Regression – Final Exam Solutions

Question 1. Multiple linear regression

1. Fit a model with the following explanatory variables: DAY_OF_WEEK, DISTANCE, AIRLINE, DESTINATION_AIRPORT, SCHEDULED_DEPARTURE, SCHEDULED_ARRIVAL. **What is the resulting coefficient of determination? What variable(s) might be excluded to reduce the complexity of the model?**

   R-squared for this model: 0.1339
   Destination airport is a factor variable with 157 levels (meaning there are flights going from ATL to 157 other airports in this sample of data). Thus, including this single explanatory variable (destination airport) involves adding 156 dummy variables to the model. This makes the model very complicated, and not very easy to use for prediction, so this variable may be excluded to reduce complexity.

   Any reasonable answer with logical explanation will be accepted.

2. Use the step function to perform a stepwise model selection minimizing AIC (using direction = "both") on the model you created in the previous step. **What variables are selected?**

   The resulting variables for this model are: DAY_OF_WEEK + AIRLINE + SCHEDULED_DEPARTURE + DESTINATION_AIRPORT

3. Create a model using these variables selected from stepwise model selection. **Did the adjusted R-squared value increase?**
   The multiple R-squared value decreased to 0.1338
   ADJUSTED R-SQUARED VALUE STAYS THE SAME

4. Use cooks distance (with distance greater than 1) on the model resulting from the stepwise model selection to detect outliers in the data, and remove any outliers you may have found. Provide the commands used to calculate cooks distance and remove any points from the data.

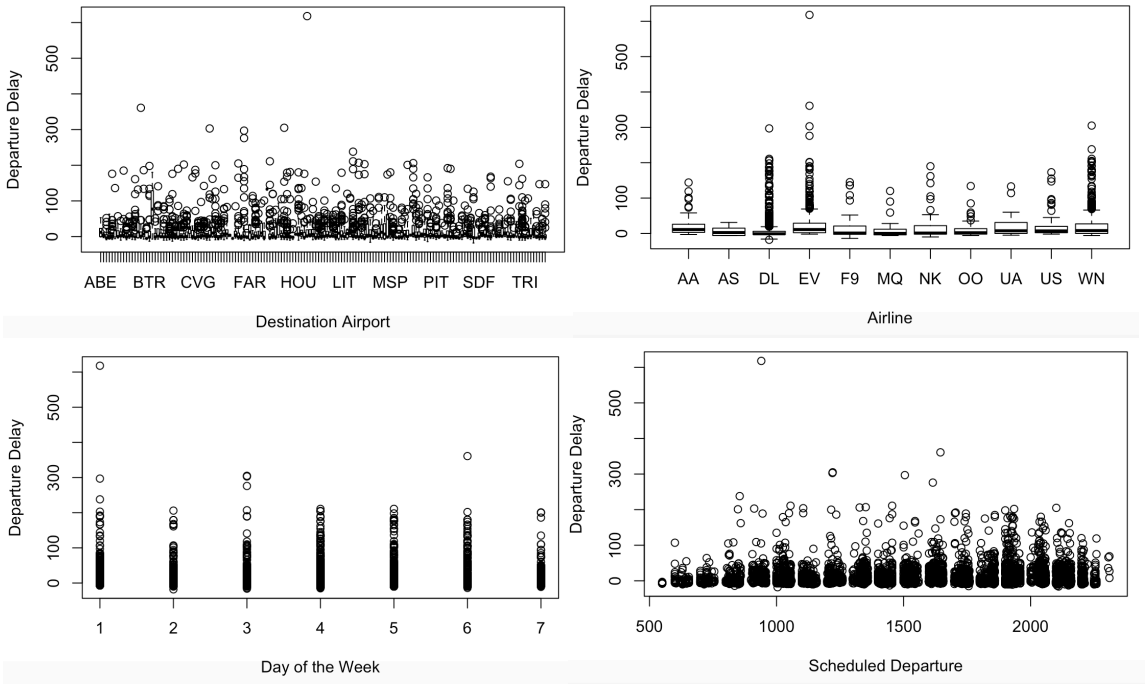
```
cook = cooks.distance(step_model)
which(cook > 1)

newdata = data[-578,]
```

