John Walker

Regression Models Course Project

21 January 2016

Executive Summary

From the 1974 Motor Trend dataset `data(mtcars)` we use a linear model to answer two questions: Question 1) Is an automatic or manual transmission better for MPG? We find that **manual transmission is better** and that the difference using this dataset is statistically significant. The parsimonious model for mpg uses engine weight, quarter mile time and automatic vs manual transmission as regressors. 2) Quantify the MPG difference between automatic and manual transmissions. Our model shows that changing from automatic to **manual transmission improves mpg by 2.9** holding other variables constant.

Exploratory Data Analysis

From `mtcars` we can do an exploratory pairs plot using `ggpairs` (see Appendix - Figure 1). From the dataset documentation, cars with automatic transmission have a zero value for the `am` variable so it might be more readable to use values of "auto" and "manual" as factors. It seems to make sense to treat the am and also the vs variable as factors, where v indicates a "V" (value zero) or straight cylinder alignment.

```
mtcars$am <- as.factor(ifelse(mtcars$am == 0, "auto", "man")) ;
mtcars$vs <- as.factor(ifelse(mtcars$vs == 0, "V", "Str"))
```

Fitting and Selecting a Model

If we fit a linear model for mpg using only am as a regresor, the coefficients show the mean mileage for an automatic at 17.147mpg and that mileage increases by 7.245 mpg changing from automatic to manual. The P value for the difference from auto to manual 0.000285 is small so we can say the difference auto/manual is significant but in this model the R^2 is 0.359 so automatic/manual only explains about 36% of the variation in mpg for the data. We need a better model than this one.

```
fit0 <- lm(mpg ~ factor(am), mtcars) ; summary(fit0)$coef
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   17.147368   1.124603 15.247492 1.133983e-15
## factor(am)man  7.244939   1.764422  4.106127 2.850207e-04
```

If we fit a model using all the variables as regressors to predict mpg (knowing from the pairs plot that several of these are highly correlated) we can see from the square root of the variable inflation factor (VIF) - the increase in standard error - that this is not the right model either. We've gone from underfit to overfit.

```
fitall <- lm(mpg ~ cyl + disp + hp + drat+ wt + qsec + factor(vs) + factor(am) + gear + carb, mtcars)
sqrt(vif(fitall))
##        cyl       disp         hp       drat         wt       qsec
##   3.920948   4.649757   3.135608   1.837014   3.894212   2.743712
## factor(vs) factor(am)       gear       carb
##   2.228424   2.156035   2.314617   2.812249
```

We can use the step function from the model with all the variables to find a more parsimonious fit at a lower standard error - resulting in a model predicting mpg from weight, qsec (acceleration) and auto/manual transmission.

```
steps <- step(fitall, k=log(nrow(mtcars)), direction = "both", trace = FALSE)
summary(steps)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + factor(am), data = mtcars)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.6178     6.9596   1.382 0.177915
## wt             -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec            1.2259     0.2887   4.247 0.000216 ***
## factor(am)man   2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```
fitbest <- lm(mpg ~ wt + qsec + factor(am), mtcars)
```

