Determinants of Wages

Stat 825
Spring 2005

Xiuyun Zhang

Hongshu Chen

Outline

- Introduction
- Data Description
- Methods
- Results
- Conclusion

Introduction

- Wages: Pretty Important in Our Lives
- Determinants:
  - Education, Age, Experience, Race, Gender, Occupation...
- Gender Difference?

Data Description

- 1985 Current Population Survey (CPS 1985)
- Random Sample of 534 Persons
- 11 Variables, 4 Continuous, 7 Categorical
- No Missing Data

Data Description

Name	Explanation	Values
EDUCATION	Number of years of education.	$2 \sim 18$
LIVES_IN_SOUTH	Indicator variable for Southern Region	Yes, No
SEX	Indicator variable for sex	Female, Male
EXPERIENCE	Number of years of work experience.	$0 \sim 55$
UNION	Indicator variable for union membership	Yes, No
WAGE	Wage (dollars per hour).	$1 \sim 26.29$
AGE	Age (years).	$18 \sim 64$
RACE	Race.	(1=Other, 2=Hispanic, 3=White)
OCCUPATION	Occupational category	Management, Sales, Clerical, Service, Professional, Other
SECTOR	Sector	(0=Other, 1=Manufacturing, 2=Construction).
MARR	Marital Status	(0=Unmarried, 1=Married)

Model Description

- Linear Regression Model:

$$y_i = x_i\beta + \epsilon_i \tag{1}$$

where $\epsilon_i \sim Normal(0, \sigma^2)$ is independent normal distributed random error.

$$Y \mid \beta, \sigma^2, X \sim Normal(X\beta,\ \sigma^2 I) \tag{2}$$

- The parameters here are $(\beta,\ \sigma^2)$, and β is what we concern about.

Bayesian Regression Methods

- Set up Prior for $(\beta,\ \sigma^2)$
  - Ridge Regression & Lasso
  - G-Prior
  - $Normal$-$Gamma$ Prior

Ridge Regression & Lasso

- RR: Normal Prior for β

$$\hat{\beta}_{ridge} = \arg\min_{\tilde{\beta}} \Big((Y - X\tilde{\beta})'(Y - X\tilde{\beta}) + \lambda \tilde{\beta}'\tilde{\beta} \Big) \tag{3}$$

$$\beta \mid \sigma^2, Y, X \sim Normal(\hat{\beta}_{ridge}, \ (X'X + \lambda I)^{-1} X'X (X'X + \lambda I)^{-1} \sigma^2) \tag{4}$$

- Lasso: Laplace Prior for β

$$\begin{aligned} \hat{\beta}_{lasso} &= \arg\min_{\tilde{\beta}} (Y - X\tilde{\beta})'(Y - X\tilde{\beta}) \\ & s.t. \ \sum_{j=1}^{m} |\beta_j| \leq s \end{aligned} \tag{5}$$

G-Prior

X and Y are inputs and outputs in the training data set

1. Get $\hat{\beta}$ and $\hat{\sigma}$ by apply Ridge Regression or Lasso to the prior data
2. Randomly draw n observations $\tilde{X}$ from X
3. Generate response for the new observations by:

   $$\tilde{Y} = \tilde{X}\hat{\beta} + \boldsymbol{\epsilon}$$

   where $\boldsymbol{\epsilon}$ is the vector of independent random variables from $Normal(0,\ \hat{\sigma}^2)$
4. $\boldsymbol{X} = [X; \tilde{X}]$, $\boldsymbol{Y} = [Y; \tilde{Y}]$
5. $\hat{\beta}_g = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$

