

TEST 1
October 16 2007

There are four questions. Each question is worth 25 points.
There are some extra pages of paper at the back.

NAME Selutians

1	
2	
3	
4	
Total	

(1) Consider these data:

X	-2	-1	0	1	2
Y	-2	-1	0	1	2

(a) We want to fit the usual simple linear regression model. Find $\hat{\beta}_0$ and $\hat{\beta}_1$. Construct the ANOVA table. Test the null hypothesis that $\beta_1 = 0$.

$$\hat{\beta}_0 = 0$$

$$\hat{\beta}_1 = 1$$

	df	SS	MS	F	P
Regression	1	10	10	∞	0
Residuals	3	0	0		
Total	4	10			

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Then $F \approx \infty \Rightarrow p\text{value} \approx 0$

So we reject H_0 .

(b) Using the same data as in (a), use generalized cross validation to choose between these two models:

Model 1: $Y = \beta_0 + \epsilon$

Model 2: $Y = \beta_0 + \beta_1 X + \epsilon$

$$GCV: \frac{1}{n} \frac{RSS(\text{model})}{\left(1 - \frac{|S|}{n}\right)^2}$$

Model 1: $|S|=0$, $RSS = 10$

$$\Rightarrow GCV = 2$$

Model 2: $|S|=1$, $RSS = 0$

$$\Rightarrow GCV = 0$$

\therefore Model 2 has a lower GCV.

(2) We are given data $(X_1, Y_1), \dots, (X_n, Y_n)$ where $X_i \in \mathbb{R}$ and $Y_i \in \mathbb{R}$. Assume that

$$Y_i = \beta X_i + \epsilon_i$$

where $\mathbb{E}(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$. Also assume that

$$X_1 < X_2 < \dots < X_n$$

and that n is even. Define

$$\hat{\beta} = \frac{(Y_2 - Y_1) + (Y_4 - Y_3) + \dots + (Y_n - Y_{n-1})}{\frac{n}{2}}$$

(a) Find the mean and variance of $\hat{\beta}$.

$$E[\hat{\beta}] = \frac{2\beta}{n} \sum_{i=1}^n (-1)^i X_i$$

$$V[\hat{\beta}] = \frac{4\sigma^2}{n}$$

Remember $V[Y_1 - Y_2] = V[Y_1] + V[Y_2]$
NOT $V[Y_1] - V[Y_2]$

(b) Assume that $X_i = i/n$. Compare the variance of $\hat{\beta}$ to the variance of the least squares estimator. Which is better?

In homework 2, Question 1, we found

$$V[\hat{\beta}_{LSE}] = \frac{\sigma^2}{\sum_{i=1}^n X_i^2}$$

$$\text{In this case: } = \frac{\sigma^2}{\sum_{i=1}^n (i/n)^2} = \frac{n^2 \sigma^2}{n(n+1)(2n+1)}$$

$$= \frac{6n \sigma^2}{2n^2 + 3n + 1}$$

$$= \frac{6 \sigma^2}{2n + 3 + 1/n}$$

which is better?

$$\min \left\{ \frac{4\sigma^2}{n}, \frac{6\sigma^2}{2n+3+1/n} \right\}$$

$$= \min \left\{ \frac{4}{n}, \frac{6}{2n+3+1/n} \right\}$$

$$= \frac{6}{2n+3+1/n} \Rightarrow \text{LSE is better.}$$

(3) We are given data $(X_1, Y_1), \dots, (X_n, Y_n)$ where $X_i \in \mathbb{R}$ and $Y_i \in \{0, 1\}$. Assume that

$$\mathbb{P}(Y = 1|X = x) = \frac{e^{\beta x}}{1 + e^{\beta x}}.$$

(a) Write down the likelihood function for β .

$$\begin{aligned} L(\beta) &= \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1 - y_i} \\ &= \prod_{i=1}^n \left(\frac{e^{\beta x_i}}{1 + e^{\beta x_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta x_i}} \right)^{1 - y_i} \\ &= \prod_{i=1}^n \frac{e^{\beta x_i y_i}}{1 + e^{\beta x_i}} \end{aligned}$$

