2008

Shrinkage, pretest and LASSO estimators in parametric and semiparametric linear models

Md. Shakhawat Hossain
University of Windsor

Shrinkage, Pretest and LASSO Estimators in Parametric and Semiparametric Linear Models

by

MD. SHAKHAWAT HOSSAIN

A Dissertation
Submitted to the Faculty of Graduate Studies
through the Department of Mathematics and Statistics
in partial fulfillment of the Requirements for
the Degree of Doctor of Philosophy at the
University of Windsor

Windsor, Ontario, Canada
2008

Your file Votre référence
ISBN: 978-0-494-42380-6
Our file Notre référence
ISBN: 978-0-494-42380-6

Declaration of Co-Authorship/Previous Publication

I. Co-Authorship Declaration

I hereby declare that this thesis incorporates the outcome of a joint research undertaken in collaboration with my supervisor, Professor S. E. Ahmed. In all cases, the key ideas, primary contributions, experimental designs, data analysis and interpretation, were performed by the author, and the contribution of co-author was primarily through the provision of some theoretical results.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from each of the co-author to include in my thesis.

I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

II. Declaration of Previous Publication

This thesis includes one original paper that has been previously published and another one is submitted for publication:

Thesis Chapter	Publication title/ full citation	Publication status
Chapter 4	Shrinkage, Pretest and Absolute Penalty Estimators in Partially Linear Models Aust. N. Z. J. Stat. 49(4),2007, 435-454	Published
Chapter 4	Positive Shrinkage, Improved Pretest and Absolute Penalty Estimators in Partially Linear Models	Submitted

I certify that I have obtained a written permission from the copyright owner to include the above published material in my thesis. I certify that the above material

describes work completed during my registration as graduate student at the University of Windsor.

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner to include such material in my thesis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University of Institution.

Abstract

The theory of pretest (Bancroft (1944)) and James-Stein(James and Stein (1961)) type shrinkage estimation has been quite well known for the last five decades though its application remains limited. In this dissertation, some contributions to different types of parametric and semiparametric linear models based on shrinkage and preliminary test estimation methods are made which improve on the maximum likelihood estimation method.

The objective of this dissertation is to study the properties of improved estimators of the parameter of interest in parametric and semiparametric linear models and compare these estimators with the least absolute shrinkage and selection operator (Tibshirani (1996)) estimator.

Chapter two contains a study of the properties of the shrinkage estimators of the parameters of interest in a Weibull regression model where the survival time may be subject to fixed censoring and the regression parameters are under linear restrictions. Asymptotic properties of the suggested estimators are established using the notion of asymptotic distributional risk. Bootstrapping procedures are used to develop confidence intervals. An extensive simulation study is conducted to assess the performance of the suggested estimators for moderate and large samples.

In chapter three, we consider generalized linear models for binary and count data. Here, we propose James-Stein type shrinkage estimators, a pretest estimator and a Park and Hastie estimator. We demonstrate the relative performances of shrinkage and pretest estimators based on the asymptotic analysis of quadratic risk functions and it is found that the shrinkage estimators outperform the maximum likelihood estimator uniformly. On the other hand, the pretest estimator dominates the maximum likelihood estimator only in a small part of the parameter space, which is consistent with the theory. A Monte Carlo simulation study has been conducted to compare shrinkage, pretest and Park and Hastie type estimators with respect to the maximum

likelihood estimator through relative efficiency.

In chapter four, we consider a partial linear model where the vector of coefficients β in the linear part can be partitioned as (β_1, β_2) where β_1 is the coefficient vector for main effects and β_2 is a vector for "nuisance" effects. In this situation, inference about β_1 may benefit from moving the least squares estimate for the full model in the direction of the least squares estimate without the nuisance variables, or from dropping the nuisance variables if there is evidence that they do not provide useful information (pre-testing). We investigate the asymptotic properties of Stein-type and pretest semiparametric estimators under quadratic loss and show that, under general conditions, a Stein-type semiparametric estimator improves on the full model conventional semiparametric least squares estimator. The relative performance of the estimators is examined using asymptotic analysis of quadratic risk functions and it is found that the Stein-type estimator outperforms the full model estimator uniformly. On the other hand, the pretest estimator dominates the least squares estimator only in a small part of the parameter space, which is consistent with the theory. We also consider an absolute penalty type estimator for partial linear models and give a Monte Carlo simulation comparison of shrinkage, pretest and the absolute penalty type estimators.

Acknowledgements

Foremost, I would like to express my gratitude to Professor Syed Ejaz Ahmed, my supervisor, for his support and encouragement in the preparation of this thesis. He involved me into research in shrinkage and LASSO methods. In my early days of graduate studies he helped augment my foundations in Statistical Theory. His insights regarding both scholarly and practical matters in statistics have always encouraged me to take the next step to becoming a statistician. I am very fortunate to have him as a supervisor who is not only enthusiastic and caring, but also friendly.

Particular thanks go to Dr. Abdul Hussein. He helped me a lot in R programming for Monte Carlo study for shrinkage estimators. He also gave me priceless advice since the beginning of my Ph.D. research

My appreciation also goes to Drs. Anne Snowdon, Myron Hlynka and Severien Nkurunziza for serving on my dissertation examination committee and to the external examiner Dr. Vijay Nair. I am thankful for their careful review of my thesis and for the valuable comments they provided.

My profound thankfulness and endless love to my wife Lucky Yeasmin and son Irfan Hossain for having the patience and understanding for the many hours and late nights. I acknowledge their support and sacrifices during my study period.

Finally, I would like to thank to all faculty, staff members and the fellow graduate students in Department of Mathematics and Statistics. Particular thanks to Christine Young and Dina Labelle, for their support at every stage during my Ph.D. study.

Contents

List of Tables

List of Figures

Abbreviations

ADB	asymptotic distributional bias
ADR	asymptotic distributional quadratic risk
APE	absolute penalty estimator
$AQDB$	asymptotic quadratic distributional bias
AR	auto regressive
BIC	Bayesian information Criteria
$ch_{min}(\mathbf{A})$	smallest eigenvalue of $\mathbf{A}$
$ch_{max}(\mathbf{A})$	largest eigenvalue of $\mathbf{A}$
CV	cross validation
GLM	generalized linear model
GCV	generalized cross validation
$IRLS$	iteratively reweighted least squares
PSE	positive shrinkage estimator
$LASSO$	Least Absolute Shrinkage and Selection Operation
MSE	mean squared error
MLE	maximum likelihood estimator
NSI	non-sample information
PH	Park and Hastie
PT	preliminary test estimator
SE	shrinkage estimator
SRE	simulated relative efficiency
$RMSE$	relative mean squared error
RE	restricted estimator
UE	unrestricted estimator
UPI	uncertain prior information
VA	Veterans' Administration

List of Symbols

$\boldsymbol{\beta}$	regression parameter vector
H_0	null hypothesis
$\mathbf{I}$	identity matrix
k	the number of regression parameter
$K_{(n)}$	Pitman type local alternative hypothesis
$\mathcal{L}$	weighted quadratic loss function
n	sample size
$\hat{\boldsymbol{\beta}}$	unrestricted estimator
$\tilde{\boldsymbol{\beta}}$	restricted estimator
$\hat{\boldsymbol{\beta}}^{PT}$	pretest estimator
$\hat{\boldsymbol{\beta}}^{S}$	shrinkage estimator
$\hat{\boldsymbol{\beta}}^{S+}$	positive shrinkage estimator
$\boldsymbol{\beta}^{I}$	integrated estimator
Λ	test statistic
$I(A)$	indicator function of A
$\mathbf{Q}$	a positive semi-definite weight matrix in the quadratic loss function
$\mathbf{H}$	$q \times k$ matrix
$\mathbf{Q}^{*}$	dispersion matrix
R	asymptotic distributional quadratic risk function
τ	tuning parameter
T	lifetime of an individual
$\Re^{p}$	p-dimensional real valued vector
$\mathbf{h}$	$q \times 1$ real valued vector
$\mathbf{I_0}$	expected information matrix
$\boldsymbol{\Sigma}$	variance covariance matrix

$\boldsymbol{\delta}$	a fixed real valued vector contained in $K_{(n)}$
$\boldsymbol{\Omega}$	variance covariance matrix of $\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h}$
Δ	non-centrality parameter
$\mathbf{U}(\boldsymbol{\beta})$	score function of $\boldsymbol{\beta}$
$\mathbf{W}$	weighting matrix
$\boldsymbol{\Lambda}$	$q \times q$ diagonal matrix
$\mathbf{D}_1$	likelihood ratio test statistic
$\mathbf{D}_2$	The Wald test statistic
$\mathbf{D}_3$	The Rao score test statistic
$\mathbf{G}(y)$	non-degenerate distribution function of $\mathbf{y}$
$\Psi_\nu(x\,;\Theta)$	noncentral χ^2 distribution function with non-centrality parameter Θ and degrees of freedom ν
T_n	test statistic for testing the semiparametric regression parameter
$\boldsymbol{\Gamma}$	asymptotic variance covariance matrix

Chapter 1

Introduction and Literature Review

1.1 Introduction

Statistical models are created in effort to ascertain knowledge about unknown population quantities. In many situations, however, consulting statisticians and biostatisticians investigate the statistical properties not only with data based on sample information but also on nonsample information. In problems of statistical inference, the use of *nonsample information(NSI) or uncertain prior information (UPI)* on some (all) of the parameters in a statistical model usually leads to an improved inference procedure for other (all) parameters of interest. However, the prior information may be known or uncertain. The known prior information is generally incorporated in the model form of a constraint, giving rise to restricted models. The statistical analysis of such restricted models leads to an improved statistical procedure when such restrictions hold in unrestricted models. The validity and efficiency of the restricted model analysis retains its properties over the restricted parameter space induced by the constraint, where the same holds for the unrestricted model analysis over the entire parameter space. Therefore, the results of an analysis based on restricted and

unrestricted models need to weigh loss of efficiency against the validity of constraints in order to choose between the two inference techniques.

When we encounter problems with uncertain prior information in statistical models, we may inflict some prior information on the model which may come from understanding data, statistical theory, previous empirical work, or other factors. Now the question arises as to how one can incorporate this uncertain prior information into the inference procedure. In this regard, Bancroft (1944) came up with idea of testing the uncertainty of the prior information in the estimation process. It is natural to perform a pretest or preliminary test on the validity of the uncertain prior information and then inference is developed based on the result of this test. The resulting estimation procedure is called pretest estimation. This estimation method is a compromised inference procedure between unrestricted and restricted rules.

The James-Stein estimator, a so called shrinkage estimator, combines the sample and the non-sample information in a way that improves the precision of the estimation process or the quality of subsequent predictions. It is easy to implement and adapt to maximum likelihood and other classical estimators. The existing literature shows that the shrinkage estimator has lower risk than the classical estimators including the maximum-likelihood estimator in the classical linear regression model, under very mild conditions.

In this dissertation, we develop shrinkage estimation methods in three different statistical models when non-sample information on the parameter of interest exists. Further, we also consider least absolute shrinkage and selection operator (Tibshirani (1996)) method.

For expository purposes, let us formulate the basic problem of estimating $\boldsymbol{\beta}$, the regression parameter of a multiple regression model which is given in vector form as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{Y} = (y_1, \cdots, y_n)$ are responses, $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_k})$ are the predictors and $\boldsymbol{\beta} =$

$(\beta_1, \ldots, \beta_k)'$ is an unknown parameter vector. The components of the error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)$ are assumed to have mean zero vector and unknown variance σ^2. For inference purposes, we assume that the errors are independently, identically and normally distributed with the above mean and variance. The maximum likelihood estimator of the parameter vectors $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Unrestricted and Restricted Estimator:

When an estimator relies on sample data only and is not a function of uncertain prior information, it is well known that the maximum likelihood estimation leads to the best estimate, at least in the class of linear unbiased estimates. Let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimator of $\boldsymbol{\beta}$ based on a sample size n. This estimator is called an unrestricted maximum likelihood estimator (UE) in the full model.

On the other hand, using the available information in models may be advantageous to obtain improved estimates. The uncertain prior information may be explicitly incorporated into the estimation scheme by modifying the parameter space. In this case, the new (restricted) parameter space is a subspace of the original one (reduced in dimension). Let $\tilde{\boldsymbol{\beta}}$ be the restricted maximum likelihood estimator (RE) of $\boldsymbol{\beta}$ when the uncertain prior information is correct. This estimator $\tilde{\boldsymbol{\beta}}$ is more efficient (or, at least, no less efficient) than the unrestricted estimator when the model satisfies the restriction. But what happens when it does not satisfy the restriction. It is easy to see that the restricted estimator will, in general, be biased.

Pretest Estimator:

Let Λ be a suitable test statistic for the null hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}$, where $\mathbf{H}$ is a $q \times k$ matrix of rank $q \leq k$ and $\mathbf{h}$ is a given $q \times 1$ vector of constants. Let $c_{n,\alpha}$ be the critical value, i.e., the $100(1-\alpha)\%$ percentile, of the distribution of Λ under the

null hypothesis. It seems natural to define an estimator of the following form:

$$\hat{\boldsymbol{\beta}}^{PT} = \hat{\boldsymbol{\beta}} - (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})I(\Lambda \leq c_{n,\alpha}), \tag{1.2}$$

where $I(A)$ is an indicator function of a set A. This is called a pretest estimator (PT). Some useful discussions about pretesting can be found in Judge and Bock (1978), Giles and Giles (1993), Ohtani *et al.* (1997), Ahmed (2001), and Ahmed *et al.* (2006*a*), among others. It is important to remark here that $\hat{\boldsymbol{\beta}}^{PT}$ performs better than $\hat{\boldsymbol{\beta}}$ in some part of the parameter space. The use of $\hat{\boldsymbol{\beta}}^{PT}$ may, however, be limited due to the large size of the pretest. The performance of this estimator is substantially better than UE when uncertain prior information is nearly correct. To overcome this shortcoming, we construct an estimation procedure based on the most celebrated James-Stein type or shrinkage estimation procedure. This rule combines the sample and non-sample information in a superior way compared to the preceding estimator.

Shrinkage and Positive Shrinkage Estimator:

Following Ahmed (2001), the shrinkage estimator (SE) of $\hat{\boldsymbol{\beta}}$ is defined as:

$$\hat{\boldsymbol{\beta}}^{S} = \tilde{\boldsymbol{\beta}} + \left[1 - \frac{(q-2)}{\Lambda}\right](\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}), \quad q \geq 3. \tag{1.3}$$

Interestingly, the above estimator is obtained by simply replacing the binary quantity $I(A)$ in (1.2) by a continuous quantity $(q-2)\Lambda^{-1}$. Hence, the above shrinkage estimator arises in a natural way. Note that $\hat{\boldsymbol{\beta}}^{S}$ is not a convex combination of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$. However, the proposed estimator $\hat{\boldsymbol{\beta}}^{S}$ may have the opposite sign of $\hat{\boldsymbol{\beta}}$. To avoid this strange behavior of $\hat{\boldsymbol{\beta}}^{S}$, we truncate $\hat{\boldsymbol{\beta}}^{S}$, which leads to a convex combination of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ and is called positive-part shrinkage estimator (PSE). This estimator can be defined as

$$\hat{\boldsymbol{\beta}}^{S+} = \tilde{\boldsymbol{\beta}} + \left[1 - \frac{(q-2)}{\Lambda}\right]^{+}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}), \tag{1.4}$$

where $z^{+} = max(0, z)$. We emphasize here that $\hat{\boldsymbol{\beta}}^{S+}$ is particularly important to

control the over-shrinking inherent in $\hat{\boldsymbol{\beta}}^S$.

In this dissertation, we consider the asymptotic set up to appraise the performance of the listed estimators. To this end, we consider the weighted quadratic loss function criterion to examine the performance of the estimators.

$$\mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\beta}; \mathbf{Q}) = \left[\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})\right]' \mathbf{Q} \left[\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})\right], \tag{1.5}$$

where $\mathbf{Q}$ is a positive semidefinite weighting matrix and $\boldsymbol{\beta}^*$ can be any one of $\hat{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}^S$, $\hat{\boldsymbol{\beta}}^{S+}$ or $\hat{\boldsymbol{\beta}}^{PT}$.

Consider the Pitman type of alternatives

$$K_n : \mathbf{H}\boldsymbol{\beta} = \mathbf{h} + \frac{\boldsymbol{\delta}}{n^{1/2}}, \tag{1.6}$$

where $\boldsymbol{\delta} = (\delta_1, \delta_2 \cdots, \delta_q) \in \Re^q$ a real fixed vector. Note that for $\boldsymbol{\delta} = \mathbf{0}$, $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ for all n.

Further, we introduce the asymptotic distribution function of $\boldsymbol{\beta}^*$ under K_n by

$$G(\mathbf{y}) = \lim_{n\to\infty} P\left[\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) \leq \mathbf{y} | K_n\right],$$

where $G(\mathbf{y})$ is a nondegenerate distribution function. Then, we define the asymptotic distributional quadratic risk (ADR) by

$$\begin{aligned} R(\boldsymbol{\beta}^*; \mathbf{Q}) &= \int \cdots \int \mathbf{y}'\mathbf{Q}\mathbf{y} dG(\mathbf{y}) \\ &= \text{trace}(\mathbf{Q}\mathbf{Q}^*), \end{aligned}$$

where $\mathbf{Q}^* = \int \cdots \int \mathbf{y}\mathbf{y}' dG(\mathbf{y})$ is the dispersion matrix for the distribution $G(\mathbf{y})$.

Based on this asymptotic risk set up, we compare the risks of the suggested estimators under the quadratic loss function. Our simulation study shows that our estimators dominate the maximum likelihood estimator in the entire parameter space.

Least Absolute Shrinkage and Selection Operator (LASSO):

The Least Absolute Shrinkage and Selection Operator, first proposed by Tibshirani (1996), regularizes ordinary least squares regression with a L_1 regularizer. This is one of many shrinkage regression methods, which all have the basic idea of shrinking the parameters towards zero. In least squares regression, these parameters are estimated by minimizing the residual sum of squares:

$$\min_{\boldsymbol{\beta}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

The LASSO imposes an additional restriction on the coefficients, namely:

$$\sum_{j=1}^{n} |\beta_j| \leq \tau,$$

where τ is a tuning parameter. If the tuning parameter $\tau \geq 0$ is large enough, this just gives the usual least squares estimates. However, smaller values of τ produce shrunken estimates $\boldsymbol{\beta}$, often with many components equal to zero. Choosing τ can be thought of as choosing the number of predictors to include in a regression model. Thus the LASSO can select predictors in a manner similar subset selection methods. However, since it is a smooth optimization problem, it is less variable than subset selection and can be applied to much larger problems (large in p).

Knight and Fu (2001) established some asymptotic results for LASSO-type estimators. Fan and Li (2002) introduced the Smoothly Clipped Absolute Deviation approach and proved its optimal properties. Efron *et al.* (2004) introduced the Least Angle Regression algorithm and discussed its close connection to LASSO. Park and Hastie (2007) developed methods for fitting the entire coefficient path for a generalized linear model with L_1 penalties.

The following subsections give the introduction and literature review for three different problems.

1.2 Delineating the effect of Misspecification in a Lifetime Censored Regression Model

Statistical analysis of failure-time data is an active and important area of research that has been received considerable attention from several applied disciplines. Survival analysis in clinical trials, reliability theory, industrial and manufacturing systems, biological sciences and social sciences provide examples where failure-time data are studied. Historically, failure times are modelled by fitting an exponential, Weibull, or log normal distribution to the data.

Failure (or response) time data usually arise with measurements of certain auxiliary variables known as covariates. For example, data on the occurrence of a heart-attack of a patient are usually coupled with measurement of blood pressure, weight, age, family history for heart diseases, life-style of patient and cholesterol level etc. Statistical analysis provides a scientific tool to investigate such relationships using data obtained from previous or current studies. The aim of statistical analysis is to identify the risk factors that contribute significantly to the presence or the occurrence of the event which is under investigation. Very often, the analysis is conducted through a statistical procedure called parametric regression modeling, where the dependence of survival time on covariates or risk factors is described explicitly through the parameters, hazard function and survival function. For example, Breslow (1974) used the exponential distribution to model the remission duration of children with leukemia and to identify important risk factors. He modelled the rate of the exponential distribution as a function of potential risk factors.

In some studies, certain covariates present a linear relation, i.e. some variables depend linearly on some others. Such phenomenon is called model misspecification and there is an uncertainty about coefficient restrictions. Since the presence of linear restrictions among covariates induces large variation and uncertainty in the regression models, the estimates of the model parameters have large variance, and prediction

based on the models may perform very poorly. Therefore the models may not serve the needs of the investigators. In this situation, one may assume that the prior information about the model consists of specification of restrictions on the regression coefficients. Such a prior information may be derived from past experience of similar investigations and from the exhibition of stability of estimates of regression coefficients in repeated studies and/or some extraneous sources and/or from some theoretical considerations; see, e.g., Judge *et al.* (1985) and Rao and Toutenburg (1995).

In this problem, we consider the shrinkage estimation (point and interval) method for the Weibull regression model where the survival time may be subject to a fixed censoring and the regression parameters are under linear restrictions.

Let T represent the lifetime of an individual. To begin with, we shall assume that there are no covariates in the study and that we are interested only in the survival probability of the individual. Suppose T has a Weibull distribution with scale parameter λ and shape parameter ν. In this situation, the survival probability is given by

$$S(t|\lambda,\nu) = exp\left[-(\lambda t)^{\nu}\right],$$

where $\lambda > 0$ and $\nu > 0$. This Weibull model can be generalized by modelling the shape parameter as a function of the covariates. In this case, λ is a function of $\mathbf{x} = (x_1, \cdots, x_k)$ involving unknown parameters and the survival probability for an individual given covariate vector $\mathbf{x}$ is

$$S(t|\lambda(\mathbf{x}),\nu) = exp\left[-(\lambda(\mathbf{x})t)^{\nu}\right].$$

We will return to this model in Chapter Two.

It is also useful to model the logarithm of the lifetime given the covariate vector $\mathbf{x}$. For Weibull distributed lifetime T with scale parameter $\lambda(\mathbf{x})$ modelled as a function of the covariate vector $\mathbf{x}$ and fixed but unknown shape parameter ν, the conditional distribution of $Y = \ln T$, given $\mathbf{x}$, has an extreme value distribution with location

parameter $\mu(\mathbf{x}) = -\ln\lambda(\mathbf{x})$ and scale parameter $\sigma = \nu^{-1}$. This corresponds to a location scale model for the logarithm of lifetime specified by

$$Y = \ln T = \mu(\mathbf{x}) + \sigma\varepsilon,$$

where ε has a standard extreme value distribution with $\mu(\mathbf{x}) \equiv \mathbf{0}$ and $\sigma \equiv 1$ corresponding to $\mathbf{x} = (0, 0, \cdots, 0)$ as $q \times 1$ vector. This is a fully parametric version of the accelerated failure time model where the distribution is specified.

Various parametric procedures for the analysis of censored data in the presence of concomitant variables have been proposed. Feigle and Zelen (1965) suggested a regression method for relating the concomitant information to the survival times of patients with cancer. They have dealt with the situation of complete information on the failure of all subjects. The method has been extended by Zippin and Armitage (1966) so that the data sets which include censored observations can be analyzed. Feigle and Zelen (1965), Glasser (1967), Zippin and Armitage (1973), Lawless and Signhal (1978) and Peduzzi *et al.* (1980) analyzed data sets from multiple myeloma patients based on exponential regression model. Odell *et al.* (1992) described a Weibull regression model for interval-censored data with fixed (e.g. baseline) covariates. Rabinowitz *et al.* (1995) extended the accelerated failure model to the interval-censored case. They presented a class of score statistics for estimating the regression coefficients without specifying the distribution function of the residuals or the joint distribution of the covariates and the interval times. Ahmed and Saleh (1999) applied James-Stein estimation method for estimating the regression coefficients in an exponential model with censoring considered when it is a priori suspected that the parameters may be restricted to a subspace. A family of penalized partial likelihood methods, such as LASSO (Tibshirani (1997)) and the smoothly clipped absolute deviation method (Fan and Li (2002)), were proposed for the Cox proportional hazards model. By shrinking some regression coefficients to zero, these methods select important variables and estimate the regression model simultaneously.

1.3 Shrinkage, Pretest and PH estimators for Generalized Linear Models

Many popular statistical methods based on mathematical models assume that data follow a normal distribution. Most obvious among these are the analysis of variance for planned experiments and multiple regression for general analysis of independent and dependent variables. In many situations, the normality assumption is not plausible. Consequently, the use of methods that assume normality may perform unsatisfactorily. In these cases, other alternatives that do not require data to have a normal distribution are attractive. Generalized linear models are an extension of the linear modelling process that allows models to be fit to data that follow probability distributions other than the normal distribution, such as the Poisson, the Binomial and the Multinomial etc. These models are defined by Nelder and Wedderburn (1972). These models also allow the mean of a population to depend on a linear predictor through a nonlinear link function and allow the response probability distribution to be any member of an exponential family of distributions. The motivation is to tailor the regression relationship by connecting the outcome to relevant independent variables so that it is appropriate to the properties of the dependent variable. These models include classical linear models with normal errors, logistic and probit models for binary data, and log-linear models for multinomial data. Many other useful statistical models can be formulated as generalized linear models by the selection of an appropriate link function and response probability distribution. Refer to McCullagh and Nelder (1989) for a thorough account of statistical modelling using generalized linear models. The books by Aitkin *et al.* (1989), Dobson (1990) and Agresti (2002) are also excellent references with many examples of applications of generalized linear models. Tibshirani (1996) briefly discussed using the LASSO to fit generalized linear models. Several other researchers have also considered using L_1 penalties to fit logistic regression models (Lokhorst (1999); Shevade and Keerthi (2003); Genkin *et al.* (2007)).

In addition Zhao and Yu (2004) and Park and Hastie (2007) developed methods for fitting the entire coefficient path for generalized linear models and other models with L_1 penalties. Park and Hastie (2007) developed efficient "glmpath" algorithms for obtaining the LASSO path for the generalized linear models. This algorithm is similar to LASSO, in which the loss function is replaced by the negative log-likelihood of any distribution in exponential family.

A classical linear model is of the form

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,$$

where y_i is the response variable for the ith observation. The quantity x_i represents the ith row of the design matrix $\mathbf{X}$, that is known from the experimental setting and is considered to be fixed or non-random. The vector of unknown coefficients $\boldsymbol{\beta}$ is estimated by the least squares fit to the data $\mathbf{y}$. The ε_i are assumed to be independent, normal random variables with zero mean and constant variance. The expected value of y_i, denoted by μ_i, is

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

When classical linear models are used extensively in statistical data analysis, there are some types of problems for which they are not appropriate. Firstly, it may not be reasonable to assume that data are normally distributed. Secondly, if the mean of the data is naturally restricted to a range of values, the traditional linear model may not be appropriate since the linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$ can take on any value. Thirdly, it may not be realistic to assume that the variance of the data is constant for all observations.

A generalized linear model extends the classical linear model and is therefore applicable to a wider range of data analysis problems. A generalized linear model consists of the following components.

- The linear component is defined just as it is for traditional linear models

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

- A monotonic differentiable link function g describes how the expected value of y_i is related to the linear predictor η_i:

$$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}.$$

- The response variables y_i are independent for $i = 1, 2, \cdots, n$ and have a probability distribution from an exponential family. This implies that the variance of the response depends on the mean μ through a variance function V :

$$var(y_i) = \frac{\phi_i V(\mu_i)}{w_i},$$

where ϕ is a constant and w_i is a known weight for each observation.

In this problem, we consider the estimation problem for the generalized linear models which may have a large collection of potential predictor variables and some of them may not have influence on the response of interest. The use of the maximum likelihood estimator is very common in the literature. These estimators are solely based on the sample information and can be extremely noisy. The shrinkage estimation method which contains the non-sample prior information can be introduced in the estimation procedure to 'improve' the quality of the estimators. The natural expectation is that the inclusion of additional information would result in a better estimator. In this dissertation, we compare three shrinkage methods with the maximum likelihood method for the estimation of generalized linear models: shrinkage type estimation method, pretest estimation method and Park and Hastie (PH) estimation method, so called "glmpath" algorithm.

A Monte Carlo simulation study has been conducted to compare shrinkage, pretest

and PH type estimators with respect to the maximum likelihood estimator through relative efficiency. This comparison shows that the shrinkage method performs better than the PH type estimation method when the dimension of the restricted parameter space is large.

1.4 Shrinkage, Pretest and Absolute Penalty Estimators in Partially Linear Models

The form of semiparametric model that has received the most attention is the partial linear model. In this model, the response y depends on two sets of regressors $(\mathbf{x}, t)$, where the mean response is linearly related to $\mathbf{x} \in \Re^p$ (parametric component), but cannot be easily parameterized in terms of $t \in [0,1]$ (nonparametric component). This model can be expressed as

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + g(t_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{1.7}$$

where $\mathbf{x}_i$ are fixed known $p \times 1$ vectors, $\boldsymbol{\beta}$ is an unknown vector of parameters, $g(\cdot)$ is an unknown (smooth) real-valued function defined on $[0,1]$, the ε_i's are unobservable random errors and the superscript $'$ denotes the transpose of a vector or matrix.

Partially linear models have many applications. Engle *et al.* (1986) were among the first to consider the partially linear model (1.7). They analyzed the relationship between temperature and electricity usage. Model (1.7) is very useful in sociology, economics, finance and biometrics. For example, in a clinical trial for the comparison of two treatments, a subject's response will depend on the treatment received and on some covariate, say age. In this case, the experimenter may be unsure of the effect of age on the response, but may want to estimate the treatment differences which are believed to be constant and independent of age, see Speckman (1988). When the ε_i are independent and identically distributed (i.i.d.) random variables, Heckman

(1986), Rice (1986), Chen (1988), Speckman (1988), Robinson (1988), Eubank *et al.* (1990), Chen and Shiau (1994), Donald and Newey (1994), Hamilton and Truong (1997) and Fan *et al.* (1998) used various estimation methods, such as the kernel method, the spline method, series estimation expansion, local linear estimation, two-stage estimation and others, to obtain estimators of the unknown quantities in (1.7). Further, the asymptotic properties of these estimators have been investigated. Shi and Li (1995) constructed M-estimators for $\boldsymbol{\beta}$ and $g(\cdot)$. When the error is an AR(1) process, Schick (1994) discussed the estimation of the autocorrelation coefficient. Schick (1996), Schick (1998) further constructed efficient estimators for the regression parameter component and autocorrelation coefficient, respectively. A survey of the estimation and application of model (1.7) can be found in the monograph of Härdle *et al.* (2000). Some recent work on semiparametric models can be found in Wang *et al.* (2004), Xue *et al.* (2004), Liang *et al.* (2004), and Bunea (2004).

Judge and Mittelhammaer (2004) eloquently argued that much empirical research proceeds in the context of partially-incomplete subject matter theories and data based on experimental designs not devised by or known to the experimenter. This generally leads to uncertainty concerning the statistical model describing the sample data. This in turn, leads to uncertainty regarding appropriate statistical inference methods. Specifying the statistical model is, as always, a critical component in estimation and inference. One typically studies the consequences of some forms of model misspecification. A common type of misspecification in the models is caused by including unnecessary predictors in the model or by leaving necessary (lurking) variables out. The validity of eliminating statistical uncertainty through the specification of a particular parametric formulation depends on information that is generally not available. The aim of this communication is to analyze some of the issues involved in the estimation of a semiparametric model that may be over-parameterized. For example, in the data analyzed by Engle *et al.* (1986) the electricity demand may be affected by weather, price, income, strikes and other factors. If we have reason to suspect that a

strike has no effect on electricity demand, we may want to decrease the influence of, or delete, this variable. Recently, Cui *et al.* (2005) developed an estimator of the error variance that can borrow information across genes using the James-Stein shrinkage concept. For linear models, Tibshirani (1996) proposed the LASSO method to shrink some coefficients and to set others to zero, and hence tries to retain good features of both subset selection and ridge regression. A penalty on the sum of the absolute ordinary least square coefficients is introduced to achieve both continuous shrinkage and automatic variable deletion. The idea of using an absolute penalty was used by Chen and Donoho (1994) and Chen *et al.* (1999) to shrink and delete basic coefficients. In this problem, we propose the absolute penalty type estimation method which is the extended version of LASSO method.

1.5 Highlights of Contributions

This dissertation extends the concept of James-Stein type shrinkage estimation methods in the context of three different linear models when the nonsample information is available. The first one of these deals with implementation of shrinkage methodology for a Weibull lifetime regression model when the parameters $\boldsymbol{\beta}$ lie in the linear subspace $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$. Further, asymptotic statistical procedures are developed for testing at and near the general linear hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}$. The problem of interval estimation is addressed by using a variety of bootstrap techniques. An extensive simulation study has been conducted to investigate the performance of the suggested methods for moderate and large sample situations. Our contribution is to study point estimation, interval estimation and testing procedures of the Weibull regression parameters when samples are drawn from arbitrary populations. We provide a total inferential package in this chapter for practitioners. Finally, a real data analysis is presented to illustrate our proposed estimation strategies.

In chapter 3, we study the application of shrinkage and pretest estimation methods

to the generalized linear model, which is the most important model for many practical situations. This chapter also addresses the comparison the shrinkage estimation method with the Park and Hastie type estimation method through maximum likelihood estimation. Asymptotic properties of the restricted, shrinkage and positive-part of shrinkage and pretest estimators are discussed and compared with the unrestricted maximum likelihood estimator. It is demonstrated that the positive part estimator utilizes sample and non-sample information in a superior way relative to the ordinary shrinkage estimator. The simulation results are presented in several figures and tables. These results reveal that the shrinkage estimators outperform the maximum likelihood estimators in the entire parameter space and the pretest estimators dominate the maximum likelihood estimators on a small part of the parameter space. On the other hand, the Park and Hastie estimator performs better than the shrinkage estimators when the restrictions on the parameter space is small. A real life data analysis is presented to compare the methods.

In chapter 4, we consider the shrinkage, pretest and absolute penalty estimator in a partial linear model. We investigate the asymptotic properties of shrinkage and pretest semiparametric estimators under quadratic loss and show that a shrinkage semiparametric estimator improves on the full model conventional semiparametric least squares estimator. The relative performance of the estimators is examined using asymptotic analysis of quadratic risk functions and it is found that the Stein-type estimator outperforms the full model estimator uniformly. On the other hand, the pretest estimator dominates the least squares estimator only in a small part of the parameter space. We also consider an absolute penalty type estimator for partially linear models and give a Monte Carlo simulation comparison of shrinkage, pretest and absolute penalty type estimators. The comparison shows that the shrinkage method performs better than the absolute penalty type estimation method when the restriction of parameter space is large.

Chapter 5 summarizes the results and concludes the dissertation with some discus-

sion of related research and the direction for future research as well. This includes a generalization of shrinkage estimation methods to the exponentiated Weibull censored regression model.

Chapter 2

Delineating the effect of Misspecification in a Lifetime Censored Regression Model

2.1 Introduction

Ascertaining the appropriate statistical model-estimator for use in representing the data sampling process is an interesting and challenging problem for statisticians. In this dissertation, we consider inference problems under linear restrictions in a Weibull lifetime regression censored model. In the classical framework, prior information may be introduced either by augmenting sample information, through likelihood function, or by modifying the parameter space. The latter is achieved through equality or inequality restrictions. In the case of exact restrictions, the new parameter space is of reduced dimensionality, which improves the precision of parameter estimates, because the available information is concentrated on a smaller set of parameters. Shrinkage methods provide useful techniques for the dealing with inference problems under such restrictions, and recent asymptotic theory has advanced the understanding of the fundamental role of the likelihood function for much the same purpose. The

important message is that when estimating many parameters (at least more than 2), there is a great advantage in shrinking the estimates. This procedure plays an important role in modern nonparametric function estimation.

We refer to Lawless (2003), Kalbfleisch and Prentice (2002), Bugaighis (1995) and Smith (1991), for detailed study and applications of the Weibull regression model. Applications of this model can be found in research in human diseases such as cancer, mortality rate for aged people and lifetime analysis of animal carcinogenesis. More applications of this model can be found in Kalbfleisch and Prentice (2002).

The main objective of this chapter is to estimate regression parameters $\boldsymbol{\beta}$ when $\boldsymbol{\beta}$ is suspected to lie in the subspace defined by

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{h}, \tag{2.1}$$

where $\mathbf{H}$ is $q \times k$ matrix of rank $q \leq k$ and $\mathbf{h}$ is a given $q \times 1$ vector of constants. The information in relation (2.1) may be regarded as *nonsample information (NSI)*. It is assumed that $\mathbf{H}$ has rank q, which implies that the q equations do not contain any redundant information about $\boldsymbol{\beta}$. This situation occurs frequently when there is over-modelling and one wishes to remove the irrelevant part of the model, which in turn will increase the efficiency of estimating $\boldsymbol{\beta}$. For instance, in some situations the interaction effects may not be present and we are interested in estimating the parameter vector on main factors only. More specifically, this research is motivated by the following data.

Clinical Trial Data: Lawless (2003) and Kalbfleisch and Prentice (2002) and others considered the Veteran's administration (VA) lung cancer data. In this trial, patients were assigned to one of two chemotherapy treatments. Several factors hypothesized to be relevant to an individual's prognosis include, performance status, age and the number of months from diagnosis of cancer to entry into the study. Further, tumors were also classified into four categories (squamous, small, adeno, and large). Only

9 of the 137 survival times were censored. Both authors suggested that the Weibull regression model is appropriate for analyzing this data. Further, it was suggested that performance status and tumor type are the most important factors and the effect of other variables may be ignored. They fitted both full and reduced models for estimation purposes.

The statistical problem here is, should we employ either the full or reduced model or both, for the further inferential purposes? We systematically address this issue and suggest the estimation strategies which improve on both components by invoking shrinkage techniques. It is well documented in the literature that when the parameter space is being reduced, estimation of regression parameters are generally improved. On the other hand, incorrect or imprecise restrictions on $\boldsymbol{\beta}$ may lead to biased (or even inconsistent) and inefficient estimators of $\boldsymbol{\beta}$. Ahmed and Saleh (1999) studied the properties of these estimators for the exponential regression censored model. Recently, a family of penalized partial likelihood methods, such as LASSO, are proposed by Tibshirani (1996) for variable selection for linear models and was further extended for the Cox proportional hazard models in Tibshirani (1997).

In this chapter we considered the integrated estimation problem for regression coefficients (both point and interval) in a Weibull regression censored model by exploiting the shrinkage estimation. Most of the reviewed literature in this arena do not deal with confidence interval problems, so we provide a total inferential package to practitioners.

The rest of Chapter 2 is organized as following. Section 2.2 introduces some notation and preliminaries for estimation of the Weibull regression model. In Section 2.3 we introduce integrated estimation. Section 2.4 showcases our main results and provides the analysis of bias and and risk comparison of the proposed estimators with the classical estimator. In Section 2.5 we present the results from our simulation study comparing the risk performance of the estimators. Interval estimation via the bootstrap method is given in Section 2.6. Finally, in Section 2.7 we apply our method

to VA lung cancer data. Concluding remarks are given in Section 2.8 to summarize the findings. Throughout this chapter, the boldface symbols represent vectors/matrices.

2.2 Notation and Preliminaries

Let $T_1, T_2, \cdots, T_n$ denote independent life length or lifetime measurements from a population modeled by Weibull distribution. Then T has the following *probability density function (pdf)* $f_T(t) = \lambda\nu(\lambda t)^{\nu-1}e^{-(\lambda t)^{\nu}}$, with $t \geq 0,\ \lambda > 0$, where λ and ν are the scale and shape parameters respectively. The survival probability of the individual is then given by

$$S(t|\lambda, \nu) = exp\left[-(\lambda t)^{\nu}\right],$$

where $\lambda \geq 0$ and $\nu \geq 0$. This is a plausible model for the lifetime T in the absence of any explanatory variable that may affect the lifetime, i.e., for the baseline distribution of T. But in the presence of concomitant information, we can extend the Weibull model by allowing the parameters λ and ν to depend on explanatory variables. Let $\mathbf{x} = (x_1, \cdots, x_k)$ be a vector of covariates for an individual. The most commonly used form of the Weibull model assumes that the covariates change only the scale of the baseline distribution while still maintaining the shape of the distribution. That is, the scale parameter λ can be modeled as a function of the covariates and the shape parameter ν is fixed but unknown. In this case the probability density function of T, given $\mathbf{x}$, for the individual with $\lambda = 1/\alpha$ is

$$f_T(t|\mathbf{x}) = \frac{\nu}{\alpha(\mathbf{x})}\left(\frac{t}{\alpha(\mathbf{x})}\right)^{\nu-1} e^{-\left(\frac{t}{\alpha(\mathbf{x})}\right)^{\nu}}, \quad t \geq 0,\ \alpha > 0, \tag{2.2}$$

where $\alpha(\mathbf{x})$ is a function of $\mathbf{x} = (x_1, \cdots, x_k)$ involving unknown parameters. We consider here the most useful form of $\alpha(\mathbf{x})$ given by

$$\alpha(\mathbf{x}) = exp(\mathbf{x}\boldsymbol{\beta}), \tag{2.3}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_k)'$ is a vector of regression parameters.

We consider a situation where lifetimes T_i may be subject to a fixed censoring. Specifically, we suppose that each individual has a lifetime and a censoring time. However, only the smaller of lifetime and censoring time is observed. In addition it is assumed that for each individual has a fixed censoring time $L_i > 0$, and a regression vector $\mathbf{x}_i = (x_{i_1}, \cdots x_{i_k})$. Therefore the censored data consists of the following pairs

$$(t_i, \gamma_i) \quad i = 1, 2, \cdots, n, \quad t_i = min(T_i, L_i),$$

and

$$\gamma_i = \begin{cases} 1 & \text{if } T_i < L_i \\ 0 & \text{if } T_i \geq L_i. \end{cases}$$

Noting that the lifetime model given in relation (2.2) is a proportional hazards model, which can be viewed as a location-scale model with a *log* transformation on the random variable T. Usually, we deal with *ln* lifetimes, $Y_i = lnT_i$. Thus, from relations (2.2) and (2.3) with $\sigma = 1/\nu$, we obtain the probability density function of Y, given $\mathbf{x}$ as follows.

$$f_Y(y|\mathbf{x}) = \frac{1}{\sigma} exp\left[\frac{y - \mathbf{x}\boldsymbol{\beta}}{\sigma} - exp\left(\frac{y - \mathbf{x}\boldsymbol{\beta}}{\sigma}\right)\right] \quad I(-\infty < y < +\infty).$$

Alternatively, the above model may be rewritten as

$$Y = \mathbf{x}\boldsymbol{\beta} + \sigma\varepsilon, \tag{2.4}$$

where ε is assumed to have a standard extreme value distribution with *pdf* given by

$$f(\varepsilon) = exp(\varepsilon - e^{\varepsilon}) \quad I(-\infty < \varepsilon < +\infty).$$

The likelihood function under the model (2.4) is based on the logarithm observations of the sample is

$$L(\boldsymbol{\beta}, \sigma) = \prod_{l=1}^{n} \left[\frac{1}{\sigma} exp\left\{ \frac{y_l - \mathbf{x}_l\boldsymbol{\beta}}{\sigma} - exp\left(\frac{y_l - \mathbf{x}_l\boldsymbol{\beta}}{\sigma} \right) \right\} \right]^{\gamma_l} \times \left[exp\left\{ -exp\left(\frac{y_l - \mathbf{x}_l\boldsymbol{\beta}}{\sigma} \right) \right\} \right]^{1-\gamma_l}. \tag{2.5}$$

Let C and D be the indexed sets including the censored and non-censored individuals respectively. Then (2.5) reduces to

$$L(\boldsymbol{\beta}, \sigma) = \prod_{l \in D} \frac{1}{\sigma} exp\left[\frac{y_l - \mathbf{x}_l\boldsymbol{\beta}}{\sigma} - exp\left(\frac{y_l - \mathbf{x}_l\boldsymbol{\beta}}{\sigma} \right) \right] \times \prod_{l \in C} exp\left\{ -exp\left(\frac{y_l - \mathbf{x}_l\boldsymbol{\beta}}{\sigma} \right) \right\},$$

and thus, the log likelihood function based on n observations for a censored sample can be written as

$$lnL(\boldsymbol{\beta}, \sigma) = \ell_n(\boldsymbol{\beta}, \sigma) = -d\ ln(\sigma) + \sum_{l \in D} \left(\frac{y_l - \mathbf{x}_l\boldsymbol{\beta}}{\sigma} \right) - \sum_{l=1}^{n} exp\left(\frac{y_l - \mathbf{x}_l\boldsymbol{\beta}}{\sigma} \right), \tag{2.6}$$

where d is the number of observed failures.