We can look at the step result using nested models. The first three variables weight, qsec and auto/manual reduce the residual sum of squares. Adding engine displacement to the fourth model does not reduce RSS much and is not statistically significant even though displacement is highly correlated to mpg.

```
fit1 <- lm(mpg ~ wt, mtcars)
fit2 <- lm(mpg ~ wt + qsec, mtcars)
fit3 <- lm(mpg ~ wt + qsec + factor(am), mtcars)
fit4 <- lm(mpg ~ wt + qsec + factor(am) + disp, mtcars)
anova(fit1, fit2, fit3, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + qsec
## Model 3: mpg ~ wt + qsec + factor(am)
## Model 4: mpg ~ wt + qsec + factor(am) + disp
##   Res.Df    RSS Df Sum of Sq       F  Pr(>F)
## 1     30 278.32
## 2     29 195.46  1    82.858 13.4762 0.00105 **
## 3     28 169.29  1    26.178  4.2576 0.04881 *
## 4     27 166.01  1     3.276  0.5328 0.47171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

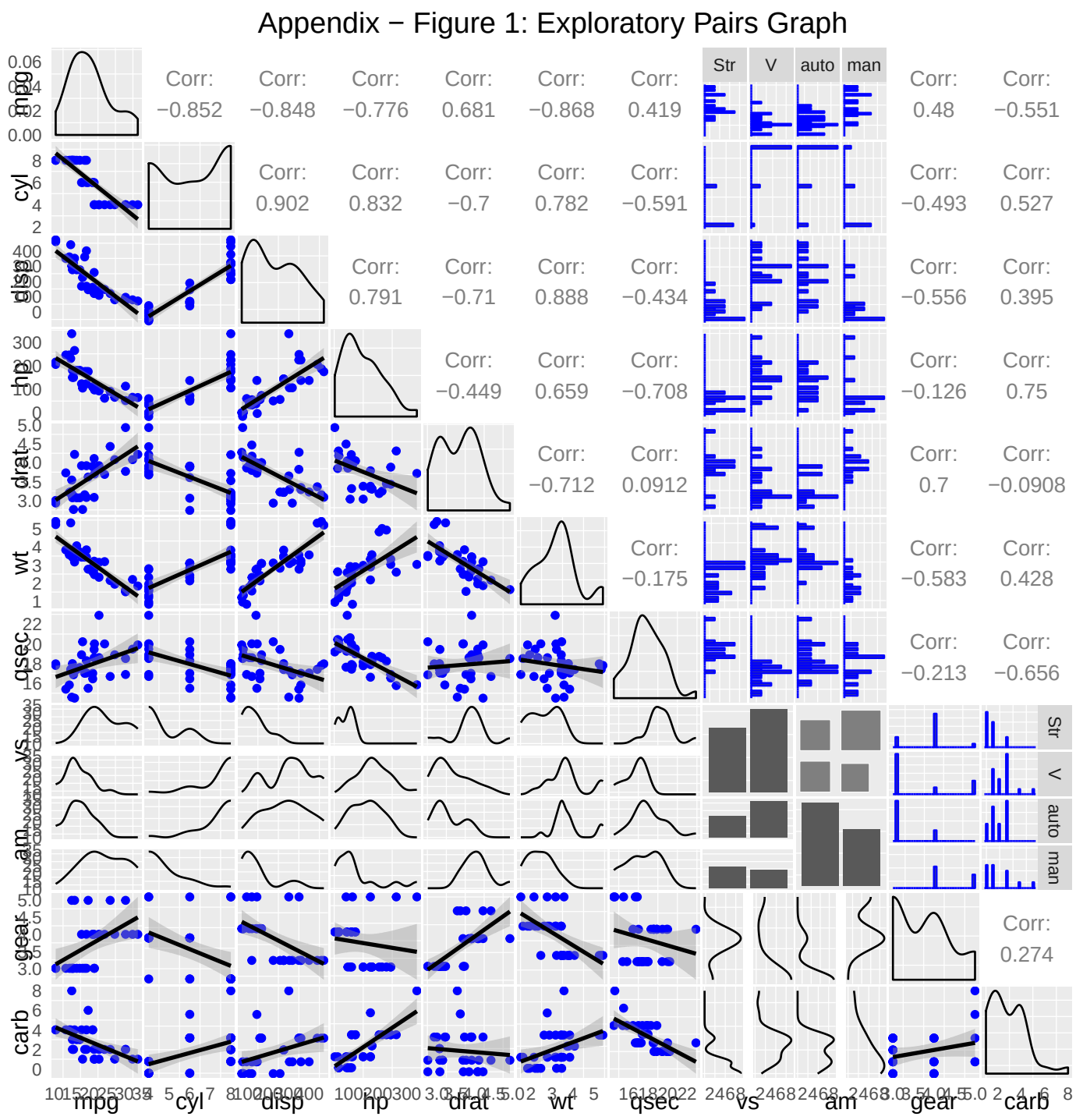
