RegressionProject

JN Gabra

4/7/2021

Executive Summary

The mtcars data set is used to see if manual transmissions are better for the miles per gallon of a vehicle. Various models are run to determine the best predictors of the miles per gallon (mpg). In general, manual transmissions are better for mpg. If transmission type was the only independent variable for mpg then a manual transmission results in 7.2 mpg increase compared to automatic transmissions. However, when we include other variables we see that there isn't that much of a difference in mpg based on transmission type.

Preprocessing, Data Loading, and Exploration

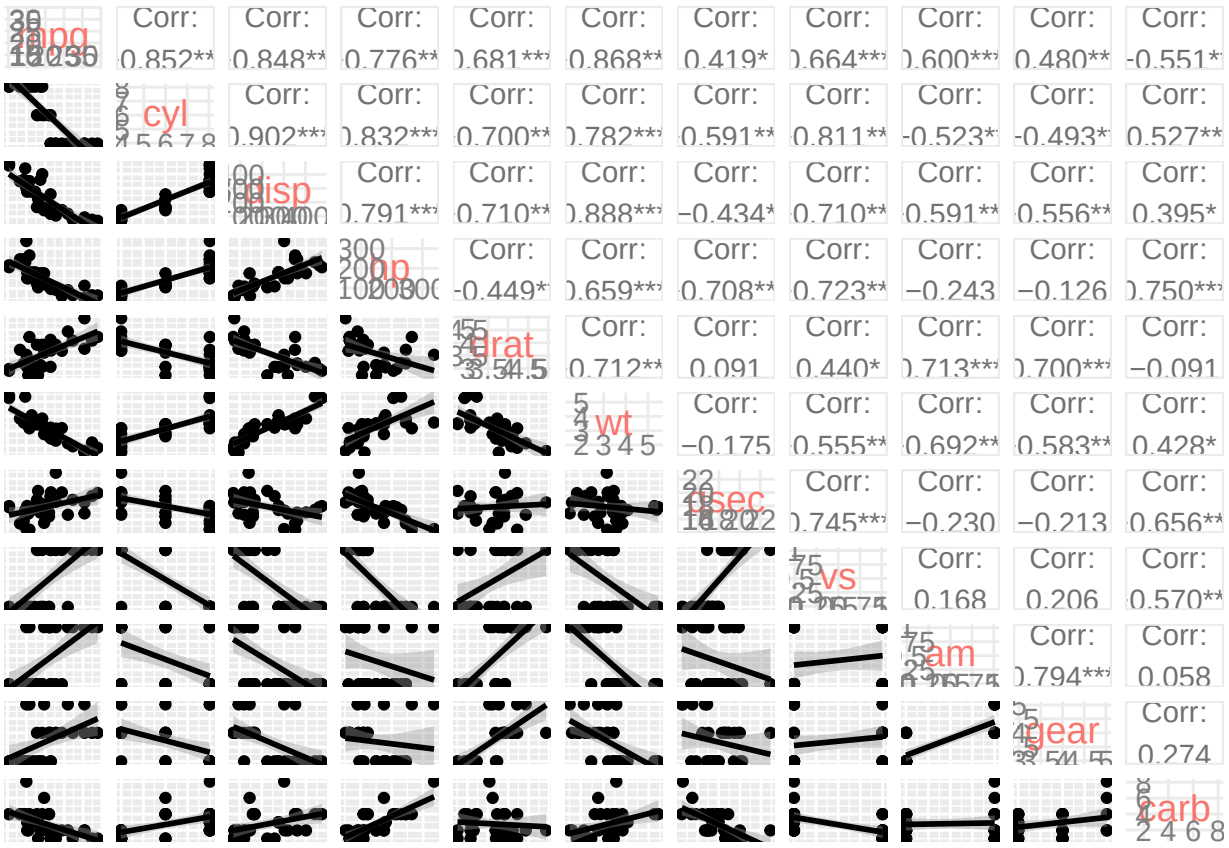
```
## Loading required package: ggplot2
```

```
## Loading required package: GGally
```

```
## Warning: package 'GGally' was built under R version 4.0.4
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
g1 = ggpairs(data,lower=list(continuous="smooth"),axisLabels = "internal")+theme(axis.line=element_blank
g1
```

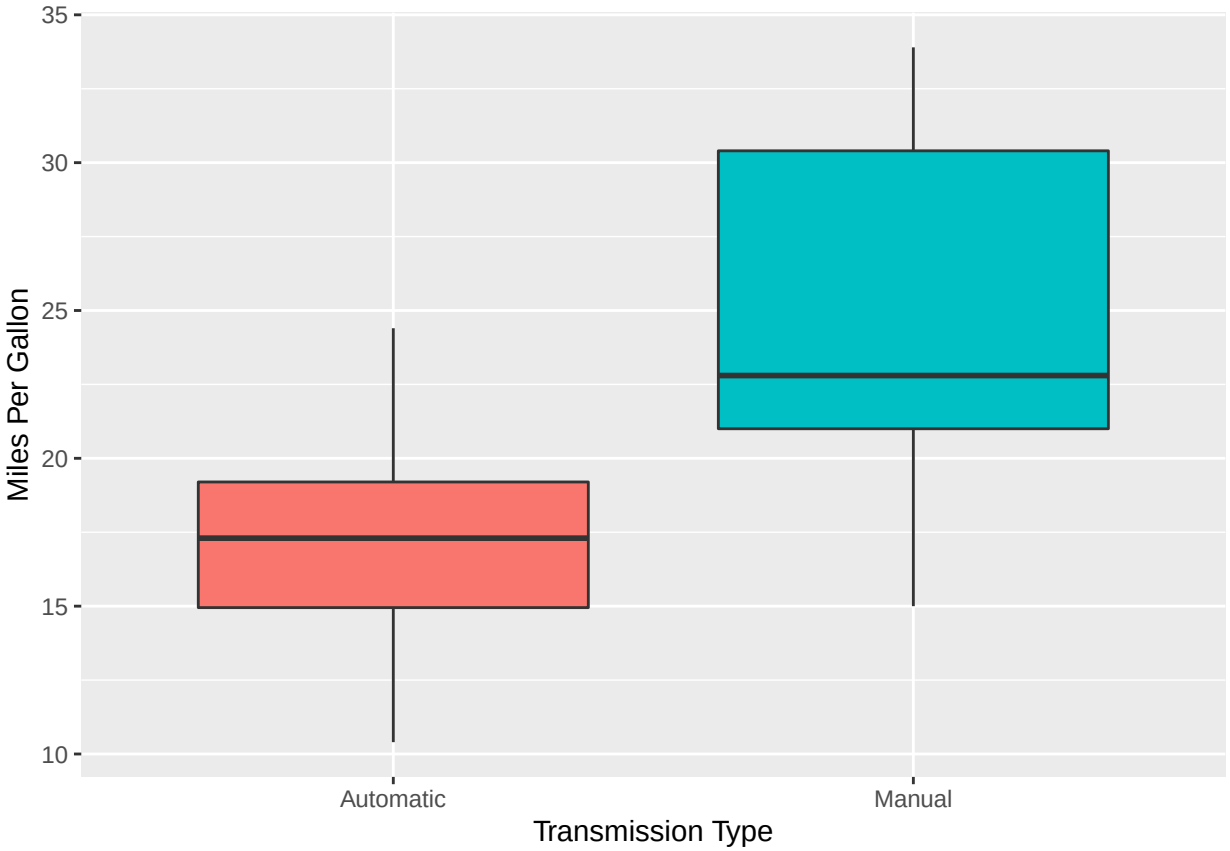


The above plot shows the correlations between all the variables in mtcars. In this, we can see that many of the variables are correlated.

Comparing Transmission Types

Now, a boxplot will be generated to directly compare the miles per gallon for the different transmission types.

```
data$am<-as.factor(data$am)
g2=ggplot(data,aes(y=mpg,x=am,fill=am))
g2=g2+geom_boxplot()+theme(legend.position = "none")
g2=g2+xlab("Transmission Type")+ylab("Miles Per Gallon")+scale_x_discrete(name ="Transmission Type", lal
g2
```



In the above box plot, we can see that Manual transmissions are associated with higher miles per gallon. Therefore, manual transmissions are better.

Model Development and Comparisions

Now we will test a few different models.

```
data<-mtcars
fit1<-lm(mpg~.,data)

fit2<-lm(mpg~am,data=data)
summary(fit2)
```

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

If we stop here, we can see that manual transmission cars (am=1) result in a 7.2 mile per gallon increase compared to automatic transmission cars.

What I will do now is try the nested testing method where variables will be added. I will first start with the transmission type variable. I will then add variables to the model in the order of decreassing correlation with mpg. Since cylinder number and displacment are extremely correlated I will only use cylinder number.

