Statistical Learning– MATH 6333
Set 3 (Linear Methods for Regression)

Tamer Oraby

UTRGV
tamer.oraby@utrgv.edu

* Last updated September 29, 2021

Linear Regression Models and Least Squares

Linear Regression Models and Least Squares

The linear regression model

$$f(X) = E(Y|X) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

For the training data $\mathcal{T} = \{(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i) : i = 1, 2, \ldots, N\}$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p} + \epsilon_i$$

for uncorrelated ϵ_i's, of variance σ^2.

In matrix form

$$y = X\beta + \epsilon,$$

where X is a $N \times (p + 1)$ matrix with ones in the first column.

Linear Regression Models and Least Squares

Each input X_j (for $j = 1, \ldots, p$) could be one of several types:

1. quantitative variable, ex: age, sales, mileage.
2. transformation of a quantitative variable, ex: log(age), $\sqrt{sales}$, $mileage^2$
3. as basis expansions, ex: in a polynomial

$$\beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3$$

Linear Regression Models and Least Squares

Each input X_j (for $j = 1, \ldots, p$) could be one of several types:

1. quantitative variable, ex: age, sales, mileage.
2. transformation of a quantitative variable, ex: log(age), $\sqrt{sales}$, $mileage^2$
3. as basis expansions, ex: in a polynomial

$$\beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3$$

Linear Regression Models and Least Squares

Each input X_j (for $j = 1, \ldots, p$) could be one of several types:

1. quantitative variable, ex: age, sales, mileage.

2. transformation of a quantitative variable, ex: log(age), $\sqrt{sales}$, $mileage^2$

3. as basis expansions, ex: in a polynomial

$$\beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3$$

Linear Regression Models and Least Squares

4. dummy variable (factor), ex: if G takes one of the levels 0, 1, or 2, then take $X_j = I(G = j)$ for $j = 1, 2$ and so

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = \begin{cases} \beta_0 & \text{if G=0,} \\ \beta_0 + \beta_1 & \text{if G=1,} \\ \beta_0 + \beta_2 & \text{if G=2.} \end{cases}$$

5. interaction between variables, ex: age x mileage

Linear Regression Models and Least Squares

4. dummy variable (factor), ex: if G takes one of the levels 0, 1, or 2, then take $X_j = I(G = j)$ for $j = 1, 2$ and so

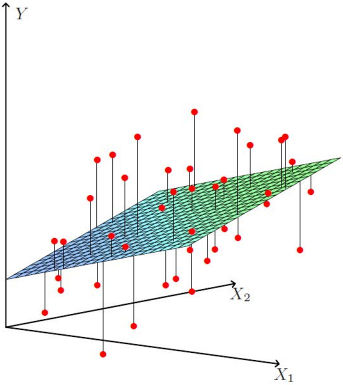
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = \begin{cases} \beta_0 & \text{if G=0,} \\ \beta_0 + \beta_1 & \text{if G=1,} \\ \beta_0 + \beta_2 & \text{if G=2.} \end{cases}$$

5. interaction between variables, ex: age x mileage

Linear Regression Models and Least Squares

The method of least squares finds β's that minimizes residual sums of squares

$$
\begin{aligned}
RSS(\beta) &= \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 \\
&= \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \\
&= (y - X\beta)^T (y - X\beta)
\end{aligned}
$$



Linear Regression Models and Least Squares

$$\text{minimize}_\beta RSS(\beta) = (y - X\beta)^T(y - X\beta)$$

- $\dfrac{\partial RSS(\beta)}{\partial\beta} = -2X^T(y - X\beta) = 0 \implies X^TX\beta = X^Ty.$
- $\dfrac{\partial^2 RSS(\beta)}{\partial\beta\partial\beta^T} = 2X^TX$
- If X is full column rank (columns are linearly independent), then X^TX is positive definite and so non-singular, then

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

- Predictions

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{H}y$$

H is called the hat matrix or the (orthogonal) projection matrix.

Linear Regression Models and Least Squares

$$\text{minimize}_\beta RSS(\beta) = (y - X\beta)^T(y - X\beta)$$

- $\dfrac{\partial RSS(\beta)}{\partial \beta} = -2X^T(y - X\beta) = 0 \implies X^TX\beta = X^Ty.$
- $\dfrac{\partial^2 RSS(\beta)}{\partial \beta \partial \beta^T} = 2X^TX$
- If X is full column rank (columns are linearly independent), then X^TX is positive definite and so non-singular, then

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

- Predictions

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{H}y$$

H is called the hat matrix or the (orthogonal) projection matrix.

Linear Regression Models and Least Squares

$$\text{minimize}_\beta RSS(\beta) = (y - X\beta)^T (y - X\beta)$$

- $\frac{\partial RSS(\beta)}{\partial \beta} = -2X^T(y - X\beta) = 0 \implies X^T X\beta = X^T y.$
- $\frac{\partial^2 RSS(\beta)}{\partial \beta \partial \beta^T} = 2X^T X$
- If X is full column rank (columns are linearly independent), then $X^T X$ is positive definite and so non-singular, then

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

- Predictions

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{H} y$$

H is called the hat matrix or the (orthogonal) projection matrix.

Linear Regression Models and Least Squares

$$\text{minimize}_\beta RSS(\beta) = (y - X\beta)^T(y - X\beta)$$

- $\dfrac{\partial RSS(\beta)}{\partial \beta} = -2X^T(y - X\beta) = 0 \implies X^TX\beta = X^Ty.$
- $\dfrac{\partial^2 RSS(\beta)}{\partial \beta \partial \beta^T} = 2X^TX$
- If X is full column rank (columns are linearly independent), then X^TX is positive definite and so non-singular, then

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

- Predictions

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{H} y$$

H is called the hat matrix or the (orthogonal) projection matrix.

Linear Regression Models and Least Squares

- If y_i's are uncorrelated and have variance σ^2, then

  $$Var(\hat{\beta}) = (X^T X)^{-1}\sigma^2$$

  if x_i's are fixed.

- An unbiased estimator of σ^2 is

$$\begin{aligned}
\widehat{\sigma^2} &= \frac{RSS(\hat{\beta})}{N-p-1} \\
&= \frac{(y - X\hat{\beta})^T(y - X\hat{\beta})}{N-p-1} \\
&= \frac{(y-\hat{y})^T(y-\hat{y})}{N-p-1} \\
&= \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2
\end{aligned}$$

Linear Regression Models and Least Squares

- If y_i's are uncorrelated and have variance σ^2, then

  $$Var(\hat{\beta}) = (X^T X)^{-1}\sigma^2$$

  if x_i's are fixed.
- An unbiased estimator of σ^2 is

  $$\begin{aligned}\widehat{\sigma^2} &= \frac{RSS(\hat{\beta})}{N-p-1} \\ &= \frac{(y - X\hat{\beta})^T(y - X\hat{\beta})}{N-p-1} \\ &= \frac{(y-\hat{y})^T(y-\hat{y})}{N-p-1} \\ &= \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\end{aligned}$$

Linear Regression Models and Least Squares

What if columns are not linearly independent?

That is, what if they are perfectly correlated

$$X_i = \text{constant} \ \times X_j$$

for some *i* and *j*.

Then, $\hat{\beta}$ is not uniquely defined.

Solutions:

- Re-code redundant qualitative inputs
- If p is much larger than N, then the number of inputs p is reduced by filtering.

Linear Regression Models and Least Squares

What if columns are not linearly independent?

That is, what if they are perfectly correlated

$$X_i = \text{constant} \ \times X_j$$

for some *i* and *j*.

Then, $\hat{\beta}$ is not uniquely defined.

Solutions:

- Re-code redundant qualitative inputs
- If *p* is much larger than *N*, then the number of inputs *p* is reduced by filtering.

Linear Regression Models and Least Squares

What if columns are not linearly independent?

That is, what if they are perfectly correlated

$$X_i = \text{constant } \times X_j$$

for some *i* and *j*.

Then, $\hat{\beta}$ is not uniquely defined.

Solutions:

- Re-code redundant qualitative inputs
- If *p* is much larger than *N*, then the number of inputs *p* is reduced by filtering.

Statistical Inference for Linear Regression

Statistical Inference for Linear Regression

If ϵ_i are iidrv such that $\epsilon_i \sim N(0, \sigma^2)$, then

$$\hat{\beta} \sim N(\beta, (X^T X)^{-1} \sigma^2),$$

$$c^T \hat{\beta} = \sum_{j=0}^{p} c_j \hat{\beta}_j \sim N(c^T \beta, c^T (X^T X)^{-1} c \, \sigma^2),$$

for a non-zero vector c, and

$$(N - p - 1) \frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{N-p-1}.$$

Moreover, $\hat{\beta}$ and $\widehat{\sigma^2}$ are statistically independent. Thus, ...

Statistical Inference for Linear Regression

...

$$\frac{c^T\hat{\beta} - c^T\beta}{\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}} \sim T_{N-p-1}$$

Example

If $c = e_j = (0, \ldots, 0, \underbrace{1}_{j^{th}}, \ldots, 0)^T$, then $c^T\beta = \beta_j$ and

$$c^T(X^TX)^{-1}c = ((X^TX)^{-1})_{jj} =: v_{jj}$$

the j^{th} diagonal element of $(X^TX)^{-1}$. Therefore,

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{v_{jj}}} \sim T_{N-p-1}$$

Statistical Inference for Linear Regression

Example

If $c = (0, \ldots, 0, \underbrace{1}_{i^{th}}, 0, \ldots, 0, \underbrace{-1}_{j^{th}}, \ldots, 0)^T$, then $c^T\beta = \beta_i - \beta_j$

and

$$c^T(X^TX)^{-1}c = v_{ii} + v_{jj} - v_{ij} - v_{ji}$$

where v_{ij} the ij^{th} element of $(X^TX)^{-1}$. Therefore,

$$\frac{(\hat{\beta}_i - \hat{\beta}_j) - (\beta_i - \beta_j)}{\hat{\sigma}\sqrt{v_{ii} + v_{jj} - v_{ij} - v_{ji}}} \sim T_{N-p-1}$$

Statistical Inference for Linear Regression

Now, since

$$\frac{c^T\hat{\beta} - c^T\beta}{\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}} \sim T_{N-p-1}$$

then

- A $(1-\alpha)100\%$ confidence interval for $c^T\beta$ is given by

$$c^T\hat{\beta} \pm t_{\alpha/2,N-p-1}\,\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}$$

- To test

$$H_0 : c^T\beta = d_0 \;\; vs \;\; H_A : c^T\beta \neq d_0,\, c^T\beta < d_0,\, \text{ or } c^T\beta > d_0$$

use a test statistic

$$t = \frac{c^T\hat{\beta} - d_0}{\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}}$$

and p-value calculated using T_{N-p-1} distribution.

Statistical Inference for Linear Regression

Now, since

$$\frac{c^T\hat{\beta} - c^T\beta}{\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}} \sim T_{N-p-1}$$

then

- A $(1-\alpha)100\%$ confidence interval for $c^T\beta$ is given by

$$c^T\hat{\beta} \pm t_{\alpha/2,N-p-1}\,\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}$$

- To test

$$H_0 : c^T\beta = d_0 \ vs \ H_A : c^T\beta \neq d_0, c^T\beta < d_0, \text{ or } c^T\beta > d_0$$

use a test statistic

$$t = \frac{c^T\hat{\beta} - d_0}{\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}}$$

and p-value calculated using T_{N-p-1} distribution.

Statistical Inference for Linear Regression

Now, since

$$\frac{c^T\hat{\beta} - c^T\beta}{\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}} \sim T_{N-p-1}$$

then

- A $(1-\alpha)100\%$ confidence interval for $c^T\beta$ is given by

$$c^T\hat{\beta} \pm t_{\alpha/2,N-p-1}\,\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}$$

- To test

$$H_0 : c^T\beta = d_0 \; vs \; H_A : c^T\beta \neq d_0, c^T\beta < d_0, \text{ or } c^T\beta > d_0$$

  use a test statistic

$$t = \frac{c^T\hat{\beta} - d_0}{\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}}$$

  and p-value calculated using T_{N-p-1} distribution.

Statistical Inference for Linear Regression

Example

To test

$$H_0 : \beta_j = 0 \; vs \; H_A : \beta_j \neq 0$$

use a test statistic

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_{jj}}}$$

and p-value calculated using T_{N-p-1} distribution.

Statistical Inference for Linear Regression

Example

To test

$$H_0 : \beta_i = \beta_j \; vs \; H_A : \beta_i \neq \beta_j$$

use a test statistic

$$t = \frac{\hat{\beta}_i - \hat{\beta}_j}{\hat{\sigma}\sqrt{v_{ii} + v_{jj} - v_{ij} - v_{ji}}}$$

and p-value calculated using T_{N-p-1} distribution.

Inference for mean response and prediction

Inference for mean response and prediction

To make a prediction for a new input vector $x_* = (x_{*1}, \ldots, x_{*p})^T$, then

- A point estimate is $\hat{y} = x_*^T \hat{\beta}$.
- A $(1-\alpha)100\%$ confidence interval for the mean response $E(y|x_*) = x_*^T \beta$ is given by

$$x_*^T \hat{\beta} \pm t_{\alpha/2, N-p-1}\, \hat{\sigma} \sqrt{x_*^T (X^T X)^{-1} x_*}$$

- A $(1-\alpha)100\%$ confidence interval for predicted response y at x_* is given by

$$x_*^T \hat{\beta} \pm t_{\alpha/2, N-p-1}\, \hat{\sigma} \sqrt{1 + x_*^T (X^T X)^{-1} x_*}$$

Inference for mean response and prediction

To make a prediction for a new input vector $x_* = (x_{*1}, \dots, x_{*p})^T$, then

- A point estimate is $\hat{y} = x_*^T \hat{\beta}$.
- A $(1-\alpha)100\%$ confidence interval for the mean response $E(y|x_*) = x_*^T \beta$ is given by

$$x_*^T \hat{\beta} \pm t_{\alpha/2, N-p-1}\, \hat{\sigma} \sqrt{x_*^T (X^T X)^{-1} x_*}$$

- A $(1-\alpha)100\%$ confidence interval for predicted response y at x_* is given by

$$x_*^T \hat{\beta} \pm t_{\alpha/2, N-p-1}\, \hat{\sigma} \sqrt{1 + x_*^T (X^T X)^{-1} x_*}$$

Inference for mean response and prediction

To make a prediction for a new input vector $x_* = (x_{*1}, \dots, x_{*p})^T$, then

- A point estimate is $\hat{y} = x_*^T \hat{\beta}$.
- A $(1-\alpha)100\%$ confidence interval for the mean response $E(y|x_*) = x_*^T \beta$ is given by

$$x_*^T \hat{\beta} \pm t_{\alpha/2, N-p-1}\, \hat{\sigma} \sqrt{x_*^T (X^T X)^{-1} x_*}$$

- A $(1-\alpha)100\%$ confidence interval for predicted response y at x_* is given by

$$x_*^T \hat{\beta} \pm t_{\alpha/2, N-p-1}\, \hat{\sigma} \sqrt{1 + x_*^T (X^T X)^{-1} x_*}$$

Inference for mean response and prediction

To make a prediction for a new input vector $x_* = (x_{*1}, \ldots, x_{*p})^T$, then

- A point estimate is $\hat{y} = x_*^T \hat{\beta}$.
- A $(1-\alpha)100\%$ confidence interval for the mean response $E(y|x_*) = x_*^T \beta$ is given by

$$x_*^T \hat{\beta} \pm t_{\alpha/2, N-p-1} \, \hat{\sigma} \sqrt{x_*^T (X^T X)^{-1} x_*}$$

- A $(1-\alpha)100\%$ confidence interval for predicted response y at x_* is given by

$$x_*^T \hat{\beta} \pm t_{\alpha/2, N-p-1} \, \hat{\sigma} \sqrt{1 + x_*^T (X^T X)^{-1} x_*}$$

Model evaluation

Model evaluation

To test

$$H_0 : \beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0 \textit{ (restricted model } M_0\textit{) vs}$$

$$H_A : \text{ At least one } \beta_{j_i} \neq 0; \text{ for } i = 1, 2, \ldots, k$$

use a test statistic

$$f = \frac{(RSS(\hat{\beta}_{\text{restricted}}) - RSS(\hat{\beta}_{\text{full}}))/k}{RSS(\hat{\beta}_{\text{full}})/(N - p - 1)}$$

and $p - value = P(F > f)$ using the F-distribution with degrees of freedom $df_1 = k$ and $df_2 = N - p - 1$.

