Ridge Regression
Readings Chapter 15 Christensen

STA721 Linear Models Duke University

Merlise Clyde

October 17, 2017
How Good are Estimators?

Quadratic loss for estimating $\beta$ using estimator $a$

$$L(\beta, a) = (\beta - a)^T(\beta - a)$$
How Good are Estimators?

Quadratic loss for estimating $\beta$ using estimator $a$

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

- Under OLS or the Reference prior the Expected Mean Square Error

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

- Similar problem with $g$ prior or mixtures of $g$-priors
How Good are Estimators?

Quadratic loss for estimating $\beta$ using estimator $a$

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

> Under OLS or the Reference prior the Expected Mean Square Error

$$E_Y[(\beta - \hat{\beta})^T (\beta - \hat{\beta}) = \sigma^2 \text{tr}[(X^T X)^{-1}]$$
How Good are Estimators?

Quadratic loss for estimating $\beta$ using estimator $a$

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

- Under OLS or the Reference prior the Expected Mean Square Error

$$E_Y[(\beta - \hat{\beta})^T (\beta - \hat{\beta})] = \sigma^2 \text{tr}[X^TX^{-1}]$$

$$= \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$$
How Good are Estimators?

Quadratic loss for estimating $\beta$ using estimator $a$

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

▶ Under OLS or the Reference prior the Expected Mean Square Error

$$E_Y[(\beta - \hat{\beta})^T (\beta - \hat{\beta})] = \sigma^2 \text{tr}[(X^T X)^{-1}]$$

$$= \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$$

▶ If smallest $\lambda_j \to 0$ then MSE $\to \infty$

▶ Similar problem with $g$ prior or mixtures of $g$-priors
Canonical Representation & Ridge Regression

Assume that $X$ has been centered and standardized so that $X^TX = \text{corr}(X)$
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 $\mathbb{R}$).
Canonical Representation & Ridge Regression

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

- Write $X = U_pL V^T$ Singular Value Decomposition
Canonical Representation & Ridge Regression

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

- Write \( X = U_p L V^T \) Singular Value Decomposition where \( U_p^T U_p = I_p \) and \( V \) is \( p \times p \) orthogonal matrix, \( L \) is diagonal
Canonical Representation & Ridge Regression

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

- Write \( X = U_pLV^T \) Singular Value Decomposition where \( U_p^T U_p = I_p \) and \( V \) is \( p \times p \) orthogonal matrix, \( L \) is diagonal

\[
Y = 1\alpha + U_pLV^T \beta + \epsilon
\]
Canonical Representation & Ridge Regression

Assume that \( X \) has been centered and standardized so that
\[ X^T X = \text{corr}(X) \] (use scale or sweep functions in \( \mathbb{R} \))

- Write \( X = U_pLV^T \) Singular Value Decomposition where
  \( U_p^T U_p = I_p \) and \( V \) is \( p \times p \) orthogonal matrix, \( L \) is diagonal

\[ Y = 1\alpha + U_pLV^T \beta + \epsilon \]

- Let \( U = [1/\sqrt{n} U_p U_{n-p-1}] \) \( n \times n \) orthogonal matrix
Canonical Representation & Ridge Regression

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

- Write $X = U_p L V^T$ Singular Value Decomposition where $U_p^T U_p = I_p$ and $V$ is $p \times p$ orthogonal matrix, $L$ is diagonal

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

- Let $U = \left[ \frac{1}{\sqrt{n}} U_p U_{n-p-1} \right]$ $n \times n$ orthogonal matrix
- Rotate by $U^T$
Canonical Representation & Ridge Regression

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

- Write $X = U_pLV^T$ Singular Value Decomposition where $U_p^TU_p = I_p$ and $V$ is $p \times p$ orthogonal matrix, $L$ is diagonal

  $$Y = 1\alpha + U_pLV^T\beta + \epsilon$$

- Let $U = [1/\sqrt{n} U_p U_{n-p-1}] n \times n$ orthogonal matrix
- Rotate by $U^T$

  $$U^TY = U^T1\alpha + U^TU_pLV^T\beta + U^T\epsilon$$
Canonical Representation & Ridge Regression

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

- Write $X = U_p L V^T$ Singular Value Decomposition where $U_p^T U_p = I_p$ and $V$ is $p \times p$ orthogonal matrix, $L$ is diagonal

\[
Y = 1\alpha + U_p L V^T \beta + \epsilon
\]

