Regression Final Assignment

Szymon Tomczyk

9/17/2020

Summary

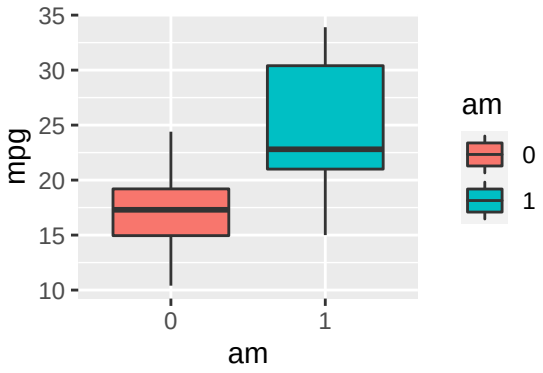
In this assgnment we are supposed to analyse the relationsip between the mpg and other variables from the mtcars dataset. Using linear regression we should answer the following questions: 1. Is an automatic or manual transmission better for MPG? 2. What is the MPG difference between automatic and manual transmissions?

Exploratory analysis

After initial look at the data I noticed that all variables were loaded with the class "numeric". The variables cyl, vs, am, gear and carb were transformed into fators

```
mtcars[,c(2, 8:11)] <- lapply(mtcars[,c(2, 8:11)], as.factor)
str(mtcars)
```

After looking at the results of two sample t-test comapring the mpg of cars with manual and automatic transmission the mean of two groups the H0 assuming the means in two groups are equal need to be rejected (p-value = 0.001374). This result shows promise for further analysis



Model selection

To try to select the optimal model I first used all available variables as regressors of mpg. Afterwards, I used vif function to look for the variables that are strongly correlated and that contribute strongly to variance inflation. In the next step I made a model with only am as a regressor. Next I started to add variables expected to have strong influence on mpg that did not excessively contribute to variance inflation.

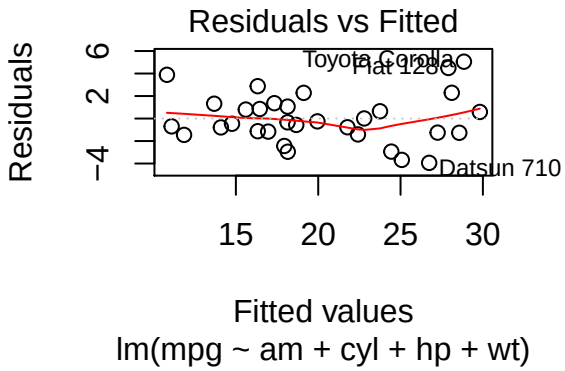
The evaluation of the nested models shows that the optimal model seems to be the one including cyl, hp and wt in addition to the am as regressors. Adding additional variables does not seem to produce significantly better results (does not decrease the deviance much further). Thus the Model 3 was selected for further analysis.

Diagnostics

Firstly, I verified that the residuals follow an approximately normal distribution using Shapiro-Wilk's normality test. The obtain results suggest that the distribution of residuals of the Model 3 was not significantly different from normal distribution.

Next, I looked for influential outliers using 'influence.measures' function. This approach revealed 3 points that could potentially affect the model. However, removing them did not change the outcome of the analysis so decided to keep them in the final model (not shown here due to the volume constrains).

Lasty, I looked at the diagnostic plots (i.e. residuals vs. fitted values) and did not notice any pattern in the the distribution of the residuals.



Conclusion

When looking simply at the mean fuel consumption of cars with manual and automatic transmission it would seam that the manual cars consume less fuel (T-test). However, when a more complex linear multivariate model is build we can see that many more variables influence the mpg than just the type of transmission. In this analysis the selected optimal model seems to indicate a modest increase of 1.81 mpg in the cars with manual transmission. The p-value for this result is 0.21 so above the alpha = 0.05 , thus, we cannot conclude that there is a difference in fuel economy between automatic and manual transmissions. Further investigation of the confidence intervals shows that the 95% confidence interval for this coefficient is wide (-1.06 to 4.68) and contains zero. In the final conclusion, given the data we cannot reject H0 assuming equal consumption between manual and automatic transmission.

```
summary(fit3)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## am1          1.80921138 1.39630450  1.295714 2.064597e-01
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
```

Apendix

Exploratory analysis

Structure of the data

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

T-test used for exploration

```
t.test(mpg ~ am, mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## 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:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Model selection

Initial model evaluation

```
vif(fit.all)
```

```
##            GVIF Df GVIF^(1/(2*Df))
## cyl  128.120962  2        3.364380
## disp  60.365687  1        7.769536
## hp    28.219577  1        5.312210
```

