Automatic or manual transmission. Which has the lowest mpg?

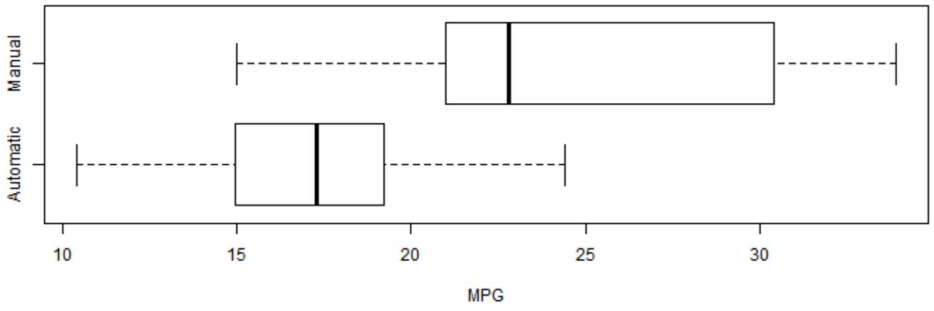
Summary

Looking at the mtcars dataset, and performing several linear models, we see that automatic cars do have a statistically significant smaller mpg than manual cars. However, there are other factors, such as weight, that play a much larger role in determining miles per gallon.

Discussion

First, let's get a boxplot of the mpg of the cars in mtcars dataset, separated into whether they're manual or automatic

```
library(datasets)
data <- mtcars
boxplot(data$mpg ~ data$am, xlab = "MPG", names = c("Automatic", "Manual"),
    horizontal = T)
```



While they do look differently distributed and manual is generally greater than automatic in mpg, are they statistically significant from each other? (Let's assume mpg in cars are normally distributed)

```
man <- data[data$am == 1, ]
aut <- data[data$am == 0, ]
t.test(man$mpg, aut$mpg, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  man$mpg and aut$mpg
## t = 3.767, df = 18.33, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.21 11.28
## sample estimates:
## mean of x mean of y
##     24.39     17.15
```

So in general, manual cars have generally higher mpg than automatic cars. But let's go into more detail about this.

While generally, automatic cars have higher mpg than manual cars, there is many other factors involved, especially when looking at the many variables in the dataset. Weight can definitely be seen as a factor of mpg, and automatic cars are generally heavier than manual cars (I would perform a statistical test here which proves it, but it takes too much room).

We can account and adjust for all these variables by creating a linear model, where we can see all the variables and how they may affect the mpg themselves, as well as levels of uncertainty about the affect.

```
a <- lm(mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb, data = data)
summary(a)$coefficients
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788  0.6573  0.51812
## cyl         -0.11144    1.04502 -0.1066  0.91609
## disp         0.01334    0.01786  0.7468  0.46349
## hp          -0.02148    0.02177 -0.9868  0.33496
## drat         0.78711    1.63537  0.4813  0.63528
## wt          -3.71530    1.89441 -1.9612  0.06325
## qsec         0.82104    0.73084  1.1234  0.27394
## vs           0.31776    2.10451  0.1510  0.88142
## am           2.52023    2.05665  1.2254  0.23399
## gear         0.65541    1.49326  0.4389  0.66521
## carb        -0.19942    0.82875 -0.2406  0.81218
```

Not too successful. We can perfect the model by performing a backward stepwise algorithm, which removes the least significant terms until only the

significant terms are left (see appendix). This leaves us with weight, qsec (time to travel ¼ mile), and am (automatic = 0, manual = 1).

```
b <- lm(mpg ~ wt + qsec + am, data = data)
summary(b)$coefficients
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    9.618     6.9596   1.382 1.779e-01
## wt            -3.917     0.7112  -5.507 6.953e-06
## qsec           1.226     0.2887   4.247 2.162e-04
## am             2.936     1.4109   2.081 4.672e-02
```

