# Regression Models Project

*February 22, 2015*

## EXECUTIVE SUMMARY

This project uses regression to address: (1)If automatic or manual transmission is better for MPG, and (2) A quantification of the MPG difference.

Preliminary conclusions: (1)Manual is worse than auto; (2)a rough quantity to represent that conclusion is the coefficient of the "am" variable in the model we end up using.

Shortcomings: More analysis needs to be done on variable interrelations. This would lead to a better "relative" importance sort of measure.

```
library(datasets)
data(mtcars)
attach(mtcars)
fit1 <- lm(mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb)
confint(fit1)
##                    2.5 %      97.5 %
## (Intercept) -26.62259745 51.22934576
## cyl          -2.28468553  2.06180457
## disp         -0.02380146  0.05047194
## hp           -0.06675236  0.02378812
## drat         -2.61383350  4.18805545
## wt           -7.65495413  0.22434628
## qsec         -0.69883421  2.34091571
## vs           -4.05880242  4.69432805
## am           -1.75681208  6.79726585
## gear         -2.44999107  3.76081711
## carb         -1.92290442  1.52406591
```

confint(fit1)'s output suggests candidates for elimination from an updated model: cyl, drat, vs, gear, and carb have 95% confidence intervals split relatively evenly by the number 0, unlike wt, whose interval lies mostly on one side.

```
fit2 <- lm(mpg ~ disp + hp + wt + qsec + am)
confint(fit2)
##                   2.5 %       97.5 %
## (Intercept) -5.66058661 34.384394537
## disp        -0.01055781  0.033033109
## hp          -0.05098537  0.008644273
## wt          -6.53883919 -1.629824922
## qsec         0.02963058  1.984163085
## am           0.41638869  6.524518102
```

We can compare fit1 and fit2 using anova() and the AIC.

```
anova(fit2, fit1)
## Analysis of Variance Table
##
## Model 1: mpg ~ disp + hp + wt + qsec + 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     26 153.44
## 2     21 147.49  5    5.9434 0.1692 0.9711
AIC(fit1, fit2)
##      df      AIC
```

```
## fit1 12 163.7098
## fit2  7 154.9740
```

Both tests favor fit2: anova because the testing added by the discarded variables to the linear prediction is insignificant (since p = 0.97); AIC because the AIC value for fit2 is much smaller than for fit1. confint(fit2) reveals that the beta coefficients for disp and hp are to p > 95% not big enough to have a noticeable effect on mpg.

```
fit3 <- lm(mpg ~ wt + qsec + am)
confint(fit3)
##                   2.5 %    97.5 %
## (Intercept) -4.63829946 23.873860
## wt          -5.37333423 -2.459673
## qsec         0.63457320  1.817199
## am           0.04573031  5.825944
```

Finally, we should consider interactions among the explaning variables in fit3. For brevity's sake, we find cor(wt,qsec), cor(wt,am), and cor(qsec,am), respectively.

```
## [1] -0.1747159
## [1] -0.6924953
## [1] -0.2298609
```

The only pair that we should integrate include is wt:am (because magnitude of cor(wt,am) much higher than other two).

```
fit4 <- lm(mpg ~ wt + qsec + am + wt:am)
summary(fit4)[8:9]
## $r.squared
## [1] 0.8958514
##
## $adj.r.squared
## [1] 0.8804219
```

summary(fit4) supports the choice of fit4 in that the un-adjusted R-squared value is 0.89 (adjusted, 0.88), which suggests fit4 "explains" about that proportion of the total variance from the population mean.

Some diagnostic info (consider along with figures in appendix)

```
dfb <- dfbetas(fit4)
dfbHigh <- dfb > 1 | dfb < -1
```

dfbHigh is just a logical vector that says whether an entry in fit4 is greater than 1 or less than -1.

