Bayes Estimators & Ridge Regression

Readings ISLR 6

STA 521 Duke University

Merlise Clyde

October 28, 2019

Model

- Model:

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

linear reg model

Model

- Model:

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Assume that we have centered and rescaled $\mathbf{X}^o$ (original $\mathbf{X}$) so that

zero mean

variance 1

$$\mathbf{X}_j = \frac{\mathbf{X}_j^o - \bar{\mathbf{X}}_j^o}{\sqrt{\sum_i (X_{ij}^o - \bar{\mathbf{X}}_j^o)^2}}$$

Model

- Model:

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Assume that we have centered and rescaled $\mathbf{X}^o$ (original $\mathbf{X}$) so that

$$\mathbf{X}_j = \frac{\mathbf{X}_j^o - \bar{\mathbf{X}}_j^o}{\sqrt{\sum_i (X_{ij}^o - \bar{\mathbf{X}}_j^o)^2}}$$

- Equivalent to using ‘r scale(X)‘ divided by $\sqrt{n-1}$

Model

- Model:

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \epsilon$$

- Assume that we have centered and rescaled $\mathbf{X}^o$ (original $\mathbf{X}$) so that

$$\mathbf{X}_j = \frac{\mathbf{X}_j^o - \bar{\mathbf{X}}_j^o}{\sqrt{\sum_i (X_{ij}^o - \bar{\mathbf{X}}_j^o)^2}}$$

- Equivalent to using ‘r scale(X)‘ divided by $\sqrt{n-1}$
- $\mathbf{X}^T\mathbf{X} = \mathrm{Cor}(\mathbf{X})$ (correlation matrix of X) ← why are center and scale

Model

- Model:

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \epsilon$$

- Assume that we have centered and rescaled $\mathbf{X}^o$ (original $\mathbf{X}$) so that

$$\mathbf{X}_j = \frac{\mathbf{X}_j^o - \bar{\mathbf{X}}_j^o}{\sqrt{\sum_i (X_{ij}^o - \bar{\mathbf{X}}_j^o)^2}}$$

- Equivalent to using ‘r scale(X)‘ divided by $\sqrt{n-1}$
- $\mathbf{X}^T\mathbf{X} = \text{Cor}(\mathbf{X})$ (correlation matrix of X)
- eigenvalue decomposition $\mathbf{X}^T\mathbf{X} = \mathbf{U}\Lambda\mathbf{U}^T$

eigenvectors

$$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix} \qquad \begin{bmatrix} u_1 & \cdots & u_p \end{bmatrix}$$

eigenvalues

Model

- Model:

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\beta + \epsilon$$

- Assume that we have centered and rescaled $\mathbf{X}^o$ (original $\mathbf{X}$) so that

$$\mathbf{X}_j = \frac{\mathbf{X}_j^o - \bar{\mathbf{X}}_j^o}{\sqrt{\sum_i (X_{ij}^o - \bar{\mathbf{X}}_j^o)^2}}$$

- Equivalent to using ‘r scale(X)‘ divided by $\sqrt{n-1}$
- $\mathbf{X}^T\mathbf{X} = \mathrm{Cor}(\mathbf{X})$ (correlation matrix of X)
- eigenvalue decomposition $\mathbf{X}^T\mathbf{X} = \mathbf{U}\Lambda\mathbf{U}^T$
- if smallest eigen value is 0, $\mathbf{X}$ has columns that are linearly dependent! condition number
- problems if largest eigenvalue/smallest eigenvalue is large!

How Good are Various Estimators

Quadratic loss for estimating β using estimator $\mathbf{a}$

$$L(\beta, \mathbf{a}) = (\beta - \mathbf{a})^T(\beta - \mathbf{a})$$

How Good are Various Estimators

Quadratic loss for estimating β using estimator $\mathbf{a}$

$$L(\beta, \mathbf{a}) = (\beta - \mathbf{a})^T(\beta - \mathbf{a})$$

- Consider our expected loss (before we see the data) of taking an "action" $\mathbf{a}$

How Good are Various Estimators

Quadratic loss for estimating $\boldsymbol{\beta}$ using estimator $\mathbf{a}$

$$L(\boldsymbol{\beta}, \mathbf{a}) = (\boldsymbol{\beta} - \mathbf{a})^T(\boldsymbol{\beta} - \mathbf{a})$$

