Regression Models

Haoyi Wei

2022-12-20

Peer-graded Assignment: regression Models Course Project

Executive summary

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Analysis

Import the dataset

```
# load the mtcars data
data(mtcars)
```

Exploratory Analysis

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

```
dim(mtcars)
```

```
## [1] 32 11
```

```
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 : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ 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  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

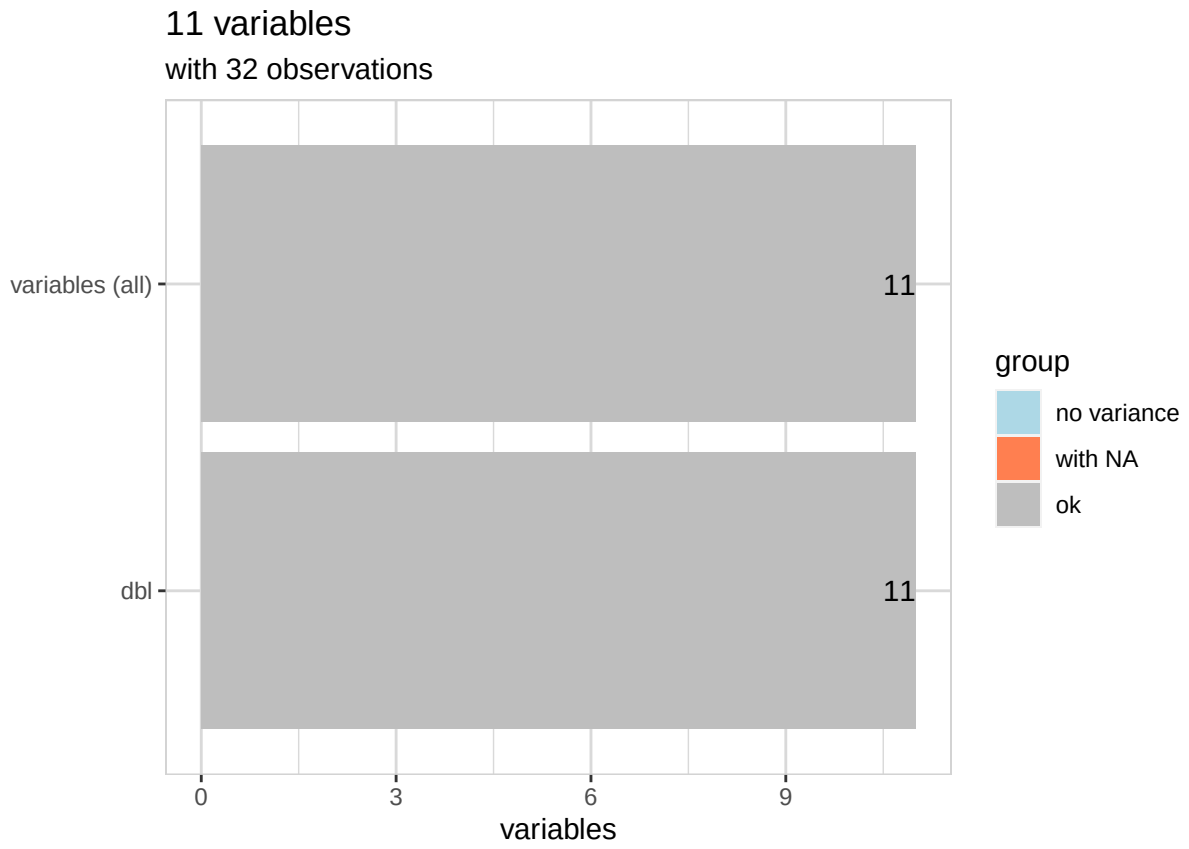
```
library(explore)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

