Regression models course project

S. Marceau

10/21/2020

Summary

During this project, we will explore a car database in order to evaluate whether an automatic or a manual transmission is better as car gaz consumption is concerned. To do so, we will review the relationship between a set of variables and miles per gallon (MPG) by fitting multiple regression models. Then we will evaluate the relevant model fit by looking at residuals in order to optimize the model accuracy and by looking for potential outliers. Finally, we will quantify the MPG difference between automatic and manual transmissions.

Exploratory analysis

Testing for automatic and manual transmission mean difference

```
##
##  Welch Two Sample t-test
##
## data:  mtcars[mtcars$am == 1, "mpg"] and mtcars[mtcars$am == 0, "mpg"]
## t = 3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.209684 11.280194
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

From figure 1 and t-test above, we can definitely tell that mpg increase for manual cars lies between 3.2096842 and 11.2801944 with 95% confidence.

Exploring the relationship between variables mpg and am

Checking whether am is an effective predictor for mpg

Assessing multiple models based on the most influent variables, looking only at RSS values, we note that model 5 and 6 have the lowest residual variance.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + disp
## Model 4: mpg ~ am + wt
## Model 5: mpg ~ am + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 48.041 1.285e-07 ***
## 3     29 300.28  0    -28.92
## 4     29 278.32  0     21.96
## 5     29 245.44  0     32.88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

comparing nested models including am variable as predictor

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + hp
## Model 3: mpg ~ am + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 245.44  1    475.46 73.841 2.445e-09 ***
## 3     28 180.29  1     65.15 10.118  0.003574 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From we should probably include hp and wt variables as model 6 reduces residual variance significantly

looking at the am coefficients for model adjusted with other variables

```
##         Estimate Std. Error     t value     Pr(>|t|)
## [1,]  7.24493927   1.764422  4.10612698 2.850207e-04
## [2,]  2.56703470   1.291428  1.98774895 5.635445e-02
## [3,]  1.83345825   1.436100  1.27669297 2.118396e-01
## [4,] -0.02361522   1.545645 -0.01527855 9.879146e-01
## [5,]  5.27708531   1.079541  4.88826953 3.460318e-05
## [6,]  2.08371013   1.376420  1.51386198 1.412682e-01
## [7,]  2.15927074   1.435176  1.50453413 1.440531e-01
```

Model 6 am beta estimate fails to reject null hyp, thus we will carry on with model 5 which is the better compromise between RSS and am beta estimate, meaning automatic cars increase mpg by 5.2770853 (which lies in the 95% confident range computed previously). It should be noted that, r-squared is 0.7820346 which is a pretty reliable regression.

Residuals review

Before validating the model above, let's review the residuals and check for potential outliers.

From the charts figure 3, we note that:
- Maserati Bora and Ford Pantera are the most levered data points
- Maserati Bora, Lotus Europa and Toyota Corolla are the most influent points beta wise
- Maserati Bora, Lotus Europa and Toyota Corolla are the most sensitive predicted values

As a result, we will consider Maserati Bora and Toyota Corolla as outliers hereafter.

```
## Loading required package: carData
```

```
##
## Call:
## lm(formula = mpg ~ am + hp, data = mtcarsNo)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9351 -2.4641  0.2883  1.5324  5.9839
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.221909   1.575272  17.281 3.99e-16 ***
## am           4.297659   1.101218   3.903 0.000572 ***
## hp          -0.062862   0.009047  -6.948 1.82e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.684 on 27 degrees of freedom
## Multiple R-squared:  0.7855, Adjusted R-squared:  0.7696
## F-statistic: 49.45 on 2 and 27 DF,  p-value: 9.406e-10
```

After removal of Toyota Corolla and Maseratti Bora, am beta estimate p value is lower than 5%, then we reject the Null hypothesis i.e. accepting am beta estimate value. Residual standard error is down from to 2.684 from 2.909 before outliers removal.

Conclusion

From the regression model without outliers fig 4, the difference in intercept is the mpg difference between manual and automatic transmission. Thus manual cars mpg is **4.3** higher than automatic cars.

Appendix