- Consider our expected loss (before we see the data) of taking an "action" $\mathbf{a}$
- Under OLS or the Independent Jeffreys Reference prior the Expected Mean Square Error

How Good are Various Estimators

Quadratic loss for estimating β using estimator $\mathbf{a}$

$$L(\beta, \mathbf{a}) = (\beta - \mathbf{a})^T(\beta - \mathbf{a})$$

- Consider our expected loss (before we see the data) of taking an "action" $\mathbf{a}$
- Under OLS or the Independent Jeffreys Reference prior the Expected Mean Square Error

$$\mathrm{E}_{\mathbf{Y}}[(\beta - \hat{\beta})^T(\beta - \hat{\beta}) \quad = \quad \sigma^2 \mathrm{tr}[(\mathbf{X}^T\mathbf{X})^{-1}]$$

$$\mathbb{E}_Y\left[(\underset{\text{true}}{\beta} - \underset{\text{OLS est.}}{\hat{\beta}})^T(\beta - \hat{\beta})\right] = \sigma^2 \mathrm{tr}[(X^TX)^{-1}]$$

$$\hat{\beta} := \hat{\beta}(Y)$$

How Good are Various Estimators

Quadratic loss for estimating β using estimator $\mathbf{a}$

$$L(\beta, \mathbf{a}) = (\beta - \mathbf{a})^T(\beta - \mathbf{a})$$

- Consider our expected loss (before we see the data) of taking an "action" $\mathbf{a}$
- Under OLS or the Independent Jeffreys Reference prior the Expected Mean Square Error

$$\begin{aligned} \mathsf{E}_{\mathbf{Y}}[(\beta - \hat{\beta})^T(\beta - \hat{\beta}) &= \sigma^2 \mathrm{tr}[(\mathbf{X}^T\mathbf{X})^{-1}] \\ &= \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1} \end{aligned}$$

small eigen values explode

How Good are Various Estimators

Quadratic loss for estimating $\boldsymbol{\beta}$ using estimator $\mathbf{a}$

$$L(\boldsymbol{\beta}, \mathbf{a}) = (\boldsymbol{\beta} - \mathbf{a})^T(\boldsymbol{\beta} - \mathbf{a})$$

- Consider our expected loss (before we see the data) of taking an "action" $\mathbf{a}$
- Under OLS or the Independent Jeffreys Reference prior the Expected Mean Square Error

$$\begin{aligned} \mathsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) &= \sigma^2 \mathrm{tr}[(\mathbf{X}^T\mathbf{X})^{-1}] \\ &= \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1} \end{aligned}$$

- If smallest $\lambda_j \to 0$ then RMSE $\to \infty$

Problems

Estimates:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

OLS

or with g-prior

$$\hat{\beta} = \frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

g-prior

may be unstable without variable selection.
Solutions:

- remove redundant variables (model selection) (AIC, BIC, other approches) 2^p models combinatorial hard problem even with MCMC
- add constant to $\mathbf{X}^T\mathbf{X}$: $\boxed{\tilde{\beta} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}}$ to stabilise eigenvalues - alternative shrinkage estimator/prior

OLS

$(X^TX)^{-1}$

$p \times p$ rank

Ridge

$(X^TX + kI)^{-1}$

X is $n \times p$

$p >> n$

pxp rank n pxp $k > 0$

Independent Prior

- Independent Jeffreys Reference prior $p(\beta_0, \phi) \propto \phi^{-1}$
- Prior Distribution on

$$\boldsymbol{\beta} \mid \phi, \beta_0, k \sim \mathrm{N}(\mathbf{0}_p, \frac{1}{\phi k}\mathbf{I}_p)$$

Independent Prior

- Independent Jeffreys Reference prior $p(\beta_0, \phi) \propto \phi^{-1}$
- Prior Distribution on

$$\boldsymbol{\beta} \mid \phi, \beta_0, k \sim \mathrm{N}(\mathbf{0}_p, \frac{1}{\phi k}\mathbf{I}_p)$$

shrinkage
regularization

- log likelihood (integrated) for $\boldsymbol{\beta}$ plus prior

$$-\frac{\phi}{2}\left(\underbrace{\|\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|^2}_{\text{OLS}} + k\|\boldsymbol{\beta}\|^2\right)$$

quadratic
penalty
on
β

- Posterior mean

$$\mathbf{b}_n = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}$$

Ridge
regression

Independent Prior

- Independent Jeffreys Reference prior $p(\beta_0, \phi) \propto \phi^{-1}$
- Prior Distribution on

$$\boldsymbol{\beta} \mid \phi, \beta_0, k \sim \mathrm{N}(\mathbf{0}_p, \frac{1}{\phi k}\mathbf{I}_p)$$