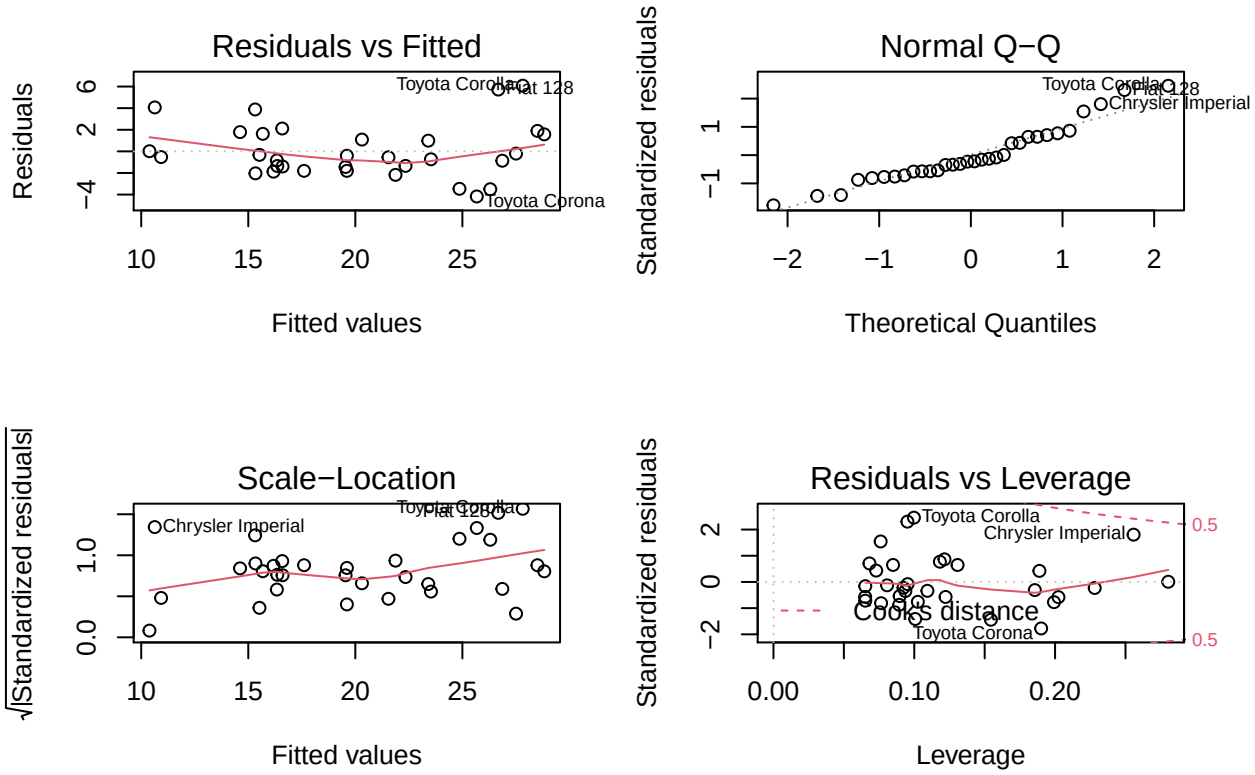
```
# Nested Testing with variables added
# Variables added in order of largest to smallest abs(Corr) with mpg
# cyl and displacement are extremely correlated and only one is chosen
fit3<-update(fit2,mpg~am+wt)
fit4<-update(fit2,mpg~am+wt+cyl)
fit5<-update(fit2,mpg~am+wt+cyl+hp)
fit6<-update(fit2,mpg~am+wt+cyl+hp+drat)
fit7<-update(fit2,mpg~am+wt+cyl+hp+drat+factor(vs))
anova(fit1,fit2,fit3,fit4,fit5,fit6,fit7)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ am
## Model 3: mpg ~ am + wt
## Model 4: mpg ~ am + wt + cyl
## Model 5: mpg ~ am + wt + cyl + hp
## Model 6: mpg ~ am + wt + cyl + hp + drat
## Model 7: mpg ~ am + wt + cyl + hp + drat + factor(vs)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     21 147.49
## 2     30 720.90 -9   -573.40  9.0711 1.779e-05 ***
## 3     29 278.32  1    442.58 63.0133 9.325e-08 ***
## 4     28 191.05  1     87.27 12.4257   0.00201 **
## 5     27 170.00  1     21.05  2.9970   0.09809 .
## 6     26 169.62  1      0.38  0.0541   0.81834
## 7     25 166.84  1      2.78  0.3961   0.53590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the outputs above we see the that the best model is the one that includes transmission, weight, and cylinders as independent variables.

Next, we will generate plots from the chosen model to inspect how well it holds up to assumptions.

```
par(mfrow=c(2,2))
plot(fit4)
```



This model does decently ok. There seems to be a trend that we are missing in the residuals vs fitted plot. Typically, I would investigate this more, but I will stop here for the scope of the class project.

Now lets look at the model details

```
summary(fit4)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
## am            0.1765     1.3045   0.135  0.89334
## wt           -3.1251     0.9109  -3.431  0.00189 **
## cyl          -1.5102     0.4223  -3.576  0.00129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

```r
confint(fit4)
```

```
##                 2.5 %     97.5 %
## (Intercept) 34.007153 44.8287134
## am          -2.495555  2.8485408
## wt          -4.991001 -1.2592836
## cyl         -2.375245 -0.6452459
```

From this we see that when we adjust the MPG model for transmission, weight, and cylinder number that manual transmissions still produce a slight increase in MPG (0.18) compared to automatic transmissions. However, we see that the transmission variable is no longer a significant predictor for mpg. This is further demonstrated with the 95% confidence interval for the slope of transmission type including 0.