Regression Models Course Progect

Summary

THis analysis was written for Motor Trend, a magazine about the automobile industry, to answer the main question, if an automatic or manual transmission better for car fuel consumption. The study could not determine the answer to the question, as all multivariate models showed the insignificance of the transmission variable.

Loading and Processing Data

First, load required libraries, download dataset and consider its stucture.

```
library(tidyverse)
data("mtcars")
```

From the structure (Appendix 1) we can see, that all the data is represented as numeric variables, although some of it should be factors. Since this is important for the regression model, change types and define levels for transmission variable (am).

```
mtcars<- mutate(mtcars,  cyl=as.factor(cyl), vs=as.factor(vs), gear=as.factor(gear),
                am=factor(mtcars$am,labels = c("automatic","manual")), carb=as.factor(carb))
```

Exploratory analysis

Briefly explore relationship between transmition and fuel consumption: build a Boxplot and the summary table (Appendix 2)

If we'd make a conclusion about the relationship of these two variables without taking into account the other variables from dataset, then data show a significantly greater efficiency of the cars with manual transmission over automatic.

Regression analysis

Initially made simple model relation fuel consumption (mpg) to transmission type (am).

```
fit<-lm(mpg~am,mtcars)
summary(fit)$coefficients
```

```
##              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
```

```
summary(fit)$adj.r.squared
```

```
## [1] 0.3384589
```

This model confirms conclusions from Exploratory analysis: coefficient of the manual transmission significantly greater than zero, p-value is 0.00029. But since adjasted r2 is relatively small, then check another models with several variables.

Lets try model that includes all variables from dataset (Appendix 3)

```
fit2<-lm(mpg~.,mtcars)
```

Here coefficient of manual transmission retain its sign, but now it very insignificant, with p-value 0.7. This model has grater adjasted r2 (compared with the first model), but noticeably overfitted (among 17 variables there are not any significant). So lets try another model with fewer variables, to avoid possible collinearity as the reason of high p-value of **am** variable. In the second model the most significant variables were horsepower and weight, so include in the next model these variables + transmission variable.

```
fit3<-lm(mpg~hp+wt+am,mtcars)
summary(fit3)$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
## ammanual     2.08371013 1.376420152  1.513862 1.412682e-01
```

```
summary(fit3)$adj.r.squared
```

```
## [1] 0.8227357
```

Third model has the biggest adjasted r2 (0.82), but transmission variabe ia also insignificant.

Residuals diagnostic

Let's build standard residuals plots

All residuals grafs looks nomal since there is no pattern in the model in fitted vs residuals plot, normal q-q close to theoritical quantiles and there is not sign of geteroscedasticity in fitted vs standardized residuals.

Conclusion

1. Since the variable did not show significance in all multivariate models, we can not determine which transmission better for MPG.
2. On average, manual transmission gives +1-2 miles per galon, but confidence interval with alpha = 0.05 includes 0.

