Regression Models: Project

Marcos Gestal

Context

You work for Motor Trend. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Data summary

```
automaticCars <- mtcars[mtcars$am=="Automatic", ]
manualCars <- mtcars[mtcars$am=="Manual", ]

summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.42   19.20   20.09   22.80   33.90
```

```
summary(automaticCars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
```

```
summary(manualCars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

See Appendix 1 for a graphical view of the relation between MPG and type of transmission (boxplot).

These informations (boxplot and mean values in summaries) seem to show a best performance about mpg values for Automatic transmissions.

Manual vs Automatic Transmission

Statistical inference: T Test

A t-Test is performed to check wether there is a statistically significant difference between the mgp values for automatic and manual transmissions. The p-value < 0.05 so the null hypothesis can be reject, so we would accept that **there are differences in the mpg values due the type of transmissions**. The t-test results give us information about **MPG mean for automatic cars is lower than the MPG for manual cars**. It confirms our previous assumption about the best performance of automatic transmissions. Now a study will be done to quantify this difference.

```
t.test(manualCars$mpg, automaticCars$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  manualCars$mpg and automaticCars$mpg
## t = 3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.209684 11.280194
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

Linear Regression

```
summary( lm(mpg ~ am, data = mtcars) )
```

```
##
## 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
```

A brief study of the data shows how for automatic transmissions, the slope 7.245 for manual transmissions vs. automatic ones, so the performace is better (lower) for the last ones. The problem of this study relies in that it only explains 36% of the variance (Multiple R-Squared value)

We repeat the linear regression model, but now with the intercepted term in the origin

```
summary( lm(mpg ~ 0 + am, data = mtcars) )
```

```
##
## Call:
## lm(formula = mpg ~ 0 + 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|)
## amAutomatic   17.147      1.125   15.25 1.13e-15 ***
## amManual      24.392      1.360   17.94  < 2e-16 ***
## ---
## 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:  0.9487, Adjusted R-squared:  0.9452
## F-statistic: 277.2 on 2 and 30 DF,  p-value: < 2.2e-16
```

Here we can observe how the performance for automatic transmissions is better than the manual ones.

Multivariate Regression

In this case, we extend the regression model to include other directly relevant variables to explain the mpg value

```
advancedModel <- lm(mpg ~ am + gear + wt + hp + cyl + carb, data = mtcars)
summary ( lm(mpg ~ am + gear + wt + hp + cyl + carb, data = mtcars) )
```

```
##
## Call:
## lm(formula = mpg ~ am + gear + wt + hp + cyl + carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1041 -1.4166 -0.5063  1.4282  5.4337
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.66187    7.16719   4.557 0.000117 ***
## amManual     1.74467    1.72343   1.012 0.321081
## gear         0.65162    1.41230   0.461 0.648510
## wt          -2.22442    1.00467  -2.214 0.036169 *
## hp          -0.01726    0.01666  -1.036 0.310030
## cyl         -0.67309    0.67059  -1.004 0.325130
## carb        -0.65048    0.57075  -1.140 0.265218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.54 on 25 degrees of freedom
## Multiple R-squared:  0.8567, Adjusted R-squared:  0.8223
## F-statistic: 24.91 on 6 and 25 DF,  p-value: 2.087e-09
```

This new regression model explains the 85.67% of the variance. We can see how the manual transmission and the number of gears increases the mpg values (transmission in a higher percentage).

```r
simpleModel <- lm(mpg ~ 0 + am, data = mtcars)
advancedModel <- lm(mpg ~ am + gear + wt + hp + cyl + carb, data = mtcars)

anova(simpleModel, advancedModel)
```

