Automatic or manual transmission: A regression analysis

Tuesday, September 16, 2014

Executive Summary

For this project, I work for Motor Trend, a magazine about the automobile industry. I am interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). I am particularly interested in the following two questions: 1. Is an automatic or manual transmission better for MPG? 2. How different is the MPG between automatic and manual transmissions??

Using OLS regression analysis, I determined that there is a signficant difference between the mean MPG for automatic and manual transmission cars. Manual transmissions achieve an increase og approximately 1.8 MPG when switching from an automatic transmission to a manual one, with all other variables held constant.

Data Analysis

I began my analysis by loading the mtcars dataset and reviewing the variables.

```
data(mtcars)
str(mtcars)
```

For more efficient analysis, I transformed transforming the following 5 variables into factors:

```
mtcars$gear <- factor(mtcars$gear,levels=c(3,4,5),labels=c("3gears","4gears","5gears"))
mtcars$cyl <- factor(mtcars$cyl,levels=c(4,6,8),labels=c("4cyl","6cyl","8cyl"))
mtcars$am <- factor(mtcars$am,levels=c(0,1),labels=c("Automatic","Manual"))
mtcars$vs <- factor(mtcars$vs)
mtcars$carb <- factor(mtcars$carb)
```

Exploratory Analysis

Next, I explored relationships between variables of interest.

a. I computed the correlation matrix with all 11 varaiables.

```
data(mtcars)
cor.out <- sort(cor(mtcars)[,1])
round(cor.out, 3)
```

b. I plotted the the relationship between all the variables of the dataset, which indicated that the variables cyl, disp, hp, drat, wt, vs and am have a strong correlation with mpg (Appendix - Figure 1).

c. I prepared a box plot for level of am (Automatic or Manual). The plot indcates that manual transmissions tend to have higher mpg. This data is further analyzed and discussed in regression analysis section by fitting a linear model (Appendix - Figure 2).

Regression Analysis

In this section, I describe how I built the linear regression models using different variables in order to find the best fit and compare it with the base model which I have using anova. After selecting a model, I performed an analysis of residuals.

I began with base model containing only am as the predictor variable.

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

Next, I developed an inclusive model that included all variables as predictors of mpg.

```
inclusivemodel <- lm(mpg ~ ., data = mtcars)
summary(inclusivemodel)
```

Next, I performed stepwise model selection in order to select significant predictors for the final, best model. The step function will perform this selection by calling lm repeatedly to build multiple regression models and select the best variables from them using both forward selection and backward elimination methods using AIC algorithm. This ensures that the useful variables are included in the model while omitting ones that do not contribute significantly to predicting mpg.

```
bestmodel <- step(inclusivemodel, direction = "both")
summary(bestmodel)
```

Finally, I compared the base model and the best model to determine if the confounder variables (cyl, hp, and wt) contribute to the accuracy of the model.

```
anova(basemodel, bestmodel)
```

Model Residuals and Diagnostics

In this section, I have prepared the residual plots (Appendix - Figure 3) of the regression model along with computation of regression diagnostics for the linear model. This analysis was comepleted in order to examine the residuals and identify leverage points. An analaysis of the residual plots indicated:

1. The points in the Residuals vs. Fitted plot are randomly scattered on the plot, which verifies the condition of independence.

2. The Normal Q-Q plot consists of the points which mostly fall on the line, which indicates that the residuals are normally distributed.

3. The Scale-Location plot consists of points scattered in a constant band pattern, which indicates constant variance.

4. There were some distinct points of interest (outliers or leverage points) in the top right of the plots that may indicate values of increased leverage of outliers.

I computed regression diagnostics of the best model to identify leverage points. I computed the top three points in each case of influence measures. The data points with the most leverage in the fit are identfied by hatvalues().

```
leverage <- hatvalues(bestmodel)
tail(sort(leverage),3)
```

The data points that influence the model coefficients the most are given by the dfbetas() function.

```
influential <- dfbetas(bestmodel)
#tail(sort(influential[,6]),3)
```

The models of vehicles identifyied above are the same models identified with the residual plots.

Statistical Inference

I performed a t-test on the two subsets of mpg data (manual and automatic transmission).

```
t.test(mpg ~ am, data = mtcars)
```

Based on the t-test results, I rejected the null hypothesis that the mpg distributions for manual and automatic transmissions are the same.

Conclusions

Based on the analysis I can conclude that:

1. Cars with Manual transmission get 1.8 (adjusted for hp, cyl, and wt) more miles per gallon compared to cars with Automatic transmission.

