Week 7 Solutions

Example 1: Hypothesis Testing

(i) `time20` $= 2.806 + 0.863$`reaction`

(ii)
- $\hat{\sigma} = 0.02455$, the residual/estimated standard error
- So $\hat{\sigma}^2 = 0.02455^2 = 0.0006027$

(iii)
- $\hat{\beta}_1 \pm t_{n-2,0.025}\sqrt{\frac{\hat{\sigma}^2}{SS_{XX}}}$
- Some of the R output can be helpful here:
- $\sqrt{\frac{\hat{\sigma}^2}{SS_{XX}}} = \text{e.s.e}(\hat{\beta}_1) = 0.3911$
- $t_{n-2,0.025} = t_{13,0.025} = 2.16$
- So 95% confidence interval is

$$\begin{aligned} \hat{\beta}_1 &\pm 2.16 \times 0.3911 \\ = 0.863 &\pm 0.845 = \underline{(0.018, 1.708)} \end{aligned}$$

(iv)
- Recall that R-output `t-value` for `reaction` is

$$2.207 \left(= \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{XX}}} \right)$$

- $t_{n-2,0.025} = t_{13,0.025} = 2.16 < 2.207$
- So we would reject $H_0 : \beta_1 = 0$ at the $\alpha = 0.05$ level
- Conclude that there is evidence that reaction time has a significant linear influences on the time to reach $20m$.
- In addition, the estimate $\hat{\beta}_1$ is positive, indicating that the shorter reactions times lead on average to significantly faster run times.

Example 2: Estimating Intervals

(i) The p-value for `Girth` is shown as `< 2e-16`. This is less than 0.01 hence we reject the null hypothesis and conclude that there is a linear association between `Girth` and `Volume` that is significant at the 1% level

(ii)
```
confint.lm(trees3.lm)
```

```
##                   2.5 %      97.5 %
## (Intercept) -75.68226247 -40.2930554
## Height        0.07264863   0.6058538
## Girth         4.16683899   5.2494820
```

This indicates that the average increase in `Volume` is highly likely to be between 0.0726 and 0.6058 cubic feet for every increase of one foot in the height of a tree.

(iii)
```
confint.lm(trees3.lm,level=0.9)
```

(iv)
```
# NOTE: this **isn't** a prediction interval, despite making
# use of the PREDICT function !!!
predict(trees3.lm, data.frame(Height = 65, Girth = 9.7),
interval = "confidence", level = 0.95)
```

Note: For a 95% interval, you do not need to specify the level argument as it produces this level of interval by default.

(v)
```
predict(trees3.lm, data.frame(Height = 65, Girth = 9.7),
interval = "prediction", level = 0.95)
```

Note: For a 95% interval, you do not need to specify the level argument as it produces this level of interval by default.

(vi) The prediction interval is wider than the confidence interval. This is because a prediction interval refers to the estimated value for an individual in the population, while a confidence interval refers to the estimated value for the population average. Since individual values are more variable than average values, the prediction interval has to be wider.

Example 3: Estimating intervals by hand using R output

(i) This is the part labelled `$fit` so 1463.966

To calculate this by hand obtain coefficients from R:

2881 - 25.02×14 - 0.005334×200000 = 1463.92 (note the difference is due to rounding)

(ii) The 95% confidence interval is calculated as follows:

$$\begin{aligned} \hat{y} &\pm t_{n-p-1,\alpha/2}\text{e.s.e}(\hat{y}) \\ 1463.966 &\pm 1.96 \times 33.62318 \\ &(1398.059, 1529.861) \end{aligned}$$

(iii) The 95% confidence interval is calculated as follows:

$$\begin{aligned} \hat{\beta}_1 &\pm t_{n-p-1,\alpha/2}\text{e.s.e}(\hat{\beta}_1) \\ -25.02 &\pm 1.96 \times 2.362 \\ &(-29.64952, -20.39048) \end{aligned}$$

This suggests that there is a significant linear association between `no2` and hospital admissions such that an increase of 1μg/l in NO_2 concentration is associated with a decrease in admissions of between 20.39048 and 29.64952. Given that pollution is bad for health (and particularly bad for respiratory health) this is surprising - it is worth noting however that the observation of this effect is likely due to some variables not being present in the model - e.g. a measure of deprivation.