Regression Models: Project 1

Summary

The goal of this analysis is to investigate the impact of the transmission type on the fuel economy of 32 vehicles in the 1974 Motor Trend dataset. We find that there is on average a 7 mpg difference between vehicles with manual transmission and vehicles with automatic transmission. A linear fit of available predictors to the fuel economy reveals that the most influential predictors are *transmission type*, *number of cylinders*, the *horsepower*, and *weight of the vehicles*. Under this regression model, a change in the trasmission from manual to automatic accounts for 1.8 more miles per gallon when all other predictors are held constant.

Introduction

The dataset for this analysis, `mtcars`, comes from the 1974 *Motor Trend* magazine and lists performance statistics of 32 vehicles (rows) for 11 categories (columns). There are six continuous numerical variables:

- `mpg`: miles per gallon
- `disp`: engine displacement (cu. in.)
- `hp`: horsepower
- `drat`: read axle ratio
- `wt`: weight (lb/1000)
- `qsec`: time per 1/4 mile

Additionally there are three discreet variables:

- `cyl`: number of cylinders in the vehicle
- `gear`: number of forward gears
- `carb`: number of carburetors

And lastly there are two categorical variables:

- `vs`: V/S
- `am`: **transmission** (automatic or manual)

In the course of this analysis we will be looking at the impact of the 10 potential predictors on the fuel economy (mpg) of the vehicles in the dataset. We are specifically interested in answering whether the type of vehicle transmission has an impact on the mpg.

Exploratory Analysis

As the **Figure 1** shows, there is a *visual* difference between the two populations, suggesting that the transmission type is at least correlated with fuel economy. There is a larger sample of manual-transmission cars than automatic transmission cars, so the sample distribution looks more normal in the former case. Since there are only 32 cars *total* we will leverage a t-test to confirm whether the populations are different. We will assume constant population variance since there was no [outward] bias in the automatic-vs-manual sampling.

```
    Two Sample t-test

data:  man_cars and aut_cars
t = -4.1061, df = 30, p-value = 0.000285
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.84837  -3.64151
sample estimates:
mean of x mean of y
 17.14737  24.39231
```

The t-test indeed shows that we can reject the null hypothesis and that the fuel economy differs across transmission types. Manual transmission cars have on average **7 fewer miles per gallon** than automatic transmission cars. Unfortunately we don't yet know what counfounding variables affect the relationship and thus need to investigate the correlations between the 10 possible predictors and mpg.

Regression Analysis

The questions we are investigating is *whether the transmission type has an effect on the fuel economy.* As such it makes sense to start with a model that just predicts mpg based off am.

```
simp.fit <- lm(mpg~am,data=mtcars)
formula(simp.fit)
summary(simp.fit)$adj.r.squared
```

```
mpg ~ am
[1] 0.3384589
```

With an R^2 of 34% the model is pretty poor at explaining the variance in mpg by transmission alone. We need to incorporate more confounders but since we don't know the interactions (correlations of factor variables are beyond the scope of this report) it is best to start with an over-fit model and step our way backwards.

Let's continue with a full-interaction model, where we consider the impact of all 10 predictors. The big and unfortunately unknown question is whether the predictors are *independent*. Perhaps the number of gears is correlated with the transmission type or the number of cylinders. These kind of interactions should be excluded from the model since they do not contribute any meaningful information and only serve to grow the variance.

```
full.fit <- lm(mpg~am+.,data=mtcars)
formula(full.fit)
summary(full.fit)$adj.r.squared
```

```
mpg ~ am + (cyl + disp + hp + drat + wt + qsec + vs + am + gear +
    carb)
[1] 0.7790215
```

Table 2 in the appendix shows the summary of the fit coefficients. The R^2 value above implies that only 78% of variation is explained by our model. We can see that no one predictor has a significant p-value, suggesting the model is already overfit. We will leverage the AIC to optimize the model by removing insignificant confounders. The `step` function will optimize on the AIC and return the best-fitting model.

```r
optim.fit <- step(full.fit,trace=0)
formula(optim.fit)
summary(optim.fit)$adj.r.squared
```

```
mpg ~ am + cyl + hp + wt
[1] 0.8400875
```

The coefficients of the new model are presented in **Table 3**, while the difference in the AIC is in **Table 4**. The new model now accounts for 84% of the variability and has much improved p-values.

