Auto Dataset Linear Regression

Michelle Bark

Descriptive Analysis This section involves the `Auto` dataset included in the "ISLR" package.

```
library(ISLR)
fix(Auto)
```

```
head(Auto)
```

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                       name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

```
str(Auto)
```

```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 1
```

```
attach(Auto)
cylinders = as.factor(cylinders)
year = as.factor(year)
origin = as.factor(origin)
```

The range of each quantitative predictor is:

```
sapply(Auto[,-c(2,7,8,9)],range)
```

```
##       mpg displacement horsepower weight acceleration
## [1,]  9.0           68         46   1613          8.0
## [2,] 46.6          455        230   5140         24.8
```

The median and variance of each quantitative predictor is:

```
sapply(Auto[,-c(2,7,8,9)],median)
```

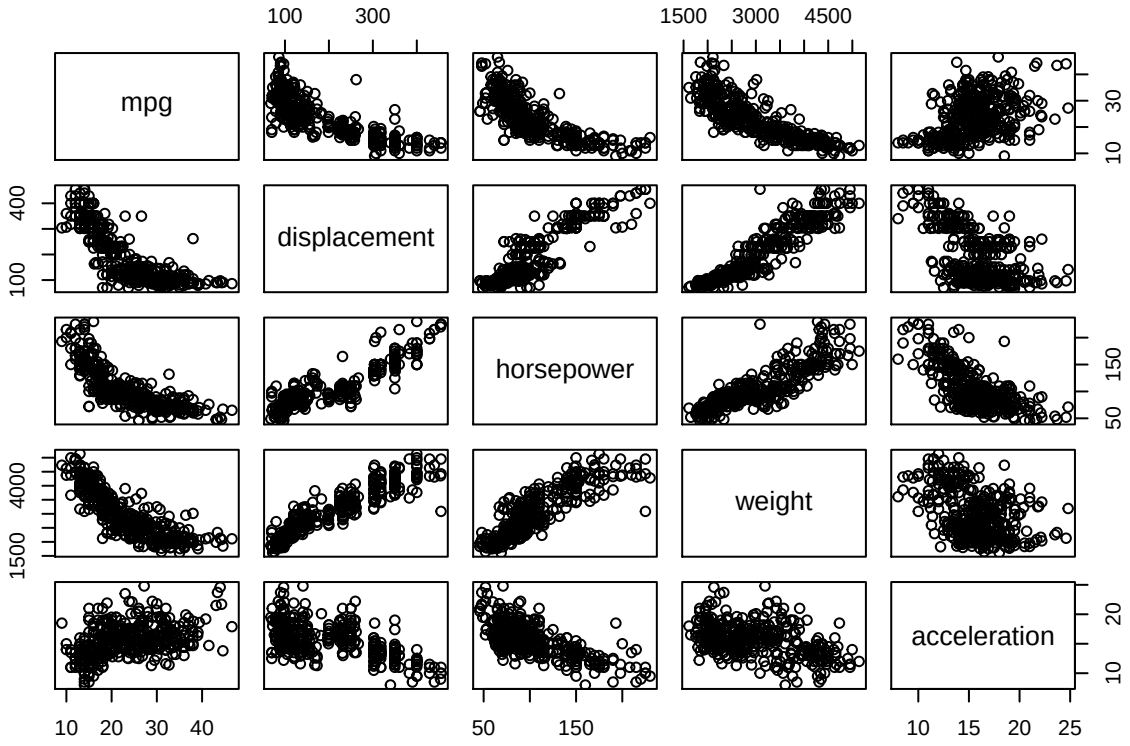
```
##          mpg displacement   horsepower       weight acceleration
##        22.75       151.00        93.50      2803.50        15.50
```

```
round(sapply(Auto[,-c(2,7,8,9)],var),3)
```

```
##          mpg displacement   horsepower       weight acceleration
##       60.918    10950.368     1481.569   721484.709        7.611
```

