Shrinkage Estimation & Ridge Regression

Readings Chapter 15 Christensen

STA721 Linear Models Duke University

Merlise Clyde

October 16, 2019

How Good are Bayes Estimators?

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

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

How Good are Bayes 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” $\mathbf{a}$ (the estimate that we report)

How Good are Bayes 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" $\mathbf{a}$ (the estimate that we report)
- Under OLS or the Reference prior the Expected Mean Square Error

How Good are Bayes 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" $\mathbf{a}$ (the estimate that we report)
- Under OLS or the Reference prior the Expected Mean Square Error

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

How Good are Bayes 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" $\mathbf{a}$ (the estimate that we report)
- Under OLS or the Reference prior the Expected Mean Square Error

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

where λ_j are eigenvalues of $\mathbf{X}^T\mathbf{X}$.

How Good are Bayes 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" $\mathbf{a}$ (the estimate that we report)
- Under OLS or the Reference prior the Expected Mean Square Error

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

  where λ_j are eigenvalues of $\mathbf{X}^T\mathbf{X}$.
- If smallest $\lambda_j \to 0$ then MSE $\to \infty$
- Note: estimate is unbiased!

Is the g-prior better?

-
- Explore Frequentist properties of using a Bayesian estimator

$$\mathsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)$$

but now $\hat{\boldsymbol{\beta}}_g = g/(1+g)\hat{\boldsymbol{\beta}}$

Is the g-prior better?

-
- Explore Frequentist properties of using a Bayesian estimator

$$\mathsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)$$

  but now $\hat{\boldsymbol{\beta}}_g = g/(1+g)\hat{\boldsymbol{\beta}}$
- Sampling distribution of $\hat{\boldsymbol{\beta}}_g = \frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

Is the g-prior better?

-
- Explore Frequentist properties of using a Bayesian estimator

$$\mathsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)$$

  but now $\hat{\boldsymbol{\beta}}_g = g/(1+g)\hat{\boldsymbol{\beta}}$
- Sampling distribution of $\hat{\boldsymbol{\beta}}_g = \frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$
- HW: show that there is a value of g prior such that the g-prior is always better than the Reference prior/OLS
- Potential problem: MSE also blows up if smallest eigenvalue goes to zero!

Estimator Properties

- Bias

Estimator Properties

- Bias
- Variance

Estimator Properties

- Bias
- Variance
- MSE $=$ Bias^2 $+$ Variance (multivariate analogs)

Estimator Properties

- Bias
- Variance
- MSE $=$ Bias^2 $+$ Variance (multivariate analogs)
- Problems with OLS, g-priors & mixtures of g-priors with collinearity

Estimator Properties

- Bias
- Variance
- MSE $=$ Bias^2 $+$ Variance (multivariate analogs)
- Problems with OLS, g-priors & mixtures of g-priors with collinearity
- Solutions:

Estimator Properties

- Bias
- Variance
- MSE $=$ Bias^2 $+$ Variance (multivariate analogs)
- Problems with OLS, g-priors & mixtures of g-priors with collinearity
- Solutions:
  - removal of terms

Estimator Properties

- Bias
- Variance
- MSE $=$ Bias^2 $+$ Variance (multivariate analogs)
- Problems with OLS, g-priors & mixtures of g-priors with collinearity
- Solutions:
  - removal of terms
  - other shrinkage estimators

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)
- Write $\mathbf{X} = \mathbf{U}_p L \mathbf{V}^T$ Singular Value Decomposition

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)
- Write $\mathbf{X} = \mathbf{U}_p L \mathbf{V}^T$ Singular Value Decomposition where $\mathbf{U}_p^T \mathbf{U}_p = \mathbf{I}_p$ and $\mathbf{V}$ is $p \times p$ orthogonal matrix, L is diagonal

