Cars Regression Model

Charles Westby

11/14/2017

Synopsis

In this paper we are going to examine a dataset that was extracted from the 1974 Motor Trend US magazine. It looks at the fuel consumption and 10 apsects of autmobile design and performance for 32 (1973-74 models) automobiles. We will use this data to try and find out if an automatic or manual transmission is better for MPG.

Executive Summary

In this paper we will show the process where we built two different regression models trying to figure out whether automatic or a manual transmission is better for MPG. In our analysis we will find that our dataset has a higher mean MPG for manual transmissions than there are for automatic transmissions. We will build our first regression model using mpg as a dependent variable and only the tranmission variable, am. From here we will see a large difference between automatic and manual transmission vehicles. However there is a lot of room for error in our first model. We will then include the other regressors from the dataset. After doing this we will find that transmission does have a small impact on the mpg of a vehicle. A manual transmission vehicle will be better than an automatic transmission vehicle by about 1.2 mpg. Factors that affect mpg more than transmission are weight, cylinders, rear axle ratio, V/S, gear and carborator. About 89.31% of the variability in mpg is explained by our second model. The actual values and the predicted values in our model only differ by about 2.83 percentage points. So although our model is fairly accurate, there is some uncertainty.

Data Processing

Loading Libraries and Data

Here we loaded packages that we will use to manipulate and graph the data

```
library(ggplot2)
library(dplyr)
library(gridExtra)
data("mtcars")
```

Previewing the Data

Here we preview the data to get an idea of how we should analyze it

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
```

```
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Here we see there are 11 different variables and 32 different observations for each variable. They all appear to be of class numeric, but they need to be converted to factor variables because they are categorical. Our variable of interest am (Transmission) is one of them. So we transform the data

Converting Number Variables to Factor Variables

Here we convert the categorical variables from number variables to factor variables. Then we look at the structure of the data again. We see that vs is now a factor variable with 2 levels, am is a factor variable with 3 levels, gear is a factor with 3 levels, cyl is a factor with 3 levels and carb is a factor with 6 levels.

```
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
mtcars$cyl <- as.factor(mtcars$cyl)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
```

```
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Results

Summary after Variable Transformation

Now we will begin to analyze the data. The first thing we do is get a new summary of the data. Here we see that there are more automatic transmission cars than thear are manual ones.

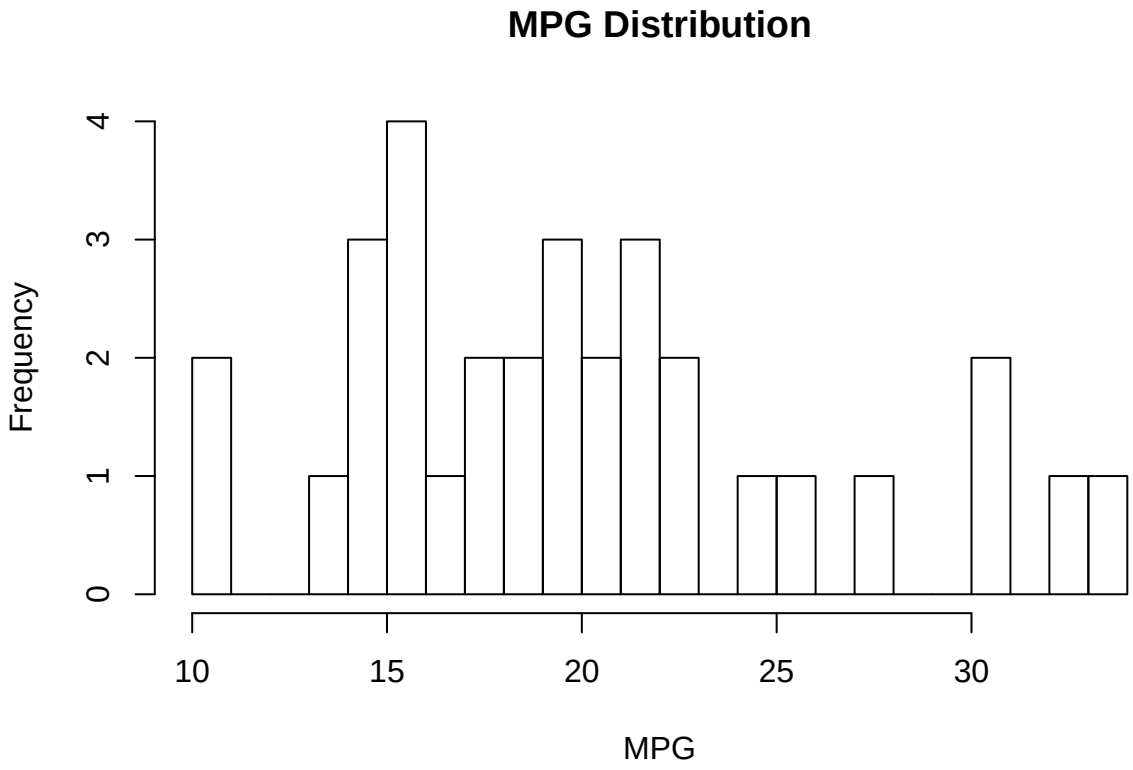
```
summary(mtcars)
```

