Regression Models

Javier

27 de octubre de 2015

Summary.

In this report we have used the dataset from the 1974 Motor Trend US magazine to answer the following questions: . Is an automatic or manual transmission better for miles per gallon (MPG)? . How different is the MPG between automatic and manual transmissions? The objective is to determine through statistical techniques if there is statistically significant difference between the mean MPG for automatic and manual transmission cars. First we make an exploratory analysis of the data set. Multivariable regression model using backward and forward is then constructed. Finally, conclusions are described.

Exploratory Data Analysis.

We want to compare the levels Automatic and Manual of the dichotomous variable using side-by-side boxplots:

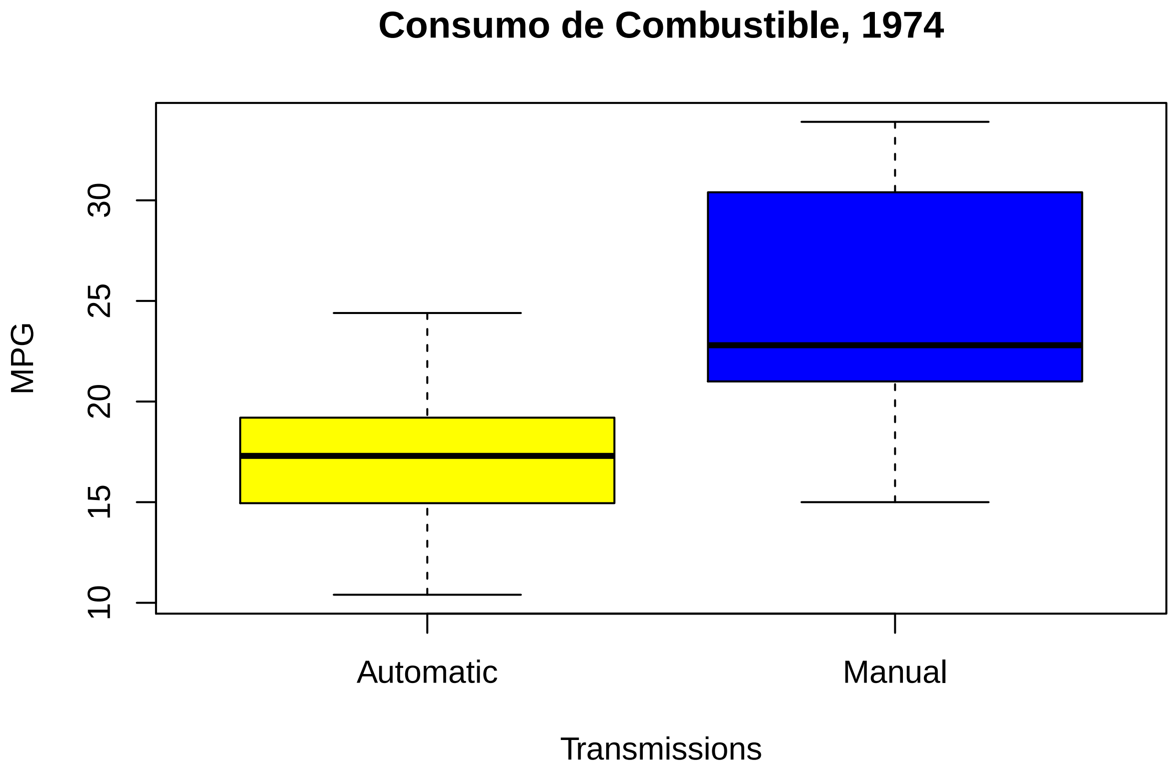
```
data(mtcars)
attach(mtcars)
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   10.40   15.42   19.20   20.09   22.80   33.90
```

