Sep 21, 2021
Dima Mikhaylov
Homework # 3

Stat 6021: Homework Set 3

(1.) "HW3_start.Rmd" is attached on UVA collab.

1. (R required) We will use the dataset "Copier.txt" for this question. The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, *Serviced* is the number of copiers serviced and *Minutes* is the total number of minutes spent by the service person.

   (a) What is the response variable in this analysis? What is predictor in this analysis?

   (b) Produce a scatterplot of the two variables. How would you describe the relationship between the number of copiers serviced and the time spent by the service person?

   (c) Use the `lm()` function to fit a linear regression for the two variables. Where are the values of $\hat{\beta}_1$, $\hat{\beta}_0$, R^2, and $\hat{\sigma}^2$ for this linear regression?

   (d) Interpret the values of $\hat{\beta}_1$, $\hat{\beta}_0$ contextually. Does the value of $\hat{\beta}_0$ make sense in this context?

   (e) Use the `anova()` function to produce the ANOVA table for this linear regression. What is the value of the ANOVA F statistic? What null and alternative hypotheses are being tested here? What is a relevant conclusion based on this ANOVA F statistic?

2. (Do not use R in this question) Suppose that for $n = 6$ students, we want to predict their scores on the second quiz using scores from the first quiz. The estimated regression line is

$$\hat{y} = 20 + 0.8x.$$

   (a) For each individual observation, calculate its predicted score on the second quiz $\hat{y}_i$ and the residual e_i. You may show your results in the table below.

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$

$\beta_0 = 20, \; \beta_1 = 0.8$

$e_i = y_i - \hat{y}_i$

x_i	70	75	80	80	85	90
y_i	75	82	80	86	90	91
$\hat{y}_i$	76	80	84	84	88	92
e_i	-1	2	-4	2	2	-1

$\bar{y} = 84$

$RSS = \sum_i (\hat{y}_i - \bar{y})^2 =$

$= 64 + 16 + 0 + 0 + 16 + 64 =$

$= 160 \Rightarrow MSR = \frac{160}{1} = 160$

ⓒ $s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{81+4+16+4+36+49}{6-1} = \frac{190}{5} = 38$ (*)

(b) Complete the ANOVA table for this dataset below. **Note:** Cells with *** in them are typically left blank.

$n=6, k=1$
$df(num) = 1$
$df(den) = 6-2=4$

	DF	SS	MS	F-stat	p-value
Regression	1	160	160	21.33	0.0099
Residual	4	30	7.5	***	***
Total	5	190	***	***	***

$SSE = \sum (y_i - \hat{y})^2 =$
$= 1+4+16+4+4+1$
$= 30 \Rightarrow MSE = \frac{30}{4} = 7.5$
$F = \frac{MSR}{MSE} = \frac{160}{7.5} = 21.33$

(c) Calculate the sample estimate of the variance σ^2 for the regression model. $s^2 = 38$ (*)

(d) What is the value of R^2 here? $R^2 = \frac{RSS}{SST} = \frac{160}{190} = 0.84$

(e) Carry out the ANOVA F test. What is an appropriate conclusion?
Reject Ho that $\beta_1 = 0 \Rightarrow$ accept Ha that β_1 is not zero with p-value = 0.009

3. (No R required) The least squares estimators of the simple linear regression model are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{2}$$

These are found by minimizing the sum of squared errors, i.e., minimize

$$SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{3}$$

Recall that fitted values and residuals from the fitted regression line are defined as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{4}$$

and

$$e_i = y_i - \hat{y}_i. \tag{5}$$

Using equations (1) to (5), show that the following equalities, (6) to (9), hold:

$$\sum_{i=1}^{n} e_i = 0 \tag{6}$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \tag{7}$$

$$\sum_{i=1}^{n} x_i e_i = 0 \tag{8}$$

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0. \tag{9}$$

Hint: Deriving the partial derivatives of the SS_{res}, (3), with respect to $\hat{\beta}_1$ and $\hat{\beta}_0$ will be useful.

Also, give a one-sentence interpretation of what the equalities (6) to (9) mean.

#6 $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)$ from (5) and (4)

Using least-squares criterion for estimating regression parameters from:

$S(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ and least-squares estimators must satisfy

$\left.\frac{\partial S}{\partial \beta_0}\right| = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$ and $\left.\frac{\partial S}{\partial \beta_1}\right| = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$

Taking derivative with respect to β_0 will give $\sum_{i=1}^{n}\underbrace{(y_i - \hat{y}_i)}_{e_i}\underbrace{\frac{\partial \hat{y}}{\partial \beta_0}}_{1} = \sum_{i=1}^{n} e_i \cdot 1 = 0$ #7

And taking derivative with respect to β_1 gives $\sum_{i=1}^{n}\underbrace{(y_i - \hat{y})}_{e_i}\underbrace{\frac{d\hat{y}}{d\beta_1}}_{x_i} = \sum_{i=1}^{n} e_i \cdot x_i = 0.$ #8

Since $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) \Rightarrow \sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$

Therefore, $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$. Finally, since $\sum_{i=1}^{n} e_i = 0 \Rightarrow$ #9

$\sum_{i=1}^{n} \hat{y} e_i = 0$. Conclusion: even without finding estimates β_0 and β_1 it follows from $\frac{\partial S}{\partial \beta_0}$ and $\frac{\partial S}{\partial \beta_1}$ that sum of the residuals be it in form of $\Sigma(y_i - \hat{y}_i)$ or $\Sigma e_i x_i$ or $\Sigma \hat{y}_i e_i$ will be zero, i.e. should cancel out by the way least-squares are fitted.