```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt             qsec       vs     am     gear   carb
##  Min.   :1.513   Min.   :14.50   0:18   0:19   3:15   1: 7
##  1st Qu.:2.581   1st Qu.:16.89   1:14   1:13   4:12   2:10
##  Median :3.325   Median :17.71                 5: 5   3: 3
##  Mean   :3.217   Mean   :17.85                        4:10
##  3rd Qu.:3.610   3rd Qu.:18.90                        6: 1
##  Max.   :5.424   Max.   :22.90                        8: 1
```

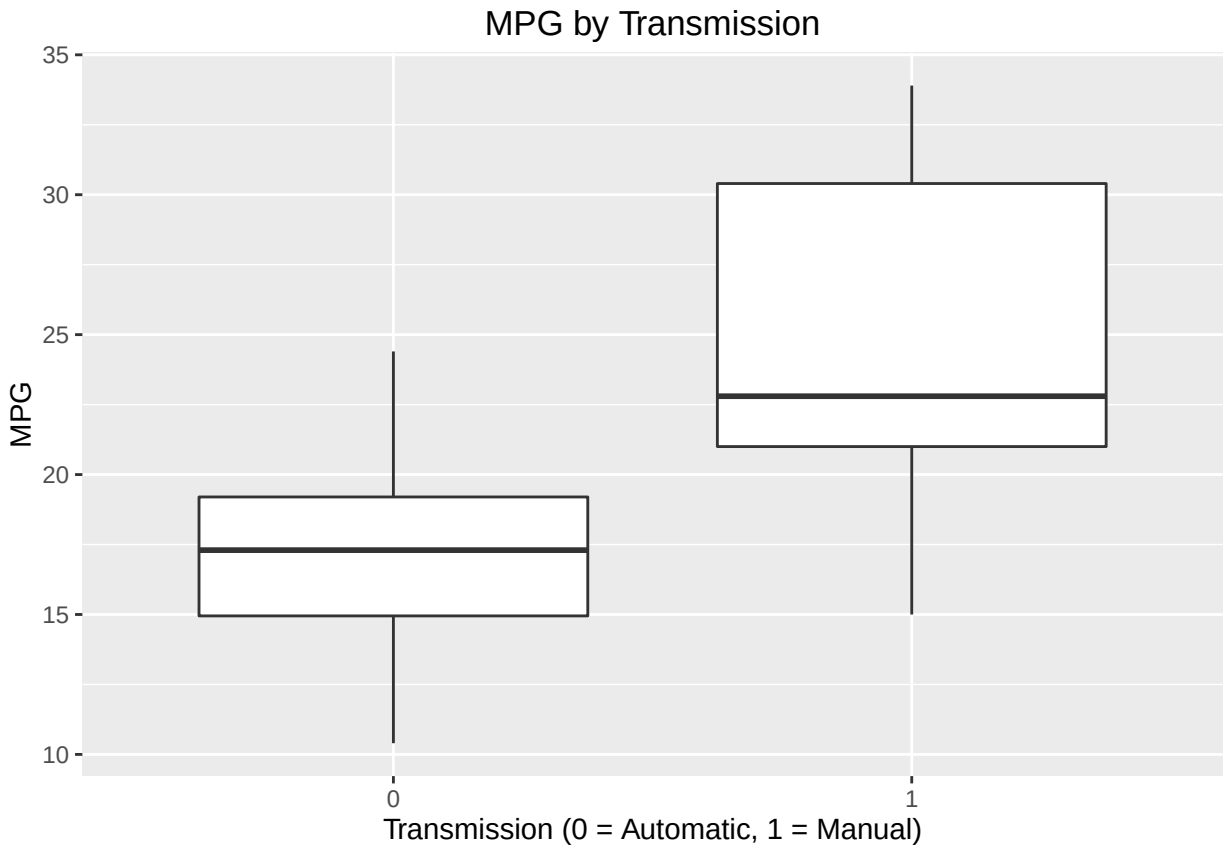
Graphing The Data

Here we will graph the data to get a better sense of how the am (Transmission) variable affects mpg. First we will look at a distribution of the mpg variable. Then we will look at the spread broken down by transmission. When looking at the spread of the distribution we see that most of the data is centered around 13-23 mpg. We also see in the graph of mpg broken down by transmission that manual transmission cars get better gas mileage.