Note: $RSS(\hat{\beta}_{\text{restricted}})$ is the residuals sum of squares of the (nested) model restricted to $\beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0$

Model evaluation

To test

$$H_0 : \beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0 \textit{ (restricted model } M_0\textit{) vs}$$

$$H_A : \text{ At least one } \beta_{j_i} \neq 0; \text{ for } i = 1, 2, \ldots, k$$

use a test statistic

$$f = \frac{(RSS(\hat{\beta}_{\text{restricted}}) - RSS(\hat{\beta}_{\text{full}}))/k}{RSS(\hat{\beta}_{\text{full}})/(N - p - 1)}$$

and $p - value = P(F > f)$ using the F-distribution with degrees of freedom $df_1 = k$ and $df_2 = N - p - 1$.

Note: $RSS(\hat{\beta}_{\text{restricted}})$ is the residuals sum of squares of the (nested) model restricted to $\beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0$

Model evaluation

To test

$$H_0 : \beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0 \textit{ (restricted model } M_0\textit{) vs}$$

$$H_A : \text{ At least one } \beta_{j_i} \neq 0; \text{ for } i = 1, 2, \ldots, k$$

use a test statistic

$$f = \frac{(RSS(\hat{\beta}_{\text{restricted}}) - RSS(\hat{\beta}_{\text{full}}))/k}{RSS(\hat{\beta}_{\text{full}})/(N - p - 1)}$$

and $p - value = P(F > f)$ using the F-distribution with degrees of freedom $df_1 = k$ and $df_2 = N - p - 1$.

Note: $RSS(\hat{\beta}_{\text{restricted}})$ is the residuals sum of squares of the (nested) model restricted to $\beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0$

Model diagnostics

Model diagnostics

1. The coefficient of determination

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

where the sums of squares of error is

$$SSE = RSS(\hat{\beta}) = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

and the total sums of squares in

$$SST = \sum_{i=1}^{N}(y_i - \bar{y})^2.$$

The regression sums of squares

$$SSR = SST - SSE = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$$

.

Model diagnostics

1. The coefficient of determination

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

where the sums of squares of error is

$$SSE = RSS(\hat{\beta}) = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

and the total sums of squares in

$$SST = \sum_{i=1}^{N}(y_i - \bar{y})^2.$$

The regression sums of squares

$$SSR = SST - SSE = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$$

.

Model diagnostics

1. The coefficient of determination

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

where the sums of squares of error is

$$SSE = RSS(\hat{\beta}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

and the total sums of squares in

$$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2.$$

The regression sums of squares

$$SSR = SST - SSE = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$

.

Model diagnostics

1. The coefficient of determination

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

where the sums of squares of error is

$$SSE = RSS(\hat{\beta}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

and the total sums of squares in

$$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2.$$

The regression sums of squares

$$SSR = SST - SSE = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$

.

Model diagnostics

2. The adjusted coefficient of determination

$$R^2_{adj} = 1 - (1 - R^2)\frac{N-1}{N-p-1} = 1 - \frac{MSE}{MST}$$

where the mean sums of squares of error is

$$MSE = \frac{SSE}{N-p-1} = \widehat{\sigma^2}$$

and the mean total sums of squares in

$$MST = \frac{SST}{N-1}.$$

The closer R^2 and R^2_{adj} are to one (or 100%), the better the fit is. (Note: $R^2_{adj} \leq R^2$.)

Model diagnostics

2. The adjusted coefficient of determination

$$R^2_{adj} = 1 - (1 - R^2)\frac{N-1}{N-p-1} = 1 - \frac{MSE}{MST}$$

where the mean sums of squares of error is

$$MSE = \frac{SSE}{N-p-1} = \widehat{\sigma^2}$$

and the mean total sums of squares in

$$MST = \frac{SST}{N-1}.$$

The closer R^2 and R^2_{adj} are to one (or 100%), the better the fit is. (Note: $R^2_{adj} \leq R^2$.)

Model diagnostics

2. The adjusted coefficient of determination

$$R^2_{adj} = 1 - (1 - R^2)\frac{N-1}{N-p-1} = 1 - \frac{MSE}{MST}$$

where the mean sums of squares of error is

$$MSE = \frac{SSE}{N-p-1} = \widehat{\sigma^2}$$

and the mean total sums of squares in

$$MST = \frac{SST}{N-1}.$$

The closer R^2 and R^2_{adj} are to one (or 100%), the better the fit is. (Note: $R^2_{adj} \leq R^2$.)

Model diagnostics

2. The adjusted coefficient of determination

$$R^2_{adj} = 1 - (1 - R^2)\frac{N-1}{N-p-1} = 1 - \frac{MSE}{MST}$$

where the mean sums of squares of error is

$$MSE = \frac{SSE}{N-p-1} = \widehat{\sigma^2}$$

and the mean total sums of squares in

$$MST = \frac{SST}{N-1}.$$

The closer R^2 and R^2_{adj} are to one (or 100%), the better the fit is. (Note: $R^2_{adj} \leq R^2$.)

Model diagnostics

3. Residual analyses to make sure of the homogeneity (to see no pattern in scatter plots of residuals vs fitted values) and normality of the residuals using Normal Q-Q plot and Shapiro-Wilk test.

4. Tests of outliers (points standing far away from the bulk of the data) and influential points (which if removed, result in significant change to the model).

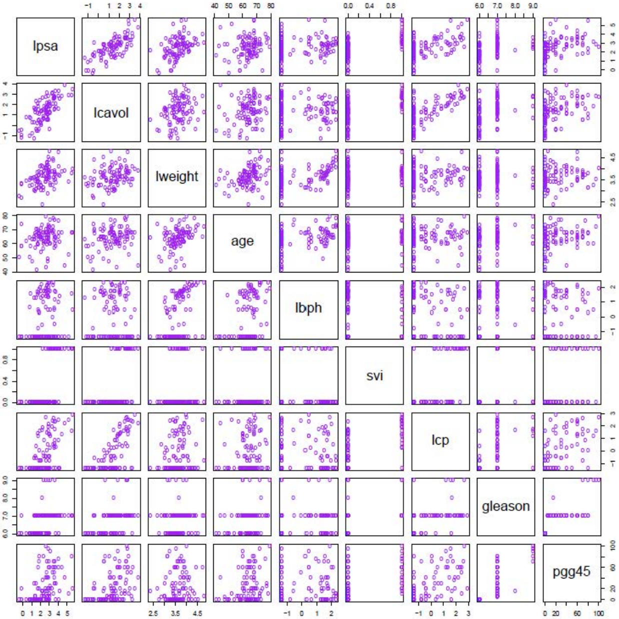
Model diagnostics

3. Residual analyses to make sure of the homogeneity (to see no pattern in scatter plots of residuals vs fitted values) and normality of the residuals using Normal Q-Q plot and Shapiro-Wilk test.

4. Tests of outliers (points standing far away from the bulk of the data) and influential points (which if removed, result in significant change to the model).

Example

Example: Prostate Cancer

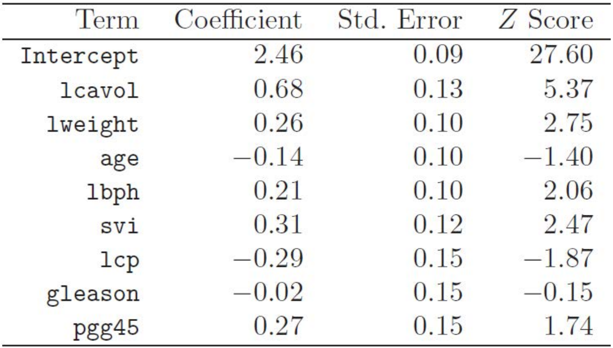


Example: Prostate Cancer

$N = 67$ and $p = 8$.

	lcavol	lweight	age	lbph	svi	lcp	gleason
lweight	0.300						
age	0.286	0.317					
lbph	0.063	0.437	0.287				
svi	0.593	0.181	0.129	−0.139			
lcp	0.692	0.157	0.173	−0.089	0.671		
gleason	0.426	0.024	0.366	0.033	0.307	0.476	
pgg45	0.483	0.074	0.276	−0.030	0.481	0.663	0.757

Example: Prostate Cancer



Term	Coefficient	Std. Error	Z Score
`Intercept`	2.46	0.09	27.60
`lcavol`	0.68	0.13	5.37
`lweight`	0.26	0.10	2.75
`age`	−0.14	0.10	−1.40
`lbph`	0.21	0.10	2.06
`svi`	0.31	0.12	2.47
`lcp`	−0.29	0.15	−1.87
`gleason`	−0.02	0.15	−0.15
`pgg45`	0.27	0.15	1.74

Example: Prostate Cancer

Dropping the least significant inputs: age, lcp, gleason, and pgg45, leads to F test statistics

$$f = \frac{(32.81 - 29.43)/4}{29.43/(67 - 8 - 1)} = 1.67$$

with $p - value = P(F_{4,58} > 1.67) = .17$ which is not significant. Thus, it is concluded to remove those inputs.

Is LS the best method for prediction?

The Gauss-Markov Theorem

Recall: $c^T\hat{\beta} = c^T(X^TX)^{-1}X^Ty =: c_0^Ty$ is unbiased (linear) estimator of $c^T\beta$ and $Var(c^T\hat{\beta}) = c^T(X^TX)^{-1}c\,\sigma^2$.

Theorem (The Gauss-Markov Theorem)

Let c_1^Ty be another unbiased (linear) estimator of $c^T\beta$,then

$$Var(c^T\hat{\beta}) \le Var(c_1^Ty)$$

□

In general, the mean squared error

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta}-\theta)^2 \\ &= Var(\hat{\theta}) + [\underbrace{E(\hat{\theta})-\theta}_{\text{Bias}(\hat{\theta})}]^2 \end{aligned}$$

The Gauss-Markov Theorem

Recall: $c^T\hat{\beta} = c^T(X^TX)^{-1}X^Ty =: c_0^Ty$ is unbiased (linear) estimator of $c^T\beta$ and $Var(c^T\hat{\beta}) = c^T(X^TX)^{-1}c\,\sigma^2$.

Theorem (The Gauss-Markov Theorem)

Let c_1^Ty be another unbiased (linear) estimator of $c^T\beta$,then

$$Var(c^T\hat{\beta}) \leq Var(c_1^Ty)$$

□

In general, the mean squared error

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &= Var(\hat{\theta}) + [\underbrace{E(\hat{\theta}) - \theta}_{\text{Bias}(\hat{\theta})}]^2 \end{aligned}$$

The Gauss-Markov Theorem

Recall: $c^T\hat{\beta} = c^T(X^TX)^{-1}X^Ty =: c_0^Ty$ is unbiased (linear) estimator of $c^T\beta$ and $Var(c^T\hat{\beta}) = c^T(X^TX)^{-1}c\,\sigma^2$.

Theorem (The Gauss-Markov Theorem)

Let c_1^Ty be another unbiased (linear) estimator of $c^T\beta$,then

$$Var(c^T\hat{\beta}) \leq Var(c_1^Ty)$$

□

In general, the mean squared error

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta}-\theta)^2 \\ &= Var(\hat{\theta}) + [\underbrace{E(\hat{\theta})-\theta}_{\text{Bias}(\hat{\theta})}]^2 \end{aligned}$$

The Gauss-Markov Theorem

How is it related to the expected prediction error (EPE) for $Y_* = f(x_*) + \epsilon_*$?

$$\begin{aligned} EPE &= E(Y_* - \hat{f}(x_*))^2 \\ &= E(\hat{f}(x_*) - f(x_*))^2 + \sigma^2 \\ &= MSE(\hat{f}(x_*)) + \sigma^2 \\ &= MSE(x_*^T \hat{\beta}) + \sigma^2 \end{aligned}$$

Thus, a small $MSE(x_*^T \hat{\beta})$ is better for prediction, even when $Bias(x_*^T \hat{\beta}) > 0$.

So, smaller number of predictors (shrinking) might be advised over a more detailed model. Also, a method other than OLS with smaller MSE, is more advisable for prediction.

The Gauss-Markov Theorem

How is it related to the expected prediction error (EPE) for $Y_* = f(x_*) + \epsilon_*$?

$$\begin{aligned} EPE &= E(Y_* - \hat{f}(x_*))^2 \\ &= E(\hat{f}(x_*) - f(x_*))^2 + \sigma^2 \\ &= MSE(\hat{f}(x_*)) + \sigma^2 \\ &= MSE(x_*^T \hat{\beta}) + \sigma^2 \end{aligned}$$

Thus, a small $MSE(x_*^T \hat{\beta})$ is better for prediction, even when $Bias(x_*^T \hat{\beta}) > 0$.

So, smaller number of predictors (shrinking) might be advised over a more detailed model. Also, a method other than OLS with smaller MSE, is more advisable for prediction.

The Gauss-Markov Theorem

How is it related to the expected prediction error (EPE) for $Y_* = f(x_*) + \epsilon_*$?

$$\begin{aligned} EPE &= E(Y_* - \hat{f}(x_*))^2 \\ &= E(\hat{f}(x_*) - f(x_*))^2 + \sigma^2 \\ &= MSE(\hat{f}(x_*)) + \sigma^2 \\ &= MSE(x_*^T\hat{\beta}) + \sigma^2 \end{aligned}$$

Thus, a small $MSE(x_*^T\hat{\beta})$ is better for prediction, even when $Bias(x_*^T\hat{\beta}) > 0$.

So, smaller number of predictors (shrinking) might be advised over a more detailed model. Also, a method other than OLS with smaller MSE, is more advisable for prediction.

Subset (Variable) Selection

Subset (Variable) Selection

- Part of model selection.
- Objective: select one of the 2^p possible subsets of variables/models (including the null regression).
- Methods:
  1. Best Subset method: search for the smallest RSS among all of the 2^p models. Note: $RSS(\hat{\beta}_{full}) < RSS(\hat{\beta}_{subset})$.

Example (Prostate Cancer)

Subset (Variable) Selection

- Part of model selection.
- Objective: select one of the 2^p possible subsets of variables/models (including the null regression).
- Methods:
  1. Best Subset method: search for the smallest RSS among all of the 2^p models. Note: $RSS(\hat{\beta}_{full}) < RSS(\hat{\beta}_{subset})$.

Example (Prostate Cancer)

Subset (Variable) Selection

- Part of model selection.
- Objective: select one of the 2^p possible subsets of variables/models (including the null regression).
- Methods:
  1. Best Subset method: search for the smallest RSS among all of the 2^p models. Note: $RSS(\hat{\beta}_{full}) < RSS(\hat{\beta}_{subset})$.

Example (Prostate Cancer)

Subset (Variable) Selection

- Part of model selection.
- Objective: select one of the 2^p possible subsets of variables/models (including the null regression).
- Methods:
  1. Best Subset method: search for the smallest RSS among all of the 2^p models. Note: $RSS(\hat{\beta}_{full}) < RSS(\hat{\beta}_{subset})$.

