Regression Models Final Project

This is an analysis of automobile data from a 1974 issue of *Motor Trend* magazine, focusing specifically on whether there is a relationship between type of transmission and miles per gallon (*mpg*) and, if so, how it can be characterized, using three regression models. The best model included no interaction terms and the overall data indicates manual transmissions have a higher mpg than automatic.

Analysis

The data can be loaded into R using the command data(mtcars). There are 11 variables: *mpg*, number of cylinders (*cyl*), displacement (*disp*) (cu. in.), gross horsepower (*hp*), rear axel ratio (*drat*), weight in 1000 lbs (*wt*), quarter second mile (*qsec*), V/S (*vs*), transmission (*am*) (0=automatic and 1=manual), number of forward gears (*gear*), and number of carburetors (*carb*). Since many of these variables are categorical (factor) variables, they need to be transformed into a more usable format using the factor function. We can plot the variables against each other to see if any strong correlations appear. The plot (see Appendix) indicates that there are some trends associated with *mpg*. Regression analysis will confirm this and model the relationship.

Regression Models

The first regression model examined is the most basic, with *mpg* depending only on type of transmission.

```
model.AMOnly <- lm(mpg ~ am, data = mtcars); summary(model.AMOnly)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

The next uses the step function to search through the combinations of variables to form the best regression model based on the Akaike information criterion (AIC). The AIC is a measure of information lost due to modeling a process, balancing the goodness of fit with the complexity of the model. The AIC=$2k$-$2\ln(L)$, where L is the maximized value of the likelihood function (describing the likelihood of the data, given the estimated parameters) and k is the number of terms. Since lower AIC scores are better, more terms results in a higher AIC.

```
model.allVariables <- lm(mpg ~ ., data = mtcars)
model.best <- step(model.allVariables, direction = "both", trace=0)
summary(model.best)$call; summary(model.best)$coefficients;summary(model.best)$r.squared
```

```
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)

##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02

## [1] 0.8496636
```

A third model uses step again to account for any interaction other variables could potentially have with *am*:

```
model.bestinteractions<-step(lm(mpg~.+am:.,data=mtcars),direction="both", trace=0)
model.bestinteractions$call
```

```
## lm(formula = mpg ~ disp + hp + drat + wt + qsec + am + gear +
##     carb + disp:am + hp:am + drat:am + wt:am + qsec:am + am:carb,
##     data = mtcars)
```

This model appears to fit the data well. However, the large number of variables in the regression equation raises the potential for overfitting, which will make it difficult to understand the relationship between *mpg* and *am*. Therefore, we will not consider this model.

To determine whether the inclusion of the additional variables in model.best are significant, an ANOVA helps us determine if the three models are statistically different:

```
anova(model.AMOnly, model.best)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This test shows that we can reject the null hypothesis that the first and second models are not different. We will select the second model, which accounts for 86.6% of the variance. Notice in this model that the *am* coefficient is positive, indicating that if the car's transmission is manual, it will have an increased *mpg*. This effect remains in the model only containing *am*. This means that the effect is likely consistent. We can further verify that this effect exists by performing a t-test on the data.

```
t.test(mpg~am,data=mtcars)$statistic;t.test(mpg~am,data=mtcars)$p.value
```

```
##         t
## -3.767123
```

```
## [1] 0.001373638
```

The p-value of this t-test is p<.001, making it statistically significant, verifying that the manual transmission has a higher *mpg* than the automatic.

Diagnostics

As shown in the appendix, the residuals vs fitted and normal QQ plot show normal and homoskedastic errors. Additionally, the scale-location and residual vs leverage plot do not indicate any extreme outliers. To be sure that the model is not influenced strongly by any individual data points, we can test leverage and influence using hatvalues and dfbetas, respectively.

```
leverage <- hatvalues(model.best);influential<-dfbetas(model.best); range(leverage); range(influential)
```

```
## [1] 0.05303857 0.29704218
```

