Regression Models Course Project

Michael Atkins

Tuesday, March 10, 2015

EXECUTIVE SUMMARY

This report was undertaken to examine two questions for Motor Trend magazine: 1) Is an automatic or manual transmission better for MPG, and 2) What is the quantifaction of MPG difference between automatic and manual transmissions? In summary, I determined that an automboile with manual over automatic transmission provides a statistically significant improvement in MPG by 1.48 MPG when accounting for the effects of other independent influencing variables. All figures described herein are included in the Appendix per the course project instructions. Note that no units were converted during this analysis.

EXPLORING THE DATA

I initially summarized the regression of the dependent MPG variable by the independent AM factor per below and noted that it appeared that MPG and AM (transmission) were correlated with an adjusted R-squared value of 33.9% and a P-value of > 0.001. I noted that the mean of the MPG for the automatic transmission (AM = 0) was 17.15 MPG while that for the manual transmission (AM = 1) was 24.4 MPG, which infers a large difference. The 95% confidence interval of -11.3 to -3.2 obtained from the T Test strongly suggets significance between AM and MPG. See "Appendix 2"

I then explored the data utilizing a collorized scatterplot matrix and determined that the biggest determinates of MPG overall other than transmission were HP, CYL, DISP, and WT, which makes sense empirically. See "Appendix 3"

I further explored the relationships between the HP, CYL, DISP, and WT variables utilizing regression regression models, plotting each model with regression lines overlayed for manual transmission (blue) and automatic transmission (green). See "Appendix 4"

FITTING THE REGRESSION MODELS

After initial data exploration, I began to generate and attempt to fit regression models, starting with HP, CYL, DISP, and WT along with AM as the independent variables. See "Appendix 5"

Then I summarized the fit analysis of the variance of the base model, noting that HP, CYL, and WT appeared to be the primary determining variables other than AM. See "Appendix 6"

The model was then refined, removing the DISP variable which ultimately offered a better fit. See "Appendix 7"

I note that the three remaining independent variables, HP, CYL, and WT may potentially be dependent upon one another and the model could be further refined, however, the three variable model has an adjusted R-Squared value of 82.7% and a significant P-Value. I then generated the final regression model. Cars with a manual transmission appear to have higher MPG by 1.48 when HP, CYL, and WT variables are held constant. See "Appendix 8"

ANALYZING THE FIT OF THE MODEL

Finally, I performed an ANOVA anlaysis of the base regression model and the final regression model below. Note that the high F-value of 29.2 and low P-value close to 0 indicating this model is a better fit than relying on tranmission type as the sole independent variable affecting MPG. See "Appendix 9"

PERFORMING MODEL DIAGNOSTICS

To close, I performed a diagnostic on the model below. The model developed above assumes normality and constant variance for the model error term. This diagnostic shows those assumptions are valid. Viewing the diagnostic plots, I note: Residuals vs Fitted - Points are scattered and there appear to be no significant

nonconstant variance (heteroscedasticity) and no nonlinearity. Normal Q-Q - Significant departures from the line suggest violations of normality. No such departures are noted here suggesting univariate normality. Scale-Location - Points are scattered in a "constant band" pattern suggesting constant variance. Residuals vs Leverage - There appear to be some outliers in this plot which may deserve further exploration. See "Appendix 10"

APPENDIX

Appendix1

```
require(ggplot2); require(stats); require(graphics); require(gclus)
```

```
## Loading required package: ggplot2
## Loading required package: gclus
## Loading required package: cluster
```

```
df <- mtcars
```

