Executive Summary

Using "mtcars" dataset, this report explores the relationship between a set of variables and miles per gallon (MPG). In particular, we do not find enough evidence for any difference in MPG between automatic and manual transmission. However, we do find enough evidence showing that increasing (1) number of engine cylinders and (2) car weight does reduce MPG.

Understanding Car Design

Engine displacement is measured by volume per cylinder times number of cylinders. This determines amount of fuel drawn into an engine to create power. The fuel is mixed with air drawn in by carburetors. Carburetors with more barrels are used for higher air flow needed in larger engine displacement. The engine/driveshaft transfers power to rear axle/wheels via connecting gears. Rear axle ratio is the number of turns the driveshaft spins in order to spin the rear axle one complete turn. A higher ratio makes it easier for the engine to turn the wheels. A straight (inline) engine has cylinders moving in opposite directions while a V-shaped engine has cylinders moving at a V-angle to each other. The former does not require balancing shafts (the motions cancel each other) while the latter is more compact in design. Automatic transmission has more components than manual transmission i.e. more weight. More gears also means more weight, but is more efficient when accelerating.

Model Selection & Linear Regression

(see appendix 1 for exploratory data analysis) To summarise the theoretical relationships

- cylinders, others -> displacement -> carburetors
- cylinders, displacement, others -> horsepower -> qsec, mpg
- cylinders, displacement, others, transmission, gears -> weight -> qsec, mpg
- rear axle ratio -> qsec, mpg
- engine shape -> qsec, mpg

We recommend **nested model testing in the following order : transmission, cylinders, displacement, rear axle ratio, gears, weight and engine shape**. This would enable us to see how adding other factors affect the coefficient on transmission. We exclude carburetors, horsepower and qsec on the basis that these are largely derivatives of other chosen factors and likely do not have much residual explanatory power on their own.

```r
library(datasets); mc <- mtcars
fit1 <- lm(mpg~am, data=mc); fit2 <- lm(mpg~am+cyl, data=mc)
fit3 <- lm(mpg~am+cyl+disp, data=mc); fit4 <- lm(mpg~am+cyl+disp+drat, data=mc)
fit5 <- lm(mpg~am+cyl+disp+drat+gear, data=mc)
fit6 <- lm(mpg~am+cyl+disp+drat+gear+wt, data=mc)
fit7 <- lm(mpg~am+cyl+disp+drat+gear+wt+vs, data=mc)
coco <- data.frame()
for (i in 1:7) {coco[1:7,i] <- round(coef(get(paste("fit",i,sep="")))[2:8],2)
                colnames(coco)[i] <- paste("fit",i,sep="")}
rownames(coco) <- c("am","cyl","disp","drat","gear","wt","vs"); coco
```

```
##      fit1  fit2  fit3  fit4  fit5  fit6  fit7
## am   7.24  2.57  1.93  1.86  3.19  1.18  1.50
## cyl    NA -2.50 -1.62 -1.60 -1.54 -1.74 -1.52
## disp   NA    NA -0.02 -0.02 -0.02  0.01  0.01
## drat   NA    NA    NA  0.13  0.72  0.33  0.35
## gear   NA    NA    NA    NA -1.53 -1.07 -1.06
## wt     NA    NA    NA    NA    NA -3.43 -3.42
## vs     NA    NA    NA    NA    NA    NA  0.87
```

```r
anova(fit1, fit2, fit3, fit4, fit5, fit6, fit7)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + disp
## Model 4: mpg ~ am + cyl + disp + drat
## Model 5: mpg ~ am + cyl + disp + drat + gear
## Model 6: mpg ~ am + cyl + disp + drat + gear + wt
## Model 7: mpg ~ am + cyl + disp + drat + gear + wt + vs
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 59.7361 5.778e-08 ***
## 3     28 252.08  1     19.28  2.5621   0.12253
## 4     27 252.03  1      0.05  0.0071   0.93341
## 5     26 238.96  1     13.07  1.7371   0.19995
## 6     25 182.10  1     56.85  7.5551   0.01118 *
## 7     24 180.61  1      1.49  0.1983   0.66011
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Not surprisingly, **number of cylinders and weight have the greatest explanatory power for miles per gallon**. Their coefficients have stable directions that are inline with car design theory i.e. more cylinders and greater weight equal more fuel consumption, and have reasonably stable values across various fits. Our selected regression model shall use transmission type, number of cylinders and weight as regressors.

Interpretation & Statistical Inference

