Coursera – Regressions: *Course Project*

Gordon Silvera, 6/21/2015

EXECUTIVE SUMMARY

- All else equal, manual transmission cars gets 5.3 mpg more than automatic.
- For my primary models, I considered the relationship between MPG, transmission and horsepower as well as that between MPG, transmission and car weight. I decided to use the relationship between MPG and horsepower in my final model because it provides a more robust relationship amongst the regression variables – both in terms of R^2 and homoscedasticity.
- I have considered adjustments and interactions between horsepower and other factor variables (as they relate to mileage). However, these additional regressors do not have a significant impact on the model, therefore I have left them out of the model. I ultimately used the following model to derive my final conclusion:

```
model3 <- lm(mpg ~ am + hp)
```

EXPLORATORY DATA ANALYSIS

To gain a foundational understanding of the data, I plotted all binary and factor variables against MPG (**Figure 1**). We can see that cars with manual transmission have an average 24.4 MPG versus 17.1 for automatic (**Figure 2**). From there, I tabulated the correlation of all variables against one another using the following code (**Figure 3**):

```
var <- mtcars[1:11]
round(cor(var, var),3)
```

I also plotted the numerical variables against one another to visualize their relationship. The results to the output can be found in **Figure 4** in the appendix.

```
pairs(~mpg + disp + hp + drat + wt + qsec, data=mtcars, main="Cars Scatterplot Matrix")
```

After reviewing the relationships of the variables, I decided to consider either weight or horsepower as the primary explanatory variables in the regression. Both of these variables have strong inverse relationships with MPG. Also, because there are highly correlated with one another (0.659), we should probably not include both variables in the regression model. I later confirmed this by testing variance inflation with a model that contained both indicators.

From here, I wanted to determine whether my models required any adjustments or interactions. I used the code in **Figure 5** to visually determine this. Note that I replicated this code for both "hp" and "wt" against all factor variables (however I've only included one output to conserve space). I concluded that the relationship between mpg ~ hp grouped by "vs" indicated that V/S may have an interaction with horsepower (**Figure 5**).

MODEL SELECTION

Armed with the knowledge I previously mentioned, I decided to test the following regression models.

```
model1 <- lm(mpg ~ am + wt)
model2 <- lm(mpg ~ am + wt + qsec)
model3 <- lm(mpg ~ am + hp)
model4 <- lm(mpg ~ am + wt + hp)
model5 <- lm(mpg ~ am + hp + hp : factor(vs))
```

Model 1. This model suggests that – with a P-value of 0.988 – there is not a significant difference between automatic and manual transmission. When adding factor variables, they either do not have an impact on the results and/or are not significant. This produces a reasonable R^2 of 0.753), however when plotting the residuals, we see a pattern in the residuals vs fitted values plot (Note: to conserve space, I have only included residual plots on my final model). Therefore I have decided not to use this model.

Model 2. When I added qsec to the model, I received the output below. This suggests that manual cars get 2.9 MPG more than transmission when accounting for the cars weight and speed (defined here as its ¼ mile time). With a P-value of 0.047, we can (barely) reject the null hypothesis and say that manual outperforms automatic in mileage. This resulted in an R^2 of 0.850. With the max variance inflation factor of 2.54, I believe this variance inflation does not negatively affect this model. However, when plotting the residuals, we observe a slight pattern (albeit less so than in model 1). For this reason, I will consider other models.

Model 3 (Final Model). For this model, I first considered the relationship between mileage and transmission type, holding horsepower constant. This produces a coefficient for transmission of 5.28, suggesting that manual cars get 5.28 MPG more than automatic (**Figure 6**). Although the R^2 of 0.782 is less than some of the other models, the plot of residuals shows no evidence of heteroscedasticity (**Figure 7**). Therefore, I have used this as my final model.

Note that I tested other factor variables and interactions to this model, however they either did not impact the results and/or were not significant. I believe that **Model 5** (written above) – that considers the interaction between horsepower and V/S – may have been the next best candidate for a final model. However, the interaction variables are not significant, therefore I have chosen the most simplistic model possible.

