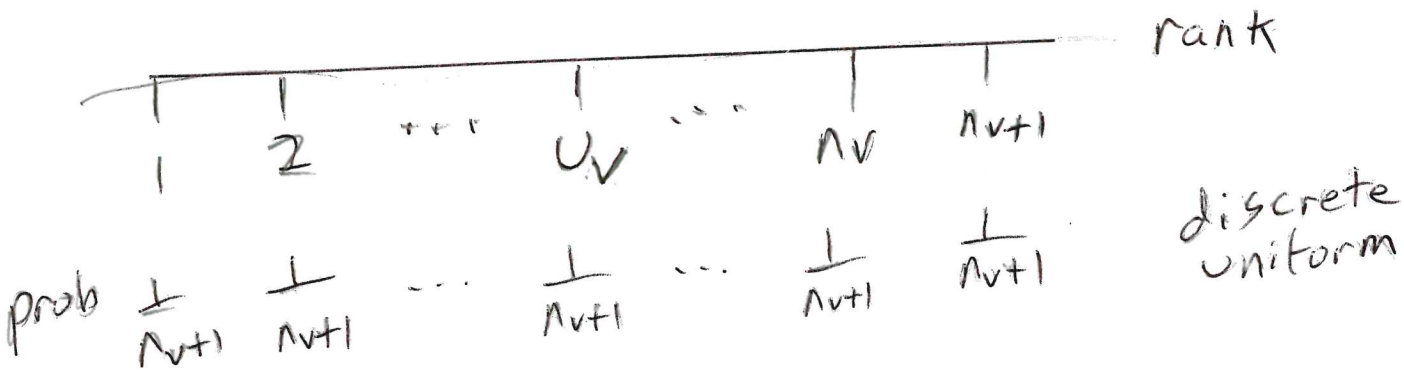
$$P(\text{rank } D^2_{\underline{x}\underline{\epsilon}} \le U_V) = \frac{U_V}{1+n_V}$$



Back to regression

47) Suppose data splitting is used resulting in model $I_H = I$, $\underset{\sim}{\beta}_I$ a $x1$, $n_V \ge Ja$, $J = 10, 20$ or higher. (sparse I)

If the model is good (check with response plot, etc)

$$\sqrt{n_V}\left(\hat{\underset{\sim}{\beta}}_I - \underset{\sim}{\beta}_I\right) \xrightarrow{D} N_a(\underline{0}, V_I)$$

and do not need $S \subseteq I$ or

$$\underset{\sim}{\beta}_{Io} = \underset{\sim}{\beta} = \underset{\sim}{\beta}_F.$$

With iid cases, $\underset{\sim}{\beta}_{I,OLS} = \underset{\underline{x}_I}{\Sigma^{-1}} \underset{\underline{x}_I Y}{\Sigma}$

$\underset{\sim}{M}_{I,OLS} = \underset{\underline{x}_I Y}{\Sigma}$, and

$$\underset{\sim}{B}_{I,OPLS} = \lambda_I \, \underset{\sim}{\Sigma}_{x_I Y} \quad \text{with}$$

$$\lambda_I = \frac{\underset{\sim}{\Sigma}_{x_I Y}^T \, \underset{\sim}{\Sigma}_{x_I Y}}{\underset{\sim}{\Sigma}_{x_I Y}^T \, \underset{\sim}{\Sigma}_{x_I} \, \underset{\sim}{\Sigma}_{x_I Y}}.$$

48] In low dimensions, want to estimate $\underset{\sim}{B} = \underset{\sim}{B}_F$ that uses all p predictors, and find predictors that are not needed. In high dimensions, can't estimate $\underset{\sim}{B}_F$ well, but with iid cases, want to greatly outperform the null model that uses iid Y with no predictors.

nontrivial predictors don't count the constant

49] we will say a fitted or POP model is sparse if a of the predictors are active ($\underset{a \times 1}{\underset{\sim}{B}_I}$ / $\underset{a \times 1}{\hat{\underset{\sim}{B}}_I}$ so a of the predictors have nonzero β_i or $\hat{\beta}_i$), and $n \geq Ja$ with $J \geq 10$.

otherwise the model is nonsparse.

A high dimensional pop reg full model is dense or abundant if $n < 5p$ and nearly all p predictors are active with nonzero β_i.