Based off the coefficients in **Table 3**, when all other predictors are held constant, a change in the transmission type from manual to automatic adds **1.8 miles per gallon**.

Unfortunately the std. error for the actual parameter we're looking for, `am`, is pretty high, resulting in a rather useless confidence interval: *with **95% confidence**, the increase in mpg for the change from manual to automatic transmission lies within **-1.06** and **4.68** mpg.* This suggests that perhaps transmission is a dependent variable of other confounders, but excluding it from the model would not allow us to quantify the change in mileage per change in transmission type.

Residual analysis

Our fit is only valid if it satisfies the requirements of a linear fit: independence, normality and homoscedasticity. We can validate all three by plotting the residuals of the fit against the predicted values and in a QQ plot.

As we can see from **Figure 2**, the residuals appear randomly and uniformly distributed around 0. The QQ plot further confirms the normal distribution of residuals, thus justifying our model.

Conclusion

It was found that the car transmission type does indeed have a meaningful impact on fuel economy of the vehicle. Of the 32 cars analyzed, manual-transmission cars on average have 7 fewer miles per gallon than automatic-transmission cars. The type of transmission alone, however, is not the only predictor of a vehicle's fuel effincy. A linear fit of mpg to the ten possible predictors showed that fuel economy is a function of *transmission*, *number of cylinders*, *horse power*, and *vehicle weight*. Under such a model, with all other factors held constant, the change of transmission from manual to automatic accounts for **1.8 more miles per gallon** on average.