$\hat{\beta}_g$ is the estimate of β by using g-prior

Normal-Gamma Prior

- Suppose β and σ^2 are independent from each other, and $P(\log \sigma) \propto 1$, then $\sigma^2 \sim Gamma(0,\ 0)$, also assume $\beta \sim\ Normal(\mu_\beta,\ \Sigma)$
  - where μ_β, Σ are estimated from the posterior of β by fitting Ridge Regression on the prior data(equation 4)
- Use Gibbs Sampling Algorithm
- The joint distribution of $(Y,\ \beta)$ is:

$$\begin{pmatrix} Y \\ \beta \end{pmatrix} \mid \sigma^2, X \sim Normal\left(\begin{pmatrix} X\mu_\beta \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} X\Sigma X' + \sigma^2 I & X\Sigma \\ \Sigma X' & \Sigma \end{pmatrix}\right) \tag{6}$$

Normal-Gamma Prior

- Thus, the conditional posterior density of β given σ^2 is

$$\beta \mid \sigma^2, Y, X \sim Normal(\xi,\ \Phi) \tag{7}$$

where

$$\xi = \mu_\beta + \Sigma X'(X\Sigma X' + \sigma^2 I)^{-1}(Y - X\mu_\beta) \tag{8}$$

$$\Phi = \Sigma - \Sigma X'(X\Sigma X' + \sigma^2 I)^{-1} X\Sigma \tag{9}$$

Normal-Gamma Prior

- The joint posterior distribution of $(\beta,\ \sigma^2)$ is:

$$\begin{aligned} P(\beta, \sigma^2 \mid Y, X) &\propto P(Y \mid \beta, \sigma^2, X) P(\beta, \sigma^2 \mid X) \qquad (10) \\ &= Normal(X\beta,\ \sigma^2 I) \times Normal(\mu_\beta,\ \Sigma) \times \frac{1}{\sigma^2} \end{aligned}$$

  Fixing β, we can get something proportional to:

$$(\sigma^2)^{-n/2+1} \exp\left(-\frac{(Y - X\beta)'(Y - X\beta)}{2\sigma^2}\right)$$

-

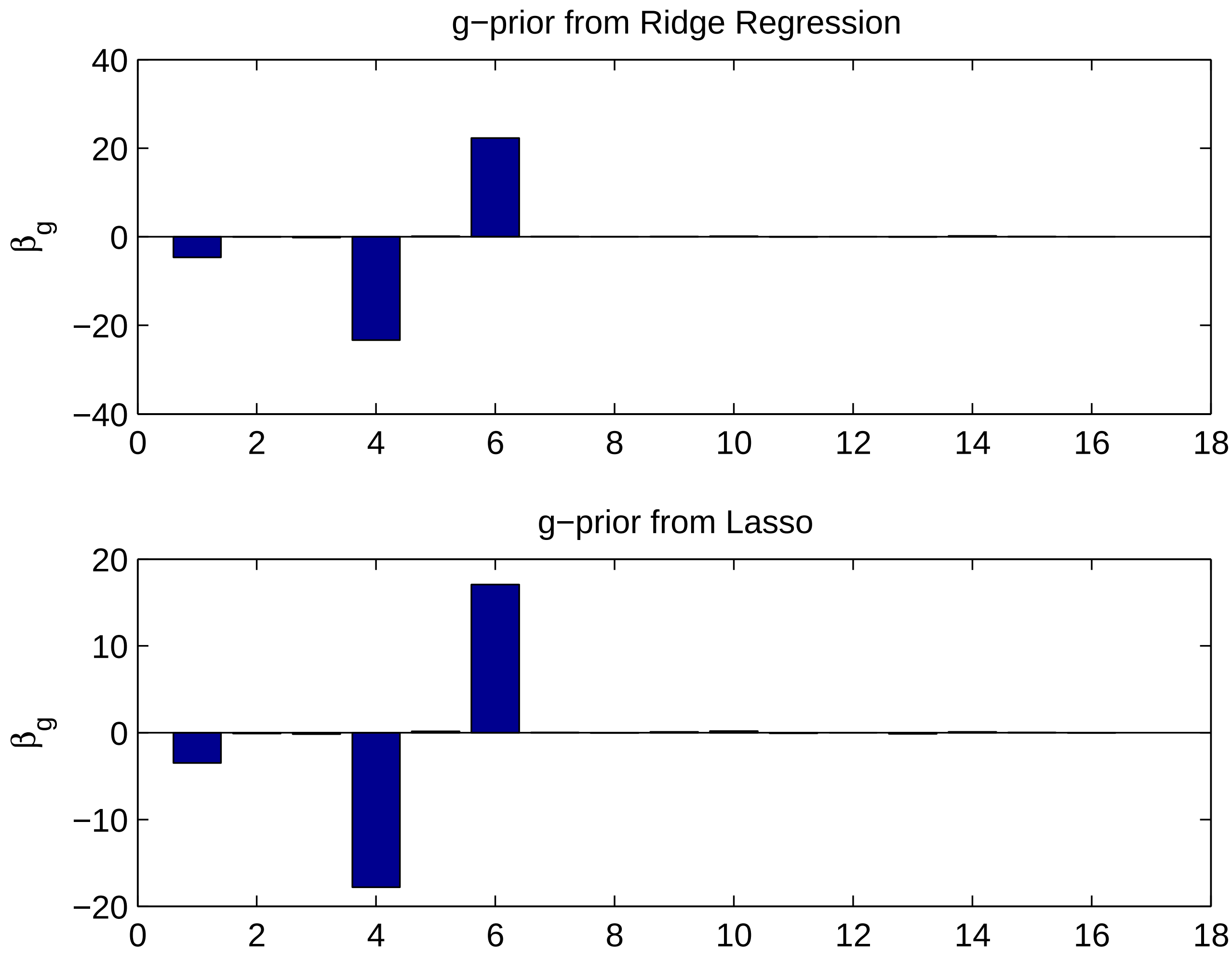
$$\sigma^2 \mid \beta, Y, X \sim Inv - Gamma\left(\frac{n}{2},\ \frac{(Y - X\beta)'(Y - X\beta)}{2}\right) \qquad (11)$$