Interpretation: Coefficients, Residuals and Uncertainty

In the best fit model changing from automatic to manual transmission account improves mpg by 2.9. From the P-value it is significant, but only just so. The standard error has increased compared to the original model but the model now accounts for 83% of the variation mpg. Looking at variable inflation for the new model the numbers look much better. A confidence interval from the new model `confint(fitbest)` says with 95% confidence that the actual value is between 0.04 and 5.8 mpg gained switching from automatic to manual transmissions.

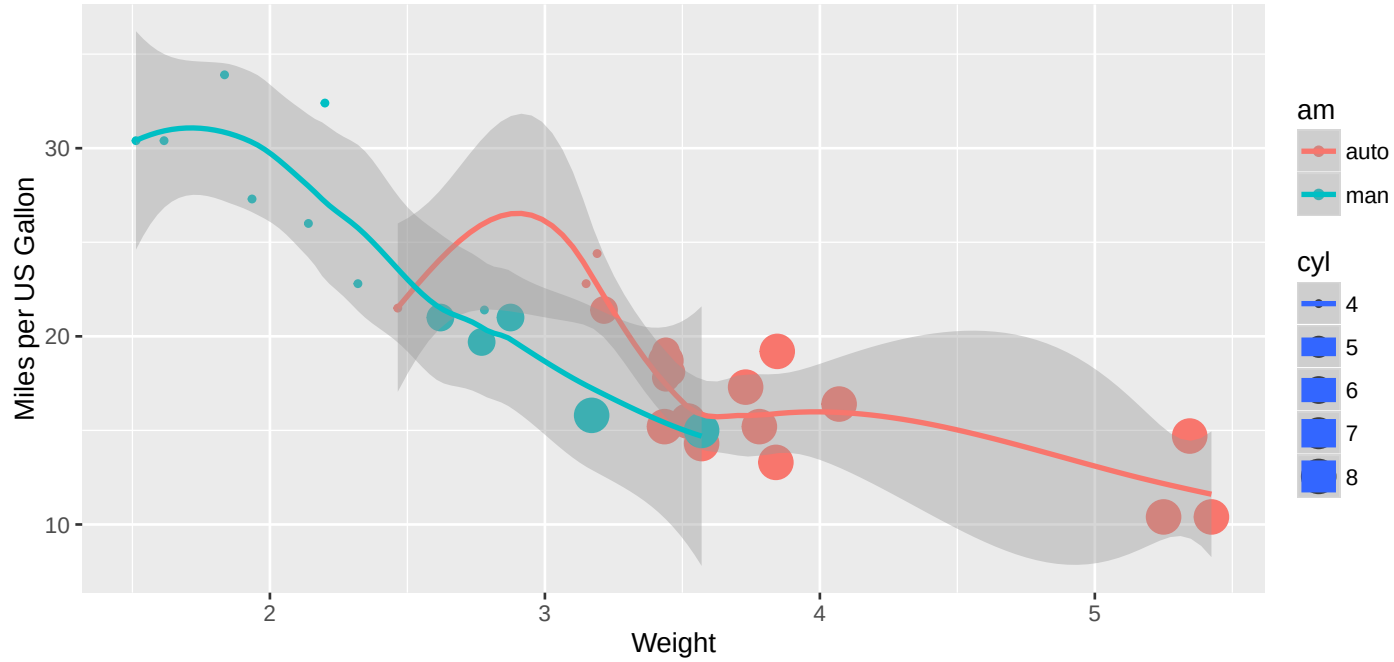
```
sqrt(vif(fitbest))
##         wt       qsec factor(am)
##   1.575738   1.168049   1.594189
```

