Jordan Rosenblum (jr3478)
COMS W4721 Homework 1 (cont'd)

Problem 3:

Part 1:

a) The signs on most of the coefficients below make intuitive sense. As number of cylinders, horsepower, and weight of the vehicle increase, miles per gallon are expected to decrease. On the other hand, higher displacement, acceleration, and a newer model year all contribute to higher miles per gallon. Higher displacement and acceleration leading to higher miles per gallon seems a bit strange but note that the coefficients are fairly close to zero (it may just be noise).

	Coefficients
intercept	23.4212116
number.of.cylinders	-0.3292773
displacement	0.6052419
horsepower	-0.2005265
weight	-5.7265694
acceleration	0.2068464
model.year	2.7908892

b) After repeating the process of splitting the data into training and testing sets 1000 times, I obtain the following mean and standard deviation of the mean absolute errors of each of the 1000 testing sets of size 20:

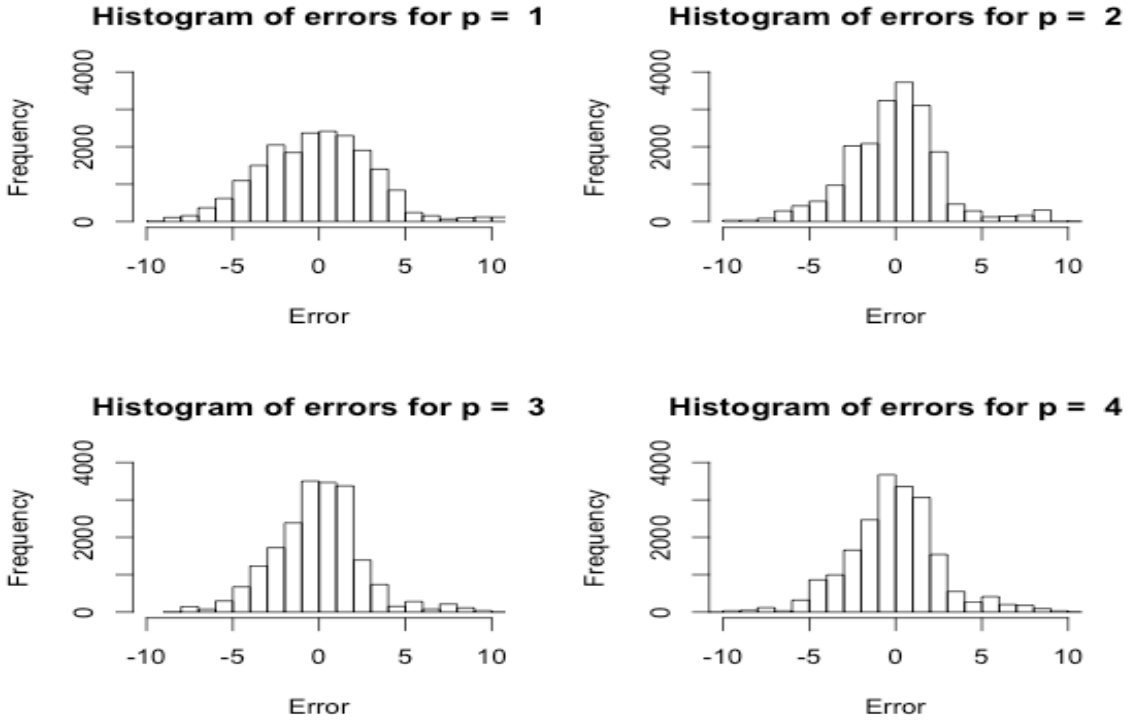
```
Mean MAE for p =  1 is:  2.675308
St. Dev. MAE for p =  1 is:  0.4781826
```

Part 2:

a) Judging by the mean and standard deviations of the root mean square error values below for each value of p, the least squares regression for p = 3 is best given that both the mean and standard deviation of the RMSE is lowest. Intuitively, this means that the model prediction for the 20 testing examples is closest to the actual value of miles per gallon and the variation across the 1000 training/testing splits is low. In reality, however, there are other factors to consider when selecting the best model (e.g. model complexity).

	Mean RMSE	St. Dev. RMSE
p = 1	3.415624	0.6765065
p = 2	2.754872	0.6254105
p = 3	2.655812	0.5942852
p = 4	2.700955	0.6184315

b) The 4 histograms below for each value of p show the distribution of errors of predicted versus actual miles per gallon in the testing sets. For p = 1, for example, there are a number of observations which are close to a zero error, but the tails of the distribution are relatively fat. As we move up to the histogram of p = 3, the distribution looks more normal with narrower tails and the bulk of the observations are closer to an error of zero, agreeing with the conclusion in part (a) above.



c) The maximum likelihood values for the mean and variance of the errors follow directly from the slides in class except it is a univariate case. Specifically,

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad and \quad \hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{ML})^2 \, , where\ x_i's\ are\ the\ errors$$

I then plugged these values (i.e. $x_i's$, $\hat{\mu}_{ML}, and\ \hat{\sigma}^2_{ML}$) into the Gaussian distribution function (i.e. pdf of the normal distribution), took the log, and summed over all x_i's to obtain the log-likelihood for each value of p (as seen in the table below). A likelihood value conveys how likely we are to see the data given our model. Hence, we want the

largest likelihood values. In the case of log-likelihood, we want the smallest absolute value. Therefore, setting p = 3 has the least negative log-likelihood value, and again looks like the most accurate model (agreeing with the conclusion in part (a) formed based on the lowest mean and standard deviation of the RMSE). This conclusion assumes that the errors follow a Gaussian distribution. Also, it assumes that there are negligible errors in the explanatory variables, the errors are independent and identically distributed, and the errors are independent of the explanatory variables.

	Log-Likelihood
p = 1	-53329.91
p = 2	-49147.52
p = 3	-48401.87
p = 4	-48761.29