Regression Models Final Project

Robert Haase

29 1 2017

Summary

A manual transmission seems to be better for gas mileage. Only regressing *mpg* on *am* shows an average increase of approx. 7mpg switching from automatic to manual transmission. This effect is lowered to about 2mpg when adding additional independent variables like displacement. Adding addtional well-selected variable to the regression proved to be statistically significant and increased the R-squared value by 50%.

Detailed Description of the Analysis

First, I took a look at the distribution of mpg differentiated by the *am* variable (manual or automatic transmission). The first two plots in the Appendix section clearly show that cars with manual transmission have a higher gas mileage than cars with automatic transmission.

```
## $automatic
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
##
## $manual
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

The summary table shows that the *median mpg* for manual cars is 22.8 and for automatic cars only 17.3. With this insight, the second step of my analysis was to perform a simple linear regression with **mpg** as the dependent variable and **am** as the only independent variable. The model summary follows next.

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## 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 ***
## ammanual       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
```

The the beta0 coef of the intercept corresponds to the mean mpg of all cars with automatic transmission, which is the same value that was shown in the previous summary table. The **ammanual** coef needs to be interpreted as the estimated increase in mpg of a car when that car were to switch from automatic to a manual transmission. The sum of beta0 and beta1 equal to the average mpg of all manual transmission

cars, as seen in the boxplot or summary table. Both coefs are highly statistically significant, as indicated by the three *. **The R-Squared value equals 0.3598, which can be interpreted as the percentage of how well the linear model describes the variance we see in the** mpg** variable.

We can try to increase the R-Squared value by adding more of the available variables to the regression. Of the remaining 9 variables, I excluded **qsec**, since it is not a direct measure or property of a car. I also excluded **V/S** and **drat** because I did not understand what they mean after checking the code book. For the remaining variables. I performed a pairwaise scatterplot (3rd plot in the appendix) to check whether the variables are correlated or not (the pdf does unfortunately squishes the figure).

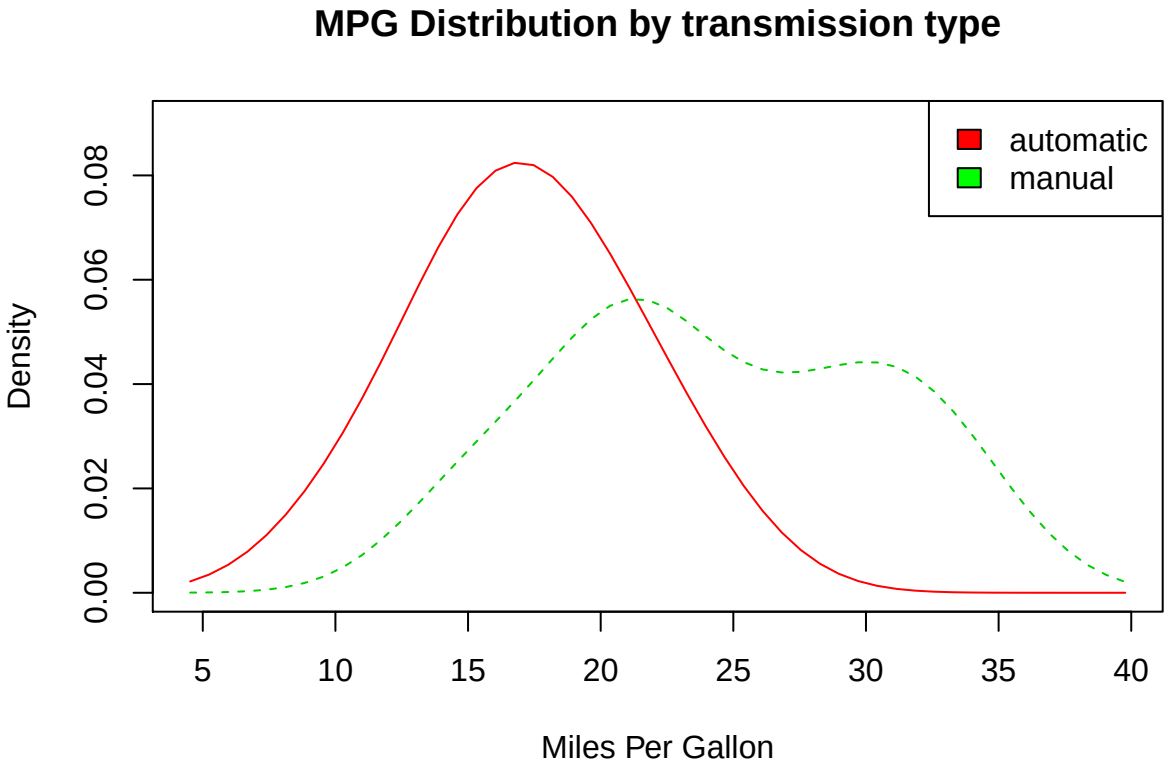
The scatterplot shows that shows clear linear correlations for all independent variables with **mpg**, except for gear. Also we can see that **cyl, disp, hp** are highly correlated among each other, which makes total sense even if you know as little about cars as I do. Therefore, I will exclude **gear** from further analysis. The second regression will be extended, compared to the first one, with the variables **wt, carb, and disp**. The latter is added as a "proxy" for the group of the three highly correlated variables **disp, hp, and cyl**. The results follow:

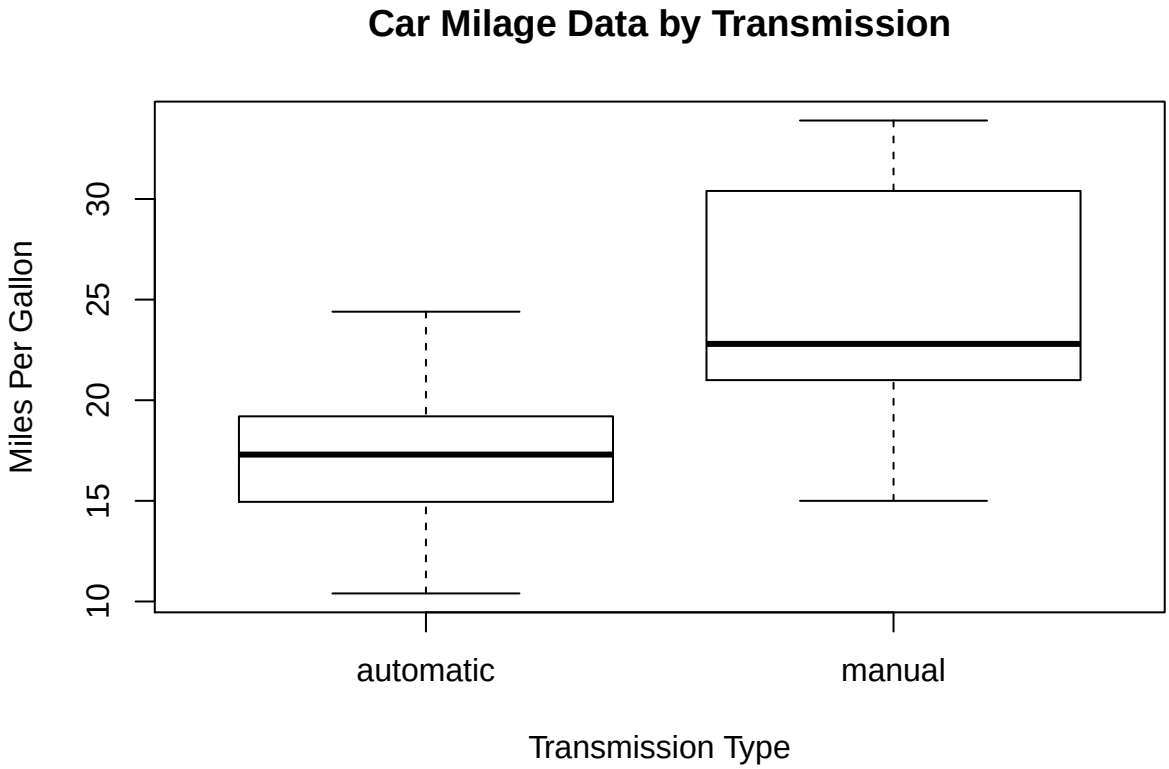
```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 31.35617156 3.050108984 10.280345 7.819619e-11
## ammanual     2.73260721 1.551617082  1.761135 8.953853e-02
## wt          -1.55035566 1.297812715 -1.194591 2.426357e-01
## disp        -0.01786457 0.008239194 -2.168242 3.912094e-02
## carb        -1.16125872 0.381831344 -3.041287 5.191351e-03
```