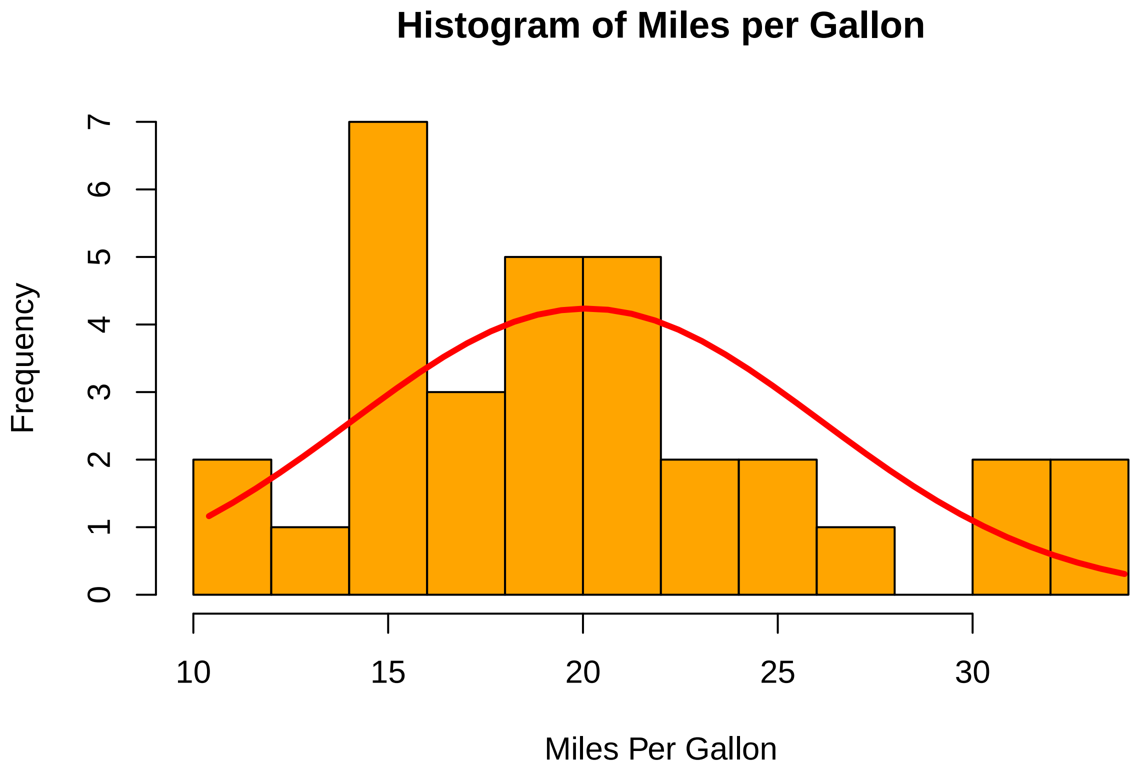
```
aggregate(mtcars$mpg,list(mtcars$am),summary)
```

```
##   Group.1 x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.
## 1       0  10.40     14.95    17.30  17.15     19.20  24.40
## 2       1  15.00     21.00    22.80  24.39     30.40  33.90
```

```
boxplot(mpg ~ am, data=mtcars,outline=TRUE,xlab = "Transmissions",
ylab = "MPG",names=c("Automatic","Manual"),col=c("yellow","blue"),
main="Consumo de Combustible, 1974")
```



The mean difference in absolute value between cars with manual transmissions and Cars with automatic transmissions is 7.24. The plot shows that the variable mpg performs better mpg in cars with manual transmissions than in cars with automatic transmissions. We will analyze whether mpg variable follows a normal distribution.

