Name:
Quiz 4, Stat 431 (Summer 2012)
August 2, 2012

You have 20 minutes to answer these questions.

In these questions, we want to study whether flights out of San Diego by United Airlines are generally delayed or not. Specifically, we want to predict the delay of the flight, a continuous number between negative infinity to positive infinity, based on its destination, a categorical variable, and time of day of the schedule flight, also a categorical variable; a negative delay means that the plane left earlier than the scheduled time while a positive delay means that the plane left later than the scheduled time. Delays are measured in minutes and a numerical summary of the variables are shown in tables 1 and table 2.

	IAD (D.C.)	DEN (Denver)	HNL (Hawaii)	ORD (Chicago)	SFO (San Fran)
Mean Delay Time	5.949749	8.463768	-2.5	9.46473	12.10373

Table 1: Table of mean delay times according to different destinations

	6AM to 9AM	9AM to Noon	Noon to 6PM	6PM to 9PM	9PM to Midnight
Mean Delay Time	3.135632	19.12551	13.71196	12.09434	-1.494845

Table 2: Table of mean delay times according to different time of the day

The mean delay is 9.691667 minutes with a standard deviation of 32.23845. There are 1200 flights under consideration.

1. Suppose we decided to only use the time of the day to predict delays. We run the regression and obtain the following output.

```
Call:
lm(formula = DepDelay ~ SchTime)

Residuals:
    Min      1Q  Median      3Q     Max
-30.126 -17.126  -8.136  -0.136 225.874

Coefficients:
                     Estimate Std. Error
(Intercept)             ?      1.511
SchTime6PMto9PM         ?      4.585
SchTime9AMtoNoon        ?      2.511
SchTime9PMtoMidnight    ?      3.539
SchTimeNoonto6PM        ?      2.232
---
```

```
Residual standard error: ? on ? degrees of freedom
Multiple R-squared: ?,     Adjusted R-squared: 0.04421
F-statistic: 14.87 on ? and ? DF,  p-value: 7.334e-12
```

Fill out the missing values in the R output above. Make sure you clearly label what values are you deriving below.

2. Fill out the following ANOVA table based on the above output.

SS	MSE	DF
SSE =	MSE =	DFE =
SSR =	MSR =	DFR =
SST =	MST =	DFT =

Table 3: ANOVA table for the regression

3. We want to include the destination into the model we built in (1) and we want to know whether destination is a factor in explaining the variation in the delay times. Conduct a hypothesis test, specify the full and the reduced model, and the degrees of freedom resulting from the F test. If the p-value for this test is 0.7636, do we reject the null at $\alpha = 0.05$?

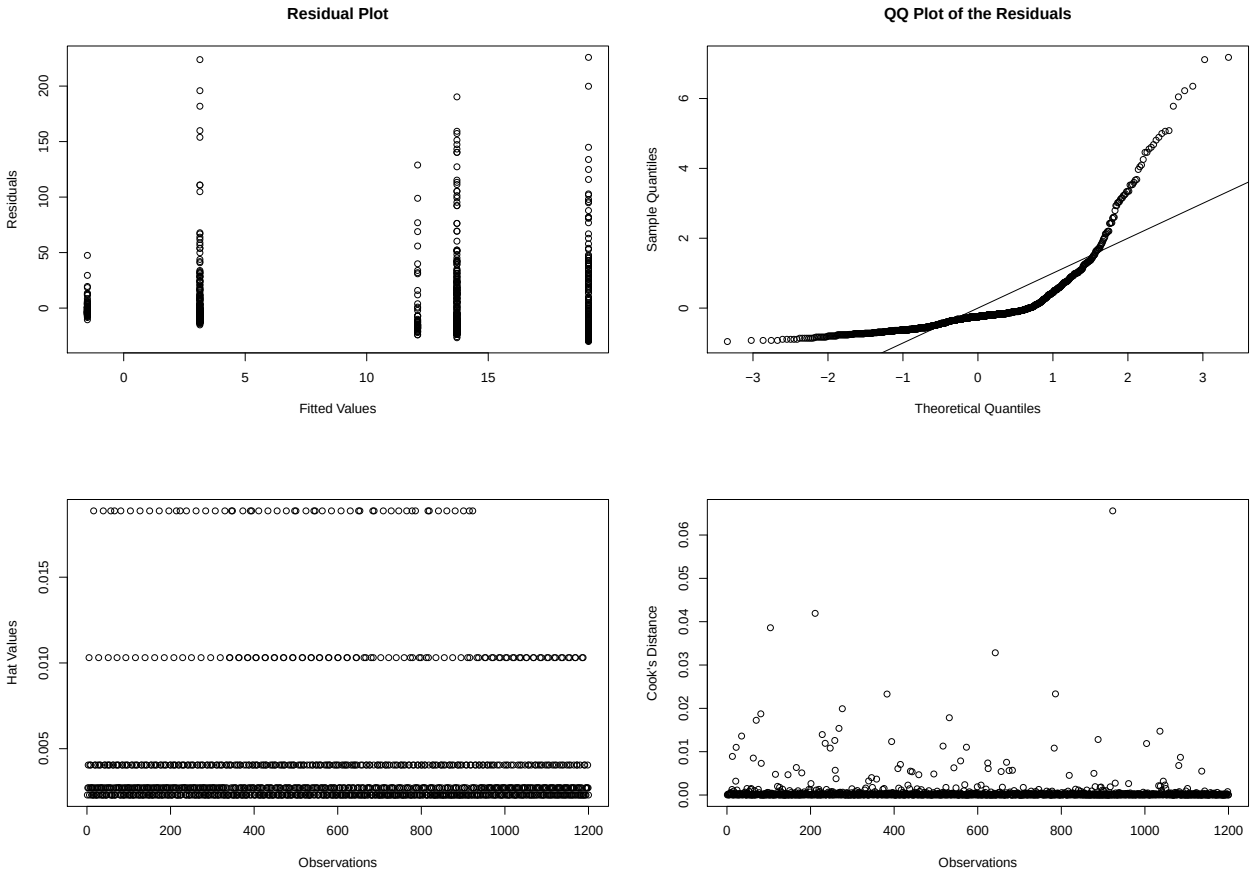

Figure 1: Diagnostic plot for model we built in question (1)

4. Consider the diagnostic plots (Figure 4) for the model we built in (1). Should we worry about homoscedasticity, linearity, or normality? If so, what transformations should we use and on which variable should we apply those transformations? How about outliers? Circle, if any, possible outliers in the data.