Appendix

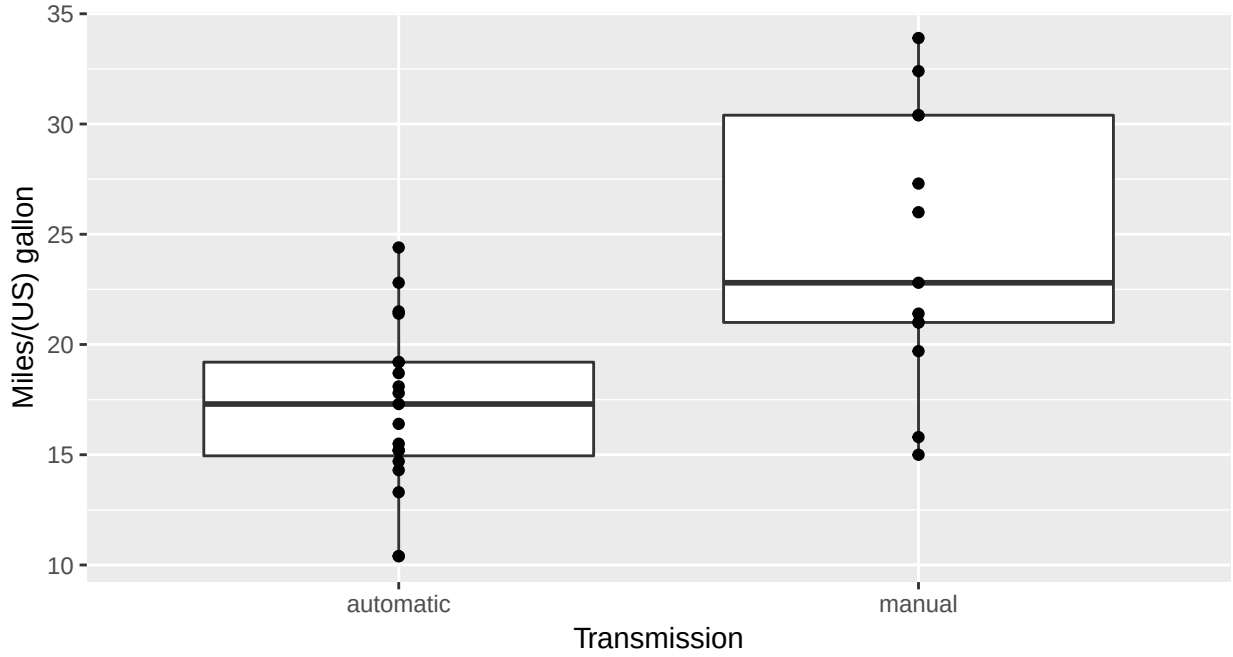
1. Sructure of dataset

```
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 "automatic","manual": 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 ...
```

2. Eploratory analysis

```
ggplot(mtcars, aes(am, mpg))+geom_boxplot()+geom_point()+xlab("Transmission")+ylab("Miles/(US) gallon")
```




```
group_by(mtcars,am) %>% summarise(mean=mean(mpg), min=min(mpg), max = max(mpg), sd=sd(mpg))
```

```
## # A tibble: 2 x 5
##   am         mean   min   max    sd
##   <fct>     <dbl> <dbl> <dbl> <dbl>
## 1 automatic  17.1  10.4  24.4  3.83
## 2 manual     24.4  15    33.9  6.17
```

3. Model 2

```
fit2<-lm(mpg~.,mtcars)
summary(fit2)$coefficients
```

```
##                Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) 23.87913244 20.06582026  1.19004018 0.25252548
## cyl6        -2.64869528  3.04089041 -0.87102622 0.39746642
## cyl8        -0.33616298  7.15953951 -0.04695316 0.96317000
## disp         0.03554632  0.03189920  1.11433290 0.28267339
## hp          -0.07050683  0.03942556 -1.78835344 0.09393155
## drat         1.18283018  2.48348458  0.47627845 0.64073922
## wt          -4.52977584  2.53874584 -1.78425732 0.09461859
## qsec         0.36784482  0.93539569  0.39325050 0.69966720
## vs1          1.93085054  2.87125777  0.67247551 0.51150791
## ammanual     1.21211570  3.21354514  0.37718957 0.71131573
## gear4        1.11435494  3.79951726  0.29328856 0.77332027
## gear5        2.52839599  3.73635801  0.67670068 0.50889747
## carb2       -0.97935432  2.31797446 -0.42250436 0.67865093
## carb3        2.99963875  4.29354611  0.69863900 0.49546781
## carb4        1.09142288  4.44961992  0.24528452 0.80956031
## carb6        4.47756921  6.38406242  0.70136677 0.49381268
## carb8        7.25041126  8.36056638  0.86721532 0.39948495
```

```r
summary(fit2)$adj.r.squared
```

```
## [1] 0.7790215
```

4. Residuals plots

```r
par(mfrow = c(2,2))
plot(fit3)
```