```
hist(mtcars$mpg, xlab = "MPG", main = "MPG Distribution", breaks = 30)
```




```
ggplot(mtcars, aes(x = factor(am), y = mpg)) +
  geom_boxplot() +
  labs(x = "Transmission (0 = Automatic, 1 = Manual)",
       y = "MPG", title = "MPG by Transmission")
```



Quantifying The Median

We see that the median mpg in a manual transmission car is greater than the median mpg in an automatic transmission car. We will calculate the median mpg's for the cars to see the numbers

```
median_summary <- mtcars %>%
  group_by(am) %>%
  summarize(median_mpg = median(mpg))
median_summary
```

```
## Source: local data frame [2 x 2]
##
##       am median_mpg
##   (fctr)      (dbl)
## 1      0       17.3
## 2      1       22.8
```

Regression Models

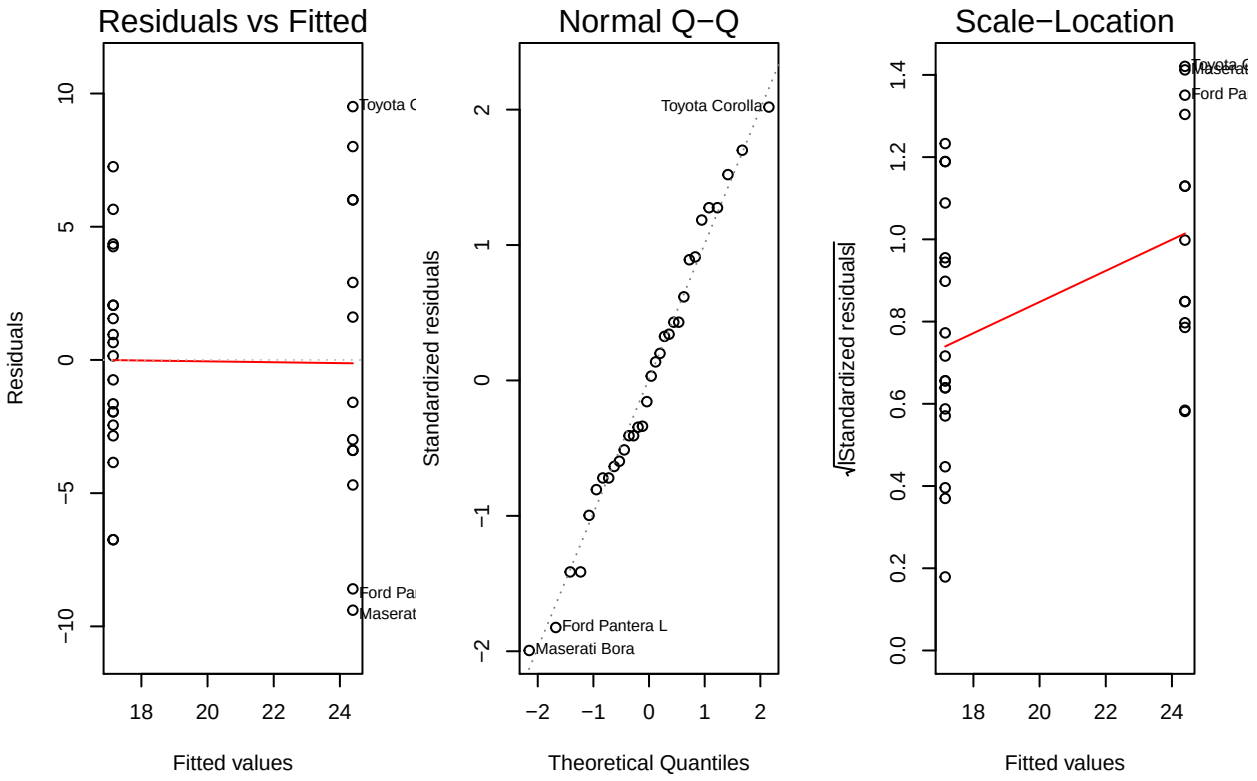
We will fit a few regression models on the data to see if there is any change in mpg based on transmission. Also we will look at how transmission affects mpg when considering the other variables.

Model 1: MPG by Transmission