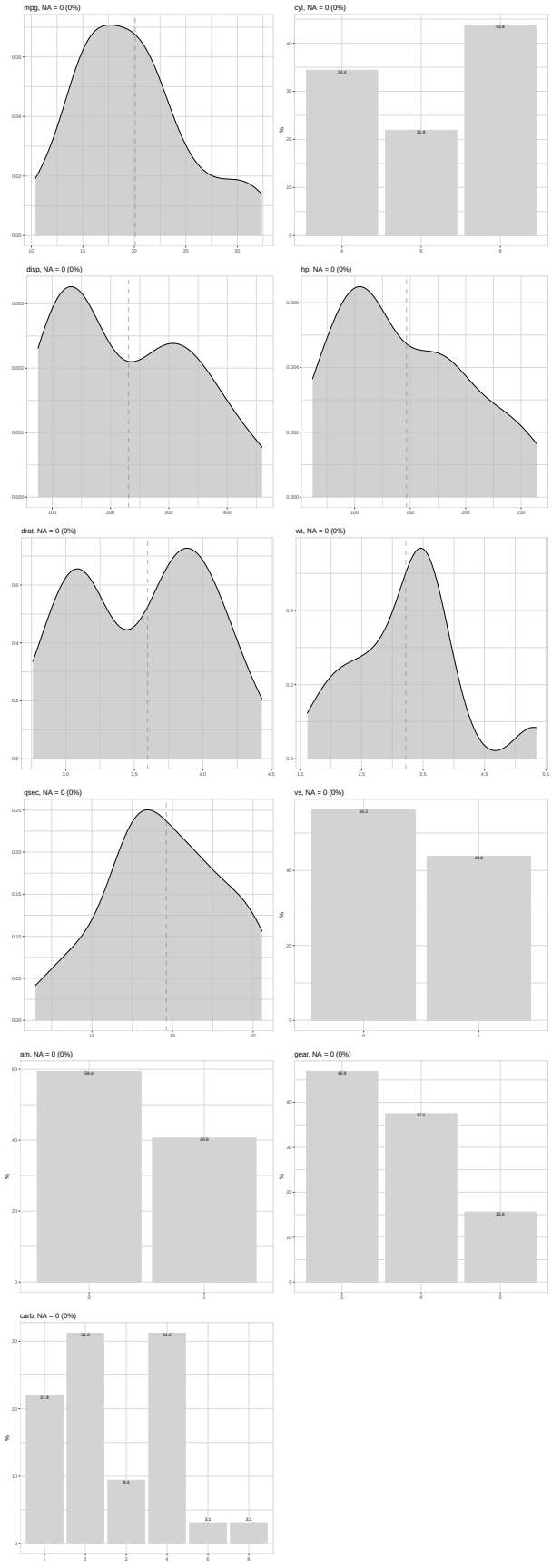
```
# how many variables?
explore_tbl(mtcars)
```




```
# describe the dataset
describe(mtcars)
```

```
## # A tibble: 11 x 8
##    variable type     na na_pct unique   min   mean    max
##    <chr>    <chr> <int>  <dbl>  <int> <dbl>  <dbl>  <dbl>
##  1 mpg      dbl       0      0     25 10.4   20.1   33.9
##  2 cyl      dbl       0      0      3  4      6.19   8
##  3 disp     dbl       0      0     27 71.1  231.   472
##  4 hp       dbl       0      0     22 52    147.   335
##  5 drat     dbl       0      0     22  2.76   3.6    4.93
##  6 wt       dbl       0      0     29  1.51   3.22   5.42
##  7 qsec     dbl       0      0     30 14.5   17.8   22.9
##  8 vs       dbl       0      0      2  0      0.44   1
##  9 am       dbl       0      0      2  0      0.41   1
## 10 gear     dbl       0      0      3  3      3.69   5
## 11 carb     dbl       0      0      6  1      2.81   8
```

```
# explore the variables
explore_all(mtcars)
```