In addition to plotting the residuals, I have checked the variance inflation of the explanatory variables (**Figure 8**) and checked for high leverage observations (**Figure 9**). While we observe low variance inflation, there are some high leverage outliers that should be removed (as indicated by the plot in **Figure 9**).

Model 4. I decided to add weight to model 3 to see whether this would provide a better indicator of the relationship between mileage and transmission type. Although this gives us an R^2 of 0.840, there is a high level of variance inflation and collinearity in this model due to the correlation between weight and horsepower. Therefore I have thrown this model out.

INTERPRETATIONS & CONCLUSIONS

In conclusion, we have determined the following:

- Manual transmissions are outperform automatic at a statistically significant level, all other factors equal. This is determined by the P-value of 3.46 x 10^-5 on the "am" coefficient.
- On average, manual transmissions outperform automatic by 5.27 MPG, as determined by the coefficient of variable "am".
- We are also 95% confident that the increase in MPG for manual cars is between 3.07 and 7.85 (as calculated in **Figure 10**).

Thank you for grading my paper, and apologies this isn't in knitr (I was having technical difficulties)!

APPENDIX

These are the primary models I have considered.

```
model1 <- lm(mpg ~ am + wt)
model2 <- lm(mpg ~ am + wt + qsec)
model3 <- lm(mpg ~ am + hp)
model4 <- lm(mpg ~ am + wt + hp)
```

Figure 1. Box Plot of Factor & Binary Variables.

```
> boxplot(mpg~cyl,data=mtcars, main="MPG ~ # Cylinders", xlab="Number of Cylinders", ylab="Miles
Per Gallon")
> boxplot(mpg~vs,data=mtcars, main="MPG ~ V/S", xlab="V/S", ylab="Miles Per Gallon")
> boxplot(mpg~am,data=mtcars, main="MPG ~ Transmission", xlab="Trans Type", ylab="Miles Per Gallo
n")
> boxplot(mpg~gear,data=mtcars, main="MPG ~ # Gears", xlab="# Gears", ylab="Miles Per Gallon")
```

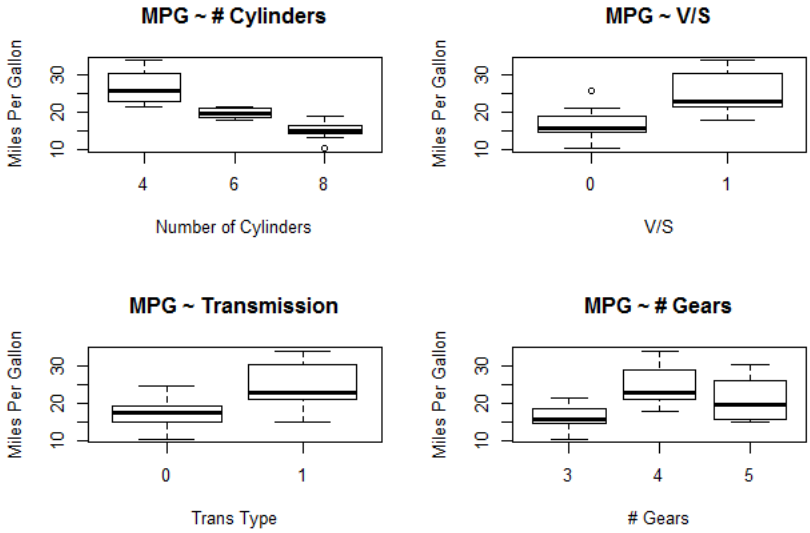

Figure 2. Determine Mean MPG for Transmission Type.

```
aggregate(mtcars$mpg, by=list(mtcars$am), FUN=mean)
  Group.1        x
1       0 17.14737
2       1 24.39231
```