Using the full data set, investigating the relationship between individual predictors with the target response gas mileage (mpg) graphically.

```
pairs(~mpg+displacement+horsepower+weight+acceleration,Auto)
```

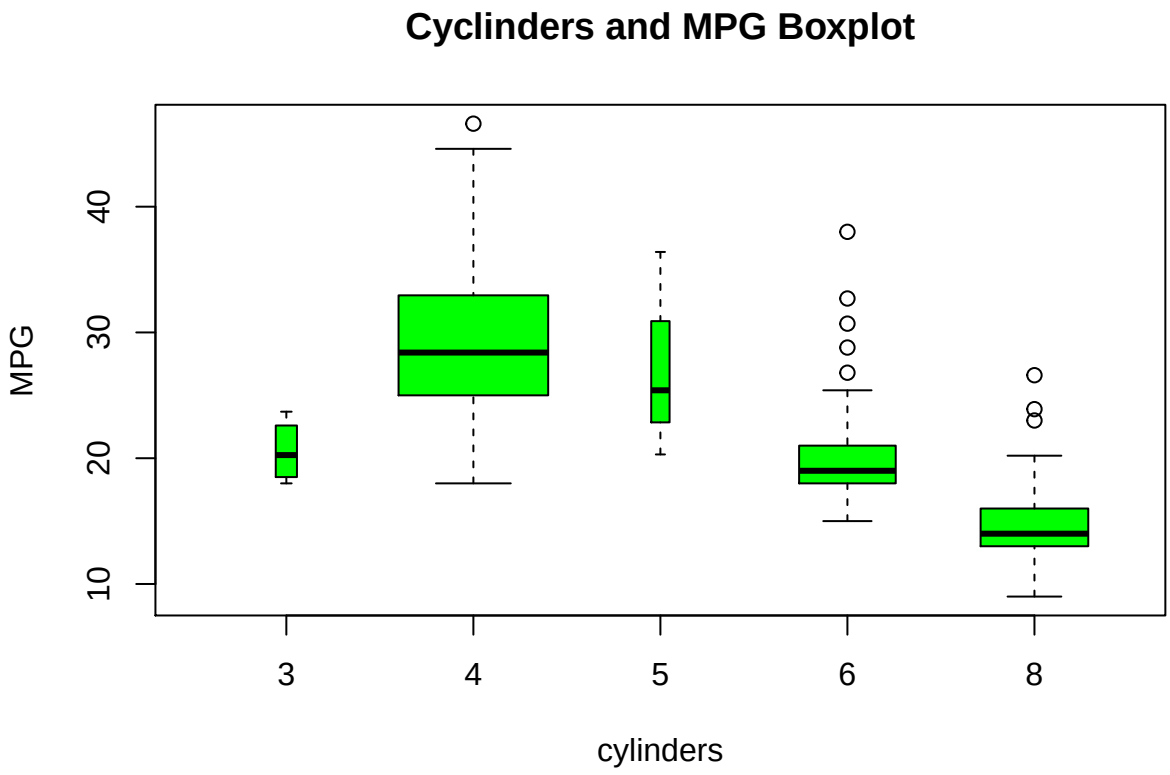


Using the pairs() function to begin identifying the relationship between mpg and the other quantitative variables. Using the top row and the first column, there seems to be a negative relationship between mpg and the 3 predictors displacement, horsepower and weight. There is possibly a positive relationship between mpg and acceleration, but the data is very clustered so it is hard to see using this particular graphical summary. In terms of shape of these relationships, the scatter plots suggest a slight curve (rather than a true linear relationship) for displacement, horsepower and weight.

