Regression Project - MPG Analysis for Motor Trend

Conrad Wong
November 7, 2015

Executive Summary

This analysis will use the mtcars dataset to analyze if there is a difference in MPG between manual and automatic cars. These are the steps to be followed:

- Data transformation and exploratory data analysis to get a sense of the shape and relationship between the different variables

- Find the regression model that maximizes percentage of explained variance in MPG

- Drive conclusions for two questions:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions

Data transformation and exploratory data analysis

```
library(ggplot2)
library(GGally)
library(car)
library(knitr)

data <- mtcars

data$trans <- as.factor(ifelse(mtcars$am==0, "Auto", "Manual"))
data <- data[, -9]
data$vs <- as.factor(ifelse(data$vs==0, "V", "Straight"))
data$cyl <- as.factor(data$cyl)
data$carb <- as.factor(data$carb)
data$gear <- as.factor(data$gear)
```

Key Findings: The pairs plot (Figure 1 in the Appendix) suggests there might be some colinearity between the variables in the mtcars dataset, as there are some pairs that are correlated, for example:

- disp and weight with corr= 0.89
- disp and hp with corr = 0.79

In addition, the boxplot (Figure 2 in the Appendix) suggests manual cars have higher mpg than automatic cars.

Model selection

Model 1: Fit of mpg by transmission

```
fit.by.trans <- lm(mpg~trans, data=data)
summary(fit.by.trans)$coefficients

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

Key Findings: Fitting by type of transmission, results in a statistical significant coefficient (p-value < 0.05), but the adjusted r2 is only 0.34. Need to look for other variables that help explain a bigger percentage of the variance in mpg. Next step is to fit a model with all variables

Model 2: Fit of mpg by all variables in the mtcars dataset

```
fit.by.all <- lm (mpg~. , data=data)
summary(fit.by.all)$coefficients

##                Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) 25.80998298 20.26412882  1.27367839 0.22216055
## cyl6        -2.64869528  3.04089041 -0.87102622 0.39746642
## cyl8        -0.33616298  7.15953951 -0.04695316 0.96317000
## disp         0.03554632  0.03189920  1.11433290 0.28267339
## hp          -0.07050683  0.03942556 -1.78835344 0.09393155
## drat         1.18283018  2.48348458  0.47627845 0.64073922
## wt          -4.52977584  2.53874584 -1.78425732 0.09461859
## qsec         0.36784482  0.93539569  0.39325050 0.69966720
## vsV         -1.93085054  2.87125777 -0.67247551 0.51150791
## gear4        1.11435494  3.79951726  0.29328856 0.77332027
## gear5        2.52839599  3.73635801  0.67670068 0.50889747
## carb2       -0.97935432  2.31797446 -0.42250436 0.67865093
## carb3        2.99963875  4.29354611  0.69863900 0.49546781
## carb4        1.09142288  4.44961992  0.24528452 0.80956031
## carb6        4.47756921  6.38406242  0.70136677 0.49381268
## carb8        7.25041126  8.36056638  0.86721532 0.39948495
## transManual  1.21211570  3.21354514  0.37718957 0.71131573

vif(fit.by.all)

##             GVIF Df GVIF^(1/(2*Df))
## cyl   128.120962  2        3.364380
## disp   60.365687  1        7.769536
## hp     28.219577  1        5.312210
## drat    6.809663  1        2.609533
## wt     23.830830  1        4.881683
## qsec   10.790189  1        3.284842
## vs      8.088166  1        2.843970
## gear   50.852311  2        2.670408
## carb  503.211851  5        1.862838
## trans   9.930495  1        3.151269
```

Key Findings: Fitting by all the variables in the mtcars dataset increases the adjusted r2 to 0.78, but non of the coefficients is statistical significant (all p-values are greater than 0.05). Looking at the VIF values suggests that there is colinearity between the variables. Next step is to remove the ones with a high GVIF^(1/(2*Df)) value.

Model 3: Fit of mpg by all variables where GVIF^(1/(2*Df)) <= 3.15

