*Gian Balsamo*

*October 2015*

## Regression Models Course Project

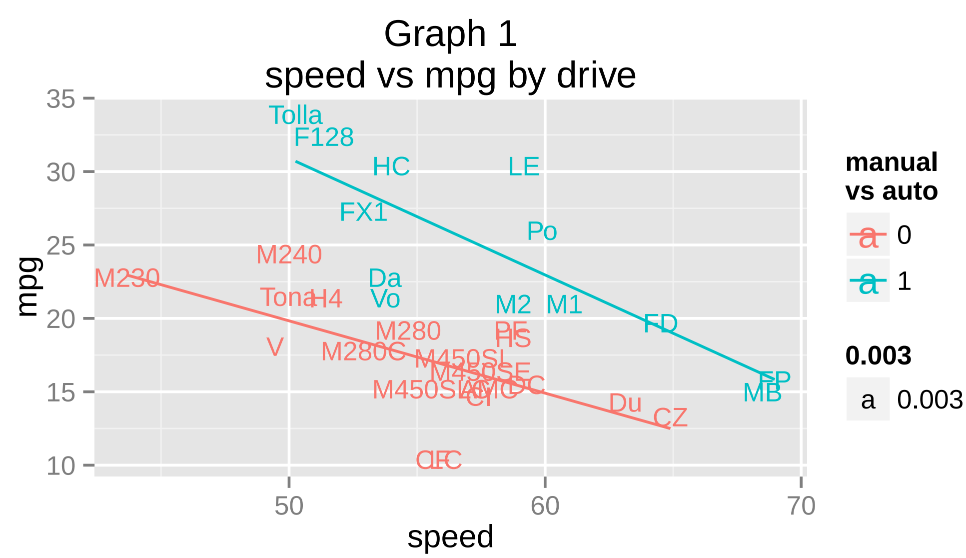
## Summary

Suppose you are given a sample of 10 dark-haired Italians and 10 blond Norwegians, and asked to use this sample to ascertain whether hair color determines nationality. In the present report I am going to disentangle an analogous conundrum. Given fuel consumption and 10 aspects of automobile design for 32 car models from 1973-1974 from the dataset "mtcars", I will show that in the Seventies a manual transmission was better than an automatic one for fuel consumption.

### Report

The key variables in our data set are: miles per gallon (mog), number of cylinders (cyl), total air displaced by the pistons (disp), horsepower (hp), weight (wt), time needed to cover 1/4 mile (qsec), automatic or manual drive (am, a binary variable where 0 stands for automatic and 1 for manual), and several other variables (cyl,vs,gear,carb) whose values may be considered proxies of either horsepower or air displaced. To these I added speed (=1/qsec) and a measure of overpower (= hp/wt) suggested by H. Henderson and P. Velleman in their 1981 "Building Multiple Regression Models Interactively." I also assigned abbreviated names to the 32 cars, as seen in the graphs.

Fuel-consumption-wise, the automatic-drive and manual-drive cars in our dataset clearly belong to distinct population with different mpg means, namely, 17.15 for automatic-drive cars and 24.39 for manual-drive cars. The P-value of the Welch Two Sample t-test rejects the null hypothesis: 0.0013736 But take a look at Graph 1:




When we scatterplot fuel consumption versus speed for automatic-drive cars (red colored) and manual-drive cars (green colored), the graph reminds us of an unrandomized multiple testing. The gap between intercepts is huge and the divergence between slopes slight. Without some drastic purging ont the data, the variable drive would have no predictive value.
Our dataset consists of three distinct groups: nimble, manual-drive speed cars whose mpg and speed-vs-weight ratio benefit from light weight of chassis and bodywork; massive, automatic-drive status cars whose mpg and speed-vs-weight ratio suffer from excessive weight, disp, and hp; subcompact and compact, manual-drive cars whose mpg and speed-vs-weight ratio benefit, just like nimble sport cars, from light weight of chassis and bodywork. In our dataset the expensive sport cars and the compacts and sub-compacts are bunched up together in the same group - which explains the looks of Graph 1. This is what I did. I split the data into 2 separate sets according to drive (manual or automatic). Based on the scatterplots of mpg versus, respectively, wt, disp, speed, and hp (see Graph 2 and Graph 3 in the Appendix), I applied to both datasets a tool of influence/leverage diagnosis, using mpg-vs-hp for manual-drive cars and mpg-vs-weight for automatic-drive cars, and got rid of outliers. I bound the two sets together again, obtaining a smaller dataset (20 rows versus the original 32). Then I went into bootstrapping mode by samplig my new dataset a thousand times with replacement; this provided me with a 1000 x 12 data frame. Then I performed stepwise model selection using backwards elimination. Then I applied the resulting AIC model to my dataset.

```r
practicalCars<-filter(myCars,am=="1")
prestigeCars<-filter(myCars,am=="0")
fitHp<- lm(mpg~hp,data=practicalCars)
hatvalues(fitHp)
practicalCars<-filter(practicalCars,model!="MB",model!="FP",model!="HC")
fitWt<-lm(mpg~wt,data=prestigeCars)
hatvalues(fitWt)
prestigeCars<-filter(prestigeCars, model!="LC" ,  model != "CI" , model != "CF",model!="Tona",model!="M230",mode
practicalCars<-mutate(practicalCars,overpower=practicalCars$hp/practicalCars$wt)
prestigeCars<-mutate(prestigeCars,overpower=prestigeCars$hp/prestigeCars$wt)
data<-rbind(practicalCars,prestigeCars)
vettore<-sample(1:20,1000,replace=T)
automobili<-data.frame()
for (i in vettore) {automobili<-rbind(automobili,data[i,])}
automobiliSmall<-select(automobili,mpg,disp,hp,overpower,am,speed)
full.model <- lm(mpg ~ ., data = automobiliSmall)
step(full.model, direction = "backward")
```

```r
modelFit3<-lm(mpg~disp+speed+am+overpower+hp,data=automobiliSmall)
summary(modelFit3)$coef
```