- log likelihood (integrated) for $\boldsymbol{\beta}$ plus prior

$$-\frac{\phi}{2}\left(\|\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|^2 + k\|\boldsymbol{\beta}\|^2\right)$$

- Posterior mean

$$\mathbf{b}_n = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}$$

ridge regression

- importance of standardizing ← our probabistic, Bayesian formulation is consistent with ridge

Independent Prior

- Independent Jeffreys Reference prior $p(\beta_0, \phi) \propto \phi^{-1}$
- Prior Distribution on

$$\boldsymbol{\beta} \mid \phi, \beta_0, k \sim \mathrm{N}(\mathbf{0}_p, \frac{1}{\phi k}\mathbf{I}_p)$$

- log likelihood (integrated) for $\boldsymbol{\beta}$ plus prior

$$-\frac{\phi}{2}\left(\|\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|^2 + k\|\boldsymbol{\beta}\|^2\right)$$

- Posterior mean

$$\mathbf{b}_n = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}$$

- importance of standardizing
- Choice of k in practice?
- $k = 0$ OLS
- $k = \infty$ estimates are $\mathbf{0}$ (intercept only)

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Control how large coefficients may grow

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Control how large coefficients may grow

P1

$$\min_{\beta} (\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{1}\bar{Y} - \mathbf{X}\beta)$$

subject to

$$\sum \beta_j^2 \leq t$$

P2

$$\min_{\beta} (Y - 1\bar{Y} - X\beta)^T (Y - 1\bar{Y} - X\beta) + k\|\beta\|^2$$

Ridge

for some K ∃ t such that β's for P1 + P2 are identical

Alternative Motivation

- If $\hat{\boldsymbol{\beta}}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Control how large coefficients may grow

$$\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{1}\bar{Y} - \mathbf{X}\beta)$$

subject to

$$\sum \beta_j^2 \leq t$$

- Equivalent Quadratic Programming Problem

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y}^c - \mathbf{X}^c \beta\|^2 + k\|\boldsymbol{\beta}\|^2$$

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Control how large coefficients may grow

p 1

$$\min_{\beta}(\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{1}\bar{Y} - \mathbf{X}\beta)$$

subject to

$$\sum \beta_j^2 \leq t$$

- Equivalent Quadratic Programming Problem

$$\min_{\beta} \|\mathbf{Y}^c - \mathbf{X}^c\beta\|^2 + k\|\beta\|^2$$

penalize likelihood

penalty can be thought of as a prior

- "penalized" likelihood

Alternative Motivation

- If $\hat{\boldsymbol{\beta}}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Control how large coefficients may grow

$$\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{1}\bar{Y} - \mathbf{X}\beta)$$

subject to

$$\sum \beta_j^2 \leq t$$

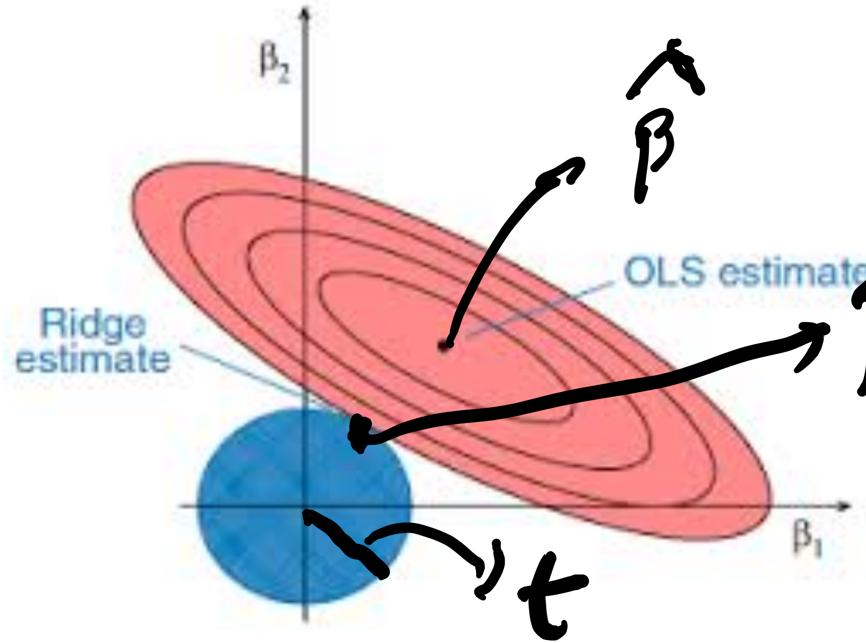
- Equivalent Quadratic Programming Problem