5. Fit a model again using the variables selected with stepwise model selection, and excluding any points you may have removed in the previous problem. **Provide the R**

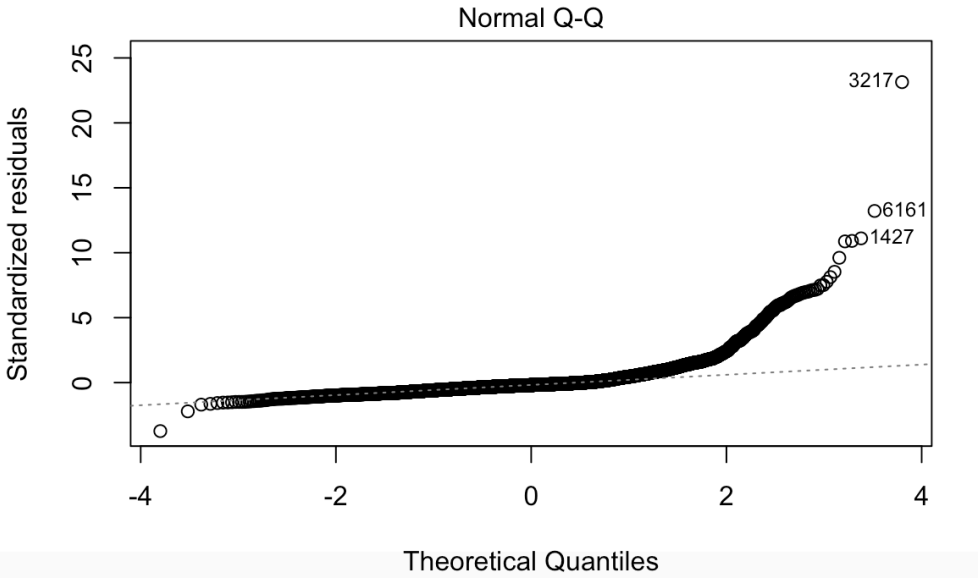
commands and the resulting residual standard error and corresponding degree of freedom for your model.

This model has a residual standard error of 25.56 on `6716` degrees of freedom

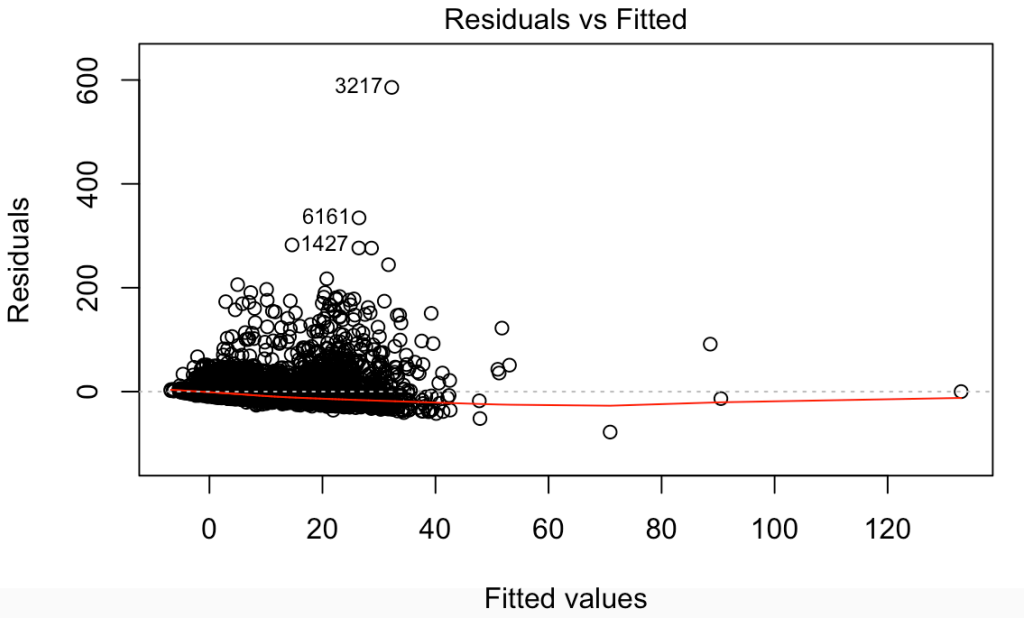
6. Check the model assumptions (for the model resulting from question 5) using the appropriate plots. **Provide the resulting plots. What can you conclude about this model?**



