Perspectives on Machine Bias Versus Human Bias: Generalized Linear Models

S. Ejaz Ahmed

Department of Mathematics and Statistics
University of Windsor, Ontario

seahmed@uwindsor.ca
www.uwindsor.ca/seahmed

Current Challenges in Statistical Learning
BIRS

December 11-16, 2011

Outline of Presentation

- Executive Summary
- Proposed Estimation Strategies
- Model Selection and Post Estimation
- Simulation Study
- Application: South African Heart Disease Data
- Envoi
- Some References

Outline of Presentation

Executive Summary

Proposed Estimation Strategies

Model Selection and Post Estimation

Simulation Study

Application: South African Heart Disease Data

Envoi

Some References

Outline of Presentation

Executive Summary

Proposed Estimation Strategies

Model Selection and Post Estimation

Simulation Study

Application: South African Heart Disease Data

Envoi

Some References

Outline of Presentation

Executive Summary

Proposed Estimation Strategies

Model Selection and Post Estimation

Simulation Study

Application: South African Heart Disease Data

Envoi

Some References

Outline of Presentation

Executive Summary

Proposed Estimation Strategies

Model Selection and Post Estimation

Simulation Study

Application: South African Heart Disease Data

Envoi

Some References

Outline of Presentation

Executive Summary

Proposed Estimation Strategies

Model Selection and Post Estimation

Simulation Study

Application: South African Heart Disease Data

Envoi

Some References

Outline of Presentation

- Executive Summary
- Proposed Estimation Strategies
- Model Selection and Post Estimation
- Simulation Study
- Application: South African Heart Disease Data
- Envoi
- Some References

Introduction and Preliminaries

- Consider a 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 $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{in})'$.
- The probability density/mass function of 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 a *scale parameter*. If ϕ is known, then the exponential-family model with canonical parameter θ_i can be written as

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

- When the parameter θ_i is modelled as a linear function of the predictors, the link function is known as canonical link.

Introduction and Preliminaries

- Consider a 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 $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{in})'$.
- The probability density/mass function of 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 a *scale parameter*. If ϕ is known, then the exponential-family model with canonical parameter θ_i can be written as

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

- When the parameter θ_i is modelled as a linear function of the predictors, the link function is known as canonical link.

Introduction and Preliminaries

- Consider a 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 $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{in})'$.
- The probability density/mass function of 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 a *scale parameter*. If ϕ is known, then the exponential-family model with canonical parameter θ_i can be written as

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

- When the parameter θ_i is modelled as a linear function of the predictors, the link function is known as canonical link.

Introduction and Preliminaries

- Consider a 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 $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{in})'$.
- The probability density/mass function of 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 a *scale parameter*. If ϕ is known, then the exponential-family model with canonical parameter θ_i can be written as

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

- When the parameter θ_i is modelled as a linear function of the predictors, the link function is known as canonical link.

Introduction and Preliminaries

Some key features for Generalized Linear Model(GLIM)

- The random component of a GLIM specifies the distribution of the response variable Y_i
- The mean and variance of the response variable Y_i are given by

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

- The systematic component of a GLIM is a linear combination of regressor variables, termed the linear predictor η,

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

where $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{in})$ is the regressor vector and β is the vector of model parameters.

Some key features for Generalized Linear Model(GLIM)

- The random component of a GLIM specifies the distribution of the response variable Y_i
- The mean and variance of the response variable Y_i are given by

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

- The systematic component of a GLIM is a linear combination of regressor variables, termed the linear predictor η,

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

where $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{in})$ is the regressor vector and β is the vector of model parameters.

Some key features for Generalized Linear Model(GLIM)

- The random component of a GLIM specifies the distribution of the response variable Y_i
- The mean and variance of the response variable Y_i are given by

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

- The systematic component of a GLIM is a linear combination of regressor variables, termed the linear predictor η,

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

where $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{in})$ is the regressor vector and β is the vector of model parameters.

Some key features for Generalized Linear Model(GLIM)

- The random component of a GLIM specifies the distribution of the response variable Y_i
- The mean and variance of the response variable Y_i are given by

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

- The systematic component of a GLIM 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.

Introduction and Preliminaries

The link function connects the random and systematic components. This connection is 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 Statistical Estimation Problem

Candidate Subspace

A Great Deal of Redundancy in the Full Model

We want to estimate β when it is plausible that β lie in the subspace

$$\mathbf{H}\beta = \mathbf{h}$$

Hence the Non-Sample information (NSI) or Uncertain prior information (UPI)is

$$NSI: \ \mathbf{H}\beta = \mathbf{h}$$

$\mathbf{H}$ is $q \times k$ matrix of rank $q \leq k$
$\mathbf{h}$ is a given $q \times 1$ vector of constants.

The Statistical Estimation Problem

Candidate Subspace

A Great Deal of Redundancy in the Full Model

We want to estimate β when it is plausible that β lie in the subspace

$$\mathbf{H}\beta = \mathbf{h}$$

Hence the Non-Sample information (NSI) or Uncertain prior information (UPI)is

$$NSI: \ \mathbf{H}\beta = \mathbf{h}$$

$\mathbf{H}$ is $q \times k$ matrix of rank $q \leq k$
$\mathbf{h}$ is a given $q \times 1$ vector of constants.

The Statistical Estimation Problem

Candidate Subspace

A Great Deal of Redundancy in the Full Model

We want to estimate β when it is plausible that β lie in the subspace

$$\mathbf{H}\beta = \mathbf{h}$$

Hence the Non-Sample information (NSI) or Uncertain prior information (UPI)is

$$NSI: \ \mathbf{H}\beta = \mathbf{h}$$

$\mathbf{H}$ is $q \times k$ matrix of rank $q \leq k$
$\mathbf{h}$ is a given $q \times 1$ vector of constants.

The Statistical Estimation Problem

Candidate Subspace

A Great Deal of Redundancy in the Full Model

We want to estimate β when it is plausible that β lie in the subspace

$$\mathbf{H}\beta = \mathbf{h}$$

Hence the Non-Sample information (NSI) or Uncertain prior information (UPI)is

$$NSI: \ \mathbf{H}\beta = \mathbf{h}$$

$\mathbf{H}$ is $q \times k$ matrix of rank $q \leq k$
$\mathbf{h}$ is a given $q \times 1$ vector of constants.

The Statistical Estimation Problem

Candidate Subspace

A Great Deal of Redundancy in the Full Model

We want to estimate β when it is plausible that β lie in the subspace

$$\mathbf{H}\beta = \mathbf{h}$$

Hence the Non-Sample information (NSI) or Uncertain prior information (UPI)is

$$NSI: \ \mathbf{H}\beta = \mathbf{h}$$

$\mathbf{H}$ is $q \times k$ matrix of rank $q \leq k$
$\mathbf{h}$ is a given $q \times 1$ vector of constants.

The Statistical Estimation Problem

Candidate Subspace

A Great Deal of Redundancy in the Full Model

We want to estimate β when it is plausible that β lie in the subspace

$$\mathbf{H}\beta = \mathbf{h}$$

Hence the Non-Sample information (NSI) or Uncertain prior information (UPI)is

$$NSI: \ \mathbf{H}\beta = \mathbf{h}$$

$\mathbf{H}$ is $q \times k$ matrix of rank $q \leq k$
$\mathbf{h}$ is a given $q \times 1$ vector of constants.

Motivating Examples

Genomics Research

The goal of this paper is to analyze some of the issues involved in the estimation of the parameters in generalized linear models that may be over-parameterized that is, too many **x**'s and thus β's are included.
For example, in genomics research it is common practice to test a candidate subset of genetic markers for association with disease. Here the candidate subset is found in a certain population by doing genome wide association studies. The candidate subset is then tested for disease association in a new population. In this new population it is possible that genetic markers not found in the first population are associated with disease.

Motivating Example

Coronary Heart Disease (CHD) Data

Consider a data set which is analyzed by Park and Hastie (2006) [this data set is originally collected by Rossouw (1983)].

The coronary heart disease (CHD) may be related to the variables:

- Systolic blood pressure
- cumulative tobacco
- Low density
- Lipoprotein cholesterol

Motivating Example

Coronary Heart Disease (CHD) Data

Consider a data set which is analyzed by Park and Hastie (2006) [this data set is originally collected by Rossouw (1983)].

The coronary heart disease (CHD) may be related to the variables:

- Systolic blood pressure
- cumulative tobacco
- Low density
- Lipoprotein cholesterol

Motivating Example

Coronary Heart Disease (CHD) Data

Consider a data set which is analyzed by Park and Hastie (2006) [this data set is originally collected by Rossouw (1983)].

The coronary heart disease (CHD) may be related to the variables:

- Systolic blood pressure
- cumulative tobacco
- Low density
- Lipoprotein cholesterol

Motivating Example

Coronary Heart Disease (CHD) Data

Consider a data set which is analyzed by Park and Hastie (2006) [this data set is originally collected by Rossouw (1983)].

The coronary heart disease (CHD) may be related to the variables:

- Systolic blood pressure
- cumulative tobacco
- Low density
- Lipoprotein cholesterol

Motivating Example

Coronary Heart Disease (CHD) Data

Consider a data set which is analyzed by Park and Hastie (2006) [this data set is originally collected by Rossouw (1983)].

The coronary heart disease (CHD) may be related to the variables:

- Systolic blood pressure
- cumulative tobacco
- Low density
- Lipoprotein cholesterol

Coronary Heart Disease (CHD) Data

Variables Inclusion and Deletion (VID)

- Adiposity
- Family history of heart disease
- Type-A behavior
- Obesity
- Alcohol
- Age
- **and many other variables**

Coronary Heart Disease (CHD) Data

Variables Inclusion and Deletion (VID)

- Adiposity
- Family history of heart disease
- Type-A behavior
- Obesity
- Alcohol
- Age
- **and many other variables**

Coronary Heart Disease (CHD) Data

Variables Inclusion and Deletion (VID)

- Adiposity
- Family history of heart disease
- Type-A behavior
- Obesity
- Alcohol
- Age
- **and many other variables**

Coronary Heart Disease (CHD) Data

Variables Inclusion and Deletion (VID)

- Adiposity
- Family history of heart disease
- Type-A behavior
- Obesity
- Alcohol
- Age
- **and many other variables**

Coronary Heart Disease (CHD) Data

Variables Inclusion and Deletion (VID)

- Adiposity
- Family history of heart disease
- Type-A behavior
- Obesity
- Alcohol
- Age
- **and many other variables**

Coronary Heart Disease (CHD) Data

Variables Inclusion and Deletion (VID)

- Adiposity
- Family history of heart disease
- Type-A behavior
- Obesity
- Alcohol
- Age
- **and many other variables**

Coronary Heart Disease (CHD) Data

Variables Inclusion and Deletion (VID)

- Adiposity
- Family history of heart disease
- Type-A behavior
- Obesity
- Alcohol
- Age
- and many other variables

Coronary Heart Disease (CHD) Data

Variables Inclusion and Deletion (VID)

- Adiposity
- Family history of heart disease
- Type-A behavior
- Obesity
- Alcohol
- Age
- **and many other variables**

Coronary Heart Disease (CHD) Data

Variables Inclusion and Deletion (VID)

The maximum likelihood analysis shows that following variables are the most important factors

- Cumulative tobacco
- Low density lipoprotein cholesterol
- Family history of heart disease
- Type-A behavior
- Age

Coronary Heart Disease (CHD) Data

Variables Inclusion and Deletion (VID)

The maximum likelihood analysis shows that following variables are the most important factors

- Cumulative tobacco
- Low density lipoprotein cholesterol
- Family history of heart disease
- Type-A behavior
- Age

Coronary Heart Disease (CHD) Data

Variables Inclusion and Deletion (VID)

The maximum likelihood analysis shows that following variables are the most important factors

- Cumulative tobacco
- Low density lipoprotein cholesterol
- Family history of heart disease
- Type-A behavior
- Age