95% confidence intervals for am

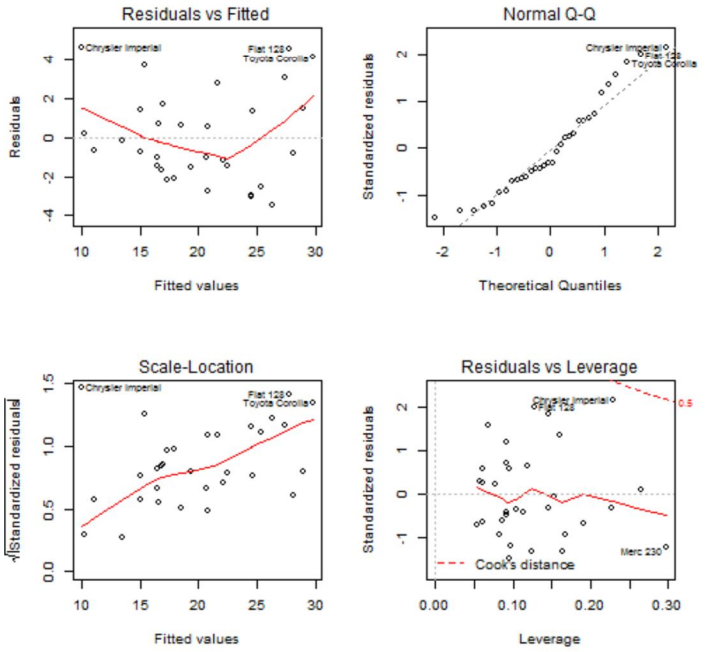
```
2.9358 + 1.41 * c(-1, 1) * pt(0.95, df = b$df)
```

```
## [1] 1.773 4.099
```

So when accounting for all the other variables in the dataset, and removing those which don't seem to affect mpg significantly, we are still left with am being an significant factor in determining mpg, with having a car with an automatic transmission taking 2.94 mpg compared to manual transmission cars (95% confidence intervals 1.7727, 4.0989). Weight, however, makes a much more significant difference to mpg (losing 3.92 mpg per 1000lb - 95% confidence intervals 3.3335, 4.5065)

Let's see if the assumptions that we used in creating the linear model hold.

```
par(mfrow = c(2, 2))
plot(b)
```



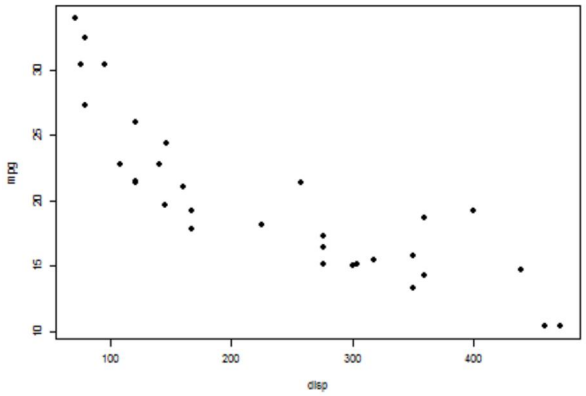
They seem to hold relatively well.

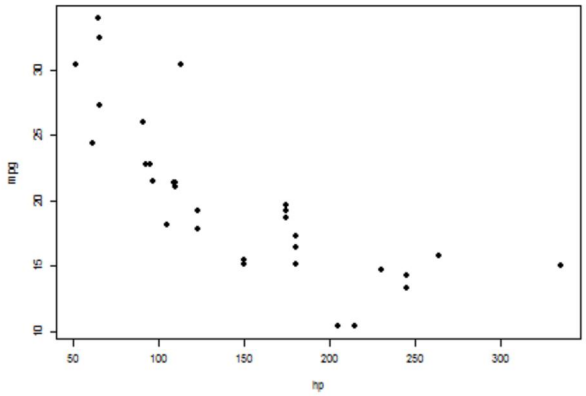
- Residuals vs Fitted: there seems surround 0 pretty evenly, with no disearnable pattern
- Normal Q-Q: seem relatively normal except at the edges
- Scale-Location: No disearnable pattern in here eariler, though the line does suggest an increase in residual variation as mpg goes up.
- Residuals vs. Leverage: appears fine, with points that hold high leverage near 0.

