STAC67: Regression Analysis

Assignment # 01 Solution

September 26, 2014

Problem 1 - Exercise 1.19: **Grade point average.** [25 Marks] The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). The results of the study follow. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	$\cdots$	120
X_i:	21	14	$\cdots$	28
Y_i:	3.897	3.885	$\cdots$	2.948

a. [10 Marks] Obtain the least squares estimates of β_0 and β_1, and state the estimated regression function.

Here, $n = 120$, $\bar{X} = 24.725$, $\bar{Y} = 3.07405$, $\sum_{i=1}^n X_i^2 = 75739$, $\sum_{i=1}^n Y_i^2 = 1183.379$, $\sum_{i=1}^n X_i Y_i = 9213.112$, $\sum_{i=1}^n (X_i - \bar{X})^2 = 2379.925$, $\sum_{i=1}^n (Y_i - \bar{Y})^2 = 49.40545$, $\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = 92.40565$.

Hence,

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = 0.03882713.$$

and

$$b_0 = \bar{Y} - b_1 \bar{X} = 2.114049$$

Hence, the estimated regression line is

$$\hat{Y} = 2.114 + 0.0388 \times X$$

b. [5 Marks] Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?

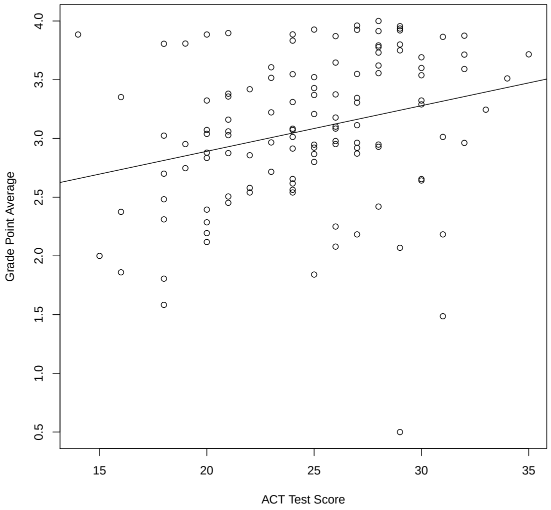

Figure 1: Plot of the estimated regression function and the data.

It seems the linear fit is not good for this data.

c. [5 Marks] Obtain a point estimate of the mean freshman GPA for students with ACT test score $X = 30$.

An estimate of the mean freshman GPA for students with ACT test score $X = 30$ is

$$\widehat{E(Y|X = 30)} = b_0 + b_1 \times 30 = 3.278863.$$

d. [5 Marks] What is the point estimate of the change in the mean response when the entrance test score increases by one point?

The change in the mean response for one point increase of the entrance test score is

$$b_1 = 0.0388.$$

Problem 2 - Exercise 1.23: Refer to **Grade point average** problem 1.19 [15 Marks]

a. [5 Marks] Obtain the residuals e_i. Do they sum to zero in accord with (1.17)?

The estimated residuals are

i:	1	2	$\cdots$	120
e_i:	0.96758105	1.2273709421	$\cdots$	-0.25320884

The sum of the residuals is zero

$$\sum_{i=1}^{120} e_i = 0.$$

b. [10 Marks] Estimate σ^2 and σ. In what units is σ expressed?

The unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{45.81761}{118} = 0.3882848.$$

The unbiased estimate of σ is

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{0.3882848} = 0.623125.$$

The unit σ is expressed is "grade points".

Problem 3 - Exercise 1.41: (Calculus needed.) Refer to the regression model $Y_i = \beta_1 X_i + \epsilon_i$, $i = 1, 2, \cdots, n$, in Exercise 1.29. [30 Marks]

a. [10 Marks] Find the least squares estimator of β_1.

We want to minimize

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2$$

We differentiate S with respect to β_1 and equate to 0 and get b_1

$$\frac{dS}{d\beta_1} = -2 \sum_{i=1}^{n} (Y_i - b_1 X_i) X_i = 0$$

Solving the above equation for b_1, we get

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

b. [5 + 5 + 2 = 12 Marks] Assume that the error terms ϵ_i are independent $N(0, \sigma^2)$ and that σ^2 is known. State the likelihood function for the n sample observations on Y and obtain the maximum likelihood estimator of β_1. It is the same as the least squares estimator?

From the stated assumption, we can write

$$Y_i \sim N(\beta_1 X_i, \sigma^2)$$

Hence,

$$f(Y_i | \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2} (Y_i - \beta_1 X_i)^2 \right\}$$

From the assumption of independent observations, we write our likelihood function as following

$$\begin{aligned} L(\beta_1, \sigma^2 | Y_1, \cdots, Y_n) &= f(Y_1, \cdots, Y_n | \beta_1, \sigma^2) \\ &= \prod_{i=1}^{n} f(Y_i | \beta_1, \sigma^2) \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2 \right\} \end{aligned}$$

The log likelihood function is written as

$$\log L(\beta_1, \sigma^2 | Y_1, \cdots, Y_n) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \left\{ \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2 \right\}$$

Differentiating this log likelihood function with respect to β_1 and equating to 0, we get

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_1 X_i) X_i = 0$$

Solving this equation for β_1, we get the maximum likelihood estimator of β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

The least squares estimator and the maximum likelihood estimator of β_1 are the same.

c. [8 Marks] Show that the maximum likelihood estimator of β_1 is unbiased.

We will show that $E(\hat{\beta}_1) = \beta_1$.
Here,

$$\begin{aligned} E(\hat{\beta}_1) &= \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2} E(Y_i) \\ &= \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2} (\beta_1 X_i) \\ &= \beta_1 \frac{\sum_{i=1}^n X_i^2}{\sum_{i=1}^n X_i^2} \\ &= \beta_1 \end{aligned}$$

Hence, $\hat{\beta}_1$ is an unbiased estimator of β_1.
Problem 4: Show that MSE is an unbiased estimator of σ^2. [20 Marks]
Solution: First show that

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

Then show that

$$E(\sum_{i=1}^n (Y_i - \bar{Y})^2) = \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 + E(\sum_{i=1}^n (\epsilon_i - \bar{\epsilon})^2) + 2\beta_1 \sum_{i=1}^n (X_i - \bar{X}) E(\epsilon_i - \bar{\epsilon}).$$

Here (please state the reasons),

$$E(\epsilon_i - \bar{\epsilon}) = 0$$

and

$$E(\sum_{i=1}^n (\epsilon_i - \bar{\epsilon})^2) = (n-1)\sigma^2.$$

Further, show that

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2.$$

Now prove that

$$E(\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2) = \sigma^2 + \beta_1^2 \sum_{i=1}^n (\hat{X}_i - \bar{X})^2$$

Combining the results, we get

$$E(\sum_{i=1}^n (Y_i - \hat{Y}_i)^2) = E(SSE) = (n-2)\sigma^2$$

Finally,

$$E(MSE) = E\left(\frac{SSE}{n-2}\right) = \sigma^2$$

This completes the proof.

Problem 5: Exercise 2.4 Refer to **Grade point average** Problem 1.19. [20 Marks]

a. [10 + 5 + 2 + 3 = 20 Marks] Obtain a 99 percent confidence interval for β_1. Interpret your confidence interval. Does it include zero? Why might the director of admissions be interested in whether the confidence interval includes zero?

The least squres estimate of β_1 is

$$b_1 = 0.03882713$$

The estimated variance of b_1 is

$$\widehat{\text{Var}\{b_1\}} = \frac{MSE}{\sum_{i=1}^{120}(X_i - \bar{X})^2} = \frac{0.3882848}{2379.925} = 0.00016315$$

For $\alpha = 0.01$ and $n = 120$, the tabulated value t_{tab} is

$$t_{tab} = t\{(1-\alpha/2),(n-2)\} = 2.618137$$

Hence, the 99% confidence interval of β_1 is

$$\left(b_1 - t_{tab} * \sqrt{\widehat{\text{Var}(b_1)}}, b_1 + t_{tab} * \sqrt{\widehat{\text{Var}(b_1)}}\right) = (0.005385614, 0.072268640).$$

Interpretation: If you draw samples of size 120 repeatedly, almost 99% of the estimates of β_1 will be in the interval $(0.005385614, 0.072268640)$.

This interval does not include zero.

The director of admission might be interested to check if the confidence interval includes 0 or not. Inclusion of 0 would indicate that the ACT test score has no significant effect on student's GPA. In this case, he/she might decide not to use ACT test score in student's admission decision.

Problem 6: Exercise 2.13 Refer to **Grade point average** Problem 1.19. [35 Marks]

a. [10 + 5 = 15 Marks] Obtain a 95 percent interval estimate of the mean freshman GPA for students whose ACT test score is 28. Interpret your confidence interval.

The estimated mean freshman GPA for students whose ACT test score is 28

$$\widehat{E(Y|X)} = b_0 + b_1 \times 28 = 3.201209$$

The estimated variance of mean freshman GPA for students whose ACT test score is 28

$$\widehat{\text{Var}(\widehat{E(Y|X)})} = \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] * MSE = 0.004985593$$

For $\alpha = 0.05$ and $n = 120$, the tabulated value t_{tab} is

$$t_{tab} = t\{(1-\alpha/2),(n-2)\} = 1.980272$$

Hence, the 95% confidence interval of $E\{Y|X = 28\}$ is

$$\left(\widehat{E(Y|X)} - t_{tab} * \sqrt{\widehat{\text{Var}(\widehat{E(Y|X)})}}, \widehat{E(Y|X)} + t_{tab} * \sqrt{\widehat{\text{Var}(\widehat{E(Y|X)})}}\right) = (3.061384, 3.341033).$$

Interpretation: If you draw samples of size 120 repeatedly, almost 95% of the estimates of $E\{Y|X = 28\}$ will be in the interval $(3.061384, 3.341033)$.

b. [10 + 5 = 15 Marks] Mary Jones obtained a score of 28 on the entrance test. Predict her freshman GPA using a 95 percent prediction interval. Interpret your prediction interval.

The prediction of Mary Jones's GPA is

$$\widehat{Y}_{h(new)} = b_0 + b_1 \times 28 = 3.201209$$

The estimated variance of prediction error for Mary Jones is

$$\widehat{\text{Var}(\widehat{Y}_{h(new)} - \widehat{Y}_h)} = \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] * MSE = 0.3932704$$

For $\alpha = 0.05$ and $n = 120$, the tabulated value t_{tab} is

$$t_{tab} = t\{(1 - \alpha/2), (n - 2)\} = 1.980272$$

Hence, the 95% prediction interval for Mary Jones GPA is

$$\left(\widehat{Y} - t_{tab} * \sqrt{\widehat{\text{Var}(\widehat{Y}_{h(new)} - \widehat{Y}_h)}}, \widehat{Y} + t_{tab} * \sqrt{\widehat{\text{Var}(\widehat{Y}_{h(new)} - \widehat{Y}_h)}}\right) = (1.959355, 4.443063).$$

Interpretation: Based on her ACT score 28, there is 95% probability that Mary Jones GPA would be in the interval $(1.959355, 4.443063)$.

c. [5 Marks] Is the prediction interval in part (b) is wider than the confidence interval in part (a)? Should it be?

YES, the prediction interval in part (b) is wider. This is expected as the prediction variability is larger than the variability of mean response.

Problem 7: Exercise 2.23 Refer to **Grade point average** Problem 1.19. [62 Marks]

a. [20] Set up the ANOVA table.

Table 1: Analysis of variance table.

Sources of Variation	SS	df	MS	F	p-value
Regression	3.587846	1	3.587846	9.240243	0.002916604
Error	45.81761	118	0.3882848		
Total	49.40545	119			

b. [4 + 4 + 2 = 10] What is estimated by MSR in your ANOVA table? by MSE? Under what condition do MSR and MSE estimate the same quantity?

MSR estimates $\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$ and MSE estimates the error variance σ^2. When $\beta_1 = 0$, then MSR and MSE both estimate the error variance σ^2.

c. [10 Marks] Conduct an F test of whether or not $\beta_1 = 0$. Control the α risk at 0.01. State the alternative, decision rule, and conclusion.

We want to test the following hypotheses

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_A : \beta_1 \neq 0.$$

The estimated F statistic is

$$F = \frac{MSR}{MSE} = 9.240243$$

At $\alpha = 0.01$ level of significance, $F(0.99, 1, 118) = 6.854641$.

Decision rule: we reject H_0 at α level of significance and conclude H_A if $F \geq F(1-\alpha, 1, 118)$. Here, we reject the null hypothesis at $\alpha = 0.01$ level of significance.

d. [5 + 5 + 2 = 12 Marks] What is the absolute magnitude of the reduction in the variation of Y when X is introduced into the regression model? What is the relative reduction? What is the name of the latter measure?

The absolute magnitude of the reduction in the variation of Y when X is introduced into the regression model is the *sum of squares of regression* or $SSR = 3.587846$.

The relative reduction in the variation of Y when X is introduced into the regression model to the variation of Y is

$$\frac{SSR}{SST} = \frac{3.587846}{49.40545} = 0.07262044$$

This latter measure is known as the *coefficient of determination.*

e. [5 Marks] Obtain r and attach the appropriate sign.

The correlation coefficient between Y (grade point average) and X (ACT test score) is

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} = 0.2694818$$

The strength of linear association between X and Y is 0.2694818.

f. [5 Marks] Which measure, R^2 or r, has the more clear-cut operational interpretation? Explain.

R^2 gives us the fraction of the total variability in Y that is explained by the regression model or X. On the other hand, r provides the strength of linear association between X and Y.

Here, R^2 provides the more clear-cut interpretation.