So] what $\hat{\beta}_F$, $\hat{\beta}_I$, or $\hat{\beta}_{I,0}$ estimate

	low dimensional	data splitting variable selection HD but sparse I	HD error
general	$\underset{\sim}{\beta}_F(x,y) = \underset{\sim}{\beta}_{I,0}(x_I, y)$	$\underset{\sim}{\beta}_I(x_I, y)$	$\underset{\sim}{\beta}_F = \underset{\sim}{\beta}(x,y) = \underset{\sim}{\beta}_{I,0}(x_I, y)$
lassoLS	$\underset{\sim}{\beta}_F(x,y) = \underset{\sim}{\beta}_{I,0}(x_I, y)$	$\underset{\sim}{\beta}_I(x_I, y)$	$\beta_F = \beta(x,y) = \beta_{I,0}(x_I, y)$
OLS	$\underset{\sim}{\beta}_{I,0}(x_I, y) = \beta_{OLS}(x,y)$	$\underset{\sim}{\beta}_I(x_I, y)$ or $\underset{\sim}{x}_I^{-1} \underset{\sim}{x}_I y$ (iid cases)	$\underset{\sim}{\beta}_E = \underset{\sim}{\beta}_{OLS} = \beta_F$ ← estimator
OPLS	$\underset{\sim}{\beta}_{OPLS} = \lambda \underset{\sim}{x} y$	$\beta_{I\,OPLS} = \lambda_I \underset{\sim}{x}_I y$	$\beta_F = \beta_{OPLS} = \beta_{OLS}$
MMLE	$\underset{\sim}{\beta}_{MMLE} = \underset{\sim}{x} y$ standardized data	$\underset{\sim}{\beta}_{I,MMLE} = \underset{\sim}{x}_I y$	$\beta_F = \beta_{MMLE} = \beta_{OLS}$

In low dimensions, $\hat{\underset{\sim}{\beta}}_F \xrightarrow{P} \underset{\sim}{\beta}_F$ and $\hat{\underset{\sim}{\beta}}_{I,0} \xrightarrow{P} \underset{\sim}{\beta}_F$ often hold (OLS, GLMs, PH).

A common error is to assume $\hat{\underset{\sim}{\beta}}_{I,0} \xrightarrow{P} \underset{\sim}{\beta}_F$ in high dimensions.

Back to ch 3

1] Let $Y = \underset{\sim}{x}^T \underset{\sim}{\beta} + e$ (MLR1).

OLS minimizes the residual sum of squares $Q_{OLS}(\underset{\sim}{\beta}) = \sum (Y_i - x_i^T \underset{\sim}{\beta})^2$

$= (\underset{\sim}{Y} - \underset{\sim}{x}\underset{\sim}{\beta})^T (\underset{\sim}{Y} - \underset{\sim}{x}\underset{\sim}{\beta}) = RSS(\underset{\sim}{\beta})$.

2] One ridge regression estimator $\hat{\underset{\sim}{\beta}}_R$

minimizes the criterion

$$Q_R(\underset{\sim}{\beta}) = \frac{1}{a} \underbrace{(\underset{\sim}{Y} - \underset{\sim}{x}\underset{\sim}{\beta})^T (\underset{\sim}{Y} - \underset{\sim}{x}\underset{\sim}{\beta})}_{RSS(\underset{\sim}{\beta})} + \frac{\lambda_n}{a} \underbrace{\underset{\sim}{\beta}^T \underset{\sim}{\beta}}_{\sum_{i=1}^{p} \beta_i^2 = \|\underset{\sim}{\beta}\|_2^2}$$

with $a = 1, 2, n, 2n$ common.

If $\lambda_n = 0$, $\hat{\underset{\sim}{\beta}}_R = \hat{\underset{\sim}{\beta}}_{OLS}$.

It can be shown that HW6

$$\hat{\underset{\sim}{\beta}}_R = (\underset{\sim}{x}^T \underset{\sim}{x} + \lambda_n I_p)^{-1} \underset{\sim}{x}^T \underset{\sim}{Y} =$$

$$(\underset{\sim}{x}^T \underset{\sim}{x} + \lambda_n I_p)^{-1} \underset{\sim}{x}^T \underset{\sim}{x} \underbrace{(\underset{\sim}{x}^T \underset{\sim}{x})^{-1} \underset{\sim}{x}^T \underset{\sim}{Y}}_{\hat{\underset{\sim}{\beta}}_{OLS}}$$

if $(X^TX)^{-1}$ exists.

