Building Regression Models

Vitalii Diakonov

2024-04-01

Introduction:

As a data analyst at Motor Trend, I'm tasked with analyzing a dataset of car specifications to uncover insights into the factors affecting miles per gallon (MPG), a crucial metric in the automotive industry. My primary objective is to determine whether there's a noticeable difference in MPG between cars equipped with automatic and manual transmissions and quantify this difference.

Data:

I'll be working with the mtcars dataset, which provides comprehensive information about various car models. My focus will primarily be on two key variables: MPG and transmission type (automatic or manual).

```
# Load necessary libraries
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

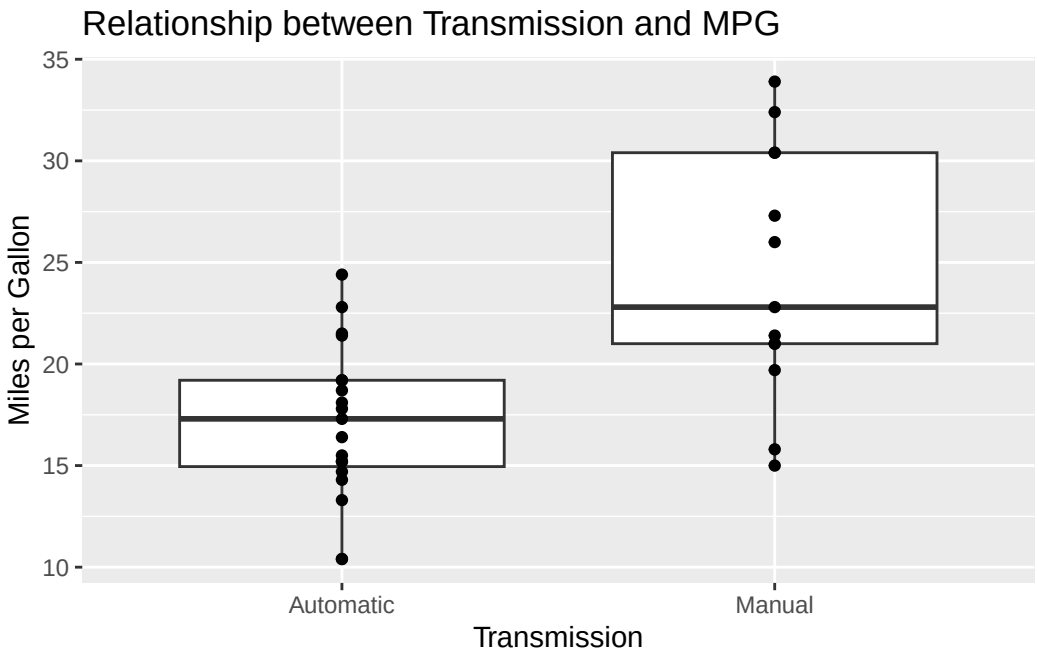
```r
# Load the mtcars dataset
data("mtcars")

# Select relevant variables for analysis (MPG and transmission)
mpg_transmission <- mtcars %>% select(mpg, am) %>% mutate(am = as.factor(am))
```

Exploratory Data Analysis (EDA):

To kick off the analysis, I'll narrow down the dataset to just MPG and transmission variables for an initial exploratory dive. Through visualizations like box plots, I'll examine the distribution of MPG across different transmission types. My goal is to discern any notable differences, especially whether manual transmission cars tend to exhibit higher average MPG.

```r
mpg_transmission <- mtcars %>%
  select(mpg, am) %>%
  mutate(am = factor(am, levels = c(0, 1), labels = c("Automatic", "Manual")))
ggplot(data = mpg_transmission, aes(x = am, y = mpg)) +
  geom_boxplot() +
  geom_point() +
  xlab('Transmission') +
  ylab('Miles per Gallon') +
  ggtitle('Relationship between Transmission and MPG')
```



Based on the box plot analysis, it appears that cars with manual transmission tend to have a higher mean miles per gallon (MPG) compared to those with automatic transmission.

Model Fitting and Selection:

Moving forward, I'll embark on the process of fitting linear regression models to the MPG data, initially incorporating all available variables. However, to ensure model robustness, I'll meticulously assess diagnostic metrics, gradually eliminating non-significant variables until I achieve a final model with significant coefficients.

```
# Fit multiple linear regression models
fitAll <- lm(mpg ~ ., data = mtcars)
summary(fitAll)
```

```
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
```

```r
# Refine the model by removing non-significant variables
fitRaw <- mtcars %>%
  select(-cyl) %>%  # Remove 'cyl' variable
  select(-disp) %>% # Remove 'disp' variable
  select(-hp) %>%   # Remove 'hp' variable
  select(-drat) %>% # Remove 'drat' variable
  select(-vs) %>%   # Remove 'vs' variable
  select(-gear) %>% # Remove 'gear' variable
  select(-carb)     # Remove 'carb' variable

fitRm <- lm(mpg ~ ., data = fitRaw)
summary(fitRm)
```

```
Call:
lm(formula = mpg ~ ., data = fitRaw)

