Solution Guide, STA302 Midterm

LEC5101, 26 October 2017

6. [a] All of the data points are in a perfectly straight line (collinear).

[b] Answers include `x <- rt(1,4)`, `x = rt(n=1,df=4)`, `x <- rt(df=4,n=1)`, or some variation on those.

[c] To express $\hat{\beta}_1 = \sum_{i=1}^n g(x_i)\, y_i$, we can start from the first line of the aid sheet:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})\, y_i - \cancelto{0}{\sum_{i=1}^n (x_i - \bar{x})\, \bar{y}}}{S_{xx}} \\
&= \sum_{i=1}^n \frac{(x_i - \bar{x})}{S_{xx}}\, y_i
\end{aligned}
$$

And let $g(x_i) = \sum_{i=1}^n \frac{(x_i - \bar{x})}{S_{xx}}$.

7. Multiple choice:

D

C

A

B

D

8. Part I:

[a]

$$\beta_1 \pm t(0.025, n-2)\ \mathrm{se}\left(\hat{\beta}_1\right) \approx 0.3993 \pm 2.4469\,(0.3427) = (-0.44,\ 1.24)$$

[b] Solution A: No, prediction intervals are wider than confidence intervals. To be within three significant figures, the $+1$ term in the variances must have very, very little effect. Compare

$$\mathrm{var}(\hat{y}^*) = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right] \qquad \text{versus} \qquad \mathrm{var}(Y^* - \hat{y}^*) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$

Because $x^* = \bar{x} = 5$, the ratio of these two variances is $1/n$ to $1 + 1/n$, or one to nine. The PI will be noticeably wider than the CI, as usual.

Solution B: No. $S \approx \sqrt{2.816} \approx 1.678$, and at $x^* = \bar{x}$, we have $\hat{y}^* = \bar{y} = 7$, $\mathrm{var}(\hat{y}^*) = 0.125\,\sigma^2$, and $\mathrm{var}(Y^* - \hat{y}^*) = 1.125\,\sigma^2$. Therefore:

$$\mathrm{CI} = \hat{y}^* \pm t(0.975, 6)\, 0.125\, S \approx (6.49,\ 7.51) \quad \text{and} \quad \mathrm{PI} = \hat{y}^* \pm t(0.975, 6)\, 1.125\, S \approx (2.4,\ 11.6)$$

giving widths of 1.03 versus 9.24.

8. Part II:

[c] Since the t-test gave a result that was not statistically significant, we do not reject the null hypothesis. Namely, we conclude that the populations behind `y1` and `y2` are unlikely to have different means.

[d] iii

[e] iv

9.

A. VI, At least one outlier

B. V, nonconstant variance

C. VI, outlier

D. II, right skew

E. VI, VII, VIII: outlier, leverage point, influential point

F. VI, outlier

10. [a] Yes. One of the GM conditions is $\mathrm{E}(e) = 0$. In all cases, we also have $\sum_{i=1}^{n} \hat{e}_i = 0$.

[b]

$$\begin{aligned}
\mathrm{cov}\left(\hat{y}_i,\, y_j - \hat{y}_j\right) &= \mathrm{cov}\left(\hat{y}_i,\, y_j\right) - \mathrm{cov}\left(\hat{y}_i,\, \hat{y}_j\right) \\
&= \mathrm{cov}\left(\sum_{k=1}^{n} h_{ik} y_k,\, y_j\right) - \mathrm{cov}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i,\, \hat{\beta}_0 + \hat{\beta}_1 x_j\right) \\
&= \mathrm{cov}\left(h_{ij} y_j,\, y_j\right) - \mathrm{var}\left(\hat{\beta}_0\right) - x_i \mathrm{cov}\left(\hat{\beta}_0,\, \hat{\beta}_1\right) - x_j \mathrm{cov}\left(\hat{\beta}_1,\, \hat{\beta}_0\right) - x_i x_j \mathrm{var}\left(\hat{\beta}_1\right) \\
&= h_{ij} \mathrm{var}(y_j) - \sigma^2 \left(\frac{1}{n} + \bar{x}^2 / S_{xx} + (x_i + x_j)\sigma^2 \bar{x} / S_{xx} - x_i x_j \sigma^2 / S_{xx}\right) \\
&= \sigma^2 h_{ij} - \sigma^2 \left[\frac{1}{n} + \frac{1}{S_{xx}}\left(\bar{x}^2 - \bar{x}(x_i + x_j) + x_i x_j\right)\right] \\
&= \sigma^2 h_{ij} - \sigma^2 \left[\frac{1}{n} + \frac{1}{S_{xx}}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)\right] \\
&= \sigma^2 h_{ij} - \sigma^2 h_{ij} \\
&= 0
\end{aligned}$$