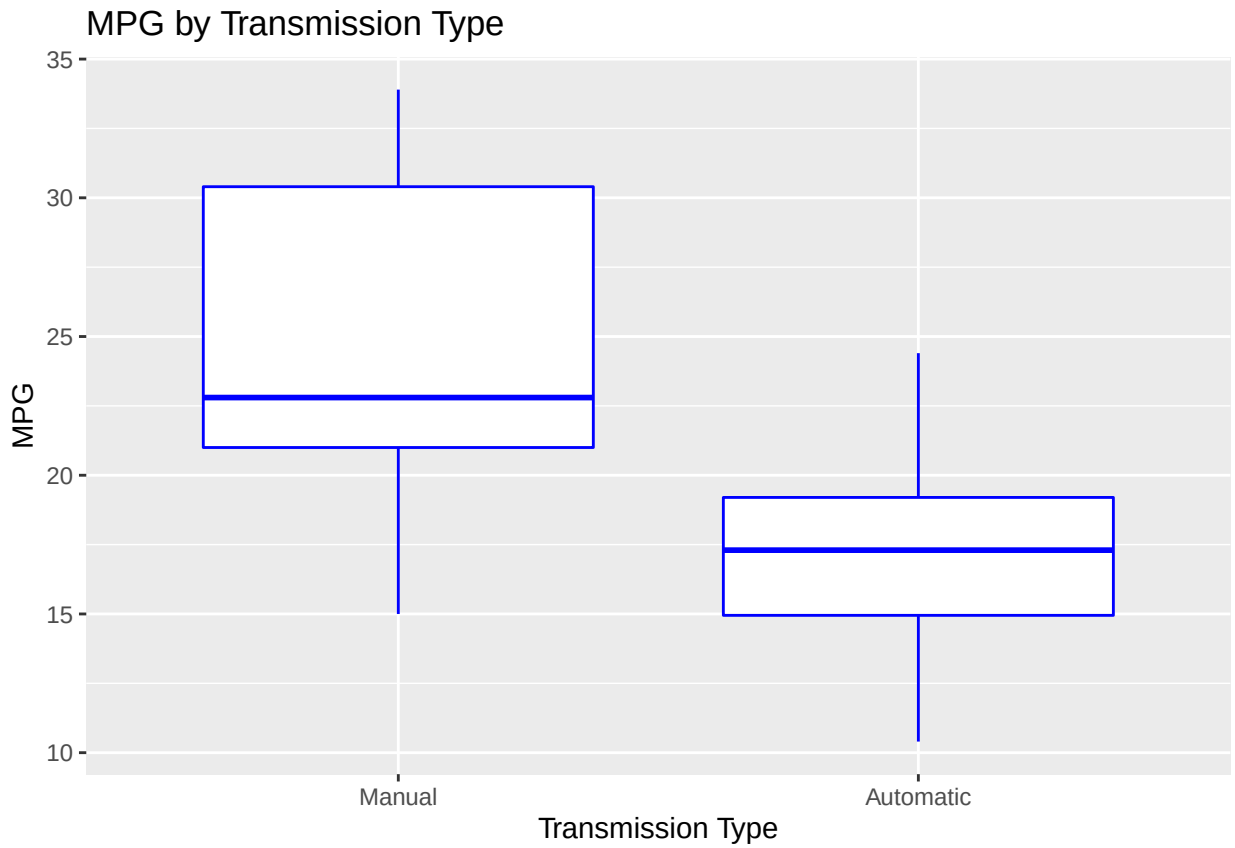

mpg, NA = 0 (0%)
cyl, NA = 0 (0%)
34.4
21.9
43.8
%
disp, NA = 0 (0%)
hp, NA = 0 (0%)
drat, NA = 0 (0%)
wt, NA = 0 (0%)
qsec, NA = 0 (0%)
vs, NA = 0 (0%)
56.2
43.8
%
am, NA = 0 (0%)
59.4
40.6
%
gear, NA = 0 (0%)
46.9
37.5
15.6
%
carb, NA = 0 (0%)
21.9
31.2
9.4
31.2
3.1
3.1
%

Descriptive Analysis

```
# descriptive analysis:is an automatic or manual transmission better for MPG

library(ggplot2)
library(dplyr)

mtcars$am <- factor(mtcars$am)
levels(mtcars$am) <- list(Manual=1, Automatic=0)
ggplot(mtcars, aes(x=am, y=mpg)) +
          geom_boxplot(color="blue") +
          ggtitle("MPG by Transmission Type") +
          labs(x="Transmission Type", y="MPG")
```



The mean and median of MPG different between manual and auto

Analytical Analysis

```
data(mtcars)
model <- lm(mpg~., data=mtcars)
summary(model)
```

```
##
```