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y}^c - \mathbf{X}^c\boldsymbol{\beta}\|^2 + k\|\boldsymbol{\beta}\|^2$$

- "penalized" likelihood
- Ridge Regression

Geometry

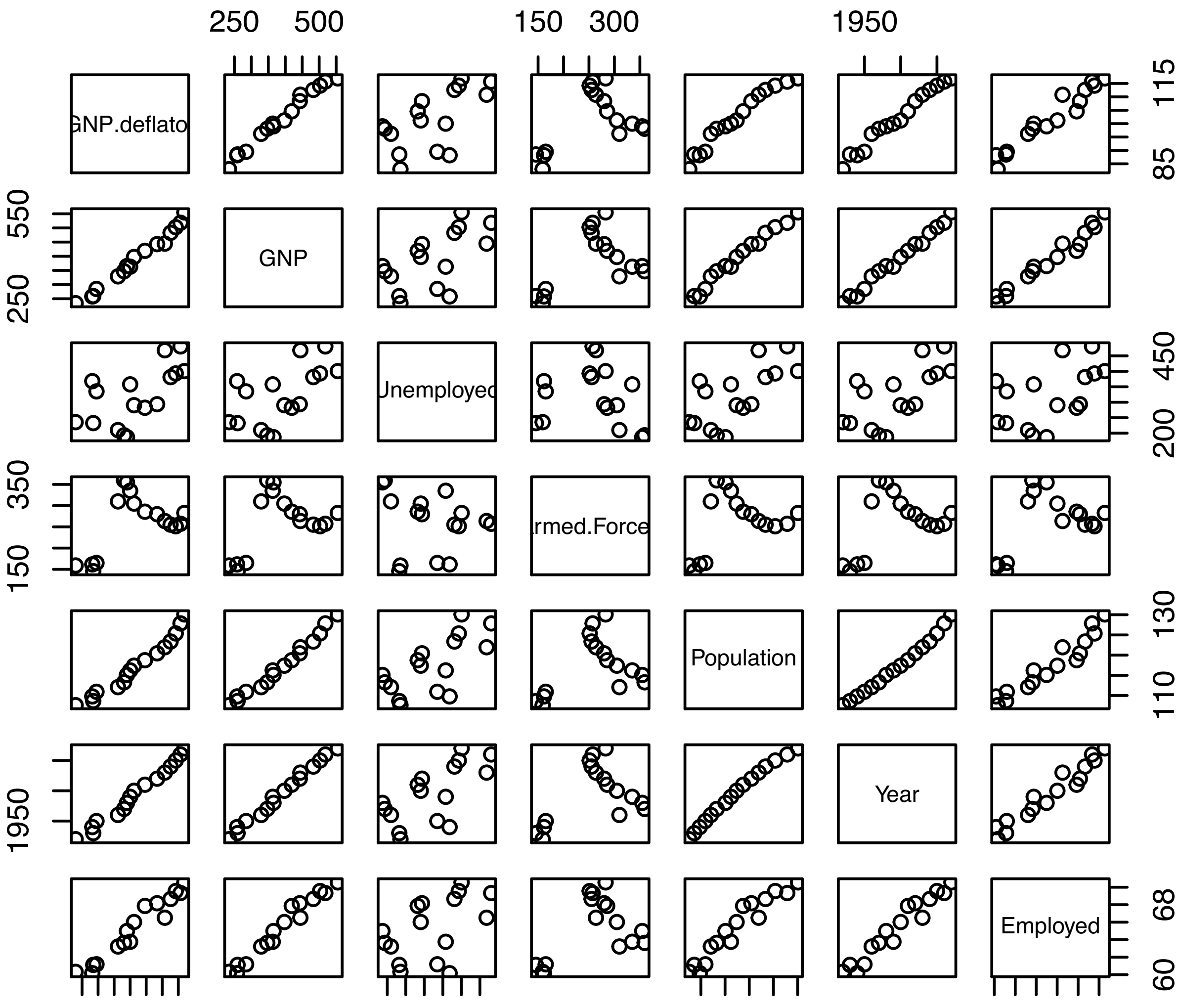
(P1) picture



[1]

[1] onlinecourses.science.pse.edu

Longley Data: library(MASS); data(longley)



OLS

```
> longley.lm = lm(Employed ~ ., data=longley)
> summary(longley.lm)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.482e+03  8.904e+02  -3.911 0.003560 **
GNP.deflator  1.506e-02  8.492e-02   0.177 0.863141
GNP          -3.582e-02  3.349e-02  -1.070 0.312681