Let $z_l = \frac{y_l - \mathbf{x}_l\boldsymbol{\beta}}{\sigma}$. Then the maximum likelihood estimates of $\boldsymbol{\beta}$ and σ are obtained by solving the following system of equations

$$\frac{\partial \ell_n}{\partial \beta_r} = -\frac{1}{\sigma} \sum_{l \in D} x_{lr} + \frac{1}{\sigma} \sum_{l=1}^{n} x_{lr} e^{z_l} = 0, \quad \text{for } r = 1, \cdots, k, \tag{2.7}$$

$$\frac{\partial \ell_n}{\partial \sigma} = -\frac{d}{\sigma} - \frac{1}{\sigma} \sum_{l \in D} z_l + \frac{1}{\sigma} \sum_{l=1}^{n} z_l e^{z_l} = 0. \tag{2.8}$$

The maximum likelihood estimates of $\boldsymbol{\beta}$ and σ are the solutions of the system of $k+1$ equations (2.7) and (2.8). Those can be solved by the Newton-Raphson iterative

algorithm.

Also, the second derivatives of log-likelihood of (2.6) are given by

$$\begin{aligned}
-\frac{\partial^2 \ell_n}{\partial \beta_r \partial \beta_s} &= \frac{1}{\sigma^2}\sum_{l=1}^{n} x_{lr}x_{ls}e^{z_l} \quad \text{for } r,s = 1,\cdots,k,\\
-\frac{\partial^2 \ell_n}{\partial \sigma^2} &= \frac{1}{\sigma^2}\left[-d - 2\sum_{l\in D} z_l + 2\sum_{l=1}^{n} z_l e^{z_l} + 2\sum_{l=1}^{n} {z_l}^2 e^{z_l}\right],\\
-\frac{\partial^2 \ell_n}{\partial \beta_r \partial \sigma} &= \frac{1}{\sigma^2}\left[-\sum_{l\in D} x_{lr} + \sum_{l=1}^{n} x_{lr}e^{z_l} + \sum_{l=1}^{n} x_{lr}z_l e^{z_l}\right] \quad \text{for } r = 1,\cdots,k.
\end{aligned}$$

The observed information matrix $\mathbf{I}$ is $(k+1)\times(k+1)$ and of partitioned form

$$\mathbf{I} = -\begin{pmatrix} \frac{\partial^2 \ell_n}{\partial \beta_r \beta_s} & \frac{\partial^2 \ell_n}{\partial \beta_r \partial \sigma} \\ \frac{\partial^2 \ell_n}{\partial \beta_r \partial \sigma} & \frac{\partial^2 \ell_n}{\partial \sigma^2} \end{pmatrix}_{(\hat{\boldsymbol{\beta}},\hat{\sigma})}.$$

Further, the expected information matrix can be calculated when fixed censoring time is known.

2.3 Integrated Estimation

The statistical objective is to estimate parameter vector $\boldsymbol{\beta}$, when NSI is available. The *unconstrained maximum likelihood (uml)* estimator of $\boldsymbol{\beta}$ can be obtained by solving the system of equations in (2.7). Note that the unrestricted estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is based on sample data only and does not incorporate nonsample information in estimating $\boldsymbol{\beta}$. However, it may be advantageous to use the available nonsample information to obtain improved estimates.

Further, for inference purposes when there is censoring, it is convenient to use the asymptotic distribution of $\hat{\boldsymbol{\beta}}$, i.e.,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, {\mathbf{I_0}}^{-1}), \quad \xrightarrow{\mathcal{L}} \text{ means converges in distribution}$$

where $\mathbf{I_0} = \lim_{n\to\infty} \frac{\mathbf{I}^0}{n}$ and suppose that $\mathbf{I_0}$ is invertible. The matrix $\mathbf{I}^0$ is the observed information of order $k \times k$, with last row and last column of the matrix $\mathbf{I}$ deleted.

Let $\tilde{\boldsymbol{\beta}}$ be the *constrained maximum likelihood (cml)* estimator of $\boldsymbol{\beta}$ when the NSI in (2.1) is correct. Since $\hat{\boldsymbol{\beta}}$ is asymptotic normal, then

$$\sqrt{n}\mathbf{H}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \longrightarrow N(0, \mathbf{H}\mathbf{I_0}^{-1}\mathbf{H}').$$

From the asymptotic normal approximation to the distribution of $\hat{\boldsymbol{\beta}}$, the first order log likelihood for $\boldsymbol{\beta}$ can be written as

$$-lnL(\boldsymbol{\beta}) \triangleq (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{I_0}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

Using the result of subsection $1f.1$ [Rao (1973), p.60], the above *ln* likelihood is minimized under constraints $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ at

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \mathbf{I_0}^{-1}\mathbf{H}'(\mathbf{H}\mathbf{I_0}^{-1}\mathbf{H}')^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h}).$$

Having defined $\tilde{\boldsymbol{\beta}}$, note that if the restrictions are correct, then $\tilde{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ and will be superior to $\hat{\boldsymbol{\beta}}$. However, this may not always be the case and the said improvement would raise the ante of imprecise estimation due to a large amount of the bias inherent in such estimator.

A natural way to balance the potential bias of the estimator under the restriction against the benchmark estimator is to take a weighted average of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$. Such integrated or composite estimators may be written as

$$\hat{\boldsymbol{\beta}}^I = \phi\hat{\boldsymbol{\beta}} + (1 - \phi)\tilde{\boldsymbol{\beta}}, \tag{2.9}$$

for a judiciously chosen weight ϕ $(0 \leq \phi \leq 1)$. Many of the estimators proposed in the reviewed literature, both design-based and model-based, have the integrated form (2.9). This can be viewed as a pure shrinkage estimator. So $\tilde{\boldsymbol{\beta}}$ is a special case

of $\hat{\boldsymbol{\beta}}^I(\phi = 0)$. Bickel (1984) showed that in parametric models, such estimates are asymptotically optimal in a minimax sense and we conjectured the same result for Weibull regression model. A major drawback of $\hat{\boldsymbol{\beta}}^I$ is that it is not uniformly better than either component estimator in terms of mean squared error (MSE). In passing we would like to remark here that integrated estimators are popular in small area estimation; refer to Falorsi *et al.* (1994) and others.

Another approach to composite estimation is to employ James-Stein or shrinkage type estimation method, which in turn yields the optimal weight for $\hat{\boldsymbol{\beta}}^I$. Stein (1956) and James and Stein (1961) presented an explicit form of an estimator which dominates the usual maximum likelihood estimator in a multi-parameter situation. This procedure has attracted a lot of attention in the mainstream statistical literature as evident by numerous publications. Efron and Morris (1975) gave an excellent expository account of the shrinkage methodology as well as examples of practical applications, including the popular example of batting averages of baseball players.

In an effort to obtain a shrinkage type estimator for the problem at hand, we use the likelihood ratio statistic as a first step, given as:

$$\Lambda = 2[lnL(\hat{\boldsymbol{\beta}}) - lnL(\tilde{\boldsymbol{\beta}})] = n(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h})'\boldsymbol{\Omega}(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h}) + o_p(1), \tag{2.10}$$

where $\boldsymbol{\Omega} = (\mathbf{H}\mathbf{I_0}^{-1}\mathbf{H}')^{-1}$.

2.3.1 Shrinkage Estimator

If $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ represents a set of $q \leq k$ independent linear restrictions on $\boldsymbol{\beta}$, then the shrinkage (SE) estimator that combines the sample and non-sample information can be defined as:

$$\hat{\boldsymbol{\beta}}^S = \tilde{\boldsymbol{\beta}} + \left(1 - (q-2)\Lambda^{-1}\right)(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}), \quad q \geq 3. \tag{2.11}$$

Since it shrinks the *uml* estimator towards $\tilde{\boldsymbol{\beta}}$, this estimator is generally called a shrinkage estimator. Clearly, if the restrictions $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ is true, then the likelihood

ratio test statistic is asymptotically distributed as χ^2_q. In this case the value of the test statistic will be small and a relatively large weight is placed on the restricted estimator $\tilde{\boldsymbol{\beta}}$. Otherwise, the value of Λ is relatively large and more weight is placed on $\hat{\boldsymbol{\beta}}$. Consequentially, $\hat{\boldsymbol{\beta}}^S$ is a special case of $\hat{\boldsymbol{\beta}}^I$ with $\phi = (q-2)\Lambda^{-1}$. Some of the salient features of the shrinkage procedure are:

a) The shrinkage estimation strategy is attractive to users wanting good estimation for the problem at hand because large gains in efficiency can be achieved in the classical full model-based framework without assuming the correctness of the reduced model.

b) This estimator is, in general, a biased estimator, although bias is accompanied by reduction in risk, and hence, does not have a serious impact on risk assessment.

c) In many situations the shrinkage estimator arises quite naturally in the empirical Bayes (EB) approach or the empirical best linear unbiased prediction (EBLUP) approach.

However, an unpleasant feature of this estimator is that it may over-shrink $\hat{\boldsymbol{\beta}}$ towards the $\tilde{\boldsymbol{\beta}}$, thus causing a possible inversion of the sign of the benchmark estimator. Here, if $\Lambda \leq (q-2)$, the proposed shrinkage estimator reverses the sign of estimator $\hat{\boldsymbol{\beta}}$. This problem is resolved by the use of the "positive rule" estimator $\hat{\boldsymbol{\beta}}^{S+}$, which preserves the sign of estimates.

2.3.2 Positive Shrinkage Estimator

A *positive shrinkage estimator (PSE)* $\hat{\boldsymbol{\beta}}^{S+}$ is obtained from (2.11) by changing the factor $1-(q-2)/\Lambda$ to 0 whenever $\Lambda \leq (q-2)$, that is,

$$\hat{\boldsymbol{\beta}}^{S+} = \tilde{\boldsymbol{\beta}} + \left(1-(q-2)\Lambda^{-1}\right)^+ (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}), \tag{2.12}$$

where $z^+ = max(0, z)$. The PSE is particularly important to control the over-shrinking inherent in $\hat{\boldsymbol{\beta}}^S$. The estimator $\hat{\boldsymbol{\beta}}^{S+}$ dominates $\hat{\boldsymbol{\beta}}^S$ in terms of total MSE and hence, is useful for practical purposes. For this reason, Ahmed (2001) recommended that the shrinkage estimator should be used as a tool for developing the PSE and should not be used as an estimator in its own right. In an effort to see that $\hat{\boldsymbol{\beta}}^{S+}$ is a special case of $\hat{\boldsymbol{\beta}}^I$, we re-write $\hat{\boldsymbol{\beta}}^{S+}$ in (2.12) in the following canonical form:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{S+} &= \hat{\boldsymbol{\beta}} - (q-2)\Lambda^{-1}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) - \\ &\quad \{1 - (q-2)\Lambda^{-1}\} I(\Lambda < q-2)(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}),\end{aligned}$$

which in turns give

$$\phi = (1 - (q-2)\Lambda^{-1})\{I((q-2)\Lambda^{-1} \leq 1)\}.$$

In parametric setups, the SE, PSE and other related estimators have been extensively studied [Judge and Bock (1978), Ahmed and Ullah (1999) and the references cited there]. Large sample properties of these estimators were studied by Sen (1986), Ahmed (2005), Ahmed (1992), Ahmed (2001), Ahmed *et al.* (2006*a*) and others. Stigler (1990) and Kubokawa (1998) provide excellent reviews of (parametric) shrinkage estimators.

2.3.3 Pretesting and LASSO

Some alternative estimators to the shrinkage estimators are based on pretesting and LASSO methods. Bancroft (1944) proposed an idea in the estimation of a regression model under a pretest for some linear restrictions (2.1) on the coefficients which is considered as a hypothesis. See Khan and Ahmed (2006) for some recent developments. In this set up, one can perform the pretest using an appropriate test statistic. If the test rejects the hypothesis, the unrestricted estimator will be used. If the

test fails to reject the hypothesis, the restricted estimator is selected. In the present context, the pretest estimator $\hat{\boldsymbol{\beta}}^{PT}$ is defined as:

$$\hat{\boldsymbol{\beta}}^{PT} = \hat{\boldsymbol{\beta}} - (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})I(\Lambda < c_\alpha),$$

where c_α is some predetermined 'critical value' for the test statistic Λ.

This "pretest" approach to estimation is unsatisfactory from several points of view. From a decision theoretic viewpoint, the discontinuity in estimation brought about by the hypothesis testing dichotomy means this method cannot be admissible. The pretest procedure often produced poor estimates. The risk function of the pretest estimator added to this poor performance. The risk exceeded the minimax bound (the risk of UE) over a substantial region of the parameter space. By contrast, procedures amongst a class of minimax estimators introduced by James and Stein (1961) achieved low risk when the reduced model was correct without sacrificing precision when the adequacy of the model was uncertain. Sclove *et al.* (1972) studied the properties of pretest estimators in linear models. They suggested another pretest estimator by replacing the restricted estimator by a shrinkage estimator. However, for bivariate data, pretest estimation is the only alternative to unrestricted and restricted estimation.

The LASSO [Tibshirani (1996)] is a method for regularizing a least squares regression. In the context of censored data, Tibshirani (1997) extended the LASSO procedure to variable selection with the Cox proportional hazard model. Huang *et al.* (2006) considered this LASSO procedure for variable selection and estimation in an accelerated failure time model with high-dimensional covariates based on Stute's weighted least squares method [Stute (1996)]. They proposed to minimize the weighted least square objective function

$$\hat{\boldsymbol{\beta}}_\tau = \min_{\boldsymbol{\beta}} \left[\frac{1}{2} \sum_{i=1}^{n} (Y_{(i)} - \mathbf{X}_{(i)}\boldsymbol{\beta})^2 \right] \quad \text{subject to} \quad \sum_{i=1}^{k} |\beta_i| \leq \tau,$$

where τ is a tuning parameter and $Y_{(i)}$ and $\mathbf{X}_{(i)}$, are defined in Huang *et al.* (2006). The tuning parameter determines how many estimated coefficients are zero. The LASSO is computed by quadratic programming techniques, and the tuning parameter is selected using cross-validation and/or generalized cross-validation. Note that the output of the LASSO resembles shrinkage and pretest methods by both shrinking and deleting coefficients. However, it is different from pretest and shrinkage procedures and it treats all the covariate coefficients equally.

In the present investigation, we are concentrating on shrinkage estimation and the LASSO method is still an ongoing research. The proposed estimators are easy to compute and implement. The objective is to produce natural adaptive estimators that are free of subjective choices and tuning parameters.

2.4 Asymptotic Bias and Risk Comparisons

We note that, as the test statistic Λ is consistent against fixed $\boldsymbol{\beta}$ such that $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$, the SE and PSE will be asymptotically equivalent in probability to $\hat{\boldsymbol{\beta}}$, for the fixed alternative (up to the order $O(n^{-1/2})$), so that asymptotically there is no shrinkage effect. Hence, in the large sample situation there is not much to investigate. This brings us to the usual Pitman type of alternatives

$$K_n : \mathbf{H}\boldsymbol{\beta} = \mathbf{h} + \frac{\boldsymbol{\delta}}{n^{1/2}}, \tag{2.13}$$

where $\boldsymbol{\delta} = (\delta_1, \delta_2 \cdots, \delta_q) \in \Re^q$ is a real fixed vector. Note that for $\boldsymbol{\delta} = \mathbf{0}$, $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ for all n. Thus, the relation (2.1) is a particular case of (2.13). Even for such local alternatives, the SE and PSE may not be asymptotically unbiased estimators of $\boldsymbol{\beta}$. With that in mind, we introduce the following:

$$\mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\beta}; \mathbf{Q}) = \left[\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})\right]' \mathbf{Q} \left[\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})\right], \tag{2.14}$$

where $\mathbf{Q}$ is a positive semidefinite weighting matrix and $\boldsymbol{\beta}^*$ is any one of $\hat{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}^S$ or $\hat{\boldsymbol{\beta}}^{S+}$. If we take the expected loss, using (2.14) and the distribution of $\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})$, that would be called the *quadratic risk* $R_n^0(\boldsymbol{\beta}^*, \boldsymbol{\beta}; \mathbf{Q})$ $(= \text{trace}(\mathbf{Q}\hat{\boldsymbol{\Sigma}}_n)$, where $\hat{\boldsymbol{\Sigma}}_n$ is the covariance matrix of $\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})$. Whenever $\lim_{n\to\infty} \hat{\boldsymbol{\Sigma}}_n = \hat{\boldsymbol{\Sigma}}$ exists, $R_n^0(\boldsymbol{\beta}^*, \boldsymbol{\beta}; \mathbf{Q}) \to R^0(\boldsymbol{\beta}^*, \boldsymbol{\beta}; \mathbf{Q}) = \text{trace}(\mathbf{Q}\hat{\boldsymbol{\Sigma}})$, which is termed the *asymptotic risk*. In our setup, we denote the distribution of $\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})$ by $\tilde{G}_n(\mathbf{u})$, $\mathbf{u} \in \Re^q$. Suppose that $\tilde{G}_n \to \tilde{G}$ (at all points of continuity) as $n \to \infty$. Let $\boldsymbol{\Sigma}_{\tilde{G}}$ be the covariance matrix of $\tilde{G}$. Then the asymptotic distributional risk (ADR) of $\boldsymbol{\beta}^*$ is defined as $R(\boldsymbol{\beta}^*; \mathbf{Q}) = \text{trace}(\mathbf{Q}\boldsymbol{\Sigma}_{\tilde{G}})$. We shall work with the ADR results in the following discussions. In this vein, we define the asymptotic bias as $\mathbf{B}^0(\boldsymbol{\beta}^*, \boldsymbol{\beta}) = E[\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})]$ and side by side, the *asymptotic distributional bias* (ADB) as the limit

$$\int \cdots \int \mathbf{x} d\tilde{G}_n(\mathbf{x}) \left(\to \mathbf{B}(\boldsymbol{\beta}^*, \boldsymbol{\beta}) = \int \cdots \int \mathbf{x} d\tilde{G}(\mathbf{x}) \right).$$

Two central results key to the study of ADR and ADB of the SE, and PSE are given in the following theorem under the following regularity conditions:

Condition 1. The failure time T is independent of the examination times given the covariates.

Condition 2. The log-likelihood function $\ell_n(\boldsymbol{\beta}, \sigma)$ is twice differentiable, and the third derivatives must be bounded.

Condition 3. The information matrix $\mathbf{I_0}$ is invertible.

Theorem 2.4.1. *Under local alternative and the usual regularity conditions, we have the following as $n \to \infty$:*

1. $\sqrt{n}(\mathbf{H}\hat{\boldsymbol{\beta}} - \boldsymbol{h}) \xrightarrow{\mathcal{L}} N(\boldsymbol{\delta}, \mathbf{H}\mathbf{I_0}^{-1}\mathbf{H}')$.

2. Λ *converges to a non-central chi-squared distribution with q degrees of freedom and non-centrality parameter $\Delta = \boldsymbol{\delta}'\boldsymbol{\Omega}\boldsymbol{\delta}$ where $\boldsymbol{\Omega}$ is defined in (2.10).*

With the above theorems, we are in a position to use the results on the parametric

model, and thereby arrive at the main results of this section. For parallel results we refer to Ahmed and Saleh (1999), further unified in Theorem 1 of Ahmed (2001, p.108). Therefore, we present (without derivation) the results on SE and PSE.

2.4.1 Asymptotic Distributional Bias

In the following theorem, we present expressions for the bias of the proposed estimators. Note that $\hat{\boldsymbol{\beta}}$ is an asymptotically unbiased estimator.

Theorem 2.4.2. *Under local alternatives K_n in (2.13) and assume that the Theorem 2.4.1 holds, we have the ADB of the proposed estimators as $n \to \infty$, in the following:*

$$
\begin{aligned}
ADB(\tilde{\boldsymbol{\beta}}) &= -\mathbf{A}\boldsymbol{\delta}, \quad \mathbf{A} = \mathbf{I_0}^{-1}\mathbf{H}'(\mathbf{H}\mathbf{I_0}^{-1}\mathbf{H}')^{-1}, && (2.15)\\
ADB(\hat{\boldsymbol{\beta}}^{S}) &= -(q-2)\mathbf{A}\boldsymbol{\delta}\; E\left[\chi^{-2}_{q+2}(\Delta)\right], && (2.16)\\
ADB(\hat{\boldsymbol{\beta}}^{S+}) &= -(q-2)\mathbf{A}\boldsymbol{\delta}\left[E(\chi^{-2}_{q+2}(\Delta)) - E(\chi^{-2}_{q+2}(\Delta)I(\chi^{2}_{q+2}(\Delta) < (q-2)))\right]\\
&\quad - \mathbf{A}\boldsymbol{\delta}\Psi_{q+2}(q-2,\Delta), && (2.17)
\end{aligned}
$$

where the notation $\Psi_\nu(q-2,\Delta)$ is the cumulative distribution function of a non-central chi-square distribution with ν degrees of freedom and non-centrality parameter Δ.

The bias expressions reveal that all three integrated estimators are asymptotically biased. However, both shrinkage estimators are bounded in Δ as opposed to $\tilde{\boldsymbol{\beta}}$.

Theorem 2.4.3. *Under local alternatives K_n in (2.13) and assume that the Theorem 2.4.1 holds, we obtain the risk function of the proposed estimators as $n \to \infty$, in the following:*

$$
\begin{aligned}
R(\tilde{\boldsymbol{\beta}};\mathbf{Q}) &= R(\hat{\boldsymbol{\beta}};\mathbf{Q}) - \mathit{trace}[\mathbf{Q}\mathbf{A}\mathbf{H}\mathbf{I_0}^{-1}] + \boldsymbol{\delta}'\mathbf{M}\boldsymbol{\delta}, \quad \mathbf{M} = \mathbf{A}'\mathbf{Q}\mathbf{A},\\
&\quad \textit{with } R(\hat{\boldsymbol{\beta}}) = \mathit{trace}[\mathbf{Q}\mathbf{I_0}^{-1}], && (2.18)\\
R(\hat{\boldsymbol{\beta}}^{S};\mathbf{Q}) &= R(\hat{\boldsymbol{\beta}};\mathbf{Q}) - (q-2)\boldsymbol{\delta}'\mathbf{M}\boldsymbol{\delta}\left[2E(\chi^{-2}_{q+4}(\Delta)) - 2E(\chi^{-2}_{q+2}(\Delta))\right]
\end{aligned}
$$

$$+ \quad (q-2)\,trace[\mathbf{QAHI_0}^{-1}]\left[(q-2)E(\chi_{q+2}^{-4}(\Delta)) - 2E(\chi_{q+2}^{-2}(\Delta))\right]$$

$$+ \quad (q-2)^2\boldsymbol{\delta}'\mathbf{M}\boldsymbol{\delta}\; E\left[(\chi_{q+4}^{-4}(\Delta))\right], \tag{2.19}$$

$$\begin{aligned} R(\hat{\boldsymbol{\beta}}^{S+};\mathbf{Q}) &= R(\hat{\boldsymbol{\beta}}^{S};\mathbf{Q}) - \boldsymbol{\delta}'\mathbf{M}\boldsymbol{\delta}\; E[(1-(q-2)\chi_{q+4}^{-2}(\Delta))^2 I(\chi_{q+4}^2(\Delta) < (q-2))] \\ &- trace[\mathbf{QAHI_0}^{-1}]E[(1-(q-2)\chi_{q+2}^{-2}(\Delta))^2 I(\chi_{q+2}^2(\Delta) < (q-2))] \\ &+ 2\;\boldsymbol{\delta}'\mathbf{M}\boldsymbol{\delta}\; E[(1-(q-2)\chi_{q+2}^{-2}(\Delta))I(\chi_{q+2}^2(\Delta) < (q-2))]. \end{aligned} \tag{2.20}$$

Proof: See details in Chapter 3.

Risk comparison:

The risk of all the integrated estimators depend on $\boldsymbol{\delta}\mathbf{M}\boldsymbol{\delta}'$. Note that $\mathbf{I_0}^{-1/2}\mathbf{H}'$ $\times$ $(\mathbf{HI_0}^{-1}\mathbf{H}')^{-1}$ $\mathbf{HI_0}^{-1/2}$ is a symmetric idempotent matrix with rank $q(\leq k)$. Thus there exists an orthogonal matrix $\boldsymbol{\Gamma}$ such that

$$\boldsymbol{\Gamma}\mathbf{I_0}^{-1/2}\mathbf{H}^T(\mathbf{HI_0}^{-1}\mathbf{H}^T)^{-1}\mathbf{HI_0}^{-1/2}\boldsymbol{\Gamma}' = \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{k-q} \end{bmatrix},$$

$$\boldsymbol{\Gamma}\mathbf{I_0}^{-1/2}\mathbf{Q}\mathbf{I_0}^{-1/2}\boldsymbol{\Gamma}' = \begin{bmatrix} \mathbf{c}_{11} & \mathbf{c}_{12} \\ \mathbf{c}_{21} & \mathbf{c}_{22} \end{bmatrix}.$$

So trace $\left[\mathbf{Q}\mathbf{I_0}^{-1}\mathbf{H}^T(\mathbf{HI_0}^{-1}\mathbf{H}^T)^{-1}\mathbf{HI_0}^{-1}\right]$

$$\begin{aligned} &= \text{trace}\left[(\boldsymbol{\Gamma}\mathbf{I_0}^{-1/2}\mathbf{Q}\mathbf{I_0}^{-1/2}\boldsymbol{\Gamma}')(\boldsymbol{\Gamma}\mathbf{I_0}^{-1/2}\mathbf{H}^T(\mathbf{HI_0}^{-1}\mathbf{H}^T)^{-1}\mathbf{HI_0}^{-1/2}\boldsymbol{\Gamma}')\right] \\ &= \text{trace}\left(\begin{bmatrix} \mathbf{c}_{11} & \mathbf{c}_{12} \\ \mathbf{c}_{21} & \mathbf{c}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{k-q} \end{bmatrix}\right) \\ &= \text{trace}(\mathbf{c}_{11}), \end{aligned}$$

where the matrices $\mathbf{c}_{11}$ and $\mathbf{c}_{12}$ are of order q and $k-q$ respectively. Further $\boldsymbol{\delta}'(\mathbf{HI_0}^{-1}\mathbf{H}^T)^{-1}\mathbf{HI_0}^{-1}\mathbf{Q}\mathbf{I_0}^{-1}\mathbf{H}'(\mathbf{HI_0}^{-1}\mathbf{H}^T)^{-1}\boldsymbol{\delta}$

$$= \left[\boldsymbol{\delta}'(\mathbf{HI_0}^{-1}\mathbf{H}^T)^{-1}\mathbf{HI_0}^{-1/2}\boldsymbol{\Gamma}'\right] \times \left[\boldsymbol{\Gamma}\mathbf{I_0}^{-1/2}\mathbf{H}'(\mathbf{HI_0}^{-1}\mathbf{H}^T)^{-1}\mathbf{HI_0}^{-1/2}\boldsymbol{\Gamma}'\right]$$

$$
\begin{aligned}
&\times \left[\boldsymbol{\Gamma}{\mathbf{I_0}}^{-1/2}\mathbf{Q}{\mathbf{I_0}}^{-1/2}\boldsymbol{\Gamma}'\right] \times \left[\boldsymbol{\Gamma}{\mathbf{I_0}}^{-1/2}\mathbf{H}'(\mathbf{H}{\mathbf{I_0}}^{-1}\mathbf{H}^T)^{-1}\mathbf{H}{\mathbf{I_0}}^{-1/2}\boldsymbol{\Gamma}'\right] \\
&\times \left[\boldsymbol{\Gamma}{\mathbf{I_0}}^{-1/2}(\mathbf{H}{\mathbf{I_0}}^{-1}\mathbf{H}^T)^{-1}\mathbf{H}\boldsymbol{\delta}\right] \\
= \; & \boldsymbol{\eta}' \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{k-q} \end{bmatrix} \begin{bmatrix} \mathbf{c}_{11} & \mathbf{c}_{12} \\ \mathbf{c}_{21} & \mathbf{c}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{k-q} \end{bmatrix} \boldsymbol{\eta} \\
= \; & (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)' \begin{bmatrix} \mathbf{c}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{pmatrix} \\
= \; & {\boldsymbol{\eta}_1}'\mathbf{c}_{11}\boldsymbol{\eta}_1,
\end{aligned}
$$

where $\boldsymbol{\eta} = \boldsymbol{\delta}'(\mathbf{H}{\mathbf{I_0}}^{-1}\mathbf{H}^T)^{-1}\mathbf{H}{\mathbf{I_0}}^{-1/2}\boldsymbol{\Gamma}' = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)'$ and $\boldsymbol{\eta}_1$ is a $q \times 1$ vector of components of $\boldsymbol{\eta}$. Hence,

$$R(\tilde{\boldsymbol{\beta}}) = \text{trace}[\mathbf{Q}{\mathbf{I_0}}^{-1}] - \text{trace}(\mathbf{c}_{11}) + \boldsymbol{\eta}_1'\mathbf{c}_{11}\boldsymbol{\eta}_1. \tag{2.21}$$

Further, by Courant theorem [Saleh (2006), Theorem 5, p.39],

$$Ch_{min}(\mathbf{c}_{11}) \leq \frac{\boldsymbol{\eta}_1'\mathbf{c}_{11}\boldsymbol{\eta}_1}{\boldsymbol{\eta}_1'\boldsymbol{\eta}_1} \leq Ch_{max}(\mathbf{c}_{11}), \tag{2.22}$$

where $Ch_{min}(\mathbf{c}_{11})$ and $Ch_{max}(\mathbf{c}_{11})$ are the minimum and maximum characteristic roots of $\mathbf{c}_{11}$ and $\Delta = \boldsymbol{\delta}'\boldsymbol{\Omega}\boldsymbol{\delta} = \boldsymbol{\eta}_1'\boldsymbol{\eta}_1$. Therefore,

$$R(\tilde{\boldsymbol{\beta}}) - \text{trace}(\mathbf{c}_{11}) + Ch_{min}(\mathbf{c}_{11}) \leq R(\hat{\boldsymbol{\beta}}) \leq R(\tilde{\boldsymbol{\beta}}) - \text{trace}(\mathbf{c}_{11}) + Ch_{max}(\mathbf{c}_{11}).$$

The bounds are equal at $\Delta = 0$. Thus, for $\Delta \in \left[0, \frac{\text{trace}(\mathbf{c}_{11})}{Ch_{max}(\mathbf{c}_{11})}\right]$, $\tilde{\boldsymbol{\beta}}$ has smaller risk than that of $\hat{\boldsymbol{\beta}}$ and outside the interval, $\hat{\boldsymbol{\beta}}$ has smaller risk than $\tilde{\boldsymbol{\beta}}$. Clearly, when Δ moves away from null vector beyond the values of $\frac{\text{trace}(\mathbf{c}_{11})}{Ch_{max}(\mathbf{c}_{11})}$, the ADR of $\tilde{\boldsymbol{\beta}}$ increases and becomes unbounded. This clearly indicates that the performance of $\tilde{\boldsymbol{\beta}}$ will strongly depend on the validity of the restriction.

The risk difference of $\hat{\boldsymbol{\beta}}^S$ and $\hat{\boldsymbol{\beta}}$ is

$$
\begin{aligned}
= \; & (q-2)\text{trace}[\mathbf{Q}\mathbf{A}\mathbf{H}{\mathbf{I_0}}^{-1}][2E(\chi_{q+2}^{-2}(\Delta)) - (q-2)E(\chi_{q+2}^{-4}(\Delta))] \\
- \; & (q-2)^2\boldsymbol{\delta}'\mathbf{M}\boldsymbol{\delta}\; E(\chi_{q+4}^{-4}(\Delta)) + 2(q-2)\boldsymbol{\delta}'\mathbf{M}\boldsymbol{\delta}\left[E(\chi_{q+4}^{-2}(\Delta)) - E(\chi_{q+2}^{-2}(\Delta))\right]
\end{aligned}
$$

$$
\begin{aligned}
&= (q-2)^2 \text{trace}(\mathbf{c}_{11}) E(\chi_{q+2}^{-4}(\Delta)) + 2\Delta(q-2)\text{trace}(\mathbf{c}_{11}) E(\chi_{q+4}^{-4}(\Delta)) \\
&- (q^2-4)(\boldsymbol{\eta}_1' \mathbf{c}_{11} \boldsymbol{\eta}_1) E(\chi_{q+4}^{-4}(\Delta)) \\
&= (q-2)^2 \text{trace}(\mathbf{c}_{11}) E(\chi_{q+2}^{-4}(\Delta)) \\
&+ \left[1 - \frac{(q+2)(\boldsymbol{\eta}_1' \mathbf{c}_{11} \boldsymbol{\eta}_1)}{2\Delta \text{trace}(\mathbf{c}_{11})}\right] 2\Delta(q-2)\text{trace}(\mathbf{c}_{11}) E(\chi_{q+4}^{-4}(\Delta)).
\end{aligned}
$$

The above risk difference is positive when

$$
\frac{\text{trace}(\mathbf{c}_{11})}{Ch_{max}(\mathbf{c}_{11})} \geq \frac{q+2}{2} \quad \text{and} \quad q \geq 3.
$$

Thus, under the above condition, the risk of $\hat{\boldsymbol{\beta}}^S$ is less than or equal to the risk of $\hat{\boldsymbol{\beta}}$ in the entire parameter space. The maximum gain in risk is achieved near the restriction.

The risk difference of $\hat{\boldsymbol{\beta}}^S$ and $\hat{\boldsymbol{\beta}}^{S+}$ is

$$
\begin{aligned}
& R(\hat{\boldsymbol{\beta}}^{S+}) - R(\hat{\boldsymbol{\beta}}^S) \\
&= -\text{trace}[\mathbf{QAHI_0}^{-1}]\ E[(1-(q-2)\chi_{q+2}^{-2}(\Delta))^2 I(\chi_{q+2}^2(\Delta) < (q-2))] \\
&- \boldsymbol{\delta}'\mathbf{M}\boldsymbol{\delta}\ E[(1-(q-2)\chi_{q+4}^{-2}(\Delta))^2 I(\chi_{q+4}^2(\Delta) < (q-2))] \\
&- 2\ \boldsymbol{\delta}'\mathbf{M}\boldsymbol{\delta}\ E[((q-2)\chi_{q+2}^{-2}(\Delta) - 1) I(\chi_{q+2}^2(\Delta) < (q-2))].
\end{aligned}
$$

The right hand side of the above expression is positive semi-definite, since the expectation of a positive random variable is positive by definition of an indicator function,

$$
[q-2-\chi_{q+2}^2(\Delta)] I\left(\chi_{q+2}^2(\Delta) < q-2\right) \geq 0,
$$

Since $P[\chi_{q+2}^2(\Delta) > 0] = 1$, $[(q-2)\chi_{q+2}^{-2}(\Delta) - 1] I\left(\chi_{q+2}^2(\Delta) < q-2\right) \geq 0$.

Thus, for all Δ and $q \geq 3$

$$
R(\hat{\boldsymbol{\beta}}^{S+}) \leq R(\hat{\boldsymbol{\beta}}^S),
$$

with strict inequality for some Δ. Hence we can conclude that the proposed estimator $\hat{\boldsymbol{\beta}}^{S+}$ is asymptotically superior to $\hat{\boldsymbol{\beta}}^{S}$ and hence to $\hat{\boldsymbol{\beta}}$.

More importantly, for practical reasons and to support our theoretical findings we conducted an extensive simulation study to investigate the performance of the proposed estimators for moderate sample sizes. Our simulation experiments have provided strong evidence that corroborates the asymptotic theory which is given in the following section.

2.5 Simulation Studies

In this section, we carry out a Monte Carlo simulation study to examine risk (namely MSE) performance of the proposed estimators. Indeed, this simulation study is based on a Weibull regression censored model with different numbers of explanatory variables. The data were generated based on the fixed censoring model through the statistical software *R* and S-PLUS.

Our sampling experiment consists of different combinations of sample sizes i.e., $n = 50, 100, 150$. The proportions of censoring (pc) in the sample are pc=10%, 20%, 30% with the shape parameter $\nu = 2/3$, i.e., $\sigma = 3/2$. For simulation, we consider the particular case of our hypothesis $H_0 : \beta_j = 0$, for $j = p+1, \cdots, k$ with $k = p+q$. Under this hypothesis, we apply the same method as Bender *et al.* (2005), i.e., generating the survival and censoring time by using

$$lnT_i = \boldsymbol{\beta}\mathbf{x}_i + \sigma\varepsilon_i, \qquad \text{for } i = 1, 2, 3, \cdots, n,$$

where ε_i is generated from an extreme value distribution. We also generated 9 covariates from normal, uniform, exponential, binomial and Weibull distributions. We set the regression coefficients including intercept $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{\beta}_1, \mathbf{0})$ with $\boldsymbol{\beta}_1 = (3, 0.5, -2.5)$ to generate survival and censoring times. Those are fixed for each realization. We provide detailed results for $(p, q) = \{(3,3), (3,6), (3,7), (1,8)\}$ and

$\alpha = 0.05$.

Table 2.1: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 10\%, n = 50$ and $q = 3$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0	2.251	1.239	1.285
0.2	2.056	1.204	1.246
0.4	1.737	1.152	1.185
0.6	1.355	1.094	1.106
0.8	1.077	1.077	1.074
1.2	0.633	1.024	1.025
1.6	0.386	1.013	1.013
2	0.267	1.007	1.007
4	0.059	0.999	0.999

The number of simulations under the null hypothesis was varied initially and it was determined that 2000 for each set of observations were adequate, since a further increase in the number of realizations did not significantly change the results. We define the parameter $\Delta^\star = ||\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}||^2$, where $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_1, \mathbf{0})'$ and $||\cdot||$ is the Euclidian norm. In order to investigate the behavior of the estimators for $\Delta^\star > 0$, further samples were generated from those distributions under local alternative hypotheses (i.e., for different $\Delta^\star$ which lies between 0 and 6).

The performance of an estimator of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^*$, will be measured in terms of its total mean squared error risk. We have numerically calculated the risk of all the estimators studied in this chapter. The performance of the estimators was evaluated in terms of absolute relative bias (ARB) and relative MSE (RMSE). The simulated RMSE $\hat{\boldsymbol{\beta}}^*$ to $\hat{\boldsymbol{\beta}}$ is defined by

$$\text{RMSE}(\hat{\boldsymbol{\beta}} : \hat{\boldsymbol{\beta}}^*) = \frac{\text{simulated risk}(\hat{\boldsymbol{\beta}})}{\text{simulated risk}(\hat{\boldsymbol{\beta}}^*)},$$

keeping in mind that a RMSE larger than one indicates the degree of superiority of

Table 2.2: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 10\%, n = 50$ and $q = 6$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	2.251	1.239	1.285
0.2	2.056	1.204	1.246
0.4	1.737	1.152	1.185
0.6	1.355	1.094	1.106
0.8	1.077	1.071	1.074
1.2	0.633	1.024	1.025
1.6	0.386	1.013	1.013
2.0	0.267	1.007	1.007
4.0	0.059	1.000	1.000

the estimator over $\hat{\boldsymbol{\beta}}$.

We report the analysis based on the RMSE. The results are reported in Tables 2.1 to 2.27 (only for $q = 3, 6$ and 8) and Figures 2.1 to 2.9. The findings can be summarized as follows:

i) For all combinations of censoring levels and sample sizes, $\tilde{\boldsymbol{\beta}}$ outshines all the estimators at and near $\Delta^\star = 0$. On the contrary, when $\Delta^\star$ deviates from the origin, the estimated risk of $\tilde{\boldsymbol{\beta}}$ increases and becomes unbounded whereas the estimated risk of all other estimators remains bounded and approaches the risk of $\hat{\boldsymbol{\beta}}$ from below. It can be safely concluded that departure from the restriction is fatal to $\tilde{\boldsymbol{\beta}}$, but it has less impact on shrinkage estimators, which is consistent with the theory.

ii) If the number of variables $q = 3$ and the sample sizes are between 50 and 150, the RMSE of shrinkage estimators vary from 1.14 to 1.29 when restriction holds, and they increase with the increase of the number of variables q (consistent with theory). In particular if $p = 1$, $q = 8$, sample size=100, and $pc = 10\%$, the RMSE's of these estimators are 3.29 and 3.64 respectively, indicating a

Table 2.3: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 10\%, n = 50$ and $q = 8$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.00	73.901	2.946	3.131
0.02	21.862	2.689	2.870
0.03	11.998	2.338	2.385
0.06	3.413	1.525	1.532
0.08	1.889	1.334	1.334
0.10	1.227	1.209	1.209
0.13	0.695	1.123	1.123
0.15	0.513	1.088	1.088

remarkable performance of the proposed estimators. On the other hand, when the value of $\Delta^\star$, increases, the RMSE's of both estimators decrease and converge to 1 irrespective of p, q and sample size n. The figures also reveal that the shrinkage estimators work better in cases with more restrictions q.

iii) For all combinations of variables p and q, the performance of the shrinkage estimators depend on the magnitude of censoring percentage. Indeed, the lower the amount the censoring, the higher the gain in reduction of MSE. In other words, the risk of shrinkage estimators increases with an increase of the percent of censored observations irrespective of the sample sizes.

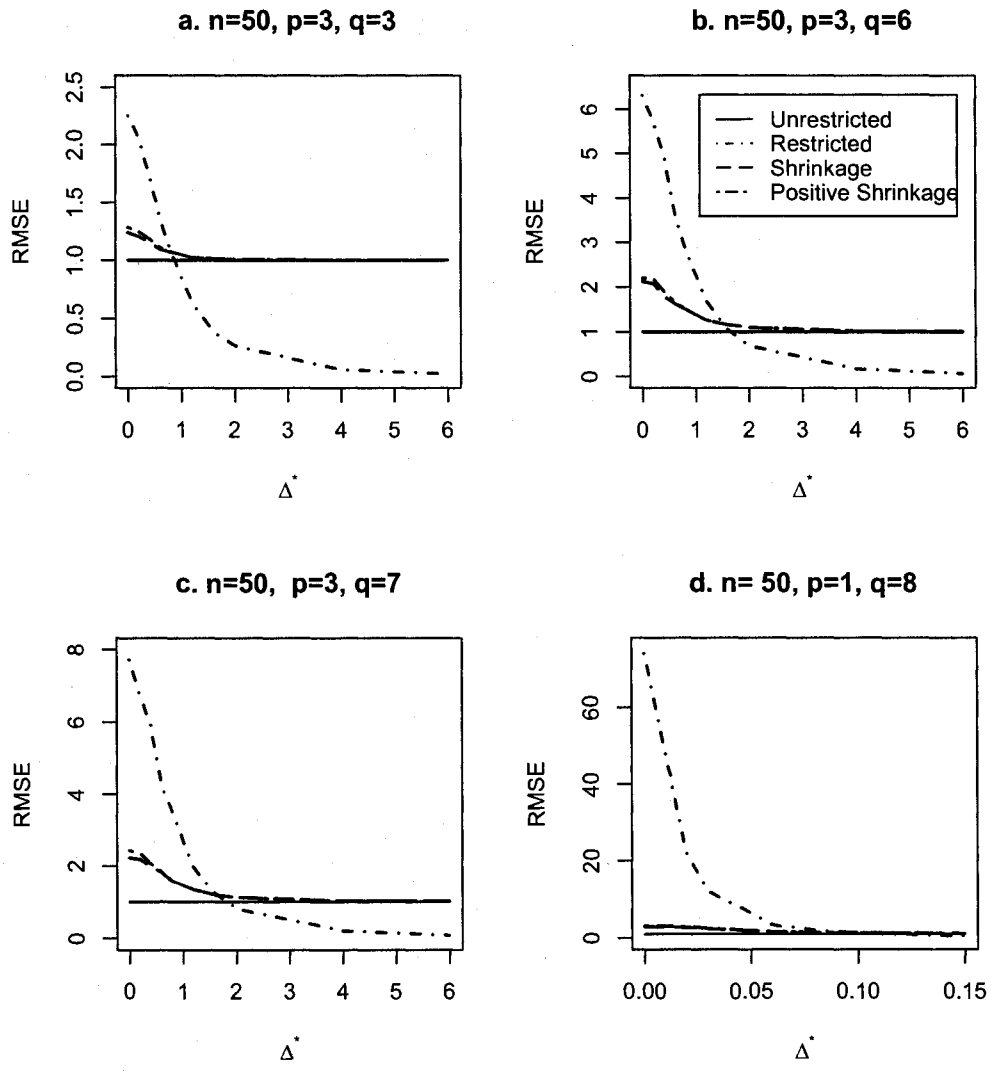

Figure 2.1: Simulated RMSE of the estimators as a function of the non-centrality parameter $\Delta^\star$ for different q and 10% censoring.