Example (Prostate Cancer)

Subset (Variable) Selection

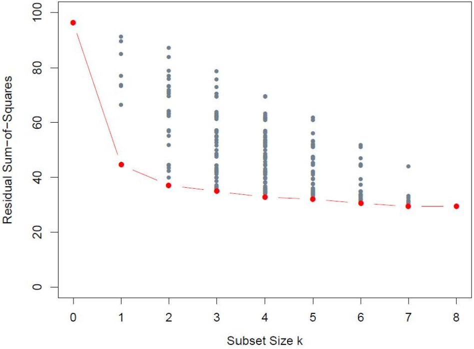
- Part of model selection.
- Objective: select one of the 2^p possible subsets of variables/models (including the null regression).
- Methods:
  1. Best Subset method: search for the smallest RSS among all of the 2^p models. Note: $RSS(\hat{\beta}_{full}) < RSS(\hat{\beta}_{subset})$.

Example (Prostate Cancer)

Subset (Variable) Selection

- Part of model selection.
- Objective: select one of the 2^p possible subsets of variables/models (including the null regression).
- Methods:
  1. Best Subset method: search for the smallest RSS among all of the 2^p models. Note: $RSS(\hat{\beta}_{full}) < RSS(\hat{\beta}_{subset})$.

Example (Prostate Cancer)

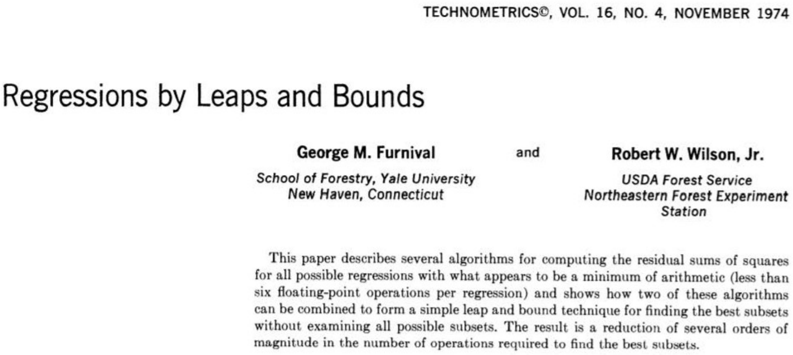


Subset (Variable) Selection

2. Leaps and bounds (good for $p \leq 40$, minimizes RSS). Also, Branch and Bounds.

Subset (Variable) Selection

2. Leaps and bounds (good for $p \leq 40$, minimizes RSS). Also, Branch and Bounds.



TECHNOMETRICS©, VOL. 16, NO. 4, NOVEMBER 1974

Regressions by Leaps and Bounds

George M. Furnival
School of Forestry, Yale University
New Haven, Connecticut

and

Robert W. Wilson, Jr.
USDA Forest Service
Northeastern Forest Experiment
Station

This paper describes several algorithms for computing the residual sums of squares for all possible regressions with what appears to be a minimum of arithmetic (less than six floating-point operations per regression) and shows how two of these algorithms can be combined to form a simple leap and bound technique for finding the best subsets without examining all possible subsets. The result is a reduction of several orders of magnitude in the number of operations required to find the best subsets.

Subset (Variable) Selection

2. Leaps and bounds (good for $p \leq 40$, minimizes RSS). Also, Branch and Bounds.

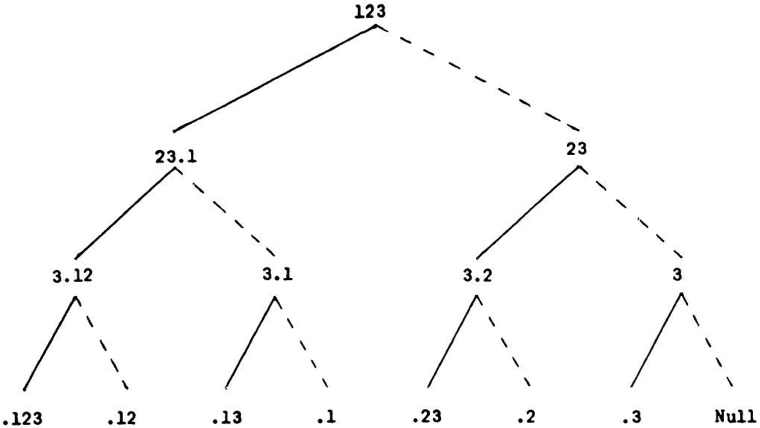


FIGURE 1—The regression tree

Subset (Variable) Selection

2. Leaps and bounds (good for $p \leq 40$, minimizes RSS). Also, Branch and Bounds.

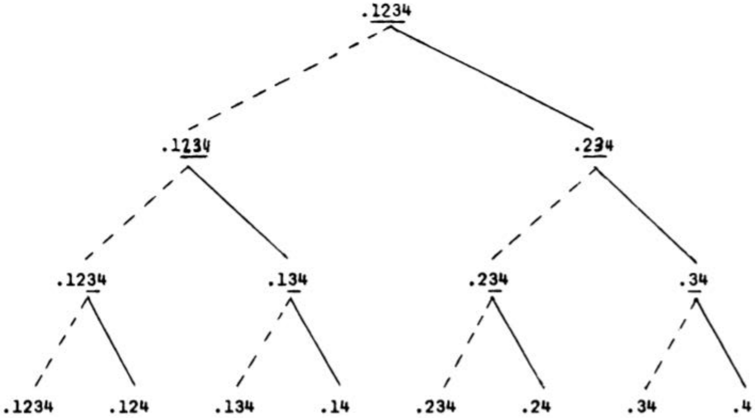


FIGURE 2—The bound tree

TECHNOMETRICS©, VOL. 16, NO. 4, NOVEMBER 1974

Subset (Variable) Selection

3. Stepwise (Forward or Backward) Selection (when $p > 40$).
   - Forward-stepwise selection (is a greedy algorithm): start with a null model (just the intercept $\hat{\beta}_0 = \bar{y}$) and then sequentially adds predictors that improves the fit. Models on the steps forward are nested. Good at all cases.
   - Backward-stepwise selection: start with a full model (all the predictors) and then sequentially removes predictors that do not alter the fit (smallest t- or z- score). Use only when $N > p$.

Subset (Variable) Selection

3. Stepwise (Forward or Backward) Selection (when $p > 40$).
   - Forward-stepwise selection (is a greedy algorithm): start with a null model (just the intercept $\hat{\beta}_0 = \bar{y}$) and then sequentially adds predictors that improves the fit. Models on the steps forward are nested. Good at all cases.
   - Backward-stepwise selection: start with a full model (all the predictors) and then sequentially removes predictors that do not alter the fit (smallest t- or z- score). Use only when $N > p$.

Subset (Variable) Selection

3. Stepwise (Forward or Backward) Selection (when $p > 40$).
   - Forward-stepwise selection (is a greedy algorithm): start with a null model (just the intercept $\hat{\beta}_0 = \bar{y}$) and then sequentially adds predictors that improves the fit. Models on the steps forward are nested. Good at all cases.
   - Backward-stepwise selection: start with a full model (all the predictors) and then sequentially removes predictors that do not alter the fit (smallest t- or z- score). Use only when $N > p$.

Subset (Variable) Selection

4. Forward-Stagewise Regression:

Stage 0: Start with $\hat{\beta}_{0,0} = \bar{y}$ and $\hat{\beta}_{j,0} = 0$ for $j = 1, 2, \ldots, p$.

Stage k: Find the most correlated variable, say X_j, with the residuals of the model in Stage $k-1$ and find the slope (b_j) of the simple linear regression between the residuals and that variable X_j.

$$\hat{\beta}_{j,k} = \hat{\beta}_{j,k-1} + b_j$$

Until: there is no correlation between the residuals and any variable.

⇓ Slow and might need more than p stages till converge.

⇑ Good for high dimensional problems.

Subset (Variable) Selection

4. Forward-Stagewise Regression:

Stage 0: Start with $\hat{\beta}_{0,0} = \bar{y}$ and $\hat{\beta}_{j,0} = 0$ for $j = 1, 2, \ldots, p$.

Stage k: Find the most correlated variable, say X_j, with the residuals of the model in Stage $k - 1$ and find the slope (b_j) of the simple linear regression between the residuals and that variable X_j.

$$\hat{\beta}_{j,k} = \hat{\beta}_{j,k-1} + b_j$$

Until: there is no correlation between the residuals and any variable.

⇓ Slow and might need more than p stages till converge.

⇑ Good for high dimensional problems.

Subset (Variable) Selection

4. Forward-Stagewise Regression:

Stage 0: Start with $\hat{\beta}_{0,0} = \bar{y}$ and $\hat{\beta}_{j,0} = 0$ for $j = 1, 2, \ldots, p$.

Stage k: Find the most correlated variable, say X_j, with the residuals of the model in Stage $k-1$ and find the slope (b_j) of the simple linear regression between the residuals and that variable X_j.

$$\hat{\beta}_{j,k} = \hat{\beta}_{j,k-1} + b_j$$

Until: there is no correlation between the residuals and any variable.

⇓ Slow and might need more than p stages till converge.

⇑ Good for high dimensional problems.

Subset (Variable) Selection

4. Forward-Stagewise Regression:

Stage 0: Start with $\hat{\beta}_{0,0} = \bar{y}$ and $\hat{\beta}_{j,0} = 0$ for $j = 1, 2, \ldots, p$.

Stage k: Find the most correlated variable, say X_j, with the residuals of the model in Stage $k-1$ and find the slope (b_j) of the simple linear regression between the residuals and that variable X_j.

$$\hat{\beta}_{j,k} = \hat{\beta}_{j,k-1} + b_j$$

Until: there is no correlation between the residuals and any variable.

⇓ Slow and might need more than p stages till converge.

⇑ Good for high dimensional problems.

Subset (Variable) Selection

4. Forward-Stagewise Regression:

Stage 0: Start with $\hat{\beta}_{0,0} = \bar{y}$ and $\hat{\beta}_{j,0} = 0$ for $j = 1, 2, \ldots, p$.

Stage k: Find the most correlated variable, say X_j, with the residuals of the model in Stage $k - 1$ and find the slope (b_j) of the simple linear regression between the residuals and that variable X_j.

$$\hat{\beta}_{j,k} = \hat{\beta}_{j,k-1} + b_j$$

Until: there is no correlation between the residuals and any variable.

$\Downarrow$ Slow and might need more than p stages till converge.

$\Uparrow$ Good for high dimensional problems.

Subset (Variable) Selection

4. Forward-Stagewise Regression:

Stage 0: Start with $\hat{\beta}_{0,0} = \bar{y}$ and $\hat{\beta}_{j,0} = 0$ for $j = 1, 2, \ldots, p$.

Stage k: Find the most correlated variable, say X_j, with the residuals of the model in Stage $k-1$ and find the slope (b_j) of the simple linear regression between the residuals and that variable X_j.

$$\hat{\beta}_{j,k} = \hat{\beta}_{j,k-1} + b_j$$

Until: there is no correlation between the residuals and any variable.

$\Downarrow$ Slow and might need more than p stages till converge.

$\Uparrow$ Good for high dimensional problems.

Subset (Variable) Selection

4. Forward-Stagewise Regression:

Stage 0: Start with $\hat{\beta}_{0,0} = \bar{y}$ and $\hat{\beta}_{j,0} = 0$ for $j = 1, 2, \ldots, p$.

Stage k: Find the most correlated variable, say X_j, with the residuals of the model in Stage $k - 1$ and find the slope (b_j) of the simple linear regression between the residuals and that variable X_j.

$$\hat{\beta}_{j,k} = \hat{\beta}_{j,k-1} + b_j$$

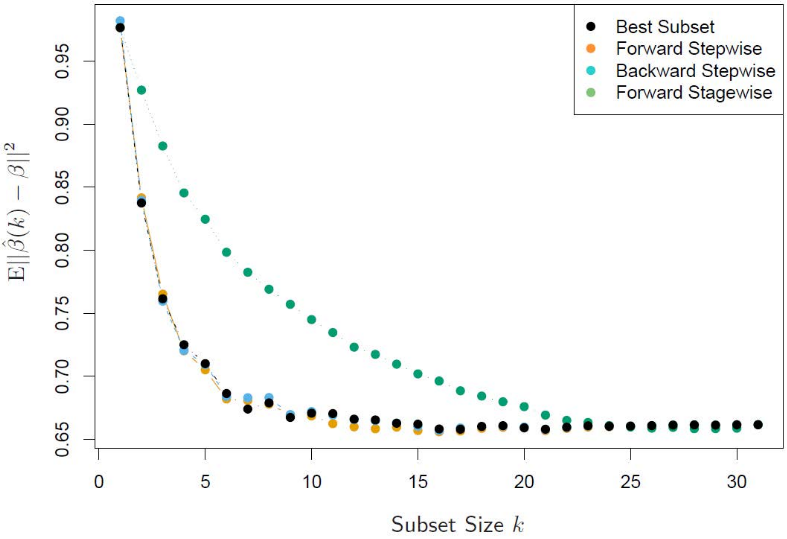
Until: there is no correlation between the residuals and any variable.

$\Downarrow$ Slow and might need more than p stages till converge.

$\Uparrow$ Good for high dimensional problems.

Subset (Variable) Selection

In a simulation study, with $N = 300$ and $p = 31$.



Subset (Variable) Selection

Measures of selection

1. Largest R^2 or R^2_{adj}.
2. Smallest RSS.
3. Smallest CV or GCV.
4. Smallest Mallow's C_p:

$$C_p = \frac{RSS_{\text{subset of } k}}{RSS(\hat{\beta}_{full})/(N - p - 1)} - (N - 2k)$$

Subset (Variable) Selection

Measures of selection

1. Largest R^2 or R^2_{adj}.
2. Smallest RSS.
3. Smallest CV or GCV.
4. Smallest Mallow's C_p:

$$C_p = \frac{RSS_{\text{subset of } k}}{RSS(\hat{\beta}_{full})/(N-p-1)} - (N-2k)$$

Subset (Variable) Selection

Measures of selection

1. Largest R^2 or R^2_{adj}.
2. Smallest RSS.
3. Smallest CV or GCV.
4. Smallest Mallow's C_p:

$$C_p = \frac{RSS_{\text{subset of } k}}{RSS(\hat{\beta}_{full})/(N-p-1)} - (N-2k)$$

Subset (Variable) Selection

Measures of selection

1. Largest R^2 or R^2_{adj}.
2. Smallest RSS.
3. Smallest CV or GCV.
4. Smallest Mallow's C_p:

$$C_p = \frac{RSS_{\text{subset of } k}}{RSS(\hat{\beta}_{full})/(N-p-1)} - (N-2k)$$

Subset (Variable) Selection

More measures of selection: (For general classes of models.) Let L be the likelihood function. $\hat{\beta}_{MLE,k}$ is the maximum likelihood estimator of size k.

1. Smallest
$$deviance = -2 \log L(\hat{\beta}_{MLE,k})$$
2. Smallest Akaike's Information Criterion
$$AIC_k = -2 \log L(\hat{\beta}_{MLE,k}) + 2k$$
3. Smallest Bayes' Information Criterion
$$BIC_k = -2 \log L(\hat{\beta}_{MLE,k}) + 2k \log(N)$$

Subset (Variable) Selection

More measures of selection: (For general classes of models.) Let L be the likelihood function. $\hat{\beta}_{MLE,k}$ is the maximum likelihood estimator of size k.

1. Smallest

$$deviance = -2\log L(\hat{\beta}_{MLE,k})$$

2. Smallest Akaike's Information Criterion

$$AIC_k = -2\log L(\hat{\beta}_{MLE,k}) + 2k$$

3. Smallest Bayes' Information Criterion

$$BIC_k = -2\log L(\hat{\beta}_{MLE,k}) + 2k\log(N)$$

Subset (Variable) Selection

More measures of selection: (For general classes of models.) Let L be the likelihood function. $\hat{\beta}_{MLE,k}$ is the maximum likelihood estimator of size k.

