Car Talk

A website reports:

> "There are 107 million U.S. households, each with an average of 1.9 cars, trucks or sport utility vehicles and 1.8 drivers, the Bureau of Transportation Statistics reported. That equals 204 million vehicles and 191 million drivers."

That's a lot of cars, not all waiting to turn left in front of you.

This project looks at gas mileage of cars. There are three data sets, two that accompany `R`, one that comes from the EPA (`http://www.fueleconomy.gov`).

Data from 1974 is contained in the built in data set `mtcars` which is loaded with the command

```
> data(mtcars)
```

Data on 93 cars manufactured in 1993 is available in the `MASS` package. The data was originally contributed by Robin Lock to the *Journal of Statistical Education*. The data is loaded into an `R` session with the commands:

```
> library(MASS)
> data(Cars93)
```

A data set for cars manufactured in 2004 is available from the Stem and Tendril website. To load the data into the variable `Cars04` issue the following command:

```
> Cars04 = read.csv("http://www.math.csi.cuny.edu/st/R/epa04.csv")
```

We will `attach()` the variable names for the `Cars93` data set, when using the other data sets, we will reference the variables using the dollar sign notation, or the `data=` argument.

```
> attach(Cars93)
> names(Cars93)
```

```
 [1] "Manufacturer"       "Model"              "Type"
 [4] "Min.Price"          "Price"              "Max.Price"
 [7] "MPG.city"           "MPG.highway"        "AirBags"
[10] "DriveTrain"         "Cylinders"          "EngineSize"
[13] "Horsepower"         "RPM"                "Rev.per.mile"
[16] "Man.trans.avail"    "Fuel.tank.capacity" "Passengers"
[19] "Length"             "Wheelbase"          "Width"
[22] "Turn.circle"        "Rear.seat.room"     "Luggage.room"
[25] "Weight"             "Origin"             "Make"
```

The different variable names are returned with the `names()` function. Most are self-explanatory, however if the variable is not clear more information can be found on the help page for the data set, which can be viewed with the command `?Cars93`.

1 Miles per gallon

An advertisement for a BMW Mini Cooper claims the car "Drinks Responsibly." A reference to its estimated 27/37 gas mileage. These numbers refer to the city and highway mileage. Cars get better gas mileage on the highway as they are not wasting gas stopping and starting, but rather are usually near the most efficient operating conditions.

The EPA reports values for city and highway mileage. Highway mileage is calculated by mixture of "non-city" driving on different kinds of rural roads and interstate highways. The test simulates a 10-mile trip and averages 48 mph. The maximum speed is 60 mph. The test is run with the engine warmed up and has little idling time and no stops (except at the end of the test). This estimate under laboratory conditions is lowered by 22% to adjust to real-world conditions.

 Question 1: To compute city estimates, the EPA simulates an 11-mile, stop-and-go trip with an average speed of 20 miles per hour (mph). The trip takes 31 minutes and has 23 stops. About 18 percent of the time is spent idling, as in waiting at traffic lights or in rush hour traffic. The maximum speed is 56 mph. The engine is initially started after being parked overnight. Vehicles are tested at 68 F to 86 F ambient temperature. The estimate under laboratory conditions is lowered by 10%.

Does this seem like a reasonable way to estimate city mileage? Why do you think the estimate is lowered 10% for city driving, but 22% for highway driving? Which estimate would you expect to have more variability?

2 Bivariate Summaries

Looking for relationships between the variables in a data set can be greatly facilitated using graphical displays and tables.

2.1 Categorical predictors of gasoline mileage

The actual mileage of a car depends on many different variables. To investigate the dependence on a single categorical variable or factor, boxplots can be used to compare distributions, and individual summaries broken up by the levels of the factor.

To make boxplots for the mileage variable for different values of `Type` can be done with syntax such as

```
> boxplot(MPG.highway[Type == "Small"])
```

It is more efficient to use the model notation if you wish to create several boxplots at once. For example, to produce boxplots of highway mileage broken up by `Type` is done with

```
> boxplot(MPG.highway ~ Type, main = "Highway mileage")
```

Figure 1 shows that there is great variation between the types of cars. Only the "small" cars are greatly skewed, even though overall the distribution of highway mileage shows skew.