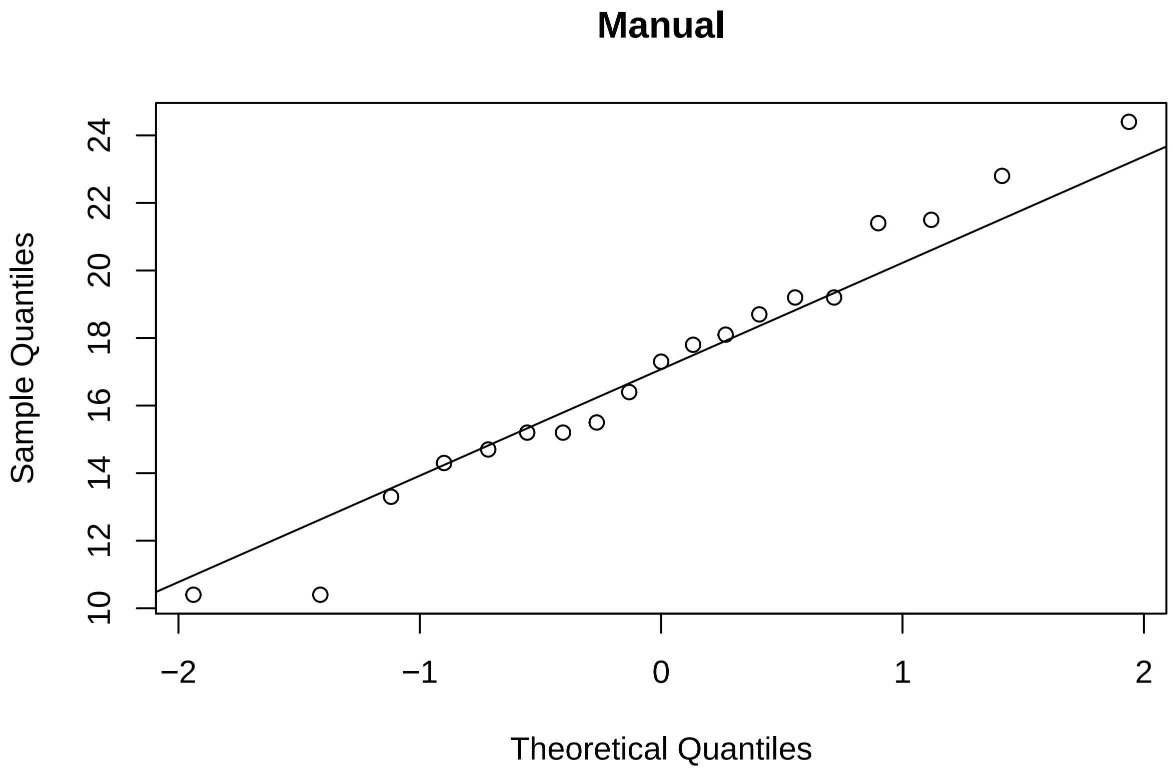


To make a QQ plot this way, R has the special qqnorm() function. As the name implies, this function plots your sample against a normal distribution. You simply give the sample you want to plot as a first argument and add any graphical parameters you like. R then creates a sample with values coming from the standard normal distribution, or a normal distribution with a mean of zero and a standard deviation of one. With this second sample, R creates the QQ plot as explained before.