1. Smallest

$$deviance = -2 \log L(\hat{\beta}_{MLE,k})$$

2. Smallest Akaike's Information Criterion

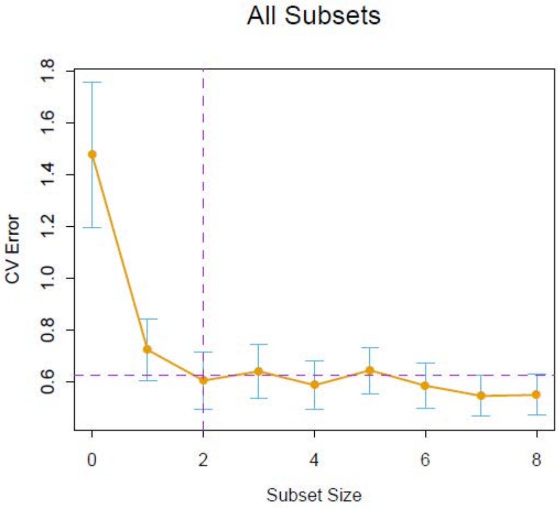
$$AIC_k = -2 \log L(\hat{\beta}_{MLE,k}) + 2k$$

3. Smallest Bayes' Information Criterion

$$BIC_k = -2 \log L(\hat{\beta}_{MLE,k}) + 2k \log(N)$$

Subset (Variable) Selection

Example (Prostate Cancer)



Shrinkage (regularization,constraints)

Shrinkage

- It includes subset selection. But, it is continuous selection rather than discrete.
- Objective: To include all of the p inputs but shrinking their coefficients towards zero. If some of them become zero, then it results in a subset. (Note: Intercept is not included in that objective.)
- It reduces variance of the estimates.

Shrinkage

- It includes subset selection. But, it is continuous selection rather than discrete.
- Objective: To include all of the p inputs but shrinking their coefficients towards zero. If some of them become zero, then it results in a subset. (Note: Intercept is not included in that objective.)
- It reduces variance of the estimates.

Shrinkage

- It includes subset selection. But, it is continuous selection rather than discrete.
- Objective: To include all of the p inputs but shrinking their coefficients towards zero. If some of them become zero, then it results in a subset. (Note: Intercept is not included in that objective.)
- It reduces variance of the estimates.

Shrinkage

To find $\hat{\beta}^{\text{shrunk}}$ that

$$\text{minimize}_{\beta}\ RSS(\beta) = \sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2$$

$$\text{subject to } \sum_{j=1}^{p} G(\beta_j) \leq t \text{ (size constraint)}$$

OR $\hat{\beta}^{\text{shrunk}} =$

$$argmin_{\beta}\left[\sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 + \lambda \sum_{j=1}^{p} G(\beta_j)\right]$$

for some positive function G. The term $\lambda \sum_{j=1}^{p} G(\beta_j)$ is called shrinkage penalty.

Shrinkage

To find $\hat{\beta}^{\text{shrunk}}$ that

$$\text{minimize}_{\beta}\ RSS(\beta) = \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2$$

$$\text{subject to } \sum_{j=1}^{p} G(\beta_j) \leq t \text{ (size constraint)}$$

OR $\hat{\beta}^{\text{shrunk}} =$

$$argmin_{\beta} \left[\sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 + \lambda \sum_{j=1}^{p} G(\beta_j) \right]$$

for some positive function G. The term $\lambda \sum_{j=1}^{p} G(\beta_j)$ is called shrinkage penalty.

Shrinkage

- Some methods:

1. Ridge regression,$G(x) = x^2$. (An L_2 shrinkage method.)

2. Least absolute shrinkage and selection operator (lasso), $G(x) = |x|$. (An L_1 shrinkage method.)

3. Bridge shrinkage,

$$G(x) = \begin{cases} |x|^q & \text{if } q > 0, \\ I(x \neq 0) & \text{if } q = 0. \end{cases}$$

(An L_q shrinkage method.) It includes both ridge and lasso.

Shrinkage

- Some methods:

1. Ridge regression,$G(x) = x^2$. (An L_2 shrinkage method.)

2. Least absolute shrinkage and selection operator (lasso), $G(x) = |x|$. (An L_1 shrinkage method.)

3. Bridge shrinkage,

$$G(x) = \begin{cases} |x|^q & \text{if } q > 0, \\ I(x \neq 0) & \text{if } q = 0. \end{cases}$$

(An L_q shrinkage method.) It includes both ridge and lasso.

Shrinkage

- Some methods:

1. Ridge regression,$G(x) = x^2$. (An L_2 shrinkage method.)

2. Least absolute shrinkage and selection operator (lasso), $G(x) = |x|$. (An L_1 shrinkage method.)

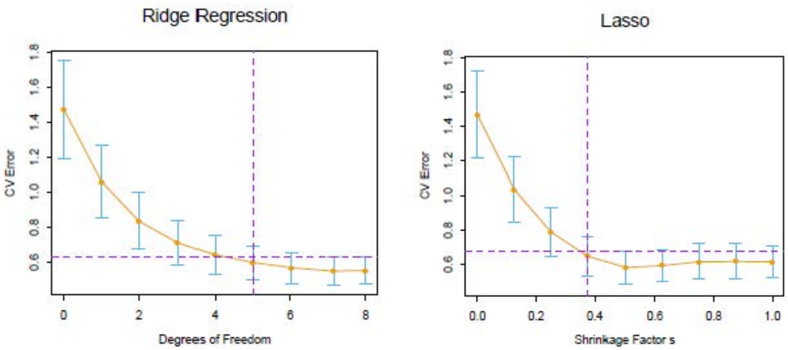
3. Bridge shrinkage,

$$G(x) = \begin{cases} |x|^q & \text{if } q > 0, \\ I(x \neq 0) & \text{if } q = 0. \end{cases}$$

   (An L_q shrinkage method.) It includes both ridge and lasso.

Shrinkage

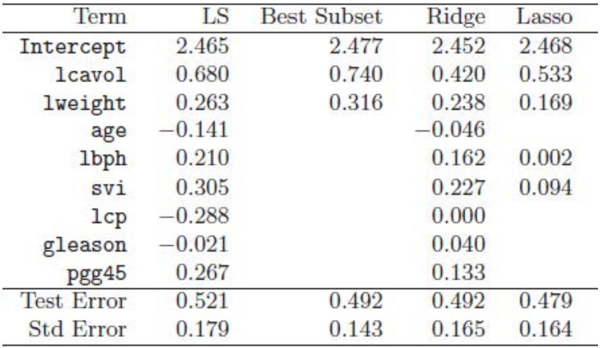
Example (Prostate Cancer)



Shrinkage

Example (Prostate Cancer)

Estimated coefficients are



Term	LS	Best Subset	Ridge	Lasso
Intercept	2.465	2.477	2.452	2.468
lcavol	0.680	0.740	0.420	0.533
lweight	0.263	0.316	0.238	0.169
age	−0.141		−0.046	
lbph	0.210		0.162	0.002
svi	0.305		0.227	0.094
lcp	−0.288		0.000	
gleason	−0.021		0.040	
pgg45	0.267		0.133	
Test Error	0.521	0.492	0.492	0.479
Std Error	0.179	0.143	0.165	0.164

Ridge Regression

Ridge Regression

To find $\hat{\beta}^{\text{ridge}}$ that

$$\text{minimize}_{\beta}\ RSS(\beta) = \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t \text{ (size constraint)}$$

OR in the Lagrangian form

$$\hat{\beta}^{\text{ridge}} = argmin_{\beta} \left[\sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right]$$

Ridge Regression

To find $\hat{\beta}^{\text{ridge}}$ that

$$\text{minimize}_{\beta}\ RSS(\beta) = \sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t \text{ (size constraint)}$$

OR in the Lagrangian form

$$\hat{\beta}^{\text{ridge}} = argmin_{\beta} \left[\sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right]$$

Ridge Regression

- The decay/tuning parameter $\lambda \geq 0$ is determined first through CV then the parameters are estimated.

- What does happen when λ increase?

Ridge Regression

Better, start with standardized data:

$$\sum_{i=1}^{N} x_{ij} = 0, \sum_{i=1}^{N} x_{ij}^2 = 1$$

which results in removing $\hat{\beta}_0$ from the optimization problem as its value would be $\bar{y}$. We are now left with a $p \times p$ matrix X.

Ridge Regression

The problem is now equivalent to find $\hat{\beta}^{\text{ridge}}$ that

$$\text{minimize}_{\beta}\ RSS(\beta) = (y - X\beta)^T(y - X\beta)$$

$$\text{subject to } \beta^T\beta \leq t$$

OR

$$\hat{\beta}^{\text{ridge}} = argmin_{\beta}\left[(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta\right]$$

Call:

$$RSS_{\lambda}(\beta) := (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

Ridge Regression

The problem is now equivalent to find $\hat{\beta}^{\text{ridge}}$ that

$$\text{minimize}_{\beta}\ RSS(\beta) = (y - X\beta)^T(y - X\beta)$$

$$\text{subject to } \beta^T\beta \leq t$$

OR

$$\hat{\beta}^{\text{ridge}} = argmin_{\beta}\left[(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta\right]$$

Call:

$$RSS_{\lambda}(\beta) := (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

Ridge Regression

The problem is now equivalent to find $\hat{\beta}^{\text{ridge}}$ that

$$\text{minimize}_{\beta}\ RSS(\beta) = (y - X\beta)^T(y - X\beta)$$

$$\text{subject to } \beta^T\beta \leq t$$

OR

$$\hat{\beta}^{\text{ridge}} = argmin_{\beta}\left[(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta\right]$$

Call:

$$RSS_{\lambda}(\beta) := (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

Ridge Regression

- $\dfrac{\partial RSS_\lambda(\beta)}{\partial\beta} = -2X^T(y - X\beta) + 2\lambda\beta = 0 \implies$

  $$(X^TX + \lambda I_p)\beta = X^Ty$$

  where I_p is the $p \times p$ identity matrix.

- $\dfrac{\partial^2 RSS_\lambda(\beta)}{\partial\beta\partial\beta^T} = 2X^TX + 2\lambda I_p$
- Even when X is not a full column rank, $X^TX + \lambda I_p$ is positive definite for $\lambda > 0$ and so non-singular, then

  $$\hat{\beta}_\lambda^{\text{ridge}} = (X^TX + \lambda I_p)^{-1}X^Ty$$

- Predictions

  $$\hat{y}_\lambda = X\hat{\beta}_\lambda^{\text{ridge}} = \underbrace{X(X^TX + \lambda I_p)^{-1}X^T}_{\text{the } \lambda\text{-hat matrix } H_\lambda} y$$

Ridge Regression

- $\dfrac{\partial RSS_\lambda(\beta)}{\partial \beta} = -2X^T(y - X\beta) + 2\lambda\beta = 0 \implies$

  $$(X^TX + \lambda I_p)\beta = X^Ty$$

  where I_p is the $p \times p$ identity matrix.
- $\dfrac{\partial^2 RSS_\lambda(\beta)}{\partial\beta\partial\beta^T} = 2X^TX + 2\lambda I_p$
- Even when X is not a full column rank, $X^TX + \lambda I_p$ is positive definite for $\lambda > 0$ and so non-singular, then

  $$\hat{\beta}_\lambda^{\text{ridge}} = (X^TX + \lambda I_p)^{-1}X^Ty$$

- Predictions

  $$\hat{y}_\lambda = X\hat{\beta}_\lambda^{\text{ridge}} = \underbrace{X(X^TX + \lambda I_p)^{-1}X^T}_{\text{the } \lambda\text{-hat matrix } H_\lambda}y$$

Ridge Regression

- $$\frac{\partial RSS_\lambda(\beta)}{\partial\beta} = -2X^T(y - X\beta) + 2\lambda\beta = 0 \implies$$

  $$(X^TX + \lambda I_p)\beta = X^Ty$$

  where I_p is the $p \times p$ identity matrix.
- $$\frac{\partial^2 RSS_\lambda(\beta)}{\partial\beta\partial\beta^T} = 2X^TX + 2\lambda I_p$$
- Even when X is not a full column rank, $X^TX + \lambda I_p$ is positive definite for $\lambda > 0$ and so non-singular, then

  $$\hat{\beta}_\lambda^{\text{ridge}} = (X^TX + \lambda I_p)^{-1}X^Ty$$

- Predictions

  $$\hat{y}_\lambda = X\hat{\beta}_\lambda^{\text{ridge}} = \underbrace{X(X^TX + \lambda I_p)^{-1}X^T}_{\text{the } \lambda\text{-hat matrix } H_\lambda} y$$

Ridge Regression

- $\dfrac{\partial RSS_\lambda(\beta)}{\partial \beta} = -2X^T(y - X\beta) + 2\lambda\beta = 0 \implies$

$$(X^TX + \lambda I_p)\beta = X^Ty$$

  where I_p is the $p \times p$ identity matrix.

- $\dfrac{\partial^2 RSS_\lambda(\beta)}{\partial\beta\partial\beta^T} = 2X^TX + 2\lambda I_p$
- Even when X is not a full column rank, $X^TX + \lambda I_p$ is positive definite for $\lambda > 0$ and so non-singular, then

$$\hat{\beta}_\lambda^{\text{ridge}} = (X^TX + \lambda I_p)^{-1}X^Ty$$

- Predictions

$$\hat{y}_\lambda = X\hat{\beta}_\lambda^{\text{ridge}} = \underbrace{X(X^TX + \lambda I_p)^{-1}X^T}_{\text{the } \lambda\text{-hat matrix } H_\lambda} y$$

Ridge Regression

Again, the solution is

$$\hat{\beta}_{\lambda}^{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T y$$

- What does happen when λ decreases to zero?
- If columns of X are orthonormal ($X^T X = I$), then

$$\hat{\beta}_{\lambda}^{\text{ridge}} = \frac{1}{1+\lambda} \hat{\beta}^{ols}$$

- In general, $\hat{\beta}_{\lambda}^{\text{ridge}}$ is a biased estimator of β. (Good problem to prove it, hint: $E(Az) = A\,E(z)$.)
- Yet, it has smaller variance than that of the OLS's. (Another good problem, hint: $Var(Az) = A\,Var(z)\,A^T$.)

Ridge Regression

Again, the solution is

$$\hat{\beta}_{\lambda}^{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T y$$

- What does happen when λ decreases to zero?
- If columns of X are orthonormal ($X^T X = I$), then

$$\hat{\beta}_{\lambda}^{\text{ridge}} = \frac{1}{1+\lambda} \hat{\beta}^{ols}$$

- In general, $\hat{\beta}_{\lambda}^{\text{ridge}}$ is a biased estimator of β. (Good problem to prove it, hint: $E(Az) = A\,E(z)$.)
- Yet, it has smaller variance than that of the OLS's. (Another good problem, hint: $Var(Az) = A\,Var(z)\,A^T$.)

Ridge Regression

Again, the solution is

$$\hat{\beta}_{\lambda}^{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T y$$

- What does happen when λ decreases to zero?
- If columns of X are orthonormal ($X^T X = I$), then

$$\hat{\beta}_{\lambda}^{\text{ridge}} = \frac{1}{1+\lambda} \hat{\beta}^{ols}$$

- In general, $\hat{\beta}_{\lambda}^{\text{ridge}}$ is a biased estimator of β. (Good problem to prove it, hint: $E(Az) = A\,E(z)$.)
- Yet, it has smaller variance than that of the OLS's. (Another good problem, hint: $Var(Az) = A\,Var(z)\,A^T$.)