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)
- Write $\mathbf{X} = \mathbf{U}_p L \mathbf{V}^T$ Singular Value Decomposition where $\mathbf{U}_p^T \mathbf{U}_p = \mathbf{I}_p$ and $\mathbf{V}$ is $p \times p$ orthogonal matrix, L is diagonal

$$\mathbf{Y} = \mathbf{1}\alpha + \mathbf{U}_p L \mathbf{V}^T \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)
- Write $\mathbf{X} = \mathbf{U}_p L \mathbf{V}^T$ Singular Value Decomposition where $\mathbf{U}_p^T \mathbf{U}_p = \mathbf{I}_p$ and $\mathbf{V}$ is $p \times p$ orthogonal matrix, L is diagonal

$$\mathbf{Y} = \mathbf{1}\alpha + \mathbf{U}_p L \mathbf{V}^T \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Let $\gamma = \mathbf{V}^T \boldsymbol{\beta}$ and create $\mathbf{U}$ an $n \times n$ orthogonal matrix

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)
- Write $\mathbf{X} = \mathbf{U}_p L \mathbf{V}^T$ Singular Value Decomposition where $\mathbf{U}_p^T \mathbf{U}_p = \mathbf{I}_p$ and $\mathbf{V}$ is $p \times p$ orthogonal matrix, L is diagonal

$$\mathbf{Y} = \mathbf{1}\alpha + \mathbf{U}_p L \mathbf{V}^T \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Let $\gamma = \mathbf{V}^T \boldsymbol{\beta}$ and create $\mathbf{U}$ an $n \times n$ orthogonal matrix

$$\mathbf{U} = [\mathbf{U}_0 \,|\, \mathbf{U}_p \,|\, \mathbf{U}_{n-p-1}]$$

where $\mathbf{U}_0 = \mathbf{1}/\sqrt{n}$

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)
- Write $\mathbf{X} = \mathbf{U}_p L \mathbf{V}^T$ Singular Value Decomposition where $\mathbf{U}_p^T\mathbf{U}_p = \mathbf{I}_p$ and $\mathbf{V}$ is $p \times p$ orthogonal matrix, L is diagonal

$$\mathbf{Y} = \mathbf{1}\alpha + \mathbf{U}_p L \mathbf{V}^T \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Let $\gamma = \mathbf{V}^T\boldsymbol{\beta}$ and create $\mathbf{U}$ an $n \times n$ orthogonal matrix

$$\mathbf{U} = [\mathbf{U}_0 \,|\, \mathbf{U}_p \,|\, \mathbf{U}_{n-p-1}]$$

  where $\mathbf{U}_0 = \mathbf{1}/\sqrt{n}$
- $\mathbf{U}_0^T\mathbf{U}_p = 0$, $\mathbf{U}_0^T\mathbf{U}_{n-p-1} = 0$ and $\mathbf{U}_p^T\mathbf{U}_{n-p-1} = 0$ (orthogonal columns)

Orthogonal Regression

Rotate by multiplyting by $\mathbf{U}^T$:

$$\mathbf{U}^T\mathbf{Y} \;=\; \mathbf{U}^T\mathbf{1}\alpha + \mathbf{U}^T\mathbf{U}_p L\mathbf{V}^T\boldsymbol{\beta} + \mathbf{U}^T\boldsymbol{\epsilon}$$

Orthogonal Regression

Rotate by multiplyting by $\mathbf{U}^T$:

$$
\begin{aligned}
\mathbf{U}^T\mathbf{Y} &= \mathbf{U}^T\mathbf{1}\alpha + \mathbf{U}^T\mathbf{U}_p L \mathbf{V}^T\boldsymbol{\beta} + \mathbf{U}^T\boldsymbol{\epsilon} \\
\mathbf{Y}^* &= \begin{bmatrix} \sqrt{n} & \mathbf{0}_p^T \\ \mathbf{0}_p & L \\ \mathbf{0}_{n-p-1} & \mathbf{0}_{n-p-1\times p} \end{bmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \boldsymbol{\epsilon}^*
\end{aligned}
$$

