Regression Analysis of the MPG Outcome Variable

Wei Wei

May 15, 2016

Executive Summary

Using mtcars, a dataset on a collection of cars, the relationship between miles per gallon (MPG) outcome variable and a set of predictor variables was explored. This analysis intends to address 2 goals/questions:

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

Using simple linear regression to model regression of MPG on transmission (am), it is found that using manual transmission is better for MPG: it yields an extra of 7.245 MPG, with an associated p-value of 0.000285. Therefore the H0 hypothesis that there is no difference between the means of MPG for cars with manual or with automatic transmissions is rejected.

Multiple linear regression analysis was used to find the covariates of the transmission variable (am) and to overcome the "omitted-variable bias". It is found that the MPG outcome variable is best predicted by the transmission variable (am) and 3 covariates: number of cylinders (cyl), the gross horsepower (hp), and the weight (wt). Quantitatively, a manual transmission will yield an extra of 1.809 MPG on a car with a 4-cylinder engine at a fixed horsepower and a fixed weight. However, with a p-value of 0.206, it is not possible to reject the H0 hypothesis that there is no difference between the means of MPG for cars with manual or with automatic transmissions. In other words, we cannot conclude with confidence that using manual transmission will yield a better MPG using the selected multiple linear regression model.

Exploratory Data Analysis and Simple Linear Regression Analysis

```
data(mtcars)
?mtcars
## simple linear regression model
fit1 <- lm(mpg ~ factor(am), data=mtcars)
summary(fit1)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

The mtcars dataset contains 32 observations of 11 variables. There are 13 observations with manual transmission (1 = manual) and 19 with automatic transmission (0 = automatic).

Exploratory data analysis shows that the median of MPG for cars with manual transmission is well separated from the median of MPG for cars with automatic transmission (see Appendix Figure 1). Simple linear regression analysis using transmission (am) as predictor and MPG as outcome variable shows that there is an extra of 7.245 MPG using manual transmission, with an associated p-value of 0.000285. Therefore the H0 hypothesis that there is no difference between the means of MPG for cars with manual or with automatic transmissions is rejected.

Multiple Linear Regression Model Selection

The MPG outcome variable may be dependent on a number of variables other than transmission (am). In fact, only 35.98% of total variance in the MPG outcome variable can be explained by the simple linear regression model (fit1, R-squared: 0.3598). Therefore, the other 9 variables were evaluated using the nested likelihood ratio tests method to see if they should be included as covariates of the transmission variable (am) in a multiple linear regression model.

```
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)

fit1 <- lm(mpg ~ am, data=mtcars) ## step-wise nested likelihood ratio test
fit2 <- update(fit1, mpg ~ am + cyl)
fit3 <- update(fit1, mpg ~ am + cyl + hp)
fit4 <- update(fit1, mpg ~ am + cyl + hp + wt)
fit5 <- update(fit1, mpg ~ am + cyl + hp + wt + disp + drat + qsec + vs + gear + carb)
anova(fit1, fit2, fit3, fit4, fit5)$Pr
```

```
## [1]           NA 7.889714e-06 1.109533e-02 2.991775e-02 9.588242e-01
```

```
## variance inflation factor analysis
library(car)
round(vif(fit4), 3)
```

```
##      GVIF Df GVIF^(1/(2*Df))
## am  2.591  1           1.610
## cyl 5.825  2           1.554
## hp  4.704  1           2.169
## wt  4.007  1           2.002
```

As the result of nested-likelihood ratio tests, cyl, hp, and wt are included as covariates, while disp, drat, qsec, vs, gear, and carb are excluded. Variance inflation factor (vif) analysis shows that am, cyl, hp, and wt are orthogonal aggressors, using vif > 10 as the cutoff value.

```
## coefficients in the multiple linear regression model
round(summary(fit4)$coef, 3)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.708      2.605  12.940    0.000
## am1            1.809      1.396   1.296    0.206
## cyl6          -3.031      1.407  -2.154    0.041
## cyl8          -2.164      2.284  -0.947    0.352
## hp            -0.032      0.014  -2.345    0.027
## wt            -2.497      0.886  -2.819    0.009
```

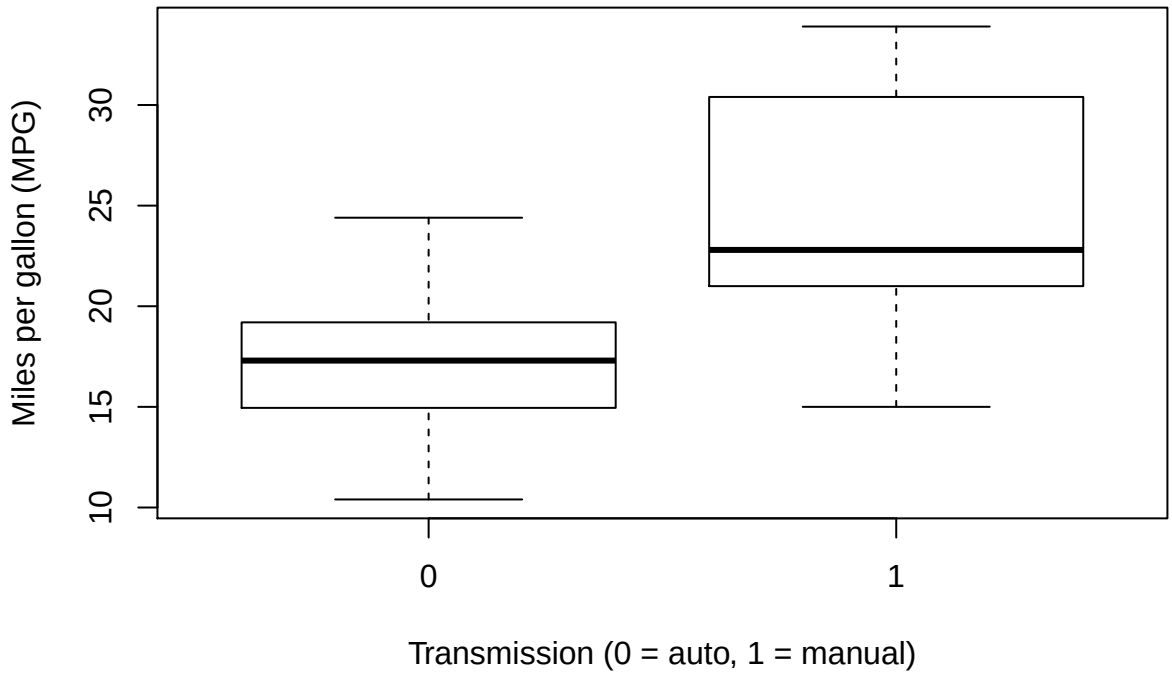
A manual transmission will yield an extra of 1.809 miles per gallon on a car with a 4-cylinder engine at a fixed horsepower and a fixed weight. However, with a p-value of 0.206, the H0 hypothesis that there is no difference between the means of MPG for cars with manual or with automatic transmissions can not be rejected.

Residuals in the multiple linear regression model were examined in the diagnostic plots (see Appendix Figure 2). No discernible patterns were detected in the "Residuals vs Fitted values" plot, suggesting that the residuals are due to noise. No high leverage and high influence data points were observed in the "Residuals vs Leverage" plot.

Appendix

```
## Appendix figure 1 exploratory analysis
par(mfrow=c(1, 1))
boxplot(mpg ~ factor(am), mtcars, xlab='Transmission (0 = auto, 1 = manual)',
        ylab='Miles per gallon (MPG)')
mtext(side=3, line=-1, text="Figure 1. Exploratory analysis of MPG ~ Transmission",
      font=2, adj=0, outer=TRUE)
```

Figure 1. Exploratory analysis of MPG ~ Transmission




```
## Appendix figure 2 diagnostic plots of regression of MPG on am + cyl + hp + wt
par(mfrow=c(2, 2))
plot(fit4)
mtext(side=3, line=-1, font=2, adj=0, outer=TRUE,
      text="Figure 2. Diagnostic plot of regression of MPG ~ am + cyl + hp + wt")
```

Figure 2. Diagnostic plot of regression of MPG ~ am + cyl + hp + wt