Ridge Regression

Again, the solution is

$$\hat{\beta}_{\lambda}^{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T y$$

- What does happen when λ decreases to zero?
- If columns of X are orthonormal ($X^T X = I$), then

$$\hat{\beta}_{\lambda}^{\text{ridge}} = \frac{1}{1+\lambda} \hat{\beta}^{ols}$$

- In general, $\hat{\beta}_{\lambda}^{\text{ridge}}$ is a biased estimator of β. (Good problem to prove it, hint: $E(Az) = A\,E(z)$.)
- Yet, it has smaller variance than that of the OLS's. (Another good problem, hint: $Var(Az) = A\,Var(z)\,A^T$.)

Ridge Regression

It handles very well the case of collinearity, as

- Originally, When a coefficient of a variable becomes large, coefficient of any correlated variables balance up with a very small and negative value. But placing a bound resolves that issue.
- It fixes the problem that X is not column full-rank.

Ridge Regression

It handles very well the case of collinearity, as

- Originally, When a coefficient of a variable becomes large, coefficient of any correlated variables balance up with a very small and negative value. But placing a bound resolves that issue.
- It fixes the problem that X is not column full-rank.

Ridge Regression

Using singular values decomposition (SVD):

$$X = UDV^T$$

Where U and V are two orthogonal matrices, $U^T U = I_p$ and $V^T V = I_p$. The columns u_j and v_j of the $N \times p$ matrix U and the $p \times p$ matrix V are spanning the columns and rows of X, respectively. D is a $p \times p$ diagonal matrix of singular values $d_1 \geq \ldots \geq d_p \geq 0$ (some might be possible 0). Then ...

Ridge Regression

Using singular values decomposition (SVD):

$$X = UDV^T$$

Where U and V are two orthogonal matrices, $U^T U = I_p$ and $V^T V = I_p$. The columns u_j and v_j of the $N \times p$ matrix U and the $p \times p$ matrix V are spanning the columns and rows of X, respectively. D is a $p \times p$ diagonal matrix of singular values $d_1 \geq \ldots \geq d_p \geq 0$ (some might be possible 0). Then ...

Ridge Regression

Then ...

$$
\begin{aligned}
\hat{\beta}_{\lambda}^{\text{ridge}} &= (X^T X + \lambda I_p)^{-1} X^T y \\
&= ((UDV^T)^T (UDV^T) + \lambda I_p)^{-1} (UDV^T)^T y \\
&= (VD^2 V^T + \lambda VV^T)^{-1} VDU^T y \\
&= V \Delta_\lambda U^T y
\end{aligned}
$$

where Δ_λ is a diagonal matrix with elements $d_j/(d_j^2 + \lambda)$, for $j = 1, \ldots, p$.

Ridge Regression

Then ...

$$
\begin{aligned}
\hat{\beta}_{\lambda}^{\text{ridge}} &= (X^TX + \lambda I_p)^{-1}X^Ty \\
&= ((UDV^T)^T(UDV^T) + \lambda I_p)^{-1}(UDV^T)^Ty \\
&= (VD^2V^T + \lambda VV^T)^{-1}VDU^Ty \\
&= V\Delta_{\lambda}U^Ty
\end{aligned}
$$

where Δ_{λ} is a diagonal matrix with elements $d_j/(d_j^2+\lambda)$, for $j = 1, \ldots, p$.

Ridge Regression

Then ...

$$\begin{aligned}\hat{\beta}_{\lambda}^{\text{ridge}} &= (X^T X + \lambda I_p)^{-1} X^T y \\ &= ((UDV^T)^T (UDV^T) + \lambda I_p)^{-1} (UDV^T)^T y \\ &= (VD^2 V^T + \lambda VV^T)^{-1} VDU^T y \\ &= V \Delta_{\lambda} U^T y\end{aligned}$$

where Δ_{λ} is a diagonal matrix with elements $d_j/(d_j^2 + \lambda)$, for $j = 1, \ldots, p$.

Ridge Regression

Then ...

$$\begin{aligned}
\hat{\beta}_\lambda^{\text{ridge}} &= (X^T X + \lambda I_p)^{-1} X^T y \\
&= ((UDV^T)^T (UDV^T) + \lambda I_p)^{-1} (UDV^T)^T y \\
&= (VD^2 V^T + \lambda V V^T)^{-1} V D U^T y \\
&= V \Delta_\lambda U^T y
\end{aligned}$$

where Δ_λ is a diagonal matrix with elements $d_j/(d_j^2 + \lambda)$, for $j = 1, \ldots, p$.

Ridge Regression

Then ...

$$\begin{aligned}
\hat{\beta}_\lambda^{\text{ridge}} &= (X^T X + \lambda I_p)^{-1} X^T y \\
&= ((UDV^T)^T (UDV^T) + \lambda I_p)^{-1} (UDV^T)^T y \\
&= (VD^2 V^T + \lambda VV^T)^{-1} VDU^T y \\
&= V \Delta_\lambda U^T y
\end{aligned}$$

where Δ_λ is a diagonal matrix with elements $d_j/(d_j^2 + \lambda)$, for $j = 1, \ldots, p$.

Ridge Regression

Thus, the prediction is

$$
\begin{aligned}
\hat{y}_\lambda = X\hat{\beta}_\lambda^{\text{ridge}} &= H_\lambda y \\
&= X(X^T X + \lambda I_p)^{-1} X^T y \\
&= (UDV^T) V \Delta_\lambda U^T y \\
&= UD\Delta_\lambda U^T y \\
&= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y
\end{aligned}
$$

Note that, $\hat{y}_0 = UU^T y = \sum_{j=1}^{p} u_j u_j^T y$ is the OLS prediction.

Ridge Regression

Thus, the prediction is

$$
\begin{aligned}
\hat{y}_\lambda = X\hat{\beta}_\lambda^{\text{ridge}} &= H_\lambda y \\
&= X(X^T X + \lambda I_p)^{-1} X^T y \\
&= (UDV^T) V \Delta_\lambda U^T y \\
&= UD\Delta_\lambda U^T y \\
&= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y
\end{aligned}
$$

Note that, $\hat{y}_0 = UU^T y = \sum_{j=1}^{p} u_j u_j^T y$ is the OLS prediction.

Ridge Regression

Thus, the prediction is

$$\begin{aligned}
\hat{y}_\lambda = X\hat{\beta}_\lambda^{\text{ridge}} &= H_\lambda y \\
&= X(X^T X + \lambda I_p)^{-1} X^T y \\
&= (UDV^T) V \Delta_\lambda U^T y \\
&= UD\Delta_\lambda U^T y \\
&= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y
\end{aligned}$$

Note that, $\hat{y}_0 = UU^T y = \sum_{j=1}^{p} u_j u_j^T y$ is the OLS prediction.

Ridge Regression

Thus, the prediction is

$$\begin{aligned}\hat{y}_\lambda = X\hat{\beta}_\lambda^{\text{ridge}} &= H_\lambda y \\ &= X(X^TX+\lambda I_p)^{-1}X^Ty \\ &= (UDV^T)V\Delta_\lambda U^Ty \\ &= UD\Delta_\lambda U^Ty \\ &= \sum_{j=1}^p u_j \frac{d_j^2}{d_j^2+\lambda} u_j^T y\end{aligned}$$

Note that, $\hat{y}_0 = UU^Ty = \sum_{j=1}^p u_j u_j^T y$ is the OLS prediction.

Ridge Regression

Consider centered data $\bar{x}_j = 0$ for all j

- The sample covariance matrix

$$S = X^T X / N = V D^2 V^T / N$$

(eigen decomposition with $V^T S V = D^2 / N$)

- With $d_1^2/N \geq d_2^2/N \geq \cdots \geq d_p^2/N$
- The eigen-vectors v_j's are called the principal components (Karhunen-Loeve) directions of X.
- Xv_1 is the (first) largest principal component since $v_1^T X^T X v_1 = d_1^2/N$ is the largest sample variance among all normalized linear combinations of the columns of X.

e.g. principal components in 2D input data

Ridge Regression

Consider centered data $\bar{x}_j = 0$ for all j

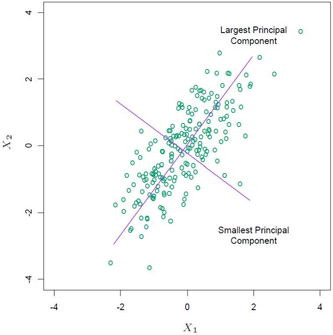
- The sample covariance matrix

$$S = X^T X / N = V D^2 V^T / N$$

(eigen decomposition with $V^T S V = D^2 / N$)

- With $d_1^2/N \geq d_2^2/N \geq \cdots \geq d_p^2/N$
- The eigen-vectors v_j's are called the principal components (Karhunen-Loeve) directions of X.
- Xv_1 is the (first) largest principal component since $v_1^T X^T X v_1 = d_1^2/N$ is the largest sample variance among all normalized linear combinations of the columns of X.

e.g. principal components in 2D input data



Ridge Regression

Consider centered data $\bar{x}_j = 0$ for all j

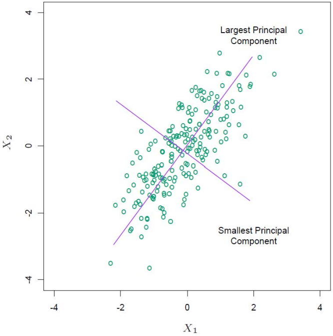
- The sample covariance matrix

$$S = X^T X/N = VD^2V^T/N$$

(eigen decomposition with $V^T SV = D^2/N$)

- With $d_1^2/N \geq d_2^2/N \geq \cdots \geq d_p^2/N$
- The eigen-vectors v_j's are called the principal components (Karhunen-Loeve) directions of X.
- Xv_1 is the (first) largest principal component since $v_1^T X^T X v_1 = d_1^2/N$ is the largest sample variance among all normalized linear combinations of the columns of X.

e.g. principal components in 2D input data



Ridge Regression

Consider centered data $\bar{x}_j = 0$ for all j

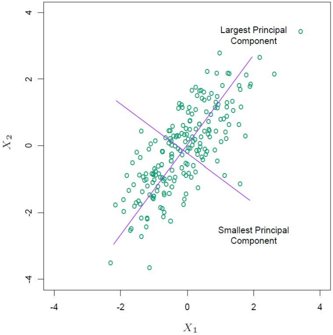
- The sample covariance matrix

$$S = X^T X / N = VD^2 V^T / N$$

(eigen decomposition with $V^T S V = D^2 / N$)

- With $d_1^2/N \geq d_2^2/N \geq \cdots \geq d_p^2/N$
- The eigen-vectors v_j's are called the principal components (Karhunen-Loeve) directions of X.
- Xv_1 is the (first) largest principal component since $v_1^T X^T X v_1 = d_1^2/N$ is the largest sample variance among all normalized linear combinations of the columns of X.

e.g. principal components in 2D input data



Ridge Regression

Thus, with

$$\hat{\beta}_{\lambda}^{\text{ridge}} = V\Delta_{\lambda}U^{T}y = \sum_{j=1}^{p} v_j \frac{d_j}{d_j^2 + \lambda} u_j^T y$$

the prediction

$$\hat{y}_{\lambda} = UD\Delta_{\lambda}U^{T}y = \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y$$

is made onto the those components and shrinks the coefficients of the low variance components more than those with high variance.

Ridge Regression

Define, the effective degrees of freedom to be

$$df(\lambda) = tr(H_\lambda) = tr(D\Delta_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \leq p$$

with $df(\lambda) = p$ at $\lambda = 0$.

Ridge Regression

Define, the effective degrees of freedom to be

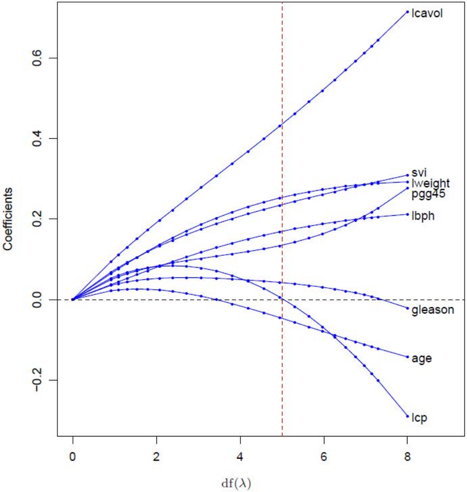
$$df(\lambda) = tr(H_\lambda) = tr(D\Delta_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \leq p$$

with $df(\lambda) = p$ at $\lambda = 0$.

Ridge Regression

Example (Prostate Cancer)

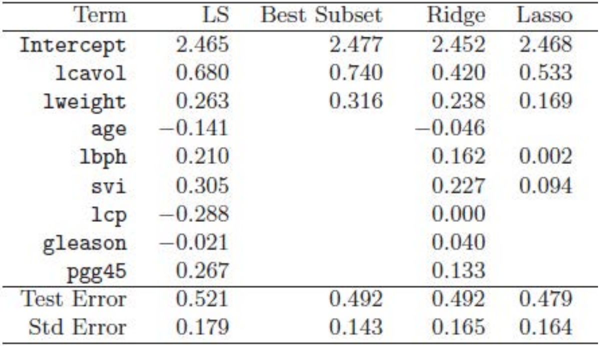
Estimated coefficients for different values of *df*(λ) with optimal *df* $= 5$ using CV.



Ridge Regression

Example (Prostate Cancer)

Estimated coefficients are



Term	LS	Best Subset	Ridge	Lasso
Intercept	2.465	2.477	2.452	2.468
lcavol	0.680	0.740	0.420	0.533
lweight	0.263	0.316	0.238	0.169
age	−0.141		−0.046	
lbph	0.210		0.162	0.002
svi	0.305		0.227	0.094
lcp	−0.288		0.000	
gleason	−0.021		0.040	
pgg45	0.267		0.133	
Test Error	0.521	0.492	0.492	0.479
Std Error	0.179	0.143	0.165	0.164

Least absolute shrinkage and selection operator (lasso) or basis pursuit

Lasso

To find $\hat{\beta}^{\text{lasso}}$ that

$$\text{minimize}_{\beta}\ RSS(\beta) = \sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2$$

$$\text{subject to } \sum_{j=1}^{p}|\beta_j| \leq t \text{ (size constraint)}$$

OR

$$\hat{\beta}^{\text{lasso}} = argmin_{\beta}\left[\sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 + \lambda \sum_{j=1}^{p}|\beta_j|\right]$$

with no closed form.

Lasso

To find $\hat{\beta}^{\text{lasso}}$ that

$$\text{minimize}_{\beta}\ RSS(\beta) = \sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t \text{ (size constraint)}$$

OR

$$\hat{\beta}^{\text{lasso}} = argmin_{\beta} \left[\sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

with no closed form.

Lasso

To find $\hat{\beta}^{\text{lasso}}$ that

$$\text{minimize}_{\beta}\ RSS(\beta) = \sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t \text{ (size constraint)}$$

OR

$$\hat{\beta}^{\text{lasso}} = argmin_{\beta} \left[\sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

with no closed form.

Lasso

Again, the solution is found using quadratic programming algorithms for each fixed λ or using the Least Angel Regression (LARS) (with computational costs comparable to the OLS).

- Standard errors are found computationally using bootstrap methods.
- What does happen when t increases beyond $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j^{ols}|$?

  Then $\hat{\beta}^{\text{lasso}} = \hat{\beta}^{ols}$.
- Thus, we use a normalized shrinkage factor $s = t/t_0$. It can be determined using CV.

Lasso

Again, the solution is found using quadratic programming algorithms for each fixed λ or using the Least Angel Regression (LARS) (with computational costs comparable to the OLS).

- Standard errors are found computationally using bootstrap methods.
- What does happen when t increases beyond $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j^{ols}|$?