```
## drat   6.809663  1        2.609533
## wt    23.830830  1        4.881683
## qsec  10.790189  1        3.284842
## vs     8.088166  1        2.843970
## am     9.930495  1        3.151269
## gear  50.852311  2        2.670408
## carb 503.211851  5        1.862838
```

Evaluation of the nested models

```
anova(fit1, fit2, fit3, fit4, fit5, fit6, fit.all)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + hp
## Model 3: mpg ~ am + cyl + hp + wt
## Model 4: mpg ~ am + cyl + hp + wt + vs
## Model 5: mpg ~ am + cyl + hp + wt + vs + gear
## Model 6: mpg ~ am + cyl + hp + wt + vs + gear + carb
## Model 7: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     27 197.20  3    523.70 21.7478 1.021e-05 ***
## 3     26 151.03  1     46.17  5.7524   0.02992 *
## 4     25 143.68  1      7.35  0.9152   0.35391
## 5     23 140.24  2      3.44  0.2141   0.80969
## 6     18 134.95  5      5.29  0.1317   0.98260
## 7     15 120.40  3     14.55  0.6043   0.62226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostics

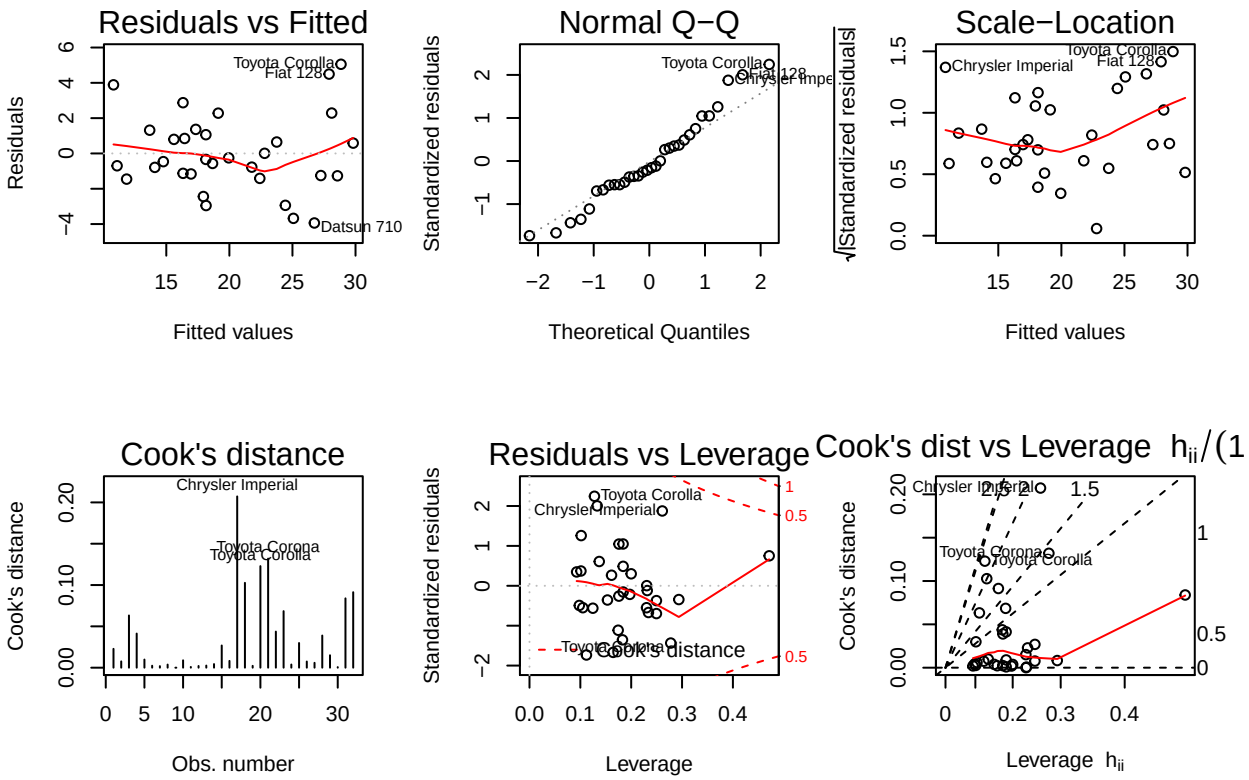
Assesment of normality of the distribution of residuals

```
shapiro.test(fit3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit3$residuals
## W = 0.96807, p-value = 0.4479
```

Plots for model diagnostics

```
par(mfrow = c(2,3))
plot(fit3, which=1:6)
```



Conclusion

Confidence intervals

```
confint(fit3)
```

```
##                    2.5 %       97.5 %
## (Intercept) 28.35390366 39.062744138
## am1         -1.06093363  4.679356394
## cyl6        -5.92405718 -0.138631806
## cyl8        -6.85902199  2.531671342
## hp          -0.06025492 -0.003963941
## wt          -4.31718120 -0.676477640
```