```
fit_am <- lm(mpg ~ am, mtcars)
summary(fit_am)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am1            7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```
par(mfrow = c(1,3))
plot(fit_am, which = 1)
plot(fit_am, which = 2)
plot(fit_am, which = 3)
```



From this model we see that the automatic transmission will have a mean of about 17.15 mpg. When interpretting the model we see the (Intercept) estimate as the first level of the factor variable am. The first level of this variable is automatic transmission. The difference in means between automatic transmission and manual transmission in our model is about 7.25 mpg. Therefore manual tranmsission cars have a mean of about 24.40 mpg. The *** next to the estimates indicate that our estimates are significant to the 0.001 confidence interval. The Resisual Standard Error is about 4.90 which means that difference between actual mpg and predicted mpg on the model differ by about 4.90 percentage points. The R-squared is 0.3598, which says that about 35.98% of the variability in mpg is explained by this model.

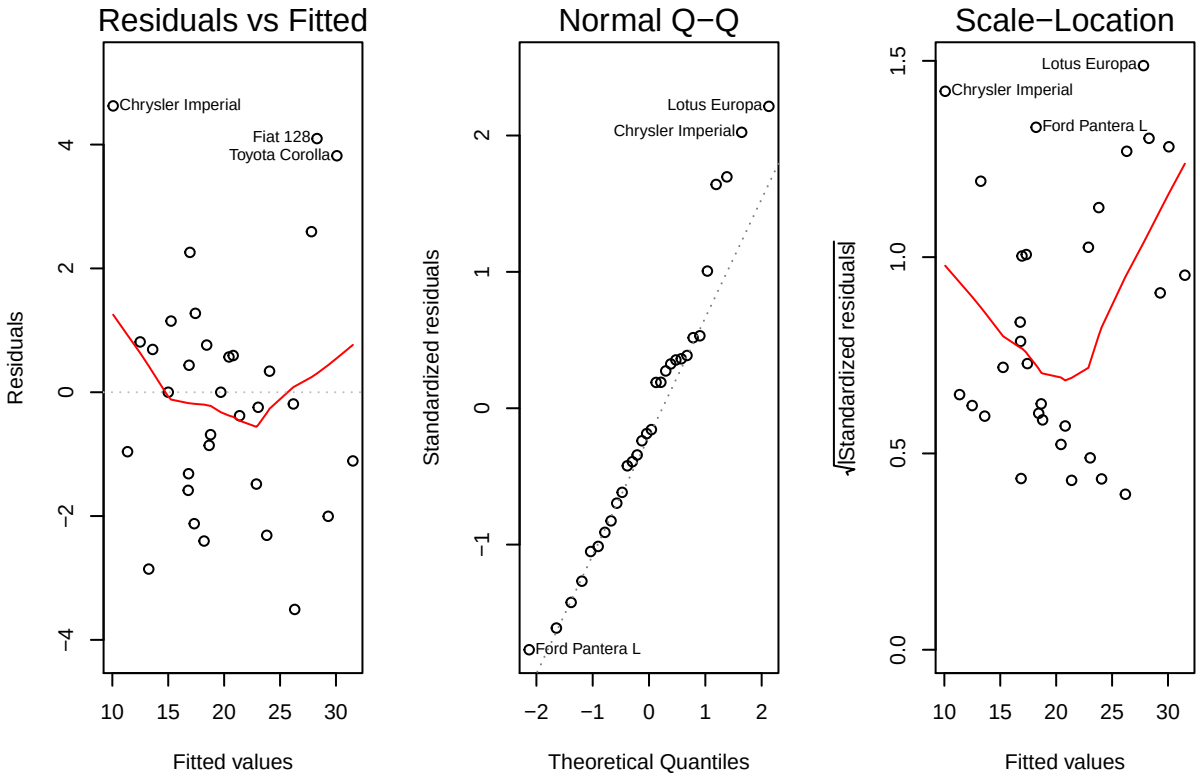
Although this model shows a relationship between mpg and transmission, there are other factors to consider when trying to determine this relationship. Each variable that is added to the model will change the coefficient for transmission. Variables correlated with transmission will have more of an effect on its coefficient than uncorrelated variables, but all added variables will have an effect. So we construct a new model containing all variables. This model will have the best fit when determining predicted mpg based on the data.

Model 2: MPG by All Variables

```
fit_all <- lm(mpg ~ ., mtcars)
summary(fit_all)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## am1          1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