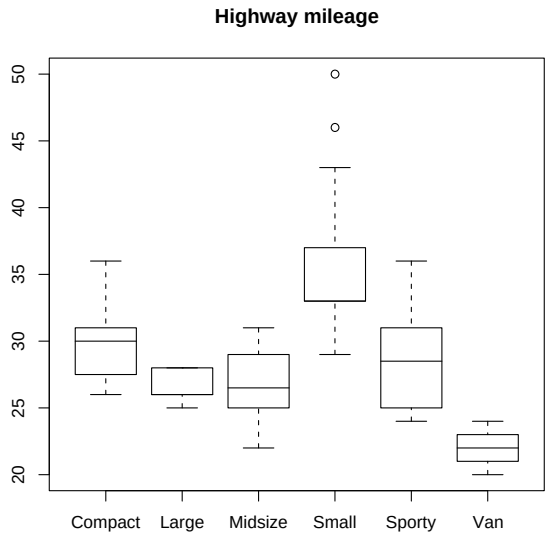

Figure 1: Boxplots of highway mileage by type of vehicle

Question 2: What class of cars is missing if the data were from 2004? Describe where it would fit in.

Question 3: Produce boxplots of city mileage broken up by `Type`. Are the relationships similar to those shown in Figure 1?

Question 4: Produce boxplots of highway mileage broken up by `Cylinders`, the number of cylinders. Is there a relationship between the two variables? Explain.

Question 5: Create boxplots of highway mileage broken up by `Origin`, a factor indicating if the car is manufactured in the United States. Does there appear to be a difference in the centers of the distributions? The shapes?

Question 6: If all cars were filled to capacity, the variable `mpg.highway*Passengers` would record the passenger miles per gallon. Make boxplots of this variable broken up by `Type`. Explain any differences between the types of cars. Are any similarities surprising?

Question 7: The variables `cty` and `hwy` in the `Cars04` data set contain the city and highway estimates computed by the EPA. For `hwy` produce boxplots broken up by `Class`. Which class has the "best" mileage? Which has the worst? Can you tell how many vehicles are in which class?

2.2 Numeric predictors of gasoline mileage

A scatterplot can show relationships between numeric variables. For example, Figure 2 shows how ciy and highway mileage are related.

```
> plot(MPG.highway ~ MPG.city)
> abline(0, 1)
```

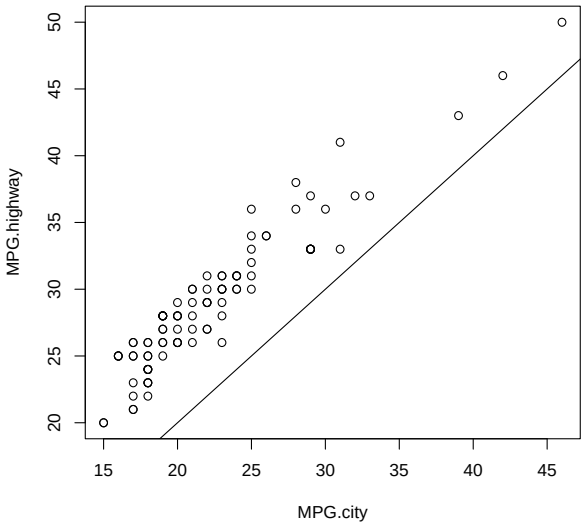

Figure 2: Scatterplot of city and highway mileage

We see in this case, that all the data points lie above the line $y = x$, which was added to the figure using `abline()`. This is because highway mileage is better than city mileage for all of the cars in the data set `Cars93`.

Question 8: Make a scatterplot of `cty` versus `hwy` for the `Cars04` data set. Are all the points above the line $y = x$? Explain.

Linear regression line

The reference line $y = x$ in Figure 2 suggests that a straight line summarizes well the relationship between the two variables. But which straight line? The regression line uses the method of least squares to summarize a linear relationship between variables. The `lm()` function returns the coefficients of the regression line. The `abline()` function will plot them.

```
> res = lm(MPG.highway ~ MPG.city, data = Cars93)
> res

Call:
lm(formula = MPG.highway ~ MPG.city, data = Cars93)

Coefficients:
(Intercept)     MPG.city
     9.0566       0.8955
```

```
> plot(MPG.highway ~ MPG.city)
> abline(res)
```

From the output, we see that the coefficients are a slope of 0.89 and an intercept of 9.05.

The argument `data=Cars93` was added, redundantly in this case, to show how to temporarily attach the data set so that the variable names in the model formula can be found within the data set.