3] X^TX is symmetric and square.

Hence $X^TX \geq 0$ (positive semidefinite)

with eigenvalues $\psi_1 \geq \psi_2 \geq \cdots \geq \psi_p \geq 0$.

If $\psi_p = 0$, then $(X^TX)^{-1}$ does not exist.

Let $(\psi, \underset{\sim}{g})$ be an eigenvalue eigenvector pair of X^TX. Then $(X^TX + \lambda_{in} I_p)\underset{\sim}{g}$

$= X^TX\underset{\sim}{g} + \lambda_{in}\underset{\sim}{g} = \psi\underset{\sim}{g} + \lambda_{in}\underset{\sim}{g} = (\psi + \lambda_{in})\underset{\sim}{g}$.

Hence $(\underbrace{\psi + \lambda_{in}}_{\geq 0 \text{ if } \lambda_{in} > 0}, \underset{\sim}{g})$ is an eigenvalue eigenvector

pair of $X^TX + \lambda_{in} I_p \underbrace{> 0 \text{ if } \lambda_{in} > 0}_{\text{positive definite}}$.

Hence $(X^TX + \lambda_{in} I_p)^{-1}$ exists if $\lambda_{in} > 0$

$\forall \lambda_{in} > 0$, even if X^TX is singular or ill conditioned.

4] The score equations for $Q_R(\beta)$ give

HW6 $\nabla Q_R(\beta) \overset{set}{=} 0$

$$-X^T(Y - X\hat{\beta}_R) + \lambda \hat{\beta}_R = 0.$$

like normal eq's for OLS

claim: $\hat{\beta}_R = X^T\hat{\gamma}$ for some $\hat{\gamma}$.

plug in ⊥

proof: $-X^T(Y - XX^T\hat{\gamma}) + \lambda X^T\hat{\gamma} = 0$

or $-X^TY + X^TXX^T\hat{\gamma} + \lambda X^T\hat{\gamma} = 0$ (X^T: $p \times n$)

or $X^TY = X^T(XX^T + \lambda I_n)\hat{\gamma}$ and

$\hat{\gamma} = (XX^T + \lambda I_n)^{-1}Y$ works. □

Thus $\hat{\beta}_R = \underset{p\times n}{X^T}\underbrace{(XX^T + \lambda I_n)^{-1}}_{n\times n \text{ matrix}}Y$ ✓

and $\hat{\beta}_R = \underbrace{(X^TX + \lambda I_p)^{-1}}_{p\times p \text{ matrix}}X^TY$.

If $n >> p$, use the 2nd formula.

If $p >> n$ use the 1st formula.

5] In H.D statistics, sometimes the method

Can be computed using either a $p \times p$ matrix B or an $n \times n$ matrix A as in 4). PCA and PCR give more examples.

6] By 2] $\hat{\beta}_R = (X^T X + \lambda_{1n} I_p)^{-1} X^T X \hat{\beta}_{OLS}$

$= A_n \hat{\beta}_{OLS} = B_n \hat{\beta}_{OLS} =$

$\hat{\beta}_{OLS} - \frac{\lambda_{1n}}{n} n(X^T X + \lambda_{1n} I_p)^{-1} \hat{\beta}_{OLS}$

where $B_n = I_p - \lambda_{1n}(X^T X + \lambda_{1n} I_p)^{-1}$.

See HW 1 show $A_n - B_n = 0$.

7] Suppose $\frac{X^T X}{n} \xrightarrow{P} V^{-1}$ so

$\sqrt{n}(\hat{\beta}_{OLS} - \beta) \xrightarrow{D} N_p(0, \sigma^2 V)$.

If $\frac{\lambda_{1n}}{n} \to 0$, then $\frac{X^T X + \lambda_{1n} I_p}{n} \xrightarrow{P} V^{-1}$ and

$n(X^T X + \lambda_{1n} I_p)^{-1} \xrightarrow{P} V$.