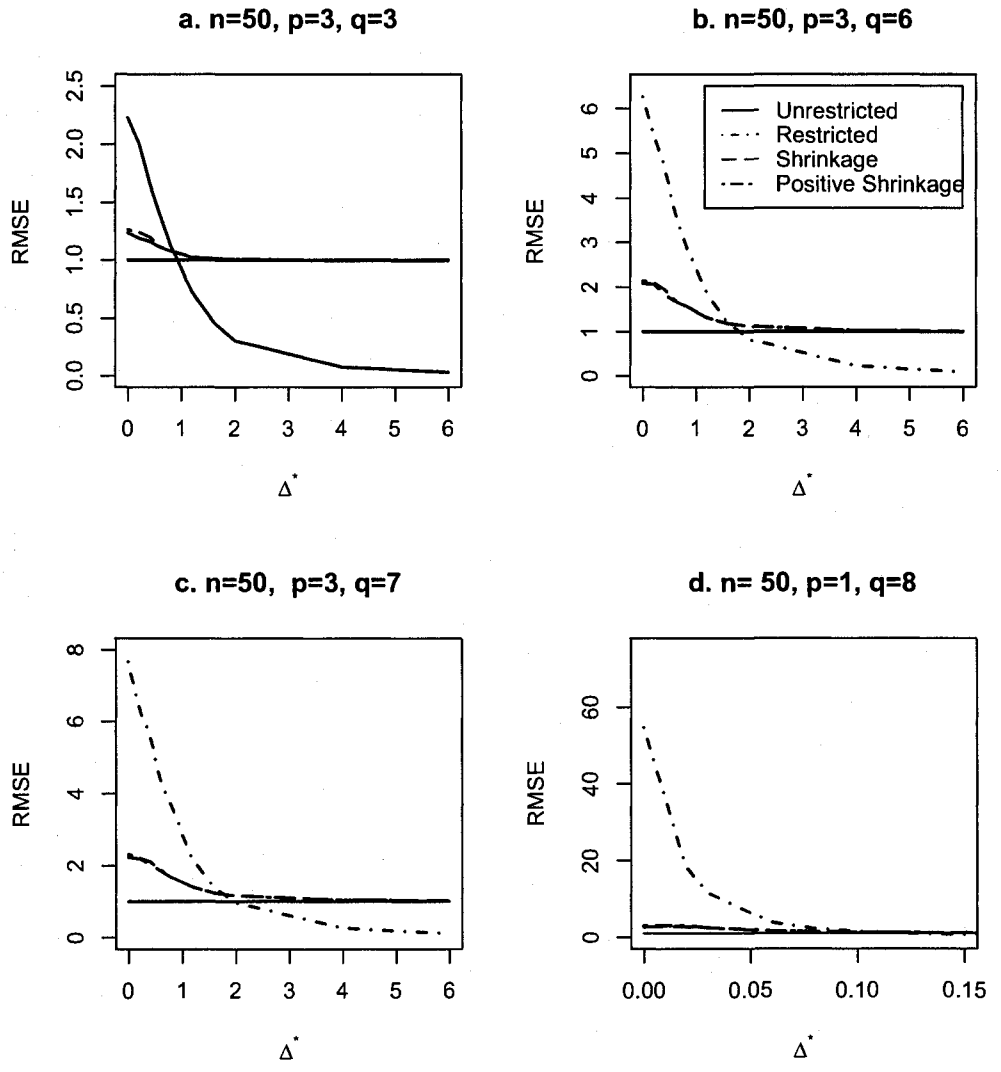

Figure 2.2: Simulated RMSE of the estimators as a function of the non-centrality parameter $\Delta^\star$ for different q and 20% censoring.

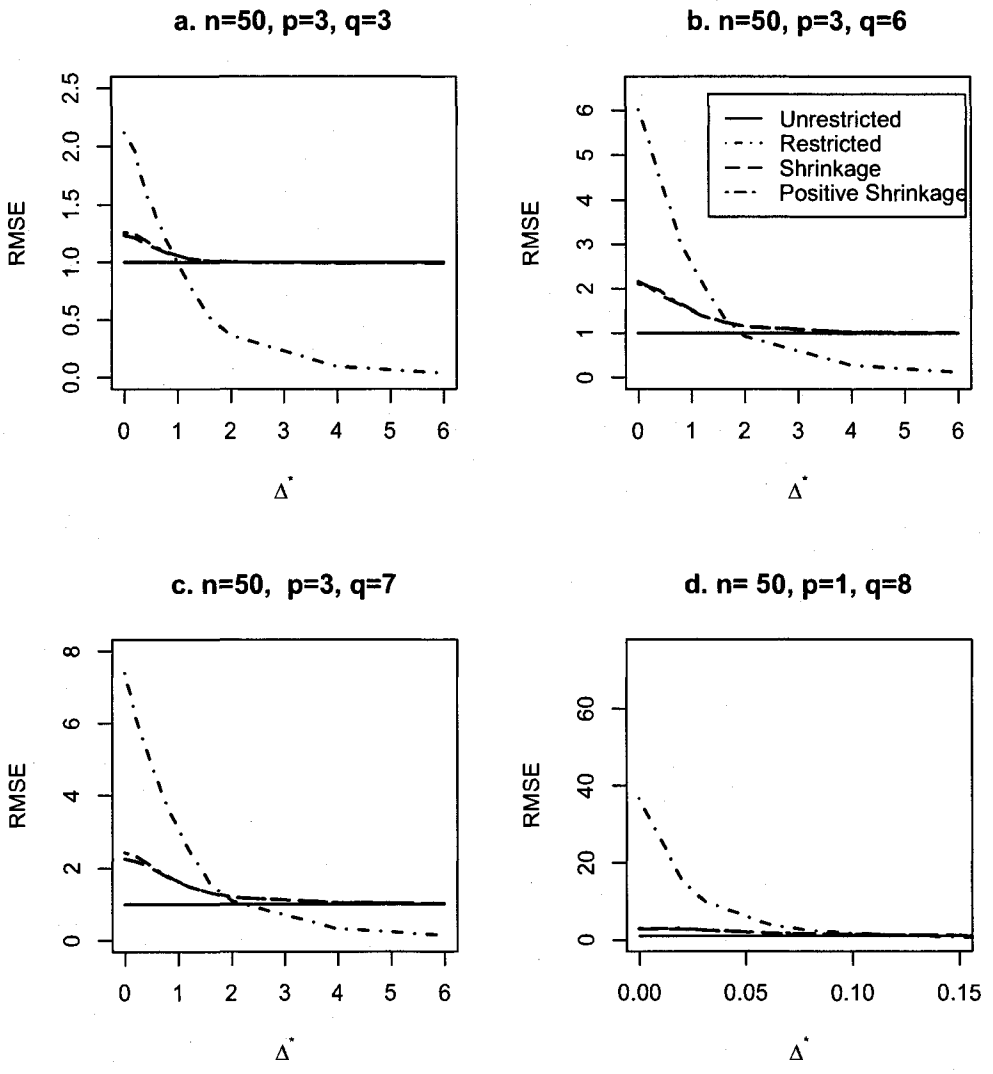

Figure 2.3: Simulated RMSE of the estimators as a function of the non-centrality parameter $\Delta^\star$ for different q and 30% censoring

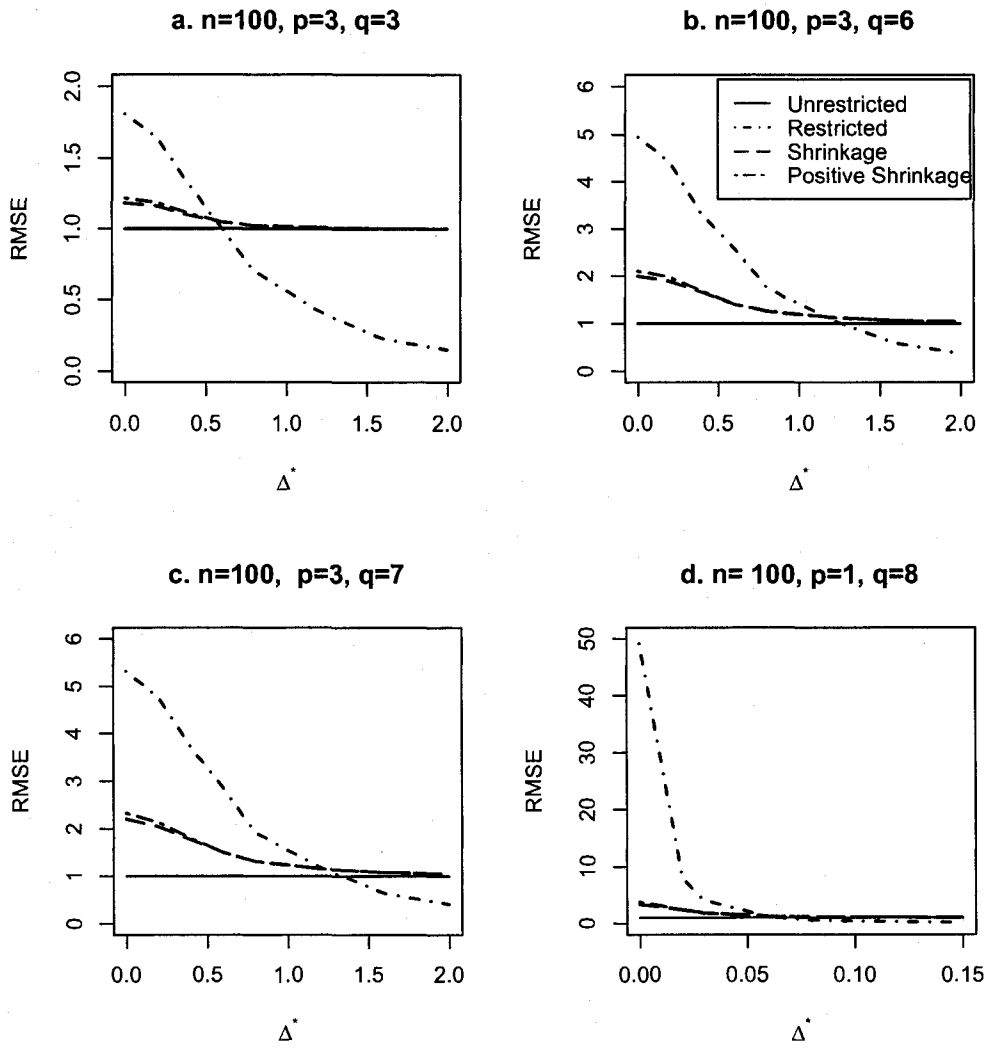

Figure 2.4: Simulated RMSE of the estimators as a function of the non-centrality parameter $\Delta^\star$ for different q and 10% censoring.

Table 2.4: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 20\%, n = 50$ and $q = 3$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	2.227	1.233	1.264
0.2	2.010	1.188	1.240
0.4	1.673	1.163	1.197
0.6	1.385	1.110	1.116
0.8	1.123	1.076	1.078
1.2	0.728	1.022	1.025
1.6	0.462	1.010	1.010
2.0	0.306	1.004	1.004
4.0	0.075	0.996	0.996

Table 2.5: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 20\%, n = 50$ and $q = 6$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	6.271	2.098	2.154
0.2	5.420	2.080	2.115
0.4	4.751	1.837	1.943
0.6	3.688	1.694	1.717
0.8	2.982	1.578	1.583
1.2	1.881	1.313	1.316
1.6	1.194	1.192	1.192
2.0	0.823	1.118	1.118
4.0	0.222	1.021	1.021

Table 2.6: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 20\%, n = 50$ and $q = 8$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.00	54.716	2.838	3.036
0.02	18.374	2.742	2.966
0.03	11.468	2.435	2.511
0.06	3.799	1.629	1.633
0.08	2.197	1.409	1.411
0.10	1.471	1.269	1.269
0.13	0.816	1.156	1.156
0.15	0.629	1.112	1.112
0.18	0.466	1.082	1.082
0.21	0.314	1.068	1.068
0.25	0.211	1.041	1.041

Table 2.7: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 30\%, n = 50$ and $q = 3$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	2.115	1.234	1.258
0.2	1.950	1.212	1.241
0.4	1.604	1.154	1.190
0.6	1.408	1.115	1.126
0.8	1.179	1.078	1.085
1.2	0.823	1.031	1.031
1.6	0.520	1.010	1.010
2.0	0.363	1.000	1.000
4.0	0.093	0.989	0.989

Table 2.8: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 30\%, n = 50$ and $q = 6$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	6.014	2.167	2.110
0.2	5.309	2.034	2.047
0.4	4.494	1.872	1.968
0.6	3.853	1.743	1.780
0.8	3.029	1.632	1.667
1.2	2.141	1.381	1.391
1.6	1.366	1.249	1.249
2.0	0.926	1.144	1.144
4.0	0.270	1.023	1.023

Table 2.9: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 30\%, n = 50$ and $q = 8$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.00	36.626	2.855	3.070
0.02	15.548	2.817	2.987
0.03	10.228	2.548	2.669
0.06	4.137	1.762	1.796
0.08	2.455	1.516	1.520
0.10	1.757	1.353	1.355
0.13	0.962	1.208	1.208
0.15	0.722	1.142	1.142
0.18	0.568	1.103	1.103
0.21	0.395	1.083	1.083
0.25	0.263	1.046	1.046

Table 2.10: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 10\%, n = 100$ and $q = 3$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	1.810	1.185	1.217
0.2	1.638	1.162	1.186
0.4	1.309	1.095	1.108
0.6	1.006	1.047	1.048
0.8	0.700	1.024	1.024
1.2	0.418	1.010	1.010
1.6	0.226	1.001	1.001
2.0	0.149	0.999	0.999

Table 2.11: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 10\%, n = 100$ and $q = 6$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	4.943	1.996	2.101
0.2	4.408	1.888	1.974
0.4	3.315	1.661	1.676
0.6	2.592	1.401	1.410
0.8	1.757	1.252	1.252
1.2	1.074	1.129	1.129
1.6	0.580	1.062	1.062
2.0	0.365	1.038	1.038

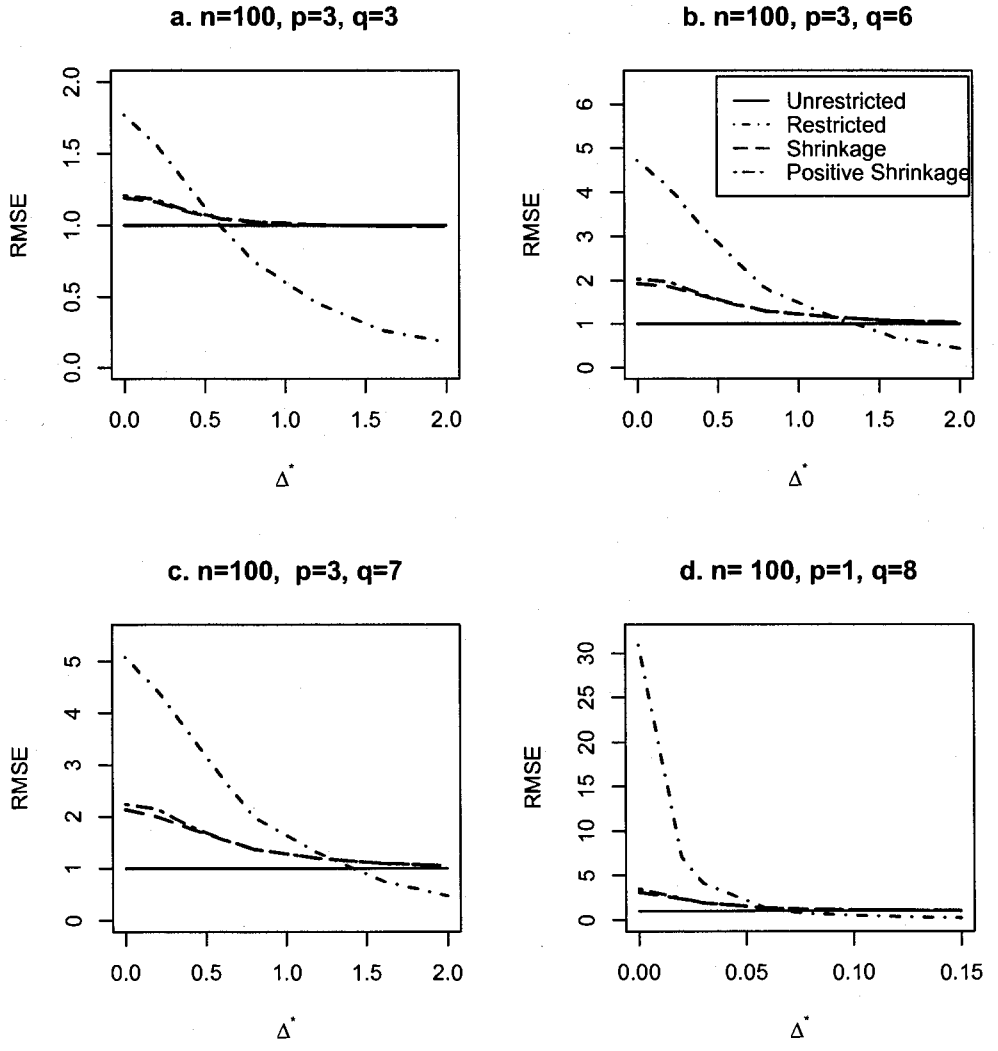

Figure 2.5: Simulated RMSE of the estimators as a function of the non-centrality parameter $\Delta^\star$ for different q and 20% censoring.

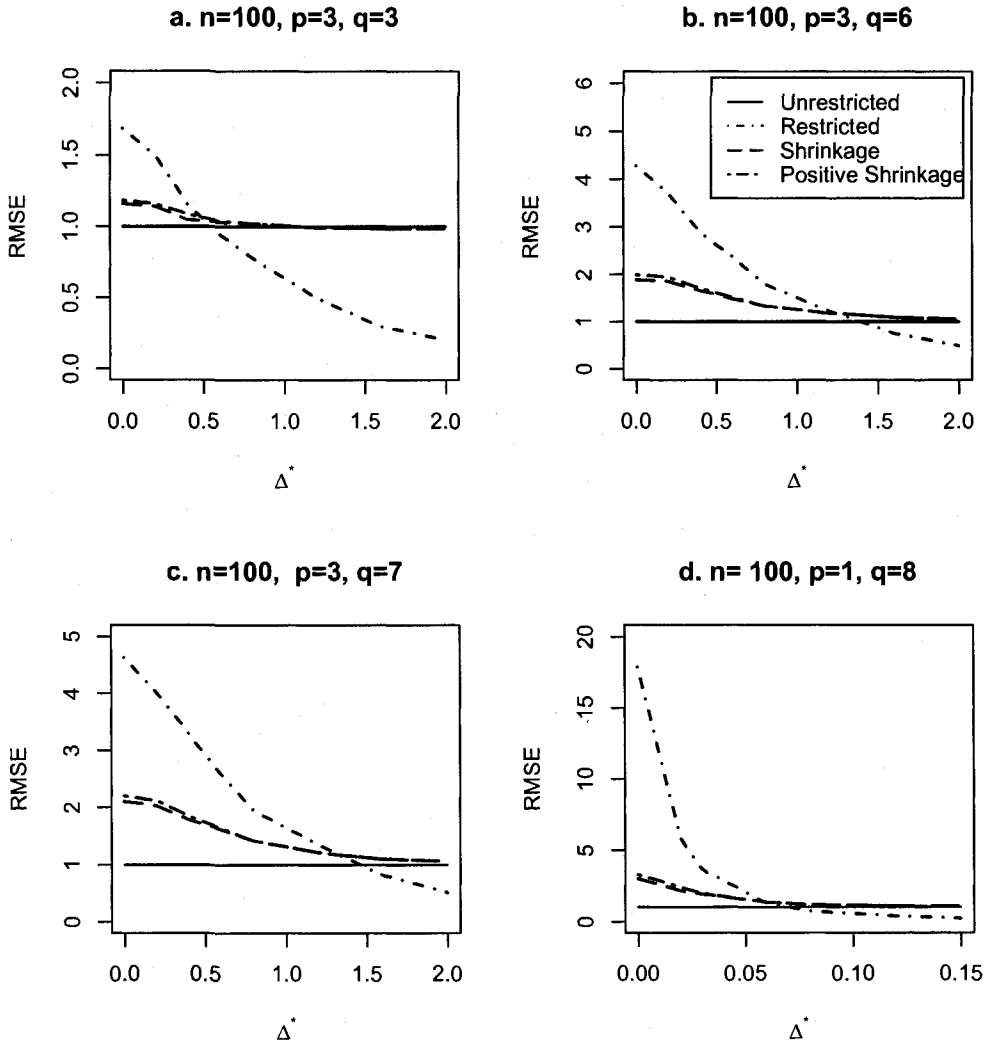

Figure 2.6: Simulated RMSE of the estimators as a function of the non-centrality parameter $\Delta^\star$ for different q and 30% censoring.

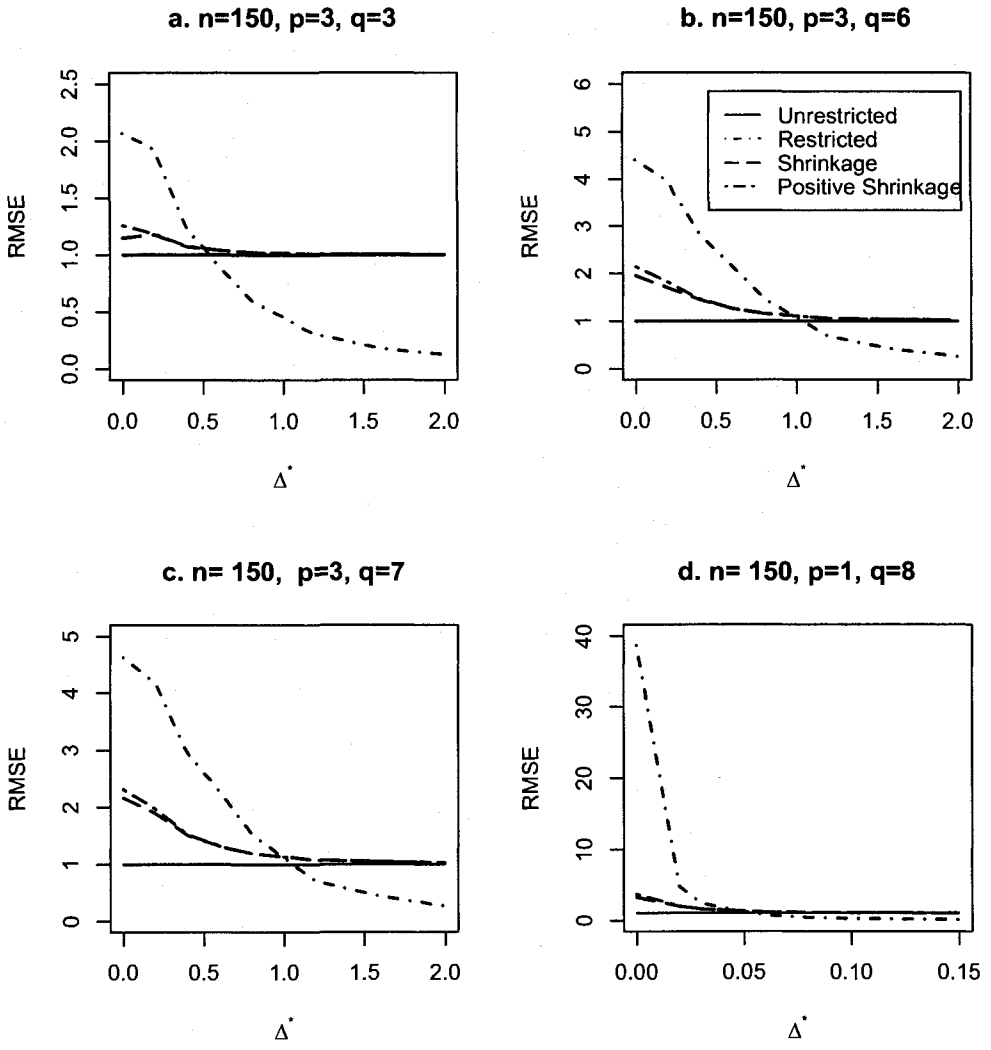

Figure 2.7: Simulated RMSE of the estimators as a function of the non-centrality parameter $\Delta^\star$ for different q and 10% censoring.

Table 2.12: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 10\%, n = 100$ and $q = 8$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.00	49.127	3.294	3.642
0.02	7.946	2.307	2.346
0.03	4.004	1.793	1.803
0.06	1.070	1.255	1.255
0.08	0.550	1.132	1.132
0.10	0.388	1.091	1.091
0.13	0.217	1.046	1.046
0.15	0.159	1.030	1.030

Table 2.13: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 20\%, n = 100$ and $q = 3$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	1.768	1.190	1.207
0.2	1.561	1.161	1.178
0.4	1.269	1.093	1.102
0.6	0.991	1.044	1.047
0.8	0.756	1.024	1.024
1.2	0.453	1.003	1.003
1.6	0.261	0.994	0.994
2.0	0.179	0.994	0.994

Table 2.14: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 20\%, n = 100$ and $q = 6$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	4.714	1.936	2.039
0.2	4.093	1.861	1.968
0.4	3.228	1.657	1.682
0.6	2.494	1.449	1.462
0.8	1.812	1.289	1.290
1.2	1.171	1.151	1.151
1.6	0.684	1.071	1.071
2.0	0.426	1.042	1.042

Table 2.15: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 20\%, n = 100$ and $q = 8$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.00	30.870	3.092	3.469
0.02	7.031	2.309	2.394
0.03	4.118	1.871	1.892
0.06	1.189	1.311	1.311
0.08	0.656	1.153	1.153
0.10	0.478	1.106	1.106
0.13	0.284	1.051	1.051
0.15	0.199	1.034	1.034

Table 2.16: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 30\%, n = 100$ and $q = 3$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	1.683	1.165	1.187
0.2	1.500	1.143	1.159
0.4	1.161	1.050	1.090
0.6	0.946	1.031	1.039
0.8	0.780	1.020	1.020
1.2	0.499	0.995	0.995
1.6	0.297	0.985	0.985
2.0	0.200	0.985	0.985

Table 2.17: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 30\%, n = 100$ and $q = 6$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	4.287	1.886	1.994
0.2	3.711	1.852	1.939
0.4	2.881	1.648	1.702
0.6	2.371	1.486	1.509
0.8	1.785	1.328	1.331
1.2	1.223	1.173	1.174
1.6	0.750	1.083	1.083
2.0	0.478	1.048	1.048

Table 2.18: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 30\%, n = 100$ and $q = 8$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.00	17.875	3.007	3.302
0.02	5.838	2.154	2.405
0.03	3.689	1.945	1.975
0.06	1.275	1.355	1.354
0.08	0.761	1.179	1.179
0.10	0.560	1.125	1.125
0.13	0.343	1.057	1.057
0.15	0.242	1.036	1.036

Table 2.19: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 10\%, n = 150$ and $q = 3$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	2.068	1.147	1.256
0.2	1.928	1.173	1.177
0.4	1.222	1.066	1.072
0.6	0.901	1.037	1.037
0.8	0.597	1.019	1.019
1.2	0.301	1.005	1.005
1.6	0.179	1.002	1.002
2.0	0.123	1.000	1.000

Table 2.20: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 10\%, n = 150$ and $q = 6$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	4.408	1.954	2.139
0.2	3.960	1.697	1.821
0.4	2.812	1.445	1.455
0.6	2.158	1.260	1.260
0.8	1.463	1.154	1.154
1.2	0.670	1.056	1.056
1.6	0.404	1.034	1.034
2.0	0.254	1.019	1.019

Table 2.21: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 10\%, n = 150$ and $q = 8$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.00	38.652	3.239	3.701
0.02	4.849	1.977	2.013
0.03	2.394	1.593	1.596
0.06	0.642	1.162	1.162
0.08	0.362	1.088	1.088
0.10	0.218	1.045	1.045
0.13	0.130	1.028	1.028
0.15	0.093	1.019	1.019

Table 2.22: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 20\%, n = 150$ and $q = 3$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	2.023	1.201	1.247
0.2	1.770	1.156	1.169
0.4	1.151	1.058	1.069
0.6	0.889	1.034	1.034
0.8	0.614	1.014	1.014
1.2	0.330	1.000	1.000
1.6	0.204	0.997	0.997
2.0	0.142	0.996	0.996

Table 2.23: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 20\%, n = 150$ and $q = 6$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	4.219	1.870	2.057
0.2	3.507	1.752	1.815
0.4	2.590	1.456	1.476
0.6	2.047	1.298	1.299
0.8	1.470	1.184	1.184
1.2	0.719	1.065	1.065
1.6	0.462	1.038	1.038
2.0	0.292	1.018	1.018

Table 2.24: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 20\%, n = 150$ and $q = 8$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.00	22.798	2.872	3.376
0.02	4.413	2.032	2.082
0.03	2.420	1.628	1.641
0.06	0.742	1.181	1.181
0.08	0.426	1.095	1.095
0.10	0.265	1.044	1.044
0.13	0.168	1.027	1.027
0.15	0.119	1.016	1.016

Table 2.25: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 30\%, n = 150$ and $q = 3$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	1.894	1.158	1.233
0.2	1.664	1.145	1.161
0.4	1.057	1.064	1.062
0.6	0.874	1.031	1.032
0.8	0.656	1.008	1.008
1.2	0.356	0.993	0.993
1.6	0.225	0.990	0.990
2.0	0.164	0.990	0.990

2.6 Bootstrap Interval Estimation

The problem of interval estimation for shrinkage estimator is frequently neglected, perhaps due to mathematical intractability of the sampling distribution of shrinkage estimators for nonnormal populations. In recent articles Ahmed *et al.* (2006*b*) and

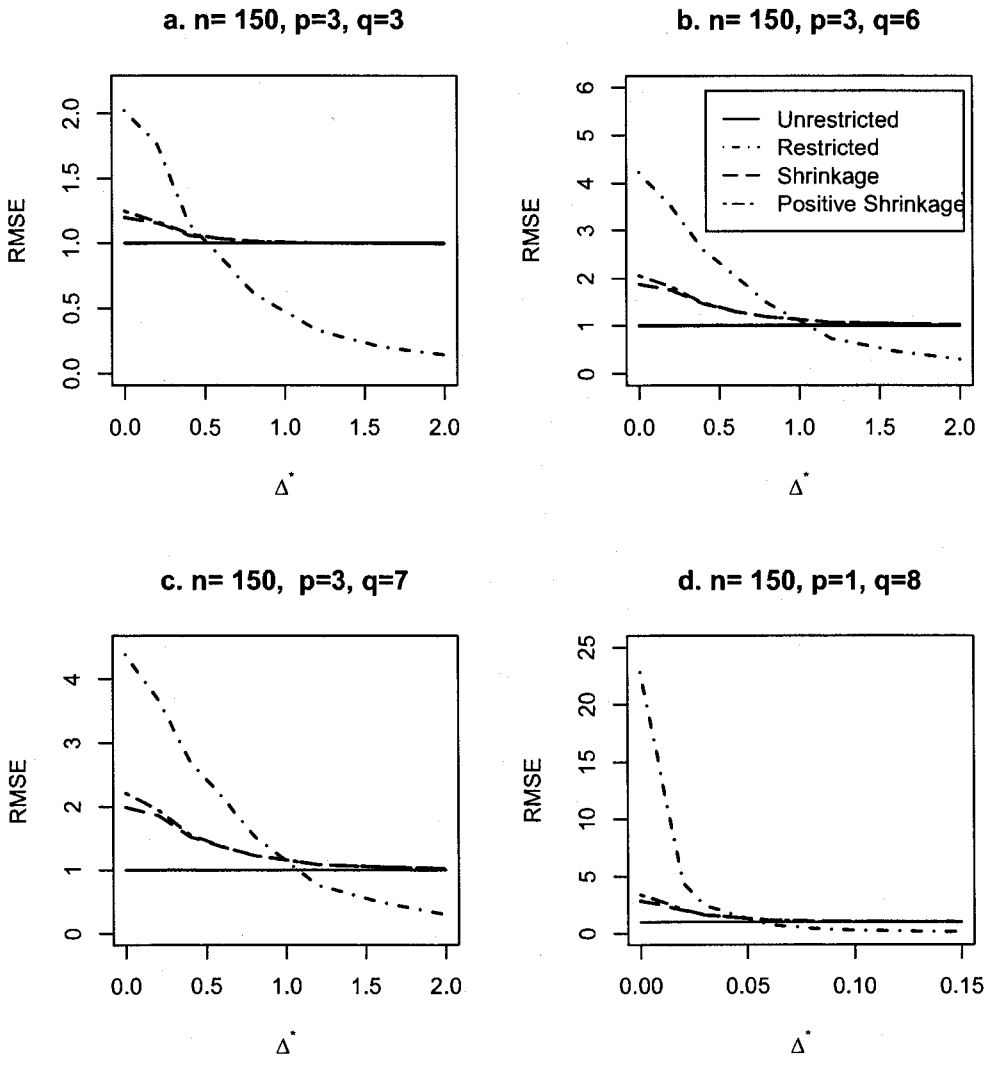

Figure 2.8: Simulated RMSE of the estimators as a function of the non-centrality parameter $\Delta^\star$ for different q and 20% censoring.

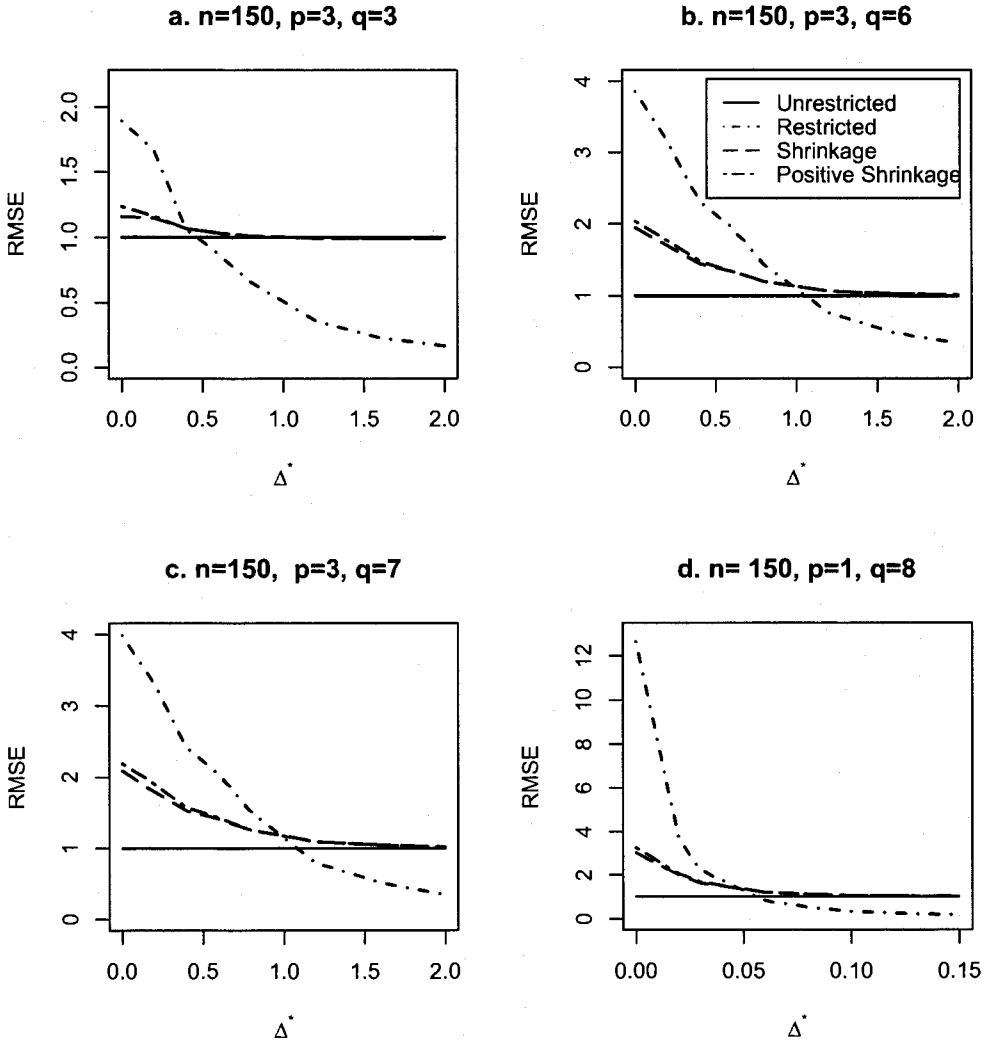

Figure 2.9: Simulated RMSE of the estimators as a function of the non-centrality parameter $\Delta^\star$ for different q and 30% censoring.

Table 2.26: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 30\%, n = 150$ and $q = 6$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.0	3.850	1.945	2.032
0.2	3.161	1.697	1.768
0.4	2.316	1.459	1.495
0.6	1.950	1.347	1.349
0.8	1.443	1.204	1.205
1.2	0.758	1.071	1.071
1.6	0.483	1.035	1.035
2.0	0.329	1.015	1.015

Efron (2006) developed a general approach to calculate the minimum volume confidence regions for the mean vector of a multivariate normal distribution. We investigate the performance of different bootstrap methods determining confidence intervals for shrinkage estimators and apply the same procedure as in Kazimi and Brownstone (1999). There are a variety of possible bootstrap sampling schemes available in the literature for survival data e.g., Davidson and Hinkley (1997). For simplicity, we consider only the case resampling bootstrap method and the remaining discussions follow. The true values of all the elements of the unknown parameter vector in the regression model are $\boldsymbol{\beta} = (2, 0.3, -2.5, 0, 0, 0)$ and the shape parameter is $\sigma = 3$. We used those values to generate survival and censoring times. We sample with replacement from the set of 100 triples $(T_k, L_k, \mathbf{X}_k)$ where $k = 1, 2, \cdots, 100$, to obtain the bootstrap data set. We then refit the Weibull regression model to these data to obtain bootstrap estimates. We conduct 1000 simulations in an iterative fashion. Within each iteration, we use 1000 bootstrap replicates to construct 95% confidence intervals for shrinkage estimators. Those intervals are based upon 95% nominal coverage ($\alpha = 0.05$).

Table 2.28 reports simulated bootstrap confidence intervals by using the asymp-

Table 2.27: Simulated RMSEs of RE, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $pc = 30\%, n = 150$ and $q = 8$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}^S$	$\hat{\boldsymbol{\beta}}^{S+}$
0.00	12.654	3.021	3.254
0.02	3.741	1.988	2.052
0.03	2.218	1.621	1.667
0.06	0.809	1.191	1.191
0.08	0.496	1.101	1.101
0.10	0.319	1.040	1.040
0.13	0.198	1.016	1.016
0.15	0.148	1.007	1.007

totic normal, percentile bootstrap and bias corrected and acceleration (BCa) method when the bootstrap samples are centered at the maximum likelihood estimator (MLE). For each method, the average upper and lower limits are reported. We also include the standard deviation of bounds over the Monte Carlo repetitions and the coverage probability computed over the repetitions. In comparing the bootstrap methods, we look for better coverage probability, lower standard error of bounds and the tightest confidence intervals.

a) It is noted that all of the intervals in Table 2.28 have lower coverage probabilities than the nominal level of 95% for lower values of $\boldsymbol{\beta}'s$. Perhaps, this is due to the systematic downward bias for the moments of the bootstrap distribution.

b) Interestingly, the asymptotic method generates the tightest confidence intervals with coverage probabilities lower than the nominal level of 95%. For example, the average intervals for β_0, β_1 and β_2 are $(-1.738,\ 6.541)$, $(0.247,\ 0.352)$ and (-3.304, -1.734) with coverage 91.4%, 88.6% and 93.6% respectively. These shorter confidence intervals were due to underestimation of parameter variability leading to a lower coverage rate for a 95% confidence interval of the MLE. This low coverage translates to an increased actual type I error over the nominal 5%

Table 2.28: 95% nominal confidence interval for the proposed estimators, with bootstrap centered at the MLE.

Method	parameter	average lower and upper bounds	standard deviation of bound	Coverage (%)
Maximum Likelihood estimator				
Asymptotic	β_0	(-1.738, 6.541)	2.541	91.4
	β_1	(0.247, 0.352)	0.033	88.6
	β_2	(-3.304, -1.734)	0.489	93.6
	β_3	(-0.960, 0.936)	0.589	90.6
	β_4	(-0.098, 0.093)	0.059	90
	β_5	(-2.887, 2.742)	1.717	91
Shrinkage estimator				
Percentile	β_0	(-2.171, 7.122)	2.403	95.4
	β_1	(0.239, 0.361)	0.033	94.6
	β_2	(-3.433, -1.590)	0.501	96.2
BCa	β_0	(-1.948, 7.487)	2.87	93.6
	β_1	(0.238, 0.360)	0.035	91.6
	β_2	(-3.441, -1.594)	0.526	94.4
Positive shrinkage estimator				
Percentile	β_0	(-2.116, 7.085)	2.423	94.6
	β_1	(0.239, 0.361)	0.033	94.4
	β_2	(3.432, -1.597)	0.502	96.2
BCa	β_0	(-1.887, 7.348)	2.428	94.4
	β_1	(0.239, 0.360)	0.035	92.2
	β_2	(-3.433, -1.593)	0.526	94.2

significant level. Hence, inference based on the MLE may not be trustworthy.

c) For percentile method, Table 2.28 reveals that the bootstrap confidence intervals perform well. This method produces a lower standard deviation of bounds. The width of the average confidence intervals for the components of SE and PSE are 9.293, 0.122, 1.853 and 9.201, 0.122, 1.835, respectively. Further, the coverage probabilities are 95.4, 94.6, 96.2 and 94.6, 94.4, 96.2, respectively. Importantly, these coverage probabilities are very close to the nominal level of 95%.

d) The BCa method not only generates wider intervals (as compare with the percentile method) but also lower coverage probabilities than the nominal level of 95%.

In summary, the percentile method performs well in this study. This method shows that the confidence intervals for the shrinkage estimators provide considerable improvement over the MLE in terms of coverage probability and produce more meaningful intervals. It is also easier to implement and its performance is better than the BCa method.

2.7 Motivating Example

We, now return to our motivating example (VA lung cancer data) and apply the proposed estimation strategies to clinical trial data. In this trial, males with advanced inoperable lung cancer were randomized to either a standard or test chemotherapy. The primary end point for therapy comparison was time to death. Only 9 of the 137 survival times were censored. As is common in such studies, there was much heterogeneity between patients in disease extent and pathology, previous treatment of the disease, demographic background, and initial health status. The response variable is the patient survival time and the covariates are the patient's performance status (PS), a measure of general fitness on a scale from 0 to 100, an indicator of histological type of the patient's tumor where large tumor cell type is the baseline. We consider squamous versus large (squamous), small versus large (small) and adeno versus large (adeno), age in years (age), prior therapy (pth), time in months from diagnosis (diag), and the treatment status (test). Including the intercept, we have nine parameters ($p = 9$). The full model is

$$\begin{aligned} Log(T_i) &= \beta_0 + \beta_1\ \mathrm{PS}_i + \beta_2\ I(\text{cell-type=squamous})_i + \beta_3\ I(\text{cell-type=small})_i \\ &+ \beta_4\ I(\text{cell-type=adeno})_i + \beta_5\ \mathrm{age}_i + \beta_6\ \mathrm{pth}_i + \beta_7\ \mathrm{diag}_i \\ &+ \beta_8\ I(\text{treatment=test})_i + \sigma\varepsilon_i. \end{aligned}$$

According to the asymptotic likelihood inference of Kalbfleisch and Prentice (2002, p.72), patient survival time does not differ significantly among treatment groups, ages, prior therapy and the time in months from diagnosis. Here we can regard those variables as NSI and use the shrinkage estimators of this chapter to evaluate the effect of performance status and tumor cell types on survival time. More formally, $H_0 : (\beta_5, \beta_6, \beta_7, \beta_8) = (0, 0, 0, 0)$ as our pivot.

Hence, the reduced model is

$$\begin{aligned} Log(T_i) &= \beta_0 + \beta_1 \text{ PS}_i + \beta_2 \, I(\text{cell-type=squamous})_i + \beta_3 \, I(\text{cell-type=small})_i \\ &+ \beta_4 \, I(\text{cell-type=adeno})_i + \sigma\varepsilon_i, \quad \text{where} \quad i = 1, 2, \cdots, 137. \end{aligned}$$

Table 2.29: Estimate (first row), standard error (second row) and bias (third row) of intercept (β_0), performance status (β_1), cell type squamous vs. large (β_2), cell type small vs. large (β_3) and Adeno vs. large (β_4) on survival time.

Estimators	β_0	β_1	β_2	β_3	β_4	RMSE
UE	2.8421	0.0308	0.3861	-0.4423	-0.7304	1.0000
	0.7422	0.0051	0.2516	0.2532	0.2073	
	-0.0222	0.0007	-0.0116	-0.0138	0.0046	
RE	3.1010	0.0299	0.3258	-0.3750	-0.7790	2.2085
	0.3642	0.0050	0.2389	0.2547	0.1979	
	0.2367	-0.0001	-0.0719	0.0535	-0.0440	
SE	2.8966	0.0305	0.3641	-0.4224	-0.7519	1.3168
	0.6167	0.0049	0.2468	0.2560	0.2000	
	0.0323	0.0005	-0.0335	0.0061	-0.0168	
PSE	2.8951	0.0306	0.3659	-0.4237	-0.7490	1.3764
	0.5991	0.0049	0.2455	0.2532	0.1997	
	0.0308	0.0005	-0.0318	0.0048	-0.0140	

The point estimates, the standard errors and relative efficiency based on case-resampling bootstrap of size B=1000 are reported in Table 2.29. The results from the example reveal shrinkage estimators are superior to the classical estimator, which is strongly in agreement with our analytical as well as simulation results. Under the null

hypothesis, the efficiency of $\tilde{\boldsymbol{\beta}}$ is higher than all other estimators but this efficiency becomes lower and lower as the hypothesis error grows.