```
#            GVIF Df GVIF^(1/(2*Df))
#drat    6.809663  1        2.609533
#vs      8.088166  1        2.843970
#gear   50.852311  2        2.670408
#carb  503.211851  5        1.862838
#trans   9.930495  1        3.151269

fit.by.some <- lm(mpg~ drat + vs + gear + carb + trans, data=data )
summary(fit.by.some)$coefficients

##               Estimate Std. Error    t value   Pr(>|t|)
## (Intercept)  11.115682   6.536428  1.7005744 0.10378643
## drat          3.074948   2.026808  1.5171382 0.14414237
## vsV          -1.483798   2.373431 -0.6251698 0.53859439
## gear4         2.593340   2.934638  0.8837000 0.38686357
## gear5         2.358218   3.568533  0.6608369 0.51590172
## carb2        -1.854541   2.163592 -0.8571583 0.40103386
## carb3        -2.771974   2.905391 -0.9540795 0.35089759
## carb4        -7.217473   2.640080 -2.7338088 0.01243960
## carb6        -5.915280   4.407054 -1.3422299 0.19385353
## carb8       -10.369284   4.427655 -2.3419360 0.02911938
## transManual   2.493866   2.565009  0.9722643 0.34198573

vif(fit.by.some)[,3]^2

##     drat       vs     gear     carb    trans
## 4.321481 5.265793 3.896788 1.517639 6.028156
```

Key Findings: The adjusted r2 decreases to 0.77, and we still have the problem of non statistical significant coefficients. Next step is to fit the model by a stepwise algorithm.

Model 4: Fit of mpg following the stepwise algorithm

```
stepModel <- step(fit.by.all, k=log(dim(data)[1]), trace=FALSE)
summary(stepModel)$coefficients

##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## transManual  2.935837  1.4109045  2.080819 4.671551e-02
```

Key Findings: This approach results in the hightest adjusted r2: 0.83, and it finds three coefficients that are statistical significant (wt, qsec and transmission). Next step is to run some diagnostics for this model

Diagnosis and residuals of stepwise model

Please refer to Figure 3 in the Appendix

Key Findings: There doesn't seem to be a pattern in the residuals, so the model is a good fit. It would be good to work with a subject matter expert to look at the lower and upper residuals in the QQ plot, and understand why they have a slight deviation from the normal distribution.

Conclusion

Is an automatic or manual transmission better for MPG?

- This analysis confirms that manual cars deliver more MPG than automatic cars

Quantify the MPG difference between automatic and manual transmissions

```
coef <- summary(stepModel)$coefficients
coef[4,1] + c(-1, 1) * qt(.975, df = stepModel$df) * coef[4, 2]

## [1] 0.04573031 5.82594408
```

- With 95% confidence, we estimate that manual cars deliver between 0.05 and 5.82 more MPG than automatic cars.

Appendix

Figure 1: GGPairs