Looking at the residuals vs fitted plot (Appendix - Figure 4a), there is no clear pattern remaining showing some quality of fit. The Q-Q plot (Appendix - Figure 4b) shows a slightly straight line - certainly problems but perhaps not a bad fit for a small number of data points. The Scale-Location and Residuals vs Leverage plots show no major problems. The influence plot (Appendix - Figure 5) shows that some points do have more significant influence.

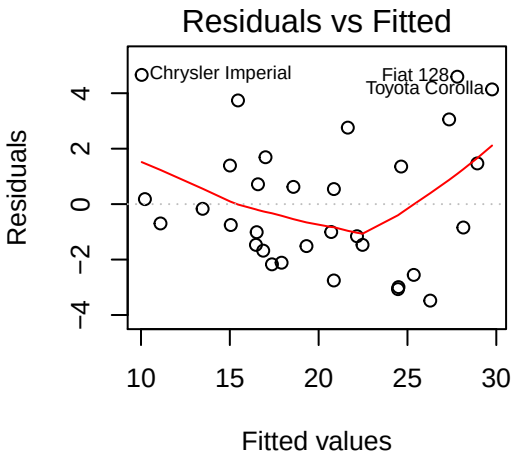
Apendix of Figures

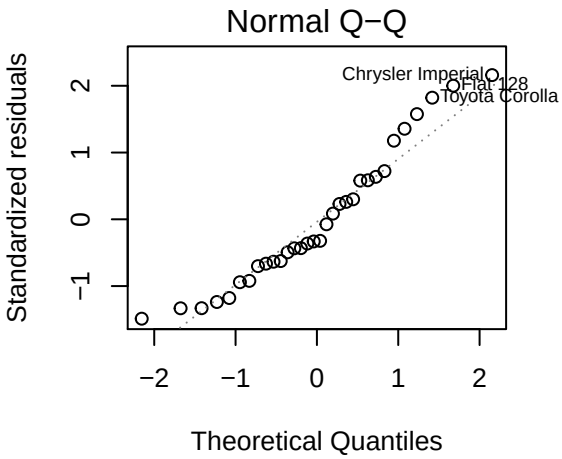


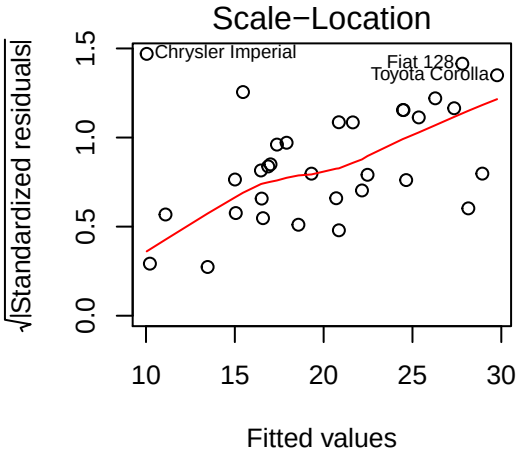
Appendix − Figure 2: MPG looking at Weight, Cylinders and Transmission Type

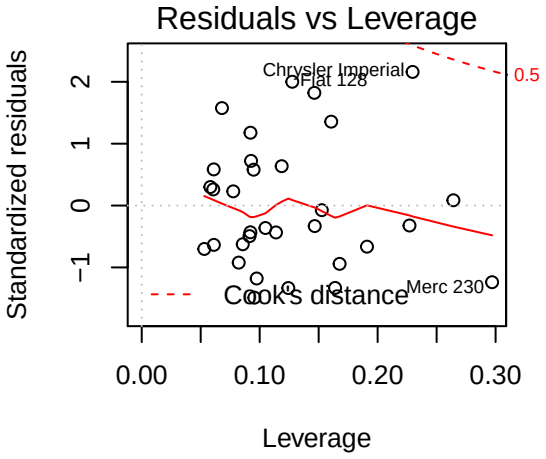


Appendix − Figures 4 a,b,c,d: Residual Plots from the fitbest linear model









Appendix − Figure 5:Influence Plot