Table 2.30: 95% bootstrap confidence interval for MLE, shrinkage and positive shrinkage estimator.

Shrinkage estimator			
Estimator	Asymptotic	Percentile	BCa
β_0	(1.549, 4.179)	(1.672, 4.096)	(2.266, 4.769)
β_1	(0.021, 0.039)	(0.021, 0.039)	(0.019, 0.037)
β_2	(-0.102, 0.880)	(-0.124, 0.804)	(-0.171, 0.785)
β_3	(-0.905, 0.048)	(-0.753, 0.081)	(-0.844, 0.232)
β_4	(-1.272,-0.198)	(-1.149, -0.357)	(-1.185, -0.409)
Positive-part Shrinkage estimator			
β_0		(1.716, 4.048)	(2.318, 4.490)
β_1		(0.021, 0.04)	(0.018, 0.037)
β_2		(-0.164, 0.797)	(-0.272, 0.723)
β_3		(-0.763, 0.064)	(-0.797, 0.254)
β_4		(-1.155, -0.373)	(-1.161, -0.396)

Finally, we calculate bootstrap confidence intervals for the regression parameter based on shrinkage estimators. Recall that shrinkage and positive shrinkage estimates are (3.162, 0.0291, 0.321, -0.386, -0.787) and (3.159, 0.029, 0.322, -0.386, -0.786). respectively. Table 2.30 summarizes 95% bootstrap confidence intervals for the shrinkage estimators using different bootstrap methods. The percentile method generates the tightest confidence interval for shrinkage estimators.

Tibshirani (1997) used the LASSO technique to choose significant variables in this data set. It was found that performance status is the dominant effect with treatment and cell type also showing the moderate effect. Our analysis is based on NSI, but not considering on tuning parameter. In chapter 3, we demonstrate that shrinkage estimators are relatively more efficient than estimates based on LASSO when q is large which is generally true and is in agreement with Tibshirani (1997). We strongly recommend the use of the suggested estimation strategy when q is large enough.

2.8 Conclusion

The objective of this study is to compare the performance of shrinkage estimators to the maximum likelihood estimator in the context of the Weibull regression model for censored data. We explored the risk properties of the estimators via asymptotic distributional risk and Monte Carlo experiments. We also conducted different bootstrap methods to generate confidence intervals for the proposed estimators. Finally, we applied shrinkage estimation to a real data set to evaluate the relative performance of the estimators at hand. It is concluded both analytically and computationally that the PSE dominates the usual shrinkage estimator. Further, both shrinkage estimators outperform the classical estimator of the regression parameter vector in the entire parameter space. In contrast, the performance of the constrained estimation heavily depends on the quality of the NSI. Not only that, the risk of the restricted estimator may become unbounded when the restriction does not hold.

Interestingly, the percentile bootstrap method yields adequate confidence intervals for shrinkage and positive shrinkage estimators. These confidence intervals permit the application of shrinkage estimators to the human disease problem like, cancer, mortality rate for aged people and the lifetime analysis of carcinogenesis where the sample size is large enough. Our simulation experiments and numerical example have provided strong evidence that corroborates with the usual asymptotic theory related to proposed estimation strategies. Importantly, we have combined the two most celebrated methods (shrinkage and Bootstrap estimation) to develop the point and interval estimation for a Weibull regression censored model. It is noted that the application of shrinkage estimators are subject to condition that $q \geq 3$. For $q = 1, 2$ one can employ the pretest approach.

Perhaps, the most important message in this chapter is that very large gains in precision may be achieved by judiciously exploiting the restriction in the parameter space which in practice will be available in any realistic problem. Our numerical findings indicate that for up to a reduction of 50% the risk seem quite realistic in

some situations. Thus, it seems conceivable to pay attention to these situations in the development of statistical inference theory. Like the statistical models underlying the statistical inferences to be made, the restriction in the parameter space will be susceptible to uncertainty and the practitioners may be reluctant to impose the restriction regarding parameters in the estimation process.

One can extend these methodologies for other accelerated failure time models such as log-normal, log-logistic etc.

Chapter 3

Shrinkage, Pretest and PH type estimators for Generalized Linear Models

3.1 Introduction

The term "generalized linear model" was first introduced in a landmark paper by Nelder and Wedderburn (1972). An important statistical development of the last thirty five years has been the advance in regression analysis provided by generalized linear models (GLMs). Much used in applications to the social sciences, biology and medicine, these models also play an important role in the area of survival analysis. These models are mathematical extensions of linear models that do not force data into unnatural scales, and thereby allow for non-linearity and non-constant variance structures in the data (Hastie and Tibshirani (1990)). They are based on an assumed relationship (called a link function; see next section) between the mean of the response variable and the linear combination of the explanatory variables. Data may be assumed to be from several families of probability distributions, including normal, binomial, Poisson, negative binomial, or gamma distributions, many of which better

fit non-normal error structures of most human ecology data.

GLM models a random variable Y that follows a distribution in the exponential family using a linear combination of the predictors, $\mathbf{x}'\boldsymbol{\beta}$ where $\mathbf{x}$ and $\boldsymbol{\beta}$ denote vectors of the predictors and the coefficients, respectively. In many cases, the parameter vector $\boldsymbol{\beta}$ is unknown and we wish to estimate it or to test hypotheses about it. These are usually done by applying the maximum likelihood method and the likelihood ratio test.

In this chapter we consider the estimation problem for the GLMs which may have a large collection of potential predictor variables and some of them may not have influence on the response of interest. In this situation, selecting the statistical model is always a vital component in estimation. One consequence of this problem is model mis-specification. The mis-specification of covariates in GLMs is a common situation. Extraneous covariates may be included in the model, but it is more likely that relevant covariates will be omitted. The latter situation may arise either because of the researchers' lack of understanding of the underlying theory, or because certain data are unavailable. With this in mind, several authors (Ahmed *et al.* (2007), Ahmed *et al.* (2006*a*), Judge and Mittelhammaer (2004) and Ahmed (1997)) have reappraised some of the standard pretest and shrinkage estimation strategies for parametric, semi-parametric and nonparametric linear models. The goal of this chapter is to analyze some of the issues involved in the estimation of generalized linear models that may be over-parameterized. For example, in the data analyzed by Park and Hastie (2007) (this data set was originally collected by Rossouw *et al.* (1983)) coronary heart disease may be related to the variables: systolic blood pressure, cumulative tobacco, low density lipoprotein cholesterol, adiposity, family history of heart disease, type-A behavior, obesity, alcohol, and age. The analysis shows that cumulative tobacco, low density lipoprotein cholesterol, family history of heart disease, type-A behavior and age are the most important factors and the effect of the other variables may be ignored. We may use those insignificant variables as non-sample information in the

shrinkage and pretest estimation procedure. The main objective is to estimate the values of unknown parameter vector $\boldsymbol{\beta}$ under a set of linear restrictions

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{h}, \tag{3.1}$$

where $\mathbf{H}$ is $q \times k$ matrix of rank $q \leq k$ and $\mathbf{h}$ is a given $q \times 1$ vector of constants. Restrictions of this kind may be regarded as *NSI*. It is assumed that $\mathbf{H}$ has rank q, which implies that the q equations do not contain any redundant information about $\boldsymbol{\beta}$.

The LASSO originally proposed by (Tibshirani (1996)) is arguably one of the most important contributions for the problem of variable selection in the past decade, and has been extensively studied in the literature. See, for example, Knight and Fu (2001), Fan and Li (2001), Leng *et al.* (2006), Yuan and Lin (2007) and Zou (2006). Efron *et al.* (2004) introduced the Least Angle regression algorithm which suggested a very fast way to draw the entire regularization path for a LASSO estimate of $\boldsymbol{\beta}$. Park and Hastie (2007) proposed an algorithm (called glmpath) that generates the coefficient paths for the L_1 regularization problems as in LASSO problems, but in which the loss function is replaced by the negetive log-likelihood of any distribution in the exponential family.

The plan of this chapter is as follows. In Section 3.2, we present the details of generalized linear models with all the relevant notations. We illustrate the properties of the maximum likelihood estimation procedure and computational details for estimating the parameters and inferences in Sections 3.3-3.4. The proposed pretest estimator, shrinkage and positive shrinkage estimators and Park and Hastie (PH) estimators are presented in Section 3.5. Asymptotic properties of the proposed estimators, bias and risk expressions and the weighted risk analysis of the estimators are contained in Section 3.6. The results of a simulation study that includes comparison with the PH estimator are reported in Section 3.7. A numerical example with nine regressor variables is presented in Section 3.8 to illustrate the methods. This chapter

concludes with some discussion in Section 3.9. Throughout this chapter, the boldface symbols represent vectors/matrices.

3.2 Description of the Generalized Linear Model

The observations belonging to a statistical model can be summarized in terms of a systematic component and a random component. In the GLM discussed by McCullagh and Nelder (1989), the random component is inherent in the exponential family distribution of the observation, while the systematic component assumes a linear structure in the predictor variables for a function of the mean. This function is known as the *link function.* When the parameter θ_i is modelled as a linear function of the predictors, the link function is known as a *canonical link.* Therefore for a given set of observations $\mathbf{Y} = (y_1, y_2, \cdots, y_n)'$, where y_i is assumed to have a distribution in the exponential family of distributions with predictor values $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{in})'$, then a probability density/mass function has the form

$$f_Y(y_i; \theta_i, \phi) = \exp\{(y_i\theta_i - b(\theta_i))/a_i(\phi) + c(y_i, \phi)\},$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions and ϕ is the *dispersion parameter* that is treated as a nuisance parameter if it is unknown. If ϕ is known, this is an exponential-family model with canonical parameter θ_i. In this chapter, we are only interested in applying our proposed estimation procedure in GLMs where the dispersion parameter ϕ is known i.e., when the responses are binary and count data. In this case, the above density function can be written as

$$f_Y(y_i; \theta_i) = c(y_i)exp\{y_i\theta_i - b(\theta_i)\}. \tag{3.2}$$

GLMs have the following key features (McCullagh and Nelder (1989)).

(1) The random component of a GLM specifies the distribution of the response

variable Y_i. The distribution has the form (3.2) and for any distribution of this form, the mean and variance of Y_i are given by

$$E[Y_i] = \mu_i = \frac{db(\theta_i)}{d\theta_i} \quad \text{and} \quad Var(Y_i) = V(\mu_i) = \frac{d^2b(\theta_i)}{d\theta_i^2}.$$

(2) The systematic component of a GLM is a linear combination of regressor variables, termed the linear predictor η,

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta},$$

where $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{in})$ is the regressor vector and $\boldsymbol{\beta}$ is the vector of model parameters. The linear form of the systematic component places the regressors on an additive scale which makes the interpretation of their effects simple. Also, the significance of each regressor can be tested with linear restrictions $H_0 : \mathbf{H}\boldsymbol{\beta} = \boldsymbol{h}$ versus $H_a : \mathbf{H}\boldsymbol{\beta} \neq \boldsymbol{h}$.

(3) The *link* function of a GLM specifies a monotonic differentiable function. This function connects the random and systematic components. This connection has been done by equating the mean response μ_i to the linear predictor η_i by $\eta_i = g(\mu_i)$, that is

$$g(\mu_i) \overset{link}{=} \eta_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

The link function $g(\mu_i) = \mu_i$ is the identity link function which equates the mean response to the linear predictor. Thus, the link function for the regression model with normally distributed response variable Y_i is the identity link. The link function which equates the linear predictor to the canonical parameter is the canonical link. That is, $\eta_i = \mathbf{x}_i'\boldsymbol{\beta} = g(\mu_i) = \theta_i$.

In practice, a given data set may be distributed according to some unknown member of the exponential family and therefore, different link functions have to be evaluated. The link is a linearizing transformation of the mean—a function

that maps the mean onto a scale on which linearity is assumed. One purpose of the link is to allow η_i to range freely while restricting the range of μ_i. For example, the inverse logit link $\mu_i = 1/1+e^{-\eta_i}$ maps $(-\infty, \infty)$ onto $(0, 1)$, which is an appropriate range if μ_i is a probability. The monotonicity of the link function guarantees that this mapping is one-to-one. Thus we can express the GLM in terms of the inverse link function,

$$E[Y_i] = \mu_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}).$$

The canonical link is in many cases a useful link function, and is a reasonable function to try, unless the subject matter suggests otherwise. The canonical link does simplify the estimation method slightly, but there is no need to restrict generalized linear modelling to canonical link functions.

In summary, generalized linear models make up a general class of probabilistic regression models with the assumptions that:

(1) the response probability distribution is a member of the exponential family of distributions;

(2) the responses $Y_i \quad i = 1, 2, \cdots, n$ form a set of independent random variables;

(3) the explanatory variables are linearly combined to explain systematic variation in a function of the mean.

In a practical data situation, GLM fitting involves the following:

- choosing an error distribution that is relevant;
- identifying the independent variables to be included in the systematic components; and
- specifying the link function

The next section presents the unrestricted maximum likelihood method for estimating the regression parameters assuming that the previous assumptions have been specified.

3.3 Unrestricted Maximum Likelihood Estimation

If the probability specifications of an exponential family model are given by $f(y_i;\theta_i)$, then the best way to fit a GLM is by maximum likelihood estimation of the parameters $\boldsymbol{\beta}$ for the observed data (Green and Silverman (1994)). With many desirable properties of maximum likelihood estimators such as consistency, efficiency, sufficiency and asymptotic normality, it is natural to consider such a method for GLMs. In general, the maximum likelihood equations which result from GLMs cannot be solved explicitly and hence recourse must be made to numerical methods. There are three methods described in this section: The Newton-Raphson method, the Fisher Scoring method, and the iteratively re-weighted least squares method. To derive likelihood equations, let the responses $y_1, y_2, \cdots, y_n$ be generated from a member of exponential family (3.2). The likelihood function is written as

$$\prod_{i=1}^{n} f(y_i;\theta_i) = \prod_{i=1}^{n} c(y_i)\exp(y_i\theta_i - b(\theta_i)). \tag{3.3}$$

Then the log-likelihood is given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [(y_i\theta_i - b(\theta_i)) + \ln c(y_i)] = \sum_{i=1}^{n} \ell_i, \tag{3.4}$$

where ℓ_i is the ith component of the log-likelihood and is therefore given by

$$\ell_i = (y_i\theta_i - b(\theta_i)) + \ln c(y_i). \tag{3.5}$$

The likelihood implicitly depends on the parameters $\beta_j, \quad j = 1, 2, \cdots, k$, firstly through the link function $g(\mu_i)$ and secondly through the linearity that it encompasses with respect to β_j values. The derivatives of the log-likelihood with respect to β_j are evaluated by the chain rule:

$$U_j(\boldsymbol{\beta}) = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = 0; \quad j = 1, 2, \cdots, k. \tag{3.6}$$

It can be seen that the score functions reduce to

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{d\mu_i}{d\eta_i} x_{ij}; \quad j = 1, 2, \cdots, k. \tag{3.7}$$

In a vector form, the score equations are given by

$$(\mathbf{Y} - \boldsymbol{\mu})' \mathbf{D}(\boldsymbol{\mu}) \mathbf{X} = \mathbf{0}, \tag{3.8}$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)'$, $\mathbf{D}(\boldsymbol{\mu}) = \text{diag}(d_{ii})$ and $d_{ii} = 1/V(\mu_i) g'(\mu_i)$.

The unrestricted maximum likelihood estimator (UE) of $\boldsymbol{\beta}$ is found by solving the score equations (3.8), for $\hat{\boldsymbol{\beta}}$. The numerical methods to solve (3.8) are essentially iterative. We need a common starting value of the estimate for all the methods. With the ultimate aim of obtaining a good starting value of the estimate, the following technique is employed using the approximate linearized form of $g(y_i)$, where

$$\begin{aligned} g(y_i) &\approx g(\mu_i) + (y_i - \mu_i) g'(\mu_i) \\ &= \eta_i + (y_i - \mu_i) \frac{d\eta_i}{d\mu_i} \\ &= z_i, \end{aligned}$$

where z_i is the adjusted dependent variable which depends on both y_i and μ_i. Given that the variance of z_i is $[g'(\mu_i)]^2 V(\mu_i)$, an initial estimate of $\boldsymbol{\beta}$ may be obtained by weighted least squares of $\boldsymbol{z}$ on $\mathbf{X}$, with variance-covariance matrix given by a diagonal

matrix $\boldsymbol{W}$ whose components are specified by

$$w_{ii} = \frac{1}{V(\mu_i)[g'(\mu_i)]^2} = \frac{1}{Var(z_i)}.$$

Clearly the score equations (3.8) can be written as

$$\sum_{i=1}^{n}(y_i - \mu_i)g'(\mu_i)w_{ii}x_{ij} = 0,$$

which transformed to the adjusted dependent variables yield the following

$$\sum_{i=1}^{n}(z_i - g(\mu_i))w_{ii}x_{ij} = 0. \tag{3.9}$$

Both $\boldsymbol{z}$ and $\boldsymbol{W}$ are used for maximum likelihood estimation through a weighted least squares regression. This process is iterative, since both $\boldsymbol{z}$ and $\boldsymbol{W}$ depend on the fitted values of current estimates available. Some scoring methods are needed to measure the iteration variations for a weighted least squares regression of a GLM, until convergence is reached.

3.3.1 The Newton-Raphson Method

The Newton-Raphson Method is a general purpose numerical method for finding the roots of an equation $\boldsymbol{U(\theta)} = \mathbf{0}$. It is derived from a first order Taylor series expansion of $\boldsymbol{U(\theta)}$ or a second order Taylor series expansion of an objection function, $l(\boldsymbol{\theta})$, about a current estimate. If $\boldsymbol{U(\theta)}$ is nonlinear in $\boldsymbol{\theta}$ then Newton-Raphson is an iterative technique. In the maximum likelihood problem, the function $\boldsymbol{U}$ is a score function. Consider a Taylor series expansion of $\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\beta}}|_{\hat{\boldsymbol{\beta}}}$, centered at $\hat{\boldsymbol{\beta}}^{(r)}$.

$$\mathbf{0} = \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\beta}}|_{\hat{\boldsymbol{\beta}}} \approx \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\beta}}|_{\hat{\boldsymbol{\beta}}^{(r)}} + \frac{\partial^2 \boldsymbol{\ell}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}|_{\hat{\boldsymbol{\beta}}^{(r)}}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(r)}),$$

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(r)} \approx \left[\left(-\frac{\partial^2 \boldsymbol{\ell}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right)^{-1} \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\beta}}\right]_{\hat{\boldsymbol{\beta}}^{(r)}}.$$

An updated estimate of $\hat{\boldsymbol{\beta}}$ is then obtained

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} + \left[\left(-\frac{\partial^2 \boldsymbol{\ell}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right)^{-1} \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\beta}}\right]_{\hat{\boldsymbol{\beta}}^{(r)}}.$$

This is iteratively repeated until convergence is met.

3.3.2 Fisher's Scoring Method

If the negative second-derivative matrix or the Hessian matrix is not positive definite at every iteration then the Newton-Raphson algorithm is no longer valid. In this case, the Hessian matrix is replaced by its expectation, giving Fisher's scoring algorithm. Thus the iterative process for Fisher's scoring algorithm is given by

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} - \left(\left[E\left(\frac{\partial^2 \boldsymbol{\ell}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right)\right]^{-1} \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\beta}}\right)_{\hat{\boldsymbol{\beta}}^{(r)}}. \tag{3.10}$$

For evaluating the derivatives in (3.10), the linear predictor η_i is used where $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$:

$$\begin{aligned} \frac{\partial \boldsymbol{\ell}}{\partial \eta_i} &= \frac{\partial \boldsymbol{\ell}}{\partial \theta_i}\frac{d\theta_i}{d\eta_i} = \left(\frac{\partial \boldsymbol{\ell}}{d\theta_i}\right)\left(\frac{d\eta_i}{d\mu_i}\frac{d\mu_i}{d\theta_i}\right)^{-1} \\ &= (y_i - \mu_i)\left(g'(\mu_i)b''(\theta_i)\right)^{-1}, \qquad (3.11) \\ \text{and } E\left(-\frac{\partial^2 \boldsymbol{\ell}}{\partial \eta_i^2}\right) &= \frac{d\mu_i}{d\eta_i} \times \left(g'(\mu_i)b''(\theta_i)\right)^{-1} \\ &= \left(g'(\mu_i)^2 b''(\theta_i)\right)^{-1}. \qquad (3.12) \end{aligned}$$

Note that $\frac{\partial^2 \boldsymbol{\ell}}{\partial \eta_i \partial \eta_j} = w_{ij} = \left(g'(\mu_i)^2 b''(\theta_i)\right)^{-1}$ if $i = j$, and it is $= 0$ if $i \neq j$.
Let $\boldsymbol{z}^*$ be the n-vector with $z_i^* = (y_i - \mu_i)g'(\mu_i)$, then we have from (3.11)

$$\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\eta}} = \mathbf{W}\boldsymbol{z}^*, \tag{3.13}$$

and from (3.12)

$$E\left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'}\right) = \mathbf{W}. \tag{3.14}$$

Since $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, then by chain rule we have

$$\begin{aligned} \frac{\partial \ell}{\partial \boldsymbol{\beta}} &= \frac{\partial \ell}{\partial \boldsymbol{\eta}} \cdot \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} = \mathbf{X}' \frac{\partial \ell}{\partial \boldsymbol{\eta}} \\ &= \mathbf{X}'\mathbf{W}\boldsymbol{z}^*, \end{aligned} \tag{3.15}$$

and

$$E\left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right) = \mathbf{X}' E\left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'}\right) \mathbf{X}. \tag{3.16}$$

Thus Fisher's scoring algorithm (3.10) yields the following sequence of updated estimates

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{z}^*. \tag{3.17}$$

3.3.3 Iteratively Reweighted Least Squares (IRLS)

Equation (3.9) can be written as

$$(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}\mathbf{X} = \mathbf{0} \quad \Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}.$$

However, the $\mathbf{z}$ and $\mathbf{W}$ depend on the unknown $\hat{\boldsymbol{\mu}}$, hence this equation gives rise to the iterative process

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)}.$$

This is known as the method of Iteratively Reweighted Least Squares. The starting value of the iteration is obtained by substituting $\hat{\boldsymbol{\mu}}^{\mathbf{0}} = \boldsymbol{y}$. At each iteration i, a weighted least squares regression of the adjusted response variable $\mathbf{z}^{(i)}$ on the design matrix $\mathbf{X}$ is obtained with the weighting matrix $\mathbf{W}^{(i)}$, where $\mathbf{z}^{(i)}$ and $\mathbf{W}^{(i)}$ are obtained by replacing $\boldsymbol{\mu}$ with $\hat{\boldsymbol{\mu}}^{(i)} = g^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}}^{(i)})$. This algorithm can be summarized as follows:

- Start with a sufficient statistic from the data to get an initial fitted vector $\hat{\boldsymbol{\mu}}^{(\mathbf{0})}$.

- From this statistic, the link function g is used to derive initial linear predictor $\hat{\boldsymbol{\eta}}^{(0)}$.
- Calculate $\left(\frac{d\eta}{d\mu}\right)_0$ and $V(\hat{\boldsymbol{\mu}}^{(0)}) = \left(\frac{d\mu}{d\theta}\right)_0$.

These statistics are used in creating the starting adjusted dependent variable and the updated weighting matrix as follows:

$$\begin{aligned} \mathbf{z}^{(0)} &= \hat{\boldsymbol{\eta}}^{(0)} + (\boldsymbol{y} - \hat{\boldsymbol{\mu}}^{(0)}) \left(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}}\right)_0, \text{ and} \\ \left(\mathbf{W}^{(0)}\right)^{-1} &= \left(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}}\right)_0^2 V^{(0)}. \end{aligned}$$

A weighted least squares regression of $\mathbf{z}^{(0)}$ on $\mathbf{X}$ is carried out for the model $E[\mathbf{z}] = \mathbf{X}\boldsymbol{\beta}$ with the adjusted weighting matrix $\mathbf{W}^{(0)}$ to obtain a first maximum likelihood estimate:

$$\hat{\boldsymbol{\beta}}^{(1)} = (\mathbf{X}'\mathbf{W}^{(0)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(0)}\mathbf{z}^{(0)},$$

which is then used to obtain updated values of $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\mu}}$:

$$\hat{\boldsymbol{\eta}}^{(1)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(1)}, \quad \hat{\mu}^{(1)} = g^{-1}(\hat{\boldsymbol{\eta}}).$$

This process is repeated to update the regression estimates at each iteration via a scoring algorithm, until the variation from one iteration to the next is sufficiently small.

An important point to note is that the weighting matrix used in IRLS, $\mathbf{W}$, is updated at each iterative step of IRLS so that each element of $\mathbf{W}$ is updated too for each observation i. Hence, $\mathbf{W}$ depends entirely on the fit of the model, and not at all on the likelihood equation $\mathbf{X}'(\boldsymbol{y} - \boldsymbol{\mu}) = \mathbf{0}$, used to determine $\hat{\boldsymbol{\beta}}$.

Under some regularity conditions [see Fahrmeir and Kaufmann (1985)], $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal with variance-covariance matrix $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$.

3.4 Restricted Estimation

In this section we consider the problem of estimating the regression parameters $\boldsymbol{\beta}$ under q linearly independent restrictions $\mathbf{H}_j'\boldsymbol{\beta} = h_j, \quad j = 1, 2, \cdots, q$, where $\mathbf{H}_j$, $j = 1, 2, \cdots, q$, are $k \times 1$ vectors and h_j, $j = 1, 2, \cdots, q$, are scalars, both consisting of known fixed numbers. The problem here is to maximize the log-likelihood function (3.2) under the linear restriction $\mathbf{H}\boldsymbol{\beta} - \mathbf{h} = 0$, where $\mathbf{H} = (\mathbf{H}_1, \cdots, \mathbf{H}_q)$ and $\mathbf{h} = (h_1, \cdots, h_q)$. One of the most popular and efficient methods, the so-called penalty function method (for details see, Fiacco and McCormick (1968)) can be applied to solve this constrained optimization problems. This method transforms a constrained problem into a non-constrained problem by adding penalty coefficients to the objective function. Cysneiros and Paula (2005) and Nyquist (1991) investigated this problem in GLMs. We will apply this methodology of penalty functions by considering the quadratic penalty function

$$F(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \sum_{i=1}^{n} [(y_i\theta_i - b(\theta_i)) + \ln c(y_i)] + \sum_{j=1}^{q} \lambda_j (h_j - \mathbf{H}_j'\boldsymbol{\beta})^2.$$

This procedure consists in finding $\text{Max}_{\boldsymbol{\beta}}\ F(\boldsymbol{\beta}, \boldsymbol{\lambda})$ for positive and fixed values of $\lambda_j, \quad j = 1, \cdots, q$. The solution for $\boldsymbol{\beta}$ will be denoted by $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ with $\boldsymbol{\lambda}$=$(\lambda_1, \cdots, \lambda_q)$. The restricted estimator of $\boldsymbol{\beta}$ is given by

$$\tilde{\boldsymbol{\beta}} = \lim_{\boldsymbol{\lambda} \to \infty} \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}), \quad \text{[See, Cysneiros and Paula (2005)]}.$$

Here $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ is an unrestricted estimator for each finite $\boldsymbol{\lambda}$ and $\hat{\boldsymbol{\beta}}(\mathbf{0})$ equals the unrestricted maximum likelihood estimator.

For computation of $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ we apply a similar approach to that of the unrestricted estimation problem that we presented in the previous section. Differentiating $F(\boldsymbol{\beta}, \boldsymbol{\lambda})$

with respect to β_j yields

$$\frac{\partial F}{\partial \beta_j} = \sum_{j=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{d\mu_i}{d\eta_i} x_{ij} + \sum_{l=1}^{q} H_{lj}\lambda_l(h_l - \mathbf{H}_l'\boldsymbol{\beta}), \quad j = 1, \cdots, k,$$

and the expectation of the matrix with second derivative above is

$$E\left(-\frac{\partial^2 F}{\partial \beta_j \partial \beta_l}\right) = \sum_{j=1}^{n} \frac{x_{ij}x_{il}}{V(\mu_i)} \left(\frac{d\mu_i}{d\eta_i}\right)^2 + \sum_{i=1}^{q} \lambda_i H_{ij} H_{il}. \tag{3.18}$$

Using Fisher's scoring method and the above equation, (3.10) can be written as

$$(\mathbf{X}'\mathbf{W}\mathbf{X} + \mathbf{H}'\boldsymbol{\Lambda}\mathbf{H})\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})^{(r+1)} = \mathbf{X}'\mathbf{W}\mathbf{z} + \mathbf{H}'\boldsymbol{\Lambda}\mathbf{h}, \tag{3.19}$$

where $\boldsymbol{\Lambda}$ is the $q \times q$ diagonal matrix with λ_j, $j = 1, \cdots, q$, as diagonal elements. If $\boldsymbol{\Lambda}$ and $\mathbf{X}'\mathbf{W}\mathbf{X}$ are invertible, then by the binomial inversion theorem [Strang (2003)], (3.19) can be written as

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})^{(r+1)} &= (\mathbf{X}'\mathbf{W}\mathbf{X} + \mathbf{H}'\boldsymbol{\Lambda}\mathbf{H})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{z} + \mathbf{H}'\boldsymbol{\Lambda}\mathbf{h}) \\
&= [(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\boldsymbol{\Lambda}\,(\boldsymbol{\Lambda} \\
&\quad + \boldsymbol{\Lambda}\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\boldsymbol{\Lambda})^{-1}\,\boldsymbol{\Lambda}\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}](\mathbf{X}'\mathbf{W}\mathbf{z} + \mathbf{H}'\boldsymbol{\Lambda}\mathbf{h}) \\
&= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z} \\
&\quad +(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\boldsymbol{\Lambda}\,(\mathbf{I} + \mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\boldsymbol{\Lambda})^{-1}\,(\mathbf{I} + \mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\boldsymbol{\Lambda})\,\mathbf{h} \\
&\quad -(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\boldsymbol{\Lambda}\,(\mathbf{I} + \mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\boldsymbol{\Lambda})^{-1}\,\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z} \\
&\quad -(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\boldsymbol{\Lambda}\,(\mathbf{I} + \mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\boldsymbol{\Lambda})^{-1}\,\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\boldsymbol{\Lambda}\mathbf{h} \\
&= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z} \\
&\quad +(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\,(\boldsymbol{\Lambda}^{-1} + \mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}')^{-1}\,[\mathbf{h} - \mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}].
\end{aligned}$$

The $(r+1)$st approximation $\tilde{\boldsymbol{\beta}}^{(r+1)}$ of the restricted maximum likelihood estimate

(RE) $\tilde{\boldsymbol{\beta}}$ is finally obtained as

$$\begin{aligned}\tilde{\boldsymbol{\beta}}^{(r+1)} &= \lim_{\boldsymbol{\lambda}\to\infty} \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})^{(r+1)} \\ &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\left[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\right]^{-1} \\ &\quad \times[\mathbf{h} - \mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}]. \qquad (3.20)\end{aligned}$$

Alteratively, (3.20) may be written as

$$\tilde{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\left[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\right]^{-1}[\mathbf{h} - \mathbf{H}\hat{\boldsymbol{\beta}}^{(r)}], \qquad (3.21)$$

for $r = 0, 1, \cdots$, where $\tilde{\boldsymbol{\beta}}^{(r+1)}$ can be considered as an unrestricted weighted least squares estimate $\hat{\boldsymbol{\beta}}^{(r)} = \mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}$ (with the weights evaluated at the restricted estimate) to which a correction term is added.

Under some regularity conditions (see for instance, Gourieroux and Monford (1995), Section 10.3), it may be showed that that $\tilde{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$, and

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N_k\left(\mathbf{0}, \tilde{\mathbf{J}}^{-}\right),$$

where $\tilde{\mathbf{J}}^{-}$ is the generalized inverse [see Rao (1962)] of matrix $\tilde{\mathbf{J}}$ and

$$\tilde{\mathbf{J}} = \lim_{\boldsymbol{\lambda}\to\infty}\left[\lim_{n\to\infty}\frac{1}{n}E\left(-\frac{\partial^2 F}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right)\right].$$

which may be evaluated at some consistent estimators of $\boldsymbol{\beta}$, such as $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$.

3.4.1 Hypothesis testing

In this section we consider the test of hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ against $H_a : \mathbf{H}\boldsymbol{\beta} \neq \mathbf{h}$. The usual methods for testing these linear hypothesis are the likelihood ratio, Wald and Rao scores tests.

The Likelihood Ratio test: The likelihood ratio test involves estimation of

both the restricted and unrestricted models and a comparison of the values of the log-likelihoods. If the difference is "small", we accept (or strictly speaking fail to reject) the restrictions on the parameters; otherwise we reject the restrictions. If $l(\hat{\boldsymbol{\beta}})$ and $l(\tilde{\boldsymbol{\beta}})$ are the values of log-likelihood at the restricted and unrestricted estimates respectively, then the deviance measure D_1, is twice the difference of the values of the log-likelihood functions i.e.,

$$\begin{aligned} D_1 &= 2[l(\hat{\boldsymbol{\beta}}; y_1, \cdots, y_n) - l(\tilde{\boldsymbol{\beta}}; y_1, \cdots, y_n)] \\ &= (\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h})' \left(\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\right)^{-1} (\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h}) + o_p(1). \end{aligned}$$

Under usual asymptotic properties, the deviance follows an approximate χ^2 distribution with q degrees of freedom when H_0 is true.

The Wald test: Under some regularity conditions [see Fahrmeir and Kaufmann (1985)], the estimator $\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h}$ has an approximate multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix, $\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'$. A Wald statistic can now be defined as

$$D_2 = (\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h})' \left(\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\right)^{-1} (\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h}).$$

Under the null hypothesis this statistic has an approximate χ^2 distribution with q degrees of freedom.

The Rao scores test: This test is computed using the score vector or gradient of the unrestricted model evaluated at the restricted estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. The score statistic is given by

$$\begin{aligned} D_3 &= \left(\mathbf{U}(\tilde{\boldsymbol{\beta}}) - \mathbf{U}(\hat{\boldsymbol{\beta}})\right)' (\mathrm{var}(\hat{\boldsymbol{\beta}}))^{-1} \left(\mathbf{U}(\tilde{\boldsymbol{\beta}}) - \mathbf{U}(\hat{\boldsymbol{\beta}})\right) \\ &= (\mathbf{z} - \boldsymbol{\eta})'\mathbf{W}'\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}(\mathbf{z} - \boldsymbol{\eta}). \end{aligned}$$

Under the null hypothesis, the statistic has asymptotically a χ^2 distribution with q

degrees of freedom.

The likelihood ratio statistic uses most information. When n tends to infinity the likelihood ratio, Wald and score tests are asymptotically equivalent. This means that, under the null hypothesis, as $n \to \infty$, the test statistics all tend to the same random variable, which has a χ^2 distribution with q degrees of freedom. For small n the Likelihood ratio statistic is more reliable than the Wald statistic. The proposed estimation strategies based on the likelihood ratio statistic will be illustrated in the next section.

3.5 Estimation Strategies

3.5.1 Pretest Estimator

The pretest estimator (PT) of $\boldsymbol{\beta}$ based on $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ is defined as

$$\hat{\boldsymbol{\beta}}^{PT} = \hat{\boldsymbol{\beta}} - (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})I(D_1 \leq \chi^2_{q,\alpha}), \quad q \geq 1,$$

where $I(A)$ is an indicator function of a set A and $\chi^2_{q,\alpha}$ is the α-level critical value of the distribution of D_1 under H_0. This estimator is a convex combination of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ via a test statistic, D_1, for testing $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ in (3.1). The PT $\hat{\boldsymbol{\beta}}^{PT}$ chooses $\hat{\boldsymbol{\beta}}$ or $\tilde{\boldsymbol{\beta}}$ according to whether H_0 is rejected or accepted. It is important to remark that $\hat{\boldsymbol{\beta}}^{PT}$ is bounded and performs better than $\tilde{\boldsymbol{\beta}}$ in some part of the parameter space. For details, see Judge and Bock (1978), Ahmed (2001), and Ahmed *et al.* (2006*a*) among others. Since the PT is a discontinuous function of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ and depends on the choice of the level of significance α, we may overcome this limitation by defining James-Stein type (shrinkage) estimator in the next section.

3.5.2 Shrinkage and Positive Shrinkage Estimator

The shrinkage estimator (SE) of $\boldsymbol{\beta}$ can be defined as:

$$\hat{\boldsymbol{\beta}}^S = \tilde{\boldsymbol{\beta}} + \left(1 - (q-2)D_1^{-1}\right)(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}), \quad q \geq 3.$$

This estimator is a weighted average of unrestricted and restricted estimators, the weight being a function of deviance statistics used to test the hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}$. The major problem with this estimator is that it may have a different sign from the unrestricted estimator, $\hat{\boldsymbol{\beta}}$, perhaps due to over-shrinking. The change of sign certainly would make researchers rather uncomfortable. To avoid the over-shrinking inherent in $\hat{\boldsymbol{\beta}}^S$, we define a positive shrinkage estimator which will control the possible over-shrinking problem, for details see Chapter 2. The positive shrinkage (PSE) estimator is defined as

$$\hat{\boldsymbol{\beta}}^{S+} = \tilde{\boldsymbol{\beta}} + \left(1 - (q-2)D_1^{-1}\right)^+ (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}),$$

where $z^+ = max(0, z)$.

3.5.3 Park and Hastie Estimators

The L_1 regularization procedure (Park and Hastie (2007)), proposed for fitting generalized linear models, is a useful tool for selecting variables according to the amount of penalization on the L_1 norm of the coefficients, in a manner less greedy than forward selection/backward deletion. It is similar to the LASSO procedure, in which the loss function is replaced by the negetive log-likelihood of any distribution in the exponential family. Since we assume that the dispersion parameter of this family is known, we are interested (in comparison with the shrinkage and pretest estimation method) in finding the maximum likelihood solution for the natural parameter $\boldsymbol{\theta}$, and thus $\boldsymbol{\beta}$,

with a penalization on the size of the L_1 norm of the coefficients ($||\boldsymbol{\beta}||_1$) i.e.,

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\lambda) &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{-l(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1\} \\ &= -\sum_{i=1}^{n}[(y_i\theta_i - b(\theta_i)) + \ln c(y_i)] + \lambda||\boldsymbol{\beta}||_1, \end{aligned} \tag{3.22}$$

where $\lambda > 0$ is the regularization parameter. If $\lambda = 0$, this just gives the maximum likelihood estimates. However, larger values of λ produce shrunken estimates of $\boldsymbol{\beta}$, often with many components equal to zero. Park and Hastie (2007) introduce an algorithm that efficiently computes solutions along the entire regularization path of the coefficient estimates as λ varies by using the predictor-corrector method of convex-optimization. Starting from $\lambda = \lambda_{max}$, where λ_{max} is the largest λ that makes $\hat{\boldsymbol{\beta}}(\lambda)$ nonzero, this algorithm computes a series of solutions, each time estimating the coefficients with a smaller λ based on previous estimate. The final estimate is denoted as the PH estimator. The regularization parameter λ is selected using k-fold cross validation. For each fold, we obtain a series of models based on BIC (Bayesian Information Criteria) corresponding to the candidate values of λ and compute log-likelihoods using the omitted fold. Then we choose the value of λ for which the average cross-validated (negative) log-likelihood is minimized. Note that the output of the L_1 regularization algorithm looks like a shrinkage and pretest methods by both shrinking and deleting coefficients. However, it is different from the shrinkage and pretest estimation procedure in that it considers all the covariates coefficients equally.

3.6 Asymptotic Results

In this section, we obtain expressions for the asymptotic distributional quadratic bias (ADB) and quadratic risks (ADR) of the proposed estimators. We define a quadratic

loss function using a positive semi-definite matrix $\mathbf{Q}$,

$$\mathcal{L}(\boldsymbol{\beta}^*; \mathbf{Q}) = \left[\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})\right]' \mathbf{Q} \left[\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})\right], \tag{3.23}$$

where $\boldsymbol{\beta}^*$ can be any one of $\hat{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}^{PT}$, $\hat{\boldsymbol{\beta}}^{S}$ or $\hat{\boldsymbol{\beta}}^{S+}$.

We note that, as the test statistics D_1, D_2 and D_3 are consistent against fixed $\boldsymbol{\beta}$ such that $\mathbf{H}\boldsymbol{\beta} \neq \mathbf{h}$, so we will investigate the properties of the estimators under local alternatives. Thus, consider the following local alternatives:

$$K_{(n)} : \mathbf{H}\boldsymbol{\beta} = \mathbf{h} + \frac{\boldsymbol{\delta}}{\sqrt{n}}, \tag{3.24}$$

where $\boldsymbol{\delta} = (\delta_1, \delta_2 \cdots, \delta_q) \in \Re^q$, a real fixed vector. Note that for $\boldsymbol{\delta} = \mathbf{0}$, $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$, for all n. Hence (3.1) is a particular case of (3.24).

Now we introduce the asymptotic distribution function of $\boldsymbol{\beta}^*$ under $K_{(n)}$ by

$$G(\mathbf{y}) = \lim_{n\to\infty} P\left[\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) \leq \mathbf{y} | K_{(n)}\right],$$

where $G(\mathbf{y})$ is nondegenerate distribution function. Then, we define the asymptotic distributional quadratic risk (ADR) by

$$\begin{aligned} R(\boldsymbol{\beta}^*; \mathbf{Q}) &= \int \cdots \int \mathbf{y}'\mathbf{Q}\mathbf{y} dG(\mathbf{y}) \\ &= \text{trace}(\mathbf{Q}\mathbf{Q}^*), \end{aligned}$$

where $\mathbf{Q}^* = \int \cdots \int \mathbf{y}\mathbf{y}' dG(\mathbf{y})$ is the dispersion matrix for the distribution $G(\mathbf{y})$.

Theorem 3.6.1. *Under local alternatives and the usual regularity conditions [see Fahrmeir and Kaufmann (1985)] and as n increases, we have the following:*

1. $\sqrt{n}(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h}) \xrightarrow{\mathcal{L}} N(\boldsymbol{\delta}, \mathbf{H}^{-1}\mathbf{H}')$, *where* $\mathbf{B} = \lim_{n\to\infty} \frac{\mathbf{X}'\mathbf{W}\mathbf{X}}{n}$ *is a nonsingular matrix of order* $k \times k$.

2. *The test statistics* D_1, D_2 *and* D_3 *converge to a non-central chi-squared distribu-*

tion with q degrees of freedom and non-centrality parameter $\Delta = \boldsymbol{\delta}'(\mathbf{H}\mathbf{B}^{-1}\mathbf{H}')^{-1}\boldsymbol{\delta}$.

Now we consider the computation of biases and risks of the proposed estimators under local alternatives $K_{(n)}$.

Theorem 3.6.2. *Under local alternatives* $K_{(n)}$ *in (3.24) and assume that the Theorem 3.6.1 holds, we have the ADB of the proposed estimators as* $n \to \infty$ *in the following:*

$$
\begin{aligned}
ADB(\hat{\boldsymbol{\beta}}) &= \mathbf{0}, && (3.25)\\
ADB(\tilde{\boldsymbol{\beta}}) &= -\mathbf{J}\boldsymbol{\delta}, \quad \mathbf{J} = \mathbf{B}^{-1}\mathbf{H}'[\mathbf{H}\mathbf{B}^{-1}\mathbf{H}']^{-1}, && (3.26)\\
ADB(\hat{\boldsymbol{\beta}}^{PT}) &= \mathbf{J}\boldsymbol{\delta}\Psi_{q+2}(q-2,\Delta), && (3.27)\\
ADB(\hat{\boldsymbol{\beta}}^{S}) &= -(q-2)\mathbf{J}\boldsymbol{\delta}E(\chi^{-2}_{q+2}(\Delta)), && (3.28)\\
ADB(\hat{\boldsymbol{\beta}}^{S+}) &= -(q-2)\mathbf{J}\boldsymbol{\delta}\left[E(\chi^{-2}_{q+2}(\Delta)) - E(\chi^{-2}_{q+2}(\Delta)I(\chi^{2}_{q+2}(\Delta) < (q-2)))\right]\\
&\quad - \mathbf{J}\boldsymbol{\delta}\Psi_{q+2}(q-2,\Delta), && (3.29)
\end{aligned}
$$

where the notation $\Psi_\nu(q-2,\Delta)$ *is the distribution function of non-central chi-square distribution with* ν *degrees of freedom and non-centrality parameter* Δ.