Orthogonal Regression

Rotate by multiplyting by $\mathbf{U}^T$:

$$
\begin{aligned}
\mathbf{U}^T\mathbf{Y} &= \mathbf{U}^T\mathbf{1}\alpha + \mathbf{U}^T\mathbf{U}_p L \mathbf{V}^T\boldsymbol{\beta} + \mathbf{U}^T\boldsymbol{\epsilon} \\
\mathbf{Y}^* &= \left[\begin{array}{cc} \sqrt{n} & \mathbf{0}_p^T \\ \mathbf{0}_p & L \\ \mathbf{0}_{n-p-1} & \mathbf{0}_{n-p-1\times p} \end{array}\right] \left(\begin{array}{c} \alpha \\ \gamma \end{array}\right) + \boldsymbol{\epsilon}^*
\end{aligned}
$$

- $y_0^* \equiv \hat{\alpha} = \bar{y}$

Orthogonal Regression

Rotate by multiplyting by $\mathbf{U}^T$:

$$\begin{aligned}
\mathbf{U}^T\mathbf{Y} &= \mathbf{U}^T\mathbf{1}\alpha + \mathbf{U}^T\mathbf{U}_p L \mathbf{V}^T\boldsymbol{\beta} + \mathbf{U}^T\boldsymbol{\epsilon} \\
\mathbf{Y}^* &= \begin{bmatrix} \sqrt{n} & \mathbf{0}_p^T \\ \mathbf{0}_p & L \\ \mathbf{0}_{n-p-1} & \mathbf{0}_{n-p-1\times p} \end{bmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \boldsymbol{\epsilon}^*
\end{aligned}$$

- $y_0^* \equiv \hat{\alpha} = \bar{y}$
- $\hat{\gamma} = (L^T L)^{-1} L^T \mathbf{U}_p^T \mathbf{Y}$ or $\hat{\gamma}_i = y_i^*/l_i$ for $i = 1, \ldots, p$

Orthogonal Regression

Rotate by multiplyting by $\mathbf{U}^T$:

$$\begin{aligned} \mathbf{U}^T\mathbf{Y} &= \mathbf{U}^T\mathbf{1}\alpha + \mathbf{U}^T\mathbf{U}_p L \mathbf{V}^T\boldsymbol{\beta} + \mathbf{U}^T\boldsymbol{\epsilon} \\ \mathbf{Y}^* &= \begin{bmatrix} \sqrt{n} & \mathbf{0}_p^T \\ \mathbf{0}_p & L \\ \mathbf{0}_{n-p-1} & \mathbf{0}_{n-p-1\times p} \end{bmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \boldsymbol{\epsilon}^* \end{aligned}$$

- $y_0^* \equiv \hat{\alpha} = \bar{y}$
- $\hat{\gamma} = (L^T L)^{-1} L^T \mathbf{U}_p^T \mathbf{Y}$ or $\hat{\gamma}_i = y_i^* / l_i$ for $i = 1, \ldots, p$
- $\mathrm{Var}(\hat{\gamma}_i) = \sigma^2 / l_i^2$

Directions in $\mathbf{X}$ space $\mathbf{U}_j$ with small eigenvectors l_i have the largest variances. Unstable directions.

Ridge Regression & Independent Prior

(Another) Normal Conjugate Prior Distribution on γ:

$$\boldsymbol{\gamma} \mid \phi \sim \mathsf{N}(\mathbf{0}_p, \frac{1}{\phi k}\mathbf{I}_p)$$

Ridge Regression & Independent Prior

(Another) Normal Conjugate Prior Distribution on γ:

$$\boldsymbol{\gamma} \mid \phi \sim \mathsf{N}(\mathbf{0}_p, \frac{1}{\phi k}\mathbf{I}_p)$$

Posterior mean

$$\tilde{\gamma} = (L^T L + k\mathbf{I})^{-1} L^T \mathbf{U}_p^T \mathbf{Y} = (L^T L + k\mathbf{I})^{-1} L^T L \hat{\gamma}$$

Ridge Regression & Independent Prior

(Another) Normal Conjugate Prior Distribution on γ:

$$\boldsymbol{\gamma} \mid \phi \sim \mathsf{N}(\mathbf{0}_p, \frac{1}{\phi k}\mathbf{I}_p)$$

