Regression Models Project

Tihana Mirkovic

Executive Summary

The *Motor Trend* magazine is interested in exploring the relationship between a number of variables and miles per gallon (MPG) (outcome), for different automobiles (1973-1974 models). The data is found in the source file `mtcars`, which was originally extracted from the 1974 edition of the *Motor Trend* US magazine. It is comprised of fuel consumption (MPG) and 10 aspects of automobile design and performance for 32 automobiles.

In particular, in this analysis, two questions are going to be addressed:

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

Hypothesis testing and linear regression confirmed a statistical differences between the means of MPG values for cars with **automatic** (am = 0) and **manual** (am = 1) transmission, where cars with a manual transmission have scored 7.2 MPG higher than those with an automatic transmission. Multivariable regression analysis was then employed to analyze the impact of confounding variables, such as weight and qsec, on the dependence of the transmission type on MPG. The best model, with the highest R-squared value, indicates that in addition to qsec, it is primarily the weight parameter that has a significant impact when quantifying the MPG difference between cars with automatic and manual transmissions.

Exploratory Data Analysis

The data set `mtcars` is loaded, as well as the libraries used in this analysis. Following the initial inspection of the variables, the class of a number of dichotomous parameters is converted from numeric to factor.

```r
library(dplyr)
library(ggplot2)
library(gridExtra)
data(mtcars)
```

```r
##str(mtcars) - output not shown for space reasons
```

```r
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

Inference: Analyzing MPG Means for Automatic and Manual Cars Insight into the two subpopulations (automatic vs manual) can be gaged from the plots in **Figures 1 and 2** of the Appendix. Then a t-test is performed to test if the two means are statistically different. We are assuming that MPG has a normal distribution, as illustrate in **Figure 1** of the Appendix. The null hypothesis, H0, is assuming that the two means are not different and that they stem from the same population.

The averages and the standard deviation of MPG values for the two subgroups of cars (automatic and manual) are compared, where the mean MPG value for manual cars is 7.24 MPGs larger.

```
t.test(mtcars$mpg~mtcars$am, conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$am
## 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 in group Automatic    mean in group Manual
##                17.14737                24.39231
```

The mean MPG value for manual cars is 24.39231 MPG (7.24 MPGs larger than for automatic transmission cars). The t-test, with a p-value=0.001374, leads to the rejection of the null hypothesis, and suggestions that the two means are indeed different and that manual cars are outperforming automatic cars, given all other factors are identical. The 95% confidence interval suggests that the difference in the means of MPG values is between 3.21 and 11.28 MPGs.

Regression Analysis In the previous section we have shown that the transmission type has an impact on the MPG value, but a further nine automobile parameters could also potentially have an influence on MPG values. The correlation between MPG and all other car parameters is investigated first, showing how some have a positive and some a negative correlation with MPG. A quick way to visualize some of the dependencies among the different variables is to examine the scattterplot matrix (**Figure 3**).

```
##data(mtcars)
##sort(cor(mtcars)[1,]) - output not shown for space reasons
```

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

Linear Regression

```
##              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
```

From a linear regression model, where only transmission type is included, we don't learn much more than in our previous analysis. The coefficients suggest that:

- intercept = the mean MPG value for automatic transmission is 17.147
- am coefficient = mean MPG value for manual transmission is: 17.147+7.245=24.392.

The p-value suggests that inclusion of the transmission type into the model is significant, but the low R-squared value means that this approach only explains about 36% of the variance. A better model is needed.

Multivariable Regression Analysis

Model with all Parameters Inclusion of all the parameters, would be modeled with `summary(lm(mpg~., data=mtcars))`, but that would probably lead to overfitting. A better model, with more carefully selected parameters is needed.

The Best Model To find the best model, we use the step() function, based on a stepwise algorithm approach, which picks the most dominant parameters needed to model MPG.

```r
bestmodel = step(lm(data = mtcars, mpg ~ .), trace=0)
summary(bestmodel)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

The three most dominant parameters to describe MPG values for automobiles are weight of the car, qsec (1/4 mile time) and transmission type. Inclusion of these three variables into the model accounts for 85% of the variance. The coefficients suggest that an increase in weight by 1000 lb will result in a decrease of MPG by 3.9165, also, in this model, on average it is predicted that a change from automatic to manual transmission would increase the MPG value by 2.9358, whereas every increase of 1/4 mile time, would also increase MPG by 1.2259. The dominance of the weight parameter is illustrated in **Figure 4**.

