Linear Regression 3

Iris Kim (6757827) and Ted Tinker (3223468)

April 22, 2017

Question 1

Because $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\frac{\sum_{i=1}^{n} x_i}{n} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$. Rearranging the second term of equation (2), we find

$$\frac{2(x_i - \overline{x})}{n \times S_{XX}} \sum_{j=1}^{n}(x_j - \overline{x}).$$

Therefore, this term should go to 0. Recalling $S_{XX} = \sum_{i=1}^{n}(x_i - \overline{x})^2$, the last term of equation (2) is

$$(x_i - \overline{x})^2 \sum_{j=1}^{n} \frac{(x_j - \overline{x})^2}{(x_j - \overline{x})^4} = (x_i - \overline{x})^2 \sum_{j=1}^{n} \frac{1}{(x_j - \overline{x})^2}$$

So, $h_{ij}^2 = \frac{1}{n} + 0 + \frac{(x_i - \overline{x})}{S_{XX}} = h_{ii}$.

Question 2

The confusion stems from the difference between a Confidence Interval and a Prediction Interval. The observations Y^* that we see for a given sample x^* are random variables, not parameters, and so a confidence interval would not be used to describe it in this way. The problem states that the report contained "associated 95% confidence intervals for the regression function at each x value," which refers to the parameters $\mathbb{E}(Y|X = x)$, not the particular set of observations displayed.

Question 3, Part A, i

We decide whether the distance is significant using a hypothesis test: The null hypothesis, H_0, is that the slope of the regression line is 0; the alternative hypothesis, H_a, is that the slope is non-zero, and therefore distance has an impact on fare. Using

```
attach(airfares) # Tells R we are using the airfares dataset
airLine <- lm(Fare~Distance) # Generates the linear regression of airfares
confint(airLine) # Prints confidence interval
```

we produce 95% confidence intervals for parameters of airLine, the regression line relating Fare and Distance in this data-set.

```
confint(airLine)
                2.5 %     97.5 %
(Intercept) 39.5816846 58.3618563
Distance     0.2102642  0.2291104
```

The parameter for the slope is said to have 95% Confidence Interval (.2102642, .2291104). This does not include 0, so by a Confidence Interval test, we reject the null hypothesis and agree that distance has an effect on air fare. We agree with the analyst, here.

ii

```
summary(airLine) # Provides details about the regression slope
```

tells us the multiple R-squared value is .994, which is what the business analyst probably means when they say that the model explains 99.4$ of the data. We may derive it in R without using summary() using the following code:

```
attach(airfares) # Tells R we are using the airfares dataset
airLine <- lm(Fare~Distance) # Generates the linear regression of airfares
y.obs <- Fare # Makes y.obs the value of the fares
y.ave <- mean(Fare) # Makes y.ave the average fare
y.hat <- airLine$fitted.values # Makes y.hat the expected fares for each observation

SST = sum((y.obs - y.ave)^2) # Find total sum of squares
RSS = sum((y.obs - y.hat)^2) # Find residual sum of squares
SSReg = sum((y.hat - y.ave)^2) # Find regression sum of squares

R = sqrt(1- RSS/SST) # Or, alternatively, sqrt(SSReg/SST)
```

Using this code, we find $R \approx .996$.

iii

The values that the analyst gives are correct, and they interpret them correctly, but there is a pattern in the data which they missed. Therefore, even if their analysis is correct, it might not very helpful, as the data cannot be said to fit a linear model. We explain this pattern in Part B:

Question 3, Part B

Looking at the graph of the residuals of our data-set, there should be no pattern; the points should be distributed above and below the regression line fairly normally. However, this is not the case: with the exception of one leverage point two standard deviations away from the regression line, the points appear to be parabolically related. A purely linear model is not appropriate for this data-set.

To improve our interpretation of the data, we recommend removing that leverage point, the point farthest on the right.