Proof:

By definition, we have

$$
\begin{aligned}
ADB(\hat{\boldsymbol{\beta}}) &= \lim_{n\to\infty} E\{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\} = \mathbf{0},\\
ADB(\tilde{\boldsymbol{\beta}}) &= \lim_{n\to\infty} E\{\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\}\\
&= \lim_{n\to\infty} E\{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) - (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}']^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})\}\\
&= \mathbf{0} - \mathbf{B}^{-1}\mathbf{H}'(\mathbf{H}\mathbf{B}^{-1}\mathbf{H}')^{-1}\lim_{n\to\infty}\sqrt{n}(\mathbf{H}\boldsymbol{\beta}-\mathbf{h})\\
&= -\mathbf{J}\boldsymbol{\delta},\\
ADB(\hat{\boldsymbol{\beta}}^{PT}) &= \lim_{n\to\infty} E\{\sqrt{n}(\hat{\boldsymbol{\beta}}^{PT}-\boldsymbol{\beta})\}\\
&= \lim_{n\to\infty} E\{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\} - \mathbf{J}\lim_{n\to\infty} E\{\sqrt{n}(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})I(D_1 < \chi^2_{q,\,\alpha}\}\\
&= \mathbf{0} - \mathbf{J}\boldsymbol{\delta}\Psi_{q+2}(q-2,\Delta) = -\mathbf{J}\boldsymbol{\delta}\Psi_{q+2}(q-2,\Delta),\\
ABD(\hat{\boldsymbol{\beta}}^{S}) &= \lim_{n\to\infty} E\{\sqrt{n}(\hat{\boldsymbol{\beta}}^{S}-\boldsymbol{\beta})\}
\end{aligned}
$$

$$
\begin{aligned}
&= \lim_{n\to\infty} E\{\sqrt{n}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \frac{(q-2)}{D_1}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})]\} \\
&= \lim_{n\to\infty} E\{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\} - (q-2)\mathbf{J}\lim_{n\to\infty} E\left[\frac{\sqrt{n}(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})}{D_1}\right] \\
&= -(q-2)\mathbf{J}\boldsymbol{\delta}E(\chi^{-2}_{q+2}(\Delta)), \text{by theorem 2.2.4 [Saleh (2006), p.32].} \\
ADB(\hat{\boldsymbol{\beta}}^{S+}) &= \lim_{n\to\infty} E\{\sqrt{n}(\hat{\boldsymbol{\beta}}^{S+}-\boldsymbol{\beta})\} \\
&= \lim_{n\to\infty} E\{\sqrt{n}[(\hat{\boldsymbol{\beta}}^{S}-\boldsymbol{\beta}) - (\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})I(D_1 < q-2) \\
&+ \frac{(q-2)}{D_1}I(D_1 < q-2)(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})]\}, \\
&= -(q-2)\mathbf{J}\boldsymbol{\delta}E(\chi^{-2}_{q+2}(\Delta)) - \mathbf{J}\lim_{n\to\infty} E\{\sqrt{n}(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})I(D_1 < q-2)\} \\
&+ \mathbf{J}\lim_{n\to\infty} E\{\sqrt{n}(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})I(D_1 < q-2)D_1^{-1}\} \\
&= -(q-2)\mathbf{J}\boldsymbol{\delta}\left[E(\chi^{-2}_{q+2}(\Delta)) - E(\chi^{-2}_{q+2}(\Delta)I(\chi^{2}_{q+2}(\Delta) < q-2)\right] \\
&- \mathbf{J}\boldsymbol{\delta}\Psi_{q+2}(q-2,\Delta).
\end{aligned}
$$

Since bias is a component of ADR, we will discuss the ADR of the estimators from here onward. Under local alternatives, the ADRs of the estimators are given in the following theorem.

Theorem 3.6.3. *Under local alternatives $K_{(n)}$ in (3.24) and assume that the Theorem 3.6.1 holds, we have the ADRs of $\hat{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}^{PT}$, $\hat{\boldsymbol{\beta}}^{S}$ and $\hat{\boldsymbol{\beta}}^{S+}$ are respectively:*

$$
\begin{aligned}
R(\hat{\boldsymbol{\beta}};\mathbf{Q}) &= trace[\mathbf{Q}\mathbf{B}^{-1}], &(3.30)\\
R(\tilde{\boldsymbol{\beta}};\mathbf{Q}) &= R(\hat{\boldsymbol{\beta}};\mathbf{Q}) - trace[\mathbf{Q}\mathbf{J}\mathbf{H}\mathbf{B}^{-1}] + \boldsymbol{\delta}'(\mathbf{J}'\mathbf{Q}\mathbf{J})\boldsymbol{\delta}, &(3.31)\\
R(\hat{\boldsymbol{\beta}}^{PT};\mathbf{Q}) &= R(\hat{\boldsymbol{\beta}};\mathbf{Q}) - trace[\mathbf{Q}\mathbf{J}\mathbf{H}\mathbf{B}^{-1}]\Psi_{q+2}(q-2,\Delta) \\
&+ \boldsymbol{\delta}'(\mathbf{J}'\mathbf{Q}\mathbf{J})\boldsymbol{\delta}[2\Psi_{q+2}(q-2,\Delta) - \Psi_{q+4}(q-2,\Delta)], &(3.32)\\
R(\hat{\boldsymbol{\beta}}^{S};\mathbf{Q}) &= R(\hat{\boldsymbol{\beta}};\mathbf{Q}) - 2(q-2)trace[\mathbf{Q}\mathbf{J}\mathbf{H}\mathbf{B}^{-1}]\{2E(\chi^{-2}_{q+2}(\Delta)) \\
&- (q-2)E(\chi^{-4}_{q+2}(\Delta))\} + (q-2)\boldsymbol{\delta}'(\mathbf{J}'\mathbf{Q}\mathbf{J})\boldsymbol{\delta}\{2E(\chi^{-2}_{q+2}(\Delta)) \\
&- 2E(\chi^{-4}_{q+2}(\Delta)) + (q-2)E(\chi^{-4}_{q+4}(\Delta))\}, &(3.33)\\
R(\hat{\boldsymbol{\beta}}^{S+};\mathbf{Q}) &= R(\hat{\boldsymbol{\beta}}^{S};\mathbf{Q}) - \boldsymbol{\delta}'(\mathbf{J}'\mathbf{Q}\mathbf{J})\boldsymbol{\delta}E[(1-(q-2)\chi^{-2}_{q+4}(\Delta))^2 I(\chi^{2}_{q+4}(\Delta) < q-2)]
\end{aligned}
$$

$$- \quad trace[\mathbf{QJHB}^{-1}]E[(1-(q-2)\chi^{-2}_{q+2}(\Delta))^2 I(\chi^2_{q+4}(\Delta) < q-2)]$$

$$+ \quad 2\boldsymbol{\delta}'(\mathbf{J}'\mathbf{QJ})\boldsymbol{\delta}E[(1-(q-2)\chi^{-2}_{q+4}(\Delta))I(\chi^2_{q+4}(\Delta) < q-2)]. \qquad (3.34)$$

Proof:

By definition we have $R(\hat{\boldsymbol{\beta}};\mathbf{Q}) = \text{trace}[\mathbf{QB}^{-1}]$. To find the risk of $\tilde{\boldsymbol{\beta}}$, we need to evaluate the mean square error (MSE) of $\tilde{\boldsymbol{\beta}}$:

$$\begin{aligned}
\text{MSE}(\tilde{\boldsymbol{\beta}}) &= \lim_{n\to\infty} E\{n(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})'\} \\
&= \lim_{n\to\infty} nE[\{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \mathbf{J}(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})\}\{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})' - (\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})'\mathbf{J}'\}] \\
&= \lim_{n\to\infty} nE[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})' + \mathbf{J}(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})\mathbf{J}' - 2\mathbf{J}(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'].
\end{aligned}$$

First term: Under $K_{(n)}$, the first term can be written as

$$\lim_{n\to\infty} E[n(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'] = \mathbf{B}^{-1}.$$

Second term: To evaluate the second term, let $\mathbf{V_1} = \sqrt{n}(\mathbf{H}\boldsymbol{\beta}-\mathbf{h}) \xrightarrow{\mathcal{L}} N_q(\boldsymbol{\delta}, \mathbf{HB}^{-1}\mathbf{H}')$ and $\mathbf{U_1} = \sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{\mathcal{L}} N_q(\mathbf{0}, \mathbf{B}^{-1})$. If $\boldsymbol{\Gamma}_1$ is a $q \times q$ symmetric and positive definite matrix, then

$$\boldsymbol{\Gamma}_1\mathbf{HB}^{-1}\mathbf{H}'\boldsymbol{\Gamma}_1' = \mathbf{I}_q \Rightarrow \mathbf{HB}^{-1}\mathbf{H}' = (\boldsymbol{\Gamma}_1'\boldsymbol{\Gamma}_1)^{-1}.$$

Now let $\mathbf{S} = \boldsymbol{\Gamma}_1\mathbf{V_1}$. Then

$$E(\mathbf{S}) = \boldsymbol{\Gamma}_1 E(\mathbf{V_1}) = \boldsymbol{\Gamma}_1\boldsymbol{\delta},$$

$$Var(\mathbf{S}) = \boldsymbol{\Gamma}_1 Var(\mathbf{V_1})\boldsymbol{\Gamma}_1' = \boldsymbol{\Gamma}_1\mathbf{HB}^{-1}\mathbf{H}'\boldsymbol{\Gamma}_1' = \mathbf{I}_q.$$

Under $K_{(n)}$, the second term can be simplified as

$$\begin{aligned}
&\lim_{n\to\infty} E[n\mathbf{J}(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})\mathbf{J}'] \\
&= \lim_{n\to\infty} \mathbf{J}E[\mathbf{V_1}\mathbf{V_1}']\mathbf{J}' \\
&= \mathbf{J}\lim_{n\to\infty} E[(\boldsymbol{\Gamma}_1^{-1}\mathbf{S})(\boldsymbol{\Gamma}_1^{-1}\mathbf{S})']\mathbf{J}' \\
&= \mathbf{J}\boldsymbol{\Gamma}_1^{-1}\lim_{n\to\infty} E[\mathbf{SS}']\boldsymbol{\Gamma}_1^{-1'}\mathbf{J}' \\
&= \mathbf{J}[(\boldsymbol{\Gamma}_1'\boldsymbol{\Gamma}_1)^{-1} + \boldsymbol{\Gamma}_1^{-1}(\boldsymbol{\Gamma}_1\boldsymbol{\delta})(\boldsymbol{\Gamma}_1\boldsymbol{\delta})'\boldsymbol{\Gamma}_1^{-1'}]\mathbf{J}'
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{J}[\mathbf{H}\mathbf{B}^{-1}\mathbf{H}' + \boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'] \\
&= \mathbf{J}\mathbf{H}\mathbf{B}^{-1}\mathbf{J}' + \mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}' \\
&= \mathbf{J}\mathbf{H}\mathbf{B}^{-1}\mathbf{H}'[\mathbf{B}^{-1}\mathbf{H}'(\mathbf{H}\mathbf{B}^{-1}\mathbf{H}')^{-1}]' + \mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}' \\
&= \mathbf{J}\mathbf{H}\mathbf{B}^{-1} + \mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'.
\end{aligned}
$$

Third term: Under $K_{(n)}$, the third term can be simplified as

$$
\begin{aligned}
&\lim_{n\to\infty} E[n\mathbf{J}(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\
&= \lim_{n\to\infty} \mathbf{J}E[(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta} + \mathbf{H}\boldsymbol{\beta} - \mathbf{h})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\
&= \lim_{n\to\infty} \mathbf{J}E[\{\mathbf{H}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathbf{H}\boldsymbol{\beta} - \mathbf{h}\}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\
&= \mathbf{J}\mathbf{H} \lim_{n\to\infty} E[n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\
&= \mathbf{J}\mathbf{H}\mathbf{B}^{-1}.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
\mathrm{MSE}(\tilde{\boldsymbol{\beta}}) &= \mathbf{B}^{-1} + \mathbf{J}\mathbf{H}\mathbf{B}^{-1} + \mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}' - 2\mathbf{J}\mathbf{H}\mathbf{B}^{-1} \\
&= \mathbf{B}^{-1} - \mathbf{J}\mathbf{H}\mathbf{B}^{-1} + \mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}', \\
\text{so that } R(\tilde{\boldsymbol{\beta}}; \mathbf{Q}) &= \mathrm{trace}[\mathbf{Q}\mathrm{MSE}(\tilde{\boldsymbol{\beta}})] \\
&= R(\hat{\boldsymbol{\beta}}; \mathbf{Q}) - \mathrm{trace}[\mathbf{Q}\mathbf{J}\mathbf{H}\mathbf{B}^{-1}] + \boldsymbol{\delta}'(\mathbf{J}'\mathbf{Q}\mathbf{J})\boldsymbol{\delta}.
\end{aligned}
$$

The MSE for the pretest estimator is

$$
\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\beta}}^{PT}) &= \lim_{n\to\infty} nE\{(\hat{\boldsymbol{\beta}}^{PT} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^{PT} - \boldsymbol{\beta})'\} \\
&= \lim_{n\to\infty} nE[\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})I(D_1 < \chi^2_{q,\,\alpha})\} \\
&\times \{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' - (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'I(D_1 < \chi^2_{q,\,\alpha})\}] \\
&= \lim_{n\to\infty} \{nE[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] + nE[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'I(D_1 < \chi^2_{q,\,\alpha})] \\
&- 2nE[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'I(D_1 < \chi^2_{q,\,\alpha})]\}.
\end{aligned}
$$

The part of third term can be written as,

$$
\begin{aligned}
& E\{E\left[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'I(D_1<\chi^2_{q,\,\alpha})\right]|(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})\} \\
= \; & E\{(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})E[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'|(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})]I(D_1<\chi^2_{q,\,\alpha})\} \\
& \text{by theorem 7.2.2 [Saleh (2006), p.343], we have} \\
= \; & E\{[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}) \\
- \; & (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}']^{-1}(\mathbf{H}\boldsymbol{\beta}-\mathbf{h})]'I(D_1<\chi^2_{q,\,\alpha})]\} \\
= \; & E[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'I(D_1<\chi^2_{q,\,\alpha})] \\
- \; & E[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'I(D_1<\chi^2_{q,\,\alpha})](\mathbf{H}\boldsymbol{\beta}-\mathbf{h})'[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}']^{-1}\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}.
\end{aligned}
$$

Now

$$
\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\beta}}^{PT}) \; = \; & \lim_{n\to\infty} nE[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'] \\
- \; & \lim_{n\to\infty} nE[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'I(D_1<\chi^2_{q,\,\alpha})] \\
+ \; & 2\lim_{n\to\infty} nE[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'I(D_1<\chi^2_{q,\,\alpha})](\mathbf{H}\boldsymbol{\beta}-\mathbf{h})'[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}']^{-1}\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}.
\end{aligned}
$$

First term:

$$
\lim_{n\to\infty} nE[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'] = \lim_{n\to\infty}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} = \mathbf{B}^{-1}.
$$

Second term:

$$
\begin{aligned}
& -\lim_{n\to\infty} nE[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'I(D_1<\chi^2_{q,\,\alpha})] \\
= \; & -\lim_{n\to\infty} nE[\mathbf{J}(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})'I(D_1<\chi^2_{q,\,\alpha})\mathbf{J}'] \\
= \; & -\lim_{n\to\infty} \mathbf{J}E\left[\mathbf{V}_1\mathbf{V}_1'I(D_1<\chi^2_{q,\,\alpha})\right]\mathbf{J}' \\
= \; & -\mathbf{J}\boldsymbol{\Gamma}_1^{-1}\lim_{n\to\infty} E[\mathbf{S}\mathbf{S}'I(\chi^2_{q+2}(\Delta)<\chi^2_{q,\,\alpha})]\boldsymbol{\Gamma}_1^{-1'}\mathbf{J}' \\
= \; & -\mathbf{J}[(\boldsymbol{\Gamma}_1'\boldsymbol{\Gamma}_1)^{-1}\Psi_{q+2}(q-2,\Delta)+\boldsymbol{\Gamma}_1^{-1}(\boldsymbol{\Gamma}_1\boldsymbol{\delta})(\boldsymbol{\Gamma}_1\boldsymbol{\delta})'\boldsymbol{\Gamma}_1'^{-1}\Psi_{q+4}(q-2,\Delta)]\mathbf{J}' \\
= \; & -\mathbf{J}[\mathbf{H}\mathbf{B}^{-1}\mathbf{H}'\Psi_{q+2}(q-2,\Delta)+\boldsymbol{\delta}\boldsymbol{\delta}'\Psi_{q+4}(q-2,\Delta)]\mathbf{J}' \\
= \; & -\mathbf{J}\mathbf{H}\mathbf{B}^{-1}\Psi_{q+2}(q-2,\Delta)-\mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'\Psi_{q+4}(q-2,\Delta).
\end{aligned}
$$

Third term:

$$
2\lim_{n\to\infty}\{E[n(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'I(D_1<\chi^2_{q,\,\alpha})](\mathbf{H}\boldsymbol{\beta}-\mathbf{h})'[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}']^{-1}\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\}
$$

$$
\begin{aligned}
&= 2 \lim_{n\to\infty} E[\sqrt{n}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})' I(D_1 < \chi^2_{q,\,\alpha})]\boldsymbol{\delta}'\mathbf{J}' \\
&= 2\mathbf{J} \lim_{n\to\infty} E\{\sqrt{n}(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h}) I(D_1 < \chi^2_{q,\,\alpha})\}\boldsymbol{\delta}'\mathbf{J}' \\
&= 2\mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'\Psi_{q+2}(q-2,\Delta).
\end{aligned}
$$

Adding first, second and third terms together we have,

$$
\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\beta}}^{PT}) &= \mathbf{B}^{-1} - \mathbf{J}\mathbf{H}\mathbf{B}^{-1}\Psi_{q+2}(q-2,\Delta) - \mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'\Psi_{q+4}(q-2,\Delta) \\
&+ 2\mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'\Psi_{q+2}(q-2,\Delta) \\
&= \mathbf{B}^{-1} - \mathbf{J}\mathbf{H}\mathbf{B}^{-1}\Psi_{q+2}(q-2,\Delta) \\
&+ \boldsymbol{\delta}'(\mathbf{J}'\mathbf{Q}\mathbf{J})\boldsymbol{\delta}\,[2\Psi_{q+2}(q-2,\Delta) - \Psi_{q+4}(q-2,\Delta)] \\
\text{so that } R(\hat{\boldsymbol{\beta}}^{PT};\mathbf{Q}) &= \mathrm{trace}[\mathbf{Q}\mathrm{MSE}(\hat{\boldsymbol{\beta}}^{PT})] \\
&= R(\hat{\boldsymbol{\beta}};\mathbf{Q}) - \mathrm{trace}[\mathbf{Q}\mathbf{J}\mathbf{H}\mathbf{B}^{-1}]\Psi_{q+2}(q-2,\Delta) \\
&+ \boldsymbol{\delta}'(\mathbf{J}'\mathbf{Q}\mathbf{J})\boldsymbol{\delta}\,[2\Psi_{q+2}(q-2,\Delta) - \Psi_{q+4}(q-2,\Delta)].
\end{aligned}
$$

The MSE for the shrinkage estimator

$$
\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\beta}}^{S}) &= \lim_{n\to\infty} nE\{(\hat{\boldsymbol{\beta}}^{S} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^{S} - \boldsymbol{\beta})'\} \\
&= \lim_{n\to\infty} \{nE[\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{q-2}{D_1}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})\}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' - \frac{q-2}{D_1}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'\}]\} \\
&= \lim_{n\to\infty} nE[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] + (q-2)^2 \lim_{n\to\infty} nE[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})' D_1^{-2}] \\
&- 2(q-2) \lim_{n\to\infty} nE[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' D_1^{-1}] \\
&= \mathbf{B}^{-1} + (q-2)^2 \lim_{n\to\infty} nE[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})' D_1^{-2}] \\
&- 2(q-2) \lim_{n\to\infty} nE\{[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \\
&- (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}']^{-1}(\mathbf{H}\boldsymbol{\beta} - \mathbf{h})]' D_1^{-1}]\} \\
&= \mathbf{B}^{-1} + (q-2)^2 \lim_{n\to\infty} \mathbf{J}E[n(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h})(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h})' D_1^{-2}]\mathbf{J}' \\
&- 2(q-2) \lim_{n\to\infty} \mathbf{J}E[n(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h})(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h})' D_1^{-1}]\mathbf{J}' \\
&+ 2(q-2) \lim_{n\to\infty} n\{E[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})' D_1^{-1}](\mathbf{H}\boldsymbol{\beta} - \mathbf{h})'[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}']^{-1}\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\}.
\end{aligned}
$$

Second term:

$$
\begin{aligned}
& (q-2)^2 \lim_{n\to\infty} \mathbf{J}E\left[\mathbf{V_1}\mathbf{V_1}'D_1^{-2}\right]\mathbf{J}' \\
= \; & (q-2)^2\mathbf{J}\boldsymbol{\Gamma}_1^{-1} \lim_{n\to\infty}\{E[\mathbf{S}\mathbf{S}'D_1^{-2}]\}\boldsymbol{\Gamma}_1^{-1'}\mathbf{J}' \\
= \; & (q-2)^2\mathbf{J}[(\boldsymbol{\Gamma}_1'\boldsymbol{\Gamma}_1)^{-1}E(\chi_{q+2}^{-4}(\Delta)) + \boldsymbol{\Gamma}_1^{-1}(\boldsymbol{\Gamma}_1\boldsymbol{\delta})(\boldsymbol{\Gamma}_1\boldsymbol{\delta})'E(\chi_{q+4}^{-4}(\Delta))\boldsymbol{\Gamma}_1^{-1'}]\mathbf{J}' \\
= \; & (q-2)^2\mathbf{J}[\mathbf{H}\mathbf{B}^{-1}\mathbf{H}'E(\chi_{q+2}^{-4}(\Delta)) + \boldsymbol{\delta}\boldsymbol{\delta}'E(\chi_{q+4}^{-4}(\Delta))]\mathbf{J}' \\
= \; & (q-2)^2\mathbf{J}\mathbf{H}\mathbf{B}^{-1}E(\chi_{q+2}^{-4}(\Delta)) + (q-2)^2\mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'E(\chi_{q+4}^{-4}(\Delta)).
\end{aligned}
$$

Third term:

$$
\begin{aligned}
& -2(q-2)\lim_{n\to\infty}\mathbf{J}E[n(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})'D_1^{-1}]\mathbf{J}' \\
= \; & -2(q-2)\mathbf{J}\mathbf{H}\mathbf{B}^{-1}E(\chi_{q+2}^{-2}(\Delta)) - 2(q-2)\mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'E(\chi_{q+4}^{-2}(\Delta)).
\end{aligned}
$$

Fourth term:

$$
\begin{aligned}
& 2(q-2)\lim_{n\to\infty} n\{E[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'D_1^{-1}](\mathbf{H}\boldsymbol{\beta}-\mathbf{h})'[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}']^{-1}\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\} \\
= \; & 2(q-2)\lim_{n\to\infty} E[\sqrt{n}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'D_1^{-1}]\boldsymbol{\delta}'\mathbf{J}' \\
= \; & 2(q-2)\mathbf{J}\lim_{n\to\infty} E\{\sqrt{n}(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})D_1^{-1}\}\boldsymbol{\delta}'\mathbf{J}' \\
= \; & 2(q-2)\mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'E(\chi_{q+2}^{-2}(\Delta)).
\end{aligned}
$$

Adding First, second, third and fourth terms we have,

$$
\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\beta}}^S) &= \mathbf{B}^{-1} + (q-2)^2\mathbf{J}\mathbf{H}\mathbf{B}^{-1}E(\chi_{q+2}^{-4}(\Delta)) + (q-2)^2\mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'E(\chi_{q+4}^{-4}(\Delta)) \\
&- 2(q-2)\mathbf{J}\mathbf{H}\mathbf{B}^{-1}E(\chi_{q+2}^{-2}(\Delta)) - 2(q-2)\mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'E(\chi_{q+4}^{-2}(\Delta)) \\
&+ 2(q-2)\mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'E(\chi_{q+2}^{-2}(\Delta)) \\
&= \mathbf{B}^{-1} - (q-2)\mathbf{J}\mathbf{H}\mathbf{B}^{-1}[E(\chi_{q+2}^{-2}(\Delta)) - (q-2)E(\chi_{q+2}^{-4}(\Delta))] \\
&+ (q-2)\mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'[2E(\chi_{q+2}^{-2}(\Delta)) - 2E(\chi_{q+4}^{-2}(\Delta)) + (q-2)E(\chi_{q+4}^{-4}(\Delta))].
\end{aligned}
$$

Now the risk of shrinkage estimator is

$$
\begin{aligned}
R(\hat{\boldsymbol{\beta}}^S;\mathbf{Q}) &= \mathrm{trace}[\mathbf{Q}\mathrm{MSE}(\hat{\boldsymbol{\beta}}^S)] \\
&= R(\hat{\boldsymbol{\beta}};\mathbf{Q}) - (q-2)\mathrm{trace}[\mathbf{Q}\mathbf{J}\mathbf{H}\mathbf{B}^{-1}]\{E(\chi_{q+2}^{-2}(\Delta)) - (q-2)E(\chi_{q+2}^{-4}(\Delta))\}
\end{aligned}
$$

$$+ \quad (q-2)\boldsymbol{\delta}'(\mathbf{J}'\mathbf{Q}\mathbf{J})\boldsymbol{\delta}\{2E(\chi^{-2}_{q+2}(\Delta)) - 2E(\chi^{-2}_{q+4}(\Delta)) + (q-2)E(\chi^{-4}_{q+4}(\Delta))\}.$$

The MSE for the positive shrinkage estimator

$$\begin{aligned}
&\mathrm{MSE}(\hat{\boldsymbol{\beta}}^{S+}) \\
&= \lim_{n\to\infty} nE\{(\hat{\boldsymbol{\beta}}^{S+} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^{S+} - \boldsymbol{\beta})'\} \\
&= \lim_{n\to\infty} nE\left[\{(\hat{\boldsymbol{\beta}}^{S} - \boldsymbol{\beta}) - \left(1 - \frac{q-2}{D_1}\right)(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})I(D_1 < q-2)\}\right. \\
&\quad \left.\{(\hat{\boldsymbol{\beta}}^{S} - \boldsymbol{\beta})' - \left(1 - \frac{q-2}{D_1}\right)(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'I(D_1 < q-2)\}\right] \\
&= \lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}}^{S} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^{S} - \boldsymbol{\beta})'\right] \\
&+ \lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'\left(1 - \frac{q-2}{D_1}\right)^2 I(D_1 < q-2)\right] \\
&- 2\lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{S} - \boldsymbol{\beta})'\left(1 - \frac{q-2}{D_1}\right) I(D_1 < q-2)\right] \\
&= \lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}}^{S} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^{S} - \boldsymbol{\beta})'\right] \\
&+ \lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'\left(1 - \frac{q-2}{D_1}\right)^2 I(D_1 < q-2)\right] \\
&- 2\lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\left(1 - \frac{q-2}{D_1}\right) I(D_1 < q-2)\right] \\
&- 2\lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'\left(1 - \frac{q-2}{D_1}\right)^2 I(D_1 < q-2)\right] \\
&+ 2\lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'\left(1 - \frac{q-2}{D_1}\right) I(D_1 < q-2)\right] \\
&= \mathrm{MSE}(\hat{\boldsymbol{\beta}}^{S}) - \lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'\left(1 - \frac{q-2}{D_1}\right)^2 I(D_1 < q-2)\right] \\
&- 2\lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\left(1 - \frac{q-2}{D_1}\right) I(D_1 < q-2)\right] \\
&+ 2\lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'\left(1 - \frac{q-2}{D_1}\right) I(D_1 < q-2)\right] \\
&= \mathrm{MSE}(\hat{\boldsymbol{\beta}}^{S}) - \lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'\left(1 - \frac{q-2}{D_1}\right)^2 I(D_1 < q-2)\right] \\
&- 2\lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'\left(1 - \frac{q-2}{D_1}\right) I(D_1 < q-2)\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \ 2\lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'\left(1-\frac{q-2}{D_1}\right)I(D_1<q-2)\right] \\
&\times \ (\mathbf{H}\boldsymbol{\beta}-\mathbf{h})'[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}']^{-1}\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \\
&+ \ 2\lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'\left(1-\frac{q-2}{D_1}\right)I(D_1<q-2)\right] \\
&= \ \mathrm{MSE}(\hat{\boldsymbol{\beta}}^S)-\lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'\left(1-\frac{q-2}{D_1}\right)^2 I(D_1<q-2)\right] \\
&+ \ 2\lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'\left(1-\frac{q-2}{D_1}\right)I(D_1<q-2)\right] \\
&\times \ (\mathbf{H}\boldsymbol{\beta}-\mathbf{h})'[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}']^{-1}\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}.
\end{aligned}
$$

Second term:

$$
\begin{aligned}
&-\lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'\left(1-\frac{q-2}{D_1}\right)^2 I(D_1<q-2)\right] \\
&= \ -\lim_{n\to\infty} \mathbf{J}E\left[n(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})'\left(1-\frac{q-2}{D_1}\right)^2 I(D_1<q-2)\right]\mathbf{J}' \\
&= \ -\lim_{n\to\infty} \mathbf{J}E\left[\mathbf{V_1}\mathbf{V_1}'\left(1-\frac{q-2}{D_1}\right)^2 I(D_1<q-2)\right]\mathbf{J}' \\
&= \ -\mathbf{J}\boldsymbol{\Gamma}_1^{-1}\lim_{n\to\infty}\left\{E\left[\mathbf{S}\mathbf{S}'\left(1-\frac{q-2}{D_1}\right)^2 I(D_1<q-2)\right]\right\}\boldsymbol{\Gamma}_1^{-1\prime}\mathbf{J}' \\
&= \ -\mathbf{J}\left\{(\boldsymbol{\Gamma}_1'\boldsymbol{\Gamma}_1)^{-1}E\left[\left(1-(q-2)\chi_{q+2}^{-2}(\Delta)\right)^2 I(\chi_{q+2}^2(\Delta)<q-2)\right]\right\}\mathbf{J}' \\
&- \ \mathbf{J}\left\{\boldsymbol{\Gamma}_1^{-1}(\boldsymbol{\Gamma}_1\boldsymbol{\delta})(\boldsymbol{\Gamma}_1\boldsymbol{\delta})'E\left[\left(1-(q-2)\chi_{q+4}^{-2}(\Delta)\right)^2 I(\chi_{q+2}^2(\Delta)<q-2)\right]\boldsymbol{\Gamma}_1'\right\}\mathbf{J}' \\
&= \ -\mathbf{J}[\mathbf{H}\mathbf{B}^{-1}\mathbf{H}'E\left[\left(1-(q-2)\chi_{q+2}^{-2}(\Delta)\right)^2 I(\chi_{q+2}^2(\Delta)<q-2)\right]\mathbf{J}' \\
&- \ \mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'E\left[\left(1-(q-2)\chi_{q+4}^{-2}(\Delta)\right)^2 I(\chi_{q+2}^2(\Delta)<q-2)\right] \\
&= \ -\mathbf{J}\mathbf{H}\mathbf{B}^{-1}E\left[\left(1-(q-2)\chi_{q+2}^{-2}(\Delta)\right)^2 I(\chi_{q+2}^2(\Delta)<q-2)\right] \\
&- \ \mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}'E\left[\left(1-(q-2)\chi_{q+4}^{-2}(\Delta)\right)^2 I(\chi_{q+2}^2(\Delta)<q-2)\right].
\end{aligned}
$$

Third term:

$$
\begin{aligned}
&2\lim_{n\to\infty} nE\left[(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'\left(1-\frac{q-2}{D_1}\right)I(D_1<q-2)\right] \\
&\times \ (\mathbf{H}\boldsymbol{\beta}-\mathbf{h})'[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}']^{-1}\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&= 2 \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})'\left(1-\frac{q-2}{D_1}\right) I(D_1 < q-2)\right]\boldsymbol{\delta}'\mathbf{J}' \\
&= 2\mathbf{J} \lim_{n\to\infty} E\left[\sqrt{n}(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{h})\left(1-\frac{q-2}{D_1}\right) I(D_1 < q-2)\right]\boldsymbol{\delta}'\mathbf{J}' \\
&= 2\mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}' E\left[\left(1-(q-2)\chi^{-2}_{q+2}(\Delta)\right) I(\chi^2_{q+2}(\Delta) < q-2)\right].
\end{aligned}
$$

Finally, the MSE of positive shrinkage estimator is

$$
\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\beta}}^{S+}) &= \mathrm{MSE}(\hat{\boldsymbol{\beta}}^{S}) - \mathbf{JHB}^{-1} E\left[\left(1-(q-2)\chi^{-2}_{q+2}(\Delta)\right)^2 I(\chi^2_{q+2}(\Delta) < q-2)\right] \\
&+ 2\mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}' E\left[\left(1-(q-2)\chi^{-2}_{q+2}(\Delta)\right) I(\chi^2_{q+2}(\Delta) < q-2)\right] \\
&- \mathbf{J}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{J}' E\left[\left(1-(q-2)\chi^{-2}_{q+4}(\Delta)\right)^2 I(\chi^2_{q+2}(\Delta) < q-2)\right].
\end{aligned}
$$

Now the risk of shrinkage estimator is

$$
\begin{aligned}
R(\hat{\boldsymbol{\beta}}^{S+};\mathbf{Q}) &= \mathrm{trace}[\mathbf{Q}\mathrm{MSE}(\hat{\boldsymbol{\beta}}^{S+})] \\
&= R(\hat{\boldsymbol{\beta}}^{S};\mathbf{Q}) - \mathrm{trace}[\mathbf{QJHB}^{-1}] E\left[\left(1-(q-2)\chi^{-2}_{q+2}(\Delta)\right)^2 I(\chi^2_{q+2}(\Delta) < q-2)\right] \\
&+ 2\boldsymbol{\delta}'(\mathbf{J}'\mathbf{QJ})\boldsymbol{\delta} E\left[\left(1-(q-2)\chi^{-2}_{q+2}(\Delta)\right) I(\chi^2_{q+2}(\Delta) < q-2)\right] \\
&- \boldsymbol{\delta}'(\mathbf{J}'\mathbf{QJ})\boldsymbol{\delta} E\left[\left(1-(q-2)\chi^{-2}_{q+4}(\Delta)\right)^2 I(\chi^2_{q+2}(\Delta) < q-2)\right].
\end{aligned}
$$

3.6.1 Risk Analysis

We now investigate the risk behavior of the proposed estimators and determine their dominance characteristics.

Risk comparison of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$:

First note that the risk of maximum likelihood estimate of $\hat{\boldsymbol{\beta}}$ is constant, since it does not depend on the non-sample information. The ADR of $\tilde{\boldsymbol{\beta}}$ depends on $\boldsymbol{\delta}'(\mathbf{J}'\mathbf{QJ})\boldsymbol{\delta}$. However, it is superior to $\hat{\boldsymbol{\beta}}$ near the null hypothesis.

Note that $\mathbf{B}^{-1/2}\mathbf{HB}^{-1}\mathbf{H}'\mathbf{B}^{-1/2}$ is a symmetric idempotent matrix with rank $q(\leq$

k). Thus there exists an orthogonal matrix $\mathbf{\Lambda}$ such that

$$\mathbf{\Lambda B}^{-1/2}[(\mathbf{HB}^{-1}\mathbf{H}']^{-1}\mathbf{B}^{-1/2}\mathbf{\Lambda}' = \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{k-q} \end{bmatrix},$$

$$\text{and } \mathbf{\Lambda B}^{-1/2}\mathbf{QB}^{-1/2}\mathbf{\Lambda}' = \begin{bmatrix} \mathbf{k}_{11} & \mathbf{k}_{12} \\ \mathbf{k}_{21} & \mathbf{k}_{22} \end{bmatrix},$$

where the matrices $\mathbf{k}_{11}$ and $\mathbf{k}_{12}$ are of order q and $k-q$ respectively. Further it can be seen that

$$\text{trace}[\mathbf{QJ}\ \mathbf{HB}^{-1}] = \text{trace}(\mathbf{k}_{11}) \text{ and } \boldsymbol{\delta}'\mathbf{J}'\mathbf{QJ}\boldsymbol{\delta} = \boldsymbol{\xi}'\mathbf{k}_{11}\boldsymbol{\xi},$$

where $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = \Lambda\mathbf{B}^{-1/2}[(\mathbf{HB}^{-1}\mathbf{H}']^{-1}\mathbf{H}\boldsymbol{\delta}$ is a $k \times 1$ vector. Hence

$$R(\tilde{\boldsymbol{\beta}}) = \text{trace}[\mathbf{QB}^{-1}] - \text{trace}(\mathbf{k}_{11}) + \boldsymbol{\xi}'\mathbf{k}_{11}\boldsymbol{\xi}. \tag{3.35}$$

Further, by Courant's theorem [Saleh (2006), Theorem 5, p.39],

$$Ch_{min}(\mathbf{k}_{11}) \leq \frac{\boldsymbol{\xi}_1'\mathbf{k}_{11}\boldsymbol{\xi}_1}{\boldsymbol{\xi}_1'\boldsymbol{\xi}_1} \leq Ch_{max}(\mathbf{k}_{11}), \tag{3.36}$$

where $Ch_{min}(\mathbf{k}_{11})$ and $Ch_{max}(\mathbf{k}_{11})$ are the minimum and maximum characteristic roots of $\mathbf{k}_{11}$ and $\Delta = \boldsymbol{\delta}'\mathbf{HB}^{-1}\mathbf{H}'\boldsymbol{\delta} = \boldsymbol{\xi}_1'\boldsymbol{\xi}_1$. Therefore,

$$R(\tilde{\boldsymbol{\beta}}) - \text{trace}(\mathbf{k}_{11}) + Ch_{min}(\mathbf{k}_{11}) \leq R(\hat{\boldsymbol{\beta}}) \leq R(\tilde{\boldsymbol{\beta}}) - \text{trace}(\mathbf{k}_{11}) + Ch_{max}(\mathbf{k}_{11}). \tag{3.37}$$

When $\Delta = 0$, the bounds of (3.37) are equal. Thus (3.37) means that for

$$\Delta \in \left[0, \frac{\text{trace}(\mathbf{k}_{11})}{Ch_{max}(\mathbf{k}_{11})}\right],$$

$\tilde{\boldsymbol{\beta}}$ has a smaller risk than $\hat{\boldsymbol{\beta}}$. Alternatively, for

$$\Delta \in \left(\frac{\text{trace}(\mathbf{k}_{11})}{Ch_{min}(\mathbf{k}_{11})}, \infty \right),$$

$\hat{\boldsymbol{\beta}}$ has smaller risk than $\tilde{\boldsymbol{\beta}}$. Clearly, when Δ moves away from H_0 beyond the value of $\frac{\text{trace}(\mathbf{k}_{11})}{Ch_{min}(\mathbf{k}_{11})}$, the ADR of $\tilde{\boldsymbol{\beta}}$ increases and become unbounded.

Risk comparison of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{PT}$:

The risk difference is given by

$$\begin{aligned} R(\hat{\boldsymbol{\beta}}) - R(\hat{\boldsymbol{\beta}}^{PT}) &= \text{trace}[\mathbf{QJHB}^{-1}]\Psi_{q+2}(q-2,\Delta) \\ &- \boldsymbol{\delta}'(\mathbf{J}'\mathbf{QJ})\boldsymbol{\delta}[2\Psi_{q+2}(q-2,\Delta) - \Psi_{q+4}(q-2,\Delta)]. \end{aligned}$$

Under H_0, $R(\hat{\boldsymbol{\beta}}) - R(\hat{\boldsymbol{\beta}}^{PT}) = \text{trace}[\mathbf{QJHB}^{-1}]\Psi_{q+2}(q-2,0)$. When Δ deviates from the null vector $\mathbf{0}$, the risk of $\hat{\boldsymbol{\beta}}^{PT}$ monotonically approaches the ADR of $\hat{\boldsymbol{\beta}}$ after achieving a maximum value. More precisely, we find that $\hat{\boldsymbol{\beta}}^{PT}$ performs better than $\hat{\boldsymbol{\beta}}$ whenever $\Delta \in [\, 0, U^0]$, where

$$U^0 = \frac{\text{trace}(\mathbf{k}_{11})\Psi_{q+2}(q-2,\Delta)}{Ch_{max}(\mathbf{k}_{11})[2\Psi_{q+2}(q-2,\Delta) - \Psi_{q+4}(q-2,\Delta)]}.$$

For $\Delta \in (\, L^0, \infty)$, where

$$L^0 = \frac{\text{trace}(\mathbf{k}_{11})\Psi_{q+2}(q-2,\Delta)}{Ch_{min}(\mathbf{k}_{11})[2\Psi_{q+2}(q-2,\Delta) - \Psi_{q+4}(q-2,\Delta)]},$$

then $\hat{\boldsymbol{\beta}}$ performs better than $\hat{\boldsymbol{\beta}}^{PT}$.

We observe that the pretest estimator which combines the unrestricted and restricted estimators to obtain a better performance of the estimators in the presence of NSI, $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$. The gain in the risk is substantial over the classical estimation procedure when restrictions (3.1) are correct.

Risk comparison of $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{PT}$:

When restrictions (3.1) are correct, then the risk difference

$$R(\hat{\boldsymbol{\beta}}^{PT}) - R(\tilde{\boldsymbol{\beta}}) = \text{trace}[\mathbf{QJHB}^{-1}]\{1 - \Psi_{q+2}(q-2,0)\} \geq 0.$$

This indicates superiority of $\tilde{\boldsymbol{\beta}}$ over $\hat{\boldsymbol{\beta}}^{PT}$ at the null hypothesis. However, under the local alternative, the risk difference is

$$\begin{aligned} R(\hat{\boldsymbol{\beta}}^{PT}) - R(\tilde{\boldsymbol{\beta}}) &= \text{trace}[\mathbf{QJHB}^{-1}][1 - \Psi_{q+2}(q-2,\Delta)] \\ &- \boldsymbol{\delta}'(\mathbf{J}'\mathbf{QJ})\boldsymbol{\delta}[1 - 2\Psi_{q+2}(q-2,\Delta) + \Psi_{q+4}(q-2,\Delta)]. \end{aligned}$$

Thus $\tilde{\boldsymbol{\beta}}$ is superior to $\hat{\boldsymbol{\beta}}^{PT}$ whenever

$$0 \leq \Delta \leq \frac{\text{trace}(\mathbf{k}_{11})[1 - \Psi_{q+2}(q-2,\Delta)]}{Ch_{max}(\mathbf{k}_{11})[1 - 2\Psi_{q+2}(q-2,\Delta) + \Psi_{q+4}(q-2,\Delta)]},$$

while the opposite holds if

$$\Delta > \frac{\text{trace}(\mathbf{k}_{11})[1 - \Psi_{q+2}(q-2,\Delta)]}{Ch_{min}(\mathbf{k}_{11})[1 - 2\Psi_{q+2}(q-2,\Delta) + \Psi_{q+4}(q-2,\Delta)]}.$$

The proposed estimators $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{PT}$ both use the sample and non-sample information, however, neither $\hat{\boldsymbol{\beta}}^{PT}$ nor $\tilde{\boldsymbol{\beta}}$ is superior with respect to each other.

