Solution Guide, STA302 Midterm

LEC0101, 24 October 2017

1. [a] `t(y^0.7)` or `t(y)^.7` etc

[b] `p <- pchisq(8,5)`, `p <- pchisq(q=8,df=5)`, `p = pchisq(df=5,q=8)`, or some variation.

[c]

$$\sum_{i=1}^{n} \hat{e}_i \hat{y}_i = \sum_{i=1}^{n} \hat{e}_i \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) = \hat{\beta}_0 \sum_{i=1}^{n} \hat{e}_i + \hat{\beta}_1 \sum_{i=1}^{n} \hat{e}_i x_i = 0$$

2. Multiple choice:

B

False

A

B

B

3. [a] [i] Yes. In the scale-location plot, $\sqrt{r_i} > \sqrt{2}$ for *at least three* data points: 16, 20, and 82. (For those interested: these happen to correspond to the Fort York, Bayview, and Fairview libraries.)

Also acceptable: The r_i values for these three points are large and separated by a noticeable gap from those of the other 97.

[ii] Yes. The threshold for leverage is often taken to be $4/n$, which here is 0.04. Three points exceed that value, according to the lower-right plot. (For those interested: these happen to be the Runnymede and Agincourt libraries and particularly the North York Central Library.)

[iii] No. The plot of residuals vs leverage reveals that Cook's distance is well below 1 for all points.

[b] The point is a 'good leverage point'.

[i] No, the y value is near $\hat{y}$.

[ii] No, because the point's contribution to RSS is not extraordinary, and MSE = S^2 = RSS / (n-2).

[iii] Yes, R^2 will decrease without the point, because the high $(x_i - \bar{x})(y_i - \bar{y})$ contribution in the numerator of R^2 will be missing. Put another way, the strength of the linear relationship will be less apparent.

[c] There are a significant number of points with low values of $\hat{y}$ and low variance. Therefore, the assumption may not be valid. Error variance generally increases with $\hat{y}$.

[d] From the normal Q-Q plot, the residuals appear to be heavy-tailed. The assumption of normality is not likely to be valid.

[e] Examples:

- The three main outliers could be investigated to see whether a new model is needed or whether they could be removed. This would improve the line of best fit.
- To address the nonconstant error variance, a gentle transformation of y may help.

4. I: A $= -(2 - .4 - .6 + 1.2) = -2.2$ B $= 2.2/3.667 \approx 0.6$

C = 5 - 2 = 3 D $= 3.667^2 \approx 13.4$

E ≈ 0.0351

II: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -3.2 + 2.2x$

III: $H_0 : \beta_0 = 0$

5. There are a few possible approaches. One method starts by substituting the formula for the estimated intercept:

$$\text{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = \text{cov}\left(\bar{y} - \hat{\beta}_1 \bar{x},\, \hat{\beta}_1\right) = \text{cov}\left(\bar{y},\, \hat{\beta}_1\right) - \bar{x}\,\text{cov}\left(\hat{\beta}_1,\, \hat{\beta}_1\right)$$

Noting that $\hat{\beta}_1 = S_{xy}/S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})y_i \,/\, S_{xx} = \sum c_i y_i$ where c_i is a function of x only, the first term above can be rewritten as

$$\text{cov}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\,,\, \sum_{i=1}^{n} c_i y_i\right) = \frac{1}{n}\sum_{i=1}^{n} c_i \text{var}(y_i) = \frac{\sigma^2}{n}\sum_{i=1}^{n} c_i = 0$$

This leaves

$$\text{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = 0 - \bar{x}\,\text{var}\left(\hat{\beta}_1\right) = \frac{-\bar{x}\,\sigma^2}{S_{xx}}$$

Another method is to rewrite $\text{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = \text{cov}\left(\sum_{i=1}^{n} d_i y_i\,,\, \sum_{i=1}^{n} c_i y_i\right) = \sigma^2 \sum_{i=1}^{n} d_i c_i$ which, after finding d_i, leads to the same result.