```
##                    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
```

Exploring the relationship between transmission and mpg

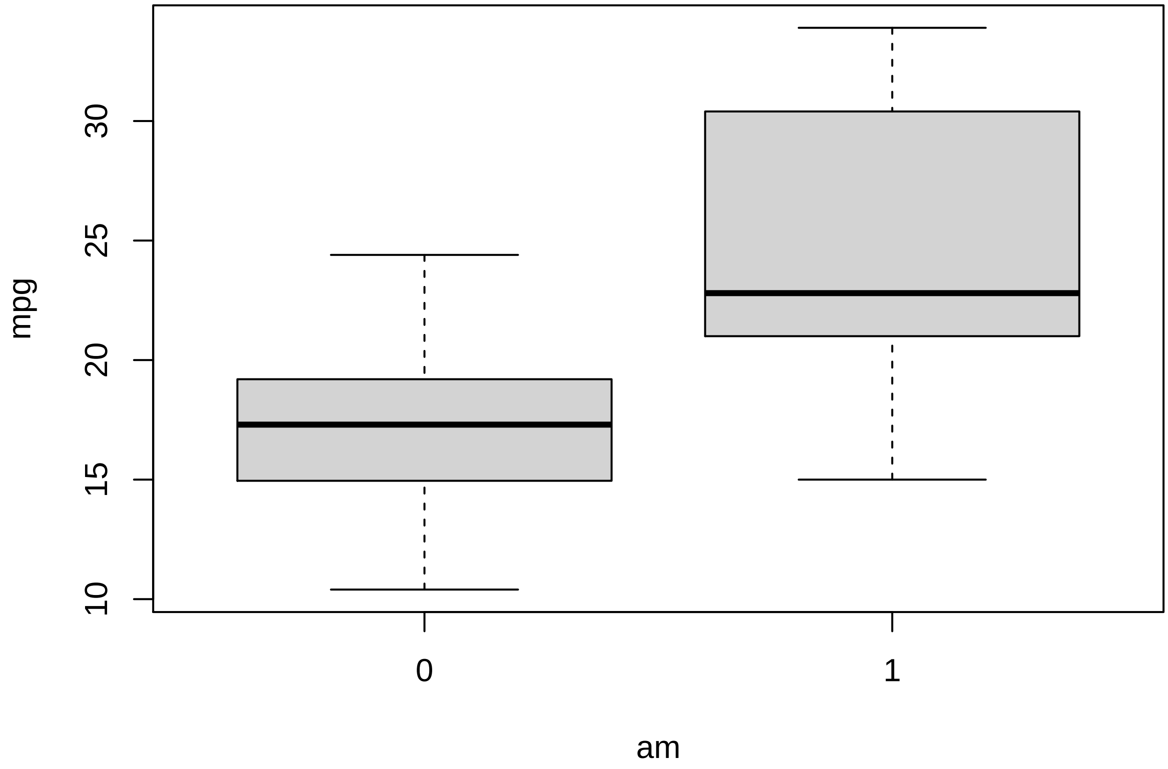

Figure 1: mpg boxplot per transmission factor

Exploring the relationship between available variables

Exploring the regression residuals

Exploring mpg as outcome with hp regressors