Using boxplots to review the qualitative predictors and mpg:

Cylinders and MPG

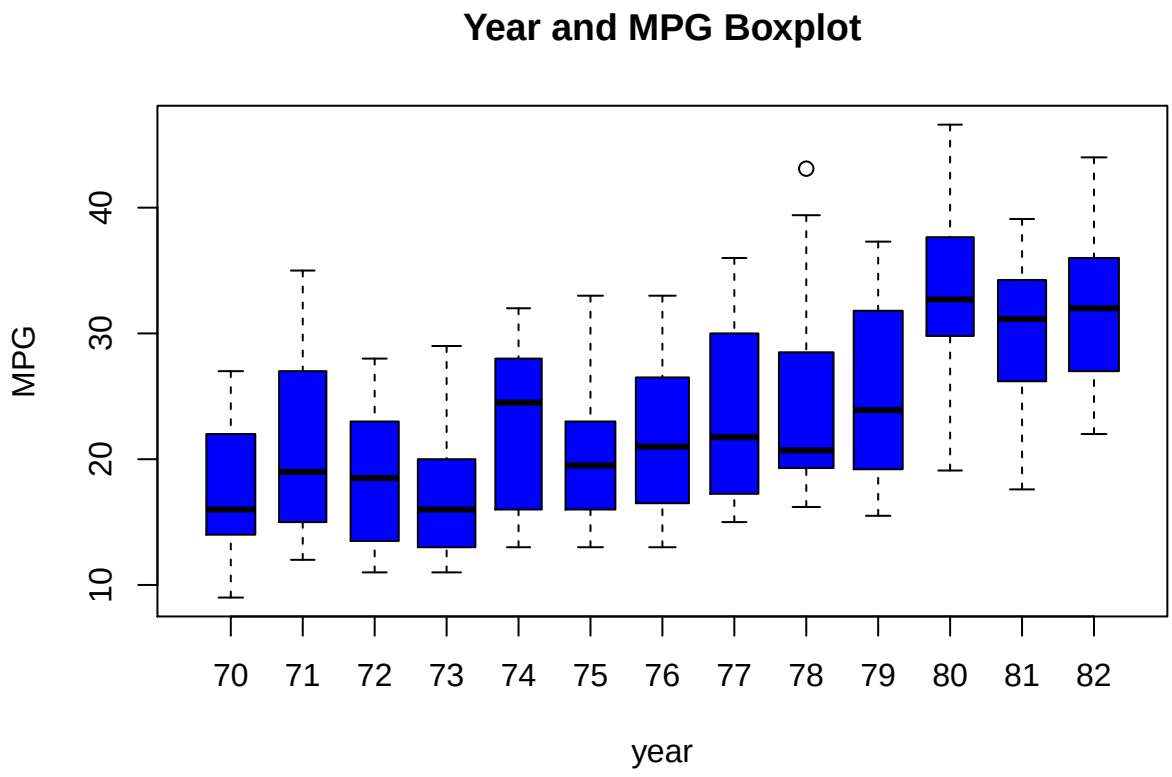
```
plot(cylinders,mpg,
     col= "green",
     xlab = "cylinders",
     ylab = "MPG",
     varwidth = T,
     main = "Cyclinders and MPG Boxplot")
```



This boxplot for number of cylinders and MPG shows that 4 cylinders seems to be optimum for the most miles per gallon. However, utilising the varwidth argument in plot(), it shows that there are a greater number observations in the data for cars that have 4 cylinders and much fewer for 3 and 5 cylinders, so it is not clear whether this is a result of lack of data for cars with 3 or 5 cylinders.