Coronary Heart Disease (CHD) Data

Variables Inclusion and Deletion (VID)

The maximum likelihood analysis shows that following variables are the most important factors

- Cumulative tobacco
- Low density lipoprotein cholesterol
- Family history of heart disease
- Type-A behavior
- Age

Coronary Heart Disease (CHD) Data

Variables Inclusion and Deletion (VID)

The maximum likelihood analysis shows that following variables are the most important factors

- Cumulative tobacco
- Low density lipoprotein cholesterol
- Family history of heart disease
- Type-A behavior
- Age

Variables Inclusion and Deletion (VID)

Nuisance Variables

The effect of some variables may be ignored

We may treat these insignificant variables as Nuisance Variables

Variables Inclusion and Deletion (VID)

Nuisance Variables

The effect of some variables may be ignored

We may treat these insignificant variables as Nuisance Variables

Variables Inclusion and Deletion (VID)

Nuisance Variables

The effect of some variables may be ignored

We may treat these insignificant variables as Nuisance Variables

VARIABLE SELECTION: OVER MODELLING

Two key aspects of variable selection methods are:

- Evaluating each potential subset of predictor variables
- Deciding on the collection of potential subsets

Evaluating Potential Subset of Predictor Variables

- R^2- Adjusted
- Akaike's Information Criterion (AIC)
- Corrected AIC
- Bayesian Information Criterion (BIC)

VARIABLE SELECTION: OVER MODELLING

Two key aspects of variable selection methods are:

- Evaluating each potential subset of predictor variables
- Deciding on the collection of potential subsets

Evaluating Potential Subset of Predictor Variables

- R^2- Adjusted
- Akaike's Information Criterion (AIC)
- Corrected AIC
- Bayesian Information Criterion (BIC)

VARIABLE SELECTION: OVER MODELLING

Two key aspects of variable selection methods are:

- Evaluating each potential subset of predictor variables
- Deciding on the collection of potential subsets

Evaluating Potential Subset of Predictor Variables

- R^2- Adjusted
- Akaike's Information Criterion (AIC)
- Corrected AIC
- Bayesian Information Criterion (BIC)

VARIABLE SELECTION: OVER MODELLING

Two key aspects of variable selection methods are:

- Evaluating each potential subset of predictor variables
- Deciding on the collection of potential subsets

Evaluating Potential Subset of Predictor Variables

- R^2- Adjusted
- Akaike's Information Criterion (AIC)
- Corrected AIC
- Bayesian Information Criterion (BIC)

VARIABLE SELECTION: OVER MODELLING

Two key aspects of variable selection methods are:

- Evaluating each potential subset of predictor variables
- Deciding on the collection of potential subsets

Evaluating Potential Subset of Predictor Variables

- R^2- Adjusted
- Akaike's Information Criterion (AIC)
- Corrected AIC
- Bayesian Information Criterion (BIC)

Likelihood Function

- Consider binary responses: $\mathbf{Y} = (y_1, y_2, \cdots, y_n)'$ and predictors $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)'$
- The log-likelihood is given by

$$l(\beta) = \sum_{i=1}^{n} [(y_i\theta_i - b(\theta_i)) + \log c(y_i)]$$

- The score equations are given by

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

where $\mathbf{D}(\mu) = \text{diag}(d_{ii})$ and $d_{ii} = 1/V(\mu_i)g'(\mu_i)$.

Likelihood Function

- Consider binary responses: $\mathbf{Y} = (y_1, y_2, \cdots, y_n)'$ and predictors $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)'$
- The log-likelihood is given by

$$l(\beta) = \sum_{i=1}^{n} [(y_i\theta_i - b(\theta_i)) + \log c(y_i)]$$

- The score equations are given by

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

where $\mathbf{D}(\mu) = \text{diag}(d_{ii})$ and $d_{ii} = 1/V(\mu_i)g'(\mu_i)$.

Likelihood Function

- Consider binary responses: $\mathbf{Y} = (y_1, y_2, \cdots, y_n)'$ and predictors $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)'$
- The log-likelihood is given by

$$l(\beta) = \sum_{i=1}^{n} [(y_i\theta_i - b(\theta_i)) + \log c(y_i)]$$

- The score equations are given by

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

where $\mathbf{D}(\mu) = \text{diag}(d_{ii})$ and $d_{ii} = 1/V(\mu_i)g'(\mu_i)$.

Likelihood Function

- Consider binary responses: $\mathbf{Y} = (y_1, y_2, \cdots, y_n)'$ and predictors $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)'$
- The log-likelihood is given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [(y_i\theta_i - b(\theta_i)) + \log c(y_i)]$$

- The score equations are given by

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

where $\mathbf{D}(\boldsymbol{\mu}) = \text{diag}(d_{ii})$ and $d_{ii} = 1/V(\mu_i)g'(\mu_i)$.

Proposed Estimation Strategies

The Candidate Estimator

- The score equations cannot be solved explicitly and hence recourse must be made numerical methods to get unrestricted maximum likelihood estimate (UE), $\hat{\beta}$.
- There are at least three methods available to solve these equations:
- The Newton-Raphson method
- Fisher's Scoring method
- Iteratively Reweighted Least Squares method

Fahrmeir and Kaufmann (1985) $\hat{\beta} \sim\ N(\beta, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1})$

Proposed Estimation Strategies

The Candidate Estimator

- The score equations cannot be solved explicitly and hence recourse must be made numerical methods to get unrestricted maximum likelihood estimate (UE), $\hat{\beta}$.
- There are at least three methods available to solve these equations:
- The Newton-Raphson method
- Fisher's Scoring method
- Iteratively Reweighted Least Squares method

Fahrmeir and Kaufmann (1985) $\hat{\beta} \sim N(\beta, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1})$

Proposed Estimation Strategies

The Candidate Estimator

- The score equations cannot be solved explicitly and hence recourse must be made numerical methods to get unrestricted maximum likelihood estimate (UE), $\hat{\beta}$.
- There are at least three methods available to solve these equations:
- The Newton-Raphson method
- Fisher's Scoring method
- Iteratively Reweighted Least Squares method

Fahrmeir and Kaufmann (1985) $\hat{\beta} \sim \ N(\beta, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1})$

Proposed Estimation Strategies

The Candidate Estimator

- The score equations cannot be solved explicitly and hence recourse must be made numerical methods to get unrestricted maximum likelihood estimate (UE), $\hat{\beta}$.
- There are at least three methods available to solve these equations:
- The Newton-Raphson method
- Fisher's Scoring method
- Iteratively Reweighted Least Squares method

Fahrmeir and Kaufmann (1985) $\hat{\beta} \sim N(\beta, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1})$

Proposed Estimation Strategies

Candidate Sub-model Estimator

- To get this estimator we need to maximize the log-likelihood under the restrictions $\mathbf{H}\beta = \mathbf{h}$.
- Using penalty function method to form a modified likelihood:

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

- Find the solution of $\text{Max}_\beta\ F(\beta, \lambda)$ for positive and fixed values of $p_j,\ j = 1, \cdots, q$.
- Using Fisher's scoring method, the solution for β will be denoted by $\hat{\beta}(\lambda)$

Proposed Estimation Strategies

Candidate Sub-model Estimator

- To get this estimator we need to maximize the log-likelihood under the restrictions $\mathbf{H}\beta = \mathbf{h}$.
- Using penalty function method to form a modified likelihood:

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

- Find the solution of $\text{Max}_\beta\, F(\beta, \lambda)$ for positive and fixed values of $p_j,\ j = 1, \cdots, q$.
- Using Fisher's scoring method, the solution for β will be denoted by $\hat{\beta}(\lambda)$

Proposed Estimation Strategies

Candidate Sub-model Estimator

- To get this estimator we need to maximize the log-likelihood under the restrictions $\mathbf{H}\beta = \mathbf{h}$.
- Using penalty function method to form a modified likelihood:

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

- Find the solution of $\text{Max}_\beta\ F(\beta, \boldsymbol{\lambda})$ for positive and fixed values of $p_j,\ j = 1, \cdots, q$.
- Using Fisher's scoring method, the solution for β will be denoted by $\hat{\beta}(\lambda)$

Proposed Estimation Strategies

Candidate Sub-model Estimator

- To get this estimator we need to maximize the log-likelihood under the restrictions $\mathbf{H}\beta = \mathbf{h}$.
- Using penalty function method to form a modified likelihood:

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

- Find the solution of $\text{Max}_\beta\ F(\beta, \boldsymbol{\lambda})$ for positive and fixed values of $p_j,\ j = 1, \cdots, q$.
- Using Fisher's scoring method, the solution for β will be denoted by $\hat{\beta}(\boldsymbol{\lambda})$

Proposed Estimation Strategies

The restricted estimator $\tilde{\beta}$ is

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

Under some regularity conditions, it may be showed that that $\tilde{\beta}$ is a consistent estimator of β, and

$$\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N_k \left(\mathbf{0}, \tilde{\mathbf{J}}^{-1}\right),$$

$$\tilde{\mathbf{J}}^{-1} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \left[\mathbf{I} - \mathbf{H}'\{\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\}^{-1}\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\right]$$

Proposed Estimation Strategies

The restricted estimator $\tilde{\beta}$ is

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

Under some regularity conditions, it may be showed that that $\tilde{\beta}$ is a consistent estimator of β, and

$$\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N_k\left(\mathbf{0}, \tilde{\mathbf{J}}^{-1}\right),$$

$$\tilde{\mathbf{J}}^{-1} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\left[\mathbf{I} - \mathbf{H}'\{\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\}^{-1}\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\right]$$

Proposed Estimation Strategies

The restricted estimator $\tilde{\beta}$ is

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

Under some regularity conditions, it may be showed that that $\tilde{\beta}$ is a consistent estimator of β, and

$$\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N_k\left(\mathbf{0}, \tilde{\mathbf{J}}^{-1}\right),$$

$$\tilde{\mathbf{J}}^{-1} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\left[\mathbf{I} - \mathbf{H}'\{\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'\}^{-1}\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\right]$$

Torturing Data Until it Confesses: Cure for the Cold

Pooling Data: Making Sense or Folly?

- Can ginseng prevent colds?
- Edmonton company CV Technologies Inc. has conducted clinical trials, with results published in the Journal of the American Geriatrics Society showing that their proprietary ginseng extract can prevent colds.
- Later, an article was published in the Vancouver Sun, in which two researchers from the UBC criticized the claims.
- They suggested that trials do not provide definite evidence that the product had any effect.

What's is going on here?

Torturing Data Until it Confesses: Cure for the Cold

Pooling Data: Making Sense or Folly?

- Can ginseng prevent colds?
- Edmonton company CV Technologies Inc. has conducted clinical trials, with results published in the Journal of the American Geriatrics Society showing that their proprietary ginseng extract can prevent colds.
- Later, an article was published in the Vancouver Sun, in which two researchers from the UBC criticized the claims.
- They suggested that trials do not provide definite evidence that the product had any effect.

What's is going on here?

Torturing Data Until it Confesses: Cure for the Cold

Pooling Data: Making Sense or Folly?

- Can ginseng prevent colds?
- Edmonton company CV Technologies Inc. has conducted clinical trials, with results published in the Journal of the American Geriatrics Society showing that their proprietary ginseng extract can prevent colds.
- Later, an article was published in the Vancouver Sun, in which two researchers from the UBC criticized the claims.
- They suggested that trials do not provide definite evidence that the product had any effect.

What's is going on here?

Torturing Data Until it Confesses: Cure for the Cold

Pooling Data: Making Sense or Folly?

- Can ginseng prevent colds?
- Edmonton company CV Technologies Inc. has conducted clinical trials, with results published in the Journal of the American Geriatrics Society showing that their proprietary ginseng extract can prevent colds.
- Later, an article was published in the Vancouver Sun, in which two researchers from the UBC criticized the claims.
- They suggested that trials do not provide definite evidence that the product had any effect.

What's is going on here?

Torturing Data Until it Confesses: Cure for the Cold

Pooling Data: Making Sense or Folly?