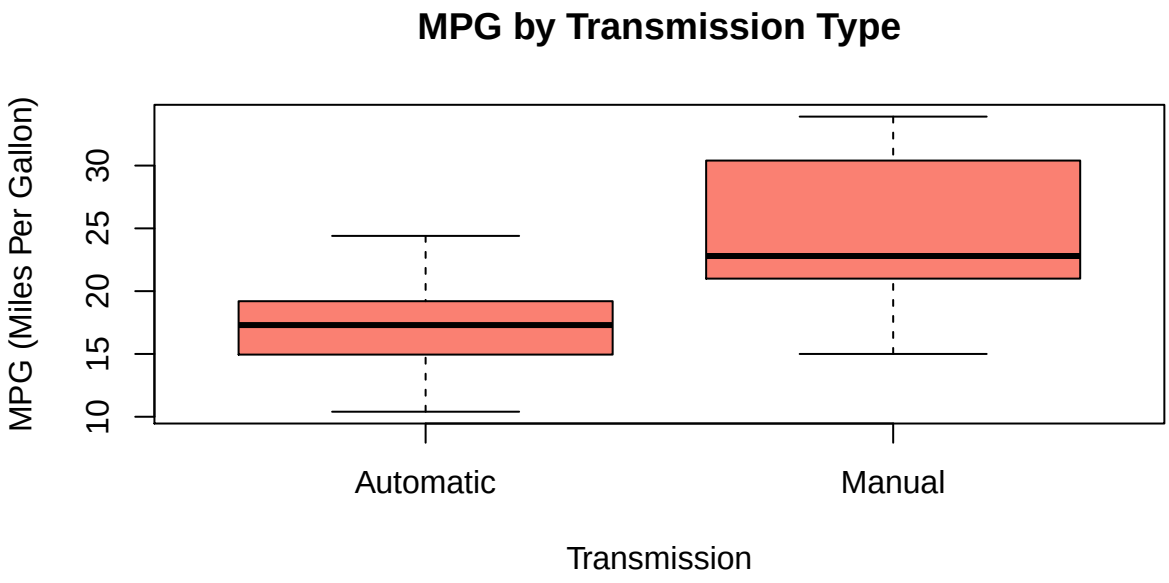
```
## Analysis of Variance Table
##
## Model 1: mpg ~ 0 + am
## Model 2: mpg ~ am + gear + wt + hp + cyl + carb
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     30 720.90
## 2     25 161.34  5    559.55 17.34 2.023e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclussions

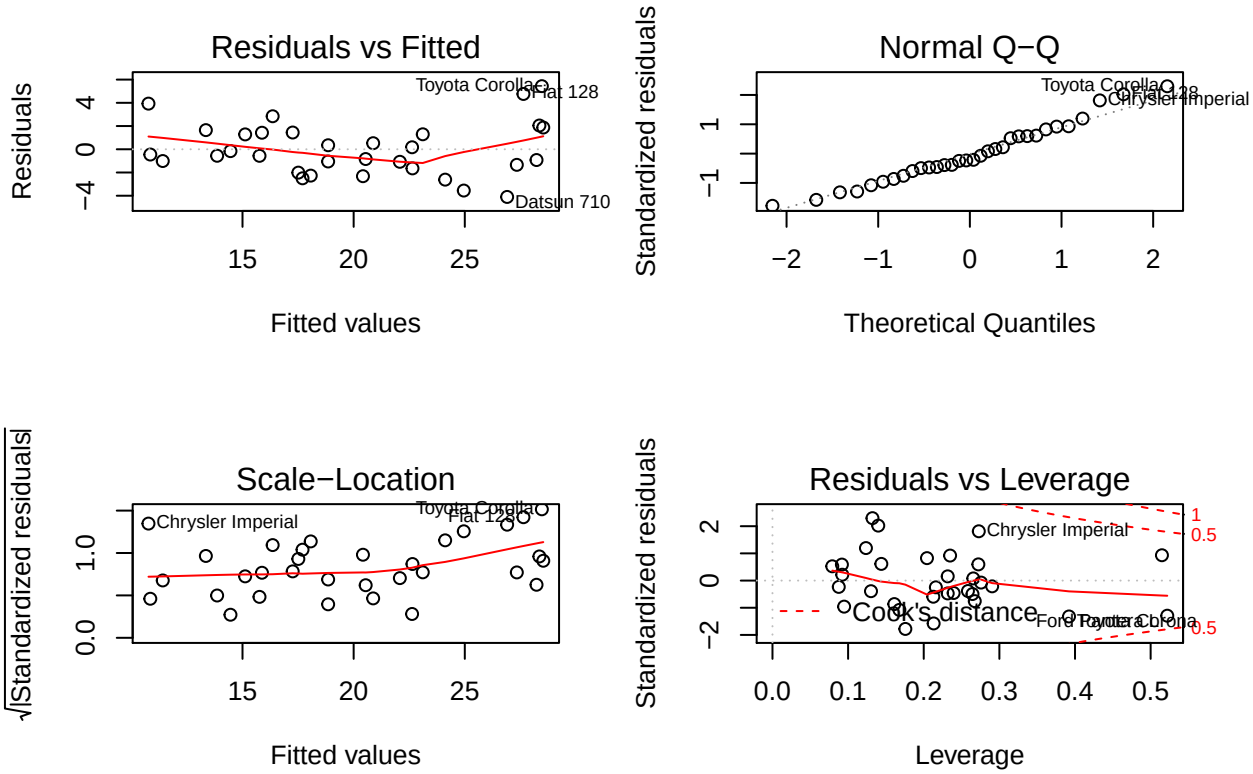
We performed a statistical study of the data and we can conclude that MPG for automatic cars presents lower values than the MPG for the manual cars.

Appendices

Appendix 1: boxplot



Appendix 2


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
Leverage
Ford Pantera L