Question 9: For the `Cars04` data set, find the regression coefficients and compare to those found using the `Cars93` data set. Have there been any improvements?

What other variables have a relationship with mileage? Certainly we expect that heavier cars have worse mileage as it take more energy to get them up to speed. Can we see this in the data? A scatterplot (Figure 3) can be produced as follows:

```
> plot(MPG.highway ~ Weight, main = "Highway mileage by weight",
+     data = Cars93)
```

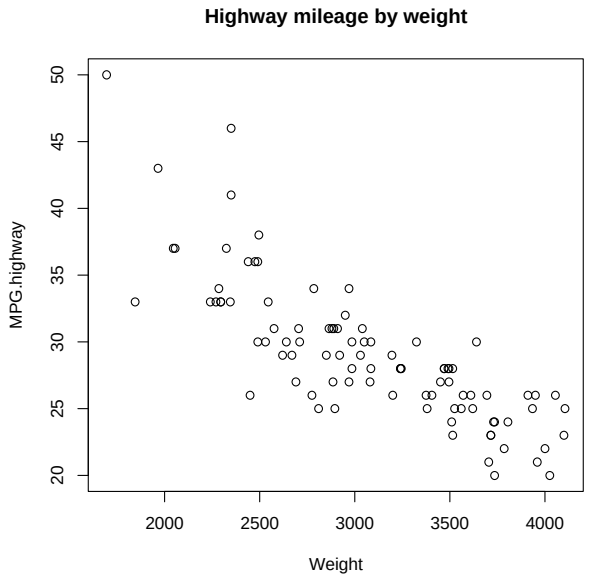

Figure 3: Highway mileage by vehicle weight

The relationship doesn't quite appear to be linear, but we try a linear model nonetheless. The regression coefficients are

```
> res = lm(MPG.highway ~ Weight, data = Cars93)
> res

Call:
lm(formula = MPG.highway ~ Weight, data = Cars93)
```

```
Coefficients:
(Intercept)       Weight
  51.601365    -0.007327
```

Question 10: Produce a scatterplot and regression line of the relationship between highway mileage and engine size in liters (`EngineSize`) for the `Cars93` data set. Explain why there should be some relationship between the two variables. Based on your graphic, does a linear relationship seem appropriate? Why?

Question 11: Produce a scatterplot and regression line of the relationship between highway mileage and `Price` for the `Cars93` data set. Explain why there might be some relationship between the two variables. Based on your graphic, does a linear relationship seem appropriate? Why?

If you fit the linear model explain why a statement like "Even though our cars are identical, mine gets better mileage as I spent $2,000 dollars less for mine." could possibly make sense. Does it really make sense?

Question 12: The variable `Min.Price` contains the price of the basic version of the vehicle, and `Max.Price` contains the price of the fully equipped version. Make a scatterplot of `Max.Price` versus `Min.Price`.

Is the relationship linear? Why would you expect it to be? Is there more variation between the two values for lower priced cars or for higher priced cars? Why might this be?

Numeric summaries of linear relationships

The regression line gives both a numeric and graphical summary of the linear relationship between two numeric variables. The Pearson correlation coefficient describes the strength of the linear relationship. The Spearman correlation coefficient can be used to describe the strength of a increasing or decreasing relationship.

Both correlations are found with `cor()`.

```
> cor(MPG.highway, Weight)
[1] -0.810658
> cor(MPG.highway, Weight, method = "spearman")
[1] -0.8381786
```

Both values should a negatively correlated relationship, as is expected from the scatterplot.

Question 13: The Spearman correlation coefficient uses the data after ranking. The value reflects if the data has a monotonic relationship, not just linear. Compare the value of the Pearson and Spearman correlation coefficients for highway mileage versus new car price (`Price`) in the `Cars93` data set. Are the two values similar in size? Explain why or why not, based on the scatterplot?

Predictions based on the regression line

We can use the regression line to make predictions for future observations. For example, what mileage would we expect a 3000 pound vehicle to get. The equation for a line can be used, or the `predict()` function can do the work.

To use the equation of the line, just plug in the weight value into the equation. The `predict()` function is used as follows:

```
> res = lm(MPG.highway ~ Weight, data = Cars93)
> predict(res, newdata = data.frame(Weight = 3000))

[1] 29.62019
```

The value of the argument `newdata=` is a data frame with column names matching the variable names used to produce the model object `res`.