- Can ginseng prevent colds?
- Edmonton company CV Technologies Inc. has conducted clinical trials, with results published in the Journal of the American Geriatrics Society showing that their proprietary ginseng extract can prevent colds.
- Later, an article was published in the Vancouver Sun, in which two researchers from the UBC criticized the claims.
- They suggested that trials do not provide definite evidence that the product had any effect.

What's is going on here?

Torturing Data Until it Confesses

- The study consisted of two randomized clinical trials (2000 and 2001), with nursing-home patients as subjects.
- In each trial, the subjects were randomly assigned to take either 200 mg of the ginseng extract or a placebo twice daily.

- The trials were conducted as double-blind studies.
- It obtained results that indicated a reduction in laboratory-confirmed respiratory illness (colds and flu).

- Results were statistically significant.

Torturing Data Until it Confesses

- The study consisted of two randomized clinical trials (2000 and 2001), with nursing-home patients as subjects.
- In each trial, the subjects were randomly assigned to take either 200 mg of the ginseng extract or a placebo twice daily.
- The trials were conducted as double-blind studies.
- It obtained results that indicated a reduction in laboratory-confirmed respiratory illness (colds and flu).
- Results were statistically significant.

Torturing Data Until it Confesses

- The study consisted of two randomized clinical trials (2000 and 2001), with nursing-home patients as subjects.
- In each trial, the subjects were randomly assigned to take either 200 mg of the ginseng extract or a placebo twice daily.

- The trials were conducted as double-blind studies.
- It obtained results that indicated a reduction in laboratory-confirmed respiratory illness (colds and flu).

- Results were statistically significant.

Torturing Data Until it Confesses

- The study consisted of two randomized clinical trials (2000 and 2001), with nursing-home patients as subjects.
- In each trial, the subjects were randomly assigned to take either 200 mg of the ginseng extract or a placebo twice daily.

- The trials were conducted as double-blind studies.
- It obtained results that indicated a reduction in laboratory-confirmed respiratory illness (colds and flu).

- Results were statistically significant.

Torturing Data Until it Confesses

- Professors criticized the claims, accusing the article's authors of **data-mining**, and saying that the trials were not definitive evidence that the product had any effect.
- The original purpose of the studies was to see whether the ginseng extract would reduce the incidence of respiratory illnesses as defined by symptoms such as cough, sore throat, and runny nose.
- A secondary purpose of the studies was to measure the difference in the incidence of laboratory-confirmed respiratory illness (influenza or respiratory syncytial virus) between the two groups.

Torturing Data Until it Confesses

- Professors criticized the claims, accusing the article's authors of **data-mining**, and saying that the trials were not definitive evidence that the product had any effect.
- The original purpose of the studies was to see whether the ginseng extract would reduce the incidence of respiratory illnesses as defined by symptoms such as cough, sore throat, and runny nose.
- A secondary purpose of the studies was to measure the difference in the incidence of laboratory-confirmed respiratory illness (influenza or respiratory syncytial virus) between the two groups.

Torturing Data Until it Confesses

- Professors criticized the claims, accusing the article's authors of **data-mining**, and saying that the trials were not definitive evidence that the product had any effect.

- The original purpose of the studies was to see whether the ginseng extract would reduce the incidence of respiratory illnesses as defined by symptoms such as cough, sore throat, and runny nose.

- A secondary purpose of the studies was to measure the difference in the incidence of laboratory-confirmed respiratory illness (influenza or respiratory syncytial virus) between the two groups.

Torturing Data Until it Confesses

- Professors criticized the claims, accusing the article's authors of **data-mining**, and saying that the trials were not definitive evidence that the product had any effect.
- The original purpose of the studies was to see whether the ginseng extract would reduce the incidence of respiratory illnesses as defined by symptoms such as cough, sore throat, and runny nose.
- A secondary purpose of the studies was to measure the difference in the incidence of laboratory-confirmed respiratory illness (influenza or respiratory syncytial virus) between the two groups.

Torturing Data Until it Confesses

- The results found no significant difference between the placebo and the (ginseng extract) groups for the number of (acute respiratory illnesses) defined by symptoms.
- They also found no significant difference in the severity or duration of symptoms related to (acute respiratory illnesses) between the two groups in either study.
- However, when the researchers pooled the data from the two studies, they did get statistically significant results.

Pooling Data: Making Sense or Folly?

Case Study: Introduction to Probability and Statistics, 2e, Mendenhall, Beaver, Beaver and Ahmed, 2010, pp 365, 406-407.

Torturing Data Until it Confesses

- The results found no significant difference between the placebo and the (ginseng extract) groups for the number of (acute respiratory illnesses) defined by symptoms.
- They also found no significant difference in the severity or duration of symptoms related to (acute respiratory illnesses) between the two groups in either study.
- However, when the researchers pooled the data from the two studies, they did get statistically significant results.

Pooling Data: Making Sense or Folly?

Case Study: Introduction to Probability and Statistics, 2e, Mendenhall, Beaver, Beaver and Ahmed, 2010, pp 365, 406-407.

Torturing Data Until it Confesses

- The results found no significant difference between the placebo and the (ginseng extract) groups for the number of (acute respiratory illnesses) defined by symptoms.
- They also found no significant difference in the severity or duration of symptoms related to (acute respiratory illnesses) between the two groups in either study.
- However, when the researchers pooled the data from the two studies, they did get statistically significant results.

Pooling Data: Making Sense or Folly?

Case Study: Introduction to Probability and Statistics, 2e, Mendenhall, Beaver, Beaver and Ahmed, 2010, pp 365, 406-407.

Torturing Data Until it Confesses

- The results found no significant difference between the placebo and the (ginseng extract) groups for the number of (acute respiratory illnesses) defined by symptoms.
- They also found no significant difference in the severity or duration of symptoms related to (acute respiratory illnesses) between the two groups in either study.
- However, when the researchers pooled the data from the two studies, they did get statistically significant results.

Pooling Data: Making Sense or Folly?

Case Study: Introduction to Probability and Statistics, 2e, Mendenhall, Beaver, Beaver and Ahmed, 2010, pp 365, 406-407.

Torturing Data Until it Confesses

- The results found no significant difference between the placebo and the (ginseng extract) groups for the number of (acute respiratory illnesses) defined by symptoms.
- They also found no significant difference in the severity or duration of symptoms related to (acute respiratory illnesses) between the two groups in either study.
- However, when the researchers pooled the data from the two studies, they did get statistically significant results.

Pooling Data: Making Sense or Folly?

Case Study: Introduction to Probability and Statistics, 2e, Mendenhall, Beaver, Beaver and Ahmed, 2010, pp 365, 406-407.

Torturing Data Until it Confesses

- The results found no significant difference between the placebo and the (ginseng extract) groups for the number of (acute respiratory illnesses) defined by symptoms.

- They also found no significant difference in the severity or duration of symptoms related to (acute respiratory illnesses) between the two groups in either study.

- However, when the researchers pooled the data from the two studies, they did get statistically significant results.

Pooling Data: Making Sense or Folly?

Case Study: Introduction to Probability and Statistics, 2e, Mendenhall, Beaver, Beaver and Ahmed, 2010, pp 365, 406-407.

Torturing Data Until it Confesses

- Combining the two studies erodes the credibility of the results: Taking two studies that do not show a benefit and then adding them together to get a positive result is a form of data-mining. It's torturing the data until it confesses.

- If the original intent had been to combine the results of the two studies, then it would be a legitimate technique, but if not, it might seem that the researchers did a second study because they did not like the initial results.

Torturing Data Until it Confesses

- Combining the two studies erodes the credibility of the results: Taking two studies that do not show a benefit and then adding them together to get a positive result is a form of data-mining. It's torturing the data until it confesses.

- If the original intent had been to combine the results of the two studies, then it would be a legitimate technique, but if not, it might seem that the researchers did a second study because they did not like the initial results.

Torturing Data Until it Confesses

- Combining the two studies erodes the credibility of the results: Taking two studies that do not show a benefit and then adding them together to get a positive result is a form of data-mining. It's torturing the data until it confesses.
- If the original intent had been to combine the results of the two studies, then it would be a legitimate technique, but if not, it might seem that the researchers did a second study because they did not like the initial results.

Proposed Estimation Strategies

Hypothesis Testing

$$H_0 : \mathbf{H}\beta = \mathbf{h} \qquad H_a : \mathbf{H}\beta \neq \mathbf{h}$$

Test Statistics

Likelihood Ratio Test (LRT)

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

Wald Test Statistic

$$D_1 = (\mathbf{H}\hat{\beta} - \mathbf{h})'\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'(\mathbf{H}\hat{\beta} - \mathbf{h})$$

Rao Score Test

$$D_2 = (\mathbf{z} - \eta)'\mathbf{W}'\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}(\mathbf{z} - \eta)$$

Proposed Estimation Strategies

Hypothesis Testing

$$H_0 : \mathbf{H}\beta = \mathbf{h} \qquad H_a : \mathbf{H}\beta \neq \mathbf{h}$$

Test Statistics

Likelihood Ratio Test (LRT)

$$D = 2[l(\hat{\beta}; y_1, \cdots, y_n) - l(\tilde{\beta}; y_1, \cdots, y_n)]$$
$$= (\mathbf{H}\hat{\beta} - \mathbf{h})'\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'(\mathbf{H}\hat{\beta} - \mathbf{h}) + o_p(1)$$

Wald Test Statistic

$$D_1 = (\mathbf{H}\hat{\beta} - \mathbf{h})'\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'(\mathbf{H}\hat{\beta} - \mathbf{h})$$

Rao Score Test

$$D_2 = (\mathbf{z} - \eta)'\mathbf{W}'\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}(\mathbf{z} - \eta)$$

Proposed Estimation Strategies

Hypothesis Testing

$$H_0 : \mathbf{H}\beta = \mathbf{h} \qquad H_a : \mathbf{H}\beta \neq \mathbf{h}$$

Test Statistics

Likelihood Ratio Test (LRT)

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

Wald Test Statistic

$$D_1 = (\mathbf{H}\hat{\beta} - \mathbf{h})'\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'(\mathbf{H}\hat{\beta} - \mathbf{h})$$

Rao Score Test

$$D_2 = (\mathbf{z} - \eta)'\mathbf{W}'\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}(\mathbf{z} - \eta)$$

Proposed Estimation Strategies

Hypothesis Testing

$$H_0 : \mathbf{H}\beta = \mathbf{h} \qquad H_a : \mathbf{H}\beta \neq \mathbf{h}$$

Test Statistics

Likelihood Ratio Test (LRT)

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

Wald Test Statistic

$$D_1 = (\mathbf{H}\hat{\beta} - \mathbf{h})'\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'(\mathbf{H}\hat{\beta} - \mathbf{h})$$

Rao Score Test

$$D_2 = (\mathbf{z} - \eta)'\mathbf{W}'\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}(\mathbf{z} - \eta)$$

Proposed Estimation Strategies

Pretest Estimator

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

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

$I(A)$ is an indicator function of a set A and $\chi^2_{q,\alpha}$ is the α-level critical value of the distribution of D under H_0.

Proposed Estimation Strategies

Pretest Estimator

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

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

$I(A)$ is an indicator function of a set A and $\chi^2_{q,\alpha}$ is the α-level critical value of the distribution of D under H_0.

Proposed Estimation Strategies

Pretest Estimator

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

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

$I(A)$ is an indicator function of a set A and $\chi^2_{q,\alpha}$ is the α-level critical value of the distribution of D under H_0.