```r
fitC <- lm(mpg~am+cyl+wt, data=mc)
sumco <- summary(fitC)$coefficients; momo <- data.frame()
for (i in 2:4) {soso <- sumco[i,1] + c(-1,0,1) * qt(.975,df=fitC$df) * sumco[i,2]
                momo <- rbind(momo, round(soso,3))}
rownames(momo) <- c("am", "cyl", "wt")
colnames(momo) <- c("95CI lower", "Cofficient", "95CI upper"); momo
```

```
##     95CI lower Cofficient 95CI upper
## am      -2.496      0.176      2.849
## cyl     -2.375     -1.510     -0.645
## wt      -4.991     -3.125     -1.259
```

(see appendix 2 for diagnosis of residuals) Looking at the coefficients, holding cylinders and weight constant, an automatic transmission would increase fuel efficiency by 0.176 miles per gallon over a manual transmission. It would appear at firsthand that automatic transmission is better for miles per gallon. However, the 95% confidence interval for this coefficient contains zero within and **we don't have enough evidence to conclude if being automatic or manual transmission would affect miles per gallon**.

The coefficients also indicate that (1) holding transmission and weight constant, an increase of one cylinder would reduce miles per gallon by 1.510, and (2) holding transmission and number of cylinders constant, an increase in weight by 1,000 pounds would reduce miles per gallon by 3.125. The 95% confidence intervals for both coefficients do not contain zero within and **we have enough evidence to conclude that increasing both cylinders and weight does reduce miles per gallon**.

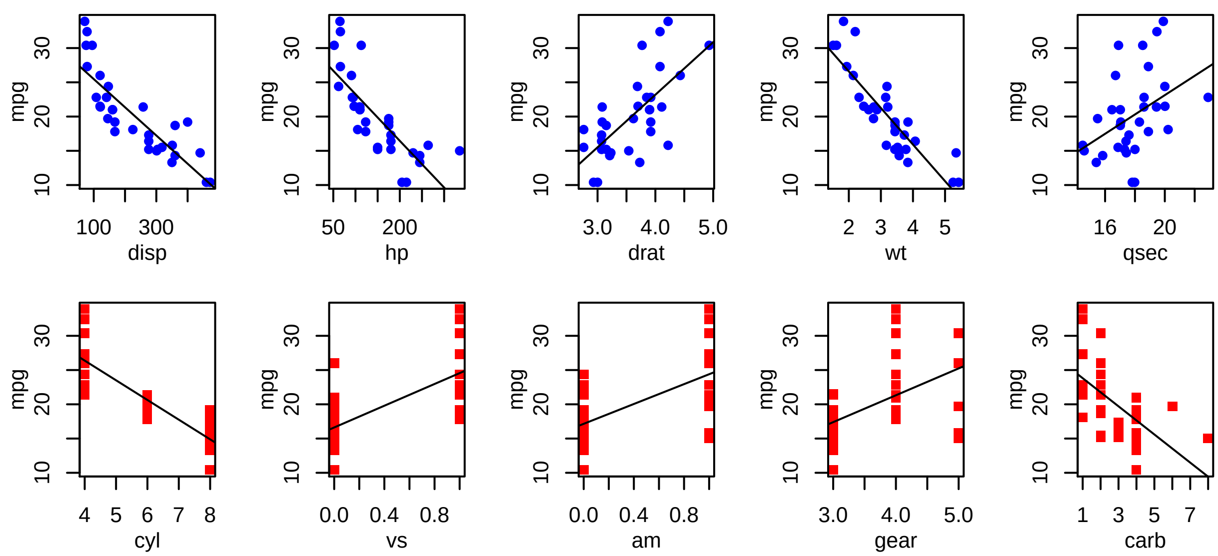
Appendix 01 : Exploratory Data Analysis

```r
for (i in c(2,8:11)) {mc[,i] <- factor(mc[,i])}
mc[sample(nrow(mc),6),]; summary(mc); mc <- mtcars
```

