coincidental differences in this regard.

4.2 Additive Hazards Model

Similar to its linear regression counterpart, the results from the application of additive hazards model methodology on the seedling survival dataset, suggested that the constructed confidence intervals could be too large. This suspicion was nevertheless disputed in the ensuing simulation study. In fact, the latter demonstrated that the constructed confidence intervals were to tight to adequately reach the desired theoretical coverage. It should be noted though, that the methodology showed consistent results for all simulated models, even in the wake of higher dimensional models. This indicates that re-scaling the confidence intervals in the methodology could potentially result in ample theory, suitable for real-life application. Parallel to the simulation results for the linear regression models, the differing values for tuning parameter λ' do not appear to cause much disparity between the additive hazards models, providing an element of robustness to the procedures.

5 Discussion

The most surprising finding of this paper, would be the invalidation of the confidence interval construction theory for linear regression models. The results indicated too wide of an interval for models with a lower number of parameters, whilst higher dimensionality caused the effectiveness of the procedures to reduce to next to nothing. In this regard, it should of course be noted that such results do not by definition have to be considered to be caused by a fault in the theory, but could instead be caused by mistakes in the programming and/or analytical derivations employed in this paper. So although there is reason for concern, by no means should the theory be definitively labeled as either wrong or incomplete.

Another one of the more uncanny results in this paper, is the difference in the results between the application and the simulations for the additive hazards model. The application seemed to suggest the confidence interval to be too wide, while the simulations showed the exact opposite. The simulations naturally are much more thorough and its results should be regarded as the most relevant accordingly. The different impression given off by the application results, could be a simple anomaly or a misinterpretation within its given context.

Additionally, in regards to the additive hazards model methodology, one should remember that there was made use of a substantiated approximation for $\bar{\mathbf{Q}}(t)$. And although the approach has been supported with arguments and derivations, whether it is applicable in practice is still up for debate. Hence, it could potentially have skewed results with no fault to the original theory as provided by Ning and Liu (2017).

For further research, I would firstly recommend a revisit to the confidence interval theory for linear regressions, as to possibly falsify my findings. On top of that I would suggest further research to be done on a partial rework or re-scaling of the theory for confidence intervals in additive hazards models, which could potentially result in a useful and applicable procedure to construct reliable confidence intervals for high-dimensional additive hazards models.

References

V. Arel-Bundock, A. Blais, and R. Dassonneville. Do voters benchmark economic performance? *British Journal of Political Science*, 2019.

S. E. Aytaç. Relative economic performance and the incumbent vote: A reference point theory. *The Journal of Politics*, 80(1), 2017.

M. Collyer, D. Sekora, and D. Adams. A method for analysis of phenotypic change for phenotypes described by high-dimensional data. *Heredity*, 115:357–365, 2015.

D. L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. Aug 2000.

J. Fan and R. Li. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. 2000.

D. Francois. *High-Dimensional Data Analysis: Optimal Metrics and Feature Selection*. PhD thesis, University catholique de ´Louvain, 2007.

Harvard Dataverse. Replication data for: Do voters benchmark economic performance?, 2018.

C. Heij, P. de Boer, P. H. Franses, T. Kloek, and H. K. van Dijk. *Econometric Methods with Applications in Business and Economics*. Oxford University Press, 2004.

M. A. Kayser and M. Peress. Benchmarking across borders: Electoral accountability and the necessity of comparison. *American Political Science Review*, 106(3):661–684, 2012.

D. G. Kleinbaum and M. Klein. *Survival Analysis*. Statistics for Biology and Health. Springer, 2012.

M. Laakso and R. Taagepera. "effective" number of parties: A measure with application to west europe. *Comparative Political Studies*, 12(1), 1979.

Y. Ning and H. Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

R. Pillay, F. Hua, B. A. Loiselle, H. Bernard, , and R. J. Fletcher. Multiple stages of tree seedling recruitment are altered in tropical forests degraded by selective logging. 2018. doi: http://doi.org/10.5281/zenodo.2020340.