Regression Analysis for the Mtcars Dataset

Max R

10/24/2017

Executive Summary (abstract)

The purpose of this report was to identify whether manual or automatic transmission cars are better for fuel economy as measured by the average number of miles a car can drive per one gallon of gasoline (MPG). While automatic transmission cars do technically have a lower MPG than manual transmission cars, this isn't a great interpretation. When weight was included in the model, the effect of transmission was made null. An explanation for this could be that manual cars tend to be smaller cars (perhaps sports cars) while automatic vehicles tend to bulkier. This means that MPG doesn't change as a function of the transmission, but rather transmission is more of a predictor of weight. The linear model of weight and quarter mile time predicting MPG is a better model that explains MPG variation that is also parsimonious.

Exploratory analysis and simple linear regression for MPG and transmission

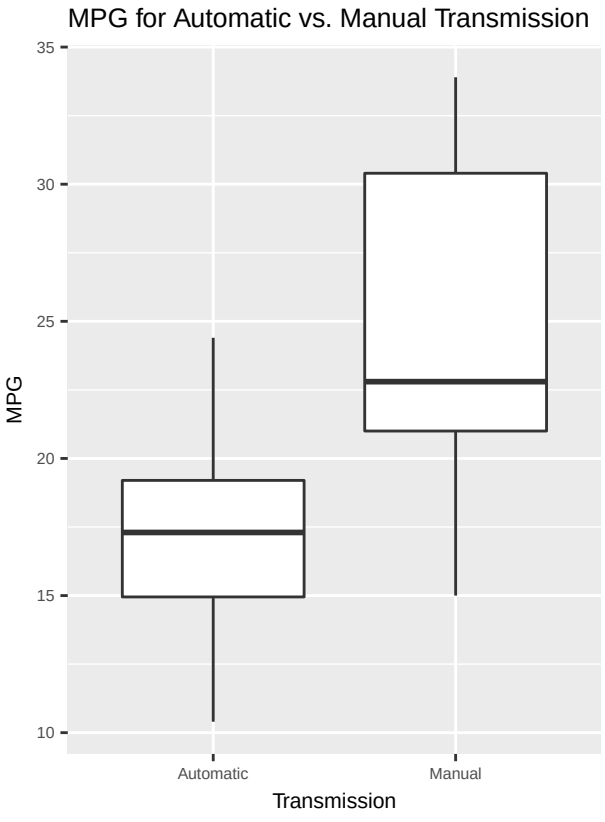

Figure 1

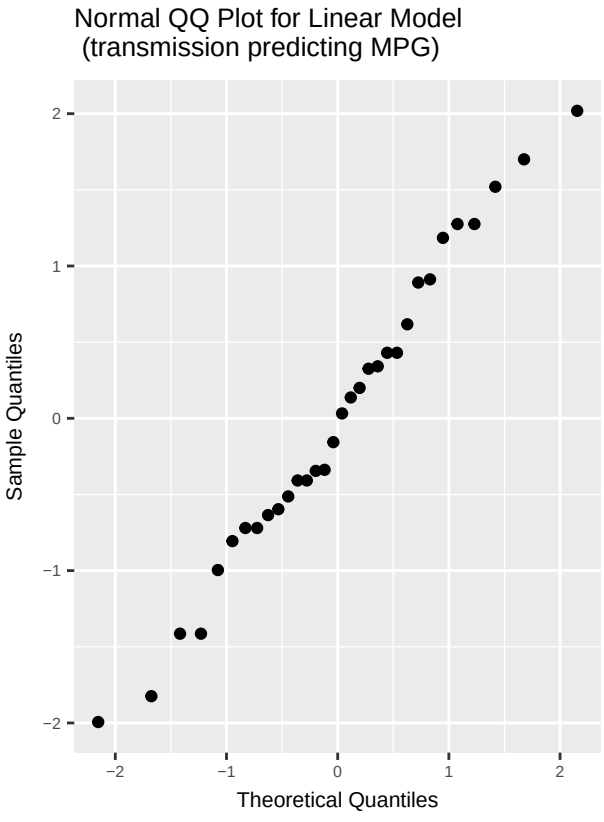

Figure 2

Table 1: Linear Model Predicting MPG with Transmission

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.147368	1.124602	15.247492	0.000000
transmissionManual	7.244939	1.764422	4.106127	0.000285

Figure 1 is an exploratory plot showing the differences in MPG between automatic and manual transmission vehicles.