  Then $\hat{\beta}^{\text{lasso}} = \hat{\beta}^{ols}$.
- Thus, we use a normalized shrinkage factor $s = t/t_0$. It can be determined using CV.

Lasso

Again, the solution is found using quadratic programming algorithms for each fixed λ or using the Least Angel Regression (LARS) (with computational costs comparable to the OLS).

- Standard errors are found computationally using bootstrap methods.
- What does happen when t increases beyond $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j^{ols}|$?

  Then $\hat{\beta}^{\text{lasso}} = \hat{\beta}^{ols}$.
- Thus, we use a normalized shrinkage factor $s = t/t_0$. It can be determined using CV.

Lasso

Again, the solution is found using quadratic programming algorithms for each fixed λ or using the Least Angel Regression (LARS) (with computational costs comparable to the OLS).

- Standard errors are found computationally using bootstrap methods.
- What does happen when t increases beyond $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j^{ols}|$?

  Then $\hat{\beta}^{\text{lasso}} = \hat{\beta}^{ols}$.
- Thus, we use a normalized shrinkage factor $s = t/t_0$. It can be determined using CV.

Lasso

Lasso tends to select more parameters, but it works very well when $p > N$. It outperforms subset selection and ridge regression in its predictive error.

Lasso

- If columns of X are orthonormal ($X^TX = I$), then

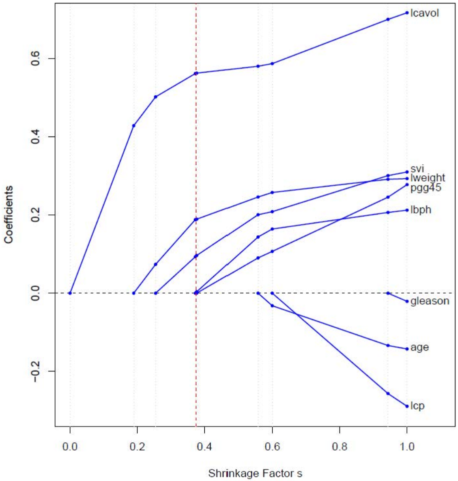
$$\hat{\beta}_{\lambda}^{\text{lasso}} = \text{sign}(\hat{\beta}^{ols})(|\hat{\beta}^{ols}| - \lambda/2)_{+}$$

It is called soft thresholding.

Lasso

Example (Prostate Cancer)

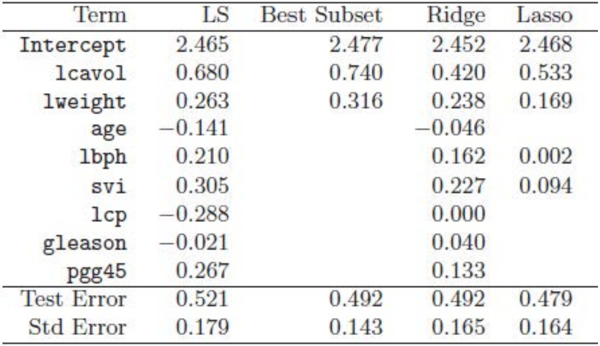
Estimated coefficients for different values of shrinkage factor *s* with optimal $s = .36$ using 10-fold CV.



Lasso

Example (Prostate Cancer)

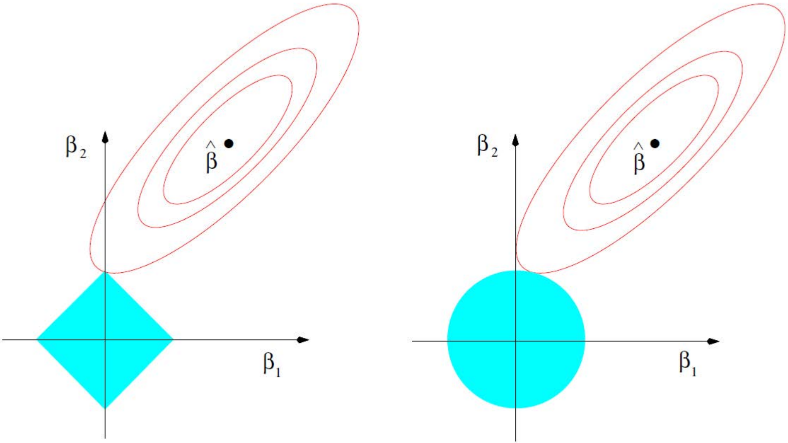
Estimated coefficients are



Term	LS	Best Subset	Ridge	Lasso
Intercept	2.465	2.477	2.452	2.468
lcavol	0.680	0.740	0.420	0.533
lweight	0.263	0.316	0.238	0.169
age	−0.141		−0.046	
lbph	0.210		0.162	0.002
svi	0.305		0.227	0.094
lcp	−0.288		0.000	
gleason	−0.021		0.040	
pgg45	0.267		0.133	
Test Error	0.521	0.492	0.492	0.479
Std Error	0.179	0.143	0.165	0.164

Lasso

Contours are for the error function around $\hat{\beta} = \hat{\beta}^{ols}$



$|\beta_1| + |\beta_2| \leq t$ vs $\beta_1^2 + \beta_2^2 \leq t$

Shrinkage+selection vs shrinkage

Elastic-net Method

Elastic-net Method

$$\hat{\beta}^{\text{elastic}} = argmin_{\beta} \left[\sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 + \lambda \sum_{j=1}^{p} \left(\alpha |\beta_j| + (1-\alpha)|\beta_j|^2 \right) \right]$$

Elastic-net selects like a lasso, shrinks like a ridge.

Example

For $\alpha = .8$, the elastic-net penalty $\sum_{j=1}^{2} \left(.8|\beta_j| + .2|\beta_j|^2\right) \leq t$

Elastic-net Method

$$\begin{aligned}\hat{\beta}^{\text{elastic}} =& argmin_{\beta} \left[\sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 \right. \\ & \left. + \lambda \sum_{j=1}^{p} \left(\alpha|\beta_j| + (1-\alpha)|\beta_j|^2 \right) \right]\end{aligned}$$

Elastic-net selects like a lasso, shrinks like a ridge.

Example

For $\alpha = .8$, the elastic-net penalty $\sum_{j=1}^{2} \left(.8|\beta_j| + .2|\beta_j|^2 \right) \leq t$

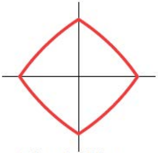
Elastic-net Method

$$\hat{\beta}^{\text{elastic}} = argmin_{\beta} \left[\sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 + \lambda \sum_{j=1}^{p} \left(\alpha |\beta_j| + (1-\alpha)|\beta_j|^2 \right) \right]$$

Elastic-net selects like a lasso, shrinks like a ridge.

Example

For $\alpha = .8$, the elastic-net penalty $\sum_{j=1}^{2} \left(.8|\beta_j| + .2|\beta_j|^2 \right) \leq t$



Elastic Net

Bridge Method

Bridge Method

To find $\hat{\beta}^{\text{bridge}}$ that

$$\text{minimize}_{\beta}\ RSS(\beta) = \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j|^q \leq t \text{ (size constraint)}$$

OR

$$\hat{\beta}^{\text{bridge}} = argmin_{\beta} \left[\sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right]$$

with no closed form for $0 < q \leq 1$.

Bridge Method

To find $\hat{\beta}^{\text{bridge}}$ that

$$\text{minimize}_{\beta}\ RSS(\beta) = \sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j|^q \leq t \text{ (size constraint)}$$

OR

$$\hat{\beta}^{\text{bridge}} = argmin_{\beta}\left[\sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q\right]$$

with no closed form for $0 < q \leq 1$.

Bridge Method

To find $\hat{\beta}^{\text{bridge}}$ that

$$\text{minimize}_{\beta}\ RSS(\beta) = \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j|^q \leq t \text{ (size constraint)}$$

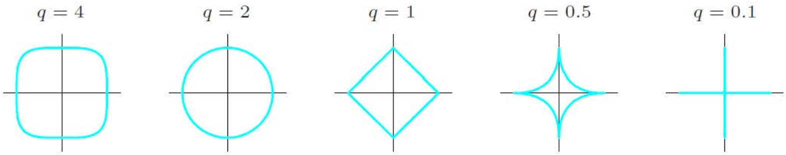
OR

$$\hat{\beta}^{\text{bridge}} = argmin_{\beta} \left[\sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p})]^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right]$$

with no closed form for $0 < q \leq 1$.

Bridge Method

$|\beta_1|^q + |\beta_2|^q \leq t$ for some q values.



Bridge Method

- When $q = 0$, the penalty term becomes $\lambda \sum_{j=1}^{p} I(\beta_j \neq 0)$
- If columns of X are orthonormal ($X^T X = I$), then

$$\hat{\beta}^{\text{bridge}} = \hat{\beta}^{ols} I(|\hat{\beta}^{ols}| \geq |\hat{\beta}^{ols}_{(M)}|)$$

  where $\hat{\beta}^{ols}_{(M)}$ is the M^{th} largest coefficient. It is called hard thresholding. It is a subset selection method.

Bridge Method

- When $q = 0$, the penalty term becomes $\lambda \sum_{j=1}^{p} I(\beta_j \neq 0)$
- If columns of X are orthonormal ($X^T X = I$), then

$$\hat{\beta}^{\text{bridge}} = \hat{\beta}^{ols} I(|\hat{\beta}^{ols}| \geq |\hat{\beta}^{ols}_{(M)}|)$$

where $\hat{\beta}^{ols}_{(M)}$ is the M^{th} largest coefficient. It is called hard thresholding. It is a subset selection method.

Bayesian Interpretation (bridge, lasso, and ridge)

Bayesian Interpretation (bridge, lasso, and ridge)

Define: The generalized Gaussian distribution $GG_q(\mu, \tau^2)$ with pdf

$$f_q(x) = \frac{1}{2\Gamma(1+\frac{1}{q})\sqrt{\frac{\Gamma(1/q)}{\Gamma(3/q)}}\,\tau} e^{-(\frac{\Gamma(3/q)}{\Gamma(1/q)})^{q/2}|\frac{x-\mu}{\tau}|^q}, \text{ for } x \in \mathbb{R}$$

with mean μ and variance τ^2.

- When $q = 1$, then $GG_1(\mu, \tau)$ is the Laplace distribution.

$$f_1(x) = \frac{1}{\sqrt{2}\tau} e^{-\sqrt{2}|\frac{x-\mu}{\tau}|}, \text{ for } x \in \mathbb{R}$$

- When $q = 2$, then $GG_2(\mu, \tau)$ is the normal distribution $N(\mu, \tau^2)$.

$$f_2(x) = \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{1}{2}(\frac{x-\mu}{\tau})^2}, \text{ for } x \in \mathbb{R}$$

Bayesian Interpretation (bridge, lasso, and ridge)

Define: The generalized Gaussian distribution $GG_q(\mu, \tau^2)$ with pdf

$$f_q(x) = \frac{1}{2\Gamma(1+\frac{1}{q})\sqrt{\frac{\Gamma(1/q)}{\Gamma(3/q)}}\,\tau} e^{-(\frac{\Gamma(3/q)}{\Gamma(1/q)})^{q/2}|\frac{x-\mu}{\tau}|^q}, \text{ for } x \in \mathbb{R}$$

with mean μ and variance τ^2.

- When $q = 1$, then $GG_1(\mu, \tau)$ is the Laplace distribution.

$$f_1(x) = \frac{1}{\sqrt{2}\tau} e^{-\sqrt{2}|\frac{x-\mu}{\tau}|}, \text{ for } x \in \mathbb{R}$$

- When $q = 2$, then $GG_2(\mu, \tau)$ is the normal distribution $N(\mu, \tau^2)$.

$$f_2(x) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{1}{2}(\frac{x-\mu}{\tau})^2}, \text{ for } x \in \mathbb{R}$$

Bayesian Interpretation (bridge, lasso, and ridge)

The generalized Gaussian distribution $GG_q(\mu, \tau^2)$ with pdf

$$f_q(x) = \frac{1}{2\Gamma(1+\frac{1}{q})\sqrt{\frac{\Gamma(1/q)}{\Gamma(3/q)}}\,\tau} e^{-(\frac{\Gamma(3/q)}{\Gamma(1/q)})^{q/2}|\frac{x-\mu}{\tau}|^q}, \text{ for } x \in \mathbb{R}$$

with mean μ and variance τ^2.

- When $q \to \infty$, then $GG_q(\mu, \tau)$ converges point-wise to the uniform distribution $\mathit{Uniform}(\mu - \sqrt{3}\tau, \mu + \sqrt{3}\tau)$.
- When $q \to 0^+$, then $GG_q(\mu, \tau)$ converges to a degenerate distribution at $x = \mu$.

Bayesian Interpretation (bridge, lasso, and ridge)

The generalized Gaussian distribution $GG_q(\mu, \tau^2)$ with pdf

$$f_q(x) = \frac{1}{2\Gamma(1+\frac{1}{q})\sqrt{\frac{\Gamma(1/q)}{\Gamma(3/q)}}\,\tau} e^{-(\frac{\Gamma(3/q)}{\Gamma(1/q)})^{q/2}|\frac{x-\mu}{\tau}|^q}, \text{ for } x \in \mathbb{R}$$

with mean μ and variance τ^2.

- When $q \to \infty$, then $GG_q(\mu, \tau)$ converges point-wise to the uniform distribution $Uniform(\mu - \sqrt{3}\tau, \mu + \sqrt{3}\tau)$.
- When $q \to 0^+$, then $GG_q(\mu, \tau)$ converges to a degenerate distribution at $x = \mu$.

Bayesian Analysis of Linear Regression

The linear regression model is

$$Y = X\beta + \epsilon,$$

where X is a $N \times (p+1)$, and $\epsilon \sim N(0, \sigma^2 I_N)$.
Then,

$$Y \sim N(X\beta, \sigma^2 I_N).$$

So the likelihood function is

$$L(\beta, \sigma | y) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{y_i - (\beta_0 + \sum_{j=1}^{p} x_{i,j}\beta_j)}{\sigma})^2}$$

Bayesian Analysis of Linear Regression

The linear regression model is

$$Y = X\beta + \epsilon,$$

where X is a $N \times (p+1)$, and $\epsilon \sim N(0, \sigma^2 I_N)$.
Then,

$$Y \sim N(X\beta, \sigma^2 I_N).$$

So the likelihood function is

$$L(\beta, \sigma | y) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{y_i - (\beta_0 + \sum_{j=1}^{p} x_{i,j}\beta_j)}{\sigma})^2}$$

Bayesian Analysis of Linear Regression

The linear regression model is

$$Y = X\beta + \epsilon,$$

where X is a $N \times (p+1)$, and $\epsilon \sim N(0, \sigma^2 I_N)$.
Then,

$$Y \sim N(X\beta, \sigma^2 I_N).$$

So the likelihood function is

$$L(\beta, \sigma | y) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y_i - (\beta_0 + \sum_{j=1}^{p} x_{i,j}\beta_j)}{\sigma}\right)^2}$$

Bayesian Analysis of Linear Regression

By Bayes' rule

$$\textit{posterior} \propto \textit{Likelihood} \cdot \textit{prior}$$

Choose $GG_q(0,\tau)$ to be a prior for each of the coefficients $\beta_1,\ldots,\beta_p$ (with the assumption that they are independent). Thus,

$$\begin{aligned}\textit{posterior} &\propto \prod_{i=1}^{N} e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\sum_{j=1}^{p} x_{i,j}\beta_j)}{\sigma}\right)^2} \cdot \prod_{j=1}^{p} e^{-\left(\frac{\Gamma(3/q)}{\Gamma(1/q)}\right)^{q/2}\left|\frac{\beta_j}{\tau}\right|^q} \\ &= e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-(\beta_0+\sum_{j=1}^{p} x_{i,j}\beta_j))^2} \cdot e^{-\left(\frac{\Gamma(3/q)}{\tau^2\Gamma(1/q)}\right)^{q/2}\sum_{j=1}^{p}|\beta_j|^q} \\ &= e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}(y_i-(\beta_0+\sum_{j=1}^{p} x_{i,j}\beta_j))^2+\lambda\sum_{j=1}^{p}|\beta_j|^q\right]}\end{aligned}$$

where $\lambda = 2\sigma^2\left(\frac{\Gamma(3/q)}{\tau^2\Gamma(1/q)}\right)^{q/2}$.

