REGRESSION MODELS PROJECT

A.M

Monday, December 8, 2015

Using this dataset `mtcars` of a collection of cars:
1.Is an automatic or manual transmission better for MPG?
2.Quantify the MPG difference between automatic and manual transmissions.

Executive Summary

Using this dataset of 32 cars with 11 variables and forward selection regression modelling, manual transmission is better for fuel efficiency. The final model uses 4 features for prediction: transmission, weight, interaction between transmission and weight, and 1/4 mile time.

Transmission type alone accounts for about 34% variability in fuel efficiency. The number of cylinders which has a good correlation with the miles consumed per gallon (-0.852162) and other predictors in the dataset do not improve the predictive power of the model. The elimination of 3 outliers(Chrysler Imperial, Masserati Bora, and Fiat 128) with high leverage and influnece from the final model increases accuracy of prediction.Any further outlier deletion does not improve the prediction, and not advisable because of the size of the sample.

The MPG difference between manual and automatic transmission after controlling for other variables is **12.5mpg** with 95% confidence interval between **5.56-19.57**

Forward model selection

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 31.416055  3.0201093 10.402291 4.001043e-11
## am          14.878423  4.2640422  3.489277 1.621034e-03
## wt          -3.785908  0.7856478 -4.818836 4.551182e-05
## am:wt       -5.298360  1.4446993 -3.667449 1.017148e-03
```

The significant relationship between transmission and *MPG* is fully mediated and moderated by *weight*. In fig. 2, it is clear that cars with automatic transmission are heavier and have lower MPG.

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

Adjusted r squared 0.8804219

Even though there is a high correlation between the number of cylinders and miles per gallon, it is not statistically significant when added to model 3. See Fig. 3. Also, the addition of other variables in the dataset neither improve the adjusted r squared nor reduce the residual standard error beyond that of model3.

Diagnostics I

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am * wt
## Model 3: mpg ~ am * wt + qsec
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 188.01  2    532.89 61.342 9.089e-11 ***
## 3     27 117.28  1     70.73 16.284  0.000403 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The addition of other variables(weight, weight and transmission interaction, and 1/4 mile time) to `mtcars$am` significantly improves the model. In fig. 4, the errors do not show any significant departure from normality, and there is no systematic variation in the residuals.

```
tail(sort(abs(round(dfbetas(model3)[,2],3))),5)
```

```
##          Merc 240D     Toyota Corolla Cadillac Fleetwood
##              0.342              0.377              0.396
##      Maserati Bora  Chrysler Imperial
##              0.398              0.608
```

```
tail(sort(abs(round(hatvalues(model3),3))),5)
```

```
##   Chrysler Imperial Lincoln Continental        Lotus Europa
##               0.285              0.320               0.326
##            Merc 230       Maserati Bora
##               0.346              0.374
```

```
tail(sort(abs(round(dffits(model3),3))),5)
```

```
##     Toyota Corolla Cadillac Fleetwood      Maserati Bora
##               0.720              0.756              0.758
##           Fiat 128  Chrysler Imperial
##               0.939              1.100
```

Both Chrysler Imperial and Maserati Bora have a high leverage and influence on the regression model. Fiat and Chrysler Imperial exert the highest influnece on the predicted response.

Remove the three cars from the dataset and run model3 again

Run model3 with the filtered dataset `mtcarsNew`

```
##
## Call:
## lm(formula = mpg ~ am * wt + qsec, data = mtcarsNew)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.7972 -1.1895 -0.3781  0.9665  3.7630
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.8451     5.4140   2.742  0.01136 *
## am           12.5648     3.3922   3.704  0.00111 **
## wt           -3.6672     0.6630  -5.531 1.09e-05 ***
## qsec          0.8745     0.2246   3.893  0.00069 ***
## am:wt        -4.1291     1.2145  -3.400  0.00236 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.805 on 24 degrees of freedom
## Multiple R-squared:  0.915,  Adjusted R-squared:  0.9008
## F-statistic: 64.59 on 4 and 24 DF,  p-value: 1.704e-12
```

Compared to Model3, model4's adjusted r squared increases by 2% and the RSE dropped from 2.084 to 1.805

Diagnostics II

In fig. 5, the errors do not show any significant departure from normality and there is no systematic variation in the residuals. Infact, they align better on QQ plot line.

Toyota Corolla seems to affect the predicted response by more than 1 point, but its removal does not affect the model's prediction accuracy.

Using the model4, the estimated difference in Miles Per Gallon between Automatic and Manual Transmission is **12.56** (**5.56-19.57**)

Appendix

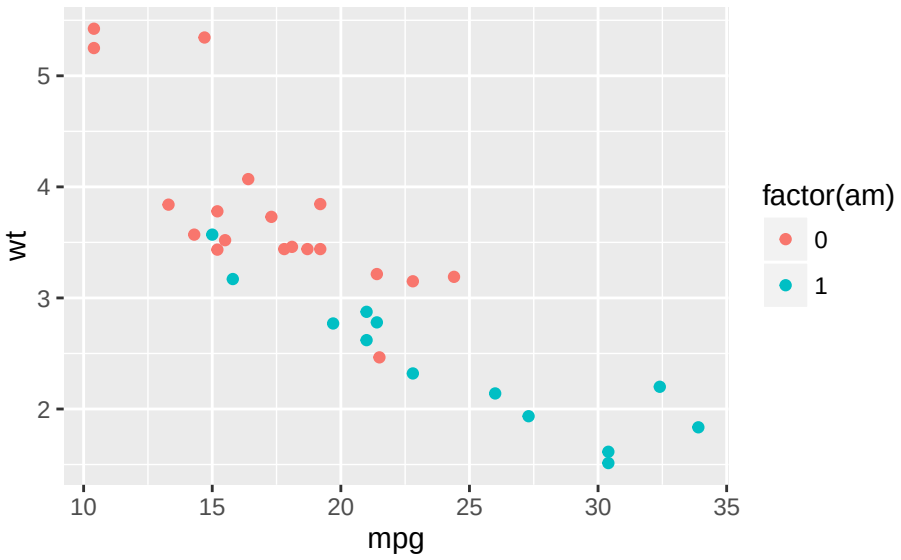

Figure 1: MPG vs Weight by Transmission

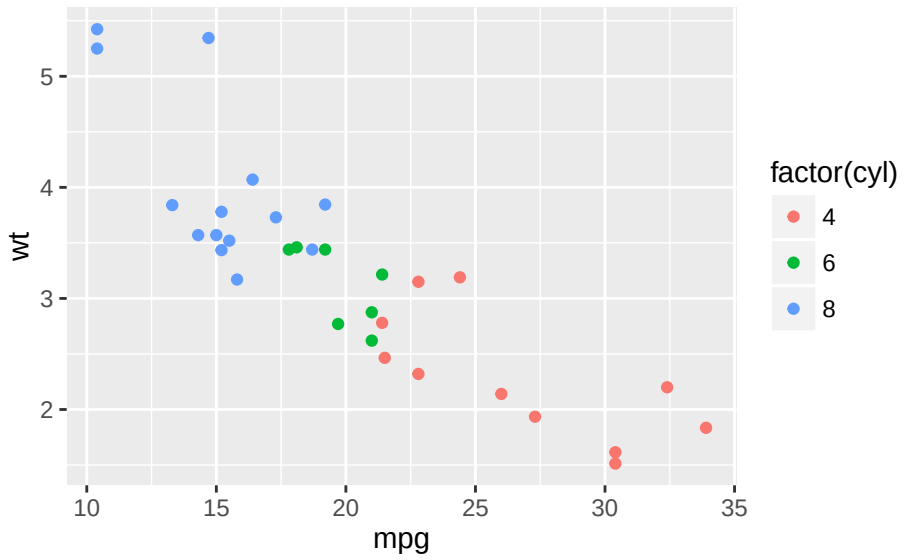

Figure 2: MPG vs weight by No. of Cylinders

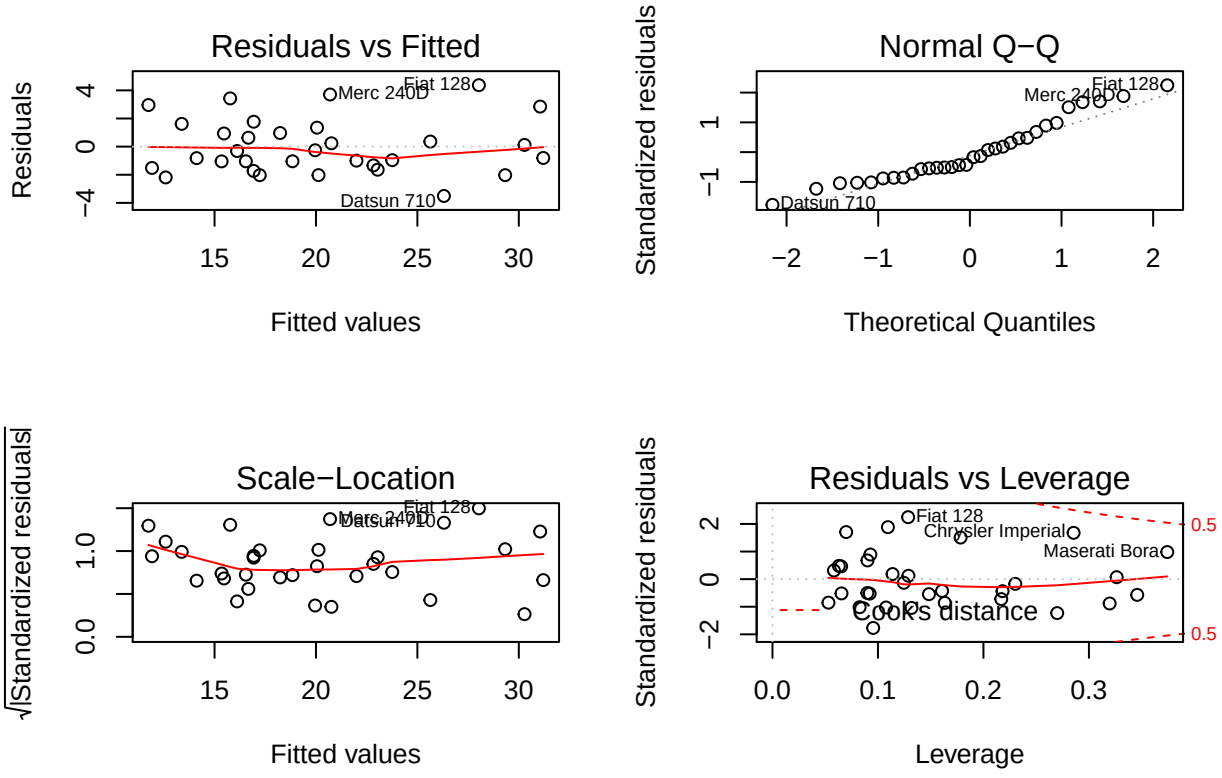

Figure 3: Model 3 diagnostics plot

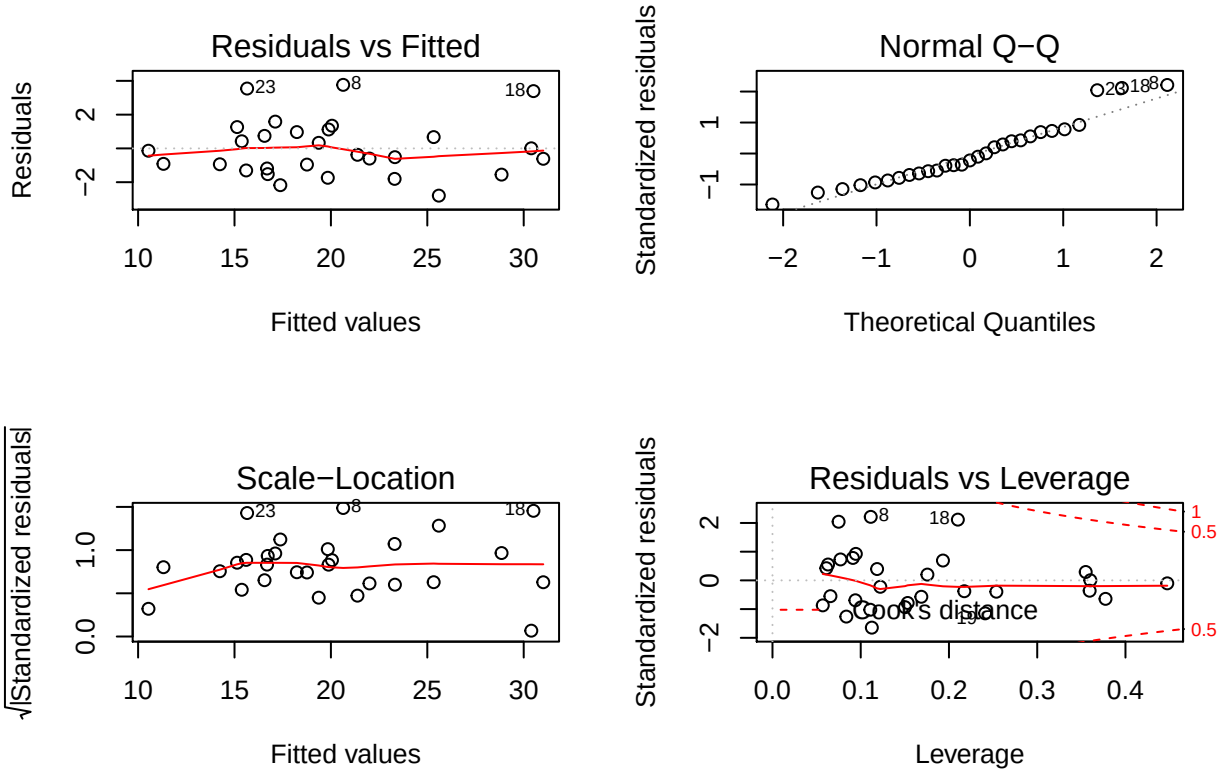

Figure 4: Model4 diagnostics plot