```
ggpairs(data[, 2:11], lower=list(continuous="smooth"))
```

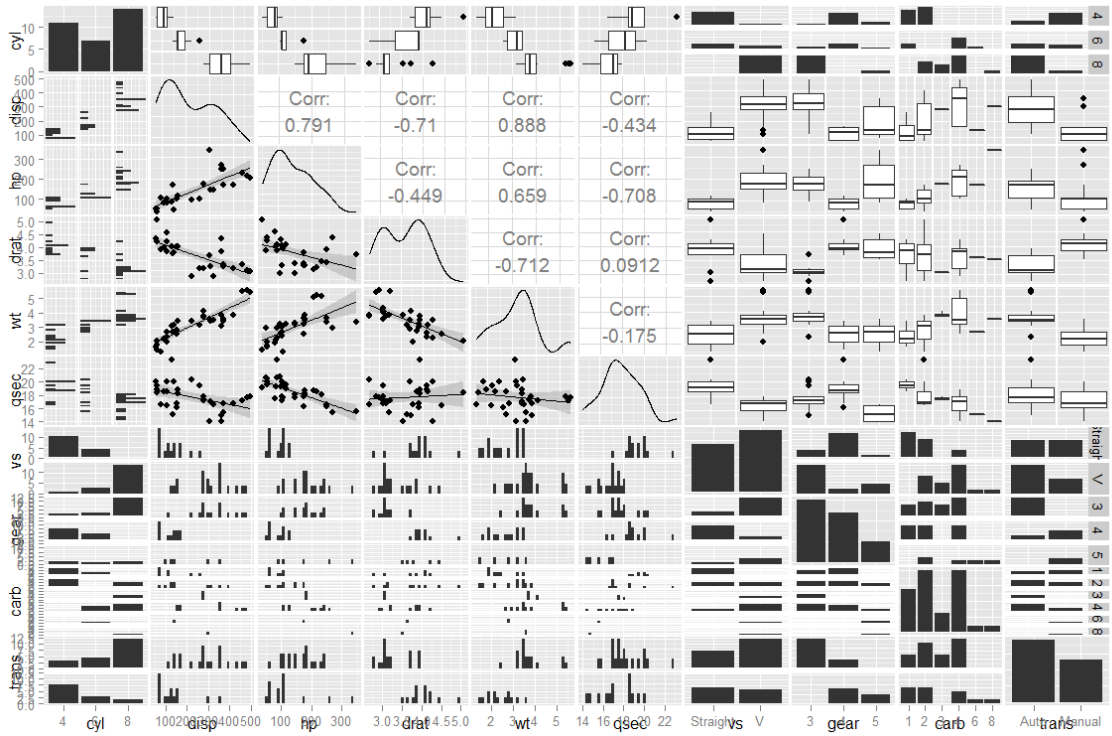

Figure 2: Boxplot of MPG by Transmission

```
plot (mpg~trans, data=data)
```

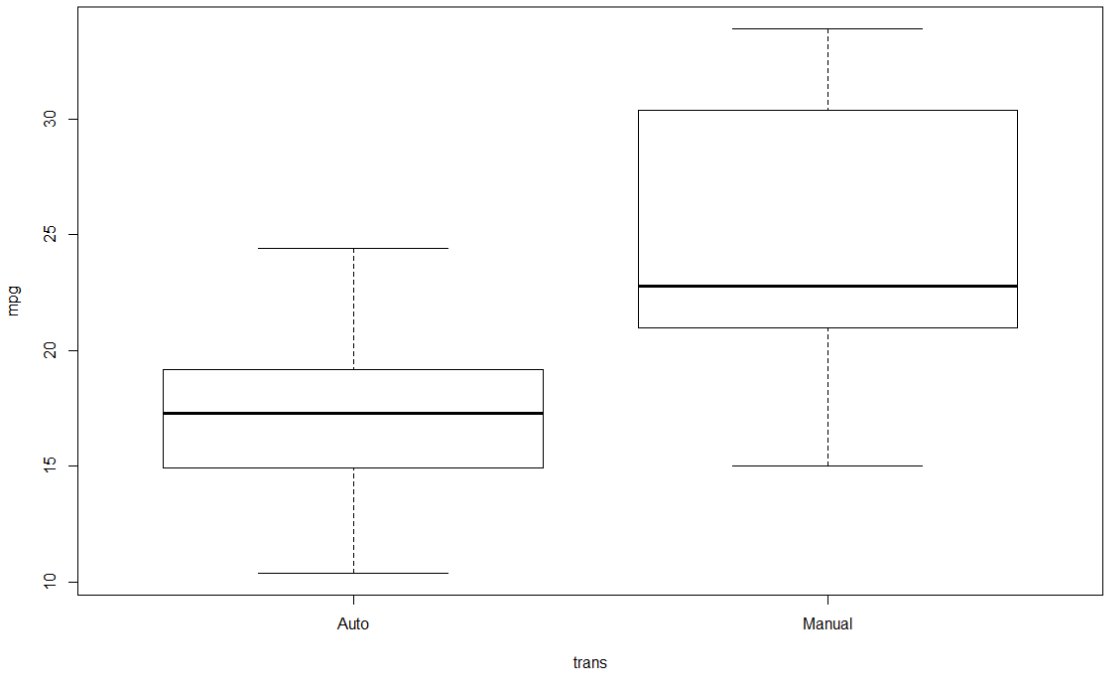

Figure 3: Residuals