07 Regression - Course Project

Joe Nguyen

20 November 2015

Executive Summary

This report explores whether an "automatic" or "manual" transmission is better for miles per gallon (MPG) and quantifies the MPG difference between the transmission types. Cars are more fuel efficient if their MPG rating is lower. A linear regression analysis shows that "automatic" achieves lower MPG with an average of 17.15 and 95% confidence interval (14.85, 19.44) compared with 24.39 and (21.62, 27.17) for "manual".

Explore `mtcars` dataset

```
##           mpg cyl disp  hp drat   wt  qsec vs am gear carb
## Mazda RX4  21   6  160 110  3.9 2.62 16.46  0  1    4    4
```

Variable `am` is transmission type (0 = automatic, 1 = manual) (from R documentation `?mtcars`). A distribution plot of `am` for MPG is shown in Fig. 1.

Model Selection and Fitting

We treat `am` as a factor variable with two levels and perform regression with the linear model: $Y_i = \beta_0 + X_{i1}\beta_1 + \epsilon_i$. Using least squares, $\hat{\beta}_0$ is the mean estimate for "automatic" and $\hat{\beta}_0 + \hat{\beta}_1$ is the mean estimate for "manual". The coefficient β_1 is the change in the mean of "manual" from "automatic".

```
##              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
```

The estimated means from the linear model for "automatic" ($\hat{\beta}_0 = 17.147$) and "manual" ($\hat{\beta}_0 + \hat{\beta}_1 = 17.147 + 7.245 = 24.392$) correspond to the means (points) in Fig. 1.

Model Selection

First, we propose that **no other regressors** are required in our linear model to answer the question *Is an automatic or manual transmission better for MPG?* The regressor `am` directly captures the affect of transmission type on MPG. However, we investigate whether additional regressors have a strong influence on explaining MPG. The goal of model building is to balance bias and variance inflation: 1) if we include unrelated regressors, the standard errors of all regressors increase, and 2) if we include addition regressors that are correlated with current regressors, the variance increases. We can evaluate the variance inflation factor (VIF) of each regressor. As a general rule,

VIF	Status of Regressors
VIF = 1	Uncorrelated (and zero inflation)
1 < VIF < 5	Moderately correlated
VIF > 5 to 10	Highly correlated

```
##       cyl      disp        hp      drat        wt      qsec        vs 
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873 
##        am      gear      carb 
##  4.648487  5.357452  7.908747
```

We would choose `disp` as an additional regressor because it is highly correlated with the other regressors, suggesting `disp` can well-explain MPG in this dataset. This correlation between `disp` and `am` is evident in Fig. 2 where "manual" cars tend to have low displacement (`disp`) and "automatic" cars have high displacement.

We can further use analysis of variance (ANOVA) and examine the F-test p-value for significance by including `disp` with `am`. From the appendix, we can see the F-test p-value for the model `mpg ~ am + disp` of 1.397e-07 strongly suggests `disp` accounts for a lot of variation in MPG, while adding all the remaining regressors has a smaller p-value (0.03157). We can also see the (explained) sum of squares of `disp` (420.62) is 64% larger than the sum of squares of adding the remaining regressors (152.79). Thus, `disp` accounts for a large porportion of the variability. Note:

Total sum of squares = Explained sum of squares + Residual sum of squares (RSS)

Diagnostics

We investigate the quality of the original model (`mpg ~ am`) fit by examining residual variation. As seen in Fig. 3, there is no systematic patterns or outlying data points. We also examine the data points in the dataset to observe whether or not individual data points have high influence (e.g. leverage) on the model. Diagnostics include `dffits`, `dfbetas`, `cooks.distance` and `hatvalues`. The results suggest all data points well-represent the model, and results are shown in the appendix.

Regression Inference

Performing a t-test, a p-value of 0.000285 (< 0.05) suggests there is significance in the difference of MPG between "automatic" and "manual". According to this model and data, "automatic" is potentially better than "manual" with a lower MPG. The very small p-value (0.000285) indicates there is little likelihood the difference is due to chance.

Confidence and Prediction Intervals

The 95% confidence intervals for the MPG mean of "automatic" and "manual" are shown below.

	Mean	Lower	Upper
Automatic	17.15	14.85	19.44
Manual	24.39	21.62	27.17

The 95% prediction intervals are similarly shown below; they are wider than the confidence intervals.

	Mean	Lower	Upper
Automatic	17.15	6.88	27.42
Manual	24.39	14.00	34.78

Appendix

Explore mtcars dataset

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
ggplot(aes(am, mpg), data = mtcars) +
    geom_violin() +
    geom_point(aes(x = levels(am), y = mpgMn$x), size = 3) +
    ggtitle("MPG Distribution for Transmission Type") +
    xlab("Transmission Type") + ylab("MPG")
```

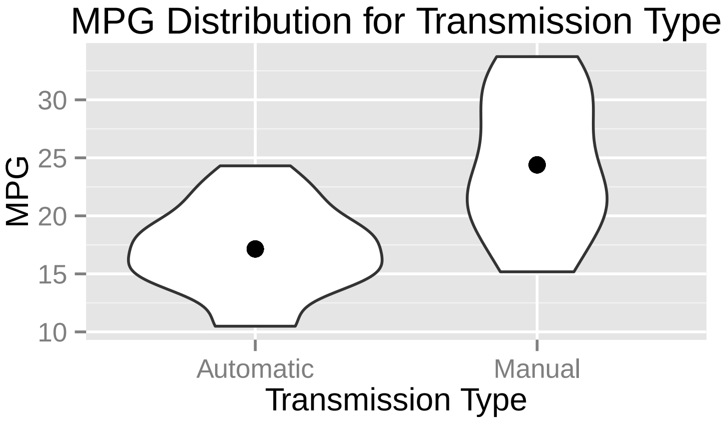

Figure 1 MPG distribution for transmission type ("automatic" or "manual"). The dots represent the mean MPG for each transmission type.

Model Selection and Fitting

Linear Regression Code