Then $A_n = A_{n\lambda} = \left(\frac{X^TX + \lambda_{1n} I_p}{n}\right)^{-1} \frac{X^TX}{n} \xrightarrow{P}$

$V V^{-1} = I_p$.

8] RR CLT: Assume 7] holds with p fixed.

for model $\underline{Y} = X\underline{\beta} + \underline{e}$.

a) If $\frac{\hat{\lambda}_{1n}}{\sqrt{n}} \xrightarrow{P} 0$, then $\sqrt{n}\left(\hat{\underline{\beta}}_R - \underline{\beta}\right) \xrightarrow{D} N_p(\underline{0}, \sigma^2 V)$.

$\hat{\underline{\beta}}_R = \underline{\beta}_{OLS}$

b) If $\frac{\hat{\lambda}_{1n}}{\sqrt{n}} \xrightarrow{P} \tau \geq 0$, then

$$\sqrt{n}\left(\hat{\underline{\beta}}_R - \underline{\beta}\right) \xrightarrow{D} N_p\left(-\tau V \underline{\beta}, \sigma^2 V\right)$$

Proof] If $\frac{\hat{\lambda}_{1n}}{\sqrt{n}} \xrightarrow{P} \tau \geq 0$, then by 6],

$$\hat{\underline{\beta}}_R = \left[I_p - \hat{\lambda}_{1n}\left(X^TX + \hat{\lambda}_{1n} I_p\right)^{-1}\right] \hat{\underline{\beta}}_{OLS}.$$

Hence $\sqrt{n}\left(\hat{\underline{\beta}}_R - \underline{\beta}\right) = \sqrt{n}\left[\hat{\underline{\beta}}_R - \underbrace{\hat{\underline{\beta}}_{OLS}}_{\underline{0}} + \underbrace{\hat{\underline{\beta}}_{OLS}}_{\underline{\beta}_{OLS}} - \underline{\beta}\right]$

$$= \sqrt{n}\left(\hat{\underline{\beta}}_{OLS} - \underline{\beta}\right) - \sqrt{n}\,\frac{\hat{\lambda}_{1n}}{n}\, n\left(X^TX + \hat{\lambda}_{1n} I_p\right)^{-1} \hat{\underline{\beta}}_{OLS}$$

$$\xrightarrow{D} N_p\left(\underline{0}, \sigma^2 V\right) - \tau V \underline{\beta} \sim N_p\left(-\tau V \underline{\beta}, \sigma^2 V\right).$$

□

the estimators could be poor for β

9] Suppose we have J estimators (of β) such as $\hat{\beta}_R(\lambda_1), \ldots, \hat{\beta}_R(\lambda_J)$ where ridge regression is computed on a grid of J λ values

$0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_J$ where $\lambda_i = \lambda_1 n i$.

Then K-fold CV (cross validation) is often used to select $\hat{\lambda}_{in} = \lambda_i^*$.

Randomly divide the (training) data into K groups or folds of approx equal size $n_j \simeq \frac{n}{K}$ for $j = 1, \ldots, K$. Leave out the 1st fold, fit the method to the K-1 remaining folds, then compute some criterion for the 1st fold. Repeat for folds $2, 3, \ldots, K$.

like data splitting but repetition causes dependence

10] For MLR, compute $\hat{y}_{i(j)}$ for each y_i in fold j left out.

Then $MSE_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (Y_i - \hat{Y}_{i(j)})^2$

and the overall criterion is

$$CV_{(K)} = \frac{1}{K} \sum_{j=1}^{K} MSE_j.$$

If each $n_j = \frac{n}{K}$) then $CV_{(K)} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(j)})^2$.

Pick the model I that minimizes

$CV_{(K)}(I)$. Usually $K = 10$ or 5.

If $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_J$,

then I_j corresponds to β_{λ_j}.

Hence for MLR, K-fold CV picks a model that is <u>good for prediction</u> of future Y_i. (Multitude of models)

II) Generate

$J = \lfloor \frac{n}{K} \rfloor + 1 = \lceil \frac{n}{K} \rceil$

1 2 3 ... K 1 2 3 ... K ... 1 2 3 ... K

get vector $\begin{matrix} \#\text{folds} \\ n \times 1 \end{matrix}$

folds 1 and 2 get 1 extra case