(b) We can also estimate β with the estimator

$$\hat{\beta} = \frac{4 \sum_i Y_i X_i}{\sum_i X_i^2}.$$

Using the approximation

$$\frac{e^a}{1+e^a} \approx \frac{1}{2} + \frac{a}{4}$$

find the mean of $\hat{\beta}$. Are there conditions under which $\mathbb{E}(\hat{\beta}) \approx \beta$?

$$E[\hat{\beta}] = \beta + \frac{2 \sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2}$$

(almost everyone got this correct so I'll not show work)

$$E[\hat{\beta}] \approx \beta \quad \text{if} \quad \sum_{i=1}^n X_i = 0 \quad \text{or} \quad \sum_{i=1}^n X_i^2 \rightarrow \infty.$$

(4) We are given data $(X_1, Y_1), \dots, (X_n, Y_n)$ where $X_i \in \mathbb{R}$ and $Y_i \in \mathbb{R}$. Assume that

$$Y_i = \beta X_i + \epsilon_i$$

Consider the lasso estimator $\hat{\beta}$ that minimizes

$$\sum_{i=1}^n (Y_i - \beta X_i)^2 + \lambda |\beta|.$$

Suppose that $Y_i = X_i$ for each X_i . Find an expression for $\hat{\beta}$. Compare the bias of $\hat{\beta}$ to the bias of the least squares estimator.

Method 1

$$\sum_{i=1}^n (Y_i - \beta X_i)^2 + \lambda |\beta|$$

$$= \sum_{i=1}^n (Y_i - \beta Y_i)^2 + \lambda |\beta|$$

$$= (1 - \beta)^2 \sum_{i=1}^n Y_i^2 + \lambda |\beta|$$

~~Subdifferential~~ Subdifferential since function is convex:

$$* \quad -2(1 - \beta) \sum_{i=1}^n Y_i^2 + \lambda \omega(\beta) \quad \text{where } \omega(\beta) = \begin{cases} 1 & \beta > 0 \\ -1 & \beta < 0 \\ [-1, 1] & \beta = 0 \end{cases}$$

Case 1: $\beta > 0$. $* = 0$

$$\Rightarrow \hat{\beta} = 1 - \frac{\lambda}{2 \sum Y_i^2}$$

but we know $\hat{\beta} > 0$ so this is true for

$$\sum Y_i^2 > \lambda/2.$$

Case 2: $\beta < 0$ $* = 0$

$$\Rightarrow \hat{\beta} = 1 + \frac{\lambda}{2 \sum Y_i^2} \quad \text{but this is never } < 0. \text{ so ignore.}$$

Case 3: $\beta = 0$ $* = 0 \Rightarrow -2 \sum Y_i^2 + [-1, 1] \lambda = 0$

4 cont'd

$$\Rightarrow \hat{\beta} = 0 \text{ for } \sum Y_i^2 \in [-\lambda/2, \lambda/2]$$

but $\sum Y_i^2 \geq 0$ so

$$\hat{\beta} = \begin{cases} 1 - \frac{\lambda}{2\sum Y_i^2} & \sum Y_i^2 > \lambda/2 \\ 0 & 0 \leq \sum Y_i^2 \leq \lambda/2 \end{cases}$$

Method 2 $X_i = Y_i$ for all $i \Rightarrow \beta \geq 0$.

$$\text{So } \sum_{i=1}^n (Y_i - \beta X_i)^2 + \lambda |\beta|$$

$$= \sum_{i=1}^n (Y_i - \beta X_i)^2 + \lambda \beta$$

$$* = (1 - \beta)^2 \sum_{i=1}^n Y_i^2 + \lambda \beta$$

$$\frac{\partial *}{\partial \beta} = 0 \Rightarrow \hat{\beta} = 1 - \frac{\lambda}{2\sum Y_i^2}$$

but we know $\hat{\beta} \geq 0$ so we say

$$\hat{\beta} = 0 \text{ if } 1 - \frac{\lambda}{2\sum Y_i^2} \leq 0 \Leftrightarrow 0 \leq \sum Y_i^2 \leq \lambda/2$$

and we have the same result as above.

BLANK PAGE