Unemployed   -2.020e-02  4.884e-03  -4.136 0.002535 **
Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
Population   -5.110e-02  2.261e-01  -0.226 0.826212
Year          1.829e+00  4.555e-01   4.016 0.003037 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3049 on 9 degrees of freedom
Multiple R-squared: 0.9955,^^IAdjusted R-squared: 0.9925
F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

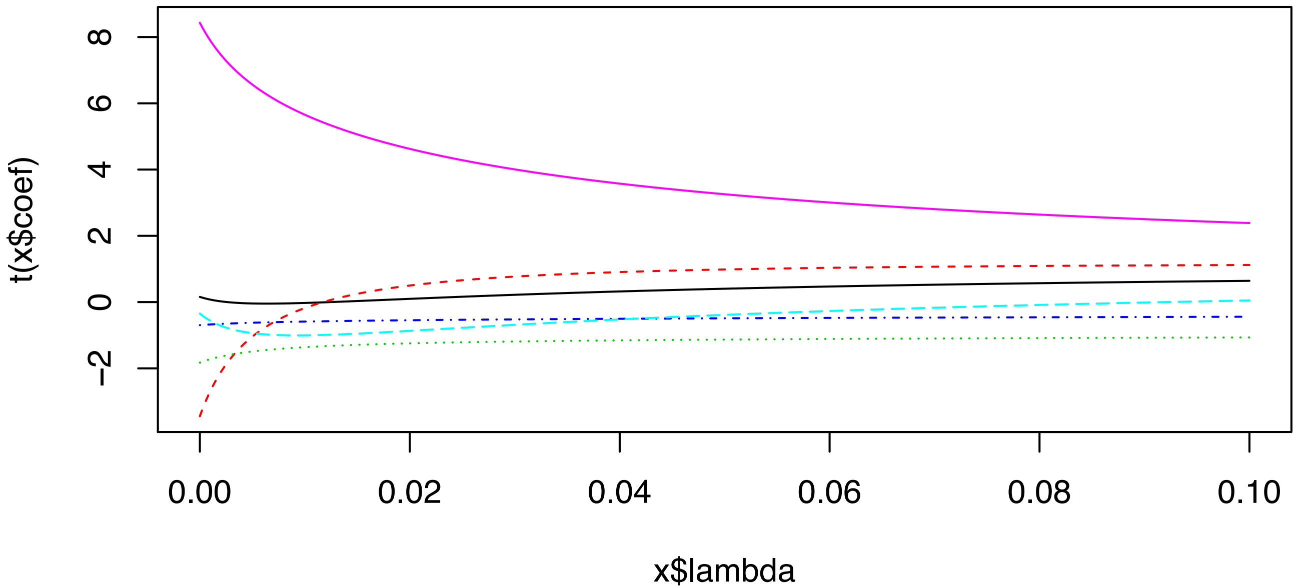
Ridge Regression

```
# from library MASS
longley.ridge = lm.ridge(Employed ~ ., data=longley,
                         lambda=seq(0, 0.1, 0.0001))
# lambda = k in notes

summary(longley.ridge)