```
par(mfrow = c(1,3))
plot(fit_all, which = 1)
plot(fit_all, which = 2)
plot(fit_all, which = 3)
```



From our model where we found the relationship between all variables and mpg, we see from the (Intercept) Estimate that if all variables are held equal to 1 that the mpg starts with 23.88. When the transmission is manual it gets about 1.21 mpg more than a manual transmission. This statistic is not highly significant in the model. The model has a residual standard error of 2.83, which means that the difference between the actual mpg and predicted mpg in the model differ by about 2.82 percentage points. We also get an R-squared of 0.8931. So about 89.31% of the variability is explained by this model. The Residual Fit Plot looks how we would expect it to look if residuals were independently and almost identically distributed with zero mean, and were uncorrelated with the fit. The highest residuals were for the outliers. The QQ Plot shows how the outliers, the Chrysler Impala, Lotus Europa and Fiat 128 affect the curve. Although they change the regression model, their impact is important and it would be unwise to remove them.

```
anova(fit_am, fit_all)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.9
## 2     15 120.4 15    600.49 4.9874 0.001759 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual Sum of Sqares (RSS) decreases in the models from 720.90 to 120.4. Therefore there is less deviance in Model 2 than there is in Model 1. The ** at the right of the table indicates that the null hypothesis is

rejected at the level of 0.01. So at least one of the additional regressors is significant. This rejection is based on applying Pr(>F) to the F statistic 4.99. All of this is evidence that the second model is a better predictor than the first model is.