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
```

1. **Intercept (Intercept):**
    - The intercept of approximately 9.62 suggests that for a car with certain characteristics (such as weight, 1/4 mile time, etc., represented by other variables), the expected MPG is around 9.62 when the transmission type is automatic and all other variables are held constant.
2. **Weight (wt):**
    - The coefficient of -3.92 indicates that for every unit increase in weight (in 1000 lbs), the MPG decreases by approximately 3.92, holding other variables constant. This suggests that heavier cars tend to have lower MPG.
3. **1/4 Mile Time (qsec):**
    - The coefficient of 1.23 suggests that for every unit increase in the 1/4 mile time, the MPG increases by approximately 1.23, when other variables are held constant. This might seem counterintuitive but could be due to various factors such as engine efficiency at different speeds.
4. **Transmission Type (am):**
    - The coefficient of 2.94 indicates that cars with manual transmission tend to have approximately 2.94 higher MPG compared to cars with automatic transmission, holding other variables constant. This suggests that manual transmission might be better for MPG.

Statistical Significance:

- The coefficients for weight (wt), 1/4 mile time (qsec), and transmission type (am) are statistically significant, as indicated by their p-values being less than 0.05.

- The overall model is statistically significant, with an extremely low p-value of 1.21e-11, suggesting that the model explains a significant amount of variance in the MPG.

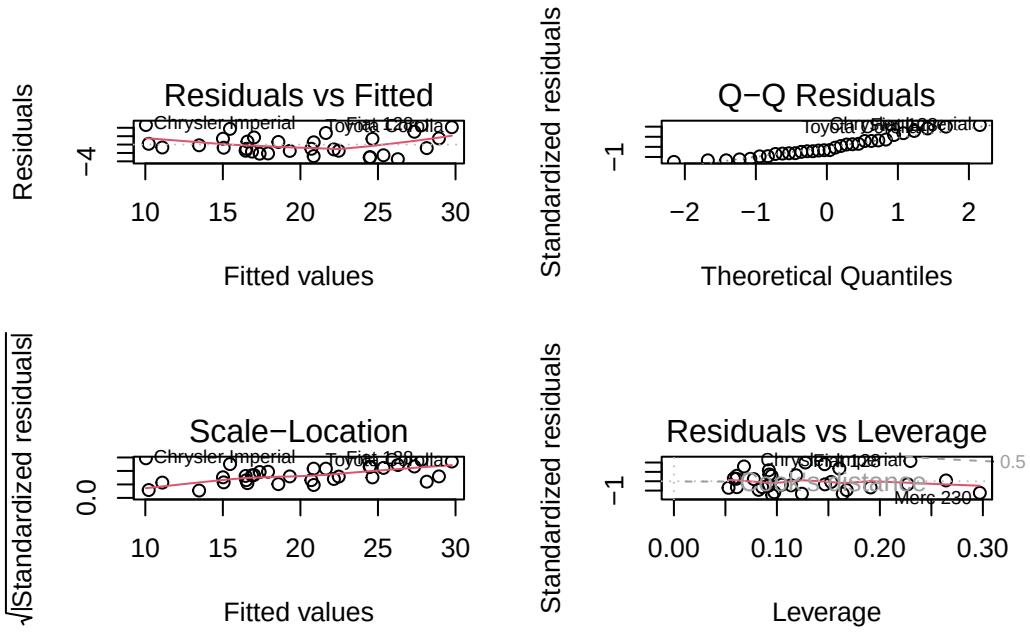
Model Fit:

- The multiple R-squared value of 0.85 indicates that approximately 85% of the variance in the MPG can be explained by the model.
- The adjusted R-squared value of 0.83 adjusts for the number of predictors in the model, providing a more accurate measure of model fit.

Diagnostics:

To validate the integrity of the model, I'll subject it to a battery of diagnostic tests. Utilizing diagnostic plots like QQ plots and residuals vs. fitted plots, I'll evaluate adherence to model assumptions and detect any potential issues that may compromise model accuracy.

```
# Conduct residual analysis and diagnostic tests
par(mfrow = c(2, 2))
plot(fitRm)
```



The QQ plot shows a pretty good correlation of the standardized and theoretical residuals. There also doesn't seem to be any significant patterns in the other three plots, indicating a good fit of the selected model.

Conclusions:

Based on the analysis conducted using linear regression, we can address the questions posed by Motor Trend:

1. **"Is an automatic or manual transmission better for MPG?"**

    - The coefficient for transmission type (am) in the model is approximately 2.94 with a p-value of 0.047. This suggests that cars with manual transmission tend to have approximately 2.94 higher MPG compared to cars with automatic transmission, after controlling for other variables such as weight and 1/4 mile time. Therefore, based on this analysis, manual transmission appears to be better for MPG.

2. **"Quantify the MPG difference between automatic and manual transmissions"**

    - The coefficient for transmission type (am) quantifies the difference in MPG between automatic and manual transmissions. In this case, the coefficient of 2.94 indicates that, on average, cars with manual transmission achieve 2.94 more miles per gallon compared to cars with automatic transmission, all else being equal.

These findings provide insights into the relationship between transmission type and MPG, suggesting that manual transmission may offer better fuel efficiency than automatic transmission. However, it's essential to consider other factors and conduct further analysis to validate these conclusions.