Regression Models in R

Ken Wood

9/9/2020

Executive Summary

Motor Trend, a magazine about the automobile industry, wants to look at a data set of a collection of cars to learn more about mileage. They are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). Specifically, they are interested in answering the following two questions:

- "Is an automatic or manual transmission better for MPG?"
- "How do we quantify the MPG difference between automatic and manual transmissions?"

```r
data(mtcars)
head(mtcars)
```

Load the 'mtcars' dataset and look at the first few rows...

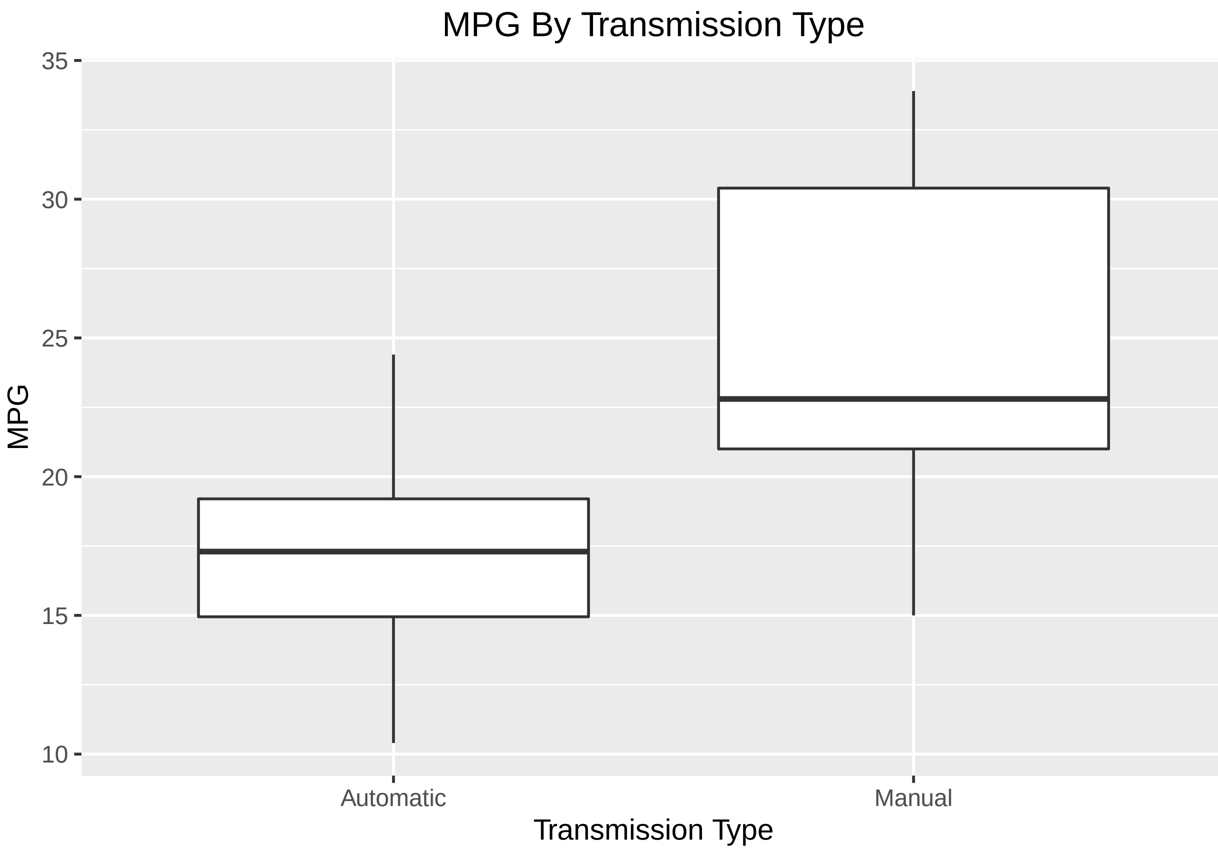
```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 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
```

Variables in the 'mtcars' dataset...

- **mpg:** Miles/(US) gallon
- **cyl:** Number of cylinders
- **disp:** Displacement (cu.in.)
- **hp:** Gross horsepower
- **drat:** Rear axle ratio
- **wt:** Weight (1000 lbs)
- **qsec:** 1/4 mile time
- **vs:** Engine (0 = V-shaped, 1 = straight)
- **am:** Transmission (0 = automatic, 1 = manual)
- **gear:** Number of forward gears
- **carb:** Number of carburetors

Question 1: "Is an automatic or manual transmission better for MPG?" First, let's generate a box plot of MPG vs. transmission type...

```r
library(ggplot2)
ggplot(mtcars, aes(x=factor(am), group=am, y=mpg)) + geom_boxplot() + scale_x_discrete(labels=c("0" = "A
```



We can perform a t-test on the mean MPG numbers for cars with automatic vs. cars with manual transmissions. Our hypotheses will be as follows:

- H_0: μ_m - $\mu_a = 0$
- H_a: μ_m - $\mu_a \neq 0$

where μ_m and μ_a are the mean MPGs for manual and automatic transmissions, respectively.

