2020

Question 1

i)

You can calculate $\bar{x}$ to be 10. So the matrix form becomes:

$$\begin{pmatrix} 3 \\ 5 \\ 24 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & -2 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

ii)

Finding b_0 first (you should take second derivative to show minimum but I leave that to you):

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1(x_i - \bar{x}))$$

$$\sum_{i=1}^{n}(y_i - b_0 - b_1(x_i - \bar{x})) = 0$$

$$\sum_{i=1}^{n} y_i - nb_0 - b_1\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

$$nb_0 = \sum_{i=1}^{n} y_i$$

$$b_0 = \bar{y}$$

Now finding b_1:

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \beta_0 - \beta_1(x_i - \bar{x}))$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - b_0 - b_1(x_i - \bar{x})) = 0$$

$$S_{xy} - \beta_0\sum_{i=1}^{n}(x_i - \bar{x}) - b_1 S_{xx} = 0$$

$$b_1 S_{xx} = S_{xy}$$

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

iii)

$$\begin{aligned} E(b_0) &= E(\bar{y}) \\ &= E\left(\beta_0 + \beta_1\sum_{i=1}^{n}(x_i - \bar{x})\right) \\ &= E(\beta_0) \\ &= \beta_0 \end{aligned}$$

$$
\begin{aligned}
E(b_1) &= E\left(\frac{S_{xy}}{S_{xx}}\right) \\
&= \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 - \beta_1(x_i - \bar{x}) + \epsilon_i)\right) \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x})^2 \beta_1 \\
&= \beta_1
\end{aligned}
$$

The β_0 term becomes 0 due to $\sum_{i=1}^{n}(x_i - \bar{x})$, and $E(\epsilon_i) = 0$.

iv)

$$
\begin{aligned}
\text{Cov}(b_0, b1) &= \text{Cov}\left(\bar{y}, \frac{S_{xy}}{S_{xx}}\right) \\
&= \frac{1}{S_{xx}} \text{Cov}\left(\bar{y}, \sum_{i=1}^{n}(x_i - \bar{x}) y_i\right) \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x}) \text{Cov}(\bar{y}, y_i) \\
&= \frac{\sigma^2}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x}) \\
&= 0
\end{aligned}
$$

For the variance part, just note that:

$$
\text{Var}(b) = \begin{pmatrix} \text{Var}(b_0) & \text{Cov}(b_0, b1) \\ \text{Cov}(b_1, b_0) & \text{Var}(b_1) \end{pmatrix}
$$

but you should just show it from first principles by doing $\text{Var}(A) = E((A - E(A))(A - E(A))^T)$.

v)

Note that since $\epsilon_i \sim N(0, \sigma^2)$, that the responses are also normally distributed, and hence b_0 and b_1 are normally distributed. Note:

$$
b_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)
$$

$$
b_0 \sim N\left(\beta_0, \frac{\sigma^2}{n}\right)
$$

Therefore, we can standardise these to:

$$
\frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim Z
$$

$$
\frac{b_0 - \beta_0}{\sigma/\sqrt{n}} \sim Z
$$

Doing some minor algebraic manipulation:

$$
\begin{aligned}
\left(\frac{b_0 - \beta_0}{\sigma/\sqrt{n}}\right)^2 + \left(\frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}\right)^2 &= Z^2 + Z^2 \\
&= \chi_1^2 + \chi_1^2 \\
&= \chi_2^2
\end{aligned}
$$

Question 2

i)

1. The residuals vs fitted graph has a shape, which violates the linear assumption
2. The QQ plot deviates, which violates the normality assumption

ii)

1. There's less shape to the residuals vs fitted graph, which restores linearity. Theres also less deviation from the QQ normal line, so normality assumption is also improved.

   The model is:

   $$\log \text{Price} = 1.449734 \cdot \log \text{Age} + 0.066160 \cdot \text{Bidders}$$

   This is easy so you can just plug in the numbers yourself

2. This is the F-test. The F-test considers the hypotheses:

   $$H_0 : \beta_{Age} = \beta_{Bidders} = 0$$
   $$H_1 : \text{Not all betas are } 0$$

   The F statistic is 167.2 for this test on $F_{2,27}$. The p-value is 6.189×10^{-16}, which indicates, with a significance value of 5%, that we can reject the null hypothesis.

3. This is testing the hypotheses:

   $$H_0 : \beta_{Bidders} = 0$$
   $$H_1 : \beta_{Bidders} \neq 0$$

   We can use the t-test for this, with a t value of 10.611, and a p-value of 3.92×10^{-11}. Using a signifiance level of 5%, we can reject the null hypothesis, and deem that the model with bidders is better.

