TEXAS A&M UNIVERSITY

Department of Statistics

STAT 608 - Regression Analysis

Homework I

Salih Kilicli

May 30, 2019

Question 1: Consider the following fitted simple linear regression line: $\hat{y} = 2.1 + 4.2x$

Solution: **If x increases by 1, then $\hat{y}$ increases by 4.2.**

Question 2: A straight line was fit by least squares to 50 pairs of points using the standard model $y = \beta_0 + \beta_1 x + e$ in which $var(e) = \sigma^2$ is independent of x. Suppose that all the usual assumptions are met. The following were part of the regression output:

- sample variance, S_y^2, of the 50 response values $= 100$
- estimate,$\hat{\sigma}^2$, of the error variance, based on 48 degrees of freedom $= 10$

Why S_y^2 and $\hat{\sigma}^2$ are so very different?

Solution : **The relation between S_y^2 and $\hat{\sigma}^2$ is given by the equation**

$$(n-1)S_y^2 = (SST = SSreg + RSS) = SSreg + (n-2)\hat{\sigma}^2$$

Since S_y^2 is large, this implies the variability in the data and as $\hat{y}_i$ differs from $\bar{y}$ for most of the values of x, SSReg increases. So this way the gap between them might be closed by SSreg due to the discrepancy in the data.

Question 3: You wish to estimate as precisely as possible the slope

in a straight line regression model $y = \alpha + \beta x + e$: Each pair of observations (x, y) costs $ 1.00 and your budget is $ 4.00. A data analyst proposes that you consider one of the following two options:

(a) Make two y-observations at $x = 1$ and a further two at $x = 4$;

(b) Make one y-observations at each of the points $x = 1, 2, 3,$ and 4.

Which of the two options would give you the most bang for your bucks? Show a relevant calculation to justify your choice.

Solution: **The answer is (a). First, notice that:**
$\bar{x}^{(a)} = \bar{x}^{(b)} = \frac{10}{4} = 2.5, \;\; SVar(x^{(a)}) = \frac{1}{3}SXX^{(a)} = 3 > \frac{5}{3} = SVar(x^{(b)}) = \frac{1}{3}SXX^{(b)}$.
Therefore $SXX^{(a)} = 9 > 5 = SXX^{(b)}$. Note that the variance of the least squares slope estimate decreases as SXX increases (i.e., as the variability in the X's increases). In conclusion, $Var(\hat{\beta}_1^{(a)}) < Var(\hat{\beta}_1^{(b)})$ and gives a better estimate.

Question 4: Work Exercise 4, page 40 in our textbook.

Solution: (a)

$$SSR = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

Since, $\hat{\beta}$ is the minimizer of SSR, and

$$\begin{aligned} \frac{\partial SSR}{\partial \beta} &= -2 \sum_{i=1}^{n} x_i (y_i - \beta x_i) = 0 \\ &= \beta \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0 \end{aligned}$$

Finally, we obtain $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

(b) (i)

$$\begin{aligned} E(\hat{\beta}|X) &= E\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} | X = x_i \right] = \frac{\sum_{i=1}^{n} x_i E[y_i | X = x_i]}{\sum_{i=1}^{n} x_i^2} \\ &= \frac{\sum_{i=1}^{n} x_i (\beta x_i)}{\sum_{i=1}^{n} x_i^2} = \frac{\beta \cancel{\sum_{i=1}^{n} x_i^2}}{\cancel{\sum_{i=1}^{n} x_i^2}} \\ &= \beta \end{aligned}$$

(ii)

$$\begin{aligned} Var(\hat{\beta}|X) &= Var\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} | X = x_i \right] = \frac{\sum_{i=1}^{n} x_i^2 \, Var[y_i | X = x_i]}{\left(\sum_{i=1}^{n} x_i^2 \right)^2} \\ &= \frac{\sum_{i=1}^{n} x_i^2 (\sigma^2)}{\left(\sum_{i=1}^{n} x_i^2 \right)^2} = \frac{\sigma^2 \cancel{\sum_{i=1}^{n} x_i^2}}{\left(\sum_{i=1}^{n} x_i^2 \right)^{\cancel{2}}} \\ &= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \end{aligned}$$