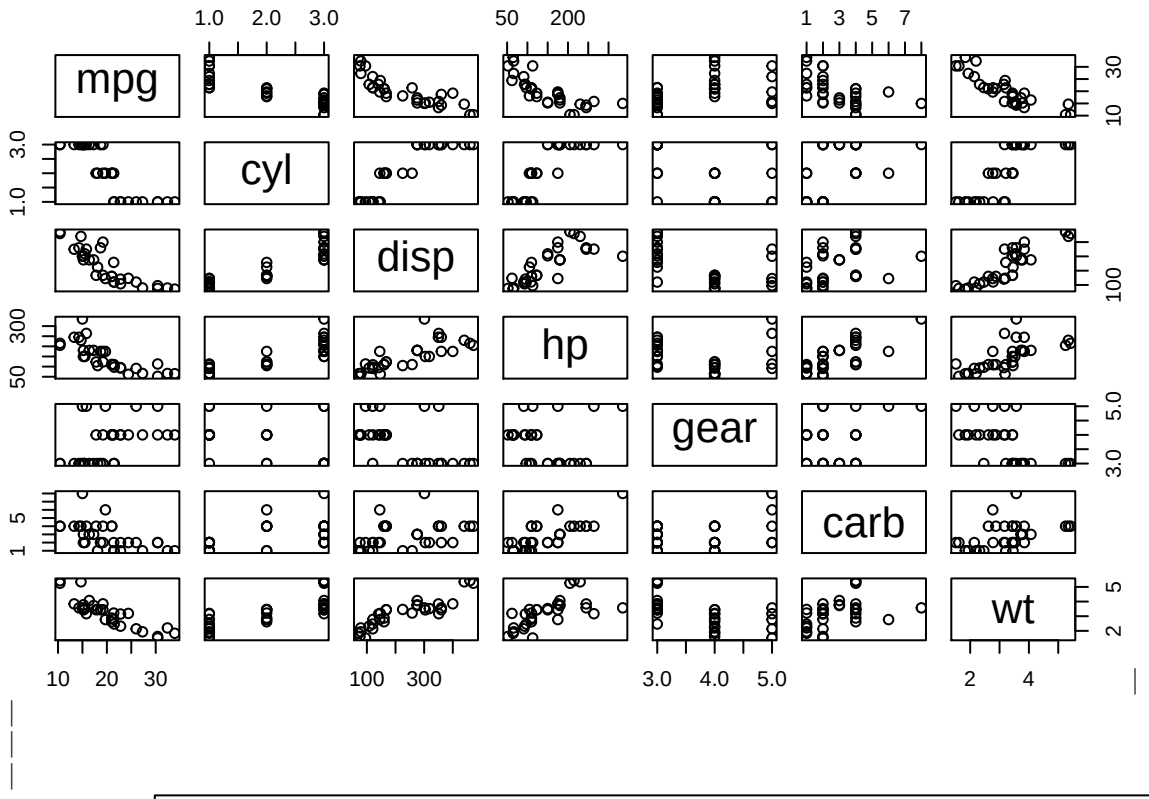
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + disp + carb
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     27 183.64  3    537.25 26.329 3.573e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