```
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
mpgMn <- with(mtcars, aggregate(mpg, by = list(am), mean))
modLm <- lm(mpg ~ am, data = mtcars)
summary(modLm)$coef
```

```
ggplot(aes(disp, mpg, colour = am), data = mtcars) +
    geom_point(size = 3, alpha = 0.7)
```

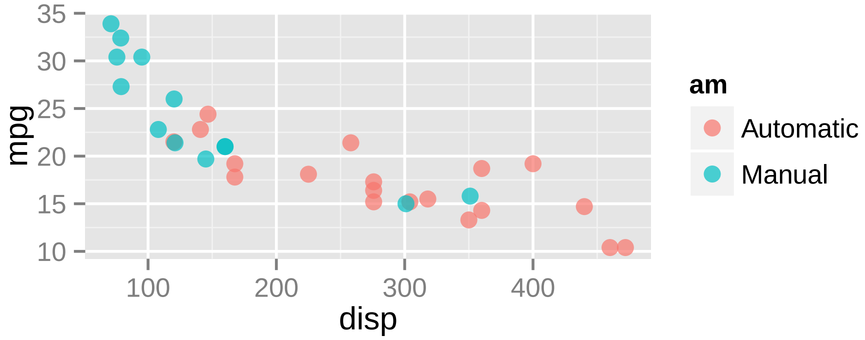

Figure 2 Engine displacement (`disp`) is correlated with transmission type (`am`) and can well-explain the outcome MPG. "Manual" cars tend to have lower displacement and higher MPG, while "automatic" cars tend to have higher displacement and lower MPG.

VIF Code

```r
library(car)
modFull <- lm(mpg ~ ., data = mtcars)
vif(modFull)
```

ANOVA Code

```r
mod_1 <- update(modLm, mpg ~ am + disp)
features <- names(subset(mtcars, select = -c(mpg, am, disp)))
form <- as.formula(paste("mpg ~ am + disp + ", paste(features, collapse = "+")))
mod_2 <- update(modLm, form)
anova(modLm, mod_1, mod_2)
```

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ am
## Model 2: mpg ~ am + disp
## Model 3: mpg ~ am + disp + cyl + hp + drat + wt + qsec + vs + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     30 720.90                                   
## 2     29 300.28  1    420.62 59.8865 1.397e-07 ***
## 3     21 147.49  8    152.79  2.7192   0.03157 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostics

```r
par(mfrow = c(2,2))
modLmDf <- data.frame(Fitted = predict(modLm),
                      Residuals = resid(modLm))
ggplot(data = modLmDf, aes(x = Fitted, y = Residuals)) +
    geom_point(size = 3, alpha = 0.7)
```

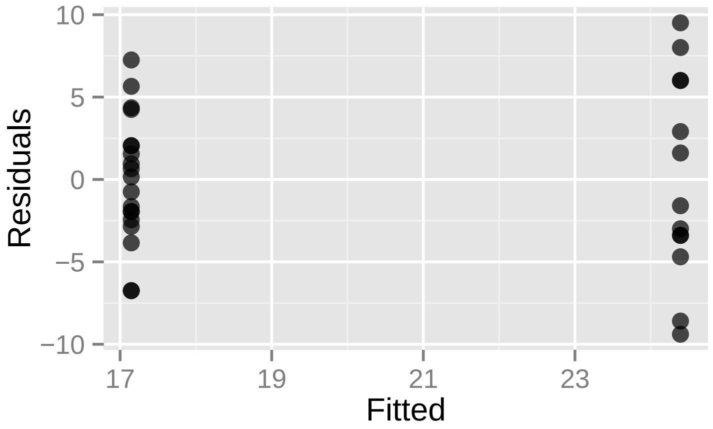

Figure 3 Residuals plot for predicted values according to the linear model `lm(mpg ~ am)`.

```
# Diagnostics when ith point is deleted in fitting the model:
# Change in outcome
dffits(modLm)
# Change in individual coefficients
dfbetas(modLm)
# Overall change in coefficients
cooks.distance(modLm)
# Leverage
hatvalues(modLm)
```

Regression Inference

```
modLmMn <- lm(mpg ~ am - 1, data = mtcars)
modCoef <- summary(modLmMn)$coef
# # vectors stored as columns by default
# modCoef[,1] + qt(0.975, df = modLmMn$df) * modCoef[,2] %*% t(c(-1,1))
predict(modLm, newdata = data.frame(am = as.factor(c("Automatic", "Manual"))),
        interval = "confidence")
```

```
##        fit      lwr      upr
## 1 17.14737 14.85062 19.44411
## 2 24.39231 21.61568 27.16894
```

```
predict(modLm, newdata = data.frame(am = as.factor(c("Automatic", "Manual"))),
        interval = "prediction")
```

```
##        fit       lwr      upr
## 1 17.14737  6.876013 27.41872
## 2 24.39231 14.003113 34.78150
```