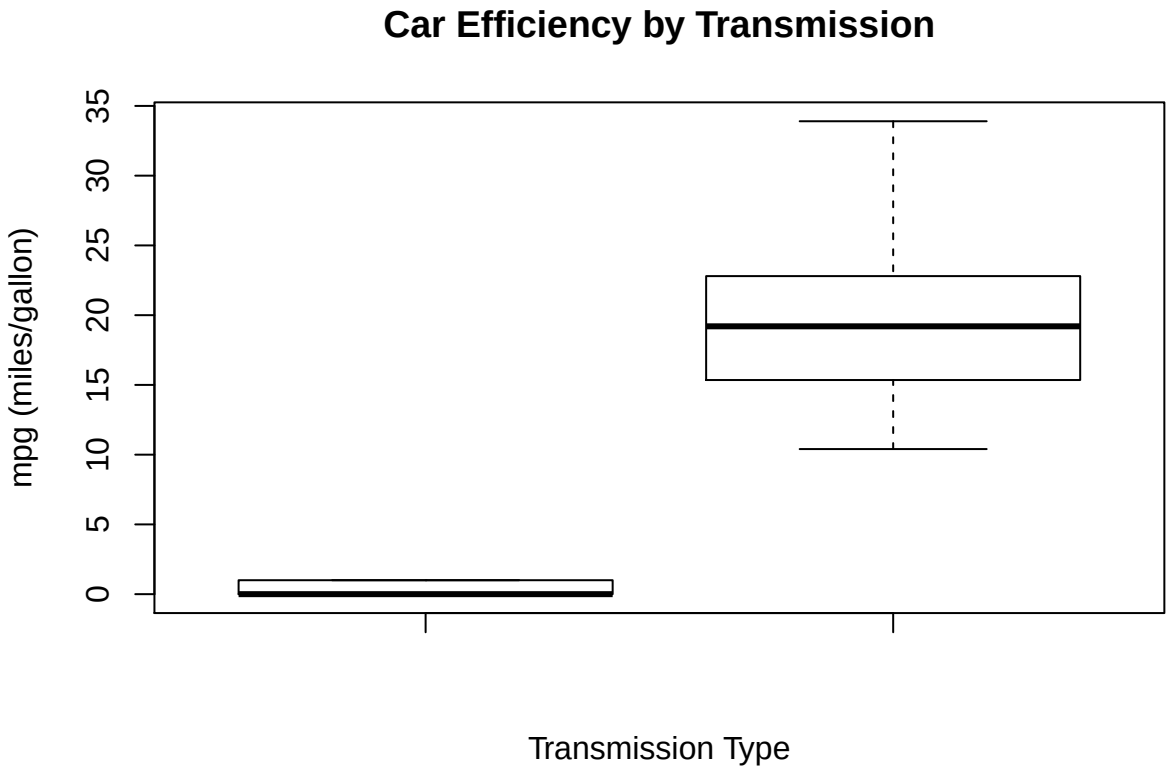
```
## [1] -0.6992918  1.0938422
```

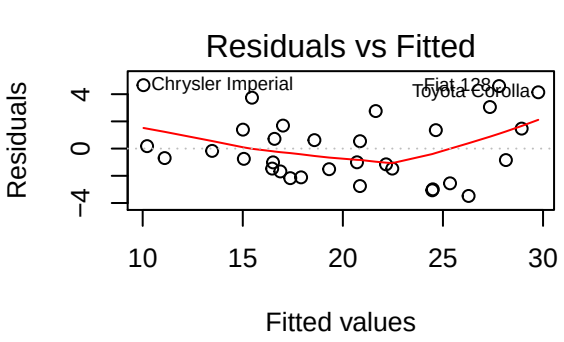
Since the values for leverage and influence are not significantly larger than any others, we can state that the model is reasonably stable.

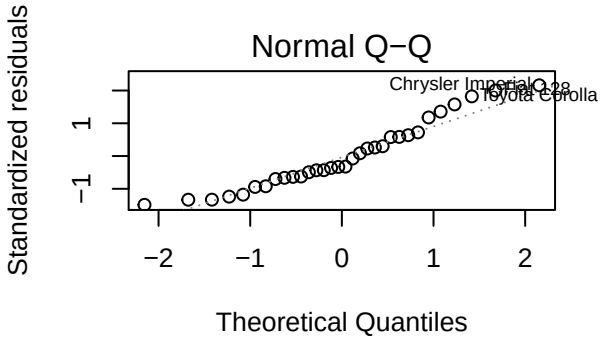
Conclusion

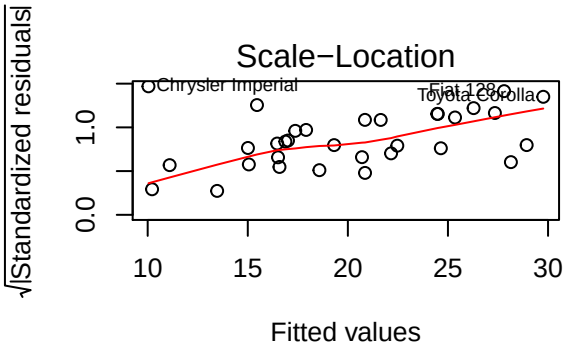
Based on the model we constructed and selected, and accompanying statistical tests and diagnostics, we can conclude that there is a relationship between *mpg* and transmission type. Specifically, manual transmission has a higher mpg than automatic transmission. Additionally, the model indicates that *mpg* has a negative relationship with horsepower, weight, and number of cylinders.