We need to separate the MPG numbers according to automatic vs. manual transmission.

```
manual <- mtcars[mtcars$am == 1,]      # get rows with manual transmission
automatic <- mtcars[mtcars$am == 0,]  # get rows with automatic transmission
manual <- manual[,"mpg"]
automatic <- automatic[,"mpg"]
row.names(manual) <- NULL             # remove row index
row.names(automatic) <- NULL
head(manual)
```

```
## [1] 21.0 21.0 22.8 32.4 30.4 33.9
```

```
head(automatic)
```

```
## [1] 21.4 18.7 18.1 14.3 24.4 22.8
```

We can now perform a t-test between 'manual' and 'automatic'.

```
result <- t.test(manual,automatic)
result
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  manual and automatic
## 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
```

Our test result shows that we should reject H_0 with a p-value very close to 0. The difference in the MPG means between manual and automatic transmissions is statistically significant at a 95% confidence level. Moreover, the mean MPG for manual transmissions (x) is significantly higher than the mean MPG for automatic transmissions (y).

Question 2: "How do we quantify the MPG difference between automatic and manual transmissions?" We start by performing a linear regression with 'mpg' as the dependent variable and the rest of the columns in 'mtcars' as the independent variables.

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

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

We see that the p-values for all of the variable coefficients are > 0.05, therefore, we cannot draw any conclusions about the statistical significance of the coefficients. To find out which independent variables are statistically significant, we will make use of R's 'step' function.

```
# run 'step' analysis with direction = "backward"
step_analysis <- step(fit,trace=FALSE)  # suppress output
```

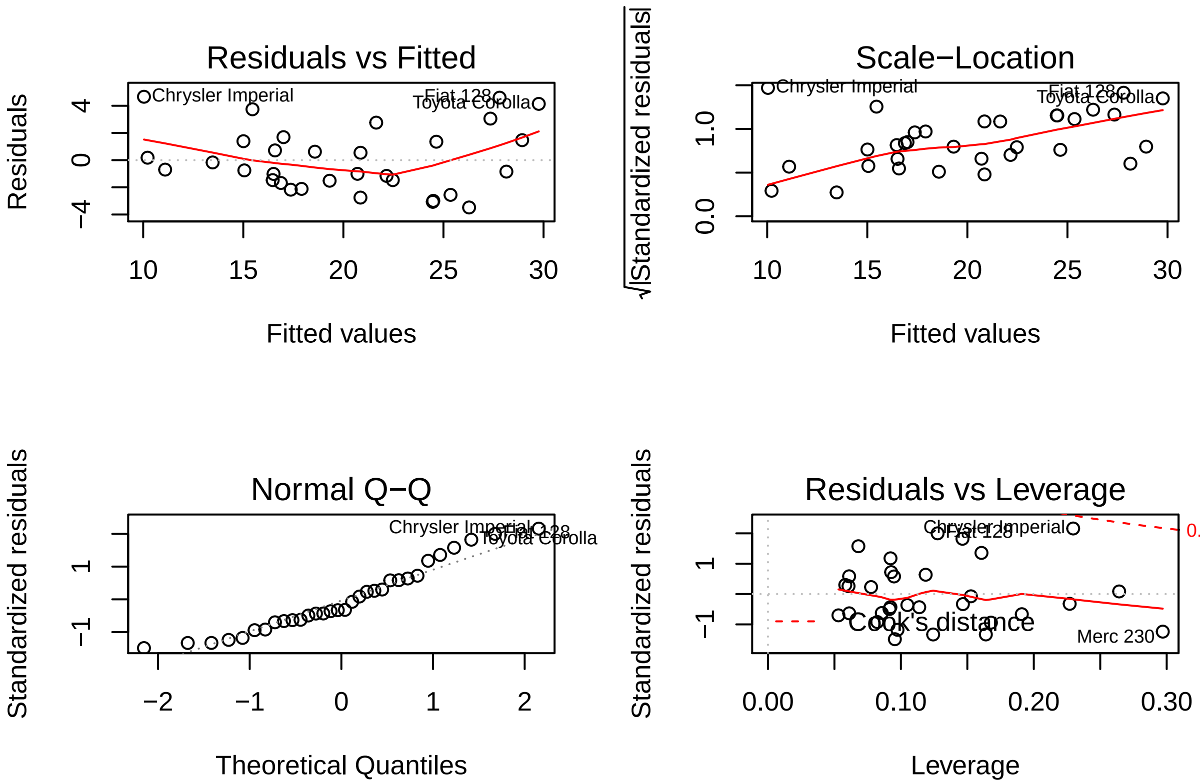
```r
summary(step_analysis)
```

```
##
## 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 ***
## am            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
```

Our step analysis results indicate that the coefficients for 'wt', 'qsec', and 'am' are statistically significant (p=values < 0.05) and our linear model comprised of these variables has a $R^2 = 0.85$, which means the model can account for about 85% of the variance in 'mpg'.

Let's generate residual plots for the simplified linear model obtained by the 'step' analysis:

```r
layout(matrix(c(1,2,3,4),2,2))      # display the plot area
plot(step_analysis)
```



We can draw the following conclusions from the residual plots:

1. The Residuals vs. Fitted plot shows no pattern consistency, thus we can conclude that the variables in our model are indeed independent.
2. The Normal Q-Q plot shows points lying very close to the line which indicates that the residuals are normally distributed.
3. The Scale-Location plot confirms our assumption of constant variance within the model, as the points are randomly distributed.
4. The values in the Residuals vs. Leverage plot all fall well within the 0.5 bands, which reveals that no outliers are present.

Conclusions

Therefore, we can safely adopt this model to provide an answer to Question #2.

- Using the final model output by the 'step' function, we see that the multiple R^2 value is sufficiently high at 0.85.
- We also see that 'wt' and 'qsec' are confounding variables in the relationship between 'mpg' and 'am'.
- The model predicts that cars with manual transmission will provide, on average, an additional 2.94 MPG compared to cars with automatic transmission.