Appendix 2

```
summary(lm(mpg~am,df))
```

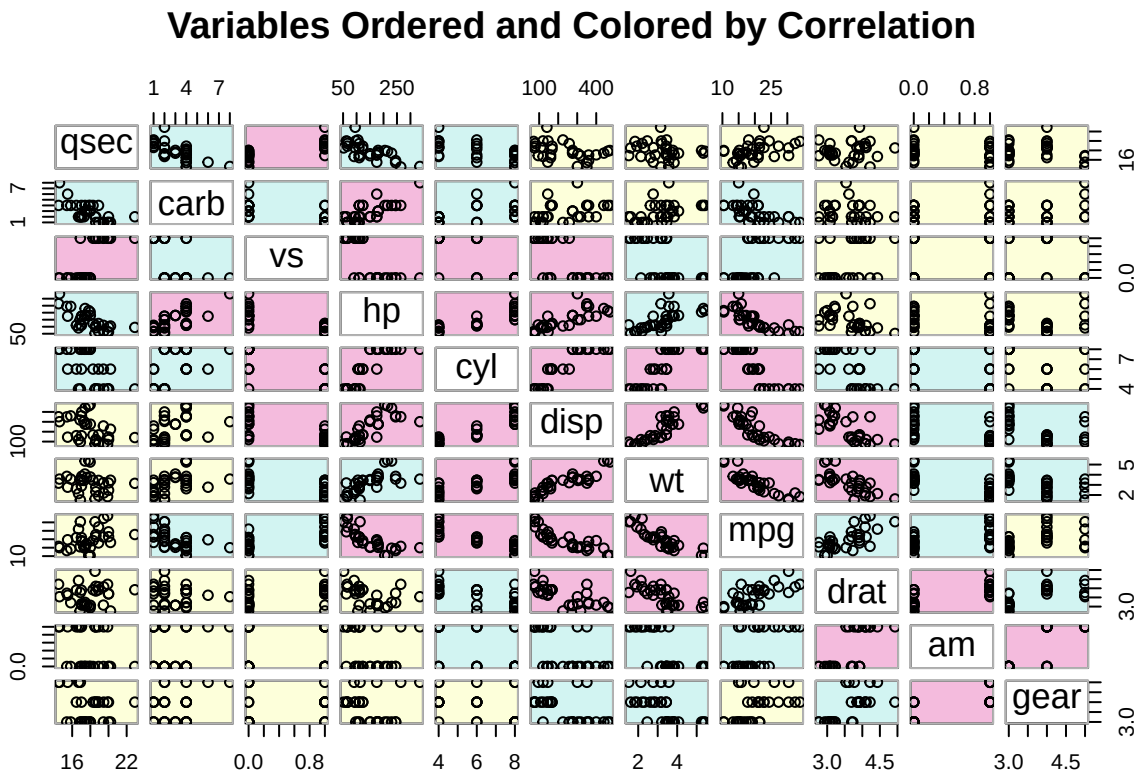
```
##
## Call:
## lm(formula = mpg ~ am, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```
t.test(mpg~am, data=df) ##
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Appendix 3

```
dta <- df[c(1,2,3,4,5,6,7,8,9,10,11)] ## explore data to determine main factors in mpg other than transm
dta.r <- abs(cor(dta))
dta.col <- dmat.color(dta.r)
dta.o <- order.single(dta.r)
cpairs(dta, dta.o, panel.colors=dta.col, gap=.5,
       main="Variables Ordered and Colored by Correlation" )
```

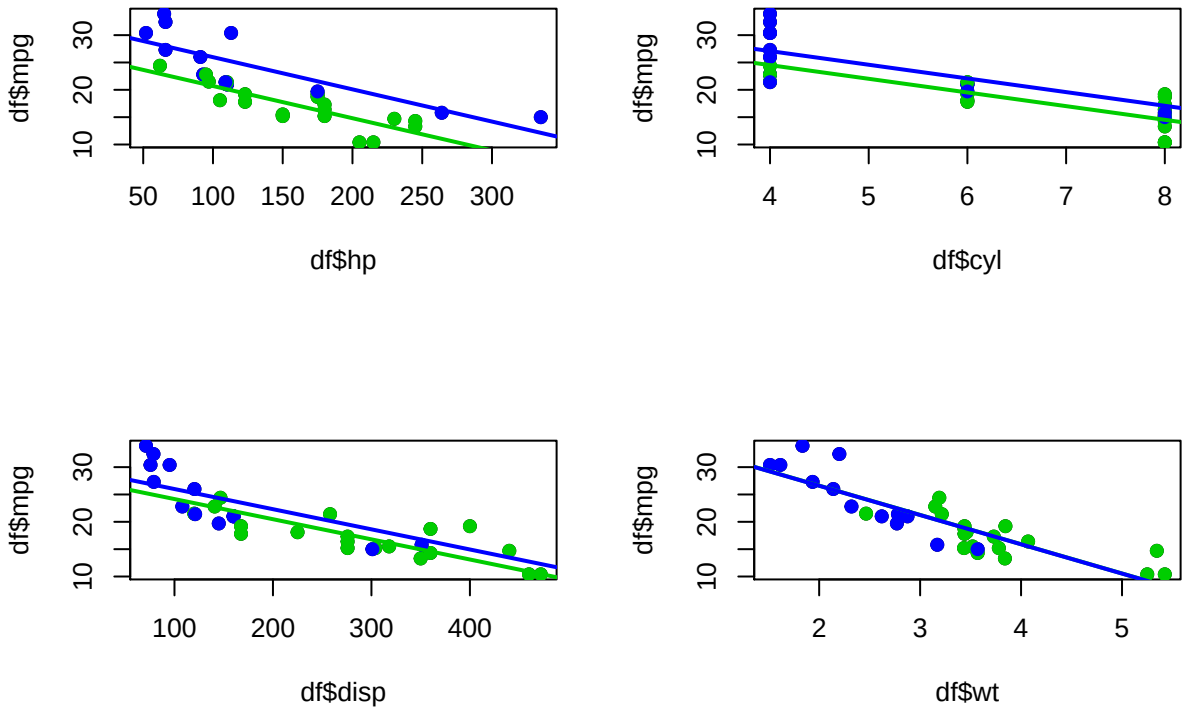


3-1.pdf

Appendix 4