- Let $U = \begin{bmatrix} 1/\sqrt{n} & U_p & U_{n-p-1} \end{bmatrix}$ $n \times n$ orthogonal matrix
- Rotate by $U^T$

\[
U^T Y = U^T 1\alpha + U^T U_p L V^T \beta + U^T \epsilon
\]

\[
Y^* = \begin{bmatrix}
\sqrt{n} & 0_p \\
0 & L \\
0_{n-p-1} & 0_{n-p-1 \times p}
\end{bmatrix}
\begin{pmatrix}
\alpha \\
\gamma
\end{pmatrix} + \epsilon^*
\]
Canonical Representation & Ridge Regression

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

- Write $X = U_p L V^T$ Singular Value Decomposition where $U_p^T U_p = I_p$ and $V$ is $p \times p$ orthogonal matrix, $L$ is diagonal

$$Y = 1\alpha + U_p L V^T \beta + \epsilon$$

- Let $U = [1/\sqrt{n} U_p U_{n-p-1}]$ $n \times n$ orthogonal matrix
- Rotate by $U^T$

$$U^T Y = U^T 1\alpha + U^T U_p L V^T \beta + U^T \epsilon$$

$$Y^* = \begin{bmatrix} \sqrt{n} & 0_p \\ 0 & L \\ 0_{n-p-1} & 0_{n-p-1 \times p} \end{bmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \epsilon^*$$
Orthogonal Regression

$$U^T Y = U^T 1\alpha + U^T U_p L V^T \beta + U^T \epsilon$$
Orthogonal Regression

\[ U^T Y = U^T 1 \alpha + U^T U_p L V^T \beta + U^T \epsilon \]

\[ Y^* = \begin{bmatrix} \sqrt{n} & 0 & 0_p \\ 0 & L & 0_{n-p-1} \\ 0_{n-p-1} & 0_{n-p-1 \times p} \end{bmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \epsilon^* \]
Orthogonal Regression

\[
U^T Y = U^T 1\alpha + U^T U_p L V^T \beta + U^T \epsilon
\]

\[
Y^* = \begin{bmatrix}
\sqrt{n} & 0_p \\
0 & L \\
0_{n-p-1} & 0_{n-p-1 \times p}
\end{bmatrix}
\begin{pmatrix}
\alpha \\
\gamma
\end{pmatrix} + \epsilon^*
\]

\[
\hat{\alpha} = \bar{y}
\]
Orthogonal Regression

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

\[ \mathbf{Y}^* = \begin{bmatrix} \sqrt{n} & 0_p \\ 0 & \mathbf{L} \\ 0_{n-p-1} & 0_{n-p-1 \times p} \end{bmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \epsilon^* \]

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

\[
U^T Y = U^T 1 \alpha + U^T U_p L V^T \beta + U^T \epsilon
\]

\[
Y^* = \begin{bmatrix}
\sqrt{n} & 0_p \\
0 & L \\
0_{n-p-1} & 0_{n-p-1 \times p}
\end{bmatrix}
\begin{pmatrix}
\alpha \\
\gamma
\end{pmatrix} + \epsilon^*
\]

- \hat{\alpha} = \bar{y}
- \hat{\gamma_i} = (L^T L)^{-1} L^T U_p Y \text{ or } \hat{\gamma_i} = y_i^*/l_i \text{ for } i = 1, \ldots, p
- \text{Var}(\hat{\gamma_i}) = \sigma^2/l_i^2

Directions in \( \mathbf{X} \) space \( U_j \) with small eigenvectors \( l_i \) have the largest variances. Unstable directions.
Ridge Regression & Independent Prior

(Another) Normal Conjugate Prior Distribution on $\gamma$:

$$\gamma \mid \phi \sim N(0_p, \frac{1}{\phi} I_p)$$

Posterior mean:

$$\tilde{\gamma} = \left( L^T L + k I_p \right)^{-1} L^T U^T p Y$$

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.
(Another) Normal Conjugate Prior Distribution on $\gamma$:

$$\gamma \mid \phi \sim \mathcal{N}(0_p, \frac{1}{\phi k} I_p)$$

Posterior mean

$$\tilde{\gamma} = (L^T L + kI)^{-1} L^T U_p^T Y = (L^T L + kI)^{-1} L^T L \hat{\gamma}$$

$\text{When } \lambda_i \to 0 \text{ then } \tilde{\gamma}_i \to 0$

$\text{When } k \to 0 \text{ we get OLS back but if } k \text{ gets too big posterior mean goes to zero.}$
Ridge Regression & Independent Prior

(Another) Normal Conjugate Prior Distribution on $\gamma$:

$$\gamma \mid \phi \sim N(0_p, \frac{1}{\phi} I_p)$$

Posterior mean

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

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

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.
Ridge Regression & Independent Prior