Data Preprocessing

- Represent Categorical Variables in Several Binary Vairables
  - Race: Hispanic, White
  - Occupation: Management, Sales, Clerical, Service, Professional
  - Sector: Manufacturing, Construction
- Logarithm Wages
- Normalize All Variables
- Randomize and Divided into Three Parts: Prior(134), Training(300), Testing(100)

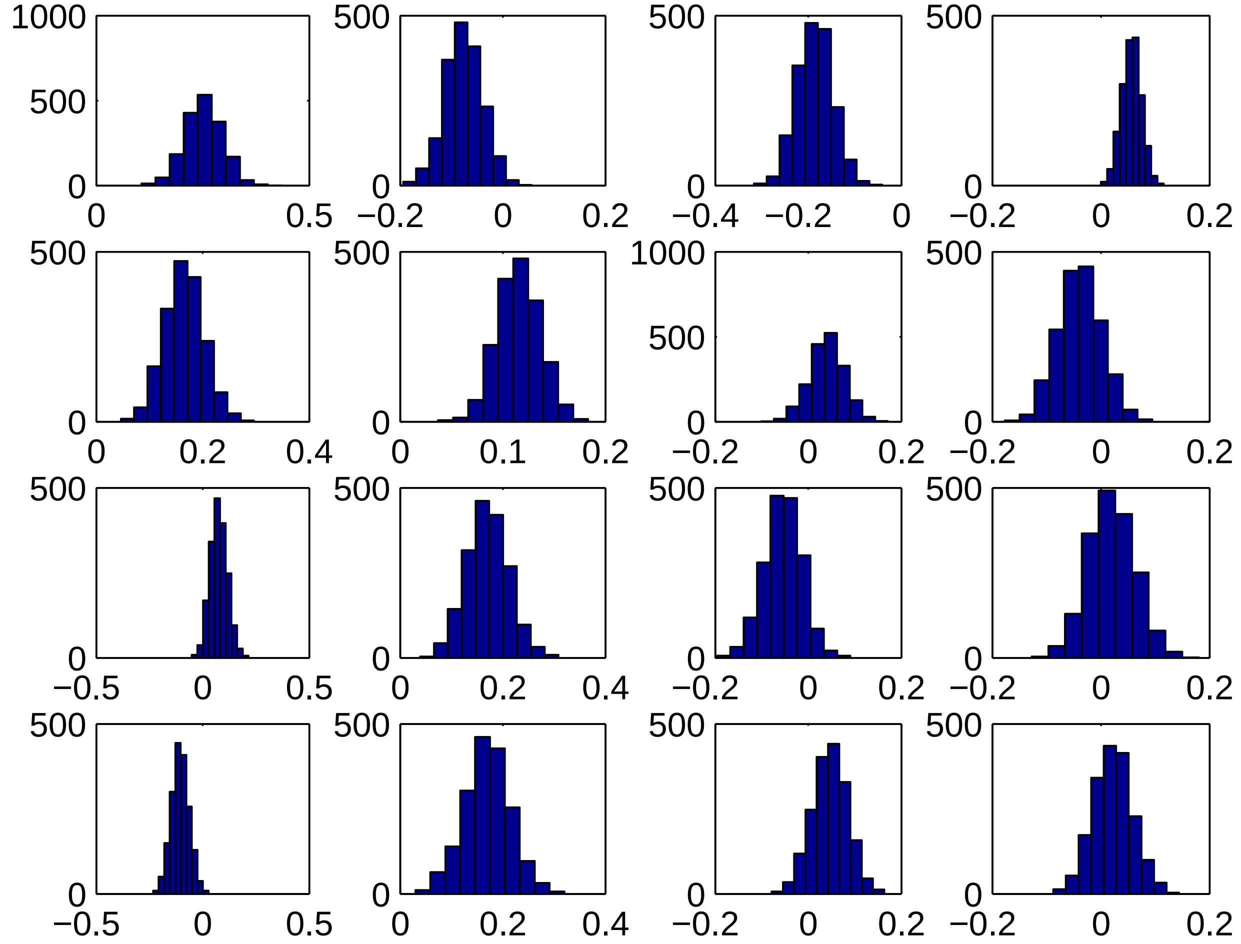
Results

Regression Parameters by Using G-Prior



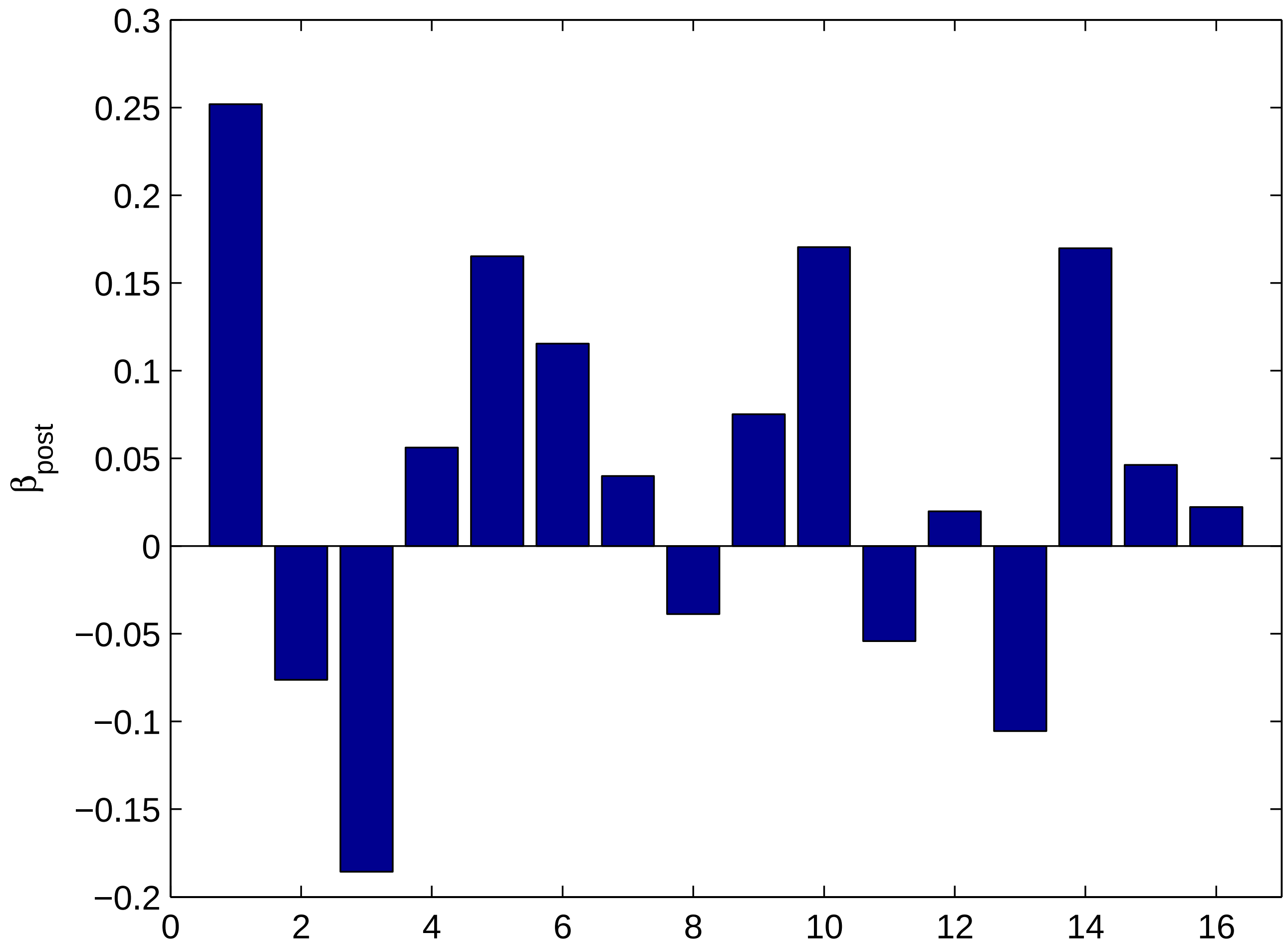
Results

Histograms of Samples of Regression Coefficients for Each Variable



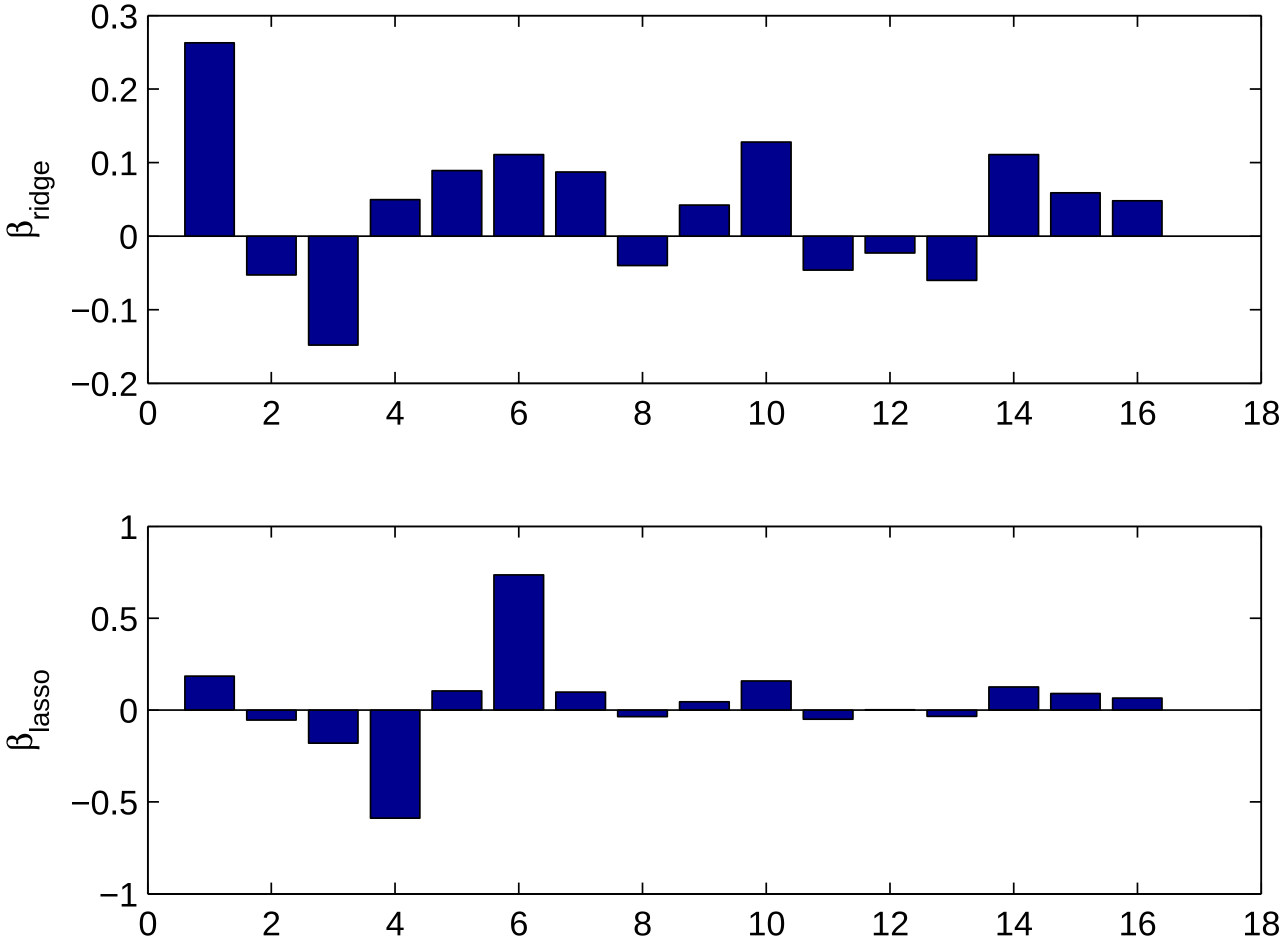
Results

Estimate of β by Using $Normal$-$Gamma$ Prior

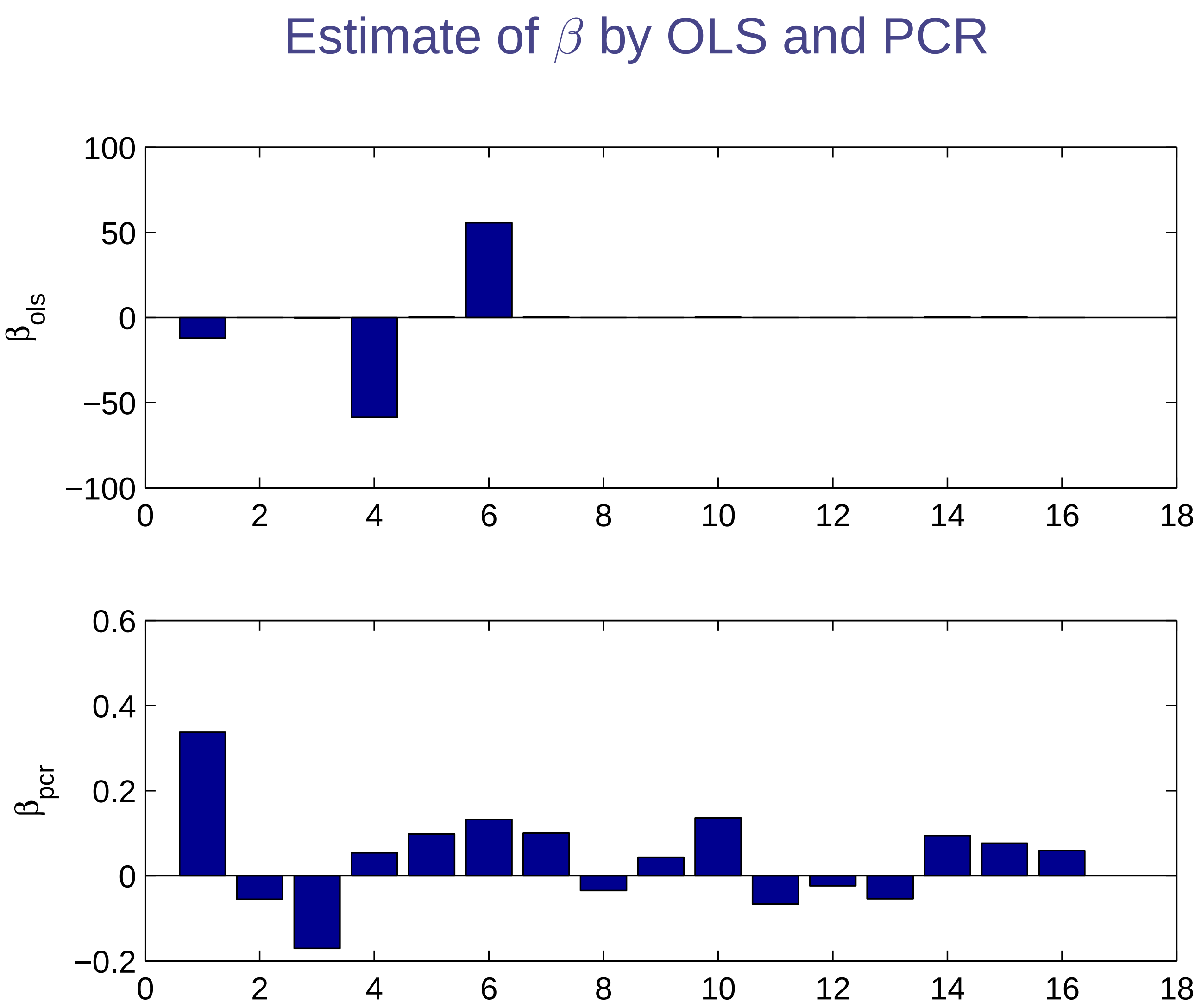


Results

Estimate of β by Ridge Regression and Lasso



Results



Estimate of β by OLS and PCR
β_{ols}
100
50
0
−50
−100
0
2
4
6
8
10
12
14
16
18
β_{pcr}
0.6
0.4
0.2
0
−0.2
0
2
4
6
8
10
12
14
16
18

Results

MSE of Testing Data

OLS	PCR	Ridge	Lasso	g-prior(ridge)	g-prior(lasso)	Normal-Gamma Prior
4.47	0.69	0.69	0.69	1.32	1.10	0.70

Results

95% Confidence Intervals of β by Using $Normal$-$Gamma$ Prior