Posterior mean

$$\tilde{\gamma} = (L^T L + k\mathbf{I})^{-1} L^T \mathbf{U}_p^T \mathbf{Y} = (L^T L + k\mathbf{I})^{-1} L^T L \hat{\boldsymbol{\gamma}}$$

$$\tilde{\gamma}_i = \frac{l_i^2}{l_i^2 + k}\hat{\gamma}_i = \frac{\lambda_i}{\lambda_i + k}\hat{\gamma}_i$$

Ridge Regression & Independent Prior

(Another) Normal Conjugate Prior Distribution on γ:

$$\boldsymbol{\gamma} \mid \phi \sim \mathsf{N}(\mathbf{0}_p, \frac{1}{\phi k}\mathbf{I}_p)$$

Posterior mean

$$\tilde{\gamma} = (L^T L + k\mathbf{I})^{-1} L^T \mathbf{U}_p^T \mathbf{Y} = (L^T L + k\mathbf{I})^{-1} L^T L \hat{\gamma}$$

$$\tilde{\gamma}_i = \frac{l_i^2}{l_i^2 + k}\hat{\gamma}_i = \frac{\lambda_i}{\lambda_i + k}\hat{\gamma}_i$$

- When $\lambda_i \to 0$ then $\tilde{\gamma}_i \to 0$
- When $k \to 0$ we get OLS back but if k gets too big posterior mean goes to zero.

Transform

- Transform back $\tilde{\boldsymbol{\beta}} = \mathbf{V}\tilde{\boldsymbol{\gamma}}$

Transform

- Transform back $\tilde{\boldsymbol{\beta}} = \mathbf{V}\tilde{\gamma}$

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

Transform

- Transform back $\tilde{\boldsymbol{\beta}} = \mathbf{V}\tilde{\gamma}$

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

- importance of standardizing

Transform

- Transform back $\tilde{\boldsymbol{\beta}} = \mathbf{V}\tilde{\boldsymbol{\gamma}}$

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

- importance of standardizing
- Is there a value of k for which ridge is better in terms of Expected MSE than OLS?

Transform

- Transform back $\tilde{\boldsymbol{\beta}} = \mathbf{V}\tilde{\gamma}$

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

- importance of standardizing
- Is there a value of k for which ridge is better in terms of Expected MSE than OLS?
- Choice of k?

MSE

Can show that

$$\mathrm{E}[(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})] = \mathrm{E}[(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})^T(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})]$$

MSE

Can show that

$$\mathrm{E}[(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})] = \mathrm{E}[(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})^T(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})]$$

- $\mathrm{Var}(\tilde{\gamma}_i) = \sigma^2 l_i^2/(l_i^2 + k)^2$

MSE

Can show that

$$\mathrm{E}[(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})] = \mathrm{E}[(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})^T(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})]$$

- $\mathrm{Var}(\tilde{\gamma}_i) = \sigma^2 l_i^2/(l_i^2 + k)^2$
- Bias of $\tilde{\boldsymbol{\gamma}}$ is $-k\gamma_i/(l_i^2 + k)$

MSE

Can show that

$$E[(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})] = E[(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})^T(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}]$$

- $\text{Var}(\tilde{\gamma}_i) = \sigma^2 l_i^2/(l_i^2 + k)^2$
- Bias of $\tilde{\boldsymbol{\gamma}}$ is $-k\gamma_i/(l_i^2 + k)$
- MSE
$$\sigma^2 \sum_i \frac{l_i^2}{(l_i^2 + k)^2} + k^2 \sum_i \frac{\gamma_i^2}{(l_i^2 + k)^2}$$

The derivative with respect to k is negative at $k = 0$, hence the function is decreasing.