```
## `geom_smooth()` using formula 'y ~ x'
```

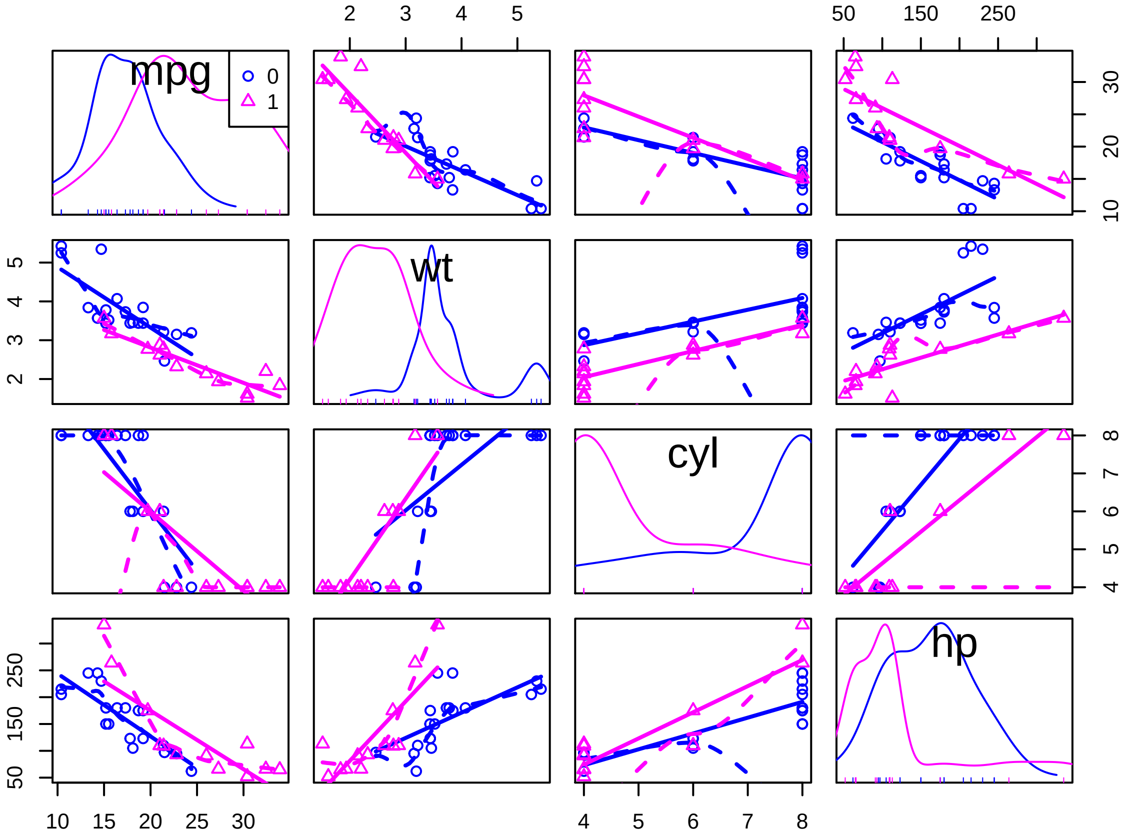

Figure 2: Variable relationships