In light of above discussions, we may conclude that none of the three estimators $\hat{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{PT}$ dominate the other two asymptotically. However, for $\Delta = 0$, the risks of the estimators may be ordered according to the magnitude of their risk as

$$R(\tilde{\boldsymbol{\beta}}) \leq R(\hat{\boldsymbol{\beta}}^{PT}) \leq R(\hat{\boldsymbol{\beta}}).$$

Risk comparison of $\hat{\boldsymbol{\beta}}^{S}$ and $\hat{\boldsymbol{\beta}}$:

The risk difference of $\hat{\boldsymbol{\beta}}^{S}$ and $\hat{\boldsymbol{\beta}}$ is

$$
\begin{aligned}
&= (q-2)\text{trace}[\mathbf{QJHB}^{-1}][2E(\chi^{-2}_{q+2}(\Delta)) - (q-2)E(\chi^{-4}_{q+2}(\Delta))] \\
&- (q-2)^2\boldsymbol{\delta}'(\mathbf{JQJ})\boldsymbol{\delta}\ E(\chi^{-4}_{q+4}(\Delta)) \\
&+ 2(q-2)\boldsymbol{\delta}'(\mathbf{JQJ})\boldsymbol{\delta}\left[E(\chi^{-2}_{q+4}(\Delta)) - E(\chi^{-2}_{q+2}(\Delta))\right]. \qquad (3.38)
\end{aligned}
$$

We know that

$$
\begin{aligned}
E(\chi^{-2}_{q+2}(\Delta)) - E(\chi^{-2}_{q+4}(\Delta)) &= 2E(\chi^{-4}_{q+4}(\Delta)), \qquad (3.39) \\
E(\chi^{-2}_{q+2}(\Delta)) - (q-2)E(\chi^{-4}_{q+2}(\Delta)) &= 2\Delta E(\chi^{-4}_{q+4}(\Delta)). \qquad (3.40)
\end{aligned}
$$

Using (3.39) and (3.40), (3.38) can be written as

$$
\begin{aligned}
& R(\hat{\boldsymbol{\beta}}^S) - R(\hat{\boldsymbol{\beta}}) \\
= \ & (q-2)\text{trace}(\mathbf{k}_{11})\left\{2\Delta E(\chi^{-4}_{q+4}(\Delta)) + (q-2)E(\chi^{-4}_{q+2}(\Delta))\right\} \\
- \ & (q-2)^2\text{trace}(\boldsymbol{\xi}'\mathbf{k}_{11}\boldsymbol{\xi})E(\chi^{-4}_{q+4}(\Delta)) - 4(q-2)\text{trace}(\boldsymbol{\xi}'\mathbf{k}_{11}\boldsymbol{\xi})E(\chi^{-4}_{q+4}(\Delta)) \\
= \ & (q-2)^2\text{trace}(\mathbf{k}_{11})E(\chi^{-4}_{q+2}(\Delta)) + 2(q-2)\Delta\text{trace}(\mathbf{k}_{11})E(\chi^{-4}_{q+4}(\Delta)) \\
- \ & (q^2-4)\text{trace}(\boldsymbol{\xi}'\mathbf{k}_{11}\boldsymbol{\xi})E(\chi^{-4}_{q+4}(\Delta)) \\
= \ & (q-2)^2\text{trace}(\mathbf{k}_{11})E(\chi^{-4}_{q+2}(\Delta)) \\
+ \ & \left[1 - \frac{(q+2)\text{trace}(\boldsymbol{\xi}'\mathbf{k}_{11}\boldsymbol{\xi})}{2\Delta\text{trace}(\mathbf{k}_{11})}\right] 2\Delta(q-2)\text{trace}(\mathbf{k}_{11})E(\chi^{-4}_{q+4}(\Delta)).
\end{aligned}
$$

The above risk difference is positive when

$$
\frac{\text{trace}(\mathbf{k}_{11})}{Ch_{max}(\mathbf{k}_{11})} \geq \frac{q+2}{2}, \quad \text{and} \quad q \geq 3.
$$

Under the above conditions, the ADR of $\hat{\boldsymbol{\beta}}^S$ is smaller than the ADR of $\hat{\boldsymbol{\beta}}$ in the entire parameter space and the upper limit is attained when $\Delta \to \infty$. It clearly indicates the asymptotic inferiority of $\hat{\boldsymbol{\beta}}$ under local alternatives and the largest gain in ADR is achieved near null hypothesis.

Risk comparison of $\hat{\boldsymbol{\beta}}^S$ and $\hat{\boldsymbol{\beta}}^{S+}$:

The risk difference of $\hat{\boldsymbol{\beta}}^{S}$ and $\hat{\boldsymbol{\beta}}^{S+}$ is

$$\begin{aligned}
& R(\hat{\boldsymbol{\beta}}^{S+}) - R(\hat{\boldsymbol{\beta}}^{S}) \\
- \;& \text{trace}[\mathbf{QJHB}^{-1}]E\left[\left(1-(q-2)\chi^{-2}_{q+2}(\Delta)\right)^2 I(\chi^2_{q+2}(\Delta) < q-2)\right] \\
+ \;& 2\boldsymbol{\delta}'(\mathbf{J}'\mathbf{QJ})\boldsymbol{\delta}E\left[\left(1-(q-2)\chi^{-2}_{q+2}(\Delta)\right) I(\chi^2_{q+2}(\Delta) < q-2)\right] \\
- \;& \boldsymbol{\delta}'(\mathbf{J}'\mathbf{QJ})\boldsymbol{\delta}E\left[\left(1-(q-2)\chi^{-2}_{q+4}(\Delta)\right)^2 I(\chi^2_{q+2}(\Delta) < q-2)\right] \\
= \;& -\text{trace}(\mathbf{k}_{11})E\left[\left(1-(q-2)\chi^{-2}_{q+2}(\Delta)\right)^2 I(\chi^2_{q+2}(\Delta) < q-2)\right] \\
+ \;& 2\boldsymbol{\xi}'\mathbf{k}_{11}\boldsymbol{\xi}\; E\left[\left(1-(q-2)\chi^{-2}_{q+2}(\Delta)\right) I(\chi^2_{q+2}(\Delta) < q-2)\right] \\
- \;& \boldsymbol{\xi}'\mathbf{k}_{11}\boldsymbol{\xi}\; E\left[\left(1-(q-2)\chi^{-2}_{q+4}(\Delta)\right)^2 I(\chi^2_{q+2}(\Delta) < q-2)\right].
\end{aligned}$$

The right hand side of above is just real number. Since the expectation of a positive random variable is positive, then by definition of an indicator function,

$$[q-2-\chi^2_{q+2}(\Delta)]I\left(\chi^2_{q+2}(\Delta) < q-2\right) \geq 0,$$

Since $P[\chi^2_{q+2}(\Delta) > 0] = 1$, $[(q-2)\chi^{-2}_{q+2}(\Delta) - 1]I\left(\chi^2_{q+2}(\Delta) < q-2\right) \geq 0$.

Thus, for all Δ and $q \geq 3$

$$R(\hat{\boldsymbol{\beta}}^{S+}) \leq R(\hat{\boldsymbol{\beta}}^{S})$$

with strict inequality holds for some Δ. Therefore, the risk of $\hat{\boldsymbol{\beta}}^{S+}$ is smaller than the risk of $\hat{\boldsymbol{\beta}}^{S}$ and hence smaller than the risk of $\hat{\boldsymbol{\beta}}$ in the entire parameter space and the upper limit is attained when $\Delta \to \infty$. Thus, we can order the risks of $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}^{S}$, and $\hat{\boldsymbol{\beta}}^{S+}$ as

$$R(\hat{\boldsymbol{\beta}}^{S+}) \leq R(\hat{\boldsymbol{\beta}}^{S}) \leq R(\hat{\boldsymbol{\beta}}),$$

for all values of Δ

3.7 Simulation studies

Now we return to the main problem of this chapter and provide a simulation study to investigate the performance of the proposed estimators for large sample sizes. We use Monte Carlo simulation experiments to examine the risk performance of the proposed estimators based on large sample methodology under various scenarios. Our simulation is based on a logistic regression model with different numbers of explanatory variables. This simulation study can be done for the log-linear model in case of count data.

Our sampling experiment consists of different combinations of sample sizes, i.e., $n = 100, 150, 200$. In this study we simulate a binary response from the following model:

$$ln\left(\frac{p_i}{1-p_i}\right) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}, \quad i = 1, \cdots, n,$$

where $p_i = P(Y = 1 |\ x_i)$ and the covariate matrix $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{in})$ has been drawn from a multivariate standard normal distribution.

For simulation, we consider the particular hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$, where $\boldsymbol{\beta}_2$ is a $k_2 \times 1$ vector with $k = k_1 + k_2$. We set the true value of $\boldsymbol{\beta}$ at $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{\beta}_1, \mathbf{0})$ with $\boldsymbol{\beta}_1 = (1.5, 2.5)$ to generate the binary response y_i. The summary of simulation result is provided for $(k_1, k_2) = \{(2,3), (2,5), (2,7)\}$ and $\alpha = 0.05$.

Under the null hypothesis, the number of simulations was varied initially and it was determined that 2000 of each set of observations were adequate, since a further increase in the number of replications did not significantly change the result. We define the parameter $\Delta^\star = ||\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}||^2$, where $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_1, \mathbf{0})'$ and $||\cdot||$ is the Euclidian norm. In order to investigate the behavior of the estimators for $\Delta^\star > 0$, further samples were generated from those distributions under local alternative hypotheses (i.e., for different $\Delta^\star$ between 0 and 4).

The performance of an estimator of $\boldsymbol{\beta}$ will be appraised using the mean squared error (MSE) criterion. All computations were conducted using the **R** statistical sys-

Table 3.1: Simulated relative efficiencies of RE, PT, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $n = 100, k_2 = 3$.

Δ^*	RE	PT	SE	PSE
0.0	2.154	1.318	1.160	1.194
0.2	1.947	1.296	1.154	1.175
0.4	1.890	1.113	1.116	1.133
0.6	1.743	1.026	1.098	1.106
0.8	1.759	0.945	1.075	1.083
1.0	1.650	0.950	1.070	1.070
1.2	1.570	0.962	1.062	1.062
2.0	1.065	0.999	1.049	1.049
4.0	0.940	0.999	1.019	1.019

tem (Ihaka and Gentleman, 1996). We have numerically calculated the relative MSE of $\tilde{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}^{PT}$, $\hat{\boldsymbol{\beta}}^{S}$, and $\hat{\boldsymbol{\beta}}^{S+}$ with respect to $\hat{\boldsymbol{\beta}}$ by simulation. The simulated relative efficiency (SRE) of the estimator $\boldsymbol{\beta}^{\diamond}$ to the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is defined by

$$\text{SRE}(\hat{\boldsymbol{\beta}} : \boldsymbol{\beta}^{\diamond}) = \frac{\text{MSE}(\hat{\boldsymbol{\beta}})}{\text{MSE}(\boldsymbol{\beta}^{\diamond})},$$

keeping in mind that the amount a SRE is larger than one indicates the degree of superiority of the estimator $\boldsymbol{\beta}^{\diamond}$ over $\hat{\boldsymbol{\beta}}$.

Our theoretical results were applied to various simulated data sets. Tables 3.1 to 3.9 and Figures 3.1 to 3.3 provide the estimated relative efficiency for various estimators over $\hat{\boldsymbol{\beta}}$ for $n = 100$, 150 and 200. The results can be summarized as follows:

(i) Simulation studies show that maximum efficiency of all the estimators relative to $\hat{\boldsymbol{\beta}}$ occurred at $\Delta = 0$. It is apparent from these tables that $\tilde{\boldsymbol{\beta}}$ dominates the other three estimators near the null hypothesis. On the contrary, as the hypothesis error i.e., Δ^* deviates from zero, the risk of $\tilde{\boldsymbol{\beta}}$ increases and becomes unbounded while the risk of shrinkage and positive shrinkage estimators remain below the risk of $\hat{\boldsymbol{\beta}}$ and merge with it as $\Delta^* \to \infty$. It can be safely concluded

Table 3.2: Simulated relative efficiencies of RE, PT, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $n = 150, k_2 = 3$.

$\Delta^\star$	RE	PT	SE	PSE
0.0	1.727	1.340	1.153	1.201
0.2	1.749	1.265	1.147	1.171
0.4	1.597	1.026	1.105	1.115
0.6	1.433	0.929	1.069	1.071
0.8	1.123	0.957	1.053	1.053
1.0	0.913	0.988	1.046	1.046
1.2	0.704	0.999	1.042	1.042
2.0	0.373	1.000	1.032	1.032
4.0	0.258	1.000	1.024	1.024

Table 3.3: Simulated relative efficiencies of RE, PT, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $n = 200, k_2 = 3$.

$\Delta^\star$	RE	PT	SE	PSE
0.0	1.773	1.369	1.171	1.218
0.2	1.543	1.181	1.115	1.153
0.4	1.292	0.950	1.052	1.073
0.6	1.046	0.887	1.029	1.033
0.8	0.809	0.942	1.025	1.026
1.0	0.655	0.981	1.025	1.025
1.2	0.531	0.996	1.026	1.026
2.0	0.246	1.000	1.025	1.025
4.0	0.154	1.000	1.021	1.021

Table 3.4: Simulated relative efficiencies of RE, PT, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $n = 100, k_2 = 5$.

$\Delta^\star$	RE	PT	SE	PSE
0.0	3.788	1.320	1.463	1.505
0.2	2.860	1.212	1.490	1.532
0.4	2.776	1.087	1.416	1.451
0.6	2.661	1.044	1.358	1.372
0.8	2.523	0.989	1.289	1.298
1.0	2.551	0.986	1.247	1.247
1.2	2.400	0.987	1.215	1.216
2.0	2.027	1.000	1.148	1.148
4.0	1.202	1.000	1.070	1.070

Table 3.5: Simulated relative efficiencies of RE, PT, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $n = 150, k_2 = 5$.

$\Delta^\star$	RE	PT	SE	PSE
0.0	2.441	1.268	1.483	1.577
0.2	2.379	1.192	1.494	1.548
0.4	2.299	1.037	1.343	1.400
0.6	2.007	0.980	1.264	1.269
0.8	1.558	0.988	1.205	1.206
1.0	1.296	0.996	1.173	1.173
1.2	1.014	0.999	1.148	1.148
2.0	0.550	1.000	1.103	1.103
4.0	0.380	1.000	1.073	1.073

Table 3.6: Simulated relative efficiencies of RE, PT, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $n = 200, k_2 = 5$.

Δ^*	RE	PT	SE	PSE
0.0	2.260	1.261	1.485	1.574
0.2	1.960	1.159	1.391	1.447
0.4	1.659	0.993	1.232	1.272
0.6	1.360	0.961	1.163	1.176
0.8	1.038	0.978	1.123	1.124
1.0	0.846	0.996	1.107	1.107
1.2	0.700	0.998	1.100	1.100
2.0	0.325	1.000	1.083	1.083
4.0	0.202	1.000	1.065	1.065

that the risk of $\tilde{\boldsymbol{\beta}}$ explodes as Δ^* increases, but it has less impact on shrinkage and positive shrinkage estimators, which is consistent with the theory.

(ii) Near the null hypothesis, the risk of the pretest estimator is less than the unrestricted maximum likelihood estimator which keeps increasing, crosses the risk of unrestricted maximum likelihood estimator, reaches a maximum, then decreases monotonically to the risk of unrestricted maximum likelihood estimator. Further, the SRE of this estimator is higher than that of the shrinkage and positive shrinkage estimator near the null hypothesis when $k_2 = 3$ and the opposite conclusion holds for larger values of k_2. Finally, we find that the performance of this estimator heavily depends on the correctness of the restrictions on the parameters.

(iii) If the number of variables $k_2 = 3$, and the sample sizes are between 100 and 200, the SRE of pretest, shrinkage and positive shrinkage estimators vary from 1.16 to 1.37 when the restriction holds, and they increase as the number of variables k_2 increases which is consistent with the theoretical results. For example, for $k_2 = 7$

Table 3.7: Simulated relative efficiencies of RE, PT, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $n = 100, k_2 = 7$.

$\Delta^\star$	RE	PT	SE	PSE
0.0	4.045	1.430	1.940	2.029
0.2	3.373	1.362	1.899	1.9474
0.4	3.145	1.209	1.777	1.817
0.6	2.726	1.069	1.654	1.686
0.8	2.123	0.994	1.511	1.519
1.0	1.811	0.981	1.430	1.431
1.2	1.393	0.992	1.353	1.353
2.0	0.856	1.000	1.163	1.163
4.0	0.612	1.000	1.088	1.088

Table 3.8: Simulated relative efficiencies of RE, PT, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $n = 150, k_2 = 7$.

$\Delta^\star$	RE	PT	SE	PSE
0.0	3.184	1.447	1.822	1.926
0.2	3.020	1.421	1.839	1.912
0.4	3.061	1.124	1.668	1.709
0.6	2.680	0.990	1.481	1.488
0.8	2.058	0.983	1.388	1.391
1.0	1.716	0.993	1.312	1.313
1.2	1.352	0.997	1.268	1.268
2.0	0.739	1.000	1.177	1.177
4.0	0.572	1.000	1.118	1.118

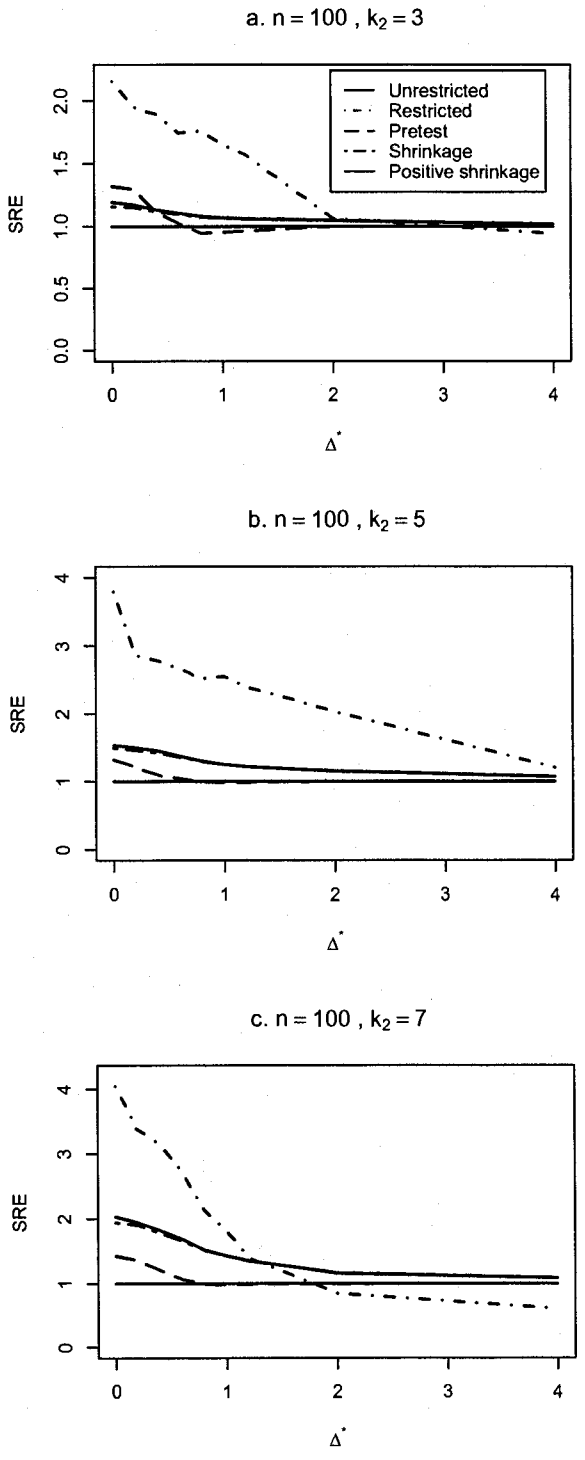

Figure 3.1: Simulated relative efficiency of the estimators as a function of non-centrality parameter Δ^* for different k_2.

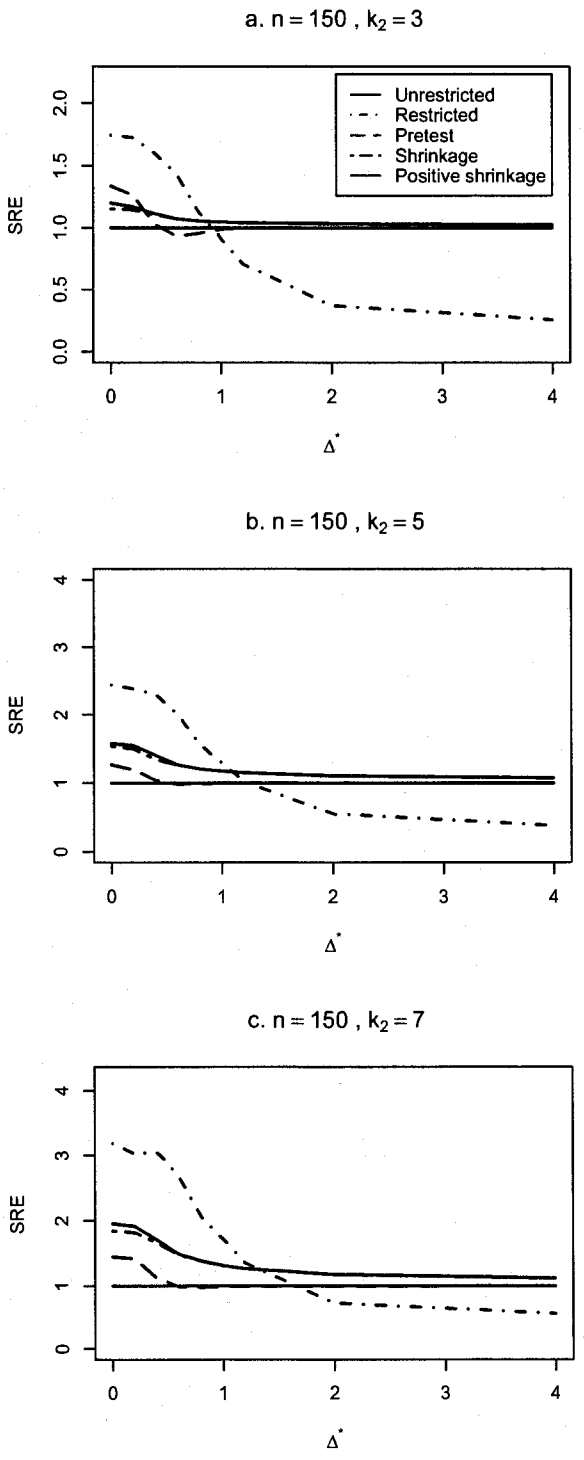

Figure 3.2: Simulated relative efficiency of the estimators as a function of non-centrality parameter Δ^* for different k_2.

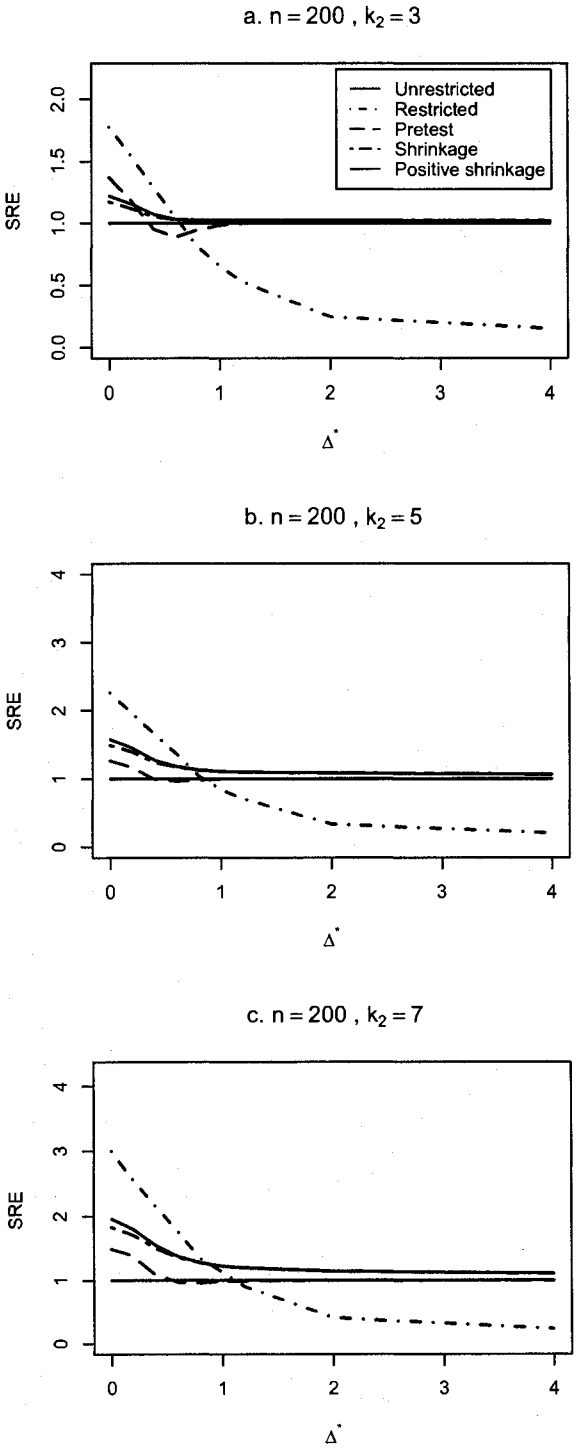

Figure 3.3: Simulated relative efficiency of the estimators as a function of non-centrality parameter Δ^* for different k_2.

Table 3.9: Simulated relative efficiencies of RE, PT, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ for $n = 200, k_2 = 7$.

Δ^*	RE	PT	SE	PSE
0.0	2.995	1.494	1.829	1.954
0.2	2.547	1.367	1.705	1.792
0.4	2.143	1.086	1.492	1.542
0.6	1.770	0.975	1.358	1.376
0.8	1.339	0.964	1.272	1.275
1.0	1.127	0.988	1.225	1.225
1.2	0.904	0.994	1.198	1.198
2.0	0.423	1.000	1.146	1.146
4.0	0.257	1.000	1.110	1.110

and $n = 150$, the SREs of these estimators are 1.44, 1.82 and 1.93 respectively, indicating the outstanding performances of the proposed estimators. On the other hand, the SRE falls sharply as Δ^* moves away from zero, and converges to one irrespective of k_1, k_2 and sample size n. Figures 3.1 to 3.3 exhibit that all the estimators dominate $\hat{\boldsymbol{\beta}}$ for small values of Δ^* and shrinkage and positive shrinkage estimators work better in case of large k_2.

3.7.1 PH versus Shrinkage and Pretest Estimator

Simulation results for PH estimator are summarized in Tables 3.10 to 3.12 for $\Delta^* = 0$. These tables shows the relative efficiencies of PH, shrinkage and pretest estimators with respect to unrestricted maximum likelihood estimator for $n = 100$, 150 and 200, when 2 out of 9 coefficients are not zero. We used the path-finding algorithm of Park and Hastie (2007) to estimate the entire solution path. We used a 10-fold cross validation procedure to choose the regularization parameter λ that achieves the lowest BIC score. For comparison, we only consider here for $\Delta^* = 0$ because the PH estimator does not take advantage of the fact that the parameter $\boldsymbol{\beta}$ lies in a subspace (3.1) and is at a disadvantage when $\Delta^* > 0$. We see from tables that when $k_2 = 3$

and 5, the performance of the PH estimator is better than the shrinkage and pretest estimation methods. On the other hand, shrinkage performs better when k_2 is large.

Table 3.10: Simulated relative efficiencies of PH, PT, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ when $\Delta^* = 0$ and $n = 100$

	$n = 100$		
Method	$k_2 = 3$	$k_2 = 5$	$k_2 = 7$
PH	1.903	2.184	1.598
PT	1.318	1.320	1.430
SE	1.160	1.463	1.940
PSE	1.194	1.505	2.029

Table 3.11: Simulated relative efficiencies of PH, PT, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ when $\Delta^* = 0$ and $n = 150$

	$n = 150$		
Method	$k_2 = 3$	$k_2 = 5$	$k_2 = 7$
PH	1.637	1.709	1.476
PT	1.340	1.268	1.447
SE	1.153	1.483	1.821
PSE	1.201	1.577	1.927

3.8 Application: South African heart disease data

The South African heart disease data was analyzed in Park and Hastie (2007) and we apply the proposed estimation strategies to this data set. This data set was collected on males in a heart disease high-risk region of western Cape, South Africa. A total of 462 individuals are included in this data set. The objective of this study is to predict CHD=1 or 0 i.e., coronary heart disease present or absent, from the set of covariates listed from below:

sbp: systolic blood pressure

Table 3.12: Simulated relative efficiencies of PH, PT, SE and PSE with respect to $\hat{\boldsymbol{\beta}}$ when $\Delta^* = 0$ and $n = 200$

	$n = 200$		
Method	$k_2 = 3$	$k_2 = 5$	$k_2 = 7$
PH	1.457	1.486	1.454
PT	1.369	1.261	1.494
SE	1.171	1.485	1.829
PSE	1.218	1.574	1.954

tobacco: cumulative tobacco (kg)

ldl: low density lipoprotein cholesterol

adiposity: Adipositylevel of fat tissue

famhist: family history of heart disease (Present, Absent)

typea: type-A behavior

obesity: Obesity level

alcohol: current alcohol intake level

age: age in years at onset disease

chd: response, coronary heart disease

Consider the full model

$$\begin{aligned} ln\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 \text{ sbp}_i + \beta_2 \text{ tobacco}_i + \beta_3 \text{ ldl}_i + \beta_4 \text{ adiposity}_i \\ &+ \beta_5 \text{ famhist}_i + \beta_6 \text{ typea}_i + \beta_7 \text{ obesity}_i + \beta_8 \text{ alcohol}_i + \beta_9 \text{ age}_i. \end{aligned}$$

Asymptotic maximum likelihood theory shows that cumulative tobacco, low density lipoprotein cholesterol, family history of heart disease, type-A behavior and age are the most important risk factors for coronary heart disease and the other four may not be risk factors for this disease. We may adopt these four factors as NSI and use the proposed estimation strategies to evaluate the effect of the other five factors on coronary heart disease.

Table 3.13: Estimate (first row), standard error (second row) and quadratic bias (third row) of tobacco (β_2), ldl (β_3), famhist (β_5), typea (β_6) and age (β_9) on coronary heart disease

Estimators	β_2	β_3	β_5	β_6	β_9	SRE
UE	0.120	0.186	0.390	0.033	0.042	1.0000
	0.059	0.141	0.482	0.024	0.024	
	0.005	0.020	0.419	0.001	0.001	
RE	0.113	0.178	0.393	0.031	0.050	1.123
	0.053	0.130	0.441	0.023	0.019	
	0.004	0.017	0.478	0.001	0.000	
PT	0.115	0.182	0.391	0.032	0.048	1.069
	0.054	0.132	0.469	0.023	0.022	
	0.004	0.018	0.496	0.001	0.000	
SE	0.117	0.183	0.390	0.033	0.045	1.039
	0.057	0.136	0.469	0.023	0.022	
	0.005	0.019	0.505	0.001	0.000	
PSE	0.117	0.184	0.392	0.032	0.045	1.044
	0.057	0.136	0.468	0.023	0.022	
	0.005	0.019	0.503	0.001	0.000	
PH	0.112	0.173	0.368	0.029	0.042	1.101
	0.057	0.132	0.433	0.022	0.022	
	0.004	0.017	0.498	0.001	0.000	

Now consider the hypothesis $H_0 : (\beta_1, \beta_4, \beta_7, \beta_8) = (0, 0, 0, 0)$. Under this hypothesis, the reduced model becomes

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_2 \text{ tobacco}_i + \beta_3 \text{ ldl}_i + \beta_5 \text{ famhist}_i + \beta_6 \text{ typea}_i + \beta_9 \text{ age}_i.$$

We draw 1000 case-resampling bootstrap samples of size $n = 150$ to evaluate the point estimates, standard errors, and relative efficiency of the proposed estimators. These results are reported in Table 3.13. It seems that the pretest, shrinkage and positive shrinkage estimators are superior to maximum likelihood estimator, which is strongly in agreement with our theoretical as well as simulation results. On the other hand, the results show that the PH estimator performs better than the shrinkage

and pretest estimator because of small number of k_2. Interestingly, the restricted estimator does better than all the estimators under the null hypothesis.

3.9 Concluding Remarks

The objective of this chapter is to compare the performance of the shrinkage and positive shrinkage estimators, a pretest estimator, a Park and Hastie estimator and the maximum likelihood estimators in the context of generalized linear models, when the parameters lie in a subspace (3.1). We examined the risk properties of the estimators in terms of ADR and Monte Carlo simulation study. It is concluded both theoretically and computationally that the risk improvement of the restricted maximum likelihood estimator is substantial at and near the null hypothesis and the improvement keeps diminishing as Δ^* moves further and further away from zero. The Park and Hastie estimator is competitive for a large number of predictors in the model with only a few of them being non-informative i.e., k_2 is small. On the other hand, the shrinkage and positive shrinkage estimators with appropriate data based weights perform well when k_2 is large. In fact, the shrinkage and positive shrinkage estimator outperforms the unrestricted maximum likelihood estimator uniformly in the entire parameter space for all k_2. In contrast, the performance of the pretest test estimator heavily depends on the quality of the NSI. The risk of the pretest estimator is smaller than that of the unrestricted maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ for small Δ^*, increases, crosses the risk of $\hat{\boldsymbol{\beta}}$, attains a maximum, then decreases monotonically to the risk of $\hat{\boldsymbol{\beta}}$ as $\Delta^* \rightarrow \infty$.

Chapter 4

Shrinkage, Pretest and Absolute Penalty Estimators in Partially Linear Models

4.1 Introduction

We consider the partially linear regression model introduced by Engle *et al.* (1986) to study the effect of weather on electricity demand, in which they assumed the relationship between temperature and electricity usage was unknown while other related factors such as income and price were parameterized linearly. A partially linear regression model is defined as

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + g(t_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{4.1}$$

where y_i's are responses, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ and $t_i \in [0, 1]$ are design points, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is an unknown parameter vector, $g(\cdot)$ is an unknown real-valued function defined on $[0, 1]$, the ε_i's are unobservable random errors.

We consider experiments where the vector of coefficients $\boldsymbol{\beta}$ in the linear part

of (4.1) can be partitioned as $(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ where $\boldsymbol{\beta}_1$ is the coefficient vector for main effects (e.g. treatment effects, genetic effects) and $\boldsymbol{\beta}_2$ is a vector for "nuisance" effects (e.g. age, laboratory). In this situation, inference about $\boldsymbol{\beta}_1$ may benefit from moving the least squares estimate for the full model in the direction of the least squares estimate without the nuisance variables (Steinian shrinkage) or from dropping the nuisance variables if there is evidence that they do not provide useful information (pretesting). In this framework, the Stein-type or shrinkage estimator combines estimation problems by shrinking a base estimator to a plausible alternative estimator. Our shrinkage estimate for the partially linear model is of the form

$$\hat{\boldsymbol{\beta}}_1^{\star} = \pi\tilde{\boldsymbol{\beta}}_1 + (1-\pi)\hat{\boldsymbol{\beta}}_1, \quad \pi \in (0,1).$$

where $\tilde{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_1$ are the semiparametric estimators of $\boldsymbol{\beta}_1$ for the model with and without the $\boldsymbol{\beta}_2$ components, respectively, and π is a shrinkage factor that shrinks the full model estimates $\hat{\boldsymbol{\beta}}_1$ towards the restricted model estimates $\tilde{\boldsymbol{\beta}}_1$. Bickel (1984) showed that, in parametric models, such estimates are asymptotically optimal in a minimax sense and conjectured the same result for semiparametric models. When π is an indicator function, $\hat{\boldsymbol{\beta}}_1^{\star}$ is a pretest estimator. For a particular data-based $\pi \in (0,1)$, we show in our semiparametric model that $\hat{\boldsymbol{\beta}}_1^{\star}$ asymptotically improves on the unrestricted least squares estimates $\hat{\boldsymbol{\beta}}_1$ and on the restricted model estimates $\tilde{\boldsymbol{\beta}}_1$. Burman and Chaudhuri (1992) considered procedures that shrink a nonparametric estimate $\hat{\mu}(\mathbf{x})$ of $\mu(\mathbf{x})$, in the model $Y = \mu(\mathbf{x}) + \varepsilon$, in the direction of a parametric estimate $g(\hat{\boldsymbol{\beta}}, \mathbf{x})$ of $\mu(\mathbf{x})$ and gave conditions under which this Steinian estimate asymptotically improves on $\hat{\mu}(\mathbf{x})$ and $g(\hat{\boldsymbol{\beta}}, \mathbf{x})$.

The rest of this chapter is organized as following. The statistical model and estimators are discussed in Section 4.2. The proposed pretest estimator, shrinkage and positive shrinkage estimators, LASSO and absolute penalty estimator are presented in Section 4.3. Some necessary assumptions and asymptotic properties of the proposed estimators are investigated in Section 4.4. The asymptotic bias and risk performance

of the proposed estimators are presented in Section 4.5, and results of a simulation study that includes a comparison with a semiparametric extension of the LASSO are given in Section 4.6. Section 4.7 presents a conclusion and some discussion. Throughout this chapter, the boldface symbols represent vectors/matrices.

4.2 Statistical Model and Estimators

Throughout this chapter we will assume that $\mathbf{1}_n = (1, \ldots, 1)'$ is not in the space spanned by the column vectors of $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$. Consequently, according to Chen (1988), under regularity conditions on g, model (4.1) is identifiable. In addition, we assume the design points $\mathbf{x}_i$ and t_i are fixed for $i = 1, \cdots, n$, and we introduce a restriction on the parameters in model (4.1),

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + g(t_i) + \varepsilon_i \quad \text{subject to} \quad \mathbf{R}\boldsymbol{\beta} = \mathbf{r}, \tag{4.2}$$

where $\mathbf{R}$ is an $s \times p$ restriction matrix, and $\mathbf{r}$ is an $s \times 1$ vector of constants.

In many applications, $\mathbf{r} = \mathbf{0}$, that is, some of the coefficients are set to zero, effectively removing the corresponding terms from the model. We let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, where $\mathbf{X}_1$ is an $n \times p_1$ submatrix containing the regressors of interest and $\mathbf{X}_2$ is an $n \times p_2$ submatrix that may or may not be relevant in the analysis of the main regressors. Accordingly, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ be the vector of parameters, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ have dimensions p_1 and p_2 respectively with $p_1 + p_2 = p$, $p_i \geq 0$ for $i = 1, 2$. We are essentially interested in the estimation of $\boldsymbol{\beta}_1$ when it is suspected that $\boldsymbol{\beta}_2$ is close to $\mathbf{0}$. Thus, we consider the restriction $\mathbf{R}\boldsymbol{\beta} = \mathbf{0}$ with $\mathbf{R} = [\mathbf{0}, \mathbf{I}]$, where $\mathbf{0}$ is a $p_2 \times p_1$ matrix of zeroes and $\mathbf{I}$ is the identity matrix. Our relevant hypothesis is

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}.$$

Let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^{'}, \hat{\boldsymbol{\beta}}_2^{'})$ be a semiparametric least squares estimator of $\boldsymbol{\beta}$ under model

(4.1) as defined previously. Then we call $\hat{\boldsymbol{\beta}}_1$ the unrestricted semiparametric least squares estimator of $\boldsymbol{\beta}_1$. If $\boldsymbol{\beta}_2 = \mathbf{0}$, then we have the restricted linear regression model

$$y_i = x_{i1}\beta_1^{(o)} + \cdots + x_{ip_1}\beta_{p_1}^{(o)} + g^{(o)}(t_i) + \varepsilon_i^{(o)},\ i = 1, \ldots, n, \tag{4.3}$$

We let $\tilde{\boldsymbol{\beta}}_1$ denotes the restricted semiparametric least squares estimator of $\boldsymbol{\beta}_1$ as defined previously. Generally speaking, $\tilde{\boldsymbol{\beta}}_1$ performs better than $\hat{\boldsymbol{\beta}}_1$ when $\boldsymbol{\beta}_2$ is close to $\mathbf{0}$. But for $\boldsymbol{\beta}_2$ away from the origin $\mathbf{0}$, $\tilde{\boldsymbol{\beta}}_1$ may be considerably biased, inefficient, and even possibly inconsistent. The estimate $\hat{\boldsymbol{\beta}}_1$, however, is consistent for a departure of $\boldsymbol{\beta}_2$ from $\mathbf{0}$. Thus, we have two extreme estimators $\hat{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_1$ suited best for the partially linear regression models (4.1) and (4.3), respectively. We attempt to strike a compromise between $\hat{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_1$ so that the compromise behaves reasonably well relative to $\hat{\boldsymbol{\beta}}_1$ as well as $\tilde{\boldsymbol{\beta}}_1$. We consider three estimators for the target $\boldsymbol{\beta}_1$ of the parametric component in the setting of the semiparametric regression model in (4.1). The first estimator is the pretest test semiparametric least squares estimator, denote by $\hat{\boldsymbol{\beta}}_1^{PT}$. This estimator is a combination of $\hat{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_1$ via the indicator function $I(T_n < T_{n,\alpha})$ where T_n is an appropriate test-function to test the null hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ vs $H_a : \boldsymbol{\beta}_2 \neq \mathbf{0}$. Further, $T_{n,\alpha}$ gives an α-level critical value using the distribution of T_n. The pretest estimator chooses $\hat{\boldsymbol{\beta}}_1$ or $\tilde{\boldsymbol{\beta}}_1$ according as H_0 is accepted or rejected. However, it is important to remember that our primary objective is to find an efficient estimator of $\boldsymbol{\beta}_1$. Thus deciding against H_a does not necessarily imply that $\boldsymbol{\beta}_2 = \mathbf{0}$, because we have no control of the probability of the type I error. Instead, we think we may get a better estimator of $\boldsymbol{\beta}_1$ by setting $\boldsymbol{\beta}_2 = \mathbf{0}$. Thus $T_{n,\alpha}$ is a threshold that determines a hard thresholding rule, and α is a tuning parameter.

The other two estimators are the James-Stein and positive James-Stein estimators, known as the shrinkage and positive shrinkage estimators respectively. The shrinkage estimator $\hat{\boldsymbol{\beta}}_1^S$ is a smooth function of the test statistic T_n.

In this chapter we confine ourselves to the partial kernel smoothing estimator of $\boldsymbol{\beta}$, which attains the usual parametric convergence rate $n^{-1/2}$ without undersmoothing

the nonparametric component $g(\cdot)$, (Speckman (1988)). Assume that $\{\mathbf{x}_i', t_i, y_i; i = 1, \cdots, n\}$ satisfy model (4.1). If $\boldsymbol{\beta}$ is known to be the true parameter, then by $E\varepsilon_i = 0$ we have $g(t_i) = E(y_i - \mathbf{x}_i'\boldsymbol{\beta})$ for $i = 1, \cdots, n$. Hence, a natural nonparametric estimator of $g(\cdot)$ given $\boldsymbol{\beta}$ is

$$\tilde{g}(t, \boldsymbol{\beta}) = \sum_{i=1}^{n} W_{ni}(t)(y_i - \mathbf{x}_i'\boldsymbol{\beta}),$$

with the weight functions $W_{ni}(\cdot)$ defined in Assumption 3 of Section 4.4. To estimate $\boldsymbol{\beta}$, we use

$$\hat{\boldsymbol{\beta}} = \arg\min\ SS(\boldsymbol{\beta}) = (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\widehat{\mathbf{Y}}, \tag{4.4}$$

with

$$SS(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i - \mathbf{x}_i'\boldsymbol{\beta} - \tilde{g}(t_i, \boldsymbol{\beta})]^2 = \sum_{i=1}^{n} (\hat{y}_i - \hat{\mathbf{x}}_i'\boldsymbol{\beta})^2,$$

where $\widehat{\mathbf{Y}} = (\hat{y}_1, \cdots, \hat{y}_n)'$, $\widehat{\mathbf{X}} = (\hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_n)'$, $\hat{y}_i = y_i - \sum_{j=1}^{n} W_{nj}(t_i)y_j$ and $\hat{\mathbf{x}}_i = \mathbf{x}_i - \sum_{j=1}^{n} W_{nj}(t_i)\mathbf{x}_j$ for $i = 1, \cdots, n$. The unrestricted estimator $\hat{\boldsymbol{\beta}}_1$ of $\boldsymbol{\beta}_1$ is

$$\hat{\boldsymbol{\beta}}_1 = (\widehat{\mathbf{X}}_1'\mathbf{Q}_{\widehat{\mathbf{X}}_2}\widehat{\mathbf{X}}_1)^{-1}\widehat{\mathbf{X}}_1'\mathbf{Q}_{\widehat{\mathbf{X}}_2}\widehat{\mathbf{Y}},$$

where $\widehat{\mathbf{X}}_1$ is composed of the first p_1 row vectors of $\widehat{\mathbf{X}}$, $\widehat{\mathbf{X}}_2$ is composed of the last p_2 row vectors of $\widehat{\mathbf{X}}$ and $\mathbf{Q}_{\widehat{\mathbf{X}}_2} = \mathbf{I} - \widehat{\mathbf{X}}_2(\widehat{\mathbf{X}}_2'\widehat{\mathbf{X}}_2)^{-1}\widehat{\mathbf{X}}_2'$. Similar to the construction of $\hat{\boldsymbol{\beta}}$, for model (4.2), the restricted estimator $\tilde{\boldsymbol{\beta}}_1$ of $\boldsymbol{\beta}_1$ has the form

$$\tilde{\boldsymbol{\beta}}_1 = (\widehat{\mathbf{X}}_1'\widehat{\mathbf{X}}_1)^{-1}\widehat{\mathbf{X}}_1'\widehat{\mathbf{Y}}.$$

4.3 Estimation Strategies

4.3.1 The Pretest Estimator

Bancroft (1944) introduced the pretest test estimation procedure as one basis for

dealing with model-estimator uncertainty [see Ahmed *et al.* (2006*a*)]. Let

$$T_n = \frac{n}{\hat{\sigma}_n^2}\hat{\boldsymbol{\beta}}_2'\widehat{\mathbf{X}}_2'\mathbf{Q}_{\widehat{\mathbf{X}}_1}\widehat{\mathbf{X}}_2\hat{\boldsymbol{\beta}}_2,$$

where

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}} - \hat{g}_n(t_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \hat{\mathbf{x}}_i'\hat{\boldsymbol{\beta}})^2,$$

with $\hat{g}_n(\cdot) = \sum_{i=1}^{n} W_{ni}(\cdot)(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})$ and $\mathbf{Q}_{\widehat{\mathbf{X}}_1} = \mathbf{I} - \widehat{\mathbf{X}}_1(\widehat{\mathbf{X}}_1'\widehat{\mathbf{X}}_1)^{-1}\widehat{\mathbf{X}}_1'$. We shall later see that the statistic T_n converges to a chi-square distribution with p_2 degrees of freedom for large n. Thus, we can choose an α-level critical value $\chi^2_{p_2,\alpha}$ and define $\hat{\boldsymbol{\beta}}_1^{PT}$ as follows:

$$\hat{\boldsymbol{\beta}}_1^{PT} = \hat{\boldsymbol{\beta}}_1 - (\hat{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_1)I(T_n \leq \chi^2_{p_2,\alpha}), \quad p_2 \geq 1.$$

Thus, $\hat{\boldsymbol{\beta}}_1^{PT}$ chooses $\tilde{\boldsymbol{\beta}}_1$ when H_0 is tenable, otherwise $\hat{\boldsymbol{\beta}}_1$ is chosen. Obviously, the dispersion of $\hat{\boldsymbol{\beta}}_1^{PT}$ is now more controlled depending on the size α of the test, but the pretest test rule makes extreme choices of either $\tilde{\boldsymbol{\beta}}_1$ or $\hat{\boldsymbol{\beta}}_1$. It is well documented in the literature that pretest test procedures are not admissible for many models, even though they may improve on unrestricted procedures. Thus, we consider another basis for resolving model-estimator uncertainty. Stein (1956) demonstrated the inadmissibility of the maximum likelihood estimator when estimating a multivariate mean vector under quadratic loss. Making use of Stein-type estimators, Sclove *et al.* (1972) demonstrated the nonoptimality of the pretest test estimator in certain multi-parametric situations.