(Another) Normal Conjugate Prior Distribution on $\gamma$:

$$\gamma | \phi \sim N(0_p, \frac{1}{\phi k} I_p)$$

Posterior mean

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

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

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

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

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

$$
\tilde{\beta} = (X^TX + kI)^{-1}X^TX\hat{\beta}
$$

- importance of standardizing
Transform back $\tilde{\beta} = V\tilde{\gamma}$

$$\tilde{\beta} = (X^TX + kI)^{-1}X^TX\hat{\beta}$$

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

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

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

- 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

\[ E[(\beta - \tilde{\beta})^T (\beta - \tilde{\beta})] = E[(\gamma - \tilde{\gamma})^T (\gamma - \tilde{\gamma})] \]
Can show that

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

\[ \text{Var}(\gamma_i - \tilde{\gamma}_i) = \sigma^2 l_i^2 / (l_i^2 + k)^2 \]
MSE

Can show that

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

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

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

- \( \text{Var}(\gamma_i - \bar{\gamma}_i) = \sigma^2 l_i^2 / (l_i^2 + k)^2 \)
- Bias of \( \bar{\gamma} \) is \(-k / (l_i^2 + k)\)
- 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

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

- \[ \text{Var}(\gamma_i - \tilde{\gamma}_i) = \sigma^2 l_i^2 / (l_i^2 + k)^2 \]
- \[ \text{Bias of } \tilde{\gamma} \text{ is } -k / (l_i^2 + k) \]
- 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 $X$
Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $X$
- Let $Y^c = (I - P_1)Y$ and $X^c$ the centered and standardized $X$ matrix
Alternative Motivation

- If \( \hat{\beta} \) is unconstrained expect high variance with nearly singular \( X \)
- Let \( Y^c = (I - P_1)Y \) and \( X^c \) the centered and standardized \( X \) matrix
- Control how large coefficients may grow
Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $X$
- Let $Y^c = (I - P_1)Y$ and $X^c$ the centered and standardized $X$ matrix
- Control how large coefficients may grow

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

subject to

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

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $X$
- Let $Y^c = (I - P_1)Y$ and $X^c$ the centered and standardized $X$ matrix
- Control how large coefficients may grow

$$\min_{\beta} (Y^c - X^c \beta)^T (Y^c - X^c \beta)$$
subject to

$$\sum \beta^2 \leq t$$

- Equivalent Quadratic Programming Problem

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

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $X$
- Let $Y^c = (I - P_1)Y$ and $X^c$ the centered and standardized $X$ matrix
- Control how large coefficients may grow

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

subject to

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

- Equivalent Quadratic Programming Problem

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

- “penalized” likelihood
Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $X$
- Let $Y^c = (I - P_1)Y$ and $X^c$ the centered and standardized $X$ matrix
- Control how large coefficients may grow

$$
\min_{\beta} (Y^c - X^c \beta)^T (Y^c - X^c \beta)
$$
subject to

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

- Equivalent Quadratic Programming Problem

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

- “penalized” likelihood
Longley Data
OLS

```r
> 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, Adjusted R-squared: 0.9925
F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
Ridge Trace
Generalized Cross-validation

```r
> 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
```
Goldstein & Smith (1974) have shown that if

1. $0 \leq h_i \leq 1$ and $\tilde{\gamma}_i = h_i \hat{\gamma}_i$

2. $\frac{\gamma_i^2}{\text{Var}(\hat{\gamma}_i)} < \frac{1+h_i}{1-h_i}$

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

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

Apply: Estimate $\sigma^2$ with SSE/(n - p - 1) and $\gamma_i$ 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 $\gamma_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2/k)$ take

$$\gamma_j^{\text{ind}} \sim \mathcal{N}(0, \sigma^2/k_i)$$
Generalized Ridge

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

\[
\gamma_j^{\text{ind}} \sim N(0, \sigma^2/k_i)
\]

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 \( \gamma_j \overset{iid}{\sim} N(0, \sigma^2/k) \) take

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

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 \( \gamma_j \sim \text{N}(0, \sigma^2/k) \) take

\[
\gamma_j^{\text{ind}} \sim \text{N}(0, \sigma^2/k_i)
\]

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 $\gamma_j \sim \text{N}(0, \sigma^2/k)$ take

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

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 $\gamma_j \sim \text{N}(0, \sigma^2/k)$ take

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

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 $\beta$?

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

which is not diagonal. Loss of invariance.
Summary

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