Linearity: Based on the scatterplots above, we don't really see a clear linear relationship between the response variable and any of the predicting variables. There is a clear relationship between Airline and departure delay in a few cases, but because this is a factor variable, we still can't say that there is a linear relationship between explanatory and predicting variables.



Normality: Based on the Q-Q Norm plot, we can see that our data is not exactly normally distributed. There are very heavy tails on the upper side, and it appears that is still at least one outlier still as well.



Constant Variance: Based on the residuals vs. fitted values we can conclude that the constant variance assumption does not hold for this data. There are a few negative values (flights that actually left before their scheduled departure), however, overall the residuals are definitely skewed higher.

Independence: Again, based on the residuals vs. fitted values we can conclude that the independence assumption probably does not hold for this data. Although this assumption is much more difficult to really check, we can see that there appears to be some correlation in the fitted values based on the way that they are all clumped together in the first third of the plot area.

7. What can you infer about the reliability of this model (from question 5)? What can you conclude about flight delays at the Atlanta airport based on this information?

   Based on the R-squared values, we can see that this model does not do a very good job in explaining the variability in the data.
   Based on this knowledge, we can probably say that flight delays at the Atlanta airport cannot be easily predicted given the data we have. Perhaps there are other factors or variables at play that were not included in this dataset.

8. What are your suggestions or curiosities about other potential ways to model flight delays at the Atlanta airport?

   Pretty open ended question, a few possible answers include:
   Explore a logistic regression to check whether or not a flight will be delayed
   Research other types of regression (gamma family, etc) to try to better explain the data
   Include other variables (such as weather/precipitation for ATL and destination airport)

9. Multiple Choice Questions (using the model resulting from question 5):
   - **a.** Which of the following variables is statistically significant at the 99% confidence level?
     i. AIRLINE=MQ
     ii. AIRLINE=UA
     iii. AIRLINE=DL
     iv. AIRLINE=F9
   - **b.** What is the p-value for the estimated regression coefficient for AIRLINE=OO?
     i. 0. 011613
     ii. 0.167571
     iii. 0.016122
     iv. 0.016564
   - **c.** What is the 99% confidence interval for the estimated regression coefficient for day of the week?
     i. (-4.650e-01 1.003e+01)
     ii. **(-9.302243e-01, -0.007901542)**
     iii. (-4.686e-01 -1.790e-01)
     iv. (-9.302243e-01, 0.007901542)

d. What is the sum of squared error for this model?

i. 171635

ii. 4387276

iii. 5123

iv. 427188000

e. What percentage of variation in the response is explained by the predicting variables used?