Proposed Estimation Strategies

Shrinkage and Positive Shrinkage Estimator

The shrinkage estimator (SE) of β can be defined as:

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

The positive shrinkage estimator which will control the possible over-shrinking problem is defined as

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

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

Proposed Estimation Strategies

Shrinkage and Positive Shrinkage Estimator

The shrinkage estimator (SE) of β can be defined as:

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

The positive shrinkage estimator which will control the possible over-shrinking problem is defined as

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

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

Proposed Estimation Strategies

Shrinkage and Positive Shrinkage Estimator

The shrinkage estimator (SE) of β can be defined as:

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

The positive shrinkage estimator which will control the possible over-shrinking problem is defined as

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

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

Proposed Estimation Strategies

Shrinkage and Positive Shrinkage Estimator

The shrinkage estimator (SE) of β can be defined as:

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

The positive shrinkage estimator which will control the possible over-shrinking problem is defined as

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

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

Executive Summary

- Bancroft (1944) suggested a preliminary test strategy for variable selection and post parameter estimation.
- Stein (1956, 1966) developed highly efficient shrinkage estimators in balanced designs. Most statisticians have ignored these (perhaps due to lack of understanding)
- Modern regularization estimators based on penalized least squares with multiple quadratic penalties extend Stein's procedures powerfully. This story, whose technical development relies on current empirical process theory, has only begun.

Moral of the Story

There is no suffering, no cause of suffering, no cessation of suffering, and no path. [R. Beran, 2010]

Executive Summary

- Bancroft (1944) suggested a preliminary test strategy for variable selection and post parameter estimation.
- Stein (1956, 1966) developed highly efficient shrinkage estimators in balanced designs. Most statisticians have ignored these (perhaps due to lack of understanding)
- Modern regularization estimators based on penalized least squares with multiple quadratic penalties extend Stein's procedures powerfully. This story, whose technical development relies on current empirical process theory, has only begun.

Moral of the Story

There is no suffering, no cause of suffering, no cessation of suffering, and no path. [R. Beran, 2010]

Executive Summary

- Bancroft (1944) suggested a preliminary test strategy for variable selection and post parameter estimation.
- Stein (1956, 1966) developed highly efficient shrinkage estimators in balanced designs. Most statisticians have ignored these (perhaps due to lack of understanding)
- Modern regularization estimators based on penalized least squares with multiple quadratic penalties extend Stein's procedures powerfully. This story, whose technical development relies on current empirical process theory, has only begun.

Moral of the Story

There is no suffering, no cause of suffering, no cessation of suffering, and no path. [R. Beran, 2010]

Executive Summary

- Bancroft (1944) suggested a preliminary test strategy for variable selection and post parameter estimation.
- Stein (1956, 1966) developed highly efficient shrinkage estimators in balanced designs. Most statisticians have ignored these (perhaps due to lack of understanding)
- Modern regularization estimators based on penalized least squares with multiple quadratic penalties extend Stein's procedures powerfully. This story, whose technical development relies on current empirical process theory, has only begun.

Moral of the Story

There is no suffering, no cause of suffering, no cessation of suffering, and no path. [R. Beran, 2010]

LEAST ABSOLUTE SHRINKAGE and SELECTION OPERATOR (LASSO)

- Variable selection via penalized estimation is appealing for dimension reduction.
- LASSO (Tibshirani, 1996) is a method that effectively (?) performs variable selection and regression coefficient simultaneously.
- The LASSO employs an L_1 type penalty on the regression coefficients which tends to produce sparse models, and thus is often used as a variable selection tool

LEAST ABSOLUTE SHRINKAGE and SELECTION OPERATOR (LASSO)

- Variable selection via penalized estimation is appealing for dimension reduction.
- LASSO (Tibshirani, 1996) is a method that effectively (?) performs variable selection and regression coefficient simultaneously.
- The LASSO employs an L_1 type penalty on the regression coefficients which tends to produce sparse models, and thus is often used as a variable selection tool

LEAST ABSOLUTE SHRINKAGE and SELECTION OPERATOR (LASSO)

- Variable selection via penalized estimation is appealing for dimension reduction.
- LASSO (Tibshirani, 1996) is a method that effectively (?) performs variable selection and regression coefficient simultaneously.
- The LASSO employs an L_1 type penalty on the regression coefficients which tends to produce sparse models, and thus is often used as a variable selection tool

LEAST ABSOLUTE SHRINKAGE and SELECTION OPERATOR (LASSO)

- Variable selection via penalized estimation is appealing for dimension reduction.
- LASSO (Tibshirani, 1996) is a method that effectively (?) performs variable selection and regression coefficient simultaneously.
- The LASSO employs an L_1 type penalty on the regression coefficients which tends to produce sparse models, and thus is often used as a variable selection tool

LASSO

- It is a constrained version of ordinary least squares. The LASSO estimate $\hat{\beta}(\lambda)$ is the solution to

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

  for some number $s \geq 0$ is a tuning parameter (shrinkage factor)

- Using a Lagrange multiplier argument, it can be shown that it equivalent to minimizing the residual sum of squares plus a penalty term on the absolute value of the regression coefficients.

LASSO

- It is a constrained version of ordinary least squares. The LASSO estimate $\hat{\beta}(\lambda)$ is the solution to

$$\hat{\beta}_{\lambda} = \min_{\beta}(\mathbf{y} - \mathbf{x}'\beta)'(\mathbf{y} - \mathbf{x}'\beta) \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \le s,$$

  for some number $s \ge 0$ is a tuning parameter (shrinkage factor)

- Using a Lagrange multiplier argument, it can be shown that it equivalent to minimizing the residual sum of squares plus a penalty term on the absolute value of the regression coefficients.

LASSO

- It is a constrained version of ordinary least squares. The LASSO estimate $\hat{\beta}(\lambda)$ is the solution to

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

  for some number $s \geq 0$ is a tuning parameter (shrinkage factor)

- Using a Lagrange multiplier argument, it can be shown that it equivalent to minimizing the residual sum of squares plus a penalty term on the absolute value of the regression coefficients.

LASSO

- It is a constrained version of ordinary least squares. The LASSO estimate $\hat{\beta}(\lambda)$ is the solution to

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

  for some number $s \geq 0$ is a tuning parameter (shrinkage factor)

- Using a Lagrange multiplier argument, it can be shown that it equivalent to minimizing the residual sum of squares plus a penalty term on the absolute value of the regression coefficients.

LASSO

Penalized Likelihood

An alternative formulation of the LASSO is to solve the penalized likelihood problem

$$\min \frac{1}{n}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\sum_{j=1}^{d}|\,\beta_j|$$

for some $\lambda \geq 0$.

- When the value of s is very large (or equivalently in $\lambda = 0$), the constraint (or equivalently the penalty term) has no effect and and the solution is just the set of LSE from the full model.

Penalized Likelihood

An alternative formulation of the LASSO is to solve the penalized likelihood problem

$$\min \frac{1}{n}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{d} | \beta_j|$$

for some $\lambda \geq 0$.

- When the value of s is very large (or equivalently in $\lambda = 0$), the constraint (or equivalently the penalty term) has no effect and and the solution is just the set of LSE from the full model.

LASSO

Penalized Likelihood

An alternative formulation of the LASSO is to solve the penalized likelihood problem

$$\min \frac{1}{n}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\sum_{j=1}^{d}|\,\beta_j|$$

for some $\lambda \geq 0$.

- When the value of s is very large (or equivalently in $\lambda = 0$), the constraint (or equivalently the penalty term) has no effect and and the solution is just the set of LSE from the full model.

LASSO

- Alternatively, for small values of s (or equivalently large values of λ) some of these resulting estimated regressions coefficient are exactly zero, effectively (?) omitting predictor variables from the model.
- LASSO performs variable selection and regression coefficients estimation simultaneously
- Knight and Fu (2000) studied the asymptotic properties of Lasso-type estimators.
- They showed that under appropriate conditions, the LASSO estimators are consistent for estimating the regression coefficients, and the limit distribution of the LASSO estimators can have positive probability mass at 0 when the true value of the parameter is 0.

LASSO

- Alternatively, for small values of *s* (or equivalently large values of λ) some of these resulting estimated regressions coefficient are exactly zero, effectively (?) omitting predictor variables from the model.
- LASSO performs variable selection and regression coefficients estimation simultaneously
- Knight and Fu (2000) studied the asymptotic properties of Lasso-type estimators.
- They showed that under appropriate conditions, the LASSO estimators are consistent for estimating the regression coefficients, and the limit distribution of the LASSO estimators can have positive probability mass at 0 when the true value of the parameter is 0.

LASSO

- Alternatively, for small values of *s* (or equivalently large values of λ) some of these resulting estimated regressions coefficient are exactly zero, effectively (?) omitting predictor variables from the model.
- LASSO performs variable selection and regression coefficients estimation simultaneously

- Knight and Fu (2000) studied the asymptotic properties of Lasso-type estimators.

- They showed that under appropriate conditions, the LASSO estimators are consistent for estimating the regression coefficients, and the limit distribution of the LASSO estimators can have positive probability mass at 0 when the true value of the parameter is 0.

LASSO

- Alternatively, for small values of s (or equivalently large values of λ) some of these resulting estimated regressions coefficient are exactly zero, effectively (?) omitting predictor variables from the model.
- LASSO performs variable selection and regression coefficients estimation simultaneously
- Knight and Fu (2000) studied the asymptotic properties of Lasso-type estimators.
- They showed that under appropriate conditions, the LASSO estimators are consistent for estimating the regression coefficients, and the limit distribution of the LASSO estimators can have positive probability mass at 0 when the true value of the parameter is 0.

LASSO

- Alternatively, for small values of s (or equivalently large values of λ) some of these resulting estimated regressions coefficient are exactly zero, effectively (?) omitting predictor variables from the model.
- LASSO performs variable selection and regression coefficients estimation simultaneously
- Knight and Fu (2000) studied the asymptotic properties of Lasso-type estimators.
- They showed that under appropriate conditions, the LASSO estimators are consistent for estimating the regression coefficients, and the limit distribution of the LASSO estimators can have positive probability mass at 0 when the true value of the parameter is 0.

Absolute Penalty Estimator (APE)

Algorithms

- Efron et al. (2004, Annals of Statistics,32) proposed an efficient algorithm called Least Angle Regression (LARS) that produce the entire Lasso solution paths in only *p steps*. In comparison, the classical Lasso require hundreds or thousands of steps.
- LARS, least angle regression provides a clever and very efficient algorithm of computing the complete LASSO sequence of solutions as s is varied from 0 to ∞
- Friedman, et al. (2007) developed the coordinate descent (CD) algorithm for penalized linear regression and penalized logistic regression and was shown to gain computational superiority.

Absolute Penalty Estimator (APE)

Algorithms

- Efron et al. (2004, Annals of Statistics,32) proposed an efficient algorithm called Least Angle Regression (LARS) that produce the entire Lasso solution paths in only *p steps*. In comparison, the classical Lasso require hundreds or thousands of steps.
- LARS, least angle regression provides a clever and very efficient algorithm of computing the complete LASSO sequence of solutions as s is varied from 0 to ∞
- Friedman, et al. (2007) developed the coordinate descent (CD) algorithm for penalized linear regression and penalized logistic regression and was shown to gain computational superiority.

Absolute Penalty Estimator (APE)

Algorithms

- Efron et al. (2004, Annals of Statistics,32) proposed an efficient algorithm called Least Angle Regression (LARS) that produce the entire Lasso solution paths in only *p steps*. In comparison, the classical Lasso require hundreds or thousands of steps.
- LARS, least angle regression provides a clever and very efficient algorithm of computing the complete LASSO sequence of solutions as s is varied from 0 to ∞
- Friedman, et al. (2007) developed the coordinate descent (CD) algorithm for penalized linear regression and penalized logistic regression and was shown to gain computational superiority.

Absolute Penalty Estimator (APE)

Algorithms

- Efron et al. (2004, Annals of Statistics,32) proposed an efficient algorithm called Least Angle Regression (LARS) that produce the entire Lasso solution paths in only *p steps*. In comparison, the classical Lasso require hundreds or thousands of steps.
- LARS, least angle regression provides a clever and very efficient algorithm of computing the complete LASSO sequence of solutions as s is varied from 0 to ∞
- Friedman, et al. (2007) developed the coordinate descent (CD) algorithm for penalized linear regression and penalized logistic regression and was shown to gain computational superiority.