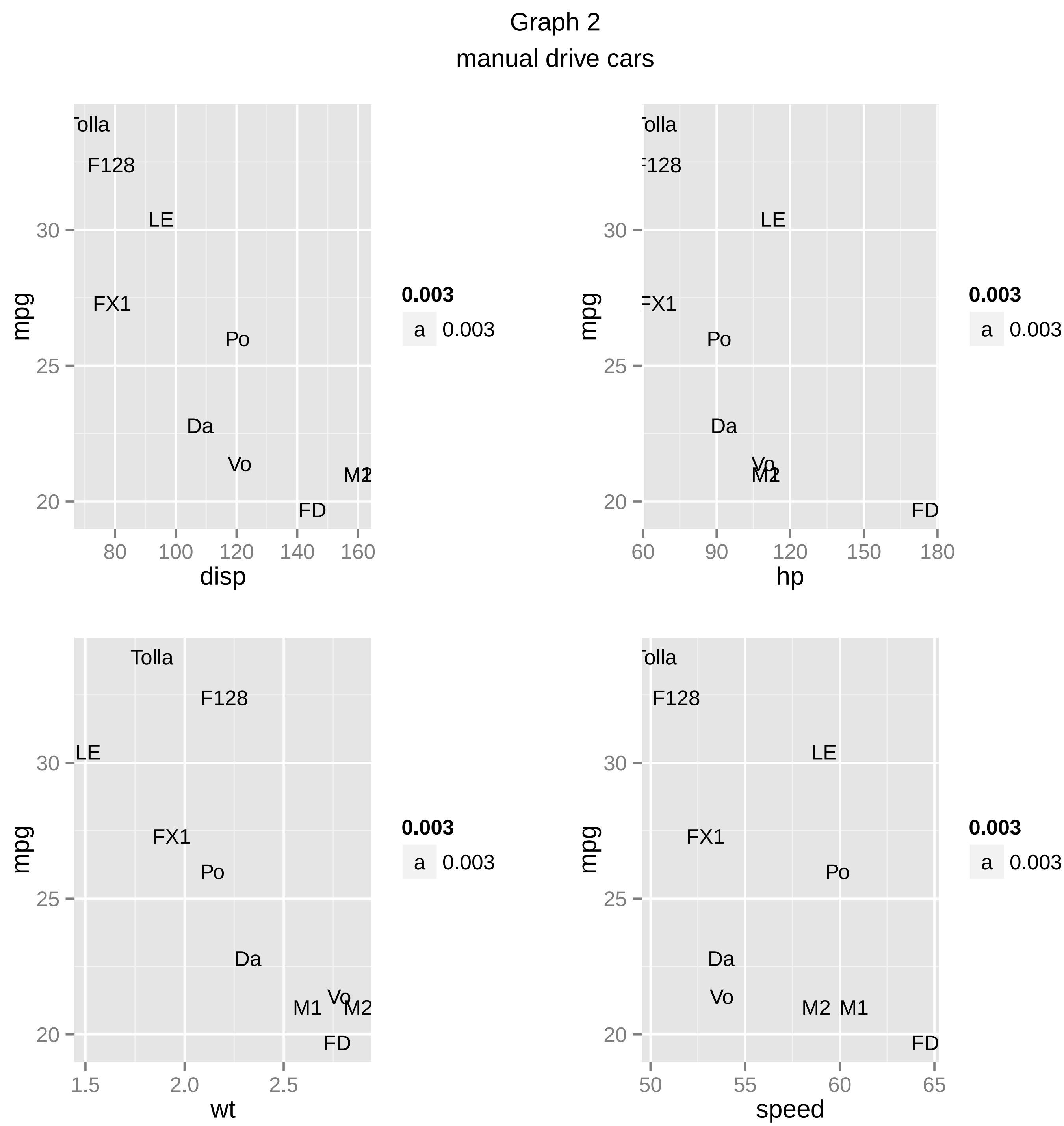
```
##                 Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept) 46.79018471 1.516765867  30.848654 3.935678e-147
## disp         0.01639408 0.002500356   6.556701  8.826485e-11
## speed       -0.58786760 0.039799612 -14.770687  8.768879e-45
## am1          6.30093043 0.469235944  13.428064  6.700563e-38
## overpower    0.27716563 0.012727169  21.777477  2.804135e-86
## hp          -0.08448595 0.004774170 -17.696468  3.952458e-61
```

The AIC model's coefficients are shown above. The adjusted R-squared is 0.8309947. I then ran 4 alternative models using combinations of am, overpower, speed, hp, disp, hp x am, and overpower x am as predictors(models not shown for lack of space), and applied the likelyhood ratio test to each versus the AIC model. The P-values of the Chi-Square distributions confirmed the AIC model as the most effective.

