Regression Models Course Project

Morgane Flauder

January 27, 2016

Summary

In this project, I will work on the dataset `mtcars` which comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. I will analyze the influence of the type of transmission (automatic or manual) on the MPG (miles per gallon) for these cars.

I found that manual transmission leads to 6.8 more miles per gallon, so the manual transmission is better for MPG.

Exploratory Analysis

First, let's visualize the repartition of MPG for the two types of transmissions.

As we can see in figure 1 (see appendix), it seems that the MPG is higher with the manual transmission compared to the automatic. The difference in means of MPG is :

```
mean(mtcars$mpg[mtcars$am == 1]) - mean(mtcars$mpg[mtcars$am == 0])
```

```
## [1] 7.244939
```

We are now going to explore this hypothesis by using linear regression models.

Analysis of the influence of the type of transmission on MPG

Initial model

First, let's create a simple model where MPG is only predicted by the type of transmission.

```
fit_am <- lm(mpg ~ am, data = mtcars)
summary(fit_am)$coefficient
```

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

We can see that according to this model, cars with manual transmissions have 7.24 more miles per gallon, which is exactly the difference in the means computed previously.

However, this model may be too simple. To verify this, let's analyze the plots of the residuals (see appendix, figure 2). The residuals should be scattered accros the horizontal line at 0 with no significant pattern. However, that is not the case. In fact, the predictive variable `am` we used is a factor variable, and shouldn't be used alone in a linear regression model.

Extended model

We are now going to construct a more complex model. To choose which variables to use, I computed the correlation between `mpg` and all the other variable. I selected the variables with a correlation over 0.5 or under -0.5.

```
abs(cor(mtcars)[1,]) > 0.5
```

```
##   mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
```

Since `qsec` and `gear` aren't strongly correlated with `mpg`, I choose to use the other 8 variables as predictors.

```
fit_8 <- lm(mpg ~ cyl + disp + hp + drat + wt + vs + am + carb, data = mtcars)
```

Since none of these features are significative at $\alpha = 5\%$ (see appendix, figure 3), I computed the Variance Inflation Factor (VIF) for this model.

```
library(car)
vif(fit_8)
```

```
##       cyl      disp        hp      drat        wt        vs        am
## 12.526040 19.633287  9.621000  3.349569 10.511066  4.302619  3.771637
##      carb
##  5.753692
```

We can see that `cyl`, `disp`, `hp` and `wt` have high VIF (relatively to the other variables), which means they might be collinear. I choose to construct a new model without these variables.

Refined model

This model use the variables `drat`, `vs`, `am` and `carb`.

```
fit_4 <- lm(mpg ~ drat + vs + am + carb, data = mtcars)
```

All the variables are significant at $\alpha = 5\%$ except `drat` (see appendix, figure 4).

Final model

I created a last model by removing `drat` from the previous one.

```
fit_3 <- lm(mpg ~ vs + am + carb, data = mtcars)
```

All the variables are significant at $\alpha = 5\%$ (see appendix, figure 5).

Choosing the best model

I performed an analysis of variance (ANOVA) in order to choose the best model (see appendix, figure 6).

We can see that Model 2 (`fit_3`) significantly improved the initial model at $\alpha = 5\%$. However, it is not the case for the other ones.

Conclusion

The variables that influence `mpg` the most, associated with `am`, are `vs` (the type of engine : V-engine or straight engine) and `carb` (the number of carburetors). This model shows that having a manual transmission result in an increase of MPG. In this model, MPG increase by 6.8 miles per gallon for a manual transmission compared to an automatic one.

Appendix

Figure 1

```
boxplot(mpg ~ am, data = mtcars,
        ylab = "Miles per gallon", xlab = "Transmission", xaxt = "n",
        main = "Difference in miles per gallon for\nautomatic and manual transmissions")
axis(1, at = c(1,2), labels = c("Automatic", "Manual"))
```