Absolute Penalty Estimator (APE)

LASSO Family – Ever Growing

- SCAD (Fan, 1997; Fan and Li, 2001)
- Lasso and Dantzig Selector(Dasso), Candes and Tao (2007)
- Relaxed Lasso (Relaxo)
- Adaptive Lasso
- Examples include the bridge regression (Frank and Friedman, 1993), the nonnegative garrote (Breiman, 1995)

Absolute Penalty Estimator (APE)

LASSO Family – Ever Growing

- SCAD (Fan, 1997; Fan and Li, 2001)
- Lasso and Dantzig Selector(Dasso), Candes and Tao (2007)
- Relaxed Lasso (Relaxo)
- Adaptive Lasso
- Examples include the bridge regression (Frank and Friedman, 1993), the nonnegative garrote (Breiman, 1995)

Absolute Penalty Estimator (APE)

LASSO Family – Ever Growing

- SCAD (Fan, 1997; Fan and Li, 2001)
- Lasso and Dantzig Selector(Dasso), Candes and Tao (2007)
- Relaxed Lasso (Relaxo)
- Adaptive Lasso
- Examples include the bridge regression (Frank and Friedman, 1993), the nonnegative garrote (Breiman, 1995)

Absolute Penalty Estimator (APE)

LASSO Family – Ever Growing

- SCAD (Fan, 1997; Fan and Li, 2001)
- Lasso and Dantzig Selector(Dasso), Candes and Tao (2007)
- Relaxed Lasso (Relaxo)
- Adaptive Lasso
- Examples include the bridge regression (Frank and Friedman, 1993), the nonnegative garrote (Breiman, 1995)

Absolute Penalty Estimator (APE)

LASSO Family – Ever Growing

- SCAD (Fan, 1997; Fan and Li, 2001)
- Lasso and Dantzig Selector(Dasso), Candes and Tao (2007)
- Relaxed Lasso (Relaxo)
- Adaptive Lasso
- Examples include the bridge regression (Frank and Friedman, 1993), the nonnegative garrote (Breiman, 1995)

Absolute Penalty Estimator (APE)

LASSO Family – Ever Growing

- SCAD (Fan, 1997; Fan and Li, 2001)
- Lasso and Dantzig Selector(Dasso), Candes and Tao (2007)
- Relaxed Lasso (Relaxo)
- Adaptive Lasso
- Examples include the bridge regression (Frank and Friedman, 1993), the nonnegative garrote (Breiman, 1995)

Absolute Penalty Estimator

Extension to Semiparametric Models

- Ahmed et al. (2008, 2009), Raheem, Ahmed and Doksum (2011) introduced Absolute Penalty Type Estimator for partially linear models.
- Fallahpour, Ahmed and Doksum (2011) introduced Absolute Penalty Type Estimator for partially linear models with Random Coefficient autoregressive Errors.
- Further, they proposed shrinkage and pretest estimators for regression parameter vector
- A relative performance of all these competitive estimators were showcased.

Absolute Penalty Estimator

Extension to Semiparametric Models

- Ahmed et al. (2008, 2009), Raheem, Ahmed and Doksum (2011) introduced Absolute Penalty Type Estimator for partially linear models.
- Fallahpour, Ahmed and Doksum (2011) introduced Absolute Penalty Type Estimator for partially linear models with Random Coefficient autoregressive Errors.
- Further, they proposed shrinkage and pretest estimators for regression parameter vector
- A relative performance of all these competitive estimators were showcased.

Absolute Penalty Estimator

Extension to Semiparametric Models

- Ahmed et al. (2008, 2009), Raheem, Ahmed and Doksum (2011) introduced Absolute Penalty Type Estimator for partially linear models.
- Fallahpour, Ahmed and Doksum (2011) introduced Absolute Penalty Type Estimator for partially linear models with Random Coefficient autoregressive Errors.
- Further, they proposed shrinkage and pretest estimators for regression parameter vector
- A relative performance of all these competitive estimators were showcased.

Absolute Penalty Estimator

Extension to Semiparametric Models

- Ahmed et al. (2008, 2009), Raheem, Ahmed and Doksum (2011) introduced Absolute Penalty Type Estimator for partially linear models.
- Fallahpour, Ahmed and Doksum (2011) introduced Absolute Penalty Type Estimator for partially linear models with Random Coefficient autoregressive Errors.
- Further, they proposed shrinkage and pretest estimators for regression parameter vector
- A relative performance of all these competitive estimators were showcased.

Absolute Penalty Estimator

Extension to Semiparametric Models

- Ahmed et al. (2008, 2009), Raheem, Ahmed and Doksum (2011) introduced Absolute Penalty Type Estimator for partially linear models.
- Fallahpour, Ahmed and Doksum (2011) introduced Absolute Penalty Type Estimator for partially linear models with Random Coefficient autoregressive Errors.
- Further, they proposed shrinkage and pretest estimators for regression parameter vector
- A relative performance of all these competitive estimators were showcased.

Absolute Penalty Estimator

Extension to Semiparametric Models

- Ahmed et al. (2008, 2009), Raheem, Ahmed and Doksum (2011) introduced Absolute Penalty Type Estimator for partially linear models.
- Fallahpour, Ahmed and Doksum (2011) introduced Absolute Penalty Type Estimator for partially linear models with Random Coefficient autoregressive Errors.
- Further, they proposed shrinkage and pretest estimators for regression parameter vector
- A relative performance of all these competitive estimators were showcased.

Least Absolute Selection and Shrinkage, Exponential Family Edition (LASSÉ)

L_1 Type Estimator

- Park and Hastie (2006)) proposed an algorithm (called glmpath) that generates the coefficient paths for the L_1 regularization problems as in LASSO problems, but in which the squared loss function is replaced by the negative log-likelihood of any distribution in the exponential family.
- We refer to the Park-Hastie procedure as LASSÉ (least absolute selection and shrinkage, Exponential family edition).
- It is a useful tool for selecting variables according to the amount of penalization on the L_1 norm of the coefficients
- It is similar to the LASSO strategy

Least Absolute Selection and Shrinkage, Exponential Family Edition (LASSÉ)

L_1 Type Estimator

- Park and Hastie (2006)) proposed an algorithm (called glmpath) that generates the coefficient paths for the L_1 regularization problems as in LASSO problems, but in which the squared loss function is replaced by the negative log-likelihood of any distribution in the exponential family.
- We refer to the Park-Hastie procedure as LASSÉ (least absolute selection and shrinkage, Exponential family edition).
- It is a useful tool for selecting variables according to the amount of penalization on the L_1 norm of the coefficients
- It is similar to the LASSO strategy

Least Absolute Selection and Shrinkage, Exponential Family Edition (LASSÉ)

L_1 Type Estimator

- Park and Hastie (2006)) proposed an algorithm (called glmpath) that generates the coefficient paths for the L_1 regularization problems as in LASSO problems, but in which the squared loss function is replaced by the negative log-likelihood of any distribution in the exponential family.
- We refer to the Park-Hastie procedure as LASSÉ (least absolute selection and shrinkage, Exponential family edition).
- It is a useful tool for selecting variables according to the amount of penalization on the L_1 norm of the coefficients
- It is similar to the LASSO strategy

Least Absolute Selection and Shrinkage, Exponential Family Edition (LASSÉ)

L_1 Type Estimator

- Park and Hastie (2006)) proposed an algorithm (called glmpath) that generates the coefficient paths for the L_1 regularization problems as in LASSO problems, but in which the squared loss function is replaced by the negative log-likelihood of any distribution in the exponential family.
- We refer to the Park-Hastie procedure as LASSÉ (least absolute selection and shrinkage, Exponential family edition).
- It is a useful tool for selecting variables according to the amount of penalization on the L_1 norm of the coefficients
- It is similar to the LASSO strategy

Least Absolute Selection and Shrinkage, Exponential Family Edition (LASSÉ)

L_1 Type Estimator

- Park and Hastie (2006)) proposed an algorithm (called glmpath) that generates the coefficient paths for the L_1 regularization problems as in LASSO problems, but in which the squared loss function is replaced by the negative log-likelihood of any distribution in the exponential family.
- We refer to the Park-Hastie procedure as LASSÉ (least absolute selection and shrinkage, Exponential family edition).
- It is a useful tool for selecting variables according to the amount of penalization on the L_1 norm of the coefficients
- It is similar to the LASSO strategy

LASSÉ

The maximum likelihood solution for the natural parameter θ, and thus β, with a penalization on the size of the L_1 norm of the coefficients ($||\beta||_1$) i.e.,

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

- $\lambda > 0$ is the regularization parameter.
- If $\lambda = 0$, this just gives the maximum likelihood estimates.
- However, larger values of λ produce shrunken estimate of β, often with many components equal to zero.

LASSÉ

The maximum likelihood solution for the natural parameter $\boldsymbol{\theta}$, and thus β, with a penalization on the size of the L_1 norm of the coefficients ($||\beta||_1$) i.e.,

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

- $\lambda > 0$ is the regularization parameter.
- If $\lambda = 0$, this just gives the maximum likelihood estimates.
- However, larger values of λ produce shrunken estimate of β, often with many components equal to zero.

LASSÉ

The maximum likelihood solution for the natural parameter $\boldsymbol{\theta}$, and thus β, with a penalization on the size of the L_1 norm of the coefficients ($||\beta||_1$) i.e.,

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

- $\lambda > 0$ is the regularization parameter.
- If $\lambda = 0$, this just gives the maximum likelihood estimates.
- However, larger values of λ produce shrunken estimate of β, often with many components equal to zero.

LASSÉ

The maximum likelihood solution for the natural parameter $\boldsymbol{\theta}$, and thus β, with a penalization on the size of the L_1 norm of the coefficients $(||\beta||_1)$ i.e.,

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

- $\lambda > 0$ is the regularization parameter.
- If $\lambda = 0$, this just gives the maximum likelihood estimates.
- However, larger values of λ produce shrunken estimate of β, often with many components equal to zero.

LASSÉ

The maximum likelihood solution for the natural parameter $\boldsymbol{\theta}$, and thus β, with a penalization on the size of the L_1 norm of the coefficients ($||\beta||_1$) i.e.,

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

- $\lambda > 0$ is the regularization parameter.
- If $\lambda = 0$, this just gives the maximum likelihood estimates.
- However, larger values of λ produce shrunken estimate of β, often with many components equal to zero.

LASSÉ

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.,

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

- $\lambda > 0$ is the regularization parameter.
- If $\lambda = 0$, this just gives the maximum likelihood estimates.
- However, larger values of λ produce shrunken estimate of $\boldsymbol{\beta}$, often with many components equal to zero.

Algorithms

- Park and Hastie (2006), 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.
- The final estimate is denoted as the LASSÉ estimator

Algorithms

- Park and Hastie (2006), 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.
- The final estimate is denoted as the LASSÉ estimator

Algorithms

- Park and Hastie (2006), 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.
- The final estimate is denoted as the LASSÉ estimator

Adaptive Lasso

The adaptive L_1GLM is the solution of

$$\hat{\boldsymbol{\beta}}_{\lambda}^{AL_1} = -\sum_{i=1}^{n} [(y_i\theta_i - b(\theta_i)) + \ln c(y_i)] + \lambda \sum_{i=1}^{k} |\beta_i| w_i,$$

where w_i's are adaptive weights defined as $w_i = |\hat{\boldsymbol{\beta}}_i|^{-\tau}$ for some positive τ, and $\hat{\beta}_i$ is the maximizer of the log likelihood.

Adaptive Lasso

- The intuition idea of the adaptive L_1GLM is that, by allowing a relatively higher penalty for coefficients inactive predictors and lower penalty for coefficients of active predictors, it is possible to reduce the estimation bias and improve variable selection accuracy, compared with the standard LASSO.
- Theoretically, adaptive L_1GLM enjoys oracle properties (Zou, 2006) that LASSO does not have.
- When k is fixed and $n \to \infty$, with some selected λ, then the adaptive L_1GLM selects the true model with probability tending to one.