##        Length Class  Mode
## coef   6006   -none- numeric
## scales    6   -none- numeric
## Inter     1   -none- numeric
## lambda 1001   -none- numeric
## ym        1   -none- numeric
## xm        6   -none- numeric
## GCV    1001   -none- numeric
## kHKB      1   -none- numeric
## kLW       1   -none- numeric
```

Ridge Trace Plot



Choice of k

cross-validation regularization - sweep K from $0 \to \infty$

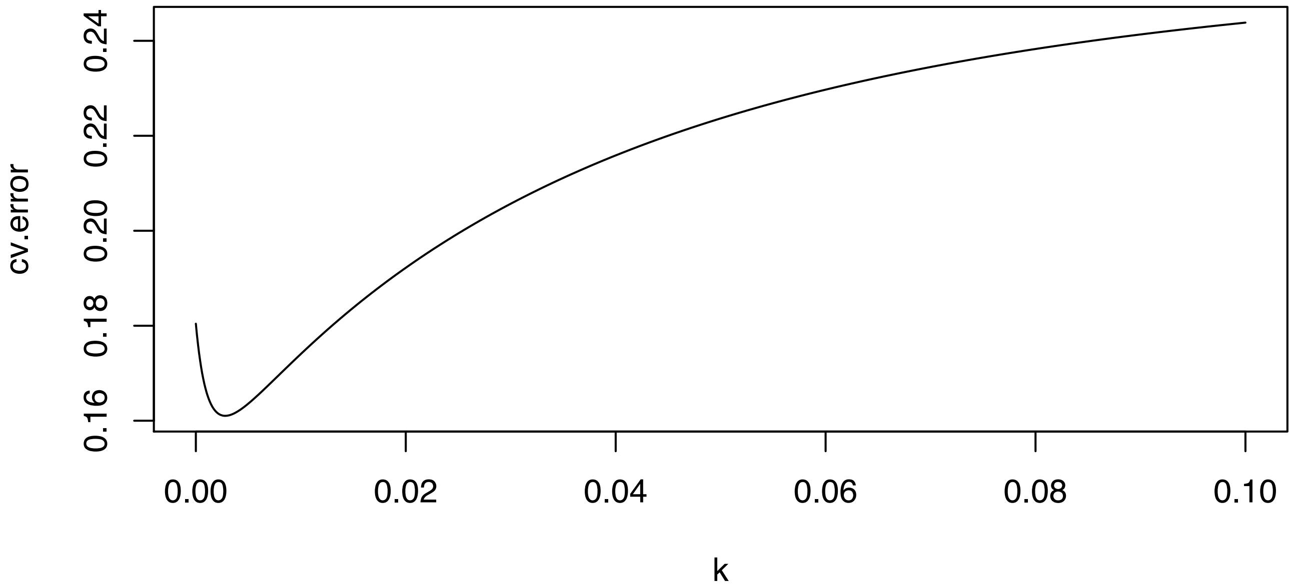
```
k = seq(0, 0.1, 0.0001)
n.k = length(k); n = nrow(longley)
cv.lambda = matrix(NA, n, n.k)

rmse.ridge = function(data, i, j, k) {
  m.ridge = lm.ridge(Employed ~ ., data = data, lambda=k[j]
                     subset = -i)
  yhat = scale(data[i,1:6, drop=F],center = m.ridge$xm,
                  scale = m.ridge$scales) %*%
           m.ridge$coef + m.ridge$ym
(yhat - data$Employed[i])^2
}

for (i  in 1:n)  {
  for (j in 1:n.k) {
    cv.lambda[i,j] = rmse.ridge(longley, i, j, k)
  }
}
```

Cross Validation Error

```
cv.error = apply(cv.lambda, 2, mean)
plot(k, cv.error, type="l")
```



Best $k = 0.0028$

Generalized Cross-validation

```
select(lm.ridge(Employed ~ ., data=longley,
         lambda=seq(0, 0.1, 0.0001)))

## modified HKB estimator is 0.004275357
## modified L-W estimator is 0.03229531
## smallest value of GCV  at 0.0028

best.k = longley.ridge$lambda[which.min(longley.ridge$GCV)]
longley.RReg = lm.ridge(Employed ~ ., data=longley,
                        lambda=best.k)
coef(longley.RReg)