Figure 3. Correlation Matrix of All Variables

```
var <- mtcars[1:11]

round(cor(var, var),3)
```

```
        mpg    cyl   disp     hp   drat     wt   qsec     vs     am   gear   carb
mpg   1.000 -0.852 -0.848 -0.776  0.681 -0.868  0.419  0.664  0.600  0.480 -0.551
cyl  -0.852  1.000  0.902  0.832 -0.700  0.782 -0.591 -0.811 -0.523 -0.493  0.527
disp -0.848  0.902  1.000  0.791 -0.710  0.888 -0.434 -0.710 -0.591 -0.556  0.395
hp   -0.776  0.832  0.791  1.000 -0.449  0.659 -0.708 -0.723 -0.243 -0.126  0.750
drat  0.681 -0.700 -0.710 -0.449  1.000 -0.712  0.091  0.440  0.713  0.700 -0.091
wt   -0.868  0.782  0.888  0.659 -0.712  1.000 -0.175 -0.555 -0.692 -0.583  0.428
qsec  0.419 -0.591 -0.434 -0.708  0.091 -0.175  1.000  0.745 -0.230 -0.213 -0.656
vs    0.664 -0.811 -0.710 -0.723  0.440 -0.555  0.745  1.000  0.168  0.206 -0.570
am    0.600 -0.523 -0.591 -0.243  0.713 -0.692 -0.230  0.168  1.000  0.794  0.058
gear  0.480 -0.493 -0.556 -0.126  0.700 -0.583 -0.213  0.206  0.794  1.000  0.274
carb -0.551  0.527  0.395  0.750 -0.091  0.428 -0.656 -0.570  0.058  0.274  1.000
```

Figure 4. Numeric Variable Scatterplot Matrix.

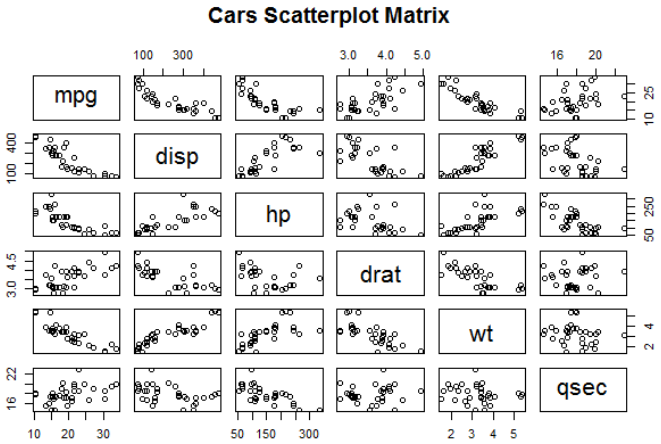

Figure 5. Scatter Plots Grouped by Factor Variables.

```
scatterplot(mpg ~ hp | vs, data=mtcars,
+             xlab="Horsepower", ylab="Miles Per Gallon",
+             main="Enhanced Scatter Plot",
+             labels=row.names(mtcars))
```

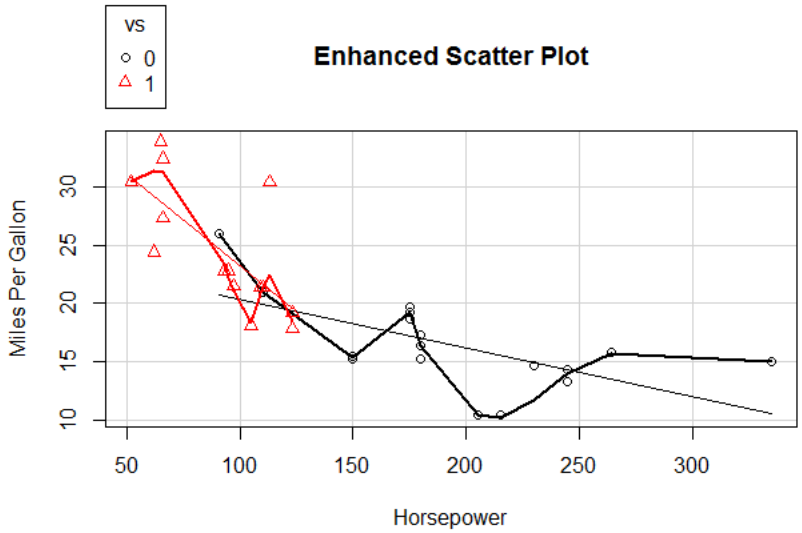

Figure 6. Model 1 Outputs.

```
summary(model3)
Call:
lm(formula = mpg ~ am + hp)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3843 -2.2642  0.1366  1.6968  5.8657

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.584914   1.425094  18.655  < 2e-16 ***
am           5.277085   1.079541   4.888 3.46e-05 ***
hp          -0.058888   0.007857  -7.495 2.92e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.909 on 29 degrees of freedom
Multiple R-squared:  0.782,       Adjusted R-squared:  0.767
```

```
F-statistic: 52.02 on 2 and 29 DF,  p-value: 2.55e-10
```