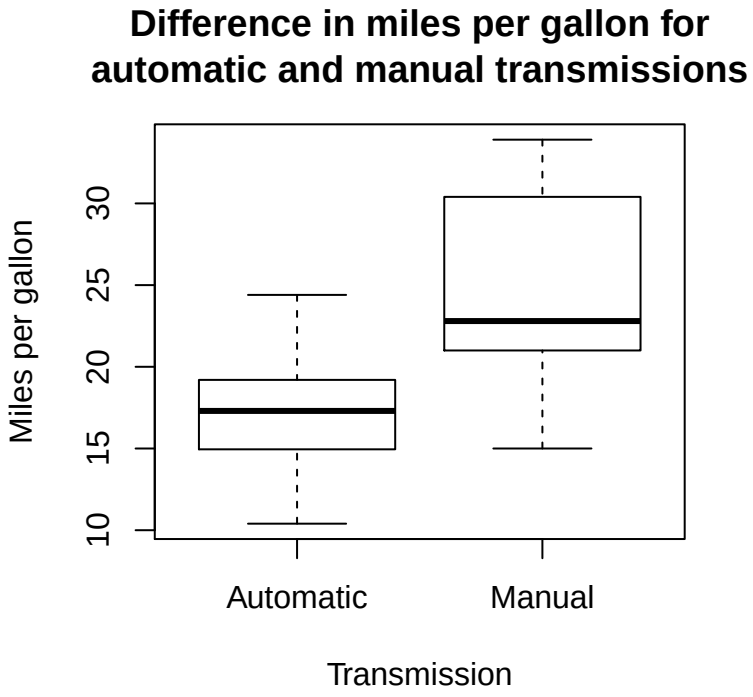
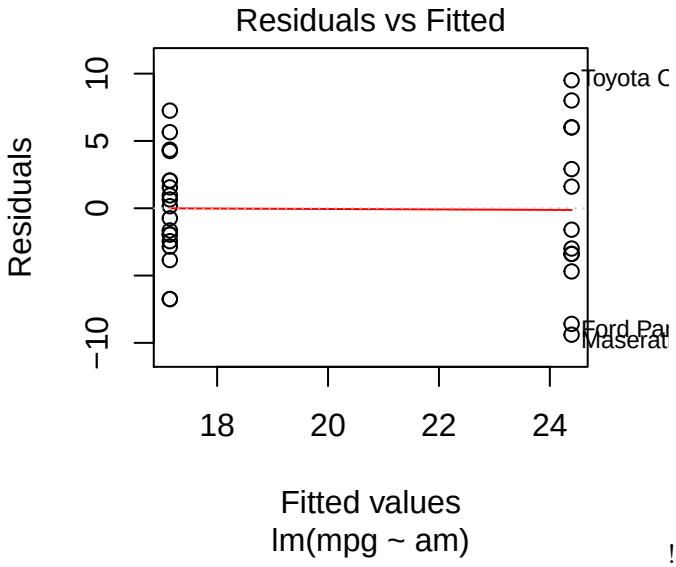


Figure 2

```
plot(fit_am)
```



!!!

Figure 3

```
summary(fit_8)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + hp + drat + wt + vs + am + carb,
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9540 -1.5150 -0.2235  1.6171  5.1623
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.769832   9.884770   3.113   0.0049 **
## cyl         -0.554718   0.930564  -0.596   0.5569
## disp         0.008244   0.016788   0.491   0.6280
## hp          -0.023384   0.021243  -1.101   0.2824
## drat         0.764939   1.607323   0.476   0.6386
## wt          -2.783772   1.555905  -1.789   0.0868 .
## vs           1.191481   1.932518   0.617   0.5436
## am           2.129301   1.827570   1.165   0.2559
## carb        -0.320044   0.697347  -0.459   0.6506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.614 on 23 degrees of freedom
## Multiple R-squared:  0.8604, Adjusted R-squared:  0.8118
## F-statistic: 17.72 on 8 and 23 DF,  p-value: 3.965e-08
```

Figure 4

```
summary(fit_4)
```

```
##
## Call:
## lm(formula = mpg ~ drat + vs + am + carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9807 -1.5218  0.3089  1.5176  4.7126
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.3201     5.0057   2.261  0.03199 *
## drat          2.6841     1.5579   1.723  0.09635 .
## vs            3.1029     1.4289   2.172  0.03884 *
## am            4.9490     1.5117   3.274  0.00291 **
## carb         -1.5115     0.3973  -3.805  0.00074 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.863 on 27 degrees of freedom
## Multiple R-squared:  0.8035, Adjusted R-squared:  0.7743
## F-statistic: 27.59 on 4 and 27 DF,  p-value: 3.431e-09
```

Figure 5

```
summary(fit_3)
```

```
##
## Call:
## lm(formula = mpg ~ vs + am + carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2803 -1.2308  0.4078  2.0519  4.8197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.5174     1.6091  12.130 1.16e-12 ***
## vs            4.1957     1.3246   3.168  0.00370 **
## am            6.7980     1.1015   6.172 1.15e-06 ***
## carb         -1.4308     0.4081  -3.506  0.00155 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.962 on 28 degrees of freedom
## Multiple R-squared:  0.7818, Adjusted R-squared:  0.7585
## F-statistic: 33.45 on 3 and 28 DF,  p-value: 2.138e-09
```

Figure 6

```
anova(fit_am, fit_3, fit_4, fit_8)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ vs + am + carb
## Model 3: mpg ~ drat + vs + am + carb
## Model 4: mpg ~ cyl + disp + hp + drat + wt + vs + am + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 245.65  2    475.25 34.7634 1.116e-07 ***
## 3     27 221.32  1     24.33  3.5596   0.07189 .
## 4     23 157.21  4     64.11  2.3446   0.08482 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```