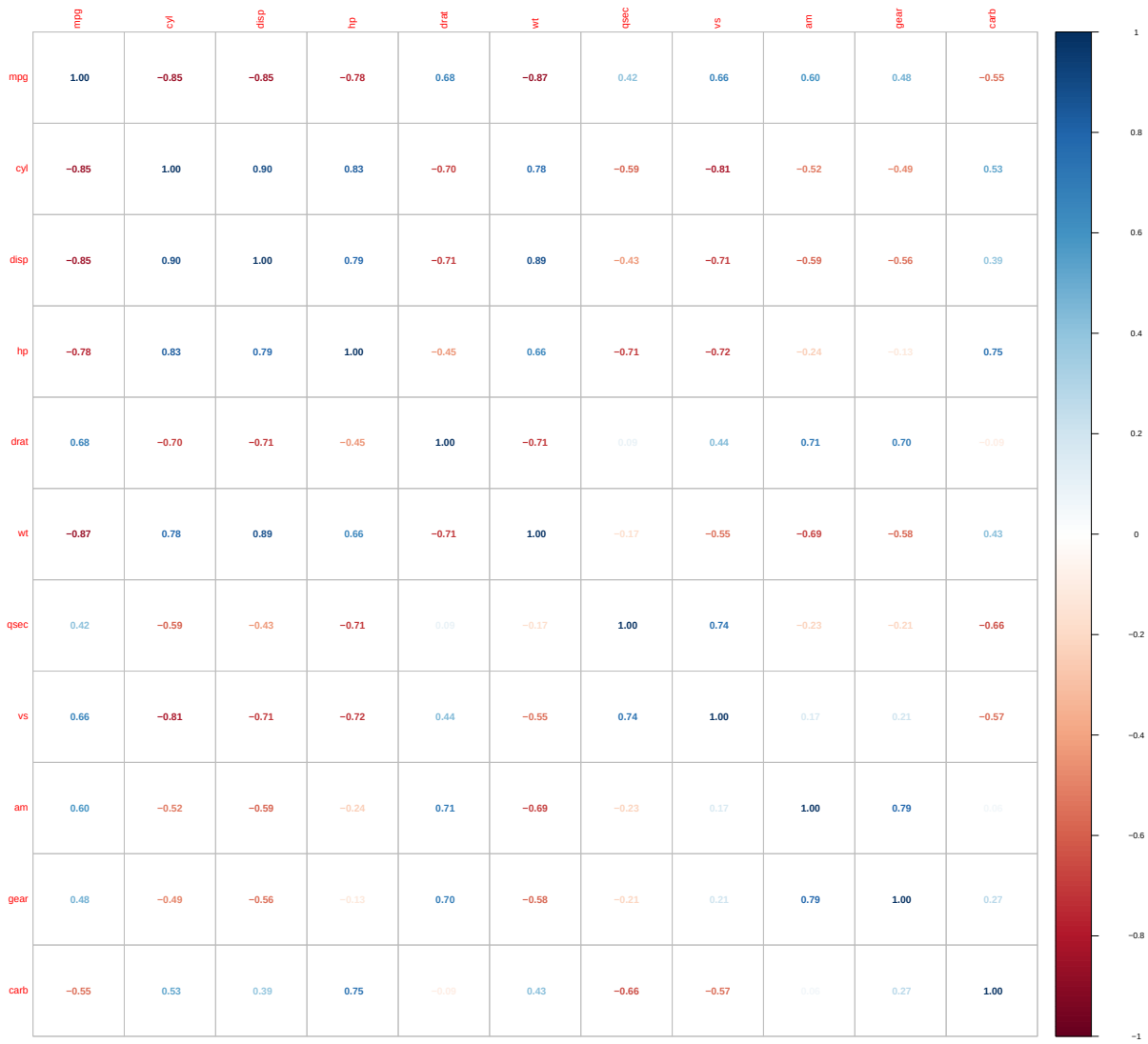
```
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Detect multicollinearity

```
# Detect multicollinearity with correlation matrix
library("corrplot")
```

```
## corrplot 0.92 loaded
```

```
mtcars$am <- as.numeric(mtcars$am)
corrplot(cor(mtcars), method = "number")
```




```
# Test for Multicollinearity with Variance Inflation Factors (VIF)

#TOLERANCE & VARIANCE INFLATION FACTOR (VIF)
library("olsrr")
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_vif_tol(model)
```