Figure 7. Plotting Residuals of Model 1

```
data(mtcars); par(mfrow = c(2, 2)); plot(model3)
```

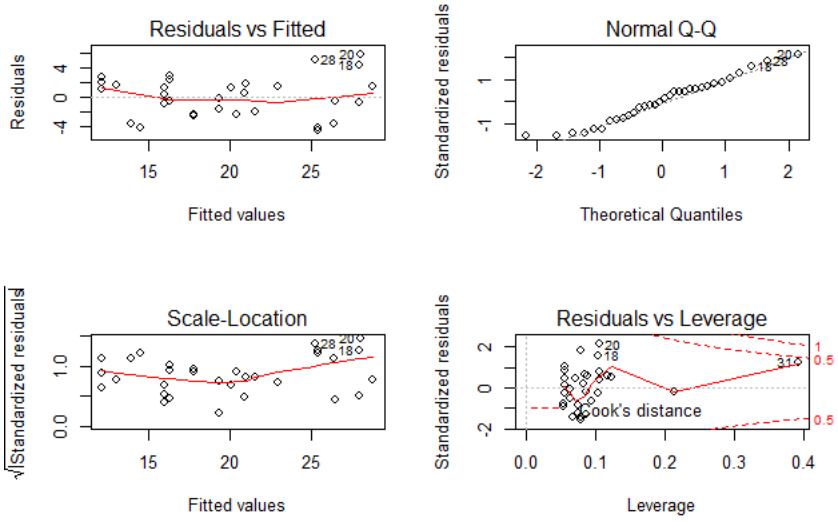

Figure 8. Variance Inflation Factor

```
vif(model3)
      am       hp
1.062867 1.062867
```

Figure 9. Leverage Plots for Model 3.

```
leveragePlots(model3)
```

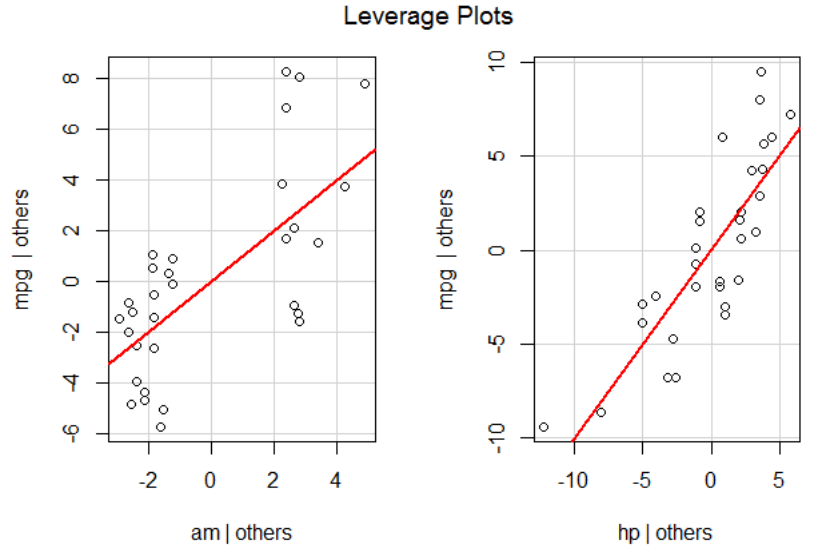

Figure 10. 95% Confidence Interval of Transmission Type in Model 3.

```
confint(model3, 'am', level=0.95)
      2.5 %   97.5 %
am 3.069177 7.484994
```