```
lmbothcyl <- lm(mpg~cyl+am,df) ## regression of mpg by cylinder and transmission
lmbothdisp <- lm(mpg~disp+am,df) ## regression of mpg by displacement and transmission
lmbothhp <- lm(mpg~hp+am,df) ## regression of mpg by horsepower and transmission
lmbothwt <- lm(mpg~wt+am,df) ## regression of mpg by weight and transmission

par(mfrow=c(2,2))
## plot mpg by horsepower with regression by transmission
plot(df$hp,df$mpg,pch=19)
points(df$hp,df$mpg,pch=19,col=((df$am=="1")+3))
abline(c(lmbothhp$coeff[1],lmbothhp$coeff[2]),col="green3",lwd=2)
abline(c(lmbothhp$coeff[1] + lmbothhp$coeff[3],lmbothhp$coeff[2]),col="blue",lwd=2)

## plot mpg by cylinder with regression by transmission
plot(df$cyl,df$mpg,pch=19)
points(df$cyl,df$mpg,pch=19,col=((df$am=="1")+3))
```

```
abline(c(lmbothcyl$coeff[1],lmbothcyl$coeff[2]),col="green3",lwd=2)
abline(c(lmbothcyl$coeff[1] + lmbothcyl$coeff[3],lmbothcyl$coeff[2]),col="blue",lwd=2)

## plot mpg by displacement with regression by transmission
plot(df$disp,df$mpg,pch=19)
points(df$disp,df$mpg,pch=19,col=((df$am=="1")+3))
abline(c(lmbothdisp$coeff[1],lmbothdisp$coeff[2]),col="green3",lwd=2)
abline(c(lmbothdisp$coeff[1] + lmbothdisp$coeff[3],lmbothdisp$coeff[2]),col="blue",lwd=2)

## plot mpg by weight with regression by transmission
plot(df$wt,df$mpg,pch=19)
points(df$wt,df$mpg,pch=19,col=((df$am=="1")+3))
abline(c(lmbothwt$coeff[1],lmbothwt$coeff[2]),col="green3",lwd=2)
abline(c(lmbothwt$coeff[1] + lmbothwt$coeff[3],lmbothwt$coeff[2]),col="blue",lwd=2)
```



4-1.pdf

Appendix 5

```
summary(lm(mpg ~ hp + cyl + disp + wt + am, df)) ## summarize base model
```

```
##
## Call:
## lm(formula = mpg ~ hp + cyl + disp + wt + am, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## hp          -0.02796    0.01392  -2.008  0.05510 .
## cyl         -1.10638    0.67636  -1.636  0.11393
## disp         0.01226    0.01171   1.047  0.30472
## wt          -3.30262    1.13364  -2.913  0.00726 **
## am           1.55649    1.44054   1.080  0.28984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

Appendix 6

```
summary(aov(mpg ~ hp + cyl + disp + wt + am, data = df)) ## summarize fit of analysis of base variance r
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## hp           1  678.4   678.4 108.127 9.34e-11 ***
## cyl          1  155.7   155.7  24.817 3.53e-05 ***
## disp         1   30.6    30.6   4.878 0.036216 *
## wt           1   90.9    90.9  14.493 0.000772 ***
## am           1    7.3     7.3   1.167 0.289843
## Residuals   26  163.1     6.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 7