Adaptive Lasso

- The intuition idea of the adaptive L_1GLM is that, by allowing a relatively higher penalty for coefficients inactive predictors and lower penalty for coefficients of active predictors, it is possible to reduce the estimation bias and improve variable selection accuracy, compared with the standard LASSO.
- Theoretically, adaptive L_1GLM enjoys oracle properties (Zou, 2006) that LASSO does not have.
- When k is fixed and $n \to \infty$, with some selected λ, then the adaptive L_1GLM selects the true model with probability tending to one.

Adaptive Lasso

- The intuition idea of the adaptive L_1GLM is that, by allowing a relatively higher penalty for coefficients inactive predictors and lower penalty for coefficients of active predictors, it is possible to reduce the estimation bias and improve variable selection accuracy, compared with the standard LASSO.
- Theoretically, adaptive L_1GLM enjoys oracle properties (Zou, 2006) that LASSO does not have.
- When k is fixed and $n \to \infty$, with some selected λ, then the adaptive L_1GLM selects the true model with probability tending to one.

Adaptive Lasso

- The intuition idea of the adaptive L_1GLM is that, by allowing a relatively higher penalty for coefficients inactive predictors and lower penalty for coefficients of active predictors, it is possible to reduce the estimation bias and improve variable selection accuracy, compared with the standard LASSO.
- Theoretically, adaptive L_1GLM enjoys oracle properties (Zou, 2006) that LASSO does not have.
- When k is fixed and $n \to \infty$, with some selected λ, then the adaptive L_1GLM selects the true model with probability tending to one.

Smoothly Clipped Absolute Deviation

- Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) method for linear and generalized linear models.
- This method selects variables and estimate parameters β simultaneously by maximizing the penalized likelihood function

$$\hat{\beta}_{\lambda}^{SCAD} = -\sum_{i=1}^{n} [(y_i\theta_i - b(\theta_i)) + \ln c(y_i)] + \lambda \sum_{i=1}^{k} p_{\lambda}(|\beta_i|),$$

where $p_{\lambda}(\cdot)$ is the SCAD penalty with a tuning parameter λ to be selected by a data-driven method.

Smoothly Clipped Absolute Deviation

- Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) method for linear and generalized linear models.
- This method selects variables and estimate parameters β simultaneously by maximizing the penalized likelihood function

$$\hat{\beta}_{\lambda}^{SCAD} = -\sum_{i=1}^{n}[(y_i\theta_i - b(\theta_i)) + \ln c(y_i)] + \lambda\sum_{i=1}^{k}p_{\lambda}(|\beta_i|),$$

where $p_{\lambda}(\cdot)$ is the SCAD penalty with a tuning parameter λ to be selected by a data-driven method.

Smoothly Clipped Absolute Deviation

- Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) method for linear and generalized linear models.
- This method selects variables and estimate parameters β simultaneously by maximizing the penalized likelihood function

$$\hat{\beta}_{\lambda}^{SCAD} = -\sum_{i=1}^{n} [(y_i\theta_i - b(\theta_i)) + \ln c(y_i)] + \lambda \sum_{i=1}^{k} p_{\lambda}(|\beta_i|),$$

where $p_{\lambda}(\cdot)$ is the SCAD penalty with a tuning parameter λ to be selected by a data-driven method.

Smoothly Clipped Absolute Deviation

- The penalty $p_\lambda(\cdot)$ satisfies $p_\lambda(0) = 0$, and its first-order derivative

$$p_\lambda^{'}(\theta) = \lambda \left[I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right],$$

  where a is some constant usually taken to be $a = 3.7$ and $(t)_+ = tI\{t > 0\}$ is the hinge loss function.
- This method consistently identifies inactive variables by producing zero solutions for their associated regression coefficients.
- Fan and Li (2001) demonstrated that as n increases, the SCAD procedure selects the true set of nonzero coefficients with probability tending to one.

Smoothly Clipped Absolute Deviation

- The penalty $p_\lambda(\cdot)$ satisfies $p_\lambda(0) = 0$, and its first-order derivative

  $$p'_\lambda(\theta) = \lambda\left[I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\right],$$

  where a is some constant usually taken to be $a = 3.7$ and $(t)_+ = tI\{t > 0\}$ is the hinge loss function.
- This method consistently identifies inactive variables by producing zero solutions for their associated regression coefficients.
- Fan and Li (2001) demonstrated that as n increases, the SCAD procedure selects the true set of nonzero coefficients with probability tending to one.

Smoothly Clipped Absolute Deviation

- The penalty $p_\lambda(\cdot)$ satisfies $p_\lambda(0) = 0$, and its first-order derivative

$$p_\lambda^{'}(\theta) = \lambda \left[I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right],$$

  where a is some constant usually taken to be $a = 3.7$ and $(t)_+ = tI\{t > 0\}$ is the hinge loss function.
- This method consistently identifies inactive variables by producing zero solutions for their associated regression coefficients.
- Fan and Li (2001) demonstrated that as n increases, the SCAD procedure selects the true set of nonzero coefficients with probability tending to one.

Smoothly Clipped Absolute Deviation

- The penalty $p_\lambda(\cdot)$ satisfies $p_\lambda(0) = 0$, and its first-order derivative

  $$p'_\lambda(\theta) = \lambda\left[I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\right],$$

  where a is some constant usually taken to be $a = 3.7$ and $(t)_+ = tI\{t > 0\}$ is the hinge loss function.
- This method consistently identifies inactive variables by producing zero solutions for their associated regression coefficients.
- Fan and Li (2001) demonstrated that as n increases, the SCAD procedure selects the true set of nonzero coefficients with probability tending to one.

Post-Model Selection Estimation Difficulties

- The variable selection process changes the properties of the estimators
- Regardless of sample size, the model selection step typically has a dramatic effect on the sampling properties of the estimators.
- As well as the properties of standard inferential procedures (tests and confidence intervals)
- The regression coefficients obtained after variable selection are biased
- Further, the p-values obtained after variable selection from standard statistics are generally much smaller than their true values.

Post-Model Selection Estimation Difficulties

- The variable selection process changes the properties of the estimators
- Regardless of sample size, the model selection step typically has a dramatic effect on the sampling properties of the estimators.
- As well as the properties of standard inferential procedures (tests and confidence intervals)
- The regression coefficients obtained after variable selection are biased
- Further, the p-values obtained after variable selection from standard statistics are generally much smaller than their true values.

Post-Model Selection Estimation Difficulties

- The variable selection process changes the properties of the estimators
- Regardless of sample size, the model selection step typically has a dramatic effect on the sampling properties of the estimators.
- As well as the properties of standard inferential procedures (tests and confidence intervals)
- The regression coefficients obtained after variable selection are biased
- Further, the p-values obtained after variable selection from standard statistics are generally much smaller than their true values.

Post-Model Selection Estimation Difficulties

- The variable selection process changes the properties of the estimators
- Regardless of sample size, the model selection step typically has a dramatic effect on the sampling properties of the estimators.
- As well as the properties of standard inferential procedures (tests and confidence intervals)
- The regression coefficients obtained after variable selection are biased
- Further, the p-values obtained after variable selection from standard statistics are generally much smaller than their true values.

Inference After Variable Selection

Post-Model Selection Estimation Difficulties

- The variable selection process changes the properties of the estimators
- Regardless of sample size, the model selection step typically has a dramatic effect on the sampling properties of the estimators.
- As well as the properties of standard inferential procedures (tests and confidence intervals)
- The regression coefficients obtained after variable selection are biased
- Further, the p-values obtained after variable selection from standard statistics are generally much smaller than their true values.

Inference After Variable Selection

Post-Model Selection Estimation Difficulties

- The variable selection process changes the properties of the estimators
- Regardless of sample size, the model selection step typically has a dramatic effect on the sampling properties of the estimators.
- As well as the properties of standard inferential procedures (tests and confidence intervals)
- The regression coefficients obtained after variable selection are biased
- Further, the p-values obtained after variable selection from standard statistics are generally much smaller than their true values.

Inference After Variable Selection

Innate Difficulties of Data Driven Model Selection

- The Data-driven model selection that do not seem to have been widely appreciated or that seem to be viewed too optimistically
- Despite some claims to contrary, no model selection procedure either implemented on a machine or not is immune to these difficulties.[Leeb and Potscher, 2005]

ARE LASSO, APE, and LASSÉ are IMMUNIZED????

Inference After Variable Selection

Innate Difficulties of Data Driven Model Selection

- The Data-driven model selection that do not seem to have been widely appreciated or that seem to be viewed too optimistically
- Despite some claims to contrary, no model selection procedure either implemented on a machine or not is immune to these difficulties.[Leeb and Potscher, 2005]

ARE LASSO, APE, and LASSÉ are IMMUNIZED????

Inference After Variable Selection

Innate Difficulties of Data Driven Model Selection

- The Data-driven model selection that do not seem to have been widely appreciated or that seem to be viewed too optimistically
- Despite some claims to contrary, no model selection procedure either implemented on a machine or not is immune to these difficulties.[Leeb and Potscher, 2005]

ARE LASSO, APE, and LASSÉ are IMMUNIZED????

Asymptotic Treatment

Consider a sequence $K_{(n)}$ of local alternatives defined by

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

$\boldsymbol{\delta} = (\delta_1, \delta_2 \cdots, \delta_q) \in \Re^q$, a real fixed vector.

Note that for $\boldsymbol{\delta} = \mathbf{0}$, $\mathbf{H}\beta = \mathbf{h}$, for all n.

We define a quadratic loss function using a positive definite matrix (p.d.m.) $\mathbf{Q}$

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

Asymptotic Analysis

- The asymptotic distribution function of β^* under $k_{(n)}$ by

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

where $G(\mathbf{y})$ is nondegenerate distribution function.

- The asymptotic distributional quadratic risk (ADR) by

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

$$\mathbf{Q}^* = \int \cdots \int \mathbf{y}\mathbf{y}' dG(\mathbf{y})$$

is the dispersion matrix for the distribution $G(\mathbf{y})$.

Asymptotic Analysis

- The asymptotic distribution function of β^* under $k_{(n)}$ by

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

where $G(\mathbf{y})$ is nondegenerate distribution function.

- The asymptotic distributional quadratic risk (ADR) by

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

$$\mathbf{Q}^* = \int \cdots \int \mathbf{y}\mathbf{y}'\, dG(\mathbf{y})$$

is the dispersion matrix for the distribution $G(\mathbf{y})$.

Asymptotic Analysis

- The asymptotic distribution function of β^* under $k_{(n)}$ by

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

where $G(\mathbf{y})$ is nondegenerate distribution function.

- The asymptotic distributional quadratic risk (ADR) by

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

$$\mathbf{Q}^* = \int \cdots \int \mathbf{y}\mathbf{y}' dG(\mathbf{y})$$

is the dispersion matrix for the distribution $G(\mathbf{y})$.

Mathematical Proof

Theorem: Under local alternatives $k_{(n)}$ and usual regularity conditions we have the ADB of the proposed estimators as $n \to \infty$ in the following:

$$
\begin{aligned}
ADB(\hat{\beta}) &= \mathbf{0}, && (1)\\
ADB(\tilde{\beta}) &= -\mathbf{J}\delta, \quad \mathbf{J} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}'[\mathbf{H}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{H}']^{-1}, && (2)\\
ADB(\hat{\beta}^{PT}) &= \mathbf{J}\delta\Psi_{q+2}(q-2,\Delta), && (3)\\
ADB(\hat{\beta}^{S}) &= -(q-2)\mathbf{J}\delta E(\chi^{-2}_{q+2}(\Delta)), && (4)\\
ADB(\hat{\beta}^{S+}) &= -(q-2)\mathbf{J}\delta\left[E(\chi^{-2}_{q+2}(\Delta)) - E(\chi^{-2}_{q+2}(\Delta)I(\chi^{2}_{q+2}(\Delta) < (q-2)))\right]\\
&\quad - \mathbf{J}\delta\Psi_{q+2}(q-2,\Delta), && (5)
\end{aligned}
$$

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.