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
2.5%	0.17	-0.15	-0.26	0.02	0.09	0.08	-0.03	-0.11
97.5%	0.34	-0.00	-0.11	0.09	0.24	0.16	0.11	0.04
includes 0		✓					✓	✓
	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
2.5%	0.00	0.09	-0.14	-0.06	-0.19	0.08	-0.03	-0.05
97.5%	0.16	0.25	0.02	0.10	-0.02	0.26	0.12	0.09
includes 0	✓		✓	✓			✓	✓

- Education, Sex, Union, Experience, Age, Management, Service, Professional

Results

Top 5 Predictors in Models

	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
Ridge	Education	Sex	Management	Age	Professional
	(0.26)	(-0.15)	(0.13)	(0.11)	(0.11)
$Normal$	Education	Sex	Management	Professional	Union
-$Gamma$	(0.25)	(-0.19)	(0.17)	(0.17)	(0.17)
PCR	Education	Sex	Management	Age	Marital Status
	(0.34)	(-0.17)	(0.14)	(0.13)	(0.10)

Conclusion

- Bayesian Methods are Better than OLS
- RR, Lasso, $Normal$-$Gamma$ Prior and PCR are Comparable
- Predictivity of Linear Model is Not Good
- Gender Discrimination: under same conditions, the females earn 83% the wages of the males
- Important Predictors: Education, Sex, Management, Age, Professioinal

Thanks!