Year and MPG

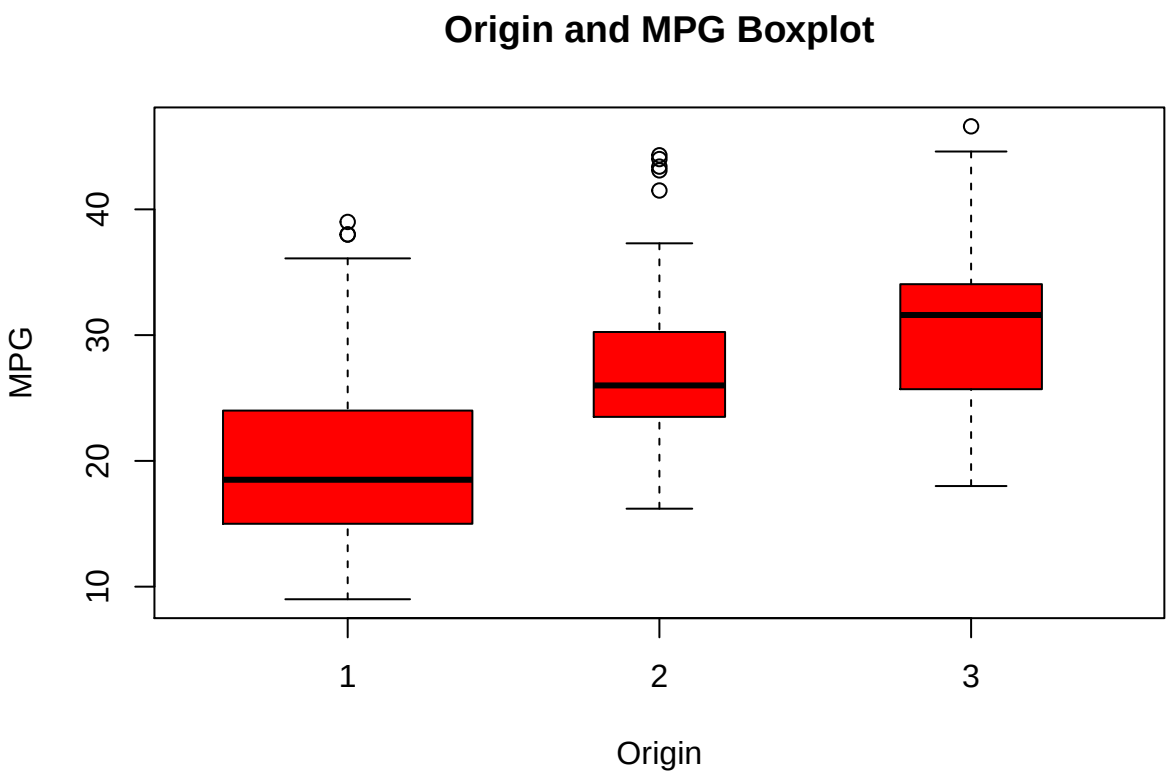
```
plot(year,mpg,
     col= "blue",
     xlab = "year",
     ylab = "MPG",
     varwidth = T,
     main = "Year and MPG Boxplot")
```



There seems to be a general trend of MPG increasing as the year increases, possibly suggesting cars became more fuel efficient in terms of more miles per gallon, over time. The interquartile range and mean moves upwards from 1970 with the mean of approximately 16 MPG to a mean of over 30 mpg for 1982, with a noticeable jump between '79 and '80.

Origin and MPG:

```
plot(origin,mpg,
     col= "red",
     xlab = "Origin",
     ylab = "MPG",
     varwidth = T,
     main = "Origin and MPG Boxplot")
```



These box plots show the origin of the vehicle plotted against MPG. Origin 3 appears to have the highest MPG, with orgin 1 the lowest, although this is a view in isolation. Considering all the previous graphical summaries, it may be that origin 3 has more cars produced with 4 cylinders and/or more produced in the later years of the observations. These summaries alone are not enough to conclude anything further. The same applies to the other predictors, when looking in isolation. There may be colinearity in the predictors.

Predicting `mpg` on the basis of the other variables:

Some of the other variables in the Auto dataset could be used to predict MPG. A combination of the other predictors would likely give the best prediction of MPG to limit bias in the prediction and create a more accurate model. As mentioned in the origin boxplots, some predictors may also be enhanced by the inclusion of other predictors. It is reasonable to conclude from this dataset, that some factors that may determine mpg work side by side to deliver that result.