(iii) **The errors $e_i|X$ are normally distributed. Since $y_i = \beta x_i + e_i, (i = 1, 2, ..., n)$, $Y_i|X$ is normally distributed. Since $\hat{\beta}|X$ is a linear combination of the y_i's:**

$$\hat{\beta}|X \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \right)$$

Question 5: Show that the least squares criterion applied to the "intercept-only" model

$$y_i = \beta_0 + e_i,\ i = 1, ..., n$$

results in the least squares estimator $\hat{\beta}_0 = \bar{y}$ of β_0.

Solution:

$$SSR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0)^2$$

Then minimizing SSR yields;

$$\frac{\partial SSR}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0) = 0$$

$$n\beta_0 = \sum_{i=1}^{n} \beta_0 = \sum_{i=1}^{n} y_i$$

Therefore the least squares estimator;

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$$

Question 6: Work Exercise 7, page 42 in our textbook.

Solution: **It is possible because the confidence interval contains the population regression line 0.95 of the time and not 95% of the observations, which is what the prediction interval is meant to do. The confidence interval illustrates the variation in inferences on the regression parameters at every observed x value. This is different from variation that can be observed for the random variable Y (condition on $X = x$).**

Question 7: This problem is based on Exercise 1 page 38 in our textbook. See that problem for a description of the data and the model. The sales data are given in dollars. Give your answers below to 3 decimal places.

(a) The value of R Square is:

(b) The F value for testing $H_0 : \beta_0$ is:

(c) What is your best estimate of β_0 ?

(d) What is your best estimate of β_1 ?

(e) The lower and upper confidence limits for a 95% confidence interval for β_1 are:

(f) Is $\beta_1 = 1$ a plausible value for β_1? Yes or No. Give a reason to support your answer.

(g) The estimated gross box office receipts for the current week for a production with $\$\ 369,000$ in gross box office receipts the previous week is:

(h) A 95% confidence interval for the expected gross box office receipts for the current week for a production with $\$\ 369,000$ in gross box office receipts the previous week is:

(i) A 95 % prediction interval for the gross box office receipts for the current week for a production with \$ $369,000$ in gross box office receipts the previous week is:

(j) Is \$ $497,000$ a plausible value for the gross box office receipts for the current week for a production with \$ $369,000$ in gross box office receipts the previous week? Yes or No. Give a reason to support your answer.

(k) Some promoters of Broadway plays use the prediction rule that next week's gross box office receipts will be equal to this week's gross box office receipts. Comment on the appropriateness of this rule.

Solution:

(a) $R^2 \approx 0.997$

(b) $F_{value} \approx 4633.718$

(c) $\hat{\beta}_0 = 6804.886$

(d) $\hat{\beta}_1 = 0.982$

(e) The lower: 0.951 **and the upper:** 1.013**.**

(f) $\beta_1 = 1$ **a plausible value since it lies inside the** 95% **confidence interval.**

(g) The estimated gross box office receipts for the current week is ≈ 369193

(h) $CI = (357321.9, 381064)$

(i) $PI = (329215.4, 409170.5)$

(j) **\$** $497,000$ **is not a plausible value since it is far outside of the prediction interval.**

(k) Predicting the current week sales from the previous week is a somewhat reasonable strategy. Reformulating the question into terms of the model, predicting current sales using the exact previous sales is assuming that $\beta_0 = 0$ **and** $\beta_1 = 1$**. From our investigation in part (f) we know that** β_1 **can be** 1 **and one can show that** β_0 **could be** 0**. However, if one continually predicts the current week from the last week then the obvious and common sense trend of decreasing sales is not being recognized. Indeed, our data reveals a decreasing trend. More sample data or another model might yield better prediction rule.**