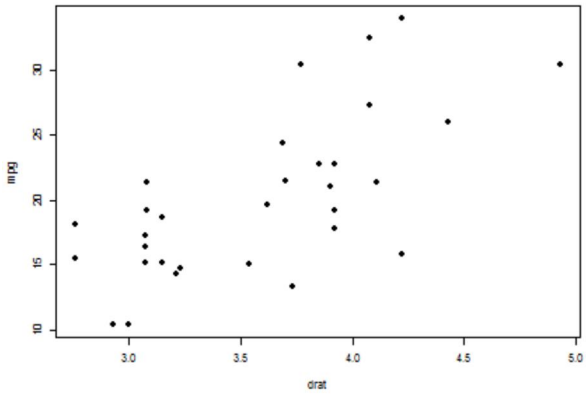
Appendix

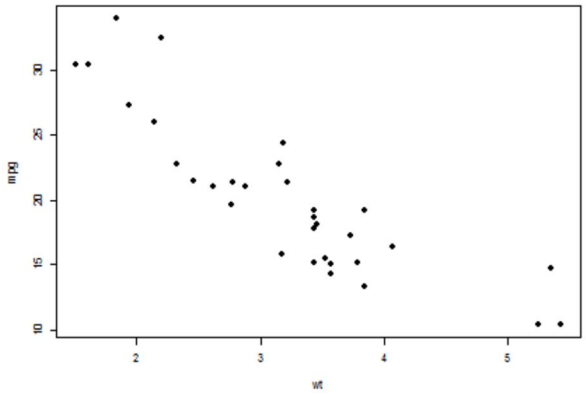
Backward Stepwise Regression Algorithm

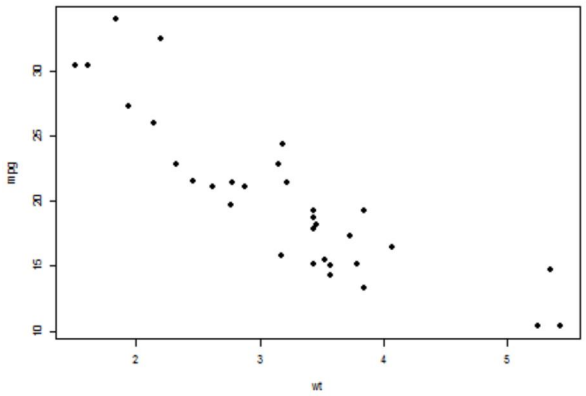
```
par(mfrow = c(4, 2), pch = 16)
plot(mtcars$disp, mtcars$mpg, xlab = "disp", ylab = "mpg")
plot(mtcars$hp, mtcars$mpg, xlab = "hp", ylab = "mpg")
plot(mtcars$drat, mtcars$mpg, xlab = "drat", ylab = "mpg")
plot(mtcars$wt, mtcars$mpg, xlab = "wt", ylab = "mpg")
plot(mtcars$wt, mtcars$mpg, xlab = "wt", ylab = "mpg")
plot(mtcars$qsec, mtcars$mpg, xlab = "qsec", ylab = "mpg")
plot(mtcars$am, mtcars$mpg, xlab = "am", ylab = "mpg")
plot(mtcars$carb, mtcars$mpg, xlab = "carb", ylab = "mpg")
```