4. There's just one trick to this question; in the fact that it's **externally** studentized. That means while the summary gives us SS_{res} with 27 DOF, we have 26 for the test. So we should use `qt(0.975, 26)`, and from this, we can see that the residual is an outlier using a significance level of 5% (as $-2.2504 < -2.055529$).

Question 3

Probably the only interesting question of this paper, and new from the previous 2 papers.

i)

$$X = \begin{pmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

ii)

$$X^T \mathbb{I} = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & \dots & 1 \\ \dots & \dots & \dots \\ 1 & \dots & 1 \end{pmatrix} = \begin{pmatrix} n & \dots & n \\ \sum_{i=1}^n x_i & \dots & \sum_{i=1}^n x_i \end{pmatrix}$$

Now:

$$X^T \mathbb{I} X = \begin{pmatrix} n & \dots & n \\ \sum_{i=1}^n x_i & \dots & \sum_{i=1}^n x_i \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n^2 & n\sum_{i=1}^n x_i \\ n\sum_{i=1}^n x_i & \left(\sum_{i=1}^n x_i\right)^2 \end{pmatrix}$$

With a scalar multiple of $\frac{1}{n}$, this becomes:

$$\begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2 \end{pmatrix}$$

iii)

Note that $X^T H X = X^T X (X^T X)^{-1} X^X = X^T X$. Therefore, we have:

$$\begin{aligned} X^T H X - \frac{1}{n} X^T \mathbb{I} X &= \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} - \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 \\ 0 & \sum_{i=1}^n x_i^2 - n\bar{x}^2 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 \\ 0 & \sum_{i=1}^n x_i^2 - \bar{x}^2 \end{pmatrix} \qquad \text{Inclusive in the sum} \\ &= \begin{pmatrix} 0 & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})(x_i + \bar{x}) \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 \\ 0 & S_{xx} \end{pmatrix} \qquad \text{One of the sums} = 0 \text{ due to } x_i - \bar{x} \end{aligned}$$

iv)

$$\begin{aligned} \operatorname{tr}\left(H - \frac{1}{n}\mathbb{I}\right) &= \operatorname{tr}(H) - \operatorname{tr}\left(\frac{1}{n}\mathbb{I}\right) \\ &= \operatorname{tr}((X^T X)^{-1} X^T X) - 1 \\ &= 2 - 1 \\ &= 1 \end{aligned}$$

v)

$$
\begin{aligned}
E(SS_{reg}) &= E\left(y^T(H - \frac{1}{n}\mathbb{I})y\right) \\
&= \text{tr}((H - \frac{1}{n}\mathbb{I})\sigma^2) + \beta^T X^T (H - \frac{1}{n}\mathbb{I})X\beta \\
&= \sigma^2 \text{tr}(H - \frac{1}{n}\mathbb{I}) + \beta^T \begin{pmatrix} 0 & 0 \\ 0 & S_{xx} \end{pmatrix} \beta \\
&= \sigma^2 + \beta_1^2 S_{xx}
\end{aligned}
$$

Question 4

i)

Observations 11 and 13 are likely influential, as they have a Cook's Distance > 0.5. This means that they have a significant impact on the β values if they are removed as observations, and thus, are likely impacting the inferences of the model.

ii)

$$D_{11} = \frac{(-2.75414799)^2(0.37847203)}{3 \cdot (1 - 0.37847203)} = 1.54$$

$$D_{13} = \frac{(2.60073093)^2(0.43661124)}{3 \cdot (1 - 0.43661124)} = 1.75$$

iii)

Cook's distance is a metric that measures the distance between the parameters b with observation i, and the parameters b_{-i}, removing observation i. If an observation was to be *influential* to inference, it follows that they should be influential to the fit parameters values. Hence, a high Cook's Distance implies an influential observation.

iv)

Remember that:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

Now,

$$\begin{aligned}
D_i &= \frac{\left(\frac{(X^TX)^{-1}x_i e_i}{1-h_{ii}}\right)^T (X^TX) \left(\frac{(X^TX)^{-1}x_i e_i}{1-h_{ii}}\right)}{p\hat{\sigma}^2} \\
&= \frac{e_i^2 x_i^T (X^TX)^{-1}(X^TX)(X^TX)^{-1}x_i}{p\hat{\sigma}^2(1-h_{ii})^2} \\
&= \frac{e_i^2 x_i^T (X^TX)^{-1}x_i}{p\hat{\sigma}^2(1-h_{ii})^2} \\
&= \frac{e_i^2}{\hat{\sigma}^2(1-h_{ii})^2}\frac{h_{ii}}{p} \\
&= \frac{r_i^2 h_{ii}}{p(1-h_{ii})} \qquad \text{Substituing the earlier identity}
\end{aligned}$$