```
##                      mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
```

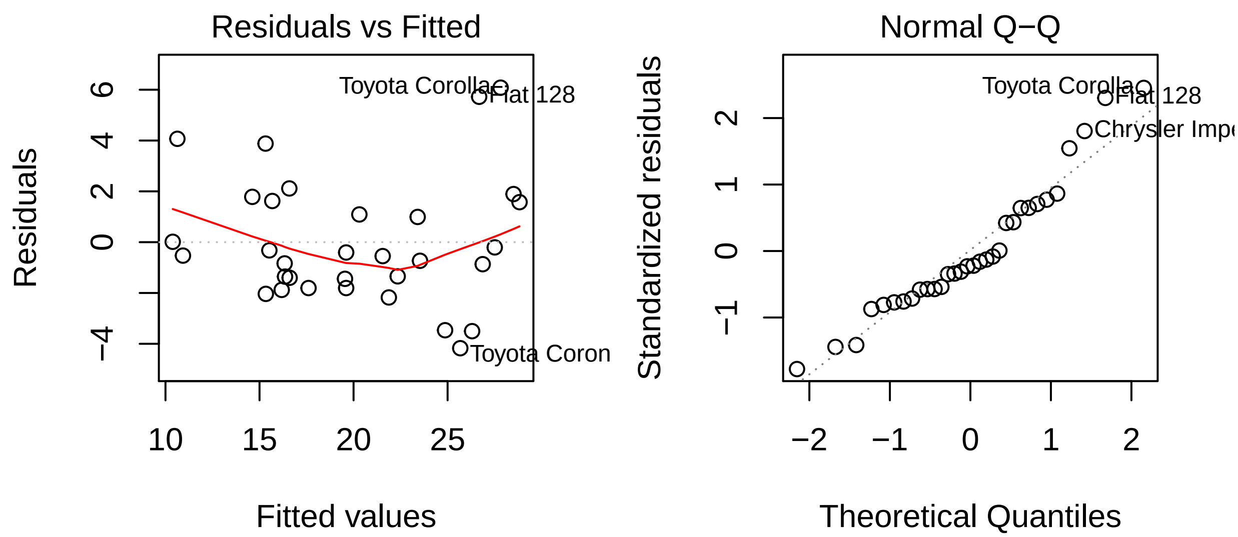
```
##       mpg        cyl        disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :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
```

```r
par(mfrow=c(2,5), mar=c(3.5,3.5,1,1), mgp=c(2,1,0))
for (i in c(3:7)) {
    plot(x=mc[,i], y=mc[,1], xlab=colnames(mc)[i], ylab="mpg", pch=16, col="blue")
    abline(lm(mc[,1]~mc[,i]))}
for (i in c(2,8:11)) {
    plot(x=mc[,i], y=mc[,1], xlab=colnames(mc)[i], ylab="mpg", pch=15, col="red")
    abline(lm(mc[,1]~mc[,i]))}
```



Appendix 02 : Diagnosis Of Residuals

```
par(mfrow=c(1,2), mar=c(4.5,4.5,2,2), oma=c(0,0,0,0))
plot(fitC, which=1:2)
```



The residual distribution seems to fit normality assumptions reasonably. However, there seems to be a hint that as fitted values get larger, residuals get smaller (or more negative) thus challenging the regression assumption of no heteroskedasticity (and also suggesting that there may be missing model terms).

```
mcZ <- mc; rownames(mcZ) <- 1:32
fitZ <- lm(mpg~am+cyl+wt, data=mcZ)
round(dfbetas(fitZ)[1:32,2],2) # influence = leverage + outlying
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12
## -0.10 -0.05 -0.22 -0.09 -0.16  0.09  0.12 -0.09  0.07  0.03  0.12 -0.04
##    13    14    15    16    17    18    19    20    21    22    23    24
## -0.08  0.01 -0.04  0.00  0.40  0.32  0.01  0.16  0.67  0.06  0.11  0.08
##    25    26    27    28    29    30    31    32
## -0.16 -0.01 -0.04  0.00 -0.24 -0.20 -0.22 -0.37
```

```
round(hatvalues(fitZ)[1:32],2) # leverage
```

```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 0.09 0.09 0.10 0.07 0.12 0.07 0.10 0.19 0.19 0.07 0.07 0.07 0.08 0.08 0.23
##   16   17   18   19   20   21   22   23   24   25   26   27   28   29   30
## 0.28 0.26 0.10 0.12 0.10 0.19 0.11 0.12 0.08 0.08 0.10 0.09 0.13 0.20 0.09
##   31   32
## 0.20 0.15
```

The influence measures (dfbetas) do not indicate any strong influential point that distorted the coefficient on transmission type. The leverage measures (hatvalues) do not indicate potential for any data entry errors.