Figure 2 is a Quantile-Quantile plot showing that the data is normally distributed.

Figure 1 shows that manual transmission cars have better gas mileage than automatic transmission cars, and it also appears that manual transmission cars have more variability in MPG. Table 1 further corroborates this by showing significant differences between the two transmissions. The model estimates an expected 7.24 increase in MPG for cars that have manual transmission compared to cars with an automatic transmission. That is, the mean MPG for automatic transmission cars is approximately 17.15 MPG and the mean MPG for manual transmission cars is approximately 24.39 MPG. These differences are significantly different according to the summary output (t-value = 4.1, $p < .001$). The R^2 (not shown in the table) is .36, which indicates 36% of the variation in MPG is explained by transmission.

Investigating weight as a regressor

Perhaps another variable better predicts fuel efficiency. The weight variable seems like it should have a strong relationship with gas mileage (heavier cars should have to utilize more gas). The correlation table below shows that weight has the highest correlation with mpg. Therefore, we can do some exploratory plotting and fit a linear model to see if transmission does predict weight.

Table 2: Correlations Table for all of the Variables

	mpg	cyl	disp	hp	drat
mpg	1.0000000	-0.8521620	-0.8475514	-0.7761684	0.6811719
cyl	-0.8521620	1.0000000	0.9020329	0.8324475	-0.6999381
disp	-0.8475514	0.9020329	1.0000000	0.7909486	-0.7102139
hp	-0.7761684	0.8324475	0.7909486	1.0000000	-0.4487591
drat	0.6811719	-0.6999381	-0.7102139	-0.4487591	1.0000000
wt	-0.8676594	0.7824958	0.8879799	0.6587479	-0.7124406
qsec	0.4186840	-0.5912421	-0.4336979	-0.7082234	0.0912048
vs	0.6640389	-0.8108118	-0.7104159	-0.7230967	0.4402785
am	0.5998324	-0.5226070	-0.5912270	-0.2432043	0.7127111
gear	0.4802848	-0.4926866	-0.5555692	-0.1257043	0.6996101
carb	-0.5509251	0.5269883	0.3949769	0.7498125	-0.0907898

	wt	qsec	vs	am	gear	carb
mpg	-0.8676594	0.4186840	0.6640389	0.5998324	0.4802848	-0.5509251
cyl	0.7824958	-0.5912421	-0.8108118	-0.5226070	-0.4926866	0.5269883
disp	0.8879799	-0.4336979	-0.7104159	-0.5912270	-0.5555692	0.3949769
hp	0.6587479	-0.7082234	-0.7230967	-0.2432043	-0.1257043	0.7498125
drat	-0.7124406	0.0912048	0.4402785	0.7127111	0.6996101	-0.0907898
wt	1.0000000	-0.1747159	-0.5549157	-0.6924953	-0.5832870	0.4276059
qsec	-0.1747159	1.0000000	0.7445354	-0.2298609	-0.2126822	-0.6562492
vs	-0.5549157	0.7445354	1.0000000	0.1683451	0.2060233	-0.5696071
am	-0.6924953	-0.2298609	0.1683451	1.0000000	0.7940588	0.0575344
gear	-0.5832870	-0.2126822	0.2060233	0.7940588	1.0000000	0.2740728
carb	0.4276059	-0.6562492	-0.5696071	0.0575344	0.2740728	1.0000000