MSE

Can show that

$$E[(\beta - \tilde{\beta})^T(\beta - \tilde{\beta})] = E[(\gamma - \tilde{\gamma})^T(\gamma - \tilde{\gamma})]$$

- $\text{Var}(\tilde{\gamma}_i) = \sigma^2 l_i^2/(l_i^2 + k)^2$
- Bias of $\tilde{\gamma}$ is $-k\gamma_i/(l_i^2 + k)$
- MSE
$$\sigma^2 \sum_i \frac{l_i^2}{(l_i^2 + k)^2} + k^2 \sum_i \frac{\gamma_i^2}{(l_i^2 + k)^2}$$

The derivative with respect to k is negative at $k = 0$, hence the function is decreasing.

Since $k = 0$ is OLS, this means that is a value of k that will always be better than OLS

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 $\mathbf{X}$
- Let $\mathbf{Y}^c = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$ and $\mathbf{X}^c$ the centered and standardized $\mathbf{X}$ matrix

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Let $\mathbf{Y}^c = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$ and $\mathbf{X}^c$ the centered and standardized $\mathbf{X}$ matrix
- Control how large coefficients may grow

Alternative Motivation

- If $\hat{\boldsymbol{\beta}}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Let $\mathbf{Y}^c = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$ and $\mathbf{X}^c$ the centered and standardized $\mathbf{X}$ matrix
- Control how large coefficients may grow

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

subject to

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

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Let $\mathbf{Y}^c = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$ and $\mathbf{X}^c$ the centered and standardized $\mathbf{X}$ matrix
- Control how large coefficients may grow

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

subject to

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

- Equivalent Quadratic Programming Problem

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

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Let $\mathbf{Y}^c = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$ and $\mathbf{X}^c$ the centered and standardized $\mathbf{X}$ matrix
- Control how large coefficients may grow

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

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

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Let $\mathbf{Y}^c = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$ and $\mathbf{X}^c$ the centered and standardized $\mathbf{X}$ matrix
- Control how large coefficients may grow

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

subject to

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

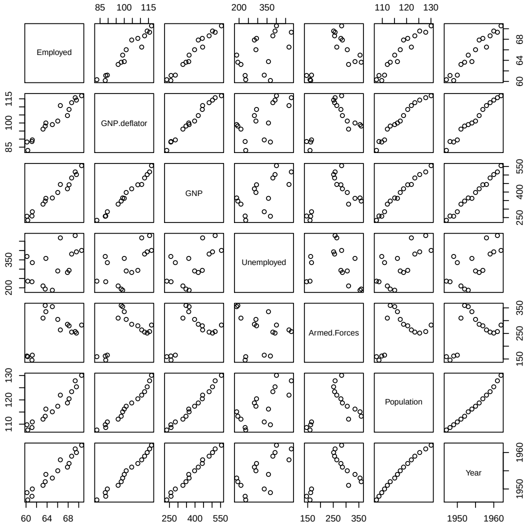
- Equivalent Quadratic Programming Problem

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

- "penalized" likelihood

Picture

Longley Data



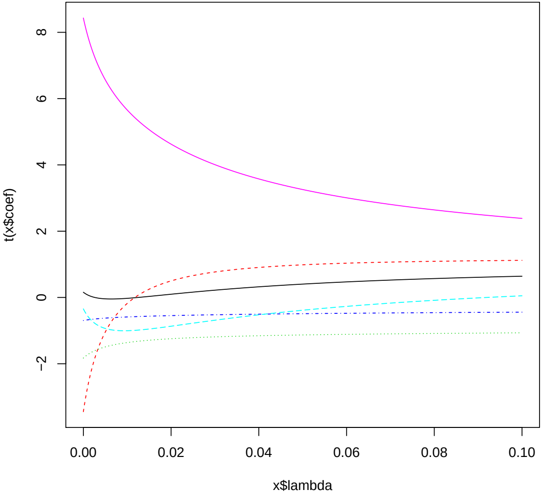
OLS

```
> 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 Trace



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

> longley.RReg = lm.ridge(Employed ~ ., data=longley,
                                 lambda=0.0028)
> coef(longley.RReg)
           GNP.deflator     GNP      Unemployed   Armed.Forces
-2.950e+03 -5.381e-04   -1.822e-02  -1.76e-02 -9.607e-03

 Population     Year
-1.185e-01  1.557e+00
```