A multiple regression model may therefore give a better result than simple regression on each predictor. The relationships however do not appear to be strictly linear, as shown in the scatter plots, so choice of a more flexible model than linear regression may be optimum for predicting mpg using the other variables. As the goal is prediction, not inference, the model should not be too flexible to maintain interpretability of those predictors.

Linear Regression The use of simple linear regression on the `Auto` dataset.

```
library(ISLR)
attach(Auto)
```

Using the `lm()` function to perform a simple linear regression with `mpg` as the response and `acceleration` as the predictor. The `summary()` function prints the results.

```
lm.fit = lm(mpg~acceleration)
print(summary(lm.fit))
```

```
##
## Call:
## lm(formula = mpg ~ acceleration)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.989  -5.616  -1.199   4.801  23.239
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.8332     2.0485   2.359   0.0188 *
## acceleration   1.1976     0.1298   9.228   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.08 on 390 degrees of freedom
## Multiple R-squared:  0.1792, Adjusted R-squared:  0.1771
## F-statistic: 85.15 on 1 and 390 DF,  p-value: < 2.2e-16
```

T-value is high and the p value is very low. This suggests there is a linear relationship between mpg and acceleration and the null hypothesis is can be rejected.

The adjusted R-squared is 0.1771. It implies a weak relationship between mpg and acceleration,in that only 17% of the variation in mpg is explained by acceleration.

The residual standard error is 7.08, so that actual mpg would be on average 7.08 miles per gallon different to the prediction. The mean mpg is 23.4459184 so the percentage error is $7.08/23.45 = 30\%$

The relationship between mpg and acceleration is minimally positive. The $\hat{\beta}_0$ intercept is 4.83 with the $\hat{\beta}_1$ at 1.19. This means the estimated mpg increases by 1.19 for each additional unit increase in acceleration. The range for acceleration is 8 to 24.8, the interpretation is only valid over that range. It is not meaningful to interpret the $\hat{\beta}_0$ intercept at 4.83 as the mpg at 0 horsepower.

The predicted mpg associated with acceleration of 14.50 is $(1.19 * 14.5) + 4.8332 = 22.0882$

97% Confidence intervals for acceleration at 14.5

```
predict(lm.fit, data.frame(acceleration = 14.5), interval = "confidence", level = 0.97)
```

```
##       fit      lwr      upr
## 1 22.1988 21.36615 23.03145
```

97% Prediction intervals for acceleration at 14.5

```
predict(lm.fit, data.frame(acceleration = 14.5), interval = "prediction", level = 0.97)
```

```
##       fit      lwr      upr
## 1 22.1988 6.755208 37.64239
```

The confidence intervals are very narrow with lower at 21.36 and the upper at 23.03 reflecting the weak relationship between mpg and acceleration. Prediction intervals are much wider showing uncertainty of prediction in mpg from acceleration and the standard error at 7.08. A simple linear model is used, it may be considered that this is not the best fit model for these variables.

Plotting the response and the predictor using the `abline()` function to display the least squares regression line. The 97% confidence intervals and prediction intervals are shown

```
plot(mpg~acceleration)
abline(lm.fit, col = "red")
p_conf <- predict(lm.fit,interval = "confidence", level = 0.97)
p_pred <- predict(lm.fit,interval = "prediction", level = 0.97)
lines(acceleration, p_conf[,"lwr"], col ="green", type = "b", pch = "-")
lines(acceleration, p_conf[,"upr"], col ="green", type = "b", pch = "-")
lines(acceleration, p_pred[,"upr"], col ="blue", type = "b", pch = "-")
lines(acceleration, p_pred[,"lwr"], col ="blue", type = "b", pch = "-")
legend("topleft",pch = c("-","-"),col =c("green", "blue"), legend = c("Confidence", "Prediction"))
```