Bayesian Analysis of Linear Regression

By Bayes' rule

$$\textit{posterior} \propto \textit{Likelihood} \cdot \textit{prior}$$

Choose $GG_q(0,\tau)$ to be a prior for each of the coefficients $\beta_1, \ldots, \beta_p$ (with the assumption that they are independent). Thus,

$$\textit{posterior} \propto \prod_{i=1}^{N} e^{-\frac{1}{2}(\frac{y_i-(\beta_0+\sum_{j=1}^{p} x_{i,j}\beta_j)}{\sigma})^2} \cdot \prod_{j=1}^{p} e^{-(\frac{\Gamma(3/q)}{\Gamma(1/q)})^{q/2}|\frac{\beta_j}{\tau}|^q}$$

$$= e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-(\beta_0+\sum_{j=1}^{p} x_{i,j}\beta_j))^2} \cdot e^{-(\frac{\Gamma(3/q)}{\tau^2\Gamma(1/q)})^{q/2}\sum_{j=1}^{p}|\beta_j|^q}$$

$$= e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}(y_i-(\beta_0+\sum_{j=1}^{p} x_{i,j}\beta_j))^2+\lambda\sum_{j=1}^{p}|\beta_j|^q\right]}$$

where $\lambda = 2\sigma^2(\frac{\Gamma(3/q)}{\tau^2\Gamma(1/q)})^{q/2}$.

Bayesian Analysis of Linear Regression

By Bayes' rule

$$\textit{posterior} \propto \textit{Likelihood} \cdot \textit{prior}$$

Choose $GG_q(0, \tau)$ to be a prior for each of the coefficients $\beta_1, \ldots, \beta_p$ (with the assumption that they are independent). Thus,

$$\textit{posterior} \propto \prod_{i=1}^{N} e^{-\frac{1}{2}(\frac{y_i - (\beta_0 + \sum_{j=1}^{p} x_{i,j}\beta_j)}{\sigma})^2} \cdot \prod_{j=1}^{p} e^{-(\frac{\Gamma(3/q)}{\Gamma(1/q)})^{q/2} |\frac{\beta_j}{\tau}|^q}$$

$$= e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - (\beta_0 + \sum_{j=1}^{p} x_{i,j}\beta_j))^2} \cdot e^{-(\frac{\Gamma(3/q)}{\tau^2\Gamma(1/q)})^{q/2} \sum_{j=1}^{p} |\beta_j|^q}$$

$$= e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}(y_i - (\beta_0 + \sum_{j=1}^{p} x_{i,j}\beta_j))^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q\right]}$$

where $\lambda = 2\sigma^2(\frac{\Gamma(3/q)}{\tau^2\Gamma(1/q)})^{q/2}$.

Bayesian Analysis of Linear Regression

By Bayes' rule

$$\textit{posterior} \propto \textit{Likelihood} \cdot \textit{prior}$$

Choose $GG_q(0,\tau)$ to be a prior for each of the coefficients $\beta_1,\ldots,\beta_p$ (with the assumption that they are independent). Thus,

$$\begin{aligned}
\textit{posterior} &\propto \prod_{i=1}^{N} e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\sum_{j=1}^{p} x_{i,j}\beta_j)}{\sigma}\right)^2} \cdot \prod_{j=1}^{p} e^{-\left(\frac{\Gamma(3/q)}{\Gamma(1/q)}\right)^{q/2}\left|\frac{\beta_j}{\tau}\right|^q} \\
&= e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-(\beta_0+\sum_{j=1}^{p} x_{i,j}\beta_j))^2} \cdot e^{-\left(\frac{\Gamma(3/q)}{\tau^2\Gamma(1/q)}\right)^{q/2}\sum_{j=1}^{p}|\beta_j|^q} \\
&= e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}(y_i-(\beta_0+\sum_{j=1}^{p} x_{i,j}\beta_j))^2+\lambda\sum_{j=1}^{p}|\beta_j|^q\right]}
\end{aligned}$$

where $\lambda = 2\sigma^2\left(\frac{\Gamma(3/q)}{\tau^2\Gamma(1/q)}\right)^{q/2}$.

Bayesian Analysis of Linear Regression

By Bayes' rule

$$\textit{posterior} \propto \textit{Likelihood} \cdot \textit{prior}$$

Choose $GG_q(0, \tau)$ to be a prior for each of the coefficients $\beta_1, \ldots, \beta_p$ (with the assumption that they are independent). Thus,

$$\begin{aligned}
\textit{posterior} &\propto \prod_{i=1}^{N} e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\sum_{j=1}^{p} x_{i,j}\beta_j)}{\sigma}\right)^2} \cdot \prod_{j=1}^{p} e^{-\left(\frac{\Gamma(3/q)}{\Gamma(1/q)}\right)^{q/2}\left|\frac{\beta_j}{\tau}\right|^q} \\
&= e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-(\beta_0+\sum_{j=1}^{p} x_{i,j}\beta_j))^2} \cdot e^{-\left(\frac{\Gamma(3/q)}{\tau^2\Gamma(1/q)}\right)^{q/2}\sum_{j=1}^{p}|\beta_j|^q} \\
&= e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}(y_i-(\beta_0+\sum_{j=1}^{p} x_{i,j}\beta_j))^2+\lambda\sum_{j=1}^{p}|\beta_j|^q\right]}
\end{aligned}$$

where $\lambda = 2\sigma^2\left(\frac{\Gamma(3/q)}{\tau^2\Gamma(1/q)}\right)^{q/2}$.

Bayesian Analysis of Linear Regression

Thus, $-$*log posterior* is a linear function in

$$\left[\sum_{i=1}^{N}(y_i - (\beta_0 + \sum_{j=1}^{p} x_{i,j}\beta_j))^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q\right]$$

and so the posterior mode (the maximum point of the posterior distribution) is the minimum of the $-$*log posterior* and so it is the bridge estimate. If $q = 2$, then it is also the mean.

Principal Component Regression (PCR) - an unsupervised technique for dimension reduction

Principal Component Regression (PCR)

Starting with standardized data ...

PCR Idea: rotate the coordinates to reflect the most variability in the inputs in X, using $z_i := Xv_i$. Then perform regression on the new coordinate system. In that manner,

- We introduce the $N \times M$ matrix $W_M = XV$ with an $p \times M$ orthonormal matrix V (with $VV^T = I_p$) for some $M \in \{1, 2, \dots, p\}$
- That is, the i^{th} column of W_M is $z_i = Xv_i$.
- Then,

$$Y = X\beta + \epsilon$$

  gives a reduced regression

$$Y = W_M\theta + \epsilon$$

  where $\theta = V^T\beta$ and so $\beta = V\theta$.

Principal Component Regression (PCR)

Starting with standardized data ...

PCR Idea: rotate the coordinates to reflect the most variability in the inputs in X, using $z_i := Xv_i$. Then perform regression on the new coordinate system. In that manner,

- We introduce the $N \times M$ matrix $W_M = XV$ with an $p \times M$ orthonormal matrix V (with $VV^T = I_p$) for some $M \in \{1, 2, \dots, p\}$
- That is, the i^{th} column of W_M is $z_i = Xv_i$.
- Then,

$$Y = X\beta + \epsilon$$

gives a reduced regression

$$Y = W_M\theta + \epsilon$$

where $\theta = V^T\beta$ and so $\beta = V\theta$.

Principal Component Regression (PCR)

Starting with standardized data ...
PCR Idea: rotate the coordinates to reflect the most variability in the inputs in X, using $z_i := Xv_i$. Then perform regression on the new coordinate system. In that manner,

- We introduce the $N \times M$ matrix $W_M = XV$ with an $p \times M$ orthonormal matrix V (with $VV^T = I_p$) for some $M \in \{1, 2, \ldots, p\}$
- That is, the i^{th} column of W_M is $z_i = Xv_i$.
- Then,

  $$Y = X\beta + \epsilon$$

  gives a reduced regression

  $$Y = W_M\theta + \epsilon$$

  where $\theta = V^T\beta$ and so $\beta = V\theta$.

Principal Component Regression (PCR)

Starting with standardized data ...
PCR Idea: rotate the coordinates to reflect the most variability in the inputs in X, using $z_i := Xv_i$. Then perform regression on the new coordinate system. In that manner,

- We introduce the $N \times M$ matrix $W_M = XV$ with an $p \times M$ orthonormal matrix V (with $VV^T = I_p$) for some $M \in \{1, 2, \ldots, p\}$
- That is, the i^{th} column of W_M is $z_i = Xv_i$.
- Then,

$$Y = X\beta + \epsilon$$

gives a reduced regression

$$Y = W_M\theta + \epsilon$$

where $\theta = V^T\beta$ and so $\beta = V\theta$.

Principal Component Regression (PCR)

Starting with standardized data ...
PCR Idea: rotate the coordinates to reflect the most variability in the inputs in X, using $z_i := Xv_i$. Then perform regression on the new coordinate system. In that manner,

- We introduce the $N \times M$ matrix $W_M = XV$ with an $p \times M$ orthonormal matrix V (with $VV^T = I_p$) for some $M \in \{1, 2, \ldots, p\}$
- That is, the i^{th} column of W_M is $z_i = Xv_i$.
- Then,

  $$Y = X\beta + \epsilon$$

  gives a reduced regression

  $$Y = W_M\theta + \epsilon$$

  where $\theta = V^T\beta$ and so $\beta = V\theta$.

Principal Component Regression (PCR)

Thus,

- The PCR estimate is

$$\hat{\beta}^{\text{pcr}} = V\hat{\theta}.$$

- If $M = p$, then

$$\hat{\beta}^{\text{pcr}} = \hat{\beta}^{\text{ols}}.$$

Principal Component Regression (PCR)

Thus,

- The PCR estimate is

$$\hat{\beta}^{\text{pcr}} = V\hat{\theta}.$$

- If $M = p$, then

$$\hat{\beta}^{\text{pcr}} = \hat{\beta}^{\text{ols}}.$$

Principal Component Regression (PCR)

- PCR starts with principal component analysis (PCA), an unsupervised learning, from X.
- PCR shares the idea of principal components with ridge regression ...
- Ridge Regression shrinks in the principal component directions of the small variance, whereas Principal Component Regression omit those directions (a number of $p - M$ smallest eigenvalues).
- Yet, PCR, like ridge regresion, is not a subset selection method, since the M components z_i's are linear combinations of the p inputs as in $z_i = Xv_i$.

Principal Component Regression (PCR)

- PCR starts with principal component analysis (PCA), an unsupervised learning, from X.
- PCR shares the idea of principal components with ridge regression ...
- Ridge Regression shrinks in the principal component directions of the small variance, whereas Principal Component Regression omit those directions (a number of $p - M$ smallest eigenvalues).
- Yet, PCR, like ridge regresion, is not a subset selection method, since the M components z_i's are linear combinations of the p inputs as in $z_i = Xv_i$.

Principal Component Regression (PCR)

- PCR starts with principal component analysis (PCA), an unsupervised learning, from X.
- PCR shares the idea of principal components with ridge regression ...
- Ridge Regression shrinks in the principal component directions of the small variance, whereas Principal Component Regression omit those directions (a number of $p - M$ smallest eigenvalues).
- Yet, PCR, like ridge regresion, is not a subset selection method, since the M components z_i's are linear combinations of the p inputs as in $z_i = Xv_i$.

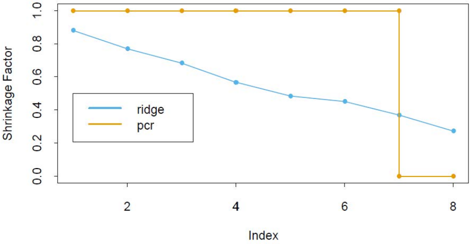
Principal Component Regression (PCR)

- PCR starts with principal component analysis (PCA), an unsupervised learning, from X.
- PCR shares the idea of principal components with ridge regression ...
- Ridge Regression shrinks in the principal component directions of the small variance, whereas Principal Component Regression omit those directions (a number of $p - M$ smallest eigenvalues).
- Yet, PCR, like ridge regresion, is not a subset selection method, since the M components z_i's are linear combinations of the p inputs as in $z_i = Xv_i$.

Principal Component Regression (PCR)

Example (Prostate Cancer)

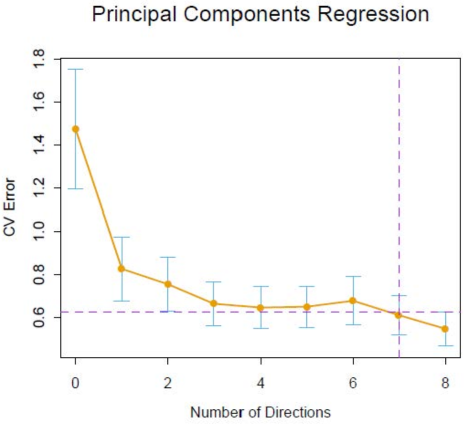
Shrinkage factor $d^2/(d^2+\lambda)$ versus the index of the component



PCR

Example (Prostate Cancer)

CV error shows optimal less complex at $M = 7$ using 10-fold CV.



Principal Component Analysis (PCA)

1. For population data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Var(X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T Var(X)\omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega,Xv_1)=0}}{argmax}\, Var(X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega,Xv_i)=0; i=1,\ldots,M-1}}{argmax}\, Var(X\omega)$

How to determine M? By CV.

Principal Component Analysis (PCA)

1. For population data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Var(X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T Var(X)\omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_1)=0}}{argmax}\, Var(X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_i)=0; i=1,\dots,M-1}}{argmax}\, Var(X\omega)$

How to determine M? By CV.

Principal Component Analysis (PCA)

1. For population data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Var(X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T Var(X)\omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1, \\ Cov(X\omega, Xv_1)=0}}{argmax}\, Var(X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1, \\ Cov(X\omega, Xv_i)=0; i=1,\ldots,M-1}}{argmax}\, Var(X\omega)$

How to determine M? By CV.

Principal Component Analysis (PCA)

1. For population data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Var(X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T Var(X)\omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_1)=0}}{argmax}\, Var(X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_i)=0; i=1,\ldots,M-1}}{argmax}\, Var(X\omega)$

How to determine M? By CV.

Principal Component Analysis (PCA)

1. For population data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Var(X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T Var(X)\omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega,Xv_1)=0}}{argmax}\, Var(X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega,Xv_i)=0;i=1,\ldots,M-1}}{argmax}\, Var(X\omega)$

How to determine M? By CV.

Principal Component Analysis (PCA)

1. For population data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Var(X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T Var(X)\omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1, \\ Cov(X\omega, Xv_1)=0}}{argmax}\, Var(X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1, \\ Cov(X\omega, Xv_i)=0; i=1,\ldots,M-1}}{argmax}\, Var(X\omega)$

How to determine M? By CV.

Principal Component Analysis (PCA)

1. For population data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\ Var(X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\ \omega^T Var(X)\omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1, \\ Cov(X\omega, Xv_1)=0}}{argmax}\ Var(X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1, \\ Cov(X\omega, Xv_i)=0; i=1,\ldots,M-1}}{argmax}\ Var(X\omega)$

How to determine M? By CV.