Testimators

Goldstein & Smith (1974) have shown that if

1. $0 \leq h_i \leq 1$ and $\tilde{\gamma}_i = h_i \hat{\gamma}_i$
2. $\frac{\gamma_i^2}{\mathrm{Var}(\hat{\gamma}_i)} < \frac{1+h_i}{1-h_i}$

then $\tilde{\gamma}_i$ has smaller MSE than $\hat{\gamma}_i$

Case: If $\gamma_j < \mathrm{Var}(\hat{\gamma}_i) = \sigma^2 / l_i^2$ then $h_i = 0$ and $\tilde{\gamma}_i$ is better.

Apply: Estimate σ^2 with SSE/(n - p - 1) and γ_i with $\hat{\gamma}_i$. Set $h_i = 0$ if t-statistic is less than 1.

"testimator" - see also Sclove (JASA 1968) and Copas (JRSSB 1983)

Generalized Ridge

Instead of $\gamma_j \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2/k)$ take

$$\gamma_j \overset{\text{ind}}{\sim} \mathsf{N}(0, \sigma^2/k_i)$$

Generalized Ridge

Instead of $\gamma_j \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2/k)$ take

$$\gamma_j \overset{\text{ind}}{\sim} \mathsf{N}(0, \sigma^2/k_i)$$

Then Condition of Goldstein & Smith becomes

$$\gamma_i^2 < \sigma^2 \left[\frac{2}{k_j} + \frac{1}{l_i^2} \right]$$

Generalized Ridge

Instead of $\gamma_j \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2/k)$ take

$$\gamma_j \stackrel{\text{ind}}{\sim} \mathsf{N}(0, \sigma^2/k_i)$$

Then Condition of Goldstein & Smith becomes

$$\gamma_i^2 < \sigma^2 \left[\frac{2}{k_j} + \frac{1}{l_i^2} \right]$$

- If l_i is small almost any k_i will improve over OLS

Generalized Ridge

Instead of $\gamma_j \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2/k)$ take

$$\gamma_j \overset{\text{ind}}{\sim} \mathsf{N}(0, \sigma^2/k_i)$$

Then Condition of Goldstein & Smith becomes

$$\gamma_i^2 < \sigma^2 \left[\frac{2}{k_j} + \frac{1}{l_i^2} \right]$$

- If l_i is small almost any k_i will improve over OLS
- if l_i^2 is large then only very small values of k_i will give an improvement

Generalized Ridge

Instead of $\gamma_j \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2/k)$ take

$$\gamma_j \stackrel{\text{ind}}{\sim} \mathsf{N}(0, \sigma^2/k_i)$$

Then Condition of Goldstein & Smith becomes

$$\gamma_i^2 < \sigma^2 \left[\frac{2}{k_j} + \frac{1}{l_i^2}\right]$$

- If l_i is small almost any k_i will improve over OLS
- if l_i^2 is large then only very small values of k_i will give an improvement
- Prior on k_i?

Generalized Ridge

Instead of $\gamma_j \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2/k)$ take

$$\gamma_j \stackrel{\text{ind}}{\sim} \mathsf{N}(0, \sigma^2/k_i)$$

Then Condition of Goldstein & Smith becomes

$$\gamma_i^2 < \sigma^2 \left[\frac{2}{k_j} + \frac{1}{l_i^2} \right]$$

- If l_i is small almost any k_i will improve over OLS
- if l_i^2 is large then only very small values of k_i will give an improvement
- Prior on k_i?
- Induced prior on $\boldsymbol{\beta}$?

$$\gamma_j \stackrel{\text{ind}}{\sim} \mathsf{N}(0, \sigma^2/k_i) \Leftrightarrow \boldsymbol{\beta} \sim \mathsf{N}(\mathbf{0}, \sigma^2 \mathbf{V} K^{-1} \mathbf{V}^T)$$

which is not diagonal. Loss of invariance.

Summary

- OLS can clearly be dominated by other estimators
- Lead to Bayes like estimators
- choice of penalities or prior hyperparameters
- hierarchical model with prior on k_i