 Question 14: Use the 1993 data to make a prediction for the highway mileage of 2004 Mini Cooper S which weighs 2,678 pounds. Compare to that given in the 2004 data:

```
> subset(Cars04, select = hwy, subset = carline.name == "MINI COOPER S")

   hwy
77  34
```

 Question 15: The 5.4 liter 2004 Ford Expedition with 8 cylinders weighs 5,671 pounds. This is beyond the range of weights in the 1993 car data, so any prediction based on this data should be considered suspicious. Nonetheless, use the 1993 data to predict the highway mileage of the a 2004 Ford Expedition. Compare to that given by

```
> subset(Cars04, select = c(Manufacturer, hwy), subset = carline.name ==
+     "EXPEDITION 4WD" & displ == 5.4)

     Manufacturer hwy
1033         FORD  17
```

If gas costs 2 dollars per gallon, how many dollars would it cost to drive the 2,780 highway miles from Los Angeles to New York City?

 Question 16: The 2004 HUMMER H2 weighs 8,600 pounds. A quirk in United States law allows HUMMER to sell this car without needing it to get evaluated by the EPA as it weighs more than 8,500 pounds. Despite the fact that 8,500 pounds is well beyond the maximum value of 4105 in the 1993 data, use the data to predict the highway mileage of a HUMMER H2. Does your answer make any sense? What would you guess the highway mileage of a H2 would be?

3 More regression analysis

3.1 resistant regression

Many statistics, such as the mean and standard deviation, can be greatly affected by just a few outlying data values. These outliers can come from long-tailed distributions, or may be errors in data collection. Resistant techniques have been developed that allow for some percentage of outlying values without dramatically affecting the value of the statistic. For example, the trimmed mean first cuts off the largest and smallest values in a data set, then finds the mean. A few arbitrarily large or small values will not affect the value as these will be cut off before the mean is found.

Linear regression is also susceptible to large outlier values. A method, least trimmed squares, can be used to find the estimated regression parameters. Rather than minimize all the squared residuals, only a percentage of the smallest ones are used. Historically, this was computationally prohibitive, but now this is not the case. The `lqs()` function will use this algorithm to find estimated parameters and is used just as `lm()` is.

For example, to visualize the difference between the two estimates for the model of highway mileage by vehicle weight we plot both regression lines in Figure 4 with the following commands:

```
> f = MPG.highway ~ Weight
> plot(f, data=Cars93)
> abline(lm( f, data=Cars93))
> abline(lqs(f, data=Cars93), lty=2)
```

We can use the resistant regression model to make predictions in the same manner as the simple regression model. For example, the predicted highway mileage of a 5,761 pound Ford Expedition is found with

```
> res = lqs(MPG.highway ~ Weight, data=Cars93)
> predict(res, newdata=data.frame(Weight=5761))
```

The same caveat about not predicting values beyond the range of the data applies.

 Question 17: Predict the highway mileage for a 2,678 pound Mini Cooper using the resistant regression line.

 Question 18: Compare the least squares regression line with the least trimmed squares one for the model of highway mileage modeled by `EngineSize`. Can you identify which residual values are being ignored by the trimming?

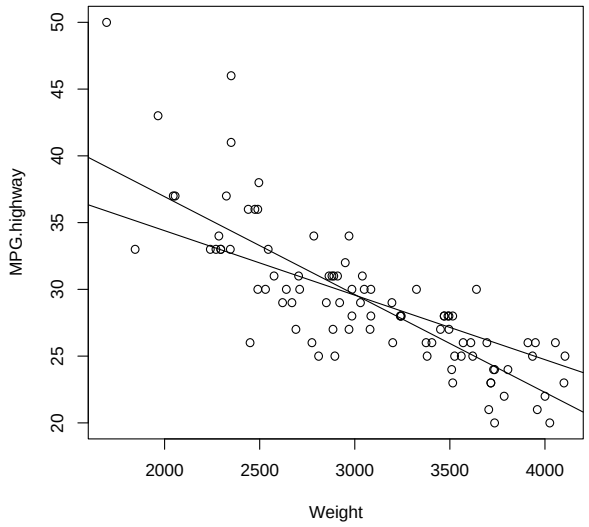

Figure 4: Highway mileage vs. vehicle weight with both least squares and least trimmed squares (dashed line) regression lines added