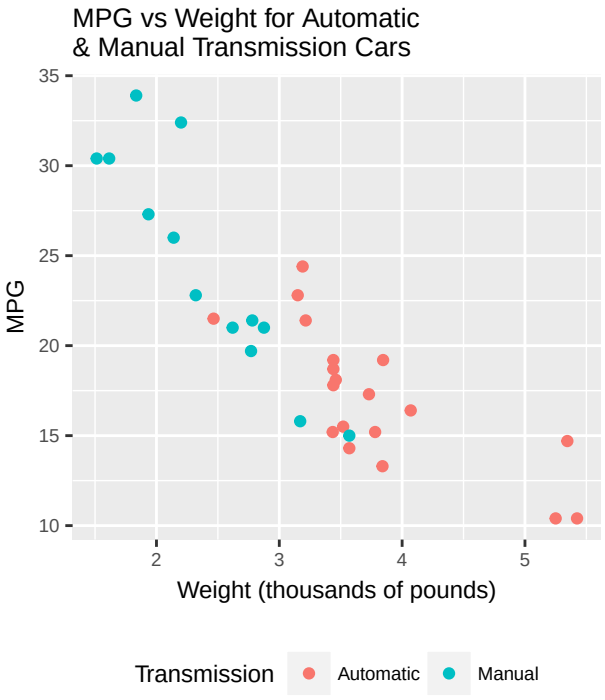

Figure 3

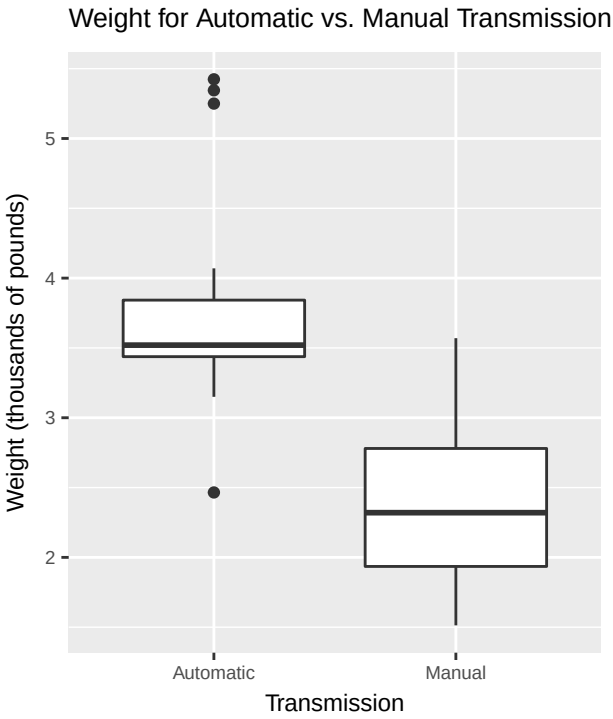

Figure 4

Table 4: Linear Model Predicting Weight with Transmission

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.768895	0.1646171	22.894914	0.00e+00
transmissionManual	-1.357895	0.2582726	-5.257603	1.13e-05

Figure 1 is a plot showing the relationship between car weight and miles per gallon. The correlation between weight and MPG is -.867.

Figure 2 is a boxplot showing the difference in weights for automatic vs. manual transmission cars.

Table 2 shows that the difference in weights between manual and automatic transmission vehicles is statistically significant.

Figure 1 shows that there is a strong negative relationship between weight and MPG. Also, it appears that manual transmission cars tend to weigh less than automatic transmission cars. This indicates that it is likely not transmission that is causing changes in MPG but rather the two transmission cars don't have equal weights. The boxplot and linear model aid with this interpretation. The table is showing that when cars have an automatic transmission, their average weight is 3.76 thousand pounds and when cars are manual transmission, their weight is 1.36 thousand pounds less than automatic transmission cars. To test to see how much transmission plays a role in predicting MPG when considering weight, we can do a multivariable linear regression model showing the coefficients and the R^2.

