Regression Models: Week 4 PGA

Artem Shevlyakov

16/02/2019

Summary

Although at first it seems that transmission type influences the mileage per gallon, it is in fact wrong. Adjustment for weight and cylinder count shows that these variables are the important ones. It's impossible to answer which type of transmission is better for MPG and quantify the difference, as it is statistically insignificant.

Exploratory data analysis and transformation

Let's load the `mtcars` dataset and look at it briefly. What are the columns made up of? Are there any missing values?

```
data(mtcars)
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
```

```
sapply(mtcars, class)
```

```
##       mpg       cyl      disp        hp      drat        wt      qsec
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##        vs        am      gear      carb
## "numeric" "numeric" "numeric" "numeric"
```

```
any(is.na(mtcars))
```

```
## [1] FALSE
```

It seems that the dataset is complete, no values are missing. However, some of the variables indicated as numeric should really be treated as factors.

Model selection

We are going to use `am` (remember it's a factor now!) as the independent variable and `mpg` as the regression. Since `mpg` is continuous, and there's no rate involved, we are going to perform linear regression.

```
model1 <- lm(mpg~as.factor(am), mtcars)
summary(model1)$coef
```

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

It seems that automatic transmission yields about 17.15 miles per gallon of gas, and manual transmission (encoded as `am==1`) is associated with a 7.245 miles higher mileage per a gallon of gas for a total of about 24.4 miles per gallon. However, there are other factors which may affect mileage that this model does not account for, namely, mass and cylinder count. Let's adjust for these variables.

```
model2.wt.add <- lm(mpg~as.factor(am)+wt, mtcars)
model3.wtcyl.add <- lm(mpg~as.factor(am)+wt+as.factor(cyl), mtcars)
anova(model1, model2.wt.add, model3.wtcyl.add)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ as.factor(am)
## Model 2: mpg ~ as.factor(am) + wt
## Model 3: mpg ~ as.factor(am) + wt + as.factor(cyl)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 65.3095 1.107e-08 ***
## 3     27 182.97  2     95.35  7.0353  0.003473 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the F-statistic it seems that the model adjusted for weight is better that the first one, and the model adjusted for both weight and number of cylinders is even better than the second one (probability of error $p<0.01$ in both cases). Let's see if the type of transmission remains a significant factor influencing mileage in the adjusted model.

```
summary(model3.wtcyl.add)$coef
```

```
##                    Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)     33.7535920  2.8134831 11.9970836 2.495549e-12
## as.factor(am)1   0.1501031  1.3002231  0.1154441 9.089474e-01
## wt              -3.1495978  0.9080495 -3.4685309 1.770987e-03
## as.factor(cyl)6 -4.2573185  1.4112394 -3.0167231 5.514697e-03
## as.factor(cyl)8 -6.0791189  1.6837131 -3.6105432 1.227964e-03
```

The model indicates that with the number of cylinders and type of transmission held constant, each 1000 lbs increase in weight results in 3.15 decrease in mileage per gallon. With the weight and type of transmission held constant, 6-cylinder engines decrease mileage by 4.26 miles, and 8-cylinder by 6.08 miles, compared to the 4-cylinder engines.

While the coefficients for the number of cylinders and weight are statistically significant, the coefficient for the transmission is not. This means that we cannot reject the null hypothesis that there is no difference between manual and automatic transmission in terms of mileage per gallon when the weight and the number of cylinders are held constant. Do we even need this variable in our model?

```
model.noam <- lm(mpg~wt+as.factor(cyl), mtcars)
anova(model.noam, model3.wtcyl.add)$"Pr(>F)"[2]
```

```
## [1] 0.9089474
```

According to the statictic, there's no statistically significant difference between the models with and without `am` as predictor. Therefore, we can conclude that it does not affect mileage and should not be included in the final model. It's the weight and the number of cylinders that affect the mileage, not the type of transmission.

The same conclusion can be made after looking at the residual plots of the two models presented in the appendix. Both plots are very similar, which confirms that there is no significant difference between the two models. The one with less variables should be used. There is no clear pattern to the residuals, so we cannot assume that there are other factors significantly influencing mileage per gallon apart from weight and the number of cylinders.

Appendix: Plots

Plot 1. Comparison of the residual plots of models including and excluding transmission type