Principal Component Analysis (PCA)

1. For population data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Var(X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T Var(X)\omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega,Xv_1)=0}}{argmax}\, Var(X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega,Xv_i)=0;i=1,\ldots,M-1}}{argmax}\, Var(X\omega)$

How to determine M? By CV.

Principal Component Analysis (PCA)

2. For sample data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T X^T X \omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ \omega^T X^T X v_1=0}}{argmax}\, \omega^T X^T X \omega$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ \omega^T X^T X v_i=0; i=1,\dots,M-1}}{argmax}\, \omega^T X^T X \omega$

But, no guarantee that the directions with the largest variance/explanation of the predictor, will also be the best for prediction. So ...

Principal Component Analysis (PCA)

2. For sample data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T X^T X \omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ \omega^T X^T X v_1=0}}{argmax}\, \omega^T X^T X \omega$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ \omega^T X^T X v_i=0; i=1,\ldots,M-1}}{argmax}\, \omega^T X^T X \omega$

But, no guarantee that the directions with the largest variance/explanation of the predictor, will also be the best for prediction. So ...

Principal Component Analysis (PCA)

2. For sample data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T X^T X \omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ \omega^T X^T X v_1=0}}{argmax}\, \omega^T X^T X \omega$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ \omega^T X^T X v_i=0; i=1,\ldots,M-1}}{argmax}\, \omega^T X^T X \omega$

But, no guarantee that the directions with the largest variance/explanation of the predictor, will also be the best for prediction. So ...

Principal Component Analysis (PCA)

2. For sample data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T X^T X \omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ \omega^T X^T X v_1=0}}{argmax}\, \omega^T X^T X \omega$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ \omega^T X^T X v_i=0; i=1,\ldots,M-1}}{argmax}\, \omega^T X^T X \omega$

But, no guarantee that the directions with the largest variance/explanation of the predictor, will also be the best for prediction. So ...

Principal Component Analysis (PCA)

2. For sample data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T X^T X \omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ \omega^T X^T X v_1=0}}{argmax}\, \omega^T X^T X \omega$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ \omega^T X^T X v_i=0; i=1,\dots,M-1}}{argmax}\, \omega^T X^T X \omega$

But, no guarantee that the directions with the largest variance/explanation of the predictor, will also be the best for prediction. So ...

Principal Component Analysis (PCA)

2. For sample data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T X^T X \omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ \omega^T X^T X v_1=0}}{argmax}\, \omega^T X^T X \omega$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ \omega^T X^T X v_i=0; i=1,\ldots,M-1}}{argmax}\, \omega^T X^T X \omega$

But, no guarantee that the directions with the largest variance/explanation of the predictor, will also be the best for prediction. So ...

Principal Component Analysis (PCA)

2. For sample data X:

Step 1: Find $v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, \omega^T X^T X \omega$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1, \\ \omega^T X^T X v_1=0}}{argmax}\, \omega^T X^T X \omega$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1, \\ \omega^T X^T X v_i=0; i=1,\ldots,M-1}}{argmax}\, \omega^T X^T X \omega$

But, no guarantee that the directions with the largest variance/explanation of the predictor, will also be the best for prediction. So ...

Partial Least Squares (PLS) - a supervised technique for dimension reduction

Partial Least Squares (PLS)

Starting with standardized data but this time including Y ...

PLS Idea: rotate the coordinates to reflect the most correlation between the output Y and the inputs in X, using PLS directions $z_i := Xv_i$. Then perform regression on the new coordinate system. In that manner,

- We introduce the $N \times M$ matrix $W_M = XV$ with an $p \times M$ orthonormal matrix V (with $VV^T = I_p$) for some $M \in \{1, 2, \ldots, p\}$
- That is, the i^{th} column of W_M is $z_i = Xv_i$.
- Then,

$$Y = X\beta + \epsilon$$

gives a reduced regression

$$Y = W_M\theta + \epsilon$$

where $\theta = V^T\beta$ and so $\beta = V\theta$.

Partial Least Squares (PLS)

Starting with standardized data but this time including Y ...
PLS Idea: rotate the coordinates to reflect the most correlation between the output Y and the inputs in X, using PLS directions $z_i := Xv_i$. Then perform regression on the new coordinate system. In that manner,

- We introduce the $N \times M$ matrix $W_M = XV$ with an $p \times M$ orthonormal matrix V (with $VV^T = I_p$) for some $M \in \{1, 2, \ldots, p\}$
- That is, the i^{th} column of W_M is $z_i = Xv_i$.
- Then,

$$Y = X\beta + \epsilon$$

gives a reduced regression

$$Y = W_M\theta + \epsilon$$

where $\theta = V^T\beta$ and so $\beta = V\theta$.

Partial Least Squares (PLS)

Starting with standardized data but this time including Y ...
PLS Idea: rotate the coordinates to reflect the most correlation between the output Y and the inputs in X, using PLS directions $z_i := Xv_i$. Then perform regression on the new coordinate system. In that manner,

- We introduce the $N \times M$ matrix $W_M = XV$ with an $p \times M$ orthonormal matrix V (with $VV^T = I_p$) for some $M \in \{1, 2, \ldots, p\}$
- That is, the i^{th} column of W_M is $z_i = Xv_i$.
- Then,

$$Y = X\beta + \epsilon$$

  gives a reduced regression

$$Y = W_M\theta + \epsilon$$

  where $\theta = V^T\beta$ and so $\beta = V\theta$.

Partial Least Squares (PLS)

Starting with standardized data but this time including Y ...
PLS Idea: rotate the coordinates to reflect the most correlation between the output Y and the inputs in X, using PLS directions $z_i := Xv_i$. Then perform regression on the new coordinate system. In that manner,

- We introduce the $N \times M$ matrix $W_M = XV$ with an $p \times M$ orthonormal matrix V (with $VV^T = I_p$) for some $M \in \{1, 2, \dots, p\}$
- That is, the i^{th} column of W_M is $z_i = Xv_i$.
- Then,

$$Y = X\beta + \epsilon$$

  gives a reduced regression

$$Y = W_M\theta + \epsilon$$

  where $\theta = V^T\beta$ and so $\beta = V\theta$.

Partial Least Squares (PLS)

Starting with standardized data but this time including Y ...
PLS Idea: rotate the coordinates to reflect the most correlation between the output Y and the inputs in X, using PLS directions $z_i := Xv_i$. Then perform regression on the new coordinate system. In that manner,

- We introduce the $N \times M$ matrix $W_M = XV$ with an $p \times M$ orthonormal matrix V (with $VV^T = I_p$) for some $M \in \{1, 2, \dots, p\}$
- That is, the i^{th} column of W_M is $z_i = Xv_i$.
- Then,

  $$Y = X\beta + \epsilon$$

  gives a reduced regression

  $$Y = W_M\theta + \epsilon$$

  where $\theta = V^T\beta$ and so $\beta = V\theta$.

Partial Least Squares (PLS)

Thus,

- The PLS estimate is

$$\hat{\beta}^{\text{pls}} = V\hat{\theta}.$$

- If $M = p$, then

$$\hat{\beta}^{\text{pls}} = \hat{\beta}^{\text{ols}}.$$

Partial Least Squares (PLS)

Thus,

- The PLS estimate is

$$\hat{\beta}^{\text{pls}} = V\hat{\theta}.$$

- If $M = p$, then

$$\hat{\beta}^{\text{pls}} = \hat{\beta}^{\text{ols}}.$$

PLS Directions $z_i = Xv_i$

For population data X:

Step 1: Find

$$v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Cov^2(Y, X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, Corr^2(Y, X\omega)\, Var(X\omega)$$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_1)=0}}{argmax} Cov^2(Y, X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_i)=0; i=1,\ldots,M-1}}{argmax} Cov^2(Y, X\omega)$

How to determine M? By CV.

PLS Directions $z_i = Xv_i$

For population data X:

Step 1: Find

$$v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Cov^2(Y, X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, Corr^2(Y, X\omega)\, Var(X\omega)$$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_1)=0}}{argmax} Cov^2(Y, X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_i)=0; i=1,\ldots,M-1}}{argmax} Cov^2(Y, X\omega)$

How to determine M? By CV.

PLS Directions $z_i = Xv_i$

For population data X:

Step 1: Find

$$v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Cov^2(Y, X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, Corr^2(Y, X\omega)\, Var(X\omega)$$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_1)=0}}{argmax} Cov^2(Y, X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_i)=0; i=1,\ldots,M-1}}{argmax} Cov^2(Y, X\omega)$

How to determine M? By CV.

PLS Directions $z_i = Xv_i$

For population data X:

Step 1: Find

$$v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Cov^2(Y, X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, Corr^2(Y, X\omega)\, Var(X\omega)$$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega,Xv_1)=0}}{argmax} \quad Cov^2(Y, X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega,Xv_i)=0;i=1,\ldots,M-1}}{argmax} \quad Cov^2(Y, X\omega)$

How to determine M? By CV.

PLS Directions $z_i = Xv_i$

For population data X:

Step 1: Find

$$v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Cov^2(Y, X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, Corr^2(Y, X\omega)\, Var(X\omega)$$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_1)=0}}{argmax} Cov^2(Y, X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_i)=0; i=1,\ldots,M-1}}{argmax} Cov^2(Y, X\omega)$

How to determine M? By CV.

PLS Directions $z_i = Xv_i$

For population data X:

Step 1: Find

$$v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Cov^2(Y, X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, Corr^2(Y, X\omega)\, Var(X\omega)$$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_1)=0}}{argmax} Cov^2(Y, X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_i)=0; i=1,\ldots,M-1}}{argmax} Cov^2(Y, X\omega)$

How to determine M? By CV.

PLS Directions $z_i = Xv_i$

For population data X:

Step 1: Find

$$v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Cov^2(Y, X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, Corr^2(Y, X\omega)\, Var(X\omega)$$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_1)=0}}{argmax} Cov^2(Y, X\omega)$

$\vdots$

Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_i)=0; i=1,\ldots,M-1}}{argmax} Cov^2(Y, X\omega)$

How to determine M? By CV.

PLS Directions $z_i = Xv_i$

For population data X:

Step 1: Find

$$v_1 = \underset{\omega:\omega^T\omega=1}{argmax}\, Cov^2(Y, X\omega) = \underset{\omega:\omega^T\omega=1}{argmax}\, Corr^2(Y, X\omega)\, Var(X\omega)$$

Step 2: Find $v_2 = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_1)=0}}{argmax} Cov^2(Y, X\omega)$

$\vdots$

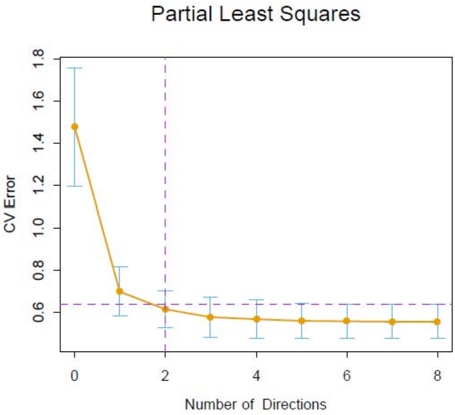
Step M: Find $v_M = \underset{\substack{\omega:\omega^T\omega=1,\\ Cov(X\omega, Xv_i)=0; i=1,\ldots,M-1}}{argmax} Cov^2(Y, X\omega)$

How to determine M? By CV.

PLS

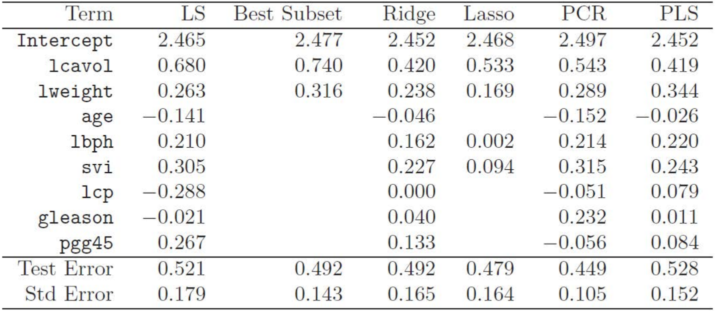
Example (Prostate Cancer)

CV error shows optimal less complex at $M = 2$ using 10-fold CV.



PLS

Example (Prostate Cancer)



Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	−0.141		−0.046		−0.152	−0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	−0.288		0.000		−0.051	0.079
gleason	−0.021		0.040		0.232	0.011
pgg45	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

K-means Regression

K-means Regression

It is a non-parametric method.

K-means Idea: the simplest is the K-nearest neighbor regression (K-NN). Thus, K-means regression is a local method. In that manner,

- The predicted response at x_* is

$$\hat{f}(x_*) = Average(y_i | x_i \in N_k(x_*)) = \frac{1}{k} \sum_{x_i \in N_k(x_*)} y_i$$

where $N_k(x_*)$is a neighborhood of x_* of size k.

K-means Regression

It is a non-parametric method.
K-means Idea: the simplest is the K-nearest neighbor regression (K-NN). Thus, K-means regression is a local method. In that manner,

- The predicted response at x_* is

$$\hat{f}(x_*) = Average(y_i | x_i \in N_k(x_*)) = \frac{1}{k} \sum_{x_i \in N_k(x_*)} y_i$$

where $N_k(x_*)$is a neighborhood of x_* of size k.

K-means Regression

It is a non-parametric method.
K-means Idea: the simplest is the K-nearest neighbor regression (K-NN). Thus, K-means regression is a local method. In that manner,

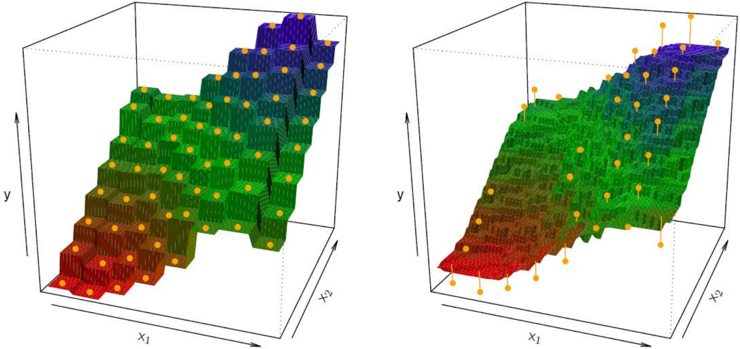
- The predicted response at x_* is

$$\hat{f}(x_*) = Average(y_i|x_i \in N_k(x_*)) = \frac{1}{k} \sum_{x_i \in N_k(x_*)} y_i$$

where $N_k(x_*)$is a neighborhood of x_* of size k.

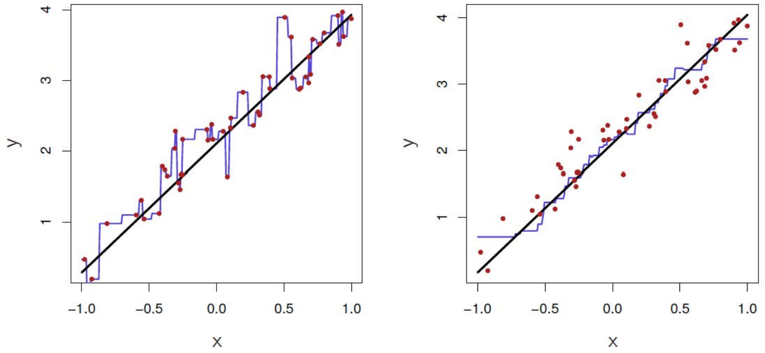
K-means Regression

$K = 1$ versus $K = 9$



K-means Regression

Parametric functions that really represent the data outperform non-parametric methods. Curse of dimensionality vs overfitting.



End of Set 3