Table 5: Linear Model Predicting MPG with Car Weight and Transmission

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.3215513	3.0546385	12.2179928	0.0000000
wt	-5.3528114	0.7882438	-6.7908072	0.0000002
transmissionManual	-0.0236152	1.5456453	-0.0152786	0.9879146

The intercept value of 37.32 indicates what the MPG would be if weight (wt) were zero and if the transmission were automatic. This is not useful by itself, but we can see from the wt variable's slope that for every 1,000 pound increase in car weight, mpg goes down by about 5.35MPG (adjusting for transmission). This is statistically significant with a t-value of -6.79 and $p < .001$. However, if we look at transmission, the estimate was very small and not statistically significant (t-value of -0.015 and $p = .987$). This indicates that once you account for weight, transmission has no effect on mpg. The R^2 is .75, indicating that 75% of the variation is explained by weight (and negligibly transmission). As a diagnostic measure, we can see if any particular data points are heavily influecing the data using the `hatvalues` function.

Table 6: Hat Values with MPG as the Outcome and Weight & Transmission as the Predictors

Mazda RX4	0.0797510
Mazda RX4 Wag	0.0908614
Datsun 710	0.0774592
Hornet 4 Drive	0.0724939
Hornet Sportabout	0.0596347
Valiant	0.0588088
Duster 360	0.0551927
Merc 240D	0.0743273
Merc 230	0.0774291
Merc 280	0.0596347
Merc 280C	0.0596347
Merc 450SE	0.0585012
Merc 450SL	0.0527295
Merc 450SLC	0.0526396
Cadillac Fleetwood	0.1946508
Lincoln Continental	0.2299796
Chrysler Imperial	0.2134536
Fiat 128	0.0798054
Honda Civic	0.1179436
Toyota Corolla	0.0984024
Toyota Corona	0.1626994
Dodge Challenger	0.0566422
AMC Javelin	0.0598492
Camaro Z28	0.0529589
Pontiac Firebird	0.0530066
Fiat X1-9	0.0915917
Porsche 914-2	0.0816777
Lotus Europa	0.1291300
Ford Pantera L	0.1142188
Ferrari Dino	0.0852669
Maserati Bora	0.1638876
Volvo 142E	0.0857382

A few of the cars do exert influence (the ones with the highest influence appear to be luxury cars) but this influence isn't excessive.

Now that we know weight is a good predictor for mpg, we can test for other variables. However, while other variables may help explain more of the variation which would increase R^2, most of these variables have high correlations with weight, indicating the effects are somewhat bounded together. A way to test to see if the regressors are orthogonal to one another is to check the variance inflation factors for the variables using the `vif` function from the `car` package. This shows the increase in variance for the `i`th regressor compared to the ideal setting of indepedent regressors. We can do a multivariable linear model for several variables predicting

MPG and then calculate the variables' VIFs. I chose a few variables that had the smallest correlations with the weight variable rather than including all of the variables.

Table 7: Multivariate Linear Model Predicting MPG with Weight, Quarter Mile Time, Number of Carburetors and Number of Forward Gears

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.5291358	8.1655270	1.656860	0.1091246
wt	-3.7018877	0.8901893	-4.158540	0.0002906
qsec	0.7613016	0.3462040	2.198997	0.0366258
carb	-0.7847940	0.5469905	-1.434749	0.1628414
gear	1.9228037	1.0973304	1.752256	0.0910813

Table 8: VIF for regressors

wt	3.635471
qsec	1.833993
carb	3.740454
gear	3.141008

Even though R^2 increased to .844, including correlated variables inflates the standard error of the model. The interpretation of the VIF table is that for each of these variables, the standard error for the MPG effect is more than triple from what it would be if the regressors were orthogonal to one another for the variables "wt", "carb" and "gear". For example, the number of carburetors variables has a VIF of 3.74, indicating its effect is 3.74 times what it would be if it were uncorrelated with the other variables. However, looking at the Table 2 correlation table and the VIF of the quarter mile time variable, we can see quarter mile time is somewhat uncorrelated with weight (VIF < 2). Also from the output, we can see "qsec" reached statistical significance, for $p < .05$ while the "carb" and "gear" variables did not reach statistical significance. We can do a final model with only weight and quarter mile times as predictor variables for mpg to see the effect on the coefficients and R^2

Table 9: Multivariate Linear Model Predicting MPG with Weight & Quarter Mile Time

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.746223	5.2520617	3.759709	0.0007650
wt	-5.047982	0.4839974	-10.429771	0.0000000
qsec	0.929198	0.2650173	3.506179	0.0014999

With an R^2 of .826 and its inclusion is considered necessary according to the p value. Including quarter mile time seems to benefit the model. The table shows that as quarter mile goes up, cars become more gas efficient.