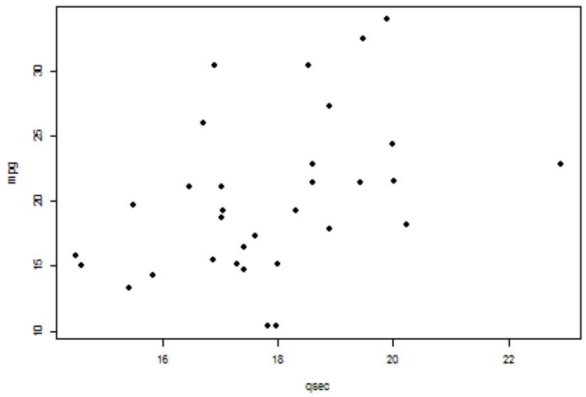


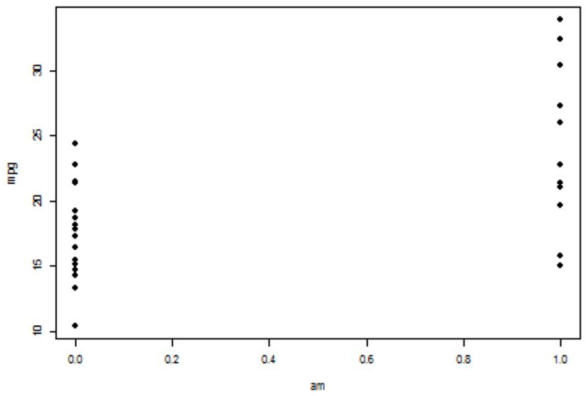


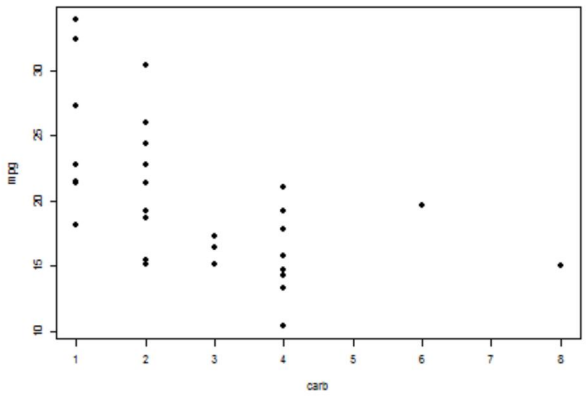













```
step(a, direction = "backward")
```

```
## Start:  AIC=70.9
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq RSS  AIC
## - cyl   1      0.08 148 68.9
## - vs    1      0.16 148 68.9
## - carb  1      0.41 148 69.0
## - gear  1      1.35 149 69.2
## - drat  1      1.63 149 69.2
## - disp  1      3.92 151 69.7
## - hp    1      6.84 154 70.3
## - qsec  1      8.86 156 70.8
## <none>               148 70.9
## - am    1     10.55 158 71.1
## - wt    1     27.01 174 74.3
##
## Step:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq RSS  AIC
## - vs    1      0.27 148 67.0
## - carb  1      0.52 148 67.0
## - gear  1      1.82 149 67.3
## - drat  1      1.98 150 67.3
## - disp  1      3.90 152 67.7
## - hp    1      7.36 155 68.5
## <none>               148 68.9
## - qsec  1     10.09 158 69.0
## - am    1     11.84 159 69.4
## - wt    1     27.03 175 72.3
##
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
##        Df Sum of Sq RSS  AIC
## - carb  1      0.69 148 65.1
## - gear  1      2.14 150 65.4
## - drat  1      2.21 150 65.4
## - disp  1      3.65 152 65.8
## - hp    1      7.11 155 66.5
## <none>               148 67.0
## - am    1     11.57 159 67.4
## - qsec  1     15.68 164 68.2
## - wt    1     27.38 175 70.4
##
## Step:  AIC=65.12
## mpg ~ disp + hp + drat + wt + qsec + am + gear
##
##        Df Sum of Sq RSS  AIC
## - gear  1       1.6 150 63.5
## - drat  1       1.9 150 63.5
## <none>               148 65.1
## - disp  1      10.1 159 65.2
## - am    1      12.3 161 65.7
## - hp    1      14.8 163 66.2
## - qsec  1      26.4 175 68.4
## - wt    1      69.1 218 75.3
##
## Step:  AIC=63.46
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq RSS  AIC
## - drat  1       3.3 153 62.2
## - disp  1       8.5 159 63.2
## <none>               150 63.5
## - hp    1      13.3 163 64.2
## - am    1      20.0 170 65.5
## - qsec  1      25.6 176 66.5
## - wt    1      67.6 218 73.4
##
## Step:  AIC=62.16
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq RSS  AIC
## - disp  1       6.6 160 61.5
## <none>               153 62.2
## - hp    1      12.6 166 62.7
## - qsec  1      26.5 180 65.3
## - am    1      32.2 186 66.3
## - wt    1      69.0 222 72.1
##
## Step:  AIC=61.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq RSS  AIC
## - hp    1       9.2 169 61.3
## <none>               160 61.5
## - qsec  1      20.2 180 63.3
## - am    1      26.0 186 64.3
## - wt    1      78.5 239 72.3
##
```

```
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq RSS  AIC
## <none>              169 61.3
## - am    1      26.2 195 63.9
## - qsec  1     109.0 278 75.2
## - wt    1     183.3 353 82.8
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = data)
##
## Coefficients:
## (Intercept)           wt         qsec           am
##        9.62        -3.92         1.23         2.94
```