Mathematical Proof

Theorem: Under local alternatives $k_{(n)}$ and usual regularity conditions we have the ADRs of $\hat{\beta}$, $\tilde{\beta}$, $\hat{\beta}^{PT}$, $\hat{\beta}^{S}$ and $\hat{\beta}^{S+}$ are respectively:

$$\begin{aligned}
R(\hat{\beta}) &= \text{trace}[\mathbf{Q}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}], \\
R(\tilde{\beta}) &= R(\hat{\beta}) - \text{trace}[\mathbf{QJH}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}] + \delta'(\mathbf{J}'\mathbf{QJ})\delta, \\
R(\hat{\beta}^{PT}) &= R(\hat{\beta}) - \text{trace}[\mathbf{QJH}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}]\Psi_{q+2}(q-2,\Delta) \\
&+ \delta'(\mathbf{J}'\mathbf{QJ})\delta[2\Psi_{q+2}(q-2,\Delta) - \Psi_{q+4}(q-2,\Delta)], \\
R(\hat{\beta}^{S}) &= R(\hat{\beta}) - 2(q-2)\text{trace}[\mathbf{QJH}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}]\{2E(\chi^{-2}_{q+2}(\Delta)) \\
&- (q-2)E(\chi^{-4}_{q+2}(\Delta))\} + (q-2)\delta'(\mathbf{J}'\mathbf{QJ})\delta\{2E(\chi^{-2}_{q+2}(\Delta)) \\
&- 2E(\chi^{-4}_{q+2}(\Delta)) + (q-2)E(\chi^{-4}_{q+4}(\Delta))\}, \\
R(\hat{\beta}^{S+}) &= R(\hat{\beta}^{S}) - \delta'(\mathbf{J}'\mathbf{QJ})\delta E[(1-(q-2)\chi^{-2}_{q+4}(\Delta))^2 I(\chi^2_{q+4}(\Delta) < (q-2))] \\
&- \text{trace}[\mathbf{QJH}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}]E[(1-(q-2)\chi^{-2}_{q+2}(\Delta))^2 I(\chi^2_{q+4}(\Delta) < (q-2))] \\
&+ 2\delta'(\mathbf{J}'\mathbf{QJ})\delta E[(1-(q-2)\chi^{-2}_{q+4}(\Delta)) I(\chi^2_{q+4}(\Delta) < (q-2))].
\end{aligned}$$

Engineering Proof: Simulation

- We use Monte Carlo simulation experiments to examine the risk performance of proposed estimators based on large sample methodology under various scenarios.

- Our sampling experiment consists of different combinations of sample sizes, i.e., $n = 100, 150, 200$.

- In this study we simulate binary response from the following model:

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

$p_i = P(Y = 1|\, x_i)$

- The covariate matrix $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{in})$ has been drawn from a multivariate standard normal distribution.

Engineering Proof: Simulation

- We use Monte Carlo simulation experiments to examine the risk performance of proposed estimators based on large sample methodology under various scenarios.

- Our sampling experiment consists of different combinations of sample sizes, i.e., $n = 100, 150, 200$.

- In this study we simulate binary response from the following model:

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

$p_i = P(Y = 1 | \, x_i)$

- The covariate matrix $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{in})$ has been drawn from a multivariate standard normal distribution.

Engineering Proof: Simulation

- We use Monte Carlo simulation experiments to examine the risk performance of proposed estimators based on large sample methodology under various scenarios.

- Our sampling experiment consists of different combinations of sample sizes, i.e., $n = 100, 150, 200$.

- In this study we simulate binary response from the following model:

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

$p_i = P(Y = 1|\, x_i)$

- The covariate matrix $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{in})$ has been drawn from a multivariate standard normal distribution.

Engineering Proof: Simulation

- We use Monte Carlo simulation experiments to examine the risk performance of proposed estimators based on large sample methodology under various scenarios.

- Our sampling experiment consists of different combinations of sample sizes, i.e., $n = 100, 150, 200$.

- In this study we simulate binary response from the following model:

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

$p_i = P(Y = 1|\, x_i)$

- The covariate matrix $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{in})$ has been drawn from a multivariate standard normal distribution.

Engineering Proof: Simulation

- We use Monte Carlo simulation experiments to examine the risk performance of proposed estimators based on large sample methodology under various scenarios.

- Our sampling experiment consists of different combinations of sample sizes, i.e., $n = 100, 150, 200$.

- In this study we simulate binary response from the following model:

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

$p_i = P(Y = 1|\, x_i)$

- The covariate matrix $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{in})$ has been drawn from a multivariate standard normal distribution.

Simulation Results

- For simulation we consider the particular case of hypothesis $H_0 : \beta_2 = \mathbf{0}$, where β_2 is a $k_2 \times 1$ vector with $k = k_1 + k_2$.

- We set the true value of β at $\beta = (\beta_1, \beta_2) = (c(1.5, 2.5), \beta_2)$ 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 $\alpha = 0.05$.

- Our NSI is $H_0 : \beta_2 = 0$ and $\Delta^\star = ||\beta - \beta^{(0)}||^2$

Simulation Results

- For simulation we consider the particular case of 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 β at $\beta = (\beta_1, \beta_2) = (c(1.5, 2.5), \beta_2)$ 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 $\alpha = 0.05$.
- Our NSI is $H_0 : \beta_2 = 0$ and $\Delta^\star = ||\beta - \beta^{(0)}||^2$

Simulation Results

- For simulation we consider the particular case of 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 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (c(1.5, 2.5), \boldsymbol{\beta}_2)$ 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 $\alpha = 0.05$.
- Our NSI is $H_0 : \boldsymbol{\beta}_2 = 0$ and $\Delta^* = ||\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}||^2$

Simulation Results

- For simulation we consider the particular case of 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 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (c(1.5, 2.5), \boldsymbol{\beta}_2)$ 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 $\alpha = 0.05$.
- Our NSI is $H_0 : \boldsymbol{\beta}_2 = 0$ and $\Delta^{\star} = ||\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}||^2$

Simulation Results

- The performance of an estimator of β will be appraised using 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 $\tilde{\beta}$, $\hat{\beta}^{PT}$, $\hat{\beta}^{S}$, and $\hat{\beta}^{S+}$ with respect to $\hat{\beta}$ by simulation.
- The simulated relative efficiency (SRE) of the estimator $\beta^{\diamond}$ to the maximum likelihood estimator $\hat{\beta}$ is denoted by

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

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

Simulation Results

- The performance of an estimator of β will be appraised using 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 $\tilde{\beta}$, $\hat{\beta}^{PT}$, $\hat{\beta}^{S}$, and $\hat{\beta}^{S+}$ with respect to $\hat{\beta}$ by simulation.

- The simulated relative efficiency (SRE) of the estimator $\beta^{\diamond}$ to the maximum likelihood estimator $\hat{\beta}$ is denoted by

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

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

Simulation Results

- The performance of an estimator of β will be appraised using 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 $\tilde{\beta}$, $\hat{\beta}^{PT}$, $\hat{\beta}^{S}$, and $\hat{\beta}^{S+}$ with respect to $\hat{\beta}$ by simulation.

- The simulated relative efficiency (SRE) of the estimator $\beta^{\diamond}$ to the maximum likelihood estimator $\hat{\beta}$ is denoted by

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

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

Simulation Results

- 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 system (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 $\beta^{\diamond}$ to the maximum likelihood estimator $\hat{\beta}$ is denoted by

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

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

Simulation Results

- The performance of an estimator of β will be appraised using 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 $\tilde{\beta}$, $\hat{\beta}^{PT}$, $\hat{\beta}^{S}$, and $\hat{\beta}^{S+}$ with respect to $\hat{\beta}$ by simulation.

- The simulated relative efficiency (SRE) of the estimator $\beta^{\diamond}$ to the maximum likelihood estimator $\hat{\beta}$ is denoted by

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

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

Simulation Results

- The performance of an estimator of β will be appraised using 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 $\tilde{\beta}$, $\hat{\beta}^{PT}$, $\hat{\beta}^{S}$, and $\hat{\beta}^{S+}$ with respect to $\hat{\beta}$ by simulation.

- The simulated relative efficiency (SRE) of the estimator $\beta^{\diamond}$ to the maximum likelihood estimator $\hat{\beta}$ is denoted by

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

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

Simulation Results

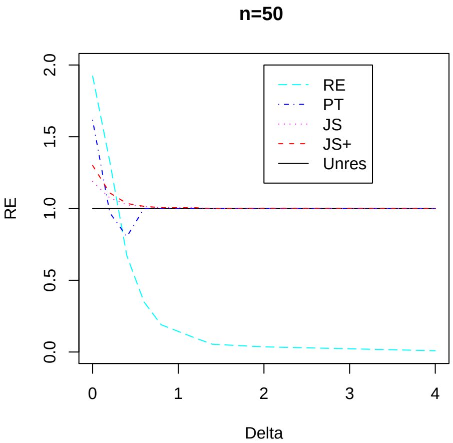

Figure: Relative efficiency of the estimators as a function of non-centrality parameter Δ^* for sample sizes $n = 150$, and insignificant parameters $k_2 = 3$

Simulation Results

Table: Simulated relative MSE with respect to $\hat{\beta}$ for $n = 150, k_2 = 3$.

$\Delta^\star$	RE	PTE	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

Simulation Results

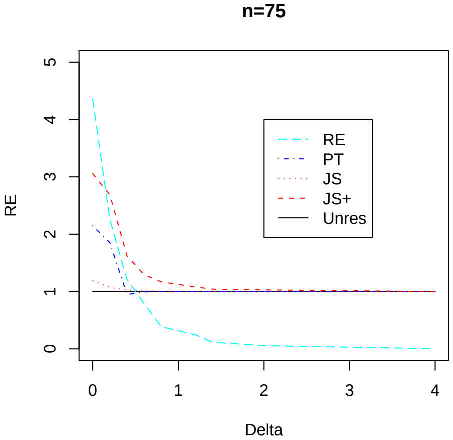

Figure: Relative MSE of the estimators as a function of non-centrality parameter Δ^* for sample sizes $n = 150$, and nuisance parameters $k_2 = 7$

Simulation Results

Table: Simulated relative MSE with respect to $\hat{\beta}$ for $n = 150, k_2 = 7$.