```
##    Variables  Tolerance       VIF
## 1        cyl 0.06504559 15.373833
## 2       disp 0.04625295 21.620241
## 3         hp 0.10170833  9.832037
## 4       drat 0.29632966  3.374620
## 5         wt 0.06594180 15.164887
## 6       qsec 0.13283814  7.527958
## 7         vs 0.20137444  4.965873
## 8         am 0.21512374  4.648487
## 9       gear 0.18665589  5.357452
## 10      carb 0.12644228  7.908747
```

As a rule of thumb, a VIF exceeding 5 requires further investigation, whereas VIFs above 10 indicate multicollinearity. Ideally, the Variance Inflation Factors are below 3. The result indicate the possibilities of multicollinearity

To address the multicollinearity issue, we use the stepwise selection method

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:olsrr':
##
##     cement

## The following object is masked from 'package:dplyr':
##
##     select
```

```
step <- stepAIC(model, direction="both", trace=FALSE)
summary(step)$coeff
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

The p-value for am is greater than 0.1, we can not conclude the coefficient on am is different from zero at convention significant levels.

```
# fit the new model
final_model <- lm(mpg ~ wt+qsec+factor(am), data = mtcars)
summary(final_model)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## factor(am)1  2.935837  1.4109045  2.080819 4.671551e-02
```

Conclusion

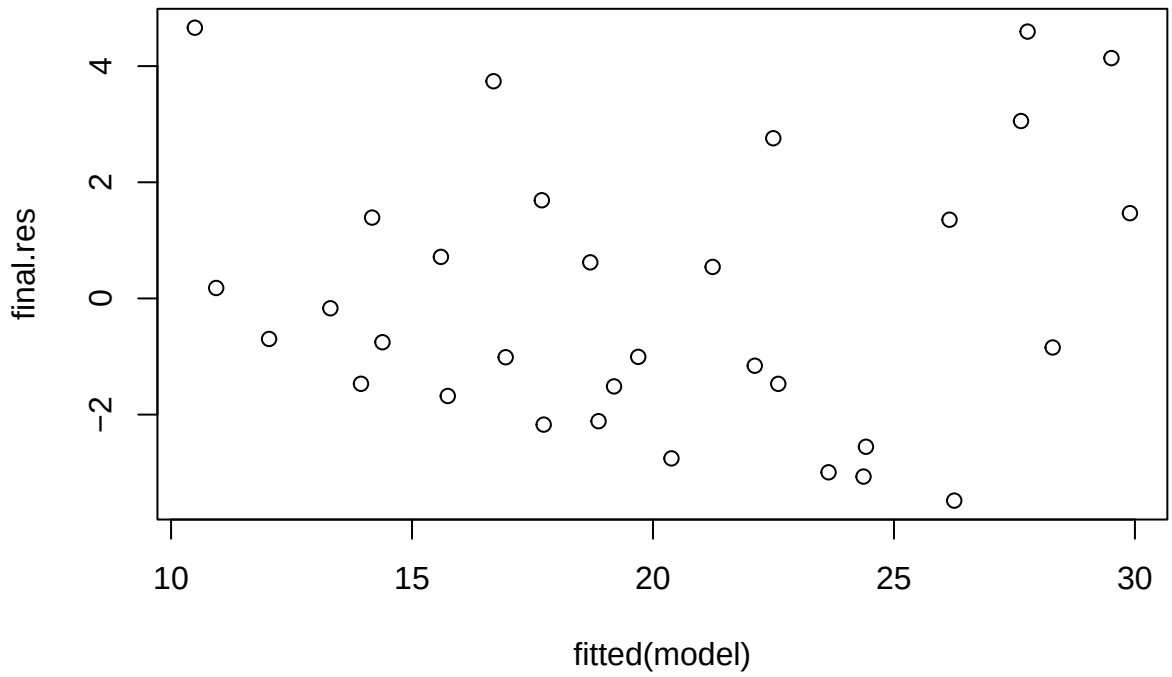
On average, manual transmission cars have 2.94 MPGs more than automatic transmission cars

Appendix

Residual plot

```
final.res = resid(final_model)

# We now plot the residual against the observed values of the variable waiting.
#produce residual vs. fitted plot
plot(fitted(model), final.res)
```



The distribution of the residul fairly consitent across different level of fitted level. No major concern about the model design