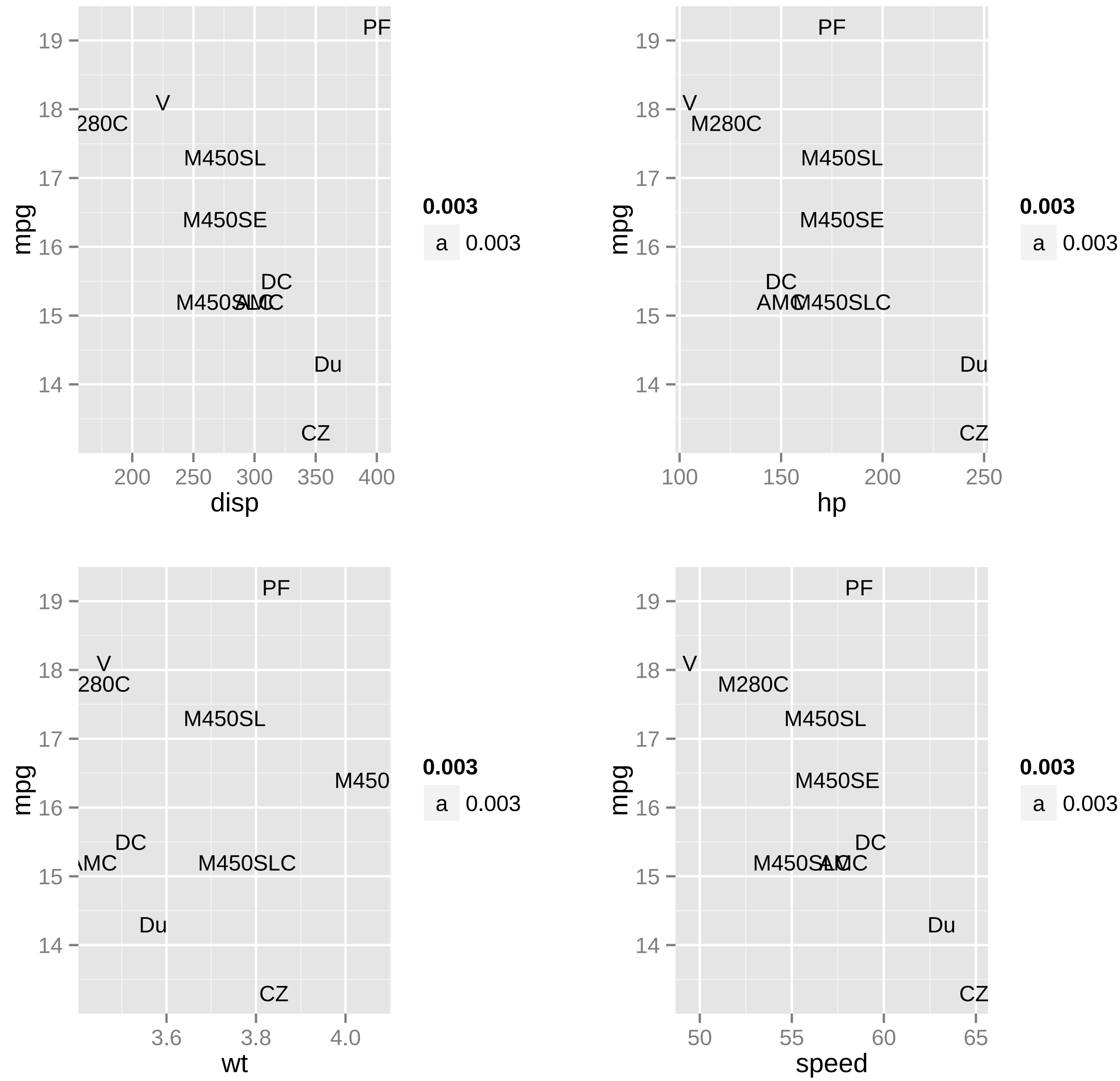
```
## [1]  4.672377e-48  6.412581e-38 5.165963e-165 2.745908e-128
```