```
summary(aov(mpg ~ hp + cyl + wt + am, data = df)) ## summarize fit of analysis of variance model
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## hp           1  678.4   678.4 107.743 6.34e-11 ***
## cyl          1  155.7   155.7  24.729 3.28e-05 ***
## wt           1  115.4   115.4  18.321  0.00021 ***
## am           1    6.6     6.6   1.052  0.31418
## Residuals   27  170.0     6.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 8

```
summary(lm(mpg ~ hp + cyl + wt + am, df)) ## summarize regression
```

```
##
## Call:
## lm(formula = mpg ~ hp + cyl + wt + am, data = df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## hp          -0.02495    0.01365  -1.828   0.0786 .
## cyl         -0.74516    0.58279  -1.279   0.2119
## wt          -2.60648    0.91984  -2.834   0.0086 **
## am           1.47805    1.44115   1.026   0.3142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

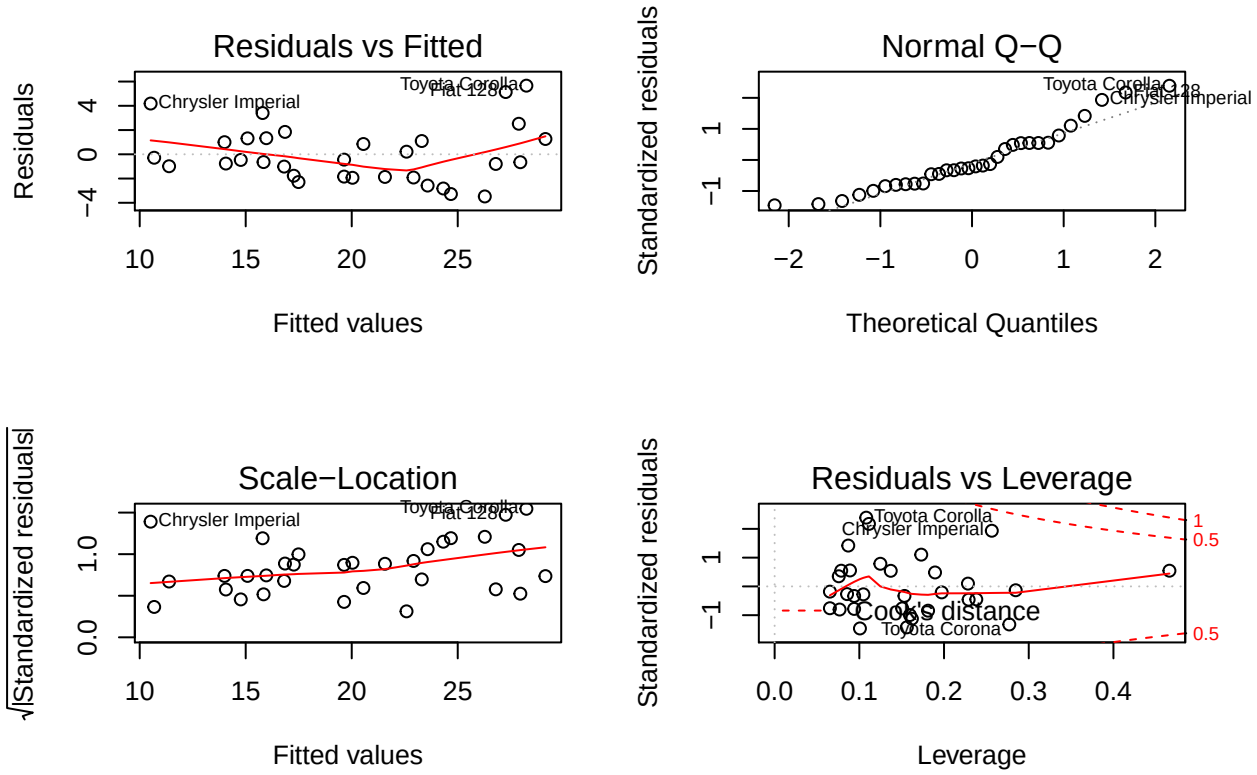
Appendix 9

```
anova(lm(mpg ~ am, df), lm(mpg ~ hp + cyl + wt + am, df)) ## (high f value and low p value)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ hp + cyl + wt + am
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.9
## 2     27 170.0  3     550.9 29.166 1.274e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 10

```
par(mfrow = c(2, 2))
plot(lm(mpg ~ hp + cyl + wt + am, df))
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
Leverage
Toyota Corona

10-1.pdf