2. mpg will decrease by 2.5 for every 1000 lb increase in wt.

3. mpg decreases only 0.32 with every increase of 10 in hp.

4. If cyl increases from 4 to 6 and 8, mpg (adjusted by hp, wt, and am) will decrease by a factor of 3 and 2.2 respectively.

Appendix

Figure 1 - Pairs plot for the dataset

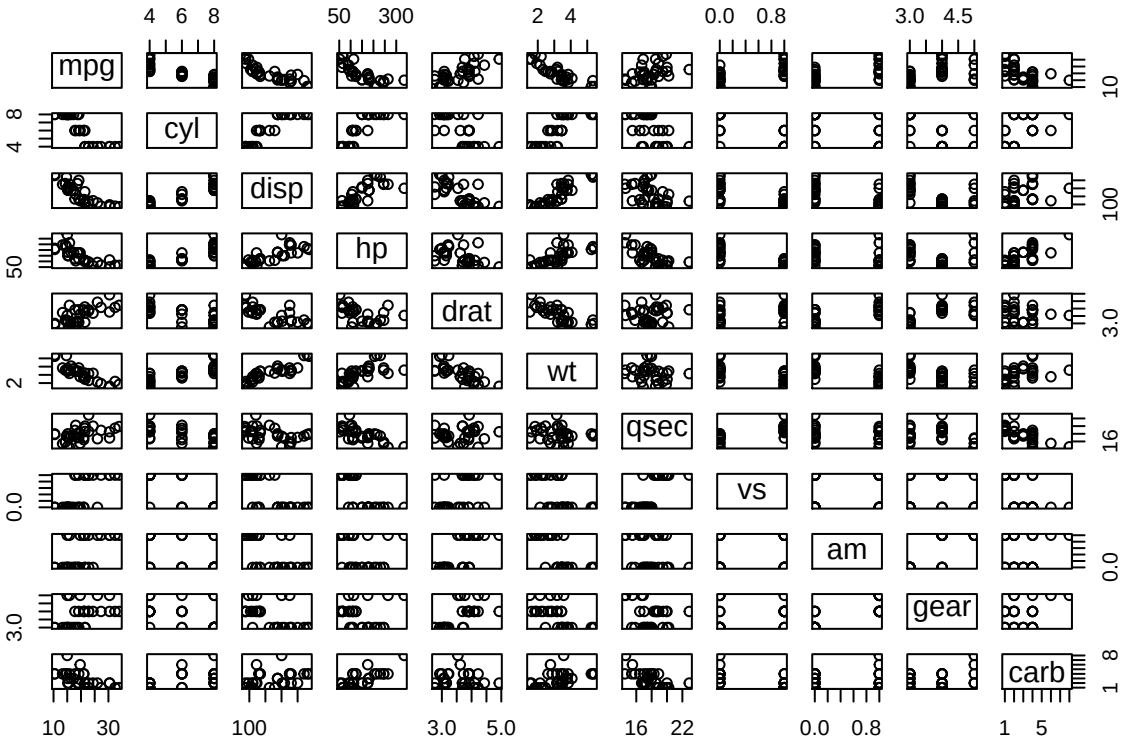


Figure 2 - Boxplot of miles per gallon by transmission type

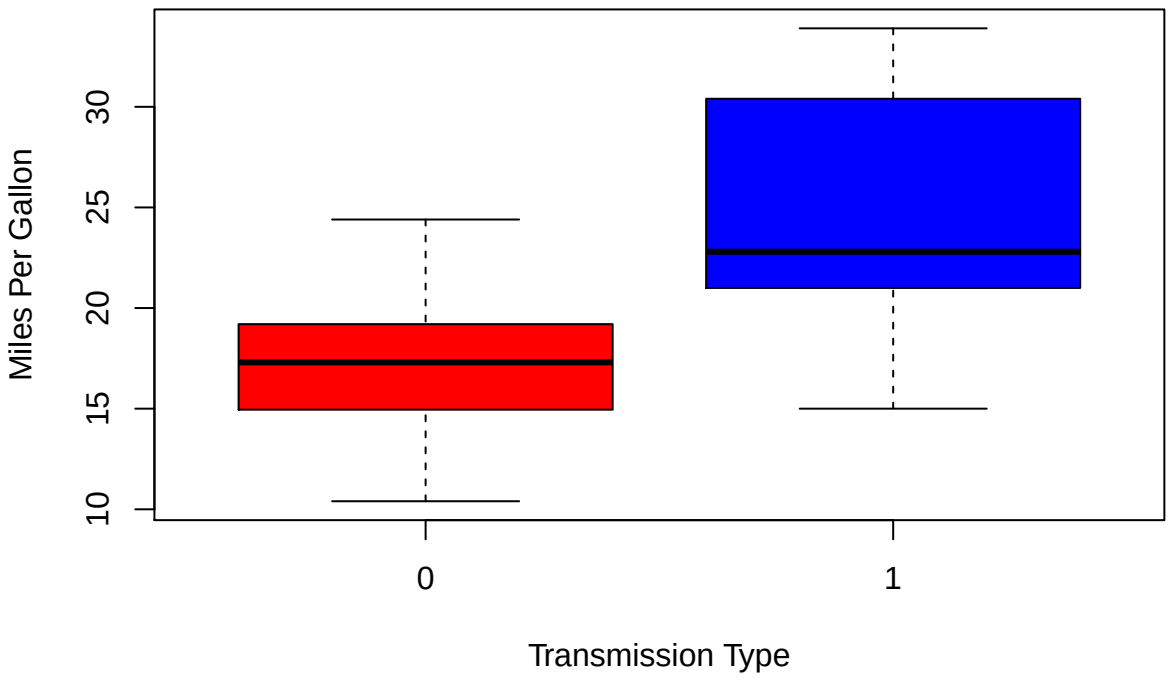
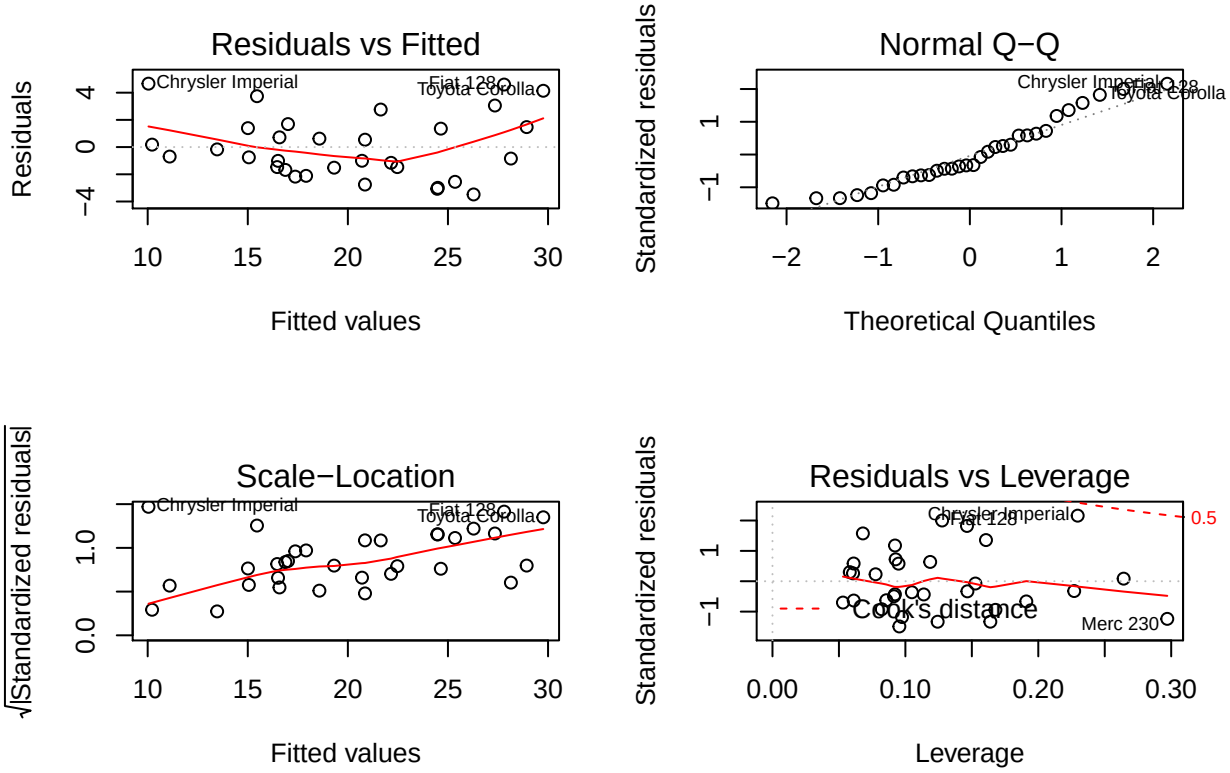


Figure 3 - Residual Plots


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
0.5
Merc 230
Leverage