4.3.2 The Shrinkage and Positive Shrinkage Estimators

The shrinkage semiparametric estimator $\hat{\boldsymbol{\beta}}_1^{S}$ is defined by

$$\hat{\boldsymbol{\beta}}_1^{S} = \hat{\boldsymbol{\beta}}_1 - (\hat{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_1)(p_2 - 2)T_n^{-1}, \quad p_2 \geq 3.$$

The estimator $\hat{\boldsymbol{\beta}}_1^S$ is in the general form of the Stein-rule family of estimators, where shrinkage of the base estimator is towards the alternative simpler estimator $\tilde{\boldsymbol{\beta}}_1$. The estimator is pulled toward the alternative estimator when the variance of the least squares estimator is large, and pulled toward the general least squares estimator when the alternative estimator has high variance, high bias, or is more highly correlated with the least squares estimator. It should be noted that $\hat{\boldsymbol{\beta}}_1^S$ is the smooth version of $\hat{\boldsymbol{\beta}}_1^{PT}$. Extending the language of Donoho and Johnstone (1998), $\hat{\boldsymbol{\beta}}_1^{PT}$ and $\hat{\boldsymbol{\beta}}_1^S$ are based on hard and smooth thresholds, respectively. However, $\hat{\boldsymbol{\beta}}_1^S$ is not a convex combination $\tilde{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_1$. We also consider the so called positive-rule shrinkage semiparametric estimator $\hat{\boldsymbol{\beta}}_1^{S+}$:

$$\hat{\boldsymbol{\beta}}_1^{S+} = \tilde{\boldsymbol{\beta}}_1 + (\hat{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_1)\left(1 - \frac{p_2 - 2}{T_n}\right)^+, \quad p_2 \geq 3,$$

For fixed models (not depending on n), $\hat{\boldsymbol{\beta}}_1^S$ adapts to the magnitude of T_n and tends to $\hat{\boldsymbol{\beta}}_1$ as T_n tends to infinity and to $\tilde{\boldsymbol{\beta}}_1$ as $T_n \to p_2 - 2$. Similar conclusions hold for $\hat{\boldsymbol{\beta}}_1^{PT}$. In the next section we will consider the intermediate case where T_n tends in probability to a constant greater than $p_2 - 2$. That is, we will consider local Pitman contiguous models where $\boldsymbol{\beta}_2$ depends on n and tends to the zero vector as $n \to \infty$. Such sequences of models have been considered in the estimation context by Bickel (1984) and Claeskens and Hjort (2003), among others.

Remark: Note that the Steinian strategy is similar in spirit to the model-averaging procedures, Bayesian or otherwise, see Bickel (1984), Hoeting *et al.* (1999), Hoeting *et al.* (2002) and Burnham and Anderson (2002).

4.3.3 LASSO and Absolute Penalty Estimator

The LASSO originally proposed for linear regression models, has become a popular model selection procedure. It is a constrained version of ordinary least squares defined

as the solution to

$$\hat{\boldsymbol{\beta}}_\tau = \min_{\boldsymbol{\beta}} (\hat{\mathbf{y}} - \hat{\mathbf{x}}'\boldsymbol{\beta})'(\hat{\mathbf{y}} - \hat{\mathbf{x}}'\boldsymbol{\beta}) \quad \text{subject to} \quad \sum_{j=1}^{n} |\beta_j| \leq \tau,$$

where τ is a tuning parameter. If τ is large enough, this just gives the usual least squares estimates. However, smaller values of τ produce shrunken estimates $\hat{\boldsymbol{\beta}}$, often with many components equal to zero. This procedure gives shrinkage, variable deletion and good prediction accuracy as well as effectively balancing variance and bias. Traditionally, the LASSO is computed by quadratic programming techniques, and τ is selected using cross-validation (CV) and generalized cross-validation (GCV). Note that the output of the LASSO resembles shrinkage and pretest methods by both shrinking and deleting coefficients. However, it is different from the pretest and shrinkage procedures of Section 3.1 in that it treats all the covariate coefficients equally. The LASSO does not single out the nuisance covariates for special scrutiny as to their usefulness in estimating main effect coefficients.

The LASSO was first introduced for linear models. We propose the absolute penalty type estimator (APE) for partially linear models, which is an extension of the LASSO method for linear models. This estimator can be obtained by applying the LASSO method to the residuals $(\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i)$, $i = 1, 2, \cdots, n$, defined in Section 4.2.

4.4 First-order Asymptotic Results

The following assumptions are required to derive the main results. These assumptions are quite general and can be easily satisfied (see Remarks 4.1–4.3 following the assumptions).

Assumption 1. There exist bounded functions $h_s(\cdot)$ over $[0, 1]$, $s = 1, \cdots, p$, such that

$$x_{is} = h_s(t_i) + u_{is}, \ i = 1, \cdots, n, s = 1, \cdots, p, \tag{4.5}$$

where $\mathbf{u}_i = (u_{i1}, \cdots, u_{ip})'$ are real vectors satisfying

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} u_{ik}u_{ij}}{n} = b_{kj}, \quad \text{for } k = 1, \cdots, p, \ j = 1, \cdots, p, \tag{4.6}$$

and the matrix $\mathbf{B} = (b_{kj})$ is nonsingular. Moreover, for any permutation $(j_1, \cdots, j_n)$ of $(1, \cdots, n)$, as $n \to \infty$,

$$\left\| \max_{1\le j\le n} \sum_{i=1}^{n} W_{ni}(t_j)\mathbf{u}_i \right\| = o(n^{-\frac{1}{6}}), \tag{4.7}$$

where $||\cdot||$ denotes the Euclidean norm and $W_{ni}(\cdot)$ satisfies Assumption 3.

Assumption 2. The functions $g(\cdot)$ and $h_s(\cdot)$ satisfy the Lipschitz condition of order 1 on $[0, 1]$ for $s = 1, \cdots, p$.

Assumption 3. The probability weight functions $W_{ni}(\cdot)$ satisfy

(i) $\max_{1\le i\le n} \sum_{j=1}^{n} W_{ni}(t_j) = O(1)$,

(ii) $\max_{1\le i,j\le n} W_{ni}(t_j) = O(n^{-2/3})$,

(iii) $\max_{1\le j\le n} \sum_{i=1}^{n} W_{ni}(t_j) I(|t_i - t_j| > c_n) = O(d_n)$,

where I is the indicator function, c_n satisfies $\limsup_{n\to\infty} nc_n^3 < \infty$, and d_n satisfies $\limsup_{n\to\infty} nd_n^3 < \infty$.

Remark 4.1. The above u_{ij} behave like zero mean, uncorrelated random variables and $h_s(t_i)$ is the regression of x_{is} on t_i. Specifically, suppose that the design points $(\mathbf{x}_i, t_i)$ are i.i.d. random variables, and let $h_s(t_i) = E(x_{is}|t_i)$ and $u_{is} = x_{is} - h_s(t_i)$ with $E[\mathbf{u}_i\mathbf{u}_i'] = \mathbf{B}$. Then by the law of large numbers, (4.6) holds with probability 1 and (4.7) holds by Lemma 1 in Shi and Lau (2000). Assumptions (4.5) and (4.6) have been used in Gao (1995a), Gao (1995b), Gao (1997), Liang and Härdle (1999), among others, and (4.7) in Shi and Lau (2000).

Remark 4.2. Assumption 2 is very mild. The usual polynomial and trigonometric functions satisfy this assumption.

Remark 4.3. Under regular conditions, the Nadaraya-Watson kernel weights, Priestley and Chao kernel weights, locally linear weights and Gasser-Müller kernel weights

satisfy Assumption 3. For example, if we take the p.d.f. of $U[-1,1]$ as the kernel function, namely,

$$K(t) = I_{[-1,1]}(t)/2,$$

with $t_i = i/n$, and the bandwidth equals to $cn^{-1/3}$, where c is a constant, then the Priestley and Chao kernel weights, which satisfies Assumption 3, are

$$W_{ni}(t) = \frac{1}{2cn^{\frac{2}{3}}} I_{\left(|t-\frac{i}{n}|\leq cn^{-\frac{1}{3}}\right)}(t).$$

Lemma 1. (i) Suppose that Assumptions 2 and 3 (iii) hold. Then as $n \to \infty$,

$$\max_{0\leq s\leq p} \max_{1\leq i\leq n} \left| G_s(t_i) - \sum_{j=1}^{n} W_{nj}(t_i) G_s(t_j) \right| = O(c_n) + O(d_n),$$

where $G_0(\cdot) = g(\cdot)$ and $G_s(\cdot) = h_s(\cdot)$, $s = 1, \ldots, p$.

(ii) Suppose that Assumptions 1 to 3 hold. Then as $n \to \infty$,

$$\max_{1\leq s\leq p} \max_{1\leq i\leq n} \left| \hat{h}_{ns}(t_i) - h_s(t_i) \right| = O(c_n) + O(d_n) + o(n^{-\frac{1}{6}})$$

where $\hat{h}_{ns}(t_i) = \sum_{j=1}^{n} W_{nj}(t_i) x_{js}$.

Lemma 2. For any sequence of independent variables $\{V_k, k = 1, \ldots, n\}$ with mean zero and finite $(2+\delta)$th moment, and for a set of positive numbers $\{a_{ki}, k, i = 1, \ldots, n\}$ such that $\max_{1\leq i,k\leq n} |a_{ki}| \leq n^{-s_1}$ for some $0 \leq s_1 \leq 1$ and $\sum_{j=k}^{n} a_{ki} = O(n^{s_2})$ for some $s_2 \geq \max\{0, 2/(2+\delta) - s_1\}$,

$$\max_{1\leq i\leq n} \left| \sum_{k=1}^{n} a_{ki} V_k \right| = O\left(n^{-\frac{s_1-s_2}{2}} \ln n\right) \text{ a.s..}$$

The proofs of Lemmas 1 and 2 can be found in Gao (1995a) and Härdle *et al.* (2000) respectively.

Lemma 3. Suppose that Assumptions 1 to 3 hold, and the ε_i are independent with

mean zero, variance σ^2 and $\mu_{3i} = E\varepsilon_i^3$ being uniformly bounded. Then we have that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{D} N(0, \sigma^2 \mathbf{B}^{-1}) \quad \text{and} \quad \max_{1\le i\le n} |\hat{g}_n(t_i) - g(t_i)| = O_p\left(n^{-\frac{1}{3}} \ln n\right),$$

where $\hat{g}_n(t) = \sum_{i=1}^n W_{ni}(t)(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_n)$ and $\mathbf{B}$ is defined in Assumption 1.

Proof. The proof of the asymptotic normality of $\hat{\boldsymbol{\beta}}_n$ is similar to that of Theorem 1 (i) of Gao (1995a). We omit the details. According to the definition of $\hat{g}_n(t_i)$, we have

$$\max_{1\le i\le n} |\hat{g}_n(t_i) - g(t_i)| \le \max_{1\le i\le n} \left|\sum_{j=1}^n W_{nj}(t_i)\mathbf{x}_j'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)\right|$$
$$+ \max_{1\le i\le n} \left|\sum_{j=1}^n W_{nj}(t_i)g(t_j) - g(t_i)\right| + \max_{1\le i\le n} \left|\sum_{j=1}^n W_{nj}(t_i)\varepsilon_j\right| \triangleq I_1 + I_2 + I_3.$$

By Assumption 1(a), I_1 can be decomposed as

$$I_1 \le \max_{1\le i\le n} \left|\sum_{s=1}^p \sum_{j=1}^n W_{nj}(t_i)u_{js}(\beta_s - \hat{\beta}_{ns})\right|$$
$$+ \max_{1\le i\le n} \left|\sum_{s=1}^p \sum_{j=1}^n W_{nj}(t_i)\left[h_s(t_j) - \hat{h}_{ns}(t_j)\right](\beta_s - \hat{\beta}_{ns})\right| \triangleq I_{11} + I_{12},$$

where $\hat{\beta}_{ns}$ and β_s are the sth components of $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}$ respectively. It is easy to see that

$$I_{11} \le p \max_{1\le s\le p} \left|\beta_s - \hat{\beta}_{ns}\right| \cdot \left[\max_{1\le i\le n} \left|\left|\sum_{j=1}^n W_{nj}(t_i)\mathbf{u}_j\right|\right|\right] = o_p\left(n^{-\frac{2}{3}}\right),$$

by Assumption 1 (c), Assumption 3 (i) and the root-n consistency of $\hat{\boldsymbol{\beta}}$. Similarly, using Lemma 1 (ii), $I_{12} = o(n^{-1/6-1/2})$. Moreover, by Lemmas 1 (i) and 2, we have

$$I_2 = O\left(n^{-\frac{1}{3}}\right), \text{ and } I_3 = O_p\left(n^{-\frac{1}{3}} \ln n\right).$$

Lemma 4. Suppose that Assumptions 1 to 3 hold. Then $n^{-1}\widehat{\mathbf{X}}'\widehat{\mathbf{X}} = \mathbf{B} + O(n^{-2/3})$.

The proof of Lemma 4 can be found in Gao (1995a).

Lemma 5. Suppose that Assumptions 1 to 3 hold. Then

$$\hat{\sigma}_n^2 = \sigma^2 + O_p\left(n^{-\frac{1}{2}}\right),\ \tilde{\boldsymbol{\beta}}_1 = (\mathbf{I}, \mathbf{B}_{11}^{-1}\mathbf{B}_{12})\hat{\boldsymbol{\beta}}_n + o_p\left(n^{-\frac{1}{2}}\right), \text{ and } T_n = n\sigma^2\hat{\boldsymbol{\beta}}_2'\mathbf{B}_{22.1}\hat{\boldsymbol{\beta}}_2 + o_p(1),$$

where $\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}$ with $\mathbf{B}$ defined in Assumption 1 and $\mathbf{B}_{22.1} = \mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}$.

Moreover, we have

$$\lim_{n\to\infty} P\{T_n \le x | K_n\} = \Psi_{p_2}(x; \Delta) \quad \text{where} \quad \Delta = (\boldsymbol{\omega}'\mathbf{B}_{22.1}\boldsymbol{\omega})\sigma^{-2}.$$

Proof. According to the definition of $\hat{\sigma}_n^2$, we have

$$\begin{aligned}
\hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 + \frac{1}{n}\sum_{i=1}^n \left[\mathbf{x}_i'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)\right]^2 + \frac{1}{n}\sum_{i=1}^n (g(t_i) - \hat{g}_n(t_i))^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^n \varepsilon_i\mathbf{x}_i'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n) + \frac{2}{n}\sum_{i=1}^n \varepsilon_i(g(t_i) - \hat{g}_n(t_i)) + \frac{2}{n}\sum_{i=1}^n \mathbf{x}_i'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)(g(t_i) - \hat{g}_n(t_i)) \\
&\triangleq I_1 + \ldots + I_6.
\end{aligned}$$

It is easy to see $I_1 = \sigma^2 + O_p(n^{-1/2})$. Based on Lemmas 1, 3, 4 and Assumptions 1 to 3, we can show that $I_i = o_p(n^{-1/2})$ for $i = 2, 3, 4$ and 6. In addition, I_4 can be decomposed as

$$I_4 = \frac{2}{n}\sum_{i=1}^n\sum_{j=1}^n W_{nj}(t_i)\varepsilon_i\varepsilon_j + \frac{2}{n}\sum_{i=1}^n \varepsilon_i\mathbf{x}_i'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n) + \frac{2}{n}\sum_{i=1}^n \varepsilon_i(g(t_i) - \hat{g}_n(t_i)).$$

By Lemma 2 of Gao (1995b),

$$\sum_{i=1}^n\sum_{j=1}^n W_{nj}(t_i)\varepsilon_i\varepsilon_j = o_p\left(n^{\frac{1}{2}}\right).$$

This implies $I_4 = o_p(n^{-1/2})$. Therefore, $\hat{\sigma}_n^2 = \sigma^2 + O_p(n^{-1/2})$. Moreover, by combining Lemmas 3 and 4, it is easy to prove the other results. We omit the details.

Proof

Using Lemma 5, we conclude that $\sqrt{n}(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)$ and T_n are asymptotically independent under K_n. This implies that

$$\lim_{n\to\infty} P\left\{\sqrt{n}(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \le \mathbf{x}, T_n \le \chi^2_{p_2,\alpha} | K_n\right\} = \Phi_{p_1}(\mathbf{x}+\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}; 0, \sigma^2\mathbf{B}_{11}^{-1})\Psi_{p_2}(\chi^2_{p_2,\alpha}; \Delta)$$

By combining Lemmas 3 and 5

$$\begin{aligned} &\lim_{n\to\infty} P\left\{\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \le \mathbf{x}, T_n \ge \chi^2_{p_2,\alpha} | K_n\right\} \\ &= \int_{E(\boldsymbol{\omega})} \Phi_{p_1}(\mathbf{x} - \mathbf{D}_{12}^{-1}\mathbf{D}_{22}\mathbf{z}; 0, \sigma^2\mathbf{D}_{11.2}^{-1}) d\Phi_{p_2}(\mathbf{z}; 0, \sigma^2\mathbf{D}_{22}), \end{aligned}$$

where

$$\mathbf{D} = \mathbf{B}^{-1} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{pmatrix} \quad \text{and} \quad E(\boldsymbol{\omega}) = \{\mathbf{z} : \sigma^{-2}(\mathbf{z}+\boldsymbol{\omega})'\mathbf{B}_{22.1}(\mathbf{z}+\boldsymbol{\omega}) \ge \chi^2_{p_2,\alpha}\}.$$

So the asymptotic cumulative distribution function of $\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{PT} - \boldsymbol{\beta}_1)$ under $\{K_n\}$ is

$$\begin{aligned} F_{p_1}(\mathbf{x}) &= \Phi_{p_1}(\mathbf{x} + \mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}; \mathbf{0}, \sigma^2\mathbf{B}_{11}^{-1})\Psi_{p_2}(\chi^2_{p_2,\alpha}; \Delta) \\ &\quad + \int_{E(\boldsymbol{\omega})} \Phi_{p_1}(\mathbf{x} - \mathbf{D}_{12}\mathbf{D}_{22}\mathbf{z}; \mathbf{0}, \sigma^2\mathbf{D}_{11.2}) d\Phi_{p_2}(\mathbf{z}; \mathbf{0}, \sigma^2\mathbf{D}_{22}) \end{aligned}$$

and under $\{K_n\}$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1^S - \boldsymbol{\beta}_1) \xrightarrow{L} \mathbf{D}_1\mathbf{U} + (p_2 - 2)\frac{\sigma^2\mathbf{B}_{11}^{-1}\mathbf{B}_{12}(\mathbf{D}_1\mathbf{U} + \boldsymbol{\omega})}{(\mathbf{D}_2\mathbf{U} + \boldsymbol{\omega})'\mathbf{B}_{22.1}(\mathbf{D}_2\mathbf{U} + \boldsymbol{\omega})}$$

as n tends to infinity, where $\mathbf{U} \sim N_p(0, \sigma^2\mathbf{B}^{-1})$. Now, using Lemma 3, the proofs of Theorems 4.1 through 4.3 follow from direct computation and the definitions of the

estimators.

4.5 Asymptotic Bias and Risk Performance

In this section, we derive expressions for asymptotic quadratic biases and quadratic risks of the estimators considered in Section 4.3. Our main concern is the performance of the five estimators when $\boldsymbol{\beta}_2$ is close to the null vector, and we consider a sequence of local alternatives $\{K_n\}$ given by

$$K_n : \boldsymbol{\beta}_{2(n)} = n^{-\frac{1}{2}}\boldsymbol{\omega},\ \boldsymbol{\omega} \neq \mathbf{0} \text{ fixed.} \tag{4.8}$$

The objective is to estimate the unknown parameter vectors by some estimator $\boldsymbol{\delta}$ when performance is evaluated by squared error loss. To study the asymptotic quadratic risks of $\hat{\boldsymbol{\beta}}_1$, $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_1^{PT}$, $\hat{\boldsymbol{\beta}}_1^{S}$ and $\hat{\boldsymbol{\beta}}_1^{S+}$, we define a quadratic loss function using a positive definite matrix (p.d.m.) $\mathbf{Q}$, by,

$$\mathcal{L}(\boldsymbol{\delta}, \boldsymbol{\beta}_1) = n(\boldsymbol{\delta} - \boldsymbol{\beta}_1)'\mathbf{Q}(\boldsymbol{\delta} - \boldsymbol{\beta}_1),$$

where $\boldsymbol{\delta}$ can be any one of $\hat{\boldsymbol{\beta}}_1$, $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_1^{PT}$, $\hat{\boldsymbol{\beta}}_1^{S}$ and $\hat{\boldsymbol{\beta}}_1^{S+}$. Now we assume that, for the estimator $\boldsymbol{\delta}$ of $\boldsymbol{\beta}_1$, the asymptotic distribution function of $\boldsymbol{\delta}$ under $\{K_n\}$ exists and is given by

$$F(\mathbf{x}) = \lim_{n\to\infty} P\{\sqrt{n}(\boldsymbol{\delta} - \boldsymbol{\beta}_1) \leq \mathbf{x} | K_n\},$$

where $F(\mathbf{x})$ is nondegenerate. Then the asymptotic distributional risk (ADR) of $\boldsymbol{\delta}$ is defined as

$$R(\boldsymbol{\delta}, \mathbf{Q}) = tr\left(\mathbf{Q}\int_{\mathbb{R}^{p_1}} \mathbf{x}\mathbf{x}^\top dF(\mathbf{x})\right) = \text{tr}(\mathbf{QV}),$$

where $\mathbf{V}$ is the dispersion matrix for the distribution $F(\mathbf{x})$.

Note that, under nonlocal (fixed) alternatives, all the estimators are asymptotically equivalent to $\hat{\boldsymbol{\beta}}_1$, while $\tilde{\boldsymbol{\beta}}_1$ has unbounded risk. To obtain the non-degenerate

asymptotic distribution F, we consider the local Pitman alternatives (4.8).

First, we present the expression for the asymptotic distributional bias (ADB) of the proposed estimators. The ADB of an estimator $\boldsymbol{\delta}$ is defined as

$$\text{ADB}(\boldsymbol{\delta}) = \lim_{n\to\infty} E\left\{n^{\frac{1}{2}}(\boldsymbol{\delta} - \boldsymbol{\beta}_1)\right\}.$$

For the next theorem, we assume that $\Psi_v(x;\Delta)$ is the cumulative distribution function of the noncentral chi-square distribution with noncentrality parameter Δ with v degrees of freedom. Further,

$$E(\chi_v^{-2j}(\Delta)) = \int_0^\infty x^{-2j} d\Psi_v(x;\Delta).$$

Theorem 4.5.1. *Suppose that Assumptions 1 to 3 defined in Section 4.4 hold and that the errors ε_i are independent with mean zero, the same variance σ^2 and that $\mu_{3i} = E\varepsilon_i^3$ is uniformly bounded. Then, under $\{K_n\}$, the ADBs of the estimators $\hat{\boldsymbol{\beta}}_1$, $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_1^{PT}$, $\hat{\boldsymbol{\beta}}_1^{S}$ and $\hat{\boldsymbol{\beta}}_1^{S+}$ are respectively*

$$\begin{aligned}
ADB(\hat{\boldsymbol{\beta}}_1) &= \mathbf{0}\ , \\
ADB(\tilde{\boldsymbol{\beta}}_1) &= -\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}, \\
ADB(\hat{\boldsymbol{\beta}}_1^{PT}) &= -\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}\Psi_{(p_2+2)}(\chi^2_{p_2,\alpha},\Delta), \\
ADB(\hat{\boldsymbol{\beta}}_1^{S}) &= -(p_2-2)\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}E(\chi^{-2}_{p_2+2}(\Delta)), \\
ADB(\hat{\boldsymbol{\beta}}^{S+}) &= -(p_2-2)\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}\left[\Psi_{(p_2+2)}(p_2-2,\Delta) + E(\chi^{-2}_{p_2+2}(\Delta))\right. \\
&- \left. E(\chi^{-2}_{p_2+2}(\Delta))I(\chi^2_{p_2+2}(\Delta) < (p_2-2))\right].
\end{aligned}$$

Proof: See section 4.4.

Since the bias expressions of the estimators are all not in scalar form, we convert them to quadratic forms. Thus, we define the asymptotic quadratic distributional

bias (AQDB) of an estimator $\boldsymbol{\delta}$ of $\boldsymbol{\beta}_1$ by

$$\text{AQDB}(\boldsymbol{\delta}) = [\text{ADB}(\boldsymbol{\delta})]'\mathbf{B}_{11.2}[\text{ADB}(\boldsymbol{\delta})],$$

where $\mathbf{B}_{11.2} = \mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21}$.

Corollary 4.1. Suppose that the assumptions of Theorem 4.5.1 hold. Then, under $\{K_n\}$, the AQDB of the estimators are

$$\begin{aligned}
AQDB(\hat{\boldsymbol{\beta}}_1) &= 0, \\
AQDB(\tilde{\boldsymbol{\beta}}_1) &= \gamma, \\
AQDB(\hat{\boldsymbol{\beta}}_1^{PT}) &= \gamma[\Psi_{(p_2+2)}(\chi^2_{p_2,\alpha};\Delta)]^2, \\
AQDB(\hat{\boldsymbol{\beta}}_1^{S}) &= (p_2-2)^2\gamma\left[E(\chi^{-2}_{p_2+2}(\Delta))\right]^2, \\
AQDB(\hat{\boldsymbol{\beta}}^{S+}) &= \gamma\left[\Psi_{(p_2+2)}(p_2-2,\Delta) + E(\chi^{-2}_{p_2+2}(\Delta))\right. \\
&- \left. E(\chi^{-2}_{p_2+2}(\Delta))I(\chi^2_{p_2+2}(\Delta) < (p_2-2))\right]^2,
\end{aligned}$$

where $\gamma = \boldsymbol{\omega}'\boldsymbol{\Upsilon}\boldsymbol{\omega}$ with $\boldsymbol{\Upsilon} = \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{11.2}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}$.

Remark 4.1. The above results establish the following results.

(i) The AQDB of $\tilde{\boldsymbol{\beta}}_1$ is an unbounded function of γ.

(ii) The bias of $\hat{\boldsymbol{\beta}}_1^{PT}$ is a function of γ and α. For fixed α, as a function of γ, the bias starts from 0, increases to a point, then decreases gradually to zero. On the other hand, as a function of α it is a decreasing function of $\alpha \in [0,1]$, achieves a maximum value at $\alpha = 0$ and is 0 at $\alpha = 1$.

(iii) In order to investigate AQDB($\hat{\boldsymbol{\beta}}_1^{S}$) and AQDB($\hat{\boldsymbol{\beta}}_1^{S+}$), we use the following result. By using a result from matrix algebra

$$Ch_{min}(\sigma^2\boldsymbol{\Upsilon}\mathbf{B}_{22.1}^{-1}) \leq \frac{\sigma^2\gamma}{\boldsymbol{\omega}'\mathbf{B}_{22.1}\boldsymbol{\omega}} \leq Ch_{max}(\sigma^2\boldsymbol{\Upsilon}\mathbf{B}_{22.1}^{-1}).$$

Therefore AQDB of $\hat{\boldsymbol{\beta}}_1^{S}$ starts from 0 at $\gamma = 0$, and increases to a point then

decreases towards 0 due to $E\left(\chi^{-2}_{p_2+2}(\Delta)\right)$ being a decreasing log convex function of Δ. The behavior of $\hat{\boldsymbol{\beta}}_1^{S+}$ is similar to $\hat{\boldsymbol{\beta}}_1^{S}$, however, the quadratic bias curve of $\hat{\boldsymbol{\beta}}_1^{S+}$ remains below the curve of $\hat{\boldsymbol{\beta}}_1^{S}$ for all values of Δ.

The asymptotic dispersion matrices of the estimators are given in the following theorem.

Theorem 4.5.2. *Suppose the assumptions of Theorem 4.1 hold. Then, under* $\{K_n\}$, *the asymptotic covariance matrices of the estimators* $\hat{\boldsymbol{\beta}}_1$, $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_1^{PT}$, $\hat{\boldsymbol{\beta}}_1^{S}$ *and* $\hat{\boldsymbol{\beta}}_1^{S+}$ *are*

$$
\begin{aligned}
\boldsymbol{\Gamma}_1(\hat{\boldsymbol{\beta}}_1) &= \sigma^2\mathbf{B}_{11.2}^{-1}, \\
\boldsymbol{\Gamma}_2(\tilde{\boldsymbol{\beta}}_1) &= \sigma^2\mathbf{B}_{11}^{-1} + \mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}\boldsymbol{\omega}'\mathbf{B}_{21}\mathbf{B}_{11}^{-1}, \\
\boldsymbol{\Gamma}_3(\hat{\boldsymbol{\beta}}_1^{PT}) &= \sigma^2\left[\mathbf{B}_{11.2}^{-1}\{1-\Psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)\} + \mathbf{B}_{11}^{-1}\Psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)\right] \\
&+ \mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}\boldsymbol{\omega}'\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\{2\Psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta) - \Psi_{p_2+4}(\chi^2_{p_2,\alpha};\Delta)\}, \\
\boldsymbol{\Gamma}_4(\hat{\boldsymbol{\beta}}_1^{S}) &= \sigma^2\mathbf{B}_{11.2}^{-1} - (p_2-2)\sigma^2\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\mathbf{B}_{22.1}^{-1}\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\left\{2E(\chi^{-2}_{p_2+2}(\Delta))\right. \\
&- \left.(p_2-2)E(\chi^{-4}_{p_2+2}(\Delta))\right\} + (p_2^2-4)\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}\boldsymbol{\omega}'\mathbf{B}_{21}\mathbf{B}_{11}^{-1}E(\chi^{-4}_{p_2+4}(\Delta)), \\
\boldsymbol{\Gamma}_5(\hat{\boldsymbol{\beta}}^{S+}) &= \boldsymbol{\Gamma}_4(\hat{\boldsymbol{\beta}}_1^{S}) + (p_2-2)\boldsymbol{\Upsilon}\left[2E(\chi^{-2}_{p_2+2}(\Delta)I(\chi^2_{p_2+2}(\Delta)\le(p_2-2))\right. \\
&- \left.(p_2-2)E(\chi^{-4}_{p_2+2}(\Delta))I(\chi^2_{p_2+2}(\Delta)\le(p_2-2))\right] \\
&- \mathbf{B}_{11}^{-1}\mathbf{B}_{12}\mathbf{B}_{22.1}^{-1}\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\Psi_{p_2+2}((p_2-2);\Delta) \\
&+ \mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}\boldsymbol{\omega}'\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\left[2\Psi_{p_2+2}((p_2-2);\Delta) - \Psi_{p_2+4}((p_2-2);\Delta)\right] \\
&- (p_2-2)\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}\boldsymbol{\omega}'\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\left[2E(\chi^{-2}_{p_2+2}(\Delta))I(\chi^2_{p_2+2}(\Delta)\le(p_2-2))\right. \\
&- 2E(\chi^{-2}_{p_2+4}(\Delta))I(\chi^2_{p_2+4}(\Delta)\le(p_2-2)) \\
&+ \left.(p_2-2)E(\chi^{-4}_{p_2+4}(\Delta))I(\chi^2_{p_2+4}(\Delta)\le(p_2-2))\right].
\end{aligned}
$$

Proof: See section 4.4.

The asymptotic distributional risk (ADR) expressions for the estimators are contained in the following theorem.

Theorem 4.5.3. *Suppose the assumptions of theorem 4.1 hold. Then under* $\{K_n\}$, *the ADRs of* $\hat{\boldsymbol{\beta}}_1$, $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_1^{PT}$, $\hat{\boldsymbol{\beta}}_1^{S}$ *and* $\hat{\boldsymbol{\beta}}_1^{S+}$ *are respectively,*

$$
\begin{aligned}
R(\hat{\boldsymbol{\beta}}_1) &= \sigma^2 tr(\mathbf{Q}\mathbf{B}_{11.2}^{-1}), \\
R(\tilde{\boldsymbol{\beta}}_1) &= \sigma^2 tr(\mathbf{Q}\mathbf{B}_{11}^{-1}) + \boldsymbol{\omega}'\boldsymbol{\Theta}\boldsymbol{\omega}, \\
R(\hat{\boldsymbol{\beta}}_1^{PT}) &= \sigma^2 \left[tr(\mathbf{Q}\mathbf{B}_{11.2}^{-1})\{1 - \Psi_{p_2+2}(\chi^2_{p_2,\alpha}; \Delta)\} + tr(\mathbf{Q}\mathbf{B}_{11}^{-1})\Psi_{p_2+2}(\chi^2_{p_2,\alpha}; \Delta) \right] \\
&+ \boldsymbol{\omega}'\boldsymbol{\Theta}\boldsymbol{\omega} \left[2\Psi_{p_2+2}(\chi^2_{p_2,\alpha}; \Delta) - \Psi_{p_2+4}(\chi^2_{p_2,\alpha}; \Delta) \right], \\
R(\hat{\boldsymbol{\beta}}_1^{S}) &= \sigma^2 \left[tr(\mathbf{Q}\mathbf{B}_{11.2}^{-1}) - (p_2 - 2) tr(\boldsymbol{\Theta}\mathbf{B}_{22.1}^{-1})(2E(\chi^{-2}_{p_2+2}(\Delta)) \right. \\
&\left. - (p_2 - 2)E(\chi^{-4}_{p_2+2}(\Delta)) \right] + (p_2^2 - 4)\boldsymbol{\omega}'\boldsymbol{\Theta}\boldsymbol{\omega} E(\chi^{-4}_{p_2+4}(\Delta)), \\
R(\hat{\boldsymbol{\beta}}^{S+}) &= \mathbf{R}_4(\hat{\boldsymbol{\beta}}_1^{S}) + (p_2 - 2) tr(\boldsymbol{\Theta}\mathbf{B}_{22.1}^{-1}) \left[2E(\chi^{-2}_{p_2+2}(\Delta) I(\chi^2_{p_2+2}(\Delta) \leq (p_2 - 2)) \right. \\
&- \left. (p_2 - 2)E(\chi^{-4}_{p_2+2}(\Delta)) I(\chi^2_{p_2+2}(\Delta) \leq (p_2 - 2)) \right] \\
&- tr(\boldsymbol{\Theta}\mathbf{B}_{22.1}^{-1})\Psi_{p_2+2}((p_2 - 2); \Delta) \\
&+ \boldsymbol{\omega}'\boldsymbol{\Theta}\boldsymbol{\omega} \left[2\Psi_{p_2+2}((p_2 - 2); \Delta) - \Psi_{p_2+4}((p_2 - 2); \Delta) \right] \\
&- (p_2 - 2)\boldsymbol{\omega}'\boldsymbol{\Theta}\boldsymbol{\omega} \left[2E(\chi^{-2}_{p_2+2}(\Delta)) I(\chi^2_{p_2+2}(\Delta) \leq (p_2 - 2)) \right. \\
&- 2E(\chi^{-2}_{p_2+4}(\Delta)) I(\chi^2_{p_2+4}(\Delta) \leq (p_2 - 2)) \\
&+ \left. (p_2 - 2)E(\chi^{-4}_{p_2+4}(\Delta)) I(\chi^2_{p_2+4}(\Delta) \leq (p_2 - 2)) \right],
\end{aligned}
$$

where $\boldsymbol{\Theta} = \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{Q}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}$.

Proof: See section 4.4.

4.5.1 Comparison of risks among the estimators

Using the following identity

$$
E(\chi^{-2}_{p_2+2}(\Delta)) - (p_2 - 2)E(\chi^{-4}_{p_2+2}(\Delta)) = \Delta E(\chi^{-4}_{p_2+4}(\Delta)),
$$

we see that $R(\hat{\boldsymbol{\beta}}_1^{S})$ satisfies

$$
\begin{aligned}
R(\hat{\boldsymbol{\beta}}_1^{S}) &= \sigma^2 \mathrm{tr}(\mathbf{Q}\mathbf{B}_{11.2}^{-1}) - (p_2 - 2)\sigma^2 \mathrm{tr}(\boldsymbol{\Theta}\mathbf{B}_{22.1}^{-1}) \\
&\left\{ (p_2 - 2)E(\chi^{-4}_{p_2+2}(\Delta)) + \left[1 - \frac{(p_2 + 2)\sigma^{-2}\boldsymbol{\omega}'\boldsymbol{\Theta}\boldsymbol{\omega}}{2\Delta \mathrm{tr}(\boldsymbol{\Theta}\mathbf{B}_{22.1}^{-1})} \right] 2E(\chi^{-4}_{p_2+4}(\Delta)) \right\} \\
&\leq R(\hat{\boldsymbol{\beta}}_1), \text{ for } p_2 \geq 3, \text{ all } \Delta > 0,
\end{aligned}
$$

for all $\mathbf{Q}$ with

$$\frac{\text{tr}(\boldsymbol{\Theta}\mathbf{B}_{22.1}^{-1})}{Ch_{max}(\boldsymbol{\Theta}\mathbf{B}_{22.1}^{-1})} \geq \frac{p_2+2}{2},$$

where $Ch_{max}(.)$ is the maximum characteristic root.

In Theorem 4.5.3, we may discard the case $\mathbf{B}_{12} = \mathbf{0}$, since in this situation $\boldsymbol{\Theta} = \mathbf{0}$ and $\mathbf{B}_{11.2} = \mathbf{B}_{11}$. Then from Theorem 4.5.3, the ADRs of all estimators are reduced to the ADR of $\hat{\boldsymbol{\beta}}_1$. In the sequel we assume that $\mathbf{B}_{12} \neq \mathbf{0}$.

Based on Theorem 4.3, the results for the estimation problem are:

(i) For any $\mathbf{Q} \in \mathbf{Q}^D$ and $\boldsymbol{\omega}$, $\text{R}(\hat{\boldsymbol{\beta}}_1^{S+}) \leq \text{R}(\hat{\boldsymbol{\beta}}_1^{S}) \leq \text{R}(\hat{\boldsymbol{\beta}}_1)$ under $\{K_n\}$ where

$$\mathbf{Q}^D = \left\{ \mathbf{Q} : \frac{\text{tr}(\boldsymbol{\Theta}\mathbf{B}_{22.1}^{-1})}{Ch_{max}(\boldsymbol{\Theta}\mathbf{B}_{22.1}^{-1})} \geq \frac{p_2+2}{2} \right\},$$

and $\hat{\boldsymbol{\beta}}_1^{S+}$ not only confirms inadmissibility of $\hat{\boldsymbol{\beta}}_1^{S}$ but also provides a simple superior estimator.

(ii) When $\boldsymbol{\omega} = \mathbf{0}$, the following holds:

$$\text{R}(\tilde{\boldsymbol{\beta}}_1) \leq \text{R}(\hat{\boldsymbol{\beta}}_1^{PT}) \leq \text{R}(\hat{\boldsymbol{\beta}}_1^{S+}) \leq \text{R}(\hat{\boldsymbol{\beta}}_1^{S}) \leq \text{R}(\hat{\boldsymbol{\beta}}_1).$$

(iii) As $\boldsymbol{\omega}$ moves away from zero, $\text{R}(\tilde{\boldsymbol{\beta}}_1)$ monotonically increases in $\lambda \equiv \boldsymbol{\omega}'\boldsymbol{\Theta}\boldsymbol{\omega}$ and goes to infinity as λ goes to infinity. The ADR of $\hat{\boldsymbol{\beta}}_1$ remains constant while $\text{R}(\hat{\boldsymbol{\beta}}_1^{PT})$ increases, crossing the line $\text{R}(\hat{\boldsymbol{\beta}}_1)$ as $\boldsymbol{\omega}$ moves away from zero. Moreover, when λ tends to infinity, the risks of $\hat{\boldsymbol{\beta}}_1^{PT}$, $\hat{\boldsymbol{\beta}}_1^{S}$ and $\hat{\boldsymbol{\beta}}^{S+}$ approach a common limit i.e., the risk of $\hat{\boldsymbol{\beta}}_1$. Thus, $\hat{\boldsymbol{\beta}}_1^{PT}$, $\hat{\boldsymbol{\beta}}_1^{S}$ and $\hat{\boldsymbol{\beta}}_1^{S+}$ have bounded ADRs, unlike the restricted estimator.

Finally, it is important to remark here that the absolute penalty estimator for our criteria with $Q \in \mathbf{Q}^D$ outperforms the conventional semiparametric least squares estimator in the entire parameter space for $p_2 \geq 3$, while this least squares estimator is admissible for $p_2 = 1$ and $p_2 = 2$.

For practical reasons and to illustrate the properties of the theoretical results, we conducted a simulation study, reported in the next section, to compare the performance of the proposed estimators for moderate and large sample sizes.

4.6 Simulation Studies

In this section, we use Monte Carlo simulation experiments to examine the quadratic risk (namely MSE) performance of the proposed estimators. Our simulation is based on a nonlinear regression model with different numbers of explanatory variables.

Our sampling experiment consists of different combinations of sample sizes, i.e., $n = 30, 50, 80$ and 100. In this study we simulate the response from the following model:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{pi}\beta_p + g(t_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the ε_i are i.i.d standard normal, $t_i = (i - 0.5)/n$, $x_{si} = (\zeta_{si}^{(1)})^2 + \zeta_i^{(1)}$ with $\zeta_{si}^{(1)}$ i.i.d. $\sim N(0,1)$ and $\zeta_i^{(2)}$ i.i.d. $\sim N(0,1)$ for all $s = 1, \ldots, p$, and $i = 1, \cdots, n$.

We consider the hypothesis $H_0 : \beta_j = 0$, for $j = p_1 + 1, \cdots, p$ with $p = p_1 + p_2$. We set the regression coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{\beta}_1, \mathbf{0})$ with $\boldsymbol{\beta}_1 = (1.5, 3, 2)$, and the nonlinear function $g(t) = \sin(4\pi t)$ to generate response y_i. Those are fixed for each realization. We provide detailed results for $(p_1, p_2) = \{(3,3), (3,5), (3,11)\}$ and $\alpha = 0.05$.

For the weight function $W_{ni}(t_j)$, we use

$$W_{ni}(t_j) = \frac{1}{nh_n} K\left(\frac{t_i - t_j}{h_n}\right) = \frac{1}{nh_n}\frac{1}{\sqrt{2\pi}} e^{-\frac{(t_i - t_j)^2}{2h_n^2}},$$

which is Priestley and Chao's weight with a Gaussian kernel. We use the cross-validation (CV) method (Bowman and Azzalini, 1997) to select the optimal band-

width h_n, which minimizes the following CV function

$$CV(h_n) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}^{-i} - \hat{x}_1^{-i}\hat{\beta}_{1n}^{-i} - \hat{x}_2^{-i}\hat{\beta}_{2n}^{-i} - \hat{x}_3^{-i}\hat{\beta}_{3n}^{-i} - \hat{x}_4^{-i}\hat{\beta}_{4n}^{-i} - \ldots - \hat{x}_p^{-i}\hat{\beta}_{pn}^{-i})^2,$$

where $(\hat{\beta}_{1n}^{-i}, \hat{\beta}_{2n}^{-i}, \hat{\beta}_{3n}^{-i}, \hat{\beta}_{4n}^{-i})' = (\hat{\boldsymbol{X}}'^{-i}\hat{\boldsymbol{X}}^{-i})^{-1}\hat{\boldsymbol{X}}'^{-i}\hat{\boldsymbol{y}}^{-i}$, $\hat{\boldsymbol{X}}^{-i} = \left(\hat{x}_{jk}^{-i}\right)'$, $1 \leq k \leq n$, $1 \leq j \leq p$, $\hat{\boldsymbol{y}}^{-i} = (\hat{y}_1^{-i}, \ldots, \hat{y}_n^{-i})$, $\hat{x}_{sk}^{-i} = x_{sk} - \sum_{j\neq i}^{n} W_{nj}(t_i)x_{sj}$, $\hat{y}_k^{-i} = y_k - \sum_{j\neq i}^{n} W_{nj}(t_i)y_j$. Here $\hat{\mathbf{y}}^{-i}$ is the predicted value of $\boldsymbol{y}= (y_1, y_2, \cdots y_n)$ at $\mathbf{x}_i = (x_{1i}, x_{2i}, \cdots, x_{pi})$ with y_i and $\mathbf{x}_i$ left out of the estimation of the β's.

The number of simulations under the null hypothesis was varied initially and it was determined that 5000 of each set of observations were adequate, since a further increase in the number of realizations did not significantly change the result. We define the parameter $\Delta^\star = ||\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}||^2$, where $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_1, \mathbf{0})'$ and $||\cdot||$ is the Euclidian norm. In order to investigate the behavior of the estimators for $\Delta^\star > 0$, further samples were generated from those distributions under local alternative hypotheses (i.e., for different $\Delta^\star$ between 0 and 4).

Table 4.1: Simulated relative efficiency with respect to $\hat{\boldsymbol{\beta}}_1$ for $n = 30, p_2 = 3$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_1^{PT}$	$\hat{\boldsymbol{\beta}}_1^{S}$	$\hat{\boldsymbol{\beta}}_1^{S+}$
0.0	2.057	1.723	1.042	1.103
0.2	1.600	1.219	1.036	1.079
0.4	0.952	0.889	1.023	1.056
0.6	0.564	0.910	1.020	1.020
0.8	0.365	0.919	1.011	1.011
1.2	0.187	1.000	1.007	1.007
1.6	0.109	1.000	1.003	1.003
2.0	0.072	1.000	1.001	1.001
4.0	0.018	1.000	1.000	1.000

The performance of an estimator of $\boldsymbol{\beta}_1$ will be based on the mean squared error (MSE) criterion. All computations were conducted using the R statistical system (Ihaka and Gentleman (1996)). We have numerically calculated the relative MSE of

Table 4.2: Simulated relative efficiency with respect to $\hat{\boldsymbol{\beta}}_1$ for $n = 30, p_2 = 5$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_1^{PT}$	$\hat{\boldsymbol{\beta}}_1^{S}$	$\hat{\boldsymbol{\beta}}_1^{S+}$
0.0	2.871	2.401	1.134	1.999
0.2	2.141	1.601	1.049	1.643
0.4	1.215	0.861	0.889	1.218
0.6	0.668	0.669	0.899	1.055
0.8	0.422	0.776	0.964	0.994
1.2	0.196	0.779	0.994	0.994
1.6	0.114	0.829	0.985	0.985
2.0	0.075	0.974	0.993	0.993

Table 4.3: Simulated relative efficiency with respect to $\hat{\boldsymbol{\beta}}_1$ for $n = 30, p_2 = 11$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_1^{PT}$	$\hat{\boldsymbol{\beta}}_1^{S}$	$\hat{\boldsymbol{\beta}}_1^{S+}$
0.0	11.782	4.899	3.051	5.422
0.2	8.557	3.342	2.098	4.152
0.4	4.422	1.518	1.468	2.490
0.6	2.464	1.076	1.279	1.757
0.8	1.519	1.023	1.288	1.439
1.2	0.721	0.969	1.193	1.195
1.6	0.419	0.987	1.111	1.111
2.0	0.266	1.000	1.071	1.071
4.0	0.069	1.000	1.018	0.019

$\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_1^{PT}$, $\hat{\boldsymbol{\beta}}_1^{S}$, and $\hat{\boldsymbol{\beta}}_1^{S+}$, with respect to $\hat{\boldsymbol{\beta}}_1$. The simulated relative efficiency (SRE) of the estimator $\hat{\boldsymbol{\beta}}_1^{\diamond}$ to the unrestricted least squares estimator $\hat{\boldsymbol{\beta}}_1$ is defined by

$$\text{SRE}(\hat{\boldsymbol{\beta}}_1 : \hat{\boldsymbol{\beta}}_1^{\diamond}) = \frac{\text{MSE}(\hat{\boldsymbol{\beta}}_1)}{\text{MSE}(\hat{\boldsymbol{\beta}}_1^{\diamond})},$$

keeping in mind that the amount by which a SRE is larger than one indicates the degree of superiority of the estimator $\hat{\boldsymbol{\beta}}_1^{\diamond}$ over $\hat{\boldsymbol{\beta}}_1$.

Our methods were applied to several simulated data sets. We report the results in Tables 4.1-4.12 and Figures 4.1 and 4.2. The findings can be summarized as follows:

Table 4.4: Simulated relative efficiency with respect to $\hat{\beta}_1$ for $n = 50, p_2 = 3$.

$\Delta^\star$	$\tilde{\beta}_1$	$\hat{\beta}_1^{PT}$	$\hat{\beta}_1^{S}$	$\hat{\beta}_1^{S+}$
0.0	2.548	1.942	1.226	1.257
0.2	1.242	0.906	1.066	1.108
0.4	0.527	0.664	1.004	1.005
0.6	0.271	0.944	1.001	1.001
0.8	0.159	1.000	1.000	1.000
1.2	0.074	1.000	0.997	0.998
1.6	0.040	1.000	0.999	0.999
2.0	0.026	1.000	1.000	1.000
4.0	0.007	1.000	1.000	1.000

(i) The restricted estimator outperforms all the other estimators when the restriction is at and near $\Delta^\star = 0$. On the contrary, when $\Delta^\star$ is larger than zero, the estimated SREs of $\tilde{\beta}_1$ increases and becomes unbounded whereas the estimated SRE of all other estimators remain bounded and approach one. It can be safely concluded that the departure from the restriction is fatal to $\tilde{\beta}_1$, but it has a much smaller impact on the absolute penalty estimator. This is consistent with the asymptotic theory.

(ii) The pretest test estimator works well near the null hypothesis, but the simulation shows that the performance heavily depends on how close β_2 is to zero, and is less efficient than the unrestricted least squares estimator $\hat{\beta}_1$, for large values of $\Delta^\star$.

(iii) When $\Delta^\star$ increases, the risk of the shrinkage and positive shrinkage estimators with respect to unrestricted least squares estimator decreases and converges to 1 irrespective of p_1, p_2 and n. Figures 1 and 2 show that the shrinkage and positive shrinkage estimators work better in cases with large p_2.

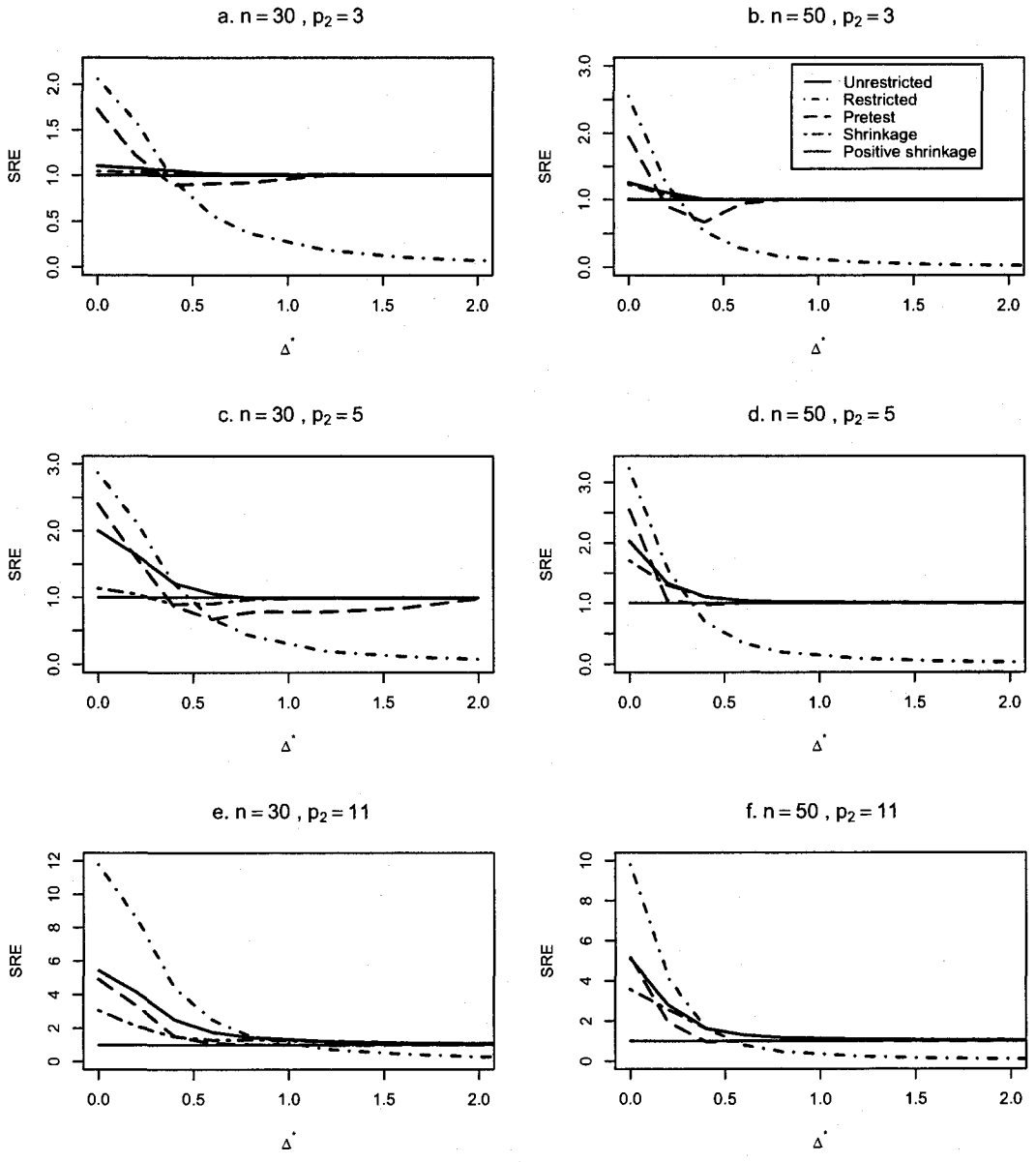

Figure 4.1: Simulated relative efficiency of the estimators as a function of non-centrality parameter Δ^* for different sample sizes n, and nuisance parameters p_2.

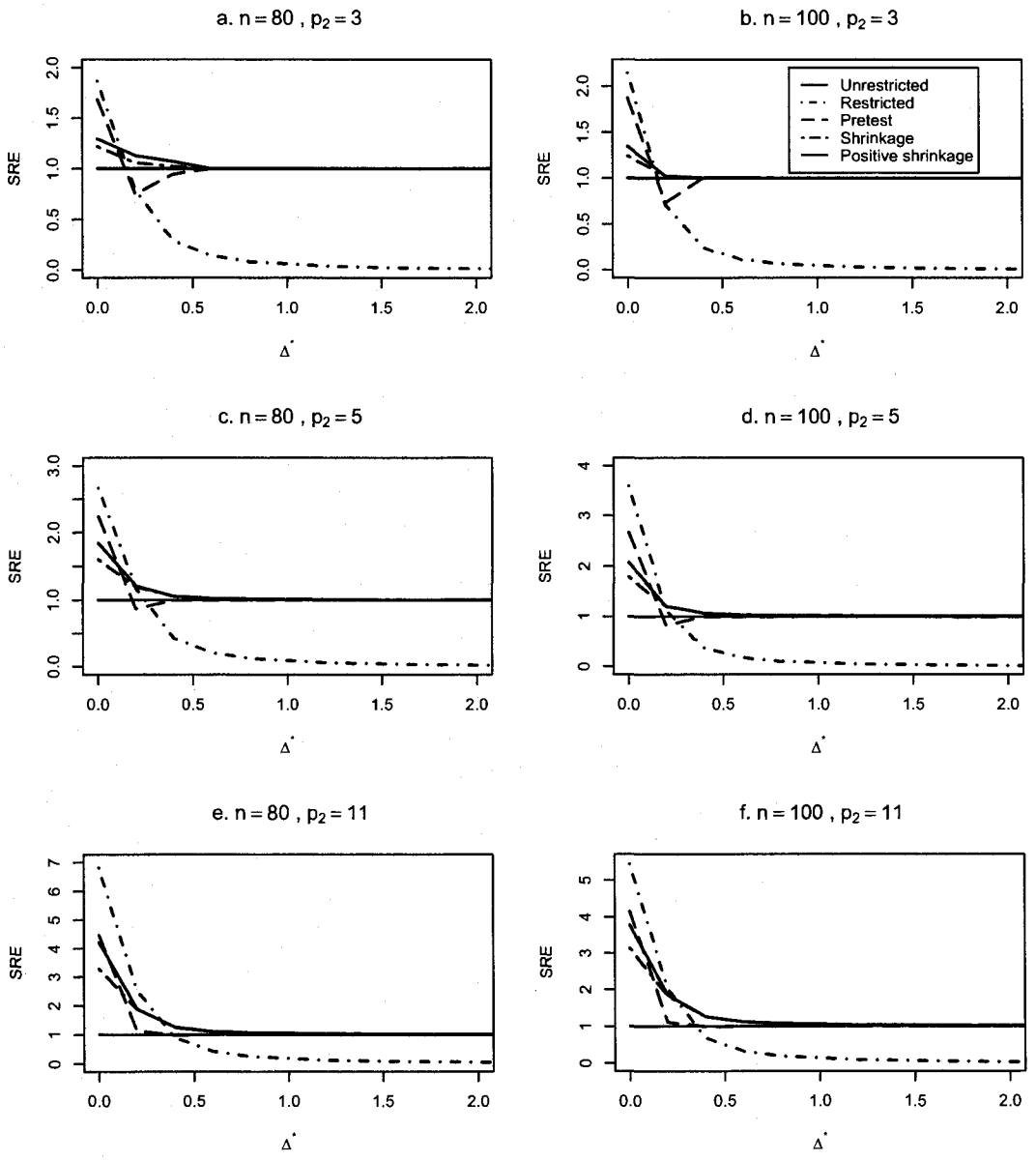

Figure 4.2: Simulated relative efficiency of the estimators as a function of non-centrality parameter Δ^* for different sample sizes n, and nuisance parameters p_2.

Table 4.5: Simulated relative efficiency with respect to $\hat{\beta}_1$ for $n = 50, p_2 = 5$.

$\Delta^\star$	$\tilde{\beta}_1$	$\hat{\beta}_1^{PT}$	$\hat{\beta}_1^{S}$	$\hat{\beta}_1^{S+}$
0.0	3.227	2.544	1.706	2.028
0.2	1.589	1.051	1.306	1.340
0.4	0.687	0.969	1.104	1.104
0.6	0.341	1.000	1.046	1.048
0.8	0.197	1.000	1.024	1.024
1.2	0.094	1.000	1.011	1.011
1.6	0.052	1.000	1.007	1.007
2.0	0.033	1.000	1.004	1.004
4.0	0.009	1.000	1.000	1.000

4.6.1 Absolute Penalty Estimators

In Tables 4.13 and 4.14, we give relative efficiencies of two absolute penalty-type estimators with respect to the unrestricted least squares estimator for $n = 30$, 50, 80 and 100 when 3 out of 14 coefficients are not zero. The penalty parameter τ is estimated using the CV and generalized CV (GCV). We only do the comparison when $\Delta^\star = 0$ because the APEs we consider here do not take advantage of the fact that $\boldsymbol{\beta}$ is partitioned into main parameters and nuisance parameters, and thus are at a disadvantage when $\Delta^\star > 0$. We see that, when $p_2 = 3$, APE performs better than the shrinkage method. On the other hand, the shrinkage method performs better when p_2 is large. Thus, we recommend using the shrinkage method when p_2 is large. Not surprisingly, the pretest estimator is the best estimator when $\Delta^* = 0$, when compared with APE and shrinkage.

4.7 Concluding Remarks

In this dissertation we compared the performance of a shrinkage estimator, a pretest estimator, an absolute penalty-type estimator and the least squares estimators in the context of a partially linear regression model with potentially irrelevant nuisance vari-

Table 4.6: Simulated relative efficiency with respect to $\hat{\boldsymbol{\beta}}_1$ for $n = 50, p_2 = 11$.

Δ^*	$\tilde{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_1^{PT}$	$\hat{\boldsymbol{\beta}}_1^{S}$	$\hat{\boldsymbol{\beta}}_1^{S+}$
0	9.808	5.145	3.561	5.064
0.2	4.188	1.951	2.569	2.817
0.4	1.611	0.949	1.612	1.622
0.6	0.791	0.994	1.295	1.295
0.8	0.465	1.000	1.169	1.169
1.2	0.213	1.000	1.078	1.079
1.6	0.119	1.000	1.042	1.042
2	0.077	1.000	1.029	1.029
4	0.019	1.000	1.006	1.000

ables. We explored the risk properties of the estimators via asymptotic distributional risk and Monte Carlo experiments. It is concluded both analytically and computationally that the restricted least squares estimator and pretest estimator dominate the usual unrestricted least squares estimators at and near the null hypothesis. The absolute penalty type estimator is competitive when the number of parameters p_2 in the nuisance parameter vector $\boldsymbol{\beta}_2$ is small, but the shrinkage estimator with appropriate data based weights performs best when p_2 is large. In fact, the shrinkage estimator outperforms the classical full model least squares estimator of the regression parameter vector in the entire parameter space for all p_2. In contrast, the performance of the reduced model least squares estimator heavily depends on the nuisance effect. Not only that, the risk of this estimator may become unbounded when the reduced model does not hold. The risk of the pretest estimator is smaller than the risk of the full model least squares estimator, $\hat{\boldsymbol{\beta}}_1$ for small $||\boldsymbol{\beta}_2||$, increases, crosses the risk of $\hat{\boldsymbol{\beta}}_1$, reaches a maximum, then decreases monotonically to the risk of $\hat{\boldsymbol{\beta}}_1$ as $||\boldsymbol{\beta}_2|| \to \infty$.

The proposed estimation strategy can be extended in various directions to more complex problems. Research on the statistical implications of proposed and related estimators is on-going. It may be worth mentioning that this is one of the two areas Bradley Efron predicted for the early 21st century (RSS News, January 1995).

Table 4.7: Simulated relative efficiency with respect to $\hat{\beta}_1$ for $n = 80, p_2 = 3$.

$\Delta^\star$	$\tilde{\beta}_1$	$\hat{\beta}_1^{PT}$	$\hat{\beta}_1^{S}$	$\hat{\beta}_1^{S+}$
0.0	1.863	1.681	1.219	1.295
0.2	0.782	0.746	1.054	1.129
0.4	0.288	0.946	1.021	1.071
0.6	0.141	1.000	1.000	1.000
0.8	0.082	1.000	0.999	0.999
1.2	0.038	1.000	0.998	0.998
1.6	0.021	1.000	0.999	0.999
2.0	0.014	1.000	0.999	0.999
4.0	0.004	1.000	0.999	0.999

Shrinkage and likelihood-based methods continue to be extremely useful tools for efficient estimation.

Table 4.8: Simulated relative efficiency with respect to $\hat{\boldsymbol{\beta}}_1$ for $n = 80, p_2 = 5$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_1^{PT}$	$\hat{\boldsymbol{\beta}}_1^{S}$	$\hat{\boldsymbol{\beta}}_1^{S+}$
0.0	2.661	2.231	1.596	1.839
0.2	1.175	0.866	1.194	1.208
0.4	0.425	0.992	1.057	1.057
0.6	0.204	1.000	1.025	1.025
0.8	0.120	1.000	1.014	1.014
1.2	0.056	1.000	1.006	1.006
1.6	0.031	1.000	1.003	1.003
2.0	0.020	1.000	1.001	1.001
4.0	0.004	1.000	1.000	1.000

Table 4.9: Simulated relative efficiency with respect to $\hat{\boldsymbol{\beta}}_1$ for $n = 80, p_2 = 11$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_1^{PT}$	$\hat{\boldsymbol{\beta}}_1^{S}$	$\hat{\boldsymbol{\beta}}_1^{S+}$
0.0	6.806	4.450	3.268	4.200
0.2	2.520	1.114	1.871	1.881
0.4	0.887	0.980	1.251	1.251
0.6	0.421	1.000	1.112	1.112
0.8	0.239	1.000	1.061	1.061
1.2	0.111	1.000	1.028	1.028
1.6	0.061	1.000	1.015	1.015
2.0	0.040	1.000	1.010	1.010
4.0	0.010	1.000	1.002	1.002

Table 4.10: Simulated relative efficiency with respect to $\hat{\boldsymbol{\beta}}_1$ for $n = 100, p_2 = 3$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_1^{PT}$	$\hat{\boldsymbol{\beta}}_1^{S}$	$\hat{\boldsymbol{\beta}}_1^{S+}$
0.0	2.139	1.866	1.238	1.343
0.2	0.709	0.728	1.021	1.023
0.4	0.240	1.000	1.003	1.003
0.6	0.115	1.000	1.001	1.001
0.8	0.066	1.000	1.000	1.000
1.2	0.031	1.000	0.999	0.999
1.6	0.016	1.000	0.999	0.999
2.0	0.011	1.000	0.999	0.999
4.0	0.002	1.000	0.999	0.999

Table 4.11: Simulated relative efficiency with respect to $\hat{\boldsymbol{\beta}}_1$ for $n = 100, p_2 = 5$.

$\Delta^\star$	$\tilde{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_1^{PT}$	$\hat{\boldsymbol{\beta}}_1^{S}$	$\hat{\boldsymbol{\beta}}_1^{S+}$
0.0	3.590	2.669	1.783	2.071
0.2	1.116	0.823	1.199	1.208
0.4	0.370	1.000	1.066	1.066
0.6	0.173	1.000	1.028	1.028
0.8	0.097	1.000	1.017	1.017
1.2	0.045	1.000	1.007	1.007
1.6	0.026	1.000	1.006	1.005
2.0	0.016	1.000	1.004	1.004
4.0	0.004	1.000	1.001	1.001

Table 4.12: Simulated relative efficiency with respect to $\hat{\boldsymbol{\beta}}_1$ for $n = 100, p_2 = 11$.

Δ^*	$\tilde{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_1^{PT}$	$\hat{\boldsymbol{\beta}}_1^{S}$	$\hat{\boldsymbol{\beta}}_1^{S+}$
0.0	5.440	4.147	3.131	3.767
0.2	2.016	1.105	1.815	1.849
0.4	0.677	0.981	1.259	1.259
0.6	0.311	1.000	1.117	1.117
0.8	0.184	1.000	1.069	1.069
1.2	0.084	1.000	1.035	1.035
1.6	0.048	1.000	1.018	1.018
2.0	0.029	1.000	1.013	1.013
4.0	0.007	1.000	1.003	1.003

Table 4.13: Simulated relative efficiency of estimators with respect to $\hat{\boldsymbol{\beta}}_1$ when $\Delta^* = 0$

	$n = 30$		$n = 50$	
Method	$p_2 = 3$	$p_2 = 11$	$p_2 = 3$	$p_2 = 11$
APE (GCV)	1.211	2.123	1.337	2.103
APE (CV)	1.179	2.316	1.387	2.208
Shrinkage	1.042	3.051	1.225	3.561
Positive Shrinkage	1.103	4.431	1.275	3.915
Pretest	1.723	4.899	1.941	5.145

Table 4.14: Simulated relative efficiency of estimators with respect to $\hat{\boldsymbol{\beta}}_1$ when $\Delta^* = 0$

	$n = 80$		$n = 100$	
Method	$p_2 = 3$	$p_2 = 11$	$p_2 = 3$	$p_2 = 11$
APE (GCV)	1.552	2.252	1.064	2.068
APE (CV)	1.373	2.186	1.184	2.054
Shrinkage	1.219	3.268	1.238	3.313
Positive Shrinkage	1.295	4.200	1.343	3.767
Pretest	1.681	4.450	1.866	4.147

Chapter 5

Conclusions and Future Research

In this dissertation we have studied estimation procedures which incorporate sample and non-sample information for some parametric and semi-parametric linear models. In some cases, we also have studied the comparison of these estimation procedures with an extended version of the LASSO procedure for different scenarios of the parameter space. The following estimation procedures are discussed in this dissertation.

(1) Unrestricted and restricted estimation.

(2) Shrinkage and positive shrinkage estimation.

(3) Pretest estimation

(4) Absolute penalty type estimation and Park and Hastie estimation.

We have applied the above estimation procedures in linear models to improve the performance of existing estimators when non-sample information is available. This can be successfully achieved by introducing shrinkage and positive shrinkage estimators which perform uniformly better than the unrestricted estimator. The estimator produced by the pretest procedure is superior to the estimators based on sample data only in some part of the parameter space induced by the non-sample prior information. The absolute penalty and Park and Hastie estimation methods perform better

than the shrinkage type estimation method when the number of restrictions on the parameter space is small.

The weighted quadratic loss function was used to calculate risk. The relative mean square error served as a criterion for comparison of the performance of the proposed shrinkage estimators. The dominance ranges of the proposed shrinkage estimators over the unrestricted estimators are discussed analytically and computationally. Our analytical findings are well supported by computational work. Several important conclusions of this study are summarized as follows:

In chapter two, we studied four estimators of Weibull regression parameters. We also applied different bootstrap methods to generate the confidence intervals for the proposed estimators. Finally a numerical example based on a real data set demonstrates how to implement and use the proposed estimation procedure. The statistical properties of the estimators were investigated analytically and numerically. The simulation study supports our theoretical findings. Based on relative mean square error, our simulation study concluded that the shrinkage and the positive shrinkage estimators outperform the classical estimator of the regression parameter vector in the entire parameter space. On the other hand, the restricted estimator is more efficient than shrinkage estimators at the null hypothesis but as it departs from this hypothesis the risk increases and becomes unbounded.

Considering the better coverage probability and lower standard error of the bounds, the percentile bootstrap method performs better than other methods. This method showed that the confidence intervals for the shrinkage estimators provide considerable improvement over the maximum likelihood estimator.

This estimation procedure can be readily extended to accommodate complex censoring patterns such as interval censoring or left truncation. The latter is the case where patients in a clinical trial have different entry times and occurs in conjunction with right and/or interval censoring.

The Weibull family of distributions have been widely used in the analysis of survival data especially in medical and engineering applications. This family is suitable in situations where the risk function is constant or monotone. It is not, however, suitable in situations where the risk function is unimodal or presents a bathtub shape. Many parametric families have been considered for modelling survival data with a more general shape for the risk function. For example, Prentice (1974) considered the generalized F distribution and Mudholkar *et al.* (1995) presented an extension of the Weibull distribution, which is called the exponentiated Weibull family of distributions, and can adequately fit lifetime data sets presenting unimodal, monotone and bathtub shaped risk functions. For future research, our regression model can be generalized to the exponentiated Weibull censored regression model and can incorporate non-sample prior information in the estimation procedure to increase the efficiency of estimates of regression parameters.

In chapter three, we consider the unrestricted, restricted, pretest, shrinkage, positive shrinkage and PH type estimators of parameters $\boldsymbol{\beta}$ for generalized linear models in the context of binary and count data. A numerical example based on real life data is used for illustration of proposed estimators presented in that chapter. It is concluded that the positive shrinkage estimator dominates the usual shrinkage type estimator and they both dominate the unrestricted estimator $\hat{\boldsymbol{\beta}}$ in terms of asymptotic distributional quadratic risk in the entire parameter space. On the other hand, the performance of the restricted estimator heavily depends on the quality of non-sample information. Under the null hypothesis, the risk of the pretest estimator keeps on increasing, crosses the risk of unrestricted maximum likelihood estimator, reaches a maximum, then decreases monotonically to the risk of the unrestricted maximum likelihood estimator. The PH estimator performs better than the shrinkage and pretest estimators when the number of restrictions on the parameter space is small and opposite conclusion holds when it is large.

The proposed estimation method for generalized linear models is the starting point of this research. This estimation method can be extended to different situations for these models, i.e., over-dispersed count data or longitudinal data when the response variable y_{ij} is related to the set of covariates x_{ij} etc. Since the logistic regression model is a special case of the multinomial logit regression model, future research will explore the properties of the shrinkage estimation method to the generalization of the logistic regression model.

In chapter 4, we compared the performance of shrinkage and positive shrinkage estimators, a pretest estimator, an absolute penalty-type estimator and least squares estimators in the context of a partially linear regression model with potentially irrelevant nuisance variables. The risk performance of the estimators is investigated through asymptotic distributional risk and Monte Carlo experiments and it is found that shrinkage estimators outperform the full model estimator uniformly. The absolute penalty type estimator performs well when the number of parameters p_2 in the nuisance parameter vector $\boldsymbol{\beta}_2$ is small, but the shrinkage estimators with appropriate data based weights perform best when p_2 is large. For all p_2, the positive shrinkage estimator dominates the usual shrinkage estimator and they both perform well relative to the classical full model least squares estimator of the regression parameter vector in the entire parameter space. On the other hand, the performance of the restricted and pretest estimators heavily depends on the quality of non-sample information.

Finally, PH and absolute penalty type estimators heavily depend on the tuning parameters but shrinkage and positive shrinkage are free from tuning parameters, and easy to compute. On the other hand, pretest estimator performs better than the PH estimator in some cases although it depends on tuning parameter α. We will recommend to the statistical community to use shrinkage estimators when the number of nuisance parameters in the linear models are large.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis.* 2nd edition, John Wiley & Sons, New York.

Ahmed, S.E. (1992). Large-sample pooling procedure for correlation. *Journal of Royal Statistical Society: Series D* **41**, 425–438.

Ahmed, S.E. (1997). *Asymptotic shrinkage estimation: the regression case.* A chapter in Applied Statistical Science II, 113-143, Nova Science Publishers, Inc, New Tork.

Ahmed, S.E. (2001). *Shrinkage estimation of regression coefficients from censored data with multiple observations. Empirical Bayes and Likelihood Inference, Lecture Notes in Statistics, 148, 103-120. Editors: S.E. Ahmed and N. Reid, Springer-Verlag: New York.* Springer-Verlag: New York.

Ahmed, S.E. (2005). Approximation-assisted estimation of eigenvalues under quadratic loss. In: *Proc. of the 14th International Workshop on Matrices and Statistics.* Institute of Information and Mathematical Sciences, Auckland, New Zealand. pp. 77–96.

Ahmed, S.E., A.A. Hussein and P.K. Sen (2006*a*). Risk comparison of some shrinkage M-estimators in linear models. *Journal of Nonparametric Statistics* **18**(4-6), 401–415.

Ahmed, S.E., A.K.Md.E. Saleh, A.I. Volodin and A.I. Volodin (2006*b*). Asymptotic expansion of the coverage probability of James-Stein estimators. *Theory of Probability and its Applications* **51**, 1–14.

Ahmed, S.E. and A.K.Md.E. Saleh (1999). Estimation of regression coefficients in an exponential regression model with censored observations. *Journal of the Japan Statistical Society* **29**, 55–64.

Ahmed, S.E. and B. Ullah (1999). To pool or not to pool: The multivariate data. *Sankhya, Series B* **61**, 266–288.

Ahmed, S.E., K.A. Doksum, S. Hossain and J. You (2007). Shrinkage, pretest and absolute penalty estimators in partially linear models. *Australian and New Zealand Journal of Statistics* **49**(4), 435–454.

Aitkin, M., D. Anderson, B. Francis and J. Hinde (1989). *Statistical Modelling in GLIM.* Oxford: Oxford Science Publications.

Bancroft, T.A. (1944). On biases in estimation due to the use of preliminary test of significance. *The Annals of Mathematical Statistics* **15**, 190–204.

Bender, R., T. Augustin and M. Blettner (2005). Generating survival times to simulate Cox's proportional hazard models. *Statistics in Medicine* **24**, 1713–1723.

Bickel, P.J. (1984). Parametric robustness: small biases can be worthwhile. *Annals of Statistics* **12**, 864–879.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.

Bugaighis, M.M. (1995). Exchange of censorship types and its impact on the estimation of parameters of a Weibull regression model. *IEEE Transactions of Reliability* **44**(3), 496–499.

Bunea, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *Annals of Statistics* **32**, 898–927.

Burman, P. and P. Chaudhuri (1992). *A Hybrid Approach to Parametric and Nonparametric Regression.* Technical Report No. 243, Division of Statistics, University of California-Davis, Davis, CA, USA.

Burnham, K.P. and D.R. Anderson (2002). *Model selection and multimodel inference.* Springer-Verlag: New York.

Chen, H. (1988). Convergence rates for parametric components in a partially linear model. *Annals of Statistics* **16**, 136–147.

Chen, H. and J. Shiau (1994). Data-driven efficient estimation for a partially linear model. *Annals of Statistics* **22**, 211–237.

Chen, S. and D.L. Donoho (1994). *On Basis Pursuit.* Technical Report, Department of Statistics, Stanford University.

Chen, S., D.L. Donoho and M.A. Saunders (1999). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* **20**(1), 33–61.

Claeskens, G. and N. Hjort (2003). The focused information criterion. *Journal of the American Statistical Association* **98**, 900–916.

Cui, X., J.T.G. Huang, J. Qiu, Blades N.J. and G.A. Churchill (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**, 59–75.

Cysneiros, F.J.A. and G.A. Paula (2005). Restricted methods in symmetrical linear regression models. *Computational Statistics and Data Analysis* **49**(3), 689–708.

Davidson, A.C. and D.V. Hinkley (1997). *Bootstrap Methods and Their Applications.* Cambridge, England: Cambridge University Press.

Dobson, A. (1990). *An Introduction to Generalized Linear Models.* London: Chapman and Hall.

Donald, G. and K. Newey (1994). Series estimation of semilinear models. *Journal of Multivariate Analysis* **50**, 30–40.

Donoho, D.L. and I.M. Johnstone (1998). Minimax estimation via wavelet shrinkage. *Annals of Statistics* **26**, 879–921.

Efron, B. (2006). Minimum volume confidence regions for a multivariate normal mean vector. *Journal of the Royal Statistical Society: Series B* **68**(4), 655–670.

Efron, B. and C. Morris (1975). Data analysis using Stein's estimator and its generlizations. *Journal of the American Statistical Association* **70**, 311–319.

Efron, B., T. Hastie, I. Johnstone and R. Tibshirani (2004). Least angle regression. *Annals of Statistics* **32**, 407–499.

Engle, R.F., Granger, J. C.W.J., Rice and A. Weiss (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of American Statistical Association* **80**, 310–319.

Eubank, R.L., J.D. Hart and P. Speckman (1990). Trigonometric series regression estimators with an application to partially linear models. *Journal of Multivariate Analysis* **32**, 70–83.

Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* **13**, 342–368.

Falorsi, P.D., S. Falorsi and A. Russo (1994). Empirical comparision of small area estimation methods for the Italian Labour Force Survey. *Survey Methodology* **20**, 171–176.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and R. Li (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics* **30**, 74–99.

Fan, J., W. Härdle and E. Mammen (1998). Direct estimation of low-dimensional components in additive models. *Annals of Statistics* **26**, 943–971.

Feigle, P. and M. Zelen (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics* **21**, 826–838.

Fiacco, A.V. and G.P. McCormick (1968). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* John Wiley & Sons, New York.

Gao, J.T. (1995a). Asymptotic theory for partially linear models. *Communications in Statistics–Theory and Methods* **24**, 1985–2009.

Gao, J.T. (1995b). The laws of the iterated logarithm of some estimates in partly linear models. *Statistics and Probability Letters* **25**, 153–162.

Gao, J.T. (1997). Adaptive parametric test in a semiparametric regression model. *Communications in Statistics–Theory and Methods* **26**, 787–800.

Genkin, A., D. Lewis and D. Madigan (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics* **49**, 291–304.

Giles, J.A. and D.E.A. Giles (1993). Preliminary-test estimation of the regression scale parameter when the loss function is asymmetric. *Communications in Statistics: Theory and Methods* **22**, 1709–1734.

Glasser, M. (1967). Exponential survival with covariance. *Journal of the American Statistical Association* **62**(318), 561–568.

Gourieroux, G. and A. Monford (1995). *Statistics and Econometric Models, vols. 1 and 2.* Cambridge University Press, Cambridge.

Green, P.J. and B.W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach.* Chapman and Hall, London.

Hamilton, A. and K. Truong (1997). Local linear estimation in partially linear models. *Journal of Multivariate Analysis* **60**, 1–19.

Härdle, W., H. Liang and J. Gao (2000). *Partially Linear Models.* Physica-Verlag, Heidelberg.

Hastie, T.J. and R.J. Tibshirani (1990). *Generalized Additive Models.* London: Chapman and Hall.

Heckman, N. (1986). Spline smoothing in a partially linear model. *Journal of the Royal Statistical Society: Series B* **48**, 244–248.

Hoeting, J.A., A.E. Rafferty and D. Madigan (2002). Bayesian variable and transformation selection in linear regression averaging: A tutorial. *Journal of Computational and Graphics Statistics* **11**, 485–507.

Hoeting, J.A., D. Madigan, A.E. Rafferty and C.T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science* **14**, 382–401.

Huang, J., S. Ma and H. Xie (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62**, 813–820.

Ihaka, R. and R. Gentleman (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314.

James, W. and C. Stein (1961). Estimation with quadratic loss. In: *Proceeding of the Fourth Berkeley Symposium On Mathematical Statistics and Probability.* University of California Press, Berkeley, CA.

Judge, G.G. and M.E. Bock (1978). *The Statistical Implications of pretest and Stein-Rule Estimators in Econometies.* Amsterdam: North Holland Publishing Company.

Judge, G.G. and R.C. Mittelhammaer (2004). A semiparametric basis for combining estimation problem under quadratic loss. *Journal of the American Statistical Association* **99**, 479–487.

Judge, G.G., W.E. Griffiths, R.C. Hill, H. Hütkepohl and Tsoung-Chao Lee (1985). *The Theory and Practice of Econometrics.* 2nd edition, John Wiley & Sons, New York.

Kalbfleisch, J.D. and R.L. Prentice (2002). *The Statistical Analysis of Failure Time Data.* John Wiley & Sons, New York.

Kazimi, C. and D. Brownstone (1999). Bootstrap confidence bands for shrinkage estimators. *Journal of Econometrics* **90**, 99–127.

Khan, B.U. and S.E. Ahmed (2006). Comparison of improved risk estimators of the multivariate mean vector. *Computational Statistics and Data Analysis* **50**, 402–421.

Knight, K. and W. Fu (2001). Asymptotics for LASSO-type estimators. *The Annals of Statistics* **28**, 1356–1378.

Kubokawa, T. (1998). The Stein phenomenon in simultaneous estimation: A review in: Applied Statistical Science III (S.E. Ahmed, M. Ahsanullah and B.K. Sinha, eds.). *NOVA Science Publishers, Inc., New York.* pp. 143–173.

Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data.* John Wiley & Sons, New York.

Lawless, J.F. and K. Signhal (1978). Efficent screening of nonnormal regression models. *Biometrics* **34**, 318–327.

Leng, C., Y. Lin and G. Wahba (2006). A note on the LASSO and related procedures in model selection. *Statistica Sinica* **16**, 1273–1284.

Liang, H. and W. Härdle (1999). Large sample theory of the estimation of the error distribution for a semiparametric model. *Communications in Statistics–Theory and Methods* **28**, 2025–2037.

Liang, H., S. Wong, J.M. Robins and R.J. Carroll (2004). Estimation in partially linear models with missing covariates. *Journal of American Statistical Association* **99**, 357–367.

Lokhorst, J. (1999). *The LASSO and generalised linear models.* Honors Project, The University of Adelaide, Australia.

McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models.* Second edition, Chapman and Hall, London.

Mudholkar, G.S., D.K. Srivastava and M. Friemer (1995). The exponentiated Weibull family: A reanalysis of the bus-motor-failure data. *Technometrics* **37**, 436–445.

Nelder, J.A. and R.W.M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A* **135**, 370–384.

Nyquist, H. (1991). Restricted estimation of generalized linear models. *Applied Statistics* **40**(1), 133–141.

Odell, P.M., K.M. Anderson and R.B. D'Agostino (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics* **48**, 951–959.

Ohtani, K., D.E.A. Giles and J.A. Giles (1997). The exact risk performance of a pretest estimator in a heteroskedastic linear regression model under a balanced loss function. *Econometric Reviews* **16**, 119–130.

Park, M.Y. and T. Hastie (2007). An L_1 regularization-path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B* **69**, 659–677.

Peduzzi, P.N., R.J. Hardy and T.R. Holdford (1980). A stepwise variable selection procedure for nonnormal regression models. *Biometrics* **36**, 511–561.

Prentice, R.L. (1974). A log-gamma model and its maximum likelihood estimation. *Biometrika* **61**, 539–544.

Rabinowitz, D., A. Tsiatis and J. Aragon (1995). Regression with interval-censored data. *Biometrika* **82**, 501–513.

Rao, C. R. (1962). A note on a generalized inverse of a matrix with applications to problems in mathematical statistics. *Journal of the Royal Statistical Society: Series B* **24**, 152–158.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* John Wiley & Sons, New York.

Rao, C.R. and H. Toutenburg (1995). *Linear Models: Least Squares and Alternatives.* Springer-Verlag: New York.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics and Probability Letters* **4**, 203–208.

Robinson, P. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica* **56**, 931–954.

Rossouw, J.E., J.P. Du Plessis, A.J.S. Benade, P.C.J. Jordaan, J.P. Kotze, P.L. Jooste and J.J. Ferreira (1983). Coronary risk factor screening in three rural communities. *South African Medical Journal* **64**, 430–436.

Saleh, A.K.Md.E. (2006). *Theory of Preliminary Test and Stein-Type Estimation with Applications.* John Wiley & Sons, New York.

Schick, A. (1994). Estimation of the autocorrelation coefficient in the presence of a regression trend. *Statistics and Probability Letters* **21**, 371–380.

Schick, A. (1996). Efficient estimation in a semiparametric additive regression model with autoregressive errors. *Stochastic Processes and their Applications* **61**, 339–361.

Schick, A. (1998). An adaptive estimator of the autocorrelation coefficient in regression models with autoregressive errors. *Stochastic Processes and their Applications* **19**, 575–589.

Sclove, S.L., C. Morris and R. Radhakrishnan (1972). Optimality of preliminary test estimation for the multinormal mean. *The Annals of Mathematical Statistics* **43**, 1481–1490.

Sen, P.K. (1986). On the asymptotic distributional risk shrinkage and preliminary test versions of maximum likelihood estimators. *Sankhya, Series A* **48**, 354–371.

Shevade, S. and S. Keerthi (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* **19**, 2246–2253.

Shi, J. and T.S. Lau (2000). Empirical likelihood for partially linear models. *Journal of Multivariate Analysis* **72**, 132–149.

Shi, P. and G. Li (1995). A note of the convergence rates of m-estimates for partially linear models. *Statistics* **26**, 27–47.

Smith, R.L. (1991). Weibull regression models for reliability data. *Reliability Engineering and System Safety* **34**(1), 55–77.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B* **50**, 413–437.

Stein, C. (1956). Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. In: *Proceeding of the Fourth Berkeley Symposium On Mathematical Statistics and Probability.* University of California Press, Berkeley, CA.. pp. 197–206.

Stigler, S.M. (1990). The 1988 Neyman Memorial Lecture, A Galtonian perpective on shrinkage estimators. *Statistical Science* **5**, 147–155.

Strang, Gilbert (2003). *Introduction to Linear Algebra, 3rd edition.* Wellesley-Cambridge Press: Wellesley, MA.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics* **23**, 461–471.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B* **58**, 267–288.

Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385–395.

Wang, Q., O. Linton and W. Härdle (2004). Semiparametric regression analysis with missing response at random. *Journal of American Statistical Association* **99**, 334–345.

Xue, H., K.F. Lam and L. Gouying (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *Journal of American Statistical Association* **99**, 346–356.

Yuan, M. and Y. Lin (2007). On the nonnegative garrote estimator. *Journal of the Royal Statistical Society: Series B* **69**, 143–161.

Zhao, P. and B. Yu (2004). *Boosted LASSO.* Technical Report, Department of Statistics, University of California at Berkeley.

Zippin, C. and P. Armitage (1966). Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. *Biometrics* **22**, 665–672.

Zippin, C. and P. Armitage (1973). Exponential survivals with censoring and explanatory variables. *Biometrika* **60**, 279–288.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* **101**(476), 1418–1429.

Vita Auctoris

Md. Shakhawat Hossain was born in 1971 in Gazipur, Bangladesh. He obtained his B.Sc(Hons.) and M.Sc. in Mathematics in 1994 and 1995 respectively from Jahangirnagar University, Dhaka, Bangladesh. In 2002, He completed his M.Sc. in Statistics from the University of Alberta, Canada. Recently he completed his Ph.D. in Statistics from the University of Windsor, Canada. He is now working as a Statistician in Population Health Research Unit, Dalhousie University, Halifax, Nova Scotia.