$\Delta^\star$	RE	PTE	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

Simulation Results

Table: Relative efficiency of RE, SE, PSE, L_1GLM, adaptive L_1GLM, and SCAD with respect to $\hat{\beta}$ when $\Delta^* = 0$ and $n = 200$

Method	$k_2 = 3$	$k_2 = 5$	$k_2 = 7$	$k_2 = 11$	$k_2 = 15$	$k_2 = 20$
Restricted	1.79	2.36	3.02	4.50	7.16	9.82
Pretest	1.53	1.81	2.19	2.65	2.67	2.72
Shrinkage	1.16	1.50	1.82	1.63	3.93	4.04
Positive Shrinkage	1.22	1.60	1.98	2.77	4.10	4.28
L_1GLM	1.24	1.53	1.69	2.51	3.38	3.92
Adaptive L_1GLM	1.34	1.55	1.77	2.53	3.51	4.02
SCAD	1.51	1.60	1.87	2.61	3.82	4.17

Application: South African heart disease data

- This data set 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 was to predict CHD (coronary heart disease)=1 or 0; present or absent, from a set of covariates listed from below:
  - **sbp:** systolic blood pressure
  - **tobacco:** cumulative tobacco (kg) **ldl:** low densiity lipoprotein cholesterol
  - **adiposity:** Adiposity level 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

Application: South African heart disease data

- This data set 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 was to predict CHD (coronary heart disease)=1 or 0; present or absent, from a set of covariates listed from below:
  - **sbp:** systolic blood pressure
  - **tobacco:** cumulative tobacco (kg) **ldl:** low densiity lipoprotein cholesterol
  - **adiposity:** Adiposity level 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

Application: South African heart disease data

- This data set 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 was to predict CHD (coronary heart disease)=1 or 0; present or absent, from a set of covariates listed from below:
  - **sbp:** systolic blood pressure
  - **tobacco:** cumulative tobacco (kg) **ldl:** low densiity lipoprotein cholesterol
  - **adiposity:** Adiposity level 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

Application: South African heart disease data

- This data set 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 was to predict CHD (coronary heart disease)=1 or 0; present or absent, from a set of covariates listed from below:
  - **sbp:** systolic blood pressure
  - **tobacco:** cumulative tobacco (kg) **ldl:** low densiity lipoprotein cholesterol
  - **adiposity:** Adiposity level 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

Application: South African heart disease data

- This data set 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 was to predict CHD (coronary heart disease)=1 or 0; present or absent, from a set of covariates listed from below:
  - **sbp:** systolic blood pressure
  - **tobacco:** cumulative tobacco (kg) **ldl:** low densiity lipoprotein cholesterol
  - **adiposity:** Adiposity level 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

Application: South African heart disease data

- This data set 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 was to predict CHD (coronary heart disease)=1 or 0; present or absent, from a set of covariates listed from below:
  - **sbp:** systolic blood pressure
  - **tobacco:** cumulative tobacco (kg) **ldl:** low densiity lipoprotein cholesterol
  - **adiposity:** Adiposity level 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

Application: South African heart disease data

- This data set 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 was to predict CHD (coronary heart disease)=1 or 0; present or absent, from a set of covariates listed from below:
    - **sbp:** systolic blood pressure
    - **tobacco:** cumulative tobacco (kg) **ldl:** low densiity lipoprotein cholesterol
    - **adiposity:** Adiposity level 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

Application: South African heart disease data

- This data set 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 was to predict CHD (coronary heart disease)=1 or 0; present or absent, from a set of covariates listed from below:
  - **sbp:** systolic blood pressure
  - **tobacco:** cumulative tobacco (kg) **ldl:** low densiity lipoprotein cholesterol
  - **adiposity:** Adiposity level 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

Application: South African heart disease data

- This data set 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 was to predict CHD (coronary heart disease)=1 or 0; present or absent, from a set of covariates listed from below:
  - **sbp:** systolic blood pressure
  - **tobacco:** cumulative tobacco (kg) **ldl:** low densiity lipoprotein cholesterol
  - **adiposity:** Adiposity level 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

Application: South African heart disease data

- This data set 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 was to predict CHD (coronary heart disease)=1 or 0; present or absent, from a set of covariates listed from below:
    - **sbp:** systolic blood pressure
    - **tobacco:** cumulative tobacco (kg) **ldl:** low densiity lipoprotein cholesterol
    - **adiposity:** Adiposity level 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

Application: South African heart disease data

- This data set 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 was to predict CHD (coronary heart disease)=1 or 0; present or absent, from a set of covariates listed from below:
  - **sbp:** systolic blood pressure
  - **tobacco:** cumulative tobacco (kg) **ldl:** low densiity lipoprotein cholesterol
  - **adiposity:** Adiposity level 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

Application: South African heart disease data

- This data set 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 was to predict CHD (coronary heart disease)=1 or 0; present or absent, from a set of covariates listed from below:
    - **sbp:** systolic blood pressure
    - **tobacco:** cumulative tobacco (kg) **ldl:** low densiity lipoprotein cholesterol
    - **adiposity:** Adiposity level 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

Application: South African Heart Disease Data

Consider the full model

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

Application: South African Heart Disease Data

Table: Estimate (first row) and standard error (second row) for tobacco (β_1), ldl (β_2), famhist (β_3), age (β_4), and typea (β_5) on coronary heart disease. The SRE column gives the relative efficiency based on bootstrap simulation of the estimators with respect to UE.

Estimators	β_1	β_2	β_3	β_4	β_5	SRE
UE	0.541 0.284	0.399 0.290	0.190 0.219	0.607 0.352	0.342 0.243	1.0000
RE	0.506 0.245	0.377 0.257	0.194 0.204	0.699 0.277	0.321 0.231	2.520
PT	0.513 0.260	0.386 0.273	0.194 0.209	0.678 0.305	0.328 0.225	1.476
SE	0.522 0.265	0.391 0.278	0.193 0.212	0.661 0.322	0.332 0.238	1.327
PSE	0.523 0.266	0.391 0.275	0.192 0.212	0.654 0.309	0.333 0.237	1.547
L_1GLM	0.407 0.233	0.285 0.238	0.133 0.162	0.538 0.266	0.203 0.198	1.789
Adaptive L_1GLM	0.407 0.231	0.284 0.224	0.133 0.164	0.538 0.272	0.207 0.192	1.808
SCAD	0.387 0.201	0.239 0.292	0.132 0.156	0.483 0.238	0.184 0.178	1.879

Envoi

- Gauss provided two justifications for least squares:
  - The maximum likelihood argument in the Gaussian error model.
  - The idea of risk, commonly known as the Gauss-Markov theorem.

- Stein's 1956 paper revealed that neither maximum likelihood estimators nor unbiased estimators have desirable risk functions when the dimension of the parameter space is not small.

- The SE and PSE outperforms the maximum likelihood estimator of the regression parameter vector in the entire parameter space.

Envoi

- Gauss provided two justifications for least squares:
  - The maximum likelihood argument in the Gaussian error model.
  - The idea of risk, commonly known as the Gauss-Markov theorem.

- Stein's 1956 paper revealed that neither maximum likelihood estimators nor unbiased estimators have desirable risk functions when the dimension of the parameter space is not small.

- The SE and PSE outperforms the maximum likelihood estimator of the regression parameter vector in the entire parameter space.

Envoi

- Gauss provided two justifications for least squares:
  - The maximum likelihood argument in the Gaussian error model.
  - The idea of risk, commonly known as the Gauss-Markov theorem.
- Stein's 1956 paper revealed that neither maximum likelihood estimators nor unbiased estimators have desirable risk functions when the dimension of the parameter space is not small.
- The SE and PSE outperforms the maximum likelihood estimator of the regression parameter vector in the entire parameter space.

Envoi

- Gauss provided two justifications for least squares:
  - The maximum likelihood argument in the Gaussian error model.
  - The idea of risk, commonly known as the Gauss-Markov theorem.

- Stein's 1956 paper revealed that neither maximum likelihood estimators nor unbiased estimators have desirable risk functions when the dimension of the parameter space is not small.

- The SE and PSE outperforms the maximum likelihood estimator of the regression parameter vector in the entire parameter space.

Envoi

Shrinkage Versus LASSÉ

- The LASSÉ dominates the SE when the number of restrictions on parameters are small.
- Shrinkage estimators outshines the LASSÉ estimation strategy for the large number of restrictions on the parameter space.
- More importantly, **Our estimators, SE and PSE are FREE from Tuning Parameters, and easy to compute**.

Envoi

Shrinkage Versus LASSÉ

- The LASSÉ dominates the SE when the number of restrictions on parameters are small.
- Shrinkage estimators outshines the LASSÉ estimation strategy for the large number of restrictions on the parameter space.
- More importantly, **Our estimators, SE and PSE are FREE from Tuning Parameters, and easy to compute**.

Envoi

Shrinkage Versus LASSÉ

- The LASSÉ dominates the SE when the number of restrictions on parameters are small.
- Shrinkage estimators outshines the LASSÉ estimation strategy for the large number of restrictions on the parameter space.
- More importantly, **Our estimators, SE and PSE are FREE from Tuning Parameters, and easy to compute**.

References

Ahmed, S. E. (2001). Shrinkage estimation of regression coefficients from censored data with multiple observations. *In S.E. Ahmed and N. Reid (Eds.), Empirical Bayes and and Likelihood inference (pp. 103–120)*. NewYork:Springer.

S. Fallahpour, S. E. Ahmed and K. Doksum (2011). L1 Penalty and Shrinkage Estimation in Partially Linear Models with Random Coefficient autoregressive Errors. To appear in Applied Stochastic Models in Business and Industry.

E. Raheem, S. E. Ahmed and K. Doksum (2011). Absolute Penalty and Shrinkage Estimation in Partially Linear Models. To appear in Computational Statistics and Data Analysis.

Ahmed, S. E., A. A. Hussein, and S. Nkurunziza, (2010). Robust inference strategy in the presence of measurements error. Statistics and Probability Letters, 80, 726-732.

References

Ahmed, S. E. (2001). Shrinkage estimation of regression coefficients from censored data with multiple observations. *In S.E. Ahmed and N. Reid (Eds.), Empirical Bayes and and Likelihood inference (pp. 103–120)*. NewYork:Springer.

S. Fallahpour, S. E. Ahmed and K. Doksum (2011). L1 Penalty and Shrinkage Estimation in Partially Linear Models with Random Coefficient autoregressive Errors. To appear in Applied Stochastic Models in Business and Industry.

E. Raheem, S. E. Ahmed and K. Doksum (2011). Absolute Penalty and Shrinkage Estimation in Partially Linear Models. To appear in Computational Statistics and Data Analysis.

Ahmed, S. E., A. A. Hussein, and S. Nkurunziza, (2010). Robust inference strategy in the presence of measurements error. Statistics and Probability Letters, 80, 726-732.

References

Ahmed, S.E, A. I. Volodin and I. N. Volodin (2009). High order approximation for the coverage probability by confident set centered at the positive-part James-stein estimator. *Statistics and Probability Letters*, 79, 1823-1828 .

Hossain, S. K. A. Doksum and S.E. Ahmed (2009). Positive Shrinkage, Improved Pretest and Absolute Penalty Estimators in Partially Linear Models. Linear Algebra and its Applications, 430 2749-2761.

Ahmed, S. E., Saleh, A. I. Volodin and I. N. Volodin (2007). Asymptotic expansion of the coverage probability of James-Stein Estimators. *Journal of Theory of Probability and its Applications*, 51, 683-695.

Ahmed, S. E., A. A. Hussein and P. K. Sen (2006). Positive-part Shrinkage M-estimation in Linear Models. *Journal of Nonparametic Statistics*, 18, 401-415.

References

Ahmed, S.E, A. I. Volodin and I. N. Volodin (2009). High order approximation for the coverage probability by confident set centered at the positive-part James-stein estimator. *Statistics and Probability Letters*, 79, 1823-1828 .

Hossain, S. K. A. Doksum and S.E. Ahmed (2009). Positive Shrinkage, Improved Pretest and Absolute Penalty Estimators in Partially Linear Models. Linear Algebra and its Applications, 430 2749-2761.

Ahmed, S. E., Saleh, A. I. Volodin and I. N. Volodin (2007). Asymptotic expansion of the coverage probability of James-Stein Estimators. *Journal of Theory of Probability and its Applications*, 51, 683-695.

Ahmed, S. E., A. A. Hussein and P. K. Sen (2006). Positive-part Shrinkage M-estimation in Linear Models. *Journal of Nonparametic Statistics*, 18, 401-415.

References

Ahmed, S. E., Doksum, K. A. and Hossain, S. and You, J. (2007). Shrinkage, Pretest and Absolute Penalty Estimators in Partially Linear Models. *Australian and New Zealand Journal of Statistics*, 49, 435-454.

BuHamra, S. Al-Kandari, N. and Ahmed, S. E.(2007). Nonparametric Inference Strategies for the Quantile Functions Under Left Truncation and Right Censoring. *Journal of Nonparametic Statistics*, 9, 189 - 198.

Ahmed, S.E. and S. N. Liu (2008). Asymptotic simultaneous estimation of Poisson means. *Accepted for publication in Linear Algebra and its Applications*.

Park, M.-Y. and Hastie, T. (2007). An L_1 regularization-path algorithm for generalized linear models. Journal of Royal Statistical Society, Series B.