Residuals and Diagnostics

The residual diagnostic plots for the bestmodel is shown in **Figure 5**. The Residuals vs Fitted plot indicates random scattering and supports the independence assumption. In the Normal Q-Q plot, points are found close to the line, indicating that the residuals are normally distributed. The values on the Scale-Location plot are within a closed band. indicate that the residuals are normally distributed (as seen on the Normal Q-Q plot), whereas no systematic pattern is observed in the Residuals vs Fitted plot. There also don't seem to be any extreme outliers. As apparent from the plots, there are some outliers which can be identified through leverage and influence measures.

```
influence<-sort(round(dfbetas(bestmodel)[, 4],3))
tail(influence)
```

```
## Cadillac Fleetwood          Mazda RX4       Lotus Europa
##              0.101              0.136              0.210
##   Dodge Challenger        AMC Javelin      Maserati Bora
##              0.295              0.350              0.525
```

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

```
##       Mazda RX4 Wag  Cadillac Fleetwood   Chrysler Imperial
##                0.250               0.250               0.261
##       Toyota Corona Lincoln Continental       Maserati Bora
##                0.278               0.294               0.471
```

Conclusion

Our analysis suggests that cars with a manual transmission have a better "miles per gallon" value than those cars with an automatic transmission. However, the best model that was found, also suggests that the car weight (wt), as well as qsec confound the relationship between MPG and the transmission type. This model, including the three parameters, was able to explain 85% of the variance, and actually showed how primarily weight, but also to a certain degree qsec was more significant for the prediction of MPG values than transmission type. The expected change in mpg by comparing automatic and manual cars is 2.9358 MPGs in this new model, considerably less than the 7.245 value extracted from linear regression when weight and qsec were not taken into account. In conclusion, quantification of the MPG difference between automatic and manual transmissions requires a multivariable regression model, and variables beyond the transmission type need to be taken into account.

Appendix: Figures

Figure 1 The distribution of MPG is illustrated on a histogram. For comparison, a density plot and a normal curve have been overlaid on top of the histogram. On the right, separate histograms illustrate the distribution of MPG values among the two subgroups of cars differentiated by transmission type.

```
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
g1<-ggplot(mtcars, aes(x=mpg)) + geom_histogram(aes(y=..density..),
breaks = seq(min(mtcars$mpg), max(mtcars$mpg), (max(mtcars$mpg)-min(mtcars$mpg))/10),
fill=I("grey87"),col=I("black"))+
labs(title="Histogram and Density Plot of Miles per Gallon", x="Miles per Gallon", y="Density")+
stat_density(geom="line", col="RosyBrown2", lwd=2, aes(linetype="Density Plot"), show.legend=TRUE)+
stat_function(fun=dnorm, col="DarkGoldenrod",args=list(mean=mean(mtcars$mpg), sd=sd(mtcars$mpg)),
lwd=2, aes(linetype="Normal Curve"), show.legend = TRUE)+
scale_color_manual(values=c("Density Plot","Normal Distribution"))+
scale_linetype_manual(name="Curves/Fits", values=c(1,1))+
guides(linetype=guide_legend(override.aes=list(colour = c("RosyBrown2","DarkGoldenrod"))))+
theme_bw()+
theme(legend.justification=c(1,1), legend.position=c(1,1))+
        theme(plot.title = element_text(size = 8))
g3<-ggplot(mtcars, aes(x=mpg)) +
        geom_histogram(breaks = seq(min(mtcars$mpg), max(mtcars$mpg),
        (max(mtcars$mpg)-min(mtcars$mpg))/10),
        fill=I("grey87"),col=I("black"))+facet_grid(am~.)+
        labs(title="Histogram of Miles per Gallon Values by Transimssion Kind",
        x="Miles per Gallon", y="Frequency")+
        theme_bw()+
        theme(plot.title = element_text(size = 8))
grid.arrange(g1, g3, ncol=2)
```

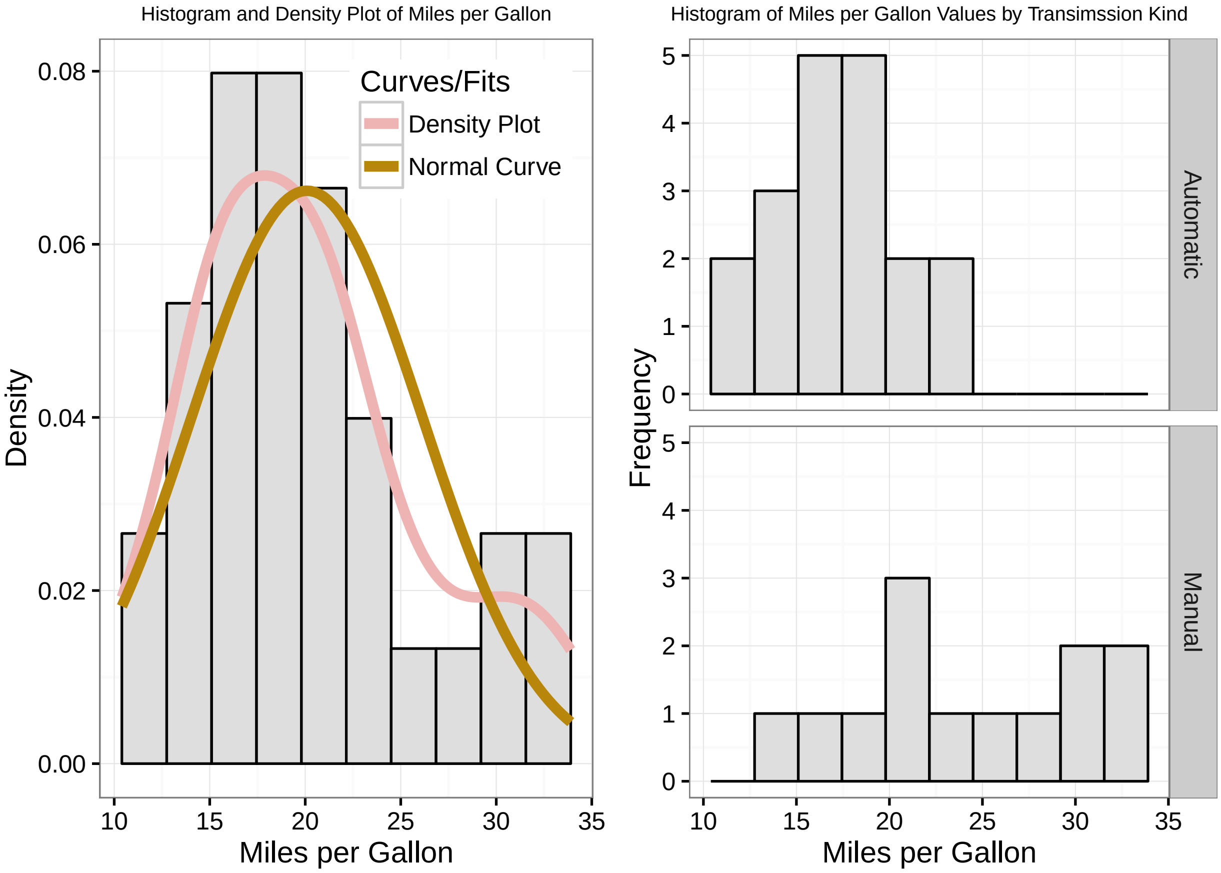


Figure 2 A violin plot and a box plot illustrate the dependence of MPG on the transmission type, indicating higher MPG values for the cars with a manual transmission.

```
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
##Violin plot
g = ggplot(data = mtcars, aes(y = mpg, x = am, fill = am))
g = g + geom_violin(colour = "black", size = 1)
g = g + xlab("Transmission type") + ylab("Miles per Gallon")+theme_bw()
g= g+ggtitle("Miles per Gallon by Transmission Type")+
        theme(plot.title = element_text(size = 10))
g2=ggplot(data = mtcars,aes(y = mpg, x = am, fill = am))
g2=g2+geom_boxplot()
g2=g2+ggtitle("Miles per Gallon Value by Transmission Type")
g2=g2+xlab("Transmission type")+ ylab("Miles per Gallon")+theme_bw()+
        theme(plot.title = element_text(size = 10))
grid.arrange(g, g2, ncol=2)
```

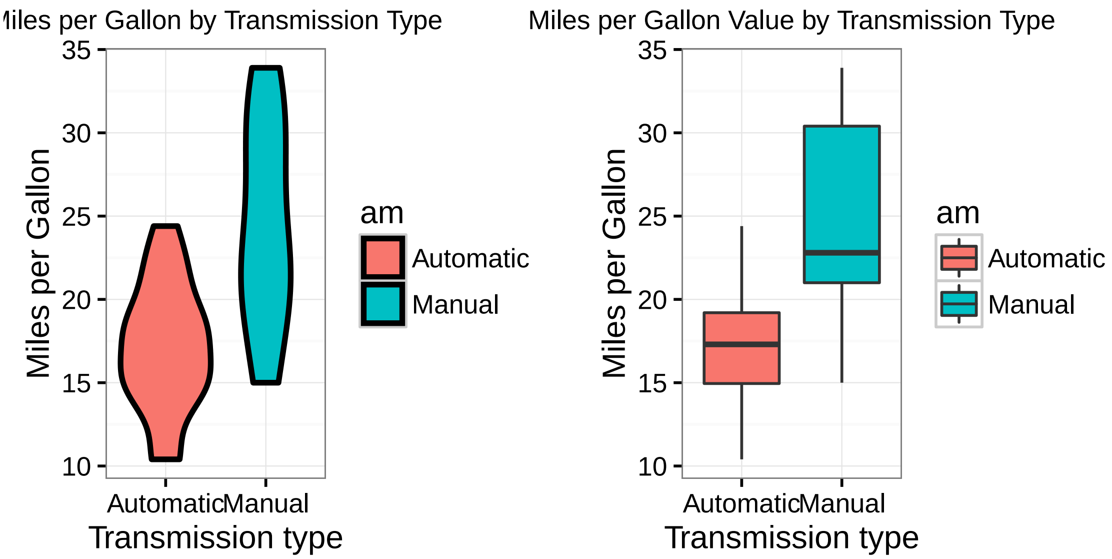


Figure 3 Scatter Plot Matrix Illustrating Correlations Among mtcars Predictor Variables

```
pairs(mtcars)
```

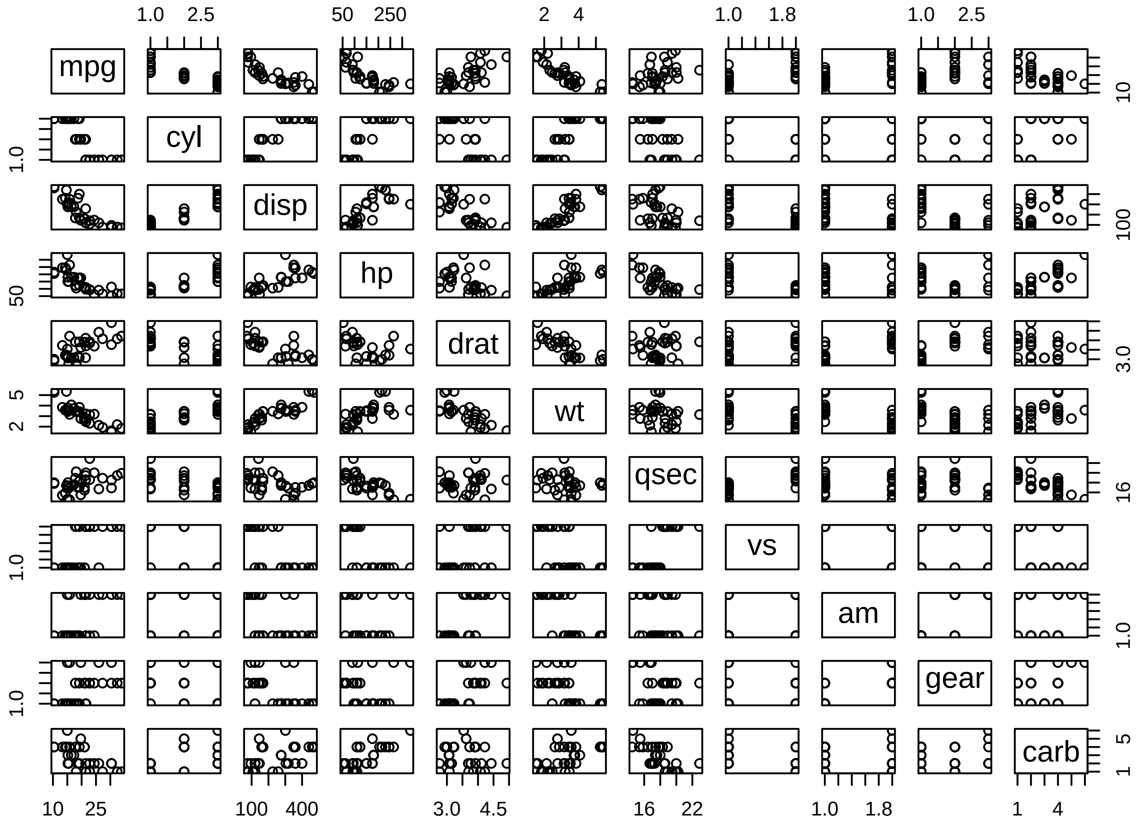


Figure 4 Showing the dominance of weight as a factor and also illustrating how manual cars tend to be lighter, whereas automatic cars tend to be heavier. Lighter cars have much better MPG values.

```
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
ggplot(mtcars, aes(x=wt, y=mpg))+geom_point(aes(colour=factor(am)), size=2)+
theme_bw()+ ggtitle("Miles per Gallon by Weight and Transmission")+ xlab("Weight(tons)")+
ylab("Miles per Gallon")+theme(plot.title = element_text(size = 10))
```

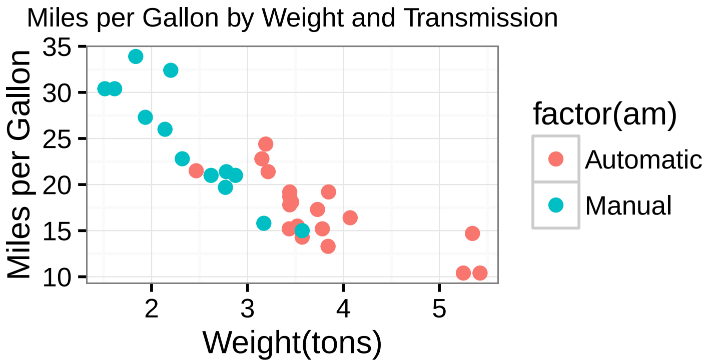
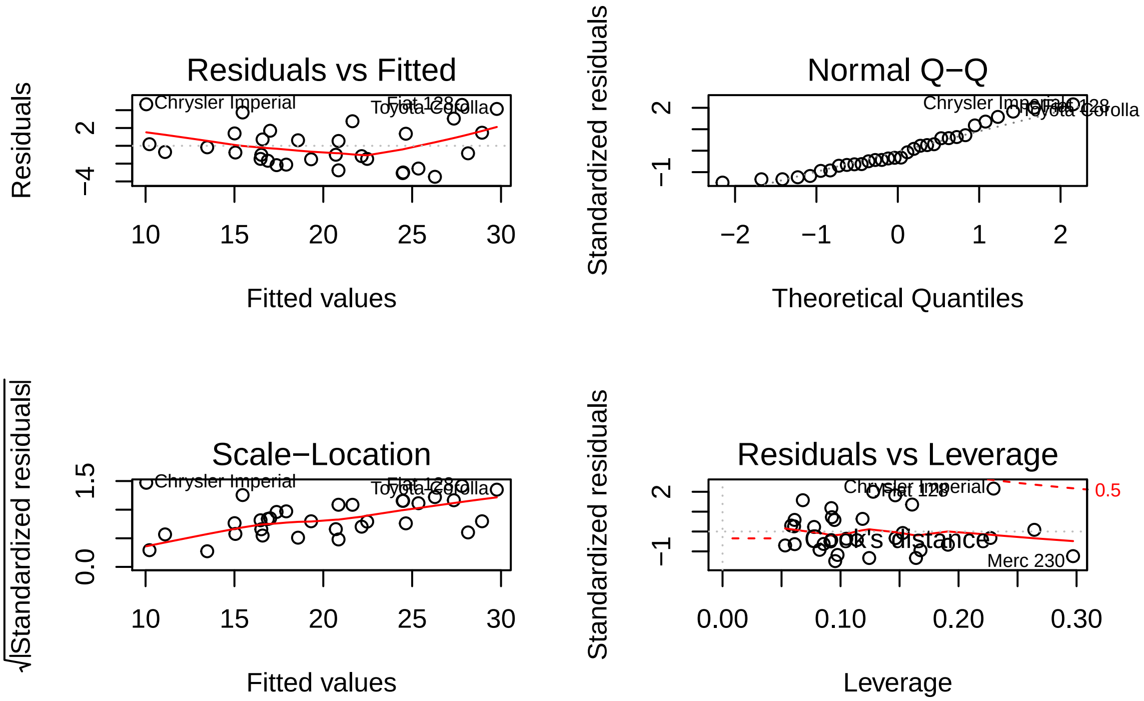


Figure 5 Residual plots for the best model.

```
data(mtcars);par(mfrow=c(2,2))
bestmodel<-lm(mpg ~ wt + qsec + am, data=mtcars)
plot(bestmodel)
```


Residuals vs Fitted
Residuals
2
−4
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
2
−1
−2
−1
0
1
2
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Toyota Corolla
Scale−Location
√|Standardized residuals|
1.5
0.0
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals vs Leverage
Standardized residuals
2
−1
0.00
0.10
0.20
0.30
Leverage
Chrysler Imperial
Fiat 128
Cook's distance
Merc 230
0.5