Appendix

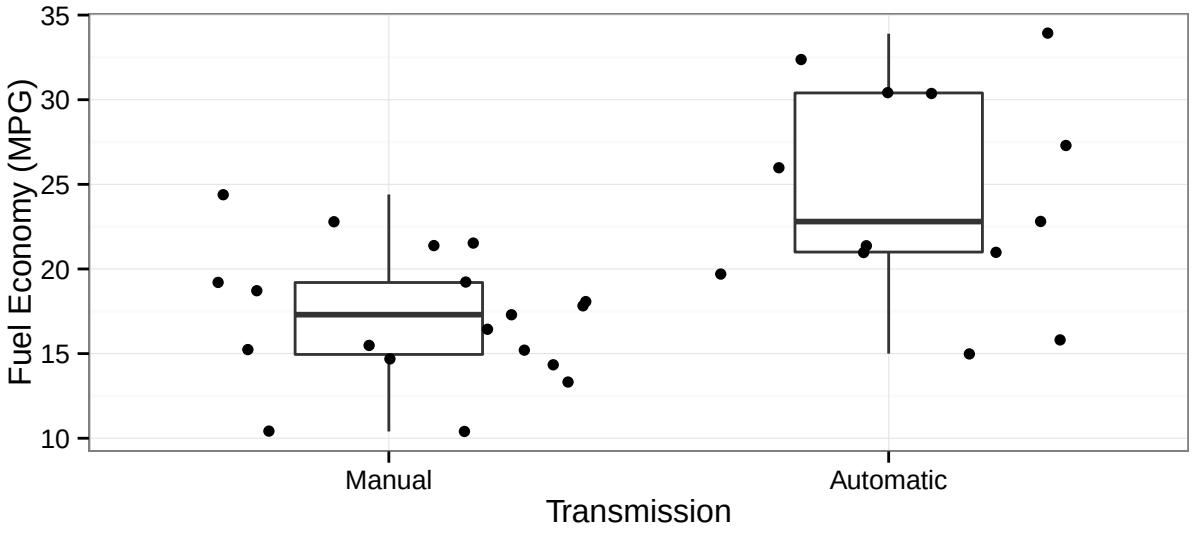

Figure 1: Distribution of fuel economy as function of transmission type

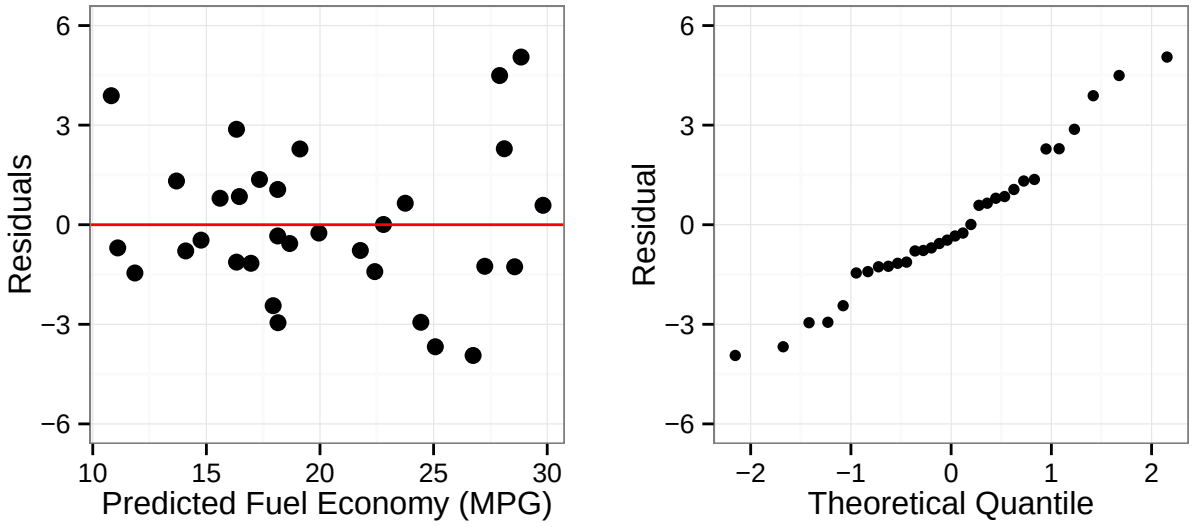

Figure 2: Residual disribution and QQ plot for optimal fit to mpg

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

Table 1: Predicted coefficients for a simple linear fit of am to mpg

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	23.8791324	20.0658203	1.1900402	0.2525255
amAutomatic	1.2121157	3.2135451	0.3771896	0.7113157
cyl6	-2.6486953	3.0408904	-0.8710262	0.3974664
cyl8	-0.3361630	7.1595395	-0.0469532	0.9631700
disp	0.0355463	0.0318992	1.1143329	0.2826734
hp	-0.0705068	0.0394256	-1.7883534	0.0939316
drat	1.1828302	2.4834846	0.4762784	0.6407392
wt	-4.5297758	2.5387458	-1.7842573	0.0946186
qsec	0.3678448	0.9353957	0.3932505	0.6996672
vsS	1.9308505	2.8712578	0.6724755	0.5115079
gear4	1.1143549	3.7995173	0.2932886	0.7733203
gear5	2.5283960	3.7363580	0.6767007	0.5088975
carb2	-0.9793543	2.3179745	-0.4225044	0.6786509
carb3	2.9996387	4.2935461	0.6986390	0.4954678
carb4	1.0914229	4.4496199	0.2452845	0.8095603
carb6	4.4775692	6.3840624	0.7013668	0.4938127
carb8	7.2504113	8.3605664	0.8672153	0.3994849

Table 2: Predicted coefficients for a full-model linear fit to mpg

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.7083239	2.6048862	12.940421	0.0000000
amAutomatic	1.8092114	1.3963045	1.295714	0.2064597
cyl6	-3.0313445	1.4072835	-2.154040	0.0406827
cyl8	-2.1636753	2.2842517	-0.947214	0.3522509
hp	-0.0321094	0.0136926	-2.345025	0.0269346
wt	-2.4968294	0.8855878	-2.819404	0.0090814

Table 3: Predicted coefficients for an optimized linear fit to mpg

	df	AIC
simp.fit	3	196.4844
full.fit	18	169.2155
optim.fit	7	154.4669

Table 4: Information Criterion of full-predictor vs optimized models