##                GNP.deflator           GNP    Unemployed  Arme
## -2.950348e+03 -5.381450e-04 -1.822639e-02 -1.761107e-02 -9.60
##    Population          Year
## -1.185103e-01  1.557856e+00
```

Priors on k

$\mathbf{X}$ is centered and standardized

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Hierarchical prior

Priors on k

$\mathbf{X}$ is centered and standardized

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Hierarchical prior

- $p(\beta_0, \phi \mid \boldsymbol{\beta}, \kappa) \propto \phi^{-1}$

Priors on k

$\mathbf{X}$ is centered and standardized

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Hierarchical prior

- $p(\beta_0, \phi \mid \boldsymbol{\beta}, \kappa) \propto \phi^{-1}$
- $\boldsymbol{\beta} \mid \phi, \kappa \sim \mathsf{N}(\mathbf{0}, \mathbf{I}(\phi\kappa)^{-1})$

Priors on k

$\mathbf{X}$ is centered and standardized

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Hierarchical prior

- $p(\beta_0, \phi \mid \boldsymbol{\beta}, \kappa) \propto \phi^{-1}$
- $\boldsymbol{\beta} \mid \phi, \kappa \sim \mathsf{N}(\mathbf{0}, \mathbf{I}(\phi\kappa)^{-1})$
- prior on κ?

$N(0, \phi^{-1} \kappa^{-1} I)$

r.v. κ

fixed

Priors on k

$\mathbf{X}$ is centered and standardized

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Hierarchical prior

- $p(\beta_0, \phi \mid \boldsymbol{\beta}, \kappa) \propto \phi^{-1}$
- $\boldsymbol{\beta} \mid \phi, \kappa \sim \mathrm{N}(\mathbf{0}, \mathbf{I}(\phi\kappa)^{-1})$
- prior on κ?
- Take

$$\kappa \mid \phi \sim \mathrm{Gamma}(1/2, 1/2)$$

Priors on k

$\mathbf{X}$ is centered and standardized

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Hierarchical prior

- $p(\beta_0, \phi \mid \boldsymbol{\beta}, \kappa) \propto \phi^{-1}$
- $\boldsymbol{\beta} \mid \phi, \kappa \sim \mathsf{N}(\mathbf{0}, \mathbf{I}(\phi\kappa)^{-1})$
- prior on κ?
- Take
$$\kappa \mid \phi \sim \mathsf{Gamma}(1/2, 1/2)$$
- What is induced prior on $\boldsymbol{\beta} \mid \phi$?

Posterior Distributions

Joint Distribution

- $\beta_0, \boldsymbol{\beta}, \phi \mid \kappa, \mathbf{Y}$ Normal-Gamma family given $\mathbf{Y}$ and κ

Posterior Distributions

Joint Distribution

- $\beta_0, \boldsymbol{\beta}, \phi \mid \kappa, \mathbf{Y}$ Normal-Gamma family given $\mathbf{Y}$ and κ
- $\kappa \mid \mathbf{Y}$ not tractable

Posterior Distributions

Joint Distribution

- $\beta_0, \boldsymbol{\beta}, \phi \mid \kappa, \mathbf{Y}$ Normal-Gamma family given $\mathbf{Y}$ and κ
- $\kappa \mid \mathbf{Y}$ not tractable

Obtain marginal for $\boldsymbol{\beta}$ via MCMC

Posterior Distributions

Joint Distribution

- $\beta_0, \boldsymbol{\beta}, \phi \mid \kappa, \mathbf{Y}$ Normal-Gamma family given $\mathbf{Y}$ and κ
- $\kappa \mid \mathbf{Y}$ not tractable

Obtain marginal for $\boldsymbol{\beta}$ via MCMC

Pick initial values $\beta_0^{(0)}, \boldsymbol{\beta}^{(0)}, \phi^{(0)}$,

Posterior Distributions

Joint Distribution

- $\beta_0, \boldsymbol{\beta}, \phi \mid \kappa, \mathbf{Y}$ Normal-Gamma family given $\mathbf{Y}$ and κ
- $\kappa \mid \mathbf{Y}$ not tractable

Obtain marginal for $\boldsymbol{\beta}$ via MCMC

Pick initial values $\beta_0^{(0)}, \boldsymbol{\beta}^{(0)}, \phi^{(0)}$,

Set $t = 1$

1. Sample $\kappa^{(t)} \sim p(\kappa \mid \beta_0^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \phi^{(t-1)}, \mathbf{Y})$