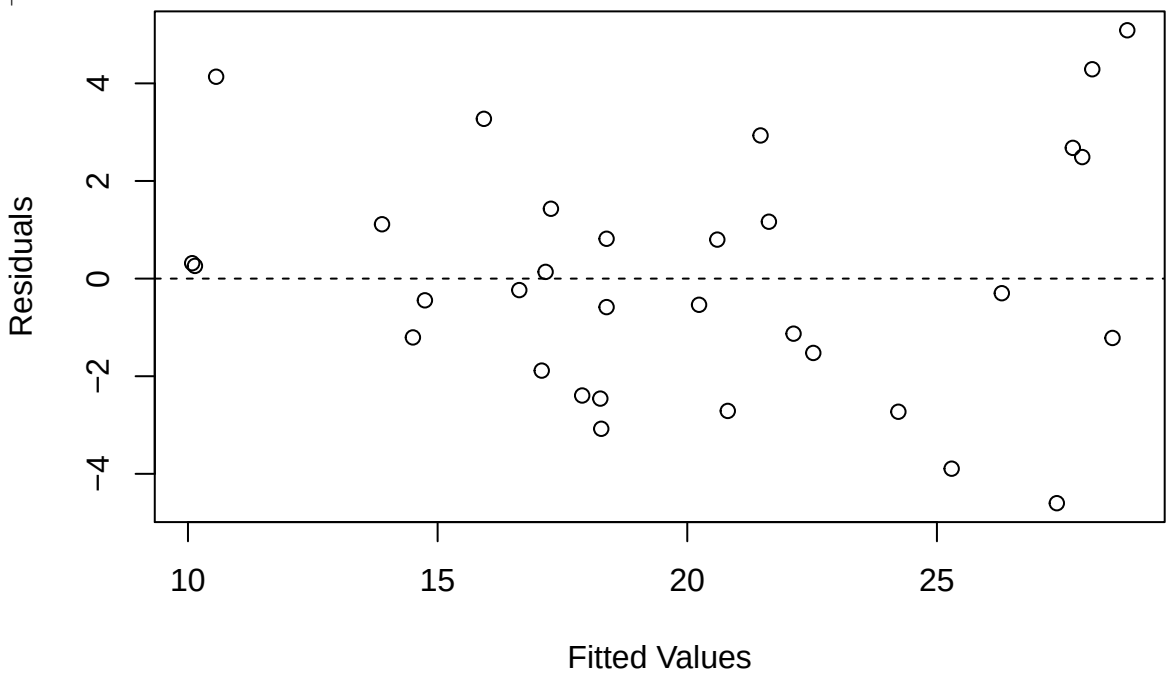
The results give us multiple insights: Surprisingly, wt is not significant at any of the significance levels, whereas common sense would say that a heavy car has a lower mpg than a light car. The remaining independent variables are significant. Looking at the coefficient of **ammanual**, we can see that the estimated effect reduces from a approx. 7mpg increase to an only approx. 2mpg increase, when comparing two cars with different transmission, the rest being equal (accounting for other factors like disp). **Disp** has a negative influence on mpg. If you increase **disp** by 1 cu.in., the mpg decreases by 0.017, ceteris paribus. The number of carburators has a negative effect, also. If the **carb** is increased by 1, the mpg decreases by 1.61, ceteris paribus.
Adding these additional indepenendent variables also boosted the R-Squared value from 0.3598 to 0.8368. That means we increased the degree of model-explained mpg variance by approx. 50%.
Performing an ANOVA (analysis of variance) confirms the R-Squared boost by telling us that the addition of these independent variables caused a significant improvement at the *** level.
Finally looking at the residual plot, there is no systematic pattern to be seen, which is concluded from the random distribution of these residuals in the plot. Therefore, it seems as if we do not lack a systematic regressor in our model.

Appendix - Plots






Scatterplot Matrix
mpg
cyl
disp
hp
gear
carb
wt


Residuals
Fitted Values