i. 19%

ii. 9%

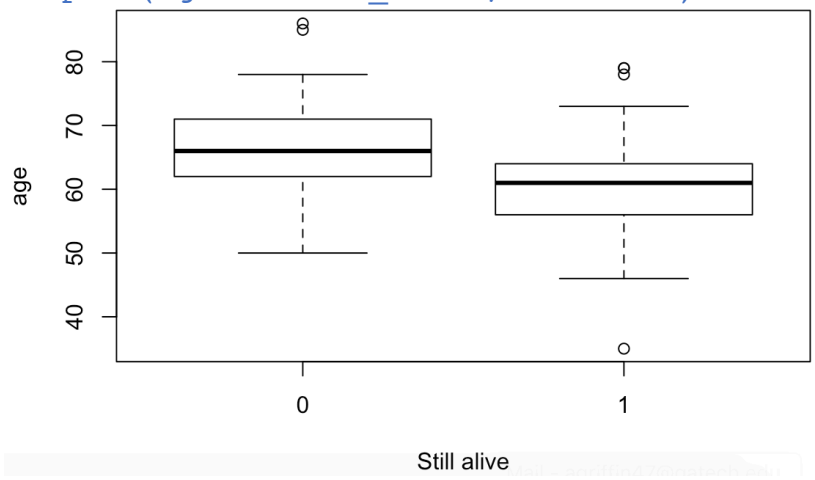
iii. 10%

iv. 12%

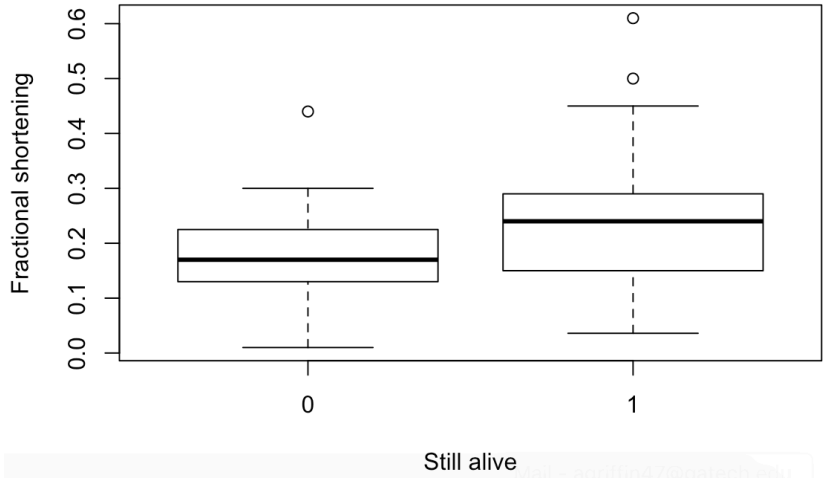
Question 2. Logistic regression

1. Plot the boxplots to describe the relationship between the response (still_alive) and the following predicting variables: age, survival, fractional_shortening, and lvdd. **Provide the R code and plots generated.**

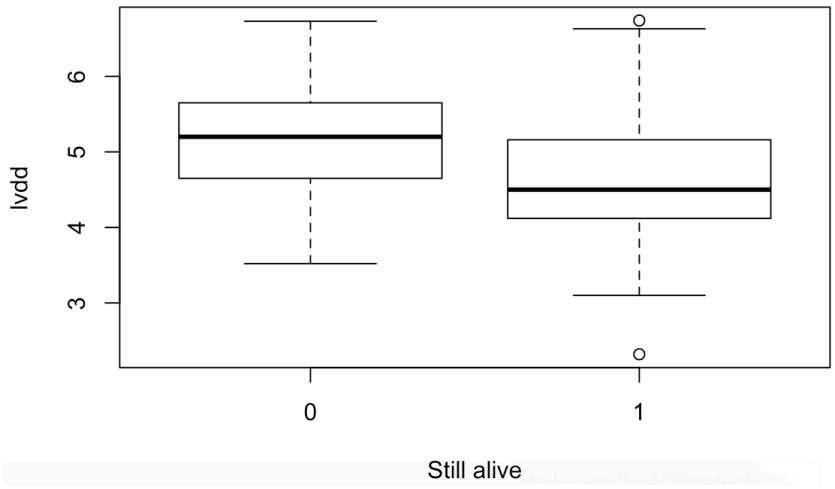
```
boxplot(age ~ still_alive,data=echo)
```