Appendix(R code)

```
library(car)
library(datasets)
library(ggplot2)
```

```r
library(knitr)
library(gridExtra)
mtcars$transmission <- as.factor(mtcars$am)
levels(mtcars$transmission) <- c("Automatic", "Manual")

p1 <- ggplot(mtcars, aes(x = transmission, y = mpg)) +
        geom_boxplot() +
        xlab("Transmission") +
        ylab("MPG") +
        labs(caption = "Figure 1") +
        ggtitle("MPG for Automatic vs. Manual Transmission") +
        theme(text = element_text(size=8),
              plot.title = element_text(size = 10))

mod <- lm(mpg ~ transmission, data = mtcars)
mpg_stdres <- rstandard(mod)
p2 <- ggplot(mtcars) +
        geom_qq(aes(sample= mpg_stdres)) +
        xlab("Theoretical Quantiles") +
        ylab("Sample Quantiles") +
        labs(caption = "Figure 2") +
         ggtitle("Normal QQ Plot for Linear Model\n (transmission predicting MPG)") +
         theme(text = element_text(size=8),
              plot.title = element_text(size = 10))

grid.arrange(p1, p2, ncol =2)

mod <- lm(mpg ~ transmission, data = mtcars)
kable(summary(mod)$coefficients, caption = "Linear Model Predicting MPG with Transmission")
mtcars$am <- as.numeric(mtcars$am)
cor_table <- cor(mtcars[,1:11], method = "pearson", use = "complete.obs")
kable(cor_table[, 1:5], caption = "Correlations Table for all of the Variables")
kable(cor_table[, 6:11])



g1 <- ggplot(mtcars, aes(x = wt, y = mpg, color = transmission)) +
        geom_point() +
        xlab("Weight (thousands of pounds)") +
        ylab("MPG") +
        labs(color = "Transmission", caption = "Figure 3") +
        scale_fill_discrete(labels = c("Automatic", "Manual")) +
        ggtitle("MPG vs Weight for Automatic\n& Manual Transmission Cars") +
        theme(text = element_text(size=9),
              plot.title = element_text(size = 10),
              legend.position = "bottom")

g2 <- ggplot(mtcars, aes(x = transmission, y = wt)) +
        geom_boxplot() +
        xlab("Transmission") +
        ylab("Weight (thousands of pounds)") +
        labs(caption = "Figure 4") +
        ggtitle("Weight for Automatic vs. Manual Transmission") +
```

```
        theme(text = element_text(size=9),
              plot.title = element_text(size = 10),
              legend.position = "bottom")

grid.arrange(g1, g2, ncol =2)
mod1 <- lm(wt ~ transmission, data = mtcars)
kable(summary(mod1)$coef, caption = "Linear Model Predicting Weight with Transmission")
mod2 <- lm(mpg ~ wt + transmission, data = mtcars)

kable(summary(mod2)$coefficients, caption = "Linear Model Predicting MPG with Car Weight and Transmissio

mod <- lm(mpg ~ wt + transmission, data = mtcars)
kable(hatvalues(mod), caption = "Hat Values with MPG as the Outcome and Weight & Transmission as the Pr
mod3 <- lm(mpg ~ wt + qsec + carb + gear, data = mtcars)
kable(summary(mod3)$coefficients, caption = "Multivariate Linear Model Predicting MPG with Weight, Quar
kable(vif(mod3), caption = "VIF for regressors")

mod4 <- lm(mpg ~ wt + qsec, data = mtcars)
kable(summary(mod4)$coefficients, caption = "Multivariate Linear Model Predicting MPG with Weight & Qua
```