You get a post. dist over shrinkage parameter

Freq. or proc. you get a $\hat{K}$

$K \mid Y, \beta_0, \beta, \phi$

Posterior Distributions

Joint Distribution

- $\beta_0, \boldsymbol{\beta}, \phi \mid \kappa, \mathbf{Y}$ Normal-Gamma family given $\mathbf{Y}$ and κ
- $\kappa \mid \mathbf{Y}$ not tractable

Obtain marginal for $\boldsymbol{\beta}$ via MCMC

Pick initial values $\beta_0^{(0)}, \boldsymbol{\beta}^{(0)}, \phi^{(0)}$,

Set $t = 1$

1. Sample $\kappa^{(t)} \sim p(\kappa \mid \beta_0^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \phi^{(t-1)}, \mathbf{Y})$
2. Sample $\beta_0^{(t)}, \boldsymbol{\beta}^{(t)}, \phi^{(t)} \mid \kappa(t), \mathbf{Y}$

Posterior Distributions

Joint Distribution

- $\beta_0, \boldsymbol{\beta}, \phi \mid \kappa, \mathbf{Y}$ Normal-Gamma family given $\mathbf{Y}$ and κ
- $\kappa \mid \mathbf{Y}$ not tractable

Obtain marginal for $\boldsymbol{\beta}$ via MCMC

Pick initial values $\beta_0^{(0)}, \boldsymbol{\beta}^{(0)}, \phi^{(0)}$,

Set $t = 1$

1. Sample $\kappa^{(t)} \sim p(\kappa \mid \beta_0^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \phi^{(t-1)}, \mathbf{Y})$
2. Sample $\beta_0^{(t)}, \boldsymbol{\beta}^{(t)}, \phi^{(t)} \mid \kappa(t), \mathbf{Y}$
3. Set $t = t + 1$ and repeat until $t > T$

Posterior Distributions

next tuesday when we do data analysis

Joint Distribution

- $\beta_0, \boldsymbol{\beta}, \phi \mid \kappa, \mathbf{Y}$ Normal-Gamma family given $\mathbf{Y}$ and κ
- $\kappa \mid \mathbf{Y}$ not tractable

Obtain marginal for $\boldsymbol{\beta}$ via MCMC

Pick initial values $\beta_0^{(0)}, \boldsymbol{\beta}^{(0)}, \phi^{(0)}$,

Set $t = 1$

1. Sample $\kappa^{(t)} \sim p(\kappa \mid \beta_0^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \phi^{(t-1)}, \mathbf{Y})$
2. Sample $\beta_0^{(t)}, \boldsymbol{\beta}^{(t)}, \phi^{(t)} \mid \kappa(t), \mathbf{Y}$
3. Set $t = t + 1$ and repeat until $t > T$

Use Samples $\beta_0^{(t)}, \boldsymbol{\beta}^{(t)}, \phi^{(t)}, \kappa^{(t)}$ for $t = B, \ldots, T$ for inference

JAGS

JAGS = Just Another Gibbs Sampler

JAGS

JAGS = Just Another Gibbs Sampler

- scripting language to express sampling models and priors

JAGS

JAGS = Just Another Gibbs Sampler

- scripting language to express sampling models and priors
- ”derives” full conditional distributions

JAGS

JAGS = Just Another Gibbs Sampler

- scripting language to express sampling models and priors
- "derives" full conditional distributions
- integrates with R

JAGS

JAGS = Just Another Gibbs Sampler

- scripting language to express sampling models and priors
- "derives" full conditional distributions
- integrates with R
- typically faster than interpreted R code

JAGS - because easy to implement Bayesian models

JAGS = Just Another Gibbs Sampler

- scripting language to express sampling models and priors
- "derives" full conditional distributions
- integrates with R
- typically faster than interpreted R code
- accounts for uncertainty about k

JAGS

STAN - more modern
faster, more
flexible JAGS

JAGS = Just Another Gibbs Sampler

- scripting language to express sampling models and priors
- ”derives” full conditional distributions
- integrates with R
- typically faster than interpreted R code
- accounts for uncertainty about k

How would you compare Bayes predictions with Ridge with Cross-validation?