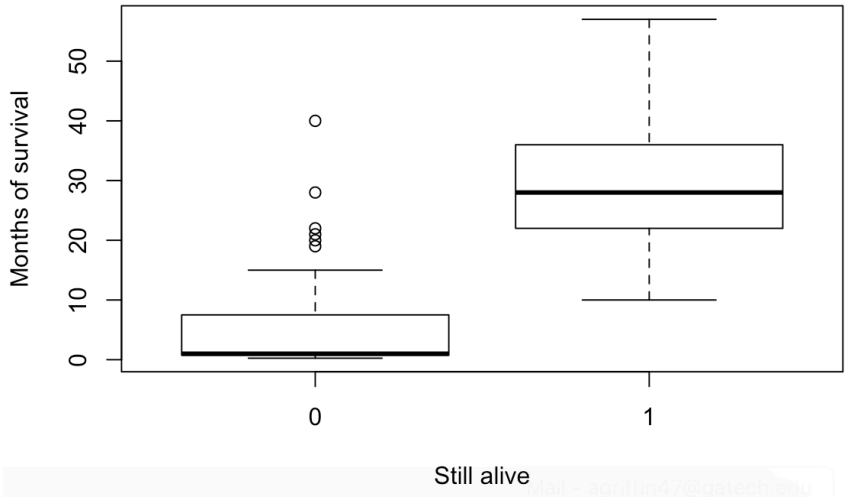

```
boxplot(fractional_shortening ~ still_alive,data=echo)
```




```
boxplot(lvdd ~ still_alive,data=echo)
```




```
boxplot(survival ~ still_alive, data=echo)
```



We would normally make boxplots the other way (Switch the order of the variables in the R code), but I realize it wasn't specified here

2. Create a model using the following predicting variables: age, fractional_shortening, and lvdd (exclude survival due to potential multicollinearity with the response variable, 'still_alive'). **What are the model parameters and what are their estimates? Which of them are statistically significant at the 95% confidence level?**

   Model parameters and their estimates:

```
(Intercept)             8.54944
age                    -0.08984
fractional_shortening   5.90295
lvdd                   -0.65886
```

   Model parameters that are statistically significant at the 95% confidence level include: intercept, age, fractional_shortening, and lvdd

3. **Interpret the estimated value of the parameter corresponding to age in the context of this problem.**

   For a one unit increase in the age of the patient, the log-odds of the patient still being alive decrease by 0.08984, holding all other variables constant. In the same way, for a one unit increase in the age of the patient, the odds of the patient still being alive increase by e^-0.08984 = 0.91407742596, holding all other variables constant.

   Be careful with your interpretation of negative parameters - if you use the word "decrease", you shouldn't include the negative sign with the coefficient

   Don't forget to include the phrase "holding all other variables constant"!

4. **Provide the equation for the estimated logit transformation of the probability of still being alive given the predicting variables.**

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 8.54944 - 0.08984X_{age,\,i} + 5.90295X_{fractional-shortening,\,i} - 0.65886X_{lvdd,i}$$

5. Multiple Choice questions:
   a. How does the log-odds change with a 1 unit increase in fractional_shortening, holding all other variables constant?
      i. The log-odds increase by 5.90295
      ii. The log-odds decrease by -5.90295
      iii. The log-odds increase by 366.1159
      iv. The log-odds decrease by 0.0027

   b. What is the null deviance and its degree of freedom?
      i. 107.12 on 106 degrees of freedom
      ii. 132.21 on 106 degrees of freedom
      iii. 107.12 on 103 degrees of freedom
      iv. 132.21 on 103 degrees of freedom

   c. What is the approximated distribution of residual deviance?
      i. Normal
      ii. Chi-squared
      iii. Binomial
      iv. Exponential
      v. Bernoulli

   d. Fill in the blanks on the following sentence: A patient with older age tend to have (a) ____ probability of still being alive and a patient with higher fractional_shortening tend to have (b)____ probability of still being alive.
      i. (a) lower & (b) lower

ii. (a) higher & (b) lower
iii. (a) lower & (b) higher
iv. (a) higher & (b) higher

e. What is the AIC for your model?
i. 312.09
ii. 65.84
iii. 98.43
iv. 115.12