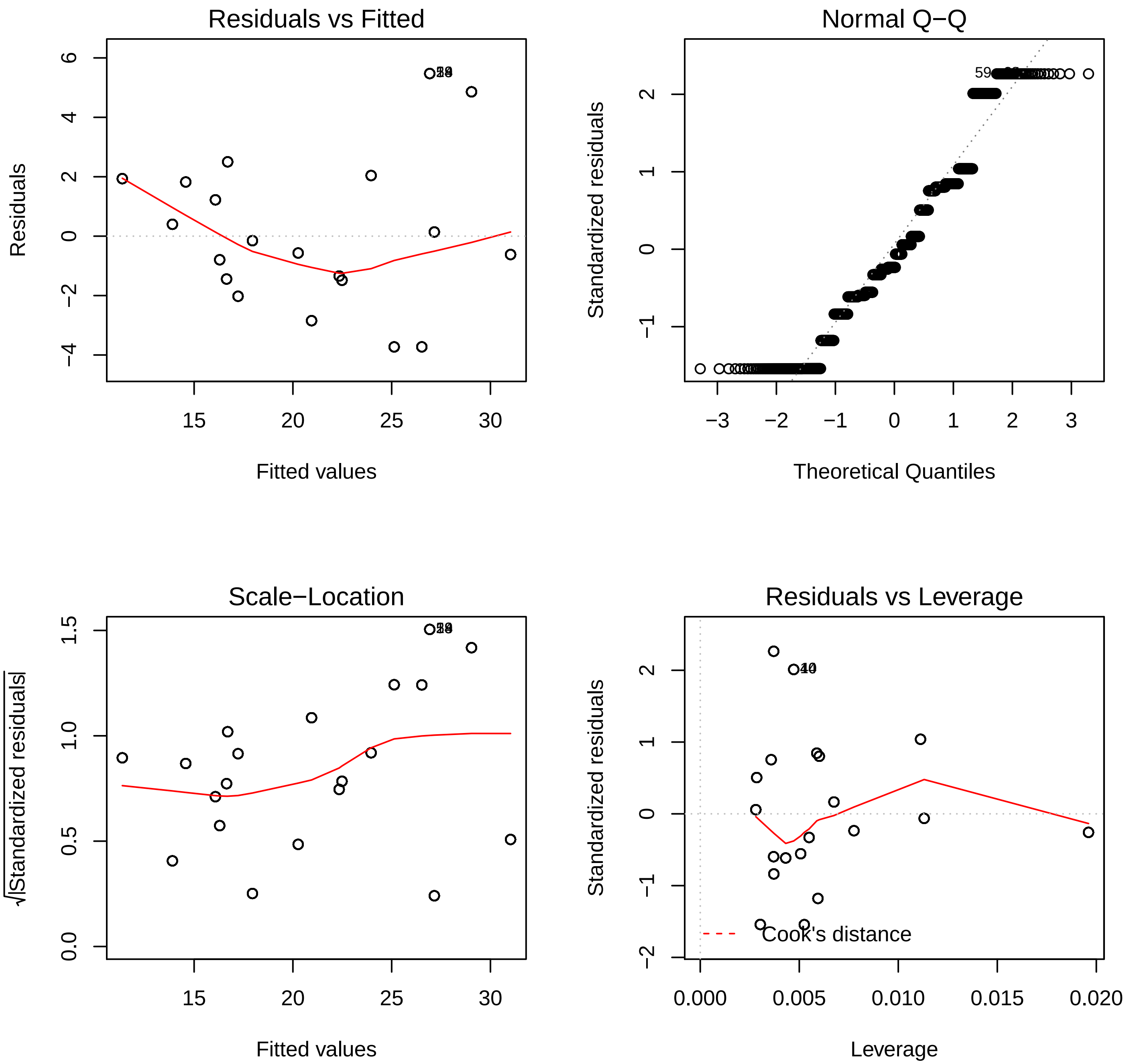
**From the above results I conclude that in the Seventies a manual transmission was better than an automatic one for MPG.**
*My model's intercept, which reflects the effect of automatic drive, is 46.79. tThe coefficient of manual drive is 6.3, which indicates that a manual drive would improve fuel consumption of 6.3 miles per gallon.*

**Appendix**


Graph 2
manual drive cars
Tolla
F128
LE
FX1
Po
Da
Vo
M1
M2
FD
mpg
30
25
20
80
100
120
140
160
disp
0.003
a 0.003
60
90
120
150
180
hp
1.5
2.0
2.5
wt
50
55
60
65
speed

Graph 3
automatic drive cars




Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage