Stat 135 Summer 2018 – Homework 6 Solutions[1]

Analysis of Variance, Simple Linear Regression

Read sections 12.1-12.2, and 14.1,14.2 of the textbook and your class notes. *Mathematical Statistics and Data Analysis, 3rd ed.* by John Rice. To obtain full credit, please **write clearly** and **show your reasoning**.

Problem 9D
It is important to remember that a random variable in standard units has expectation 0, and variance 1, and expected square 1:

$$E(X^*) = \frac{E(X) - E(X)}{SD(X)} = 0 \; Var(X^*) = \frac{Var(X)}{(SD(X))^2} = 1 \Rightarrow E(X^{*2}) = 1$$

a) The first equality is true because multiplication is commutative. Next, note that Since $E(X) = E(Y) = 0$ and $SD(X) = SD(Y) = 1$, the definition says that $R(X,Y) = E(XY) = R(X,Y)$.

b) Using the hint,

$$E[(X^* + Y^*)^2] \geq 0 \Rightarrow E[X^{*2} + 2X^*Y^* + Y^{*2}] = 1 + 2\rho + 1 \geq 0 \Rightarrow \rho \geq -1$$

$$E[(X^* - Y^*)^2] \geq 0 \Rightarrow E[X^{*2} - 2X^*Y^* + Y^{*2}] = 1 - 2\rho + 1 \geq 0 \Rightarrow \rho \leq 1$$

Problem 9E Use the fact that $\sigma^2 = \frac{n-1}{n}s^2$.

$$r = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{s_x\sqrt{\frac{n-1}{n}}}\right)\left(\frac{y_i - \bar{y}}{s_y\sqrt{\frac{n-1}{n}}}\right) = \frac{1}{n}\frac{n}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Problem 9F

a) By definition, $r = Cov(x,y)/(\sigma_x\sigma_y)$. So $Cov(x,y) = r\sigma_x\sigma_y$. In class we derived a formula for the slope in terms of the covariance, and so:

$$\hat{b} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{Cov(x,y)}{\sigma_x^2} = \frac{r\sigma_x\sigma_y}{\sigma_x^2} = r\frac{\sigma_y}{\sigma_x} = r\frac{s_y}{s_x}$$

b) The formula $\hat{a} = \bar{y} - \hat{b}\bar{x} + \hat{b}x$. If you plut in $\bar{x}$ for x then you find $\hat{y} = \bar{y}$ and so the line goes through the point $(\bar{x}, \bar{y})$.

c) Look at the equation of the line in the previous part:

$$\hat{y}_i - \bar{y} = \hat{b}(x_i - \bar{x}) \Rightarrow \hat{y}_i - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x_i - \bar{x}) \Rightarrow \left(\frac{\hat{y}_i - \bar{y}}{\sigma_y}\right) = r\left(\frac{x_i - \bar{x}}{\sigma_x}\right)$$

d) If a student is 1.5 SDs above average on the midterm, then the students midterm score in standard units is 1.5. Therefore by part c the predicted final score in standard units will be 1.5r. For midterm and final scores r will be a number strictly between 0 and 1 (because you expect the association to be positive but not perfectly linear). So the predicted final score will be less than 1.5 SDs above average. Similarly if the students midterm score in standard units is 1.5, then the predicted final score will be 1.5r in standard units, that is, fewer than 1.5 SDs below the mean.

[1] attributed in part to Prof. Ani Adhikari

Problem 9G

a)

$$\frac{1}{n}\sum \hat{y}_i = \frac{1}{n}\sum(\hat{a}+\hat{b}x_i) = \hat{a}+\hat{b}\bar{x} = \bar{y}-\hat{b}\bar{x}+\hat{b}\bar{x} = \bar{y}$$

b) By the definition of variance,

$$\sigma^2_{\hat{y}} = \frac{1}{n}\sum(\hat{y}_i-\bar{\hat{y}})^2 = \frac{1}{n}\sum(\hat{y}_i-\bar{y})^2$$

by the result of part a. So

$$\sigma^2_{\hat{y}} = \frac{1}{n}\sum(\hat{a}+\hat{b}x_i-\bar{y})^2 = \frac{1}{n}\sum(\bar{y}-\hat{b}\bar{x}+\hat{b}x_i-\bar{y})^2 = \frac{1}{n}\sum(\hat{b}(x_i-\bar{x}))^2 = \hat{b}^2\sigma^2_x = r^2\sigma^2_y$$

When $r=0$ the regression line is flat; its equation is $\hat{y}=\bar{y}$. So fitted values are all the same and hence their variance is zero. When $r=1$ then the points all fall exactly on a line, which must be the same as the regression line. Hence the fitted values are the same as the observed values of y. So the variances are identical.

Problem 9H

a)

$$\bar{\hat{e}} = \frac{1}{n}\sum(\hat{y}_i-y_i) = \bar{\hat{y}}-\bar{y} = \bar{y}-\bar{y} = 0$$

b) Because $\bar{\hat{e}}=0$,

$$\sigma^2_{\hat{e}} = \frac{1}{n}\sum\hat{e}^2_i = \frac{1}{n}\sum(\hat{y}_i-y_i)^2 = \frac{1}{n}\sum(\bar{y}-\hat{b}\bar{x}+\hat{b}x_i-y_i)^2 = \frac{1}{n}\sum[\hat{b}(x_i-\bar{x})-(y_i-\bar{y})]^2$$

Expand the square to see that the expression becomes

$$\hat{b}^2\sigma^2_x - 2\hat{b}r\sigma_x\sigma_y + \sigma^2_y = \sigma^2_y(r^2-2r^2+1) = (1-r^2)\sigma^2_y$$

When $r=0$ we have the constant prediction $\bar{y}$, no matter what the value of x. The residuals are then just the deviations of the data (y) from their average ($\bar{y}$) and therefore the mean squared residual is just the variance σ^2_y. When $r=1$ then all the points lie on the regression line and the residuals are all zero. Hence their variance is also zero.

Problem 9I

a) Add the results of 9G part b and 9H part b:

$$\sigma^2_{\hat{e}} + \sigma^2_{\hat{y}} = (1-r^2)\sigma^2_y + r^2\sigma^2_y = \sigma^2_y.$$

b) This is the same equality as in part a. Just multiply each element in part a by 1/n.

Problem 9J

a) This is just 9I part b restated.

$$(y_i-\bar{y})^2 = [(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})]^2 = (y_i-\hat{y}_i)^2 + 2(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) + (\hat{y}_i-\bar{y})^2$$

Sum both sides and notice that the result of 9I part b implies that the cross-product term must be 0.

b) “Uncorrelated” means “correlation $= 0$” which is equivalent to “covariance $= 0$”’. Part a shows that the residuals are uncorrelated with the fitted values. Since the fitted values are a linear function of x, the residuals must also be uncorrelated with x. In order to prove this by algebra its enough to show that $Cov(x, \hat{e}) = 0$, as follows.

$$Cov(x, \hat{e}) = \frac{1}{n}\sum(x_i - \bar{x})(\hat{e}_i - \bar{\hat{e}}) = \frac{1}{n}\sum(x_i - \bar{x})(\hat{y}_i - y_i) = \frac{1}{\hat{b}n}\sum(\hat{b}x_i - \hat{b}\bar{x})(\hat{y}_i - y_i)$$

By a calculation that should now be familiar, the first term in the product is $\hat{y}_i - \bar{y}$. And therefore

$$Cov(x, \hat{e}) = \frac{1}{\hat{b}n}\sum(\hat{y}_i - \bar{y})(\hat{y}_i - y_i) = 0$$

by the result of part a.