+5/5

High-dimensional data analysis, Video class, 12/13(Tue) 14:00-16:50

Video from Lecture 2 - Part a - Statistical Learning with Applications in R - Linear Regression https://www.youtube.com/watch?v=QRzaKZRqens

Name Yuan - Hsin Lin.

Step 1: Read all questions before the video starts (**15minutes**)
(5 minutes break)
Step 2: See the video (**1 hour**). You can write answer during the video.
(5 minutes break)
Step 3: Complete answer (**up to 16:50**).

+2

A. Simple Linear Regression (*y*=sales and *x*=TV example)

1. Define RSS

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

2. Derive the LS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize RSS.

$$\frac{d\,RSS}{d\beta_0} = \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i) \overset{let}{=} 0$$

$$\frac{d\,RSS}{d\beta_1} = \sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_1 x_i) \overset{let}{=} 0$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2} = \frac{Sxy}{Sxx}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

LSE:

$$\sum x_i y_i = \sum x_i \beta_0 + \sum x_i^2 \beta_1$$

$$\sum y_i - n\beta_0 - \sum x_i \beta_1 = 0$$

$$\Rightarrow \sum y_i = n\beta_0 - n\beta_1\bar{x} \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$\sum x_i y_i = \sum x_i(\bar{y} - \hat{\beta}_1\bar{x}) + \sum x_i^2 \beta_1$$

$$\Rightarrow \beta_1 \sum(x_i-\bar{x})^2 = \sum(x_i-\bar{x})(y_i-\bar{y})$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}$$

3. Derive $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum(x_i-\bar{x})^2}$$

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}\right) = \frac{1}{\sum(x_i-\bar{x})^2} Var(y_i)$$

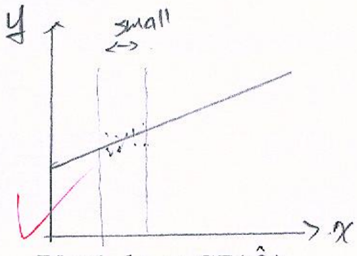
$$SE(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right]$$

$$Var(\hat{\beta}_0) = Var(\bar{y}) - \bar{x}\,Var(\hat{\beta}_1) = \frac{\sigma^2}{n} + \frac{\sigma^2\bar{x}}{\sum(x_i-\bar{x})^2} = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right]$$

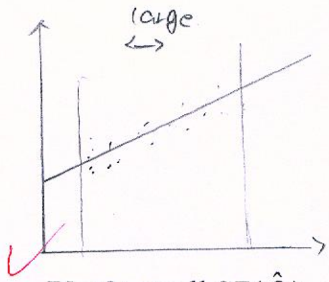
4. To make $SE(\hat{\beta}_1)$ small, how one can do for *x* ?

$$\sum(x_i-\bar{x})^2 = (n-1)Var(x) \quad \text{Let } Var(x)\uparrow \Rightarrow SE(\hat{\beta}_1)\downarrow$$

5. Draw the 2 plots of *x* and *y*:



Plot1: large $SE(\hat{\beta}_1)$



Plot2: small $SE(\hat{\beta}_1)$

6. Write an approximate 95% confidence interval for β_1.

$$\beta_1 \in [\hat{\beta}_1 - 2SE(\hat{\beta}_1), \hat{\beta}_1 + 2SE(\hat{\beta}_1)] \overset{\text{for TV}}{=} [0.042, 0.053]$$

7. What is the meaning of R^2 ?

R^2 measures how much of the variability of your data is captured by your linear model

8. Define R^2 by a formula.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \quad \text{where } TSS = \sum(\bar{y} - y_i)^2$$

9. Write R^2 in terms of the correlation between x and y?

$R^2 = r^2$ where $r = \dfrac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2}\sqrt{\sum (y_i-\bar{y})^2}}$

+1

10. Derive the above formula.

$TSS = \sum (y_i-\bar{y})^2 = \sum (y_i-\hat{y}+\hat{y}-\bar{y})^2 = \sum (y_i-\hat{y})^2 + 2\sum (y_i-\hat{y})(\hat{y}-\bar{y}) + \sum (\hat{y}-\bar{y})^2 = \sum (y_i-\hat{y})^2 + \sum \hat{\beta}_1^2 (x_i-\bar{x})^2$

$TSS-RSS = \sum \hat{\beta}_1^2 (x_i-\bar{x})^2 = \dfrac{[\sum x_i(y_i-\bar{y})]^2}{\sum (x_i-\bar{x})^2}$ $\quad \dfrac{TSS-RSS}{TSS} = \dfrac{[\sum (x_i-\bar{x})(y_i-\bar{y})]^2}{\sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2}$

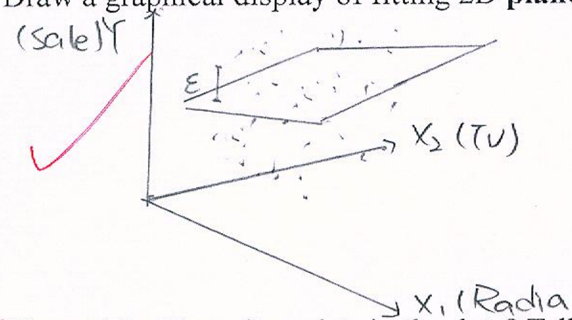
+2

B. Multiple Linear Regression

1. Write an interpretation of regression coeffcients in words.

A regression coefficient β_j estimates the expected change in Y per unit change in X_j, with other predictors held fixed.

2. Draw a graphical display of fitting 2D **plane** for X_1=Radia and X_2=TV.



$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

3. The multicollinearity exists in the data? Tell details (what variables, how much).

	TV	Radia	newspaper	sale
TV	1	0.0648	0.0567	0.7822
Radia		1	0.3541	0.5762
newspaper			1	0.2283
sale				1

~~Cov~~ Corr (X_1, X_3) = ~~Cov~~ Corr (Radia, TV) = 0.3541.

Exist multicollinearity but no stronger.

C. Variable selection

1. List up all 2-subset models for variables (X_1, X_2, X_3, X_4).

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ $\quad y = \beta_0 + \beta_1 X_2 + \beta_2 X_3$

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_3$ $\quad y = \beta_0 + \beta_1 X_2 + \beta_2 X_4$

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_4$ $\quad y = \beta_0 + \beta_1 X_3 + \beta_2 X_4$

2. How many combinations of subset models are possible for p variables?

$\binom{p}{1} + \binom{p}{2} + \dots + \binom{p}{p} = 2^p$

3. How many model fitting steps are necessary in forward selection with p=40 variables.

$\dfrac{P(P+1)}{2} = \dfrac{40(41)}{2} = 820$

4. Describe the backward selection.

1 Start with all variables in the model

2 Remove the variable with the largest p-value (the variable that is the least statistical significant

3 The new (p-1) variable model is fit, and the variable with the largest p value is removed

4 Continuous until a stopping rule is reached.