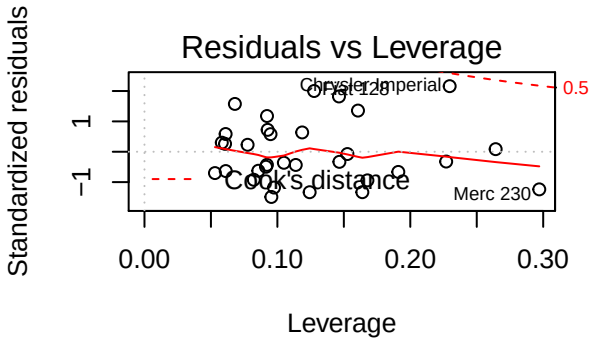
Appendix

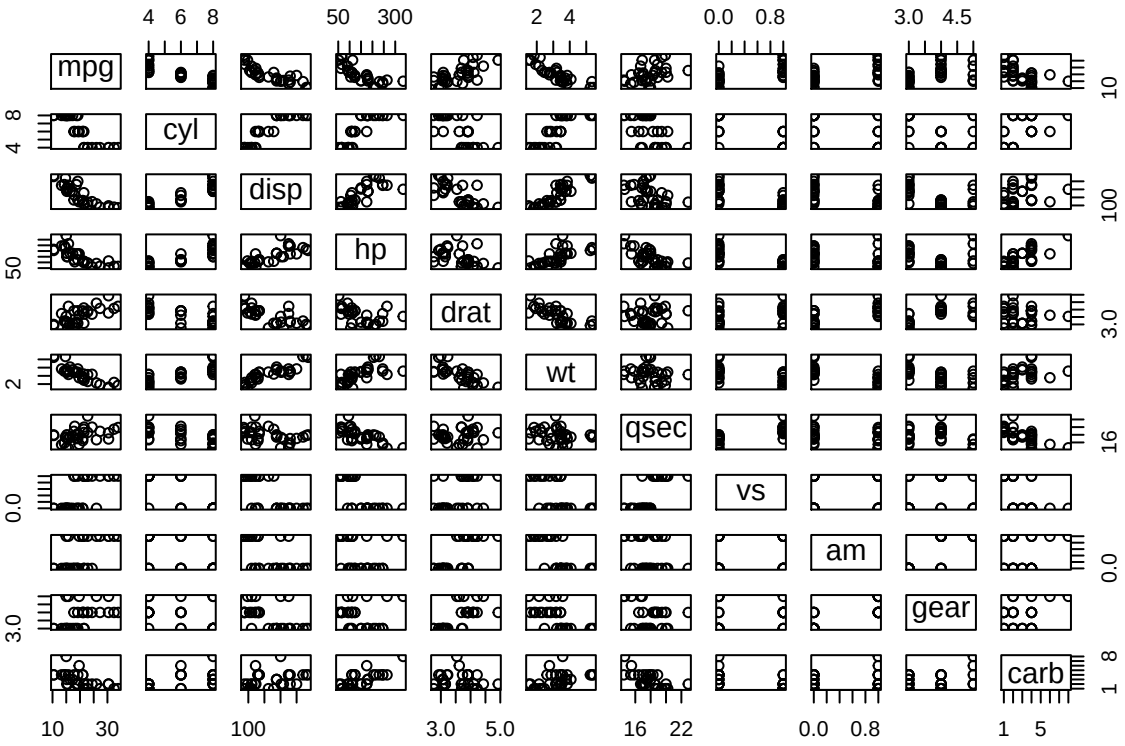

Car Efficiency by Transmission
mpg (miles/gallon)
0
5
10
15
20
25
30
35
Transmission Type


Residuals vs Fitted
Residuals
Chrysler Imperial
Fiat 128
Toyota Corolla
−4
0
4
10
15
20
25
30
Fitted values


Normal Q−Q
Standardized residuals
Chrysler Imperial
Fiat 128
Toyota Corolla
−1
1
−2
−1
0
1
2
Theoretical Quantiles


Scale−Location
√|Standardized residuals|
Chrysler Imperial
Fiat 128
Toyota Corolla
0.0
1.0
10
15
20
25
30
Fitted values


Residuals vs Leverage
Standardized residuals
Chrysler Imperial
Fiat 128
Cook's distance
0.5
Merc 230
−1
1
0.00
0.10
0.20
0.30
Leverage


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb