Regression Model - Motor Trend Analysis

Author: John Joyce

Date: November 21, 2017

Summary

This analysis investigates the relationship between a set of variables and miles per gallon (MPG) for a collection of cars (mtcars dataset in R).
In particular, this analysis answers the following questions:

```
1) Is an automatic or manual transmission better for MPG?
2) What is the MPG delta between automatic and manual transmissions?
```

This analysis concludes that the manual transmission MPG is greater than the automatic transmission MPG by 2.935.

Data Analysis

This section loads the ggplot2 and MASS libraries as well as the mtcars data set.
This section also provides a high-level review of the mtcars data set and illustrates the delta of the Automatic Transmission mean MPG and the Manual Transmission mean MPG.

```
##  Load the ggplot2 library to manipulate and display plots.
##  Load the MASS library to perform a stepAIC multivariate regression analysis.

    library(ggplot2)
    library(MASS)

##  Load the mtcars data set.

    data(mtcars)
```

```
##  Review the data variable types in the mtcars data set.

    str(mtcars)
```

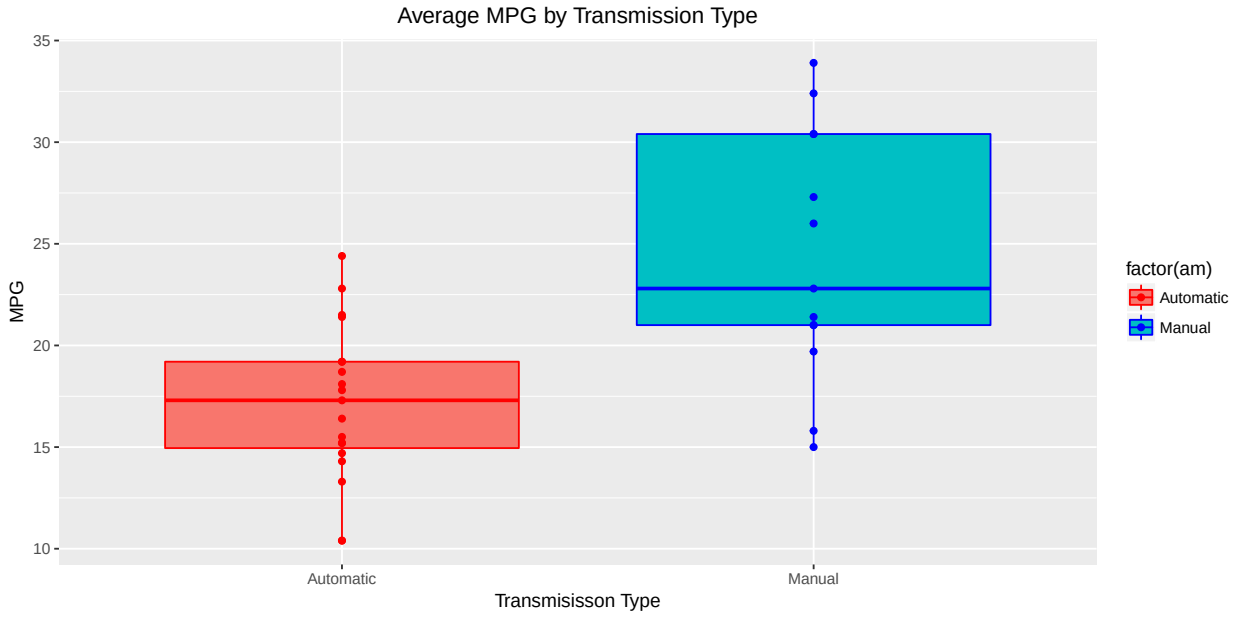
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
##  Review the first few rows of data in the data set.

    head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
##  Create a factor variable for various metrics in the mtcars data set.

    mtcars$cyl <- factor(mtcars$cyl,labels=c("4","6","8"))
    mtcars$am <- factor(mtcars$am,labels=c("Automatic", "Manual"))

    automatic_df <- mtcars[mtcars$am=="Automatic",]
    manual_df <- mtcars[mtcars$am=="Manual",]

##  Aggregate the data to find the overall average MPG by Transmission type.

    mpg_avg <- aggregate(mpg~am,mtcars,mean)
```

The following chart illustrates the average MPG for automatic and manual transmission types:



Based on this high-level review, we suspect that the Manual Transmission achieves higher MPG than Automatic Transmission by 7.245.

Hypothesis

This section provides an analysis to determine if the hypothesis that Manual Transmission achieves higher MPG than Automatic Transmission by 7.245. Perform a t-test to determine the significance of the delta.

```
##  Perform a t-test using the Automatic MPG and the Manual MPG data to determine the significance of
##  this delta.

    ttest_df <- t.test(automatic_df$mpg,manual_df$mpg)
    ttest_df
```

```
##
##  Welch Two Sample t-test
##
## data:  automatic_df$mpg and manual_df$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
```

Based upon this t test, the calculated p-value is 0.0013736.
This is considered to be significant since it is less than .05 using 95% confidence intervals. Since the delta in MPG data is found to be significant, the next section proceeds with determining a model to explain the data.

Regression Analysis - Initial Linear Model

This section provides a regression analysis aimed at providing a model to explain the mtcars data.

```
##  Create a linear model for the Automatic MPG and the Manual MPG data.
##  Provide a summary of the linear model.

    mpg_model <- lm(mpg~am,mtcars)
    summary(mpg_model)
```

```
##
## 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
```

```
##  Calculate and report the 95% confidence interval to verify the significance of the MPG delta.

    confint_mpg_model <- confint(mpg_model,level=.95)
    confint_mpg_model[2,]
```

```
##    2.5 %   97.5 %
##  3.64151 10.84837
```

Based on these linear model and confidence interval tests, we observe the following:

1) Average Automatic Transmission MPG < Manual Transmission by 7.245.

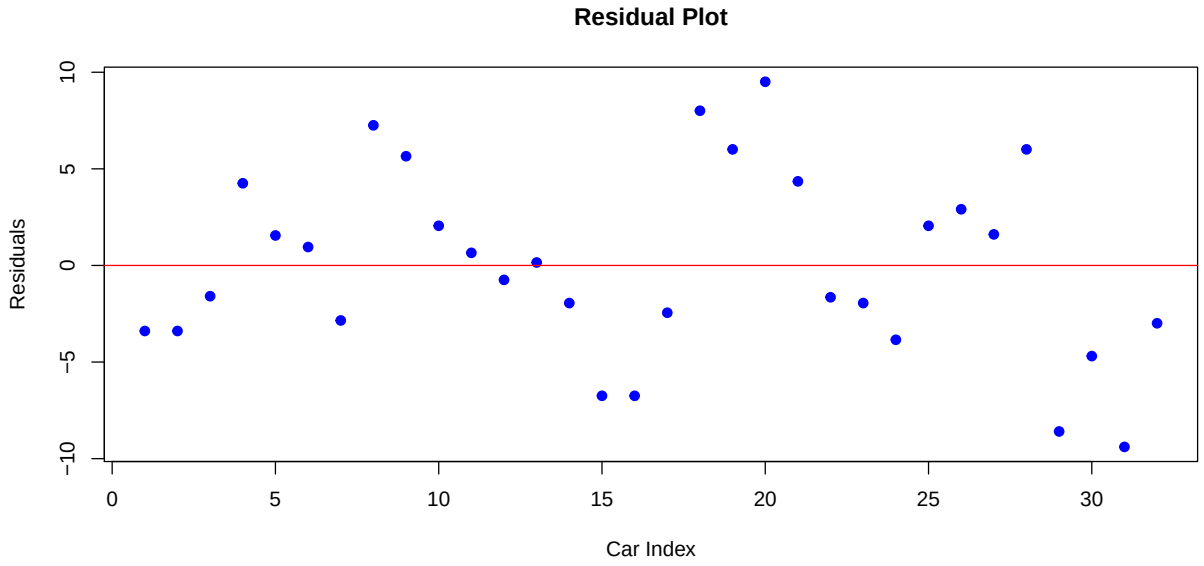
2) The p value is small (<.05).

3) The confidence interval does not contain zero.

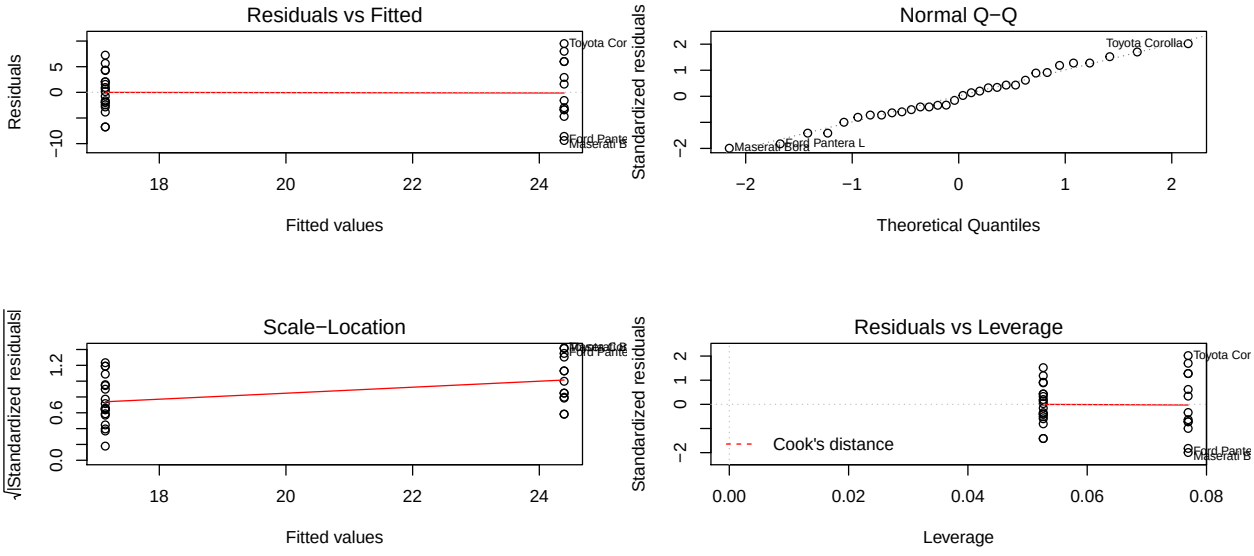
4) Based upon p-value and confidence interval, the alternate hypothesis (delta in MPG between automatic and manual transmission) is accepted.

5) $R^2 = 0.36$. It appears that this model could be improved so we will explore the relationship of other factors.

```
##  Create a variable to store the mpg_model residual values.

    mpg_model_resid <- mpg_model$residuals

##  Plot the residuals for the linear model.

    plot(mpg_model_resid,col="blue",pch=19,main="Residual Plot",xlab="Car Index",ylab="Residuals")
    abline(0,0,col="red")
```




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



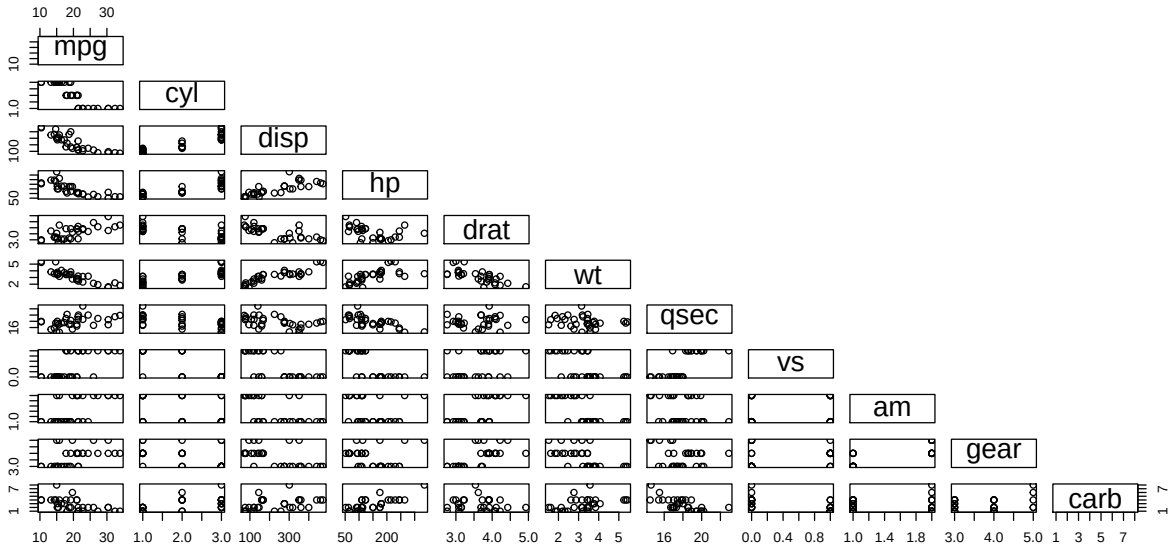
```
##  Convert the mpg_model residuals to a data frame to view various data points.

    mpg_model_df <- data.frame(mpg_model$residuals)

##  Review the first few rows of residual data.

    head(mpg_model_df)
```

```
##                   mpg_model.residuals
## Mazda RX4                  -3.3923077
## Mazda RX4 Wag              -3.3923077
## Datsun 710                 -1.5923077
## Hornet 4 Drive              4.2526316
## Hornet Sportabout           1.5526316
## Valiant                     0.9526316
```

Regression Analysis - Improved Multivariate Model Estimation

To try and improve the model, we will perform a multivariate regression analysis that includes other factors of mtcars data. First, we will look to see which independent variables are correlated with the MPG variable.

```
##  Create a pairs plot to find the variables which would have a dependency on the MPG.

    pairs(mpg~.,mtcars,upper.panel=NULL)
```



At first glance, it appears like the transmission (am), number of cylinders (cyl), displacement (disp), horsepower (hp), and weight (wt) have the strongest correlation to MPG so these parameters are used to derive a new multivariate model.

```
##  Create a multivariate model using transmission type, cylinders, and weight as independent
##  variables.

    mpg_model2 <- lm(mpg~am+cyl+disp+hp+wt,data=mtcars)
    summary(mpg_model2)
```

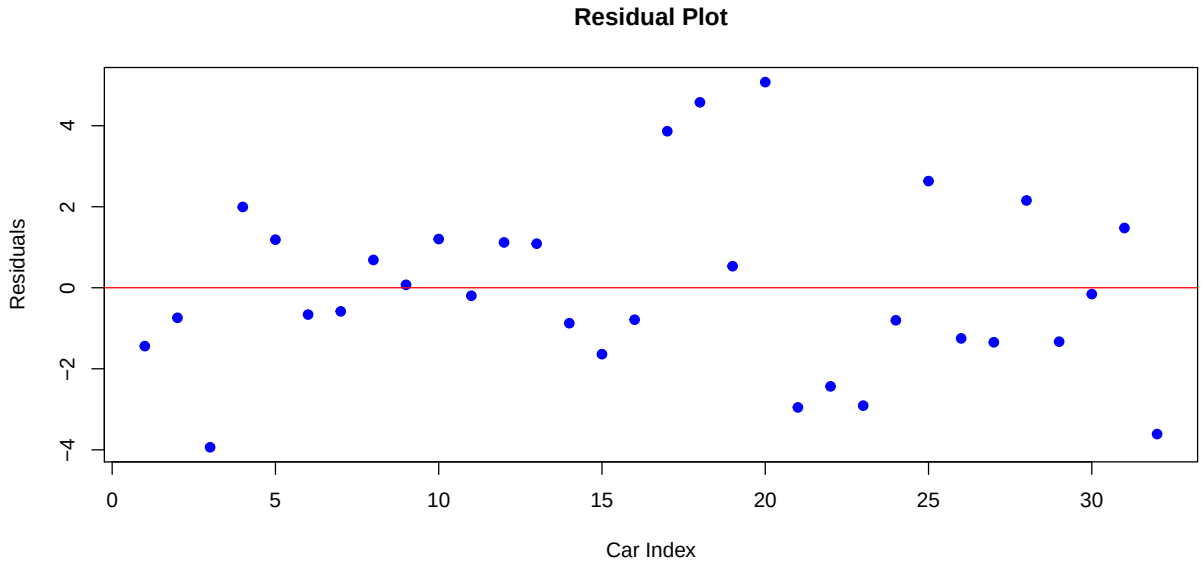
```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## amManual     1.806099   1.421079   1.271   0.2155
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
```

```
## hp          -0.032480   0.013983  -2.323   0.0286 *
## wt          -2.738695   1.175978  -2.329   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

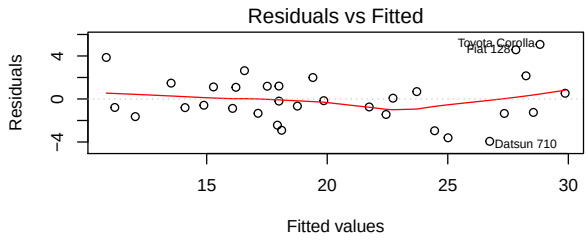
```
##  Use the anova function to compare the mpg_model and mpg_model2 models.

    anova(mpg_model,mpg_model2)
```

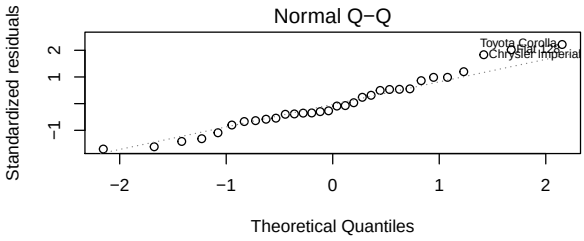
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     25 150.41  5    570.49 18.965 8.637e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

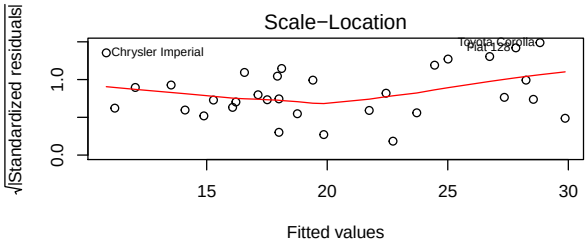
```
##  Create a variable to store the mpg_model2 residual values.

    mpg_model2_resid <- mpg_model2$residuals

##  Plot the residuals for the multivariate model mpg_model2.

    plot(mpg_model2_resid,col="blue",pch=19,main="Residual Plot",xlab="Car Index",ylab="Residuals")
    abline(0,0,col="red")
```

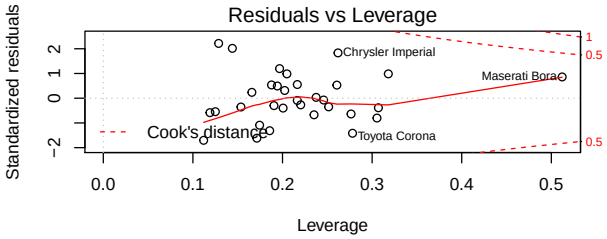



```
    par(mfrow=c(2,2))
    plot(mpg_model2)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710


Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial


Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128


Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Maserati Bora
Toyota Corona
Cook's distance
1
0.5
0.5

Regression Analysis - Best Fit Model using stepAIC

One final alternative approach to finding the best fit model utilizes the stepAIC function within the MASS library.

```
##  Create a multivariate model using stepAIC function and report the analysis of variance results.

    mpg_model3 <- lm(mpg~.,data=mtcars)
    step <- stepAIC(mpg_model3)
```

```
##  Report the stepAIC analysis of variance results.

    step$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
## Final Model:
## mpg ~ wt + qsec + am
##
##
##     Step Df     Deviance Resid. Df Resid. Dev      AIC
## 1                               20   133.3235 69.66535
## 2 - drat  1  0.001646814        21   133.3251 67.66575
## 3 - gear  1  1.857511109        22   135.1826 66.10850
## 4   - vs  1  4.250437656        23   139.4330 65.09916
## 5 - carb  1  2.897542867        24   142.3306 63.75733
## 6 - disp  1  1.651140725        25   143.9817 62.12642
## 7  - cyl  2 16.084729691        27   160.0665 61.51530
## 8   - hp  1  9.219469347        28   169.2859 61.30730
```

Using the stepAIC model results in using the weight (wt), 1/4 mile time (qsec), and transmission type (am) variables to find the best fit model. Next, we will re-run the analysis using these three regressors.

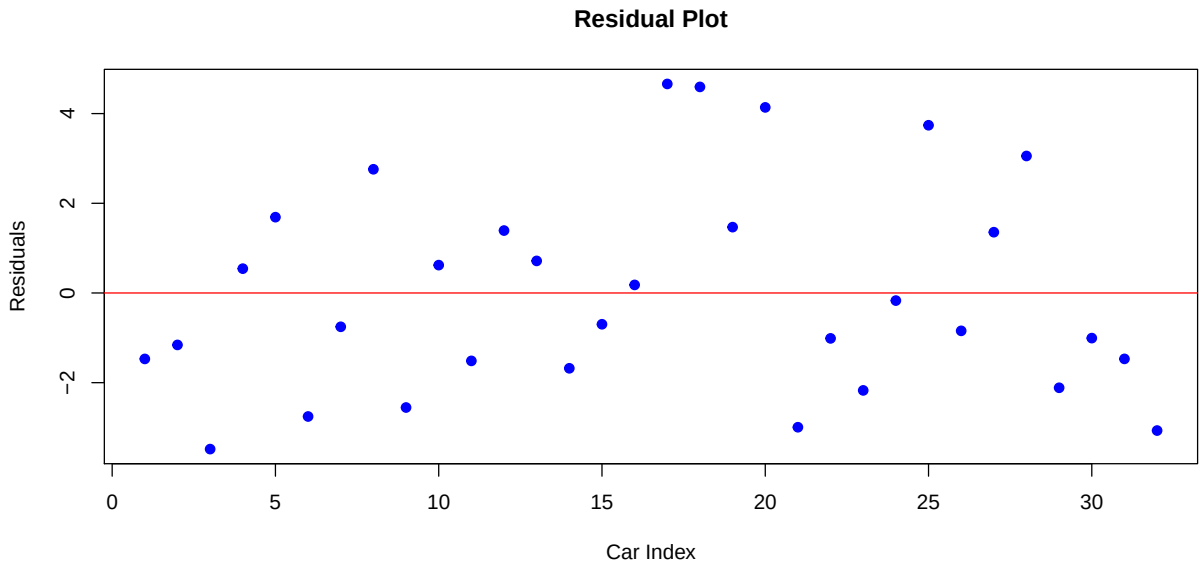
```
##  Create a multivariate model using weight (wt), 1/4 mile time (qsec), and transmission
##  type (am) variables.
##  Use the anova function to compare the mpg_model, mpg_model2, and mpg_model3 models.

    mpg_model3 <- lm(mpg~wt+qsec+am,data=mtcars)
    summary(mpg_model3)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

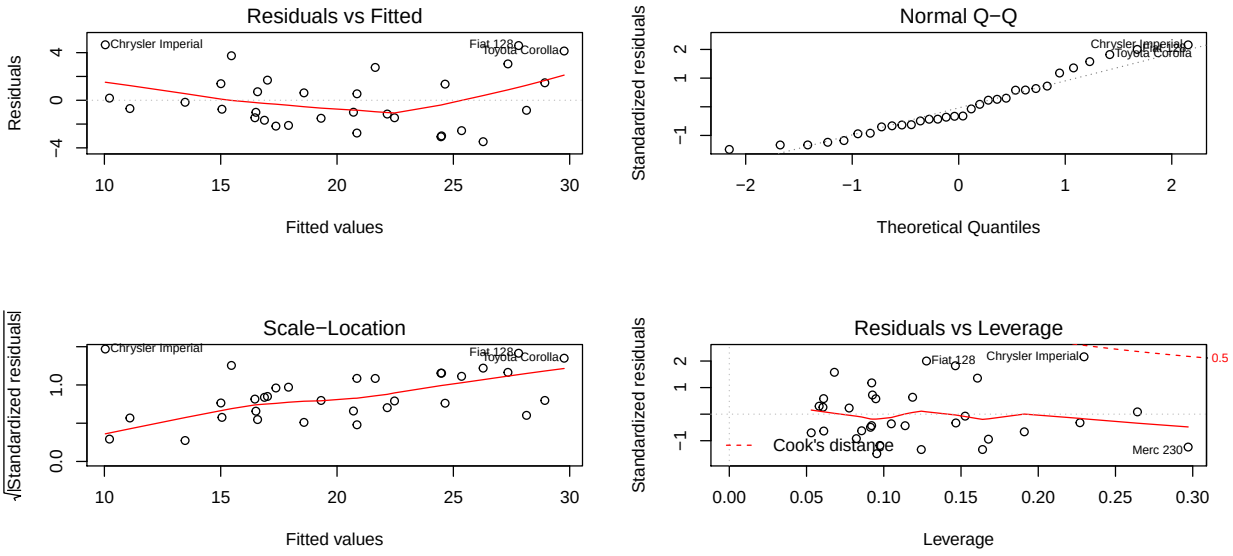
```
    anova(mpg_model,mpg_model2,mpg_model3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + hp + wt
## Model 3: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     25 150.41  5    570.49 18.9646 8.637e-08 ***
## 3     28 169.29 -3    -18.88  1.0459    0.3896
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##  Create a variable to store the mpg_model3 residual values.

    mpg_model3_resid <- mpg_model3$residuals

##  Plot the residuals for the multivariate model mpg_model3.

    plot(mpg_model3_resid,col="blue",pch=19,main="Residual Plot",xlab="Car Index",ylab="Residuals")
    abline(0,0,col="red")
```




```
par(mfrow=c(2,2))
plot(mpg_model3)
```



Conclusion

This study performed three different regression analyses to answer the following questions:
1) Is an automatic or manual transmission better for MPG?
2) What is the MPG delta between automatic and manual transmissions?

Based upon the various R^2 and p values, it appears that analysis #3 has the best fit:

```
##  Store and report all the R^2 and p-value data for all three regression analyses into a
##  summary data frame.

    r_squared <- cbind(summary(mpg_model)$adj.r.squared,
        summary(mpg_model2)$adj.r.squared,summary(mpg_model3)$adj.r.squared)
    p_value <- cbind(summary(mpg_model)$coef[2,4],
        summary(mpg_model2)$coef[2,4],summary(mpg_model3)$coef[4,4])
    summary_df <- data.frame(rbind(r_squared,p_value))
    colnames(summary_df) <- c("Analysis #1", "Analysis #2", "Analysis #3")
    rownames(summary_df) <- c("R Squared", "p Value")
    summary_df
```

```
##            Analysis #1 Analysis #2 Analysis #3
## R Squared 0.3384589082   0.8343702  0.83355608
## p Value   0.0002850207   0.2154510  0.04671551
```

```
##  Calculate the delta in MPG based upon the coefficients listed in the mpg_model3 summary.

     mpg_delta <- round(summary(mpg_model3)$coefficients[4,1],4)
```

Therefore, this study concludes that that manual transmission MPG is greater than the automatic transmission MPG by 2.9358.

Raw Code

```
##  Load the ggplot2 library to manipulate and display plots.
##  Load the MASS library to perform a stepAIC multivariate regression analysis.

    library(ggplot2)
    library(MASS)

##  Load the mtcars data set.

    data(mtcars)

##  Review the data variable types in the mtcars data set.

    str(mtcars)

##  Review the first few rows of data in the data set.

    head(mtcars)

##  Create a factor variable for various metrics in the mtcars data set.

    mtcars$cyl <- factor(mtcars$cyl,labels=c("4","6","8"))
    mtcars$am <- factor(mtcars$am,labels=c("Automatic", "Manual"))

    automatic_df <- mtcars[mtcars$am=="Automatic",]
    manual_df <- mtcars[mtcars$am=="Manual",]

##  Aggregate the data to find the overall average MPG by Transmission type.

    mpg_avg <- aggregate(mpg~am,mtcars,mean)

##  Generate a ggplot2 using the following parameters for the plot:
##      -   g:                      variable for initial plot layer
##      -   aes:                    aesthetics (x: transmission - am, y: mpg)
##      -   geom_boxplot color:      automatic = "red", manual = "blue"
##      -   geom_point color:        automatic = "red", manual = "blue"
##      -   x axis label (labs):    "Transmission Type"
##      -   y axis label (labs):    "MPG"
##      -   main title (labs):      "Average MPG by Transmission Type"
##      -   theme element:          center: element_text(hjust = 0.5)

    g <- ggplot(mtcars, aes(factor(am),mpg,fill=factor(am))) +
        geom_boxplot(aes(colour=factor(am))) +
        geom_point(aes(colour=factor(am))) +
        theme(plot.title = element_text(hjust = 0.5)) +
        labs(title="Average MPG by Transmission Type") +
        labs(x="Transmisisson Type",y="MPG") +
        scale_colour_manual(values = c("red", "blue"))
    print(g)

##  Perform a t-test using the Automatic MPG and the Manual MPG data to determine the significance of
##  this delta.
```

```
    ttest_df <- t.test(automatic_df$mpg,manual_df$mpg)
    ttest_df

##  Create a linear model for the Automatic MPG and the Manual MPG data.
##  Provide a summary of the linear model.

    mpg_model <- lm(mpg~am,mtcars)
    summary(mpg_model)

##  Calculate and report the 95% confidence interval to verify the significance of the MPG delta.

    confint_mpg_model <- confint(mpg_model,level=.95)
    confint_mpg_model[2,]

##  Create a variable to store the mpg_model residual values.

    mpg_model_resid <- mpg_model$residuals

##  Plot the residuals for the linear model.

    plot(mpg_model_resid,col="blue",pch=19,main="Residual Plot",xlab="Car Index",ylab="Residuals")
    abline(0,0,col="red")

    par(mfrow=c(2,2))
    plot(mpg_model)

##  Convert the mpg_model residuals to a data frame to view various data points.

    mpg_model_df <- data.frame(mpg_model$residuals)

##  Review the first few rows of residual data.

    head(mpg_model_df)

##  Create a pairs plot to find the variables which would have a dependency on the MPG.

    pairs(mpg~.,mtcars,upper.panel=NULL)

##  Create a multivariate model using transmission type, cylinders, and weight as independent
##  variables.

    mpg_model2 <- lm(mpg~am+cyl+disp+hp+wt,data=mtcars)
    summary(mpg_model2)

##  Use the anova function to compare the mpg_model and mpg_model2 models.

    anova(mpg_model,mpg_model2)

##  Create a variable to store the mpg_model2 residual values.

    mpg_model2_resid <- mpg_model2$residuals

##  Plot the residuals for the multivariate model mpg_model2.
```

```
    plot(mpg_model2_resid,col="blue",pch=19,main="Residual Plot",xlab="Car Index",ylab="Residuals")
    abline(0,0,col="red")

    par(mfrow=c(2,2))
    plot(mpg_model2)

##  Create a multivariate model using stepAIC function and report the analysis of variance results.

    mpg_model3 <- lm(mpg~.,data=mtcars)
    step <- stepAIC(mpg_model3)

##  Report the stepAIC analysis of variance results.

    step$anova

##  Create a multivariate model using weight (wt), 1/4 mile time (qsec), and transmission
##  type (am) variables.
##  Use the anova function to compare the mpg_model, mpg_model2, and mpg_model3 models.

    mpg_model3 <- lm(mpg~wt+qsec+am,data=mtcars)
    summary(mpg_model3)
    anova(mpg_model,mpg_model2,mpg_model3)

##  Create a variable to store the mpg_model3 residual values.

    mpg_model3_resid <- mpg_model3$residuals

##  Plot the residuals for the multivariate model mpg_model3.

    plot(mpg_model3_resid,col="blue",pch=19,main="Residual Plot",xlab="Car Index",ylab="Residuals")
    abline(0,0,col="red")

    par(mfrow=c(2,2))
    plot(mpg_model3)

##  Store and report all the R^2 and p-value data for all three regression analyses into a summary data

    r_squared <- cbind(summary(mpg_model)$adj.r.squared,
        summary(mpg_model2)$adj.r.squared,summary(mpg_model3)$adj.r.squared)
    p_value <- cbind(summary(mpg_model)$coef[2,4],
        summary(mpg_model2)$coef[2,4],summary(mpg_model3)$coef[4,4])
    summary_df <- data.frame(rbind(r_squared,p_value))
    colnames(summary_df) <- c("Analysis #1", "Analysis #2", "Analysis #3")
    rownames(summary_df) <- c("R Squared", "p Value")
    summary_df

##  Calculate the delta in MPG based upon the coefficients listed in the mpg_model3 summary.

     mpg_delta <- round(summary(mpg_model3)$coefficients[4,1],4)
```