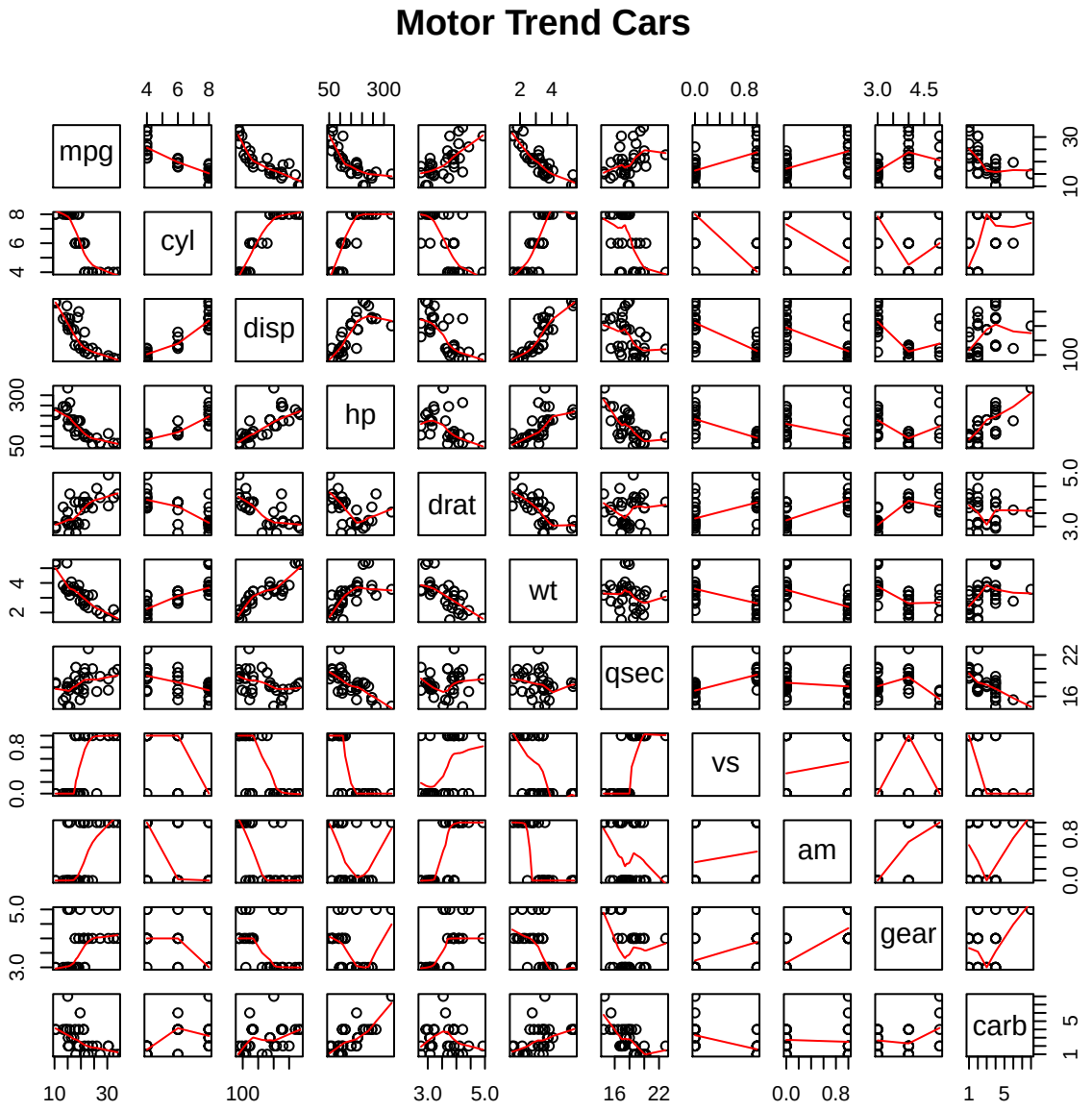
```
sum(dfbHigh)
## [1] 0
```

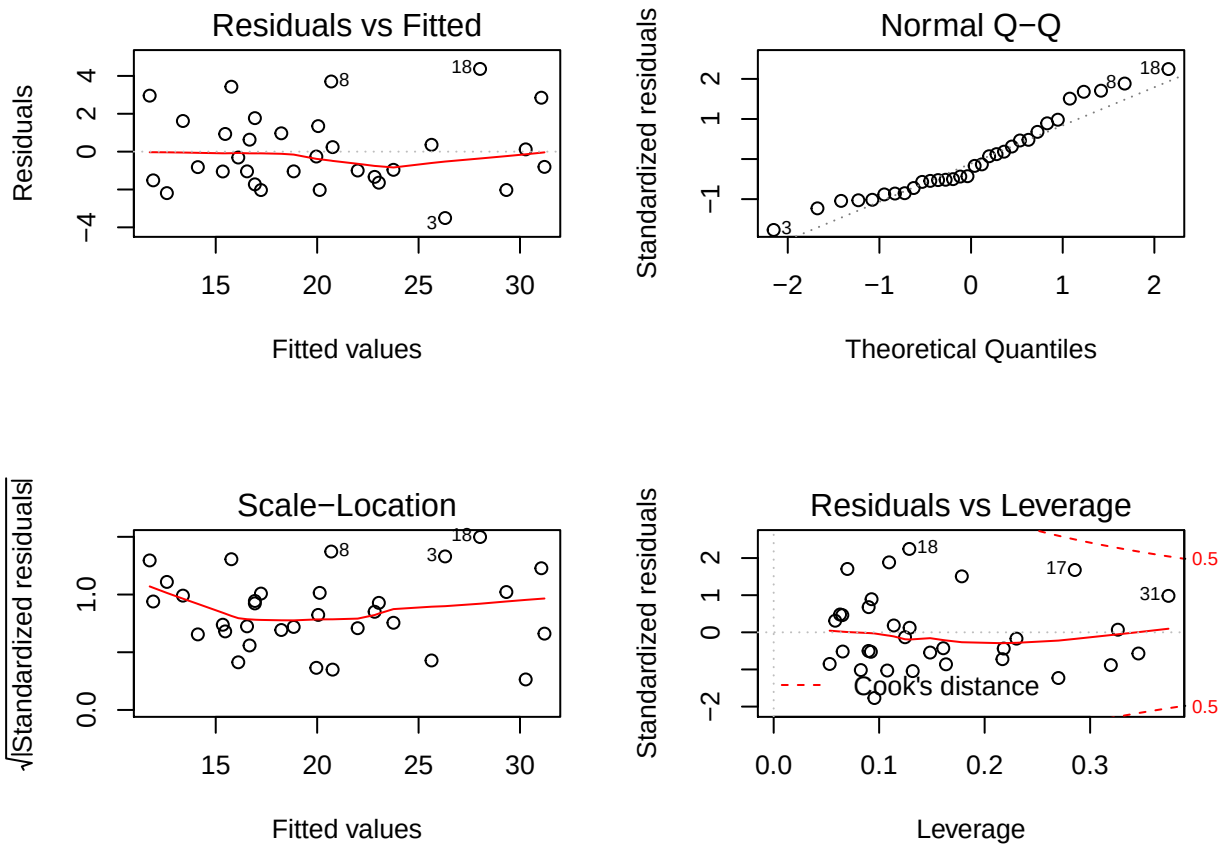
So each entry in dfbHigh was FALSE, i.e., dfbetas(fit4)'s values were all within (-1,1).

```
summary(fit4)$coef
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## am          14.079428  3.4352512  4.098515 0.0003408693
## wt:am       -4.141376  1.1968119 -3.460340 0.0018085763
```

The estimated coefficient for am in fit4 is by far the highest. So we'd expect that going from 0 ("manual") to 1 ("auto") would increase mpg by about 14.1.

PAIRWISE PLOTS






Another representation of the Residuals vs. Leverage figure