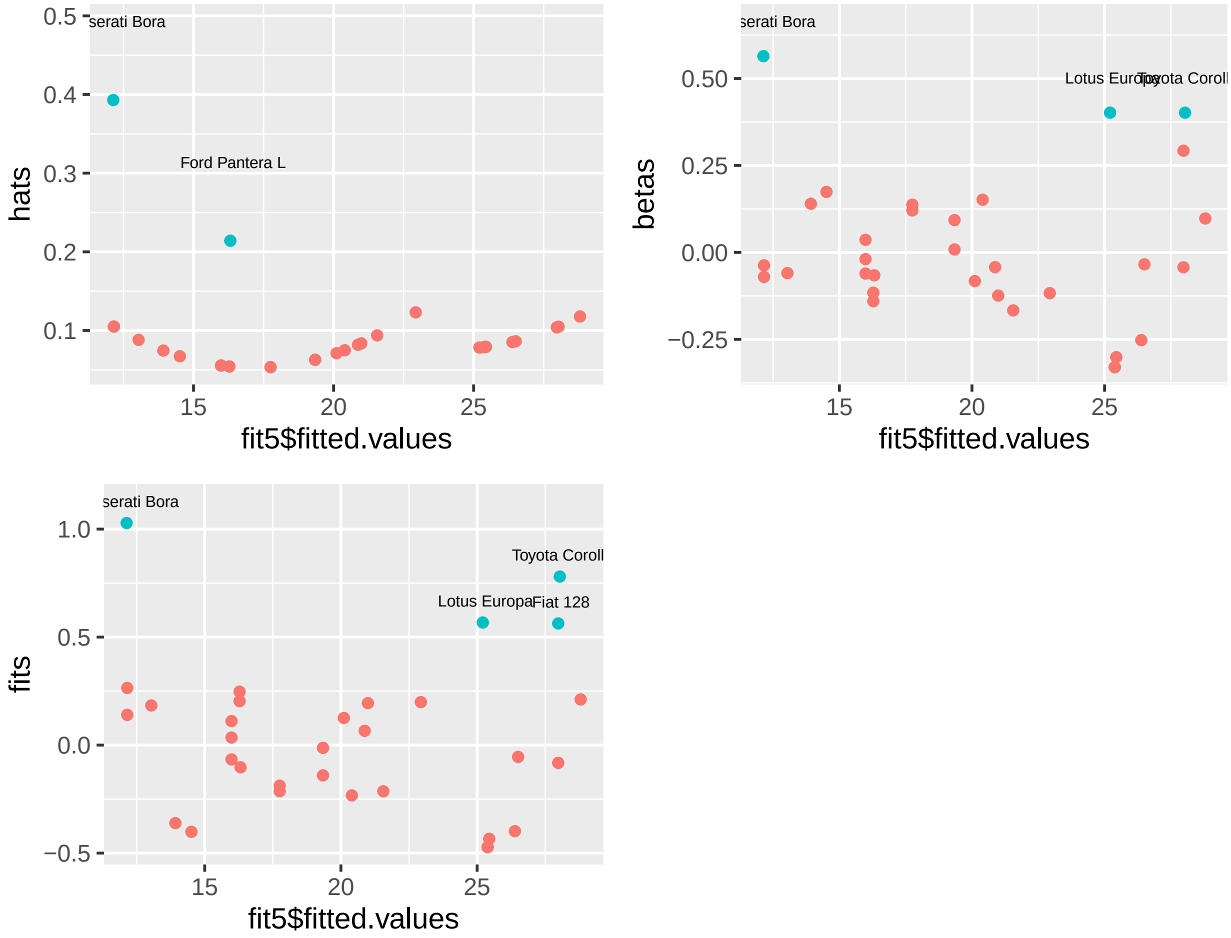

Figure 3: Residual charts

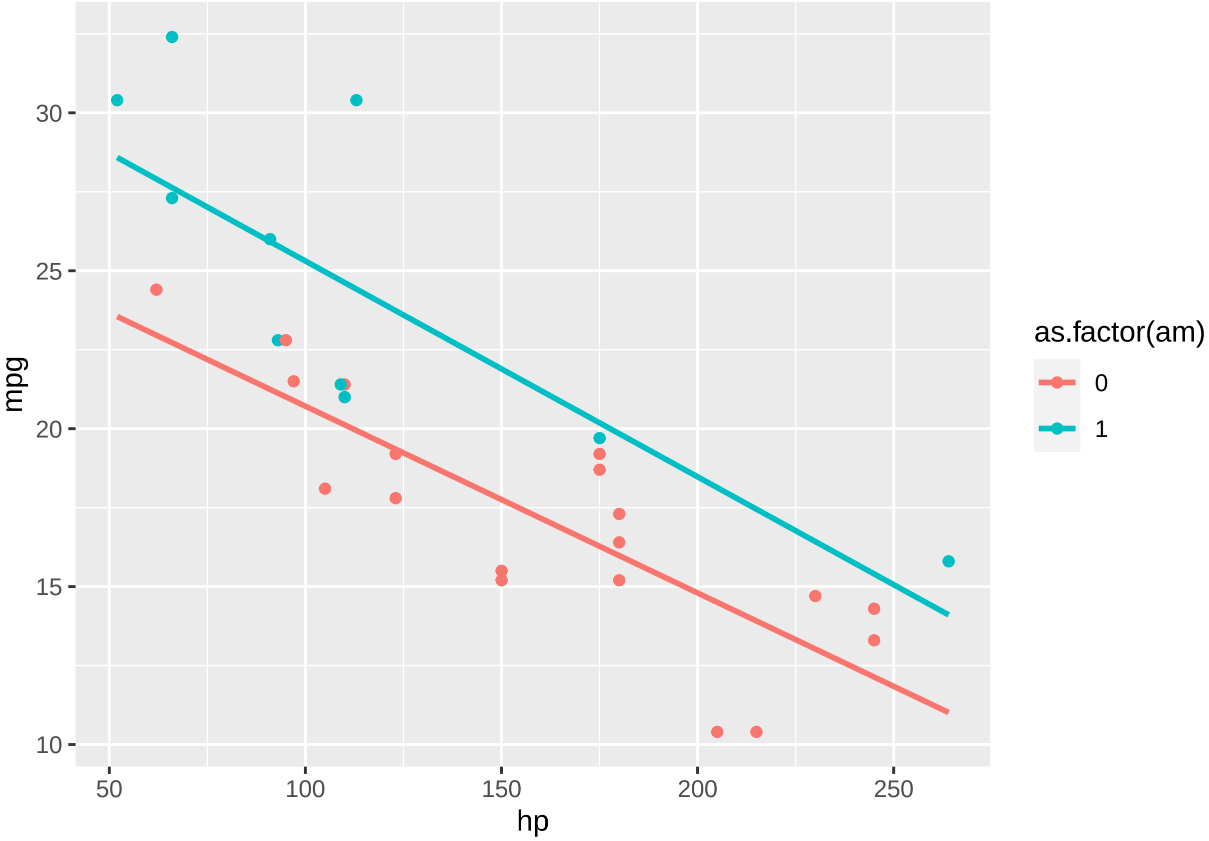

Figure 4: mpg regression