Regression Models Class Project by T. Bhat

Regression Models Class Project by T. Bhat

Executive Summary:

Do cars with manual transmission behave more favorably than automatic transmission cars with respect to fuel efficiency? It is a common belief that changing gears manually results in better fuel management. We will answer: - Is an automatic or manual transmission better for miles per gallon (MPG)? - How different is the MPG between automatic and manual transmissions? Using hypothesis testing and simple linear regression, we determine that there is a signficant difference between the mean MPG for automatic and manual transmission cars, with the latter having **7.245** more MPGs on average using **coefficients** from the regression model. However, in order to adjust for other confounding variables such as the wt, hp and vs of the car, we run a multivariate regression to get a better estimate the impact of transmission type on MPG. After validating the models using ANOVA, the results from the multivariate regression model reveal that, on average, manual transmission cars get **1.976** miles per gallon more than automatic transmission cars using **coeffients**.

Uncertainty and Limitations:

The data is from 1974 and may not be valid for current car models. Also, the size of the dataset is small with 32 observations.

Exploratory Data Analysis:

The details of exploratory data analysis is in the appendix under Exploratory Data Analysis. We find that there are 11 variables describing 32 vehicles. Also the mpg for cars with manual transmission is higher than the cars with automated transmission. Is this a significant difference? We set our alpha value at 0.5 and run the t.test. With a p-value of 0.001374, we reject the null hypothesis and claim that there is a significant difference in the mean MPG between manual transmission and automatic cars.

Building Different Models

The detailed R code and plots are noted in the appendix for this section as mandated by the project course.

Linear Regression Model Using One Variable

Simple One Variable Linear Regression Model. In this model, we model mpg being predicted using transmission type. This model accounts for 36% of variance, so having just one variable is not sufficient as it doesnt account for all the variance.

Multivariate Linear Regression

We look for which variables correlate best with mpg, we look at the correlation table. This can be obtained using the R commands data(mtcars), cor(mtcars). From the table, we find that wt, cyl, disp, hp, am correlate highly with mpg. We see that cyl and disp are highly correlated with each other. We take out confounding variables and build out model with five variables am. cyl, wt, hp and vs. Similarly, vs and qsec are highly correlated with each other. Am and drat are also highly correlated with each other. This model with five variables accounts for 85% of the variance. This is much better than the model wiht one variable. Let us analyze the variances of the two models using Anova. Looking at the p-value of 6.227e-08, we see that the multivariate model is significantly different than the model with a single variable. Lastly, we check the residuals for non-normality and also examine the residuals vs. fitted values plot to spot for any signs of heteroskedasticity

Coeffients

The coefficients show the intercept and the difference in mpg between manual and automatic at 7.245 with one variable. The coeffients for multivariate model shows that manual transmission gets 1.976 mpg more than the automatic model.

Conclusion

Looking at the plots under multivariate regression models in the **Appendix**, we can see that our residuals are normally distributed and homoskedastic. This shows that our multivariate model is a reasonably good one and R^2 accounts for 85% of variance. We also found that in addition to am, we also have weight and vs contributing heavily as predictors for mpg and better than the univariate model.

Appendix

Exploratory Data Analysis

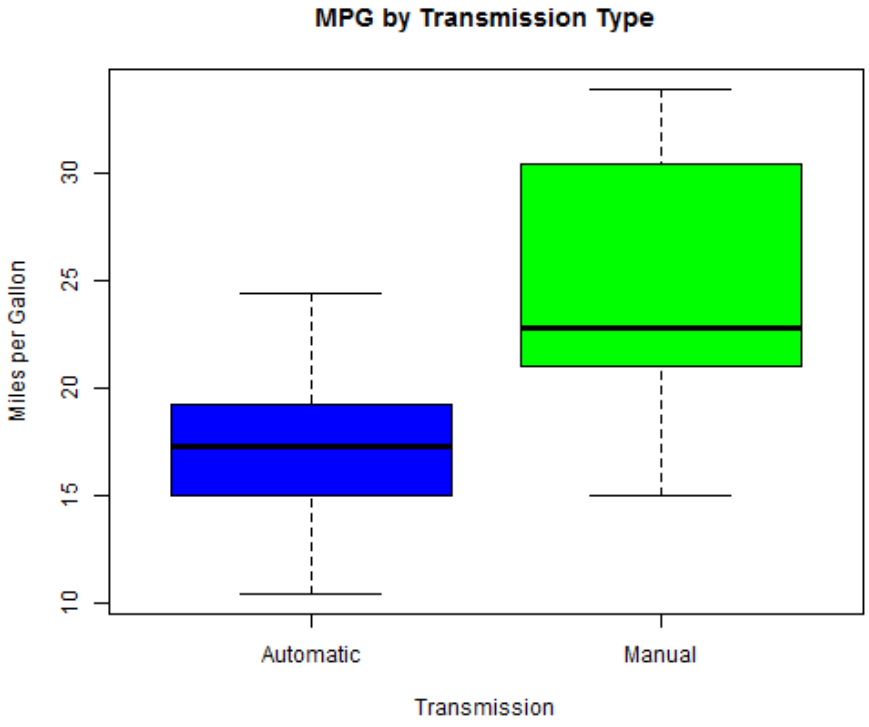
```
data(cars)
dim(mtcars)
```

```
## [1] 32 11
```

```
head(mtcars, n=1)
```

```
##           mpg cyl disp  hp drat   wt  qsec vs     am gear carb
## Mazda RX4  21   6  160 110  3.9 2.62 16.46  0 Manual    4    4
```

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
boxplot(mpg~am,data=mtcars,col = c("blue", "green"), xlab="Transmission", ylab="Miles per Ga
```




```
aggregate(mpg~am, data=mtcars, mean)
```

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

```
autoData <- mtcars[mtcars$am == "Automatic",]
manData <- mtcars[mtcars$am == "Manual",]
t.test(autoData$mpg, manData$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  autoData$mpg and manData$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:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

Building Different Models

Linear Regression Model Using One Variable

```
fit <- lm(mpg~am, data=mtcars)
summary(fit)$coefficients
```

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

Multivariate Regression Models

```
mfit <- lm(mpg~am+cyl+wt+hp+vs, data=mtcars)
summary(mfit)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 33.24159905 5.48527074  6.0601565 2.108602e-06
## amManual     1.97574750 1.64825342  1.1986916 2.414597e-01
## cyl         -0.40178727 0.79364098 -0.5062582 6.169415e-01
## wt          -2.54331718 0.93506164 -2.7199460 1.148231e-02
## hp          -0.02588661 0.01387176 -1.8661377 7.334148e-02
## vs           1.17066640 1.81282822  0.6457680 5.240886e-01
```

```
anova(fit,mfit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + wt + hp + vs
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 167.31  4    553.58 21.506 6.227e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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