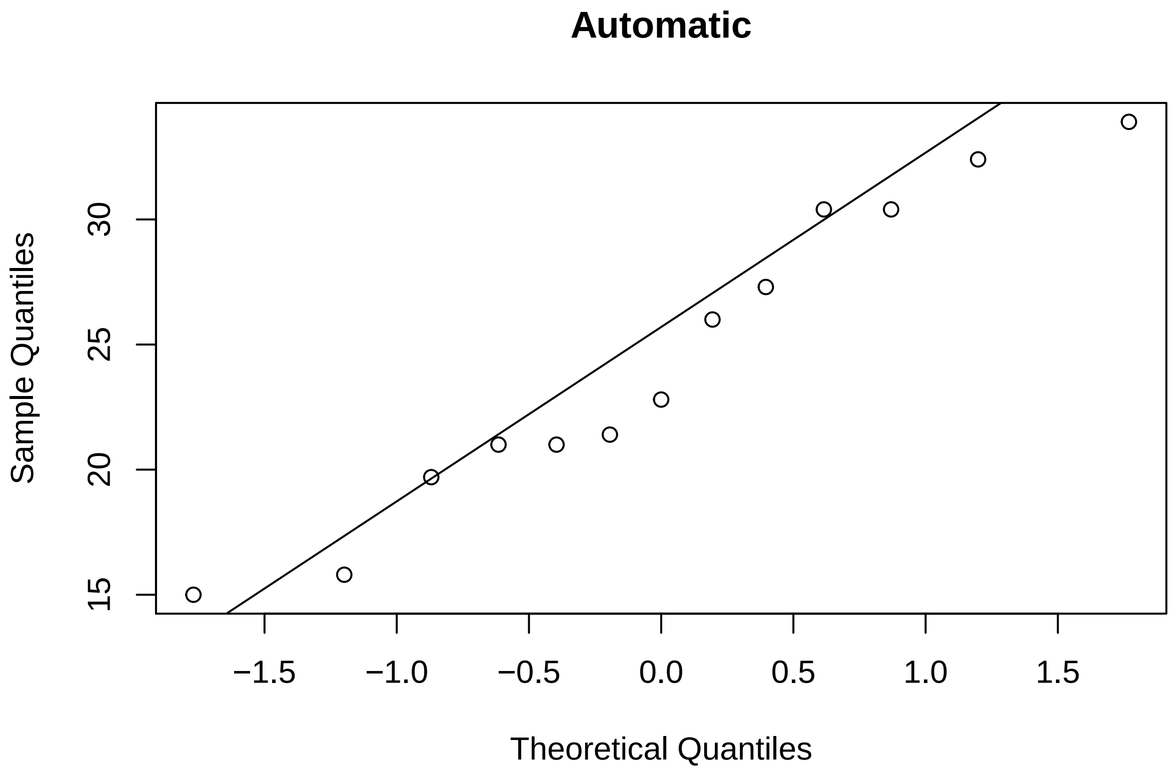
We can use the qqnorm() function twice to create both plots. For the variable am, you can use the following code:

```
library(stats)

qqnorm( mtcars$mpg[mtcars$am==0], main='Manual')
qqline( mtcars$mpg[mtcars$am==0] )
```



```
qqnorm( mtcars$mpg[mtcars$am==1], main='Automatic')
qqline( mtcars$mpg[mtcars$am==1] )
```



Now we study if there is a signficiant difference in the mean MPG between manual transmission and automatic transmission with the T-Student test:

```
t.test(mtcars$mpg[mtcars$am==0],mtcars$mpg[mtcars$am==1])
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg[mtcars$am == 0] and mtcars$mpg[mtcars$am == 1]
## 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 of x mean of y
##  17.14737  24.39231
```

This test shows the existence of signficiant difference in the mean in the mean MPG between manual transmission and automatic transmission with a p-value of 0.001374.

Building our Model.

Correlation

Before applying the regression model we studied the correlation of predictors:

```r
data(mtcars)
cor(mtcars)
```

```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

```r
sort(cor(mtcars)[1,])
```

```
##         wt        cyl       disp         hp       carb       qsec
## -0.8676594 -0.8521620 -0.8475514 -0.7761684 -0.5509251  0.4186840
##       gear         am         vs       drat        mpg
##  0.4802848  0.5998324  0.6640389  0.6811719  1.0000000
```

We note that the variables wt, cyl, disp, and hp are very correlated with the dependent variable mpg. Instead the variable is not strongly correlated am.

The variables cyl, vs, am, gear and carb should be treated as discrete variables.

```r
cars <- mtcars

cars$am <- as.factor(cars$am)
cars$cyl <- as.factor(cars $cyl)
cars$vs <- as.factor(cars $vs)
cars$am <- as.factor(cars $am)
cars$gear <- as.factor(cars $gear)
cars$carb <- as.factor(cars $carb)
```

To begin our model testing, we fit a multiple linear regression for mpg and we perform Backward Elimination and Forward Selection. Also based on Model AIC (not individual regression coefficients). fit1 and fit2 represent “extreme” models.

```r
library(MASS)
fit1 <- lm(mpg ~ ., data=cars)
fit2 <- lm(mpg ~ 1, data=cars)
met1 <- stepAIC(fit1,direction="backward")
```

```
## Start:  AIC=76.4
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  5   13.5989 134.00 69.828
## - gear  2    3.9729 124.38 73.442
## - am    1    1.1420 121.55 74.705
## - qsec  1    1.2413 121.64 74.732
## - drat  1    1.8208 122.22 74.884
## - cyl   2   10.9314 131.33 75.184
## - vs    1    3.6299 124.03 75.354
## <none>              120.40 76.403
## - disp  1    9.9672 130.37 76.948
## - wt    1   25.5541 145.96 80.562
## - hp    1   25.6715 146.07 80.588
##
## Step:  AIC=69.83
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  2    5.0215 139.02 67.005
## - disp  1    0.9934 135.00 68.064
## - drat  1    1.1854 135.19 68.110
## - vs    1    3.6763 137.68 68.694
## - cyl   2   12.5642 146.57 68.696
## - qsec  1    5.2634 139.26 69.061
## <none>              134.00 69.828
## - am    1   11.9255 145.93 70.556
## - wt    1   19.7963 153.80 72.237
## - hp    1   22.7935 156.79 72.855
##
## Step:  AIC=67
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1    0.9672 139.99 65.227
## - cyl   2   10.4247 149.45 65.319
## - disp  1    1.5483 140.57 65.359
## - vs    1    2.1829 141.21 65.503
## - qsec  1    3.6324 142.66 65.830
## <none>              139.02 67.005
## - am    1   16.5665 155.59 68.608
## - hp    1   18.1768 157.20 68.937
## - wt    1   31.1896 170.21 71.482
##
## Step:  AIC=65.23
## mpg ~ cyl + disp + hp + wt + qsec + vs + am
##
```

```
##        Df Sum of Sq    RSS    AIC
## - disp  1    1.2474 141.24 63.511
## - vs    1    2.3403 142.33 63.757
## - cyl   2   12.3267 152.32 63.927
## - qsec  1    3.1000 143.09 63.928
## <none>              139.99 65.227
## - hp    1   17.7382 157.73 67.044
## - am    1   19.4660 159.46 67.393
## - wt    1   30.7151 170.71 69.574
##
## Step:  AIC=63.51
## mpg ~ cyl + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - qsec  1     2.442 143.68 62.059
## - vs    1     2.744 143.98 62.126
## - cyl   2    18.580 159.82 63.466
## <none>              141.24 63.511
## - hp    1    18.184 159.42 65.386
## - am    1    18.885 160.12 65.527
## - wt    1    39.645 180.88 69.428
##
## Step:  AIC=62.06
## mpg ~ cyl + hp + wt + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - vs    1     7.346 151.03 61.655
## <none>              143.68 62.059
## - cyl   2    25.284 168.96 63.246
## - am    1    16.443 160.12 63.527
## - hp    1    36.344 180.02 67.275
## - wt    1    41.088 184.77 68.108
##
## Step:  AIC=61.65
## mpg ~ cyl + hp + wt + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              151.03 61.655
## - am    1     9.752 160.78 61.657
## - cyl   2    29.265 180.29 63.323
## - hp    1    31.943 182.97 65.794
## - wt    1    46.173 197.20 68.191
```

```
met2<-stepAIC(fit2,direction="forward",scope=list(upper=fit1,lower=fit2))
```

```
## Start:  AIC=115.94
## mpg ~ 1
##
##        Df Sum of Sq     RSS     AIC
## + wt    1    847.73  278.32  73.217
## + disp  1    808.89  317.16  77.397
## + cyl   2    824.78  301.26  77.752
## + hp    1    678.37  447.67  88.427
## + drat  1    522.48  603.57  97.988
## + vs    1    496.53  629.52  99.335
## + gear  2    483.24  642.80 102.003
## + am    1    405.15  720.90 103.672
## + carb  5    500.56  625.49 107.129
## + qsec  1    197.39  928.66 111.776
## <none>              1126.05 115.943
##
## Step:  AIC=73.22
## mpg ~ wt
##
##        Df Sum of Sq    RSS    AIC
## + cyl   2    95.263 183.06 63.810
## + hp    1    83.274 195.05 63.840
## + qsec  1    82.858 195.46 63.908
## + vs    1    54.228 224.09 68.283
## + disp  1    31.639 246.68 71.356
## + gear  2    40.372 237.95 72.202
## <none>              278.32 73.217
## + drat  1     9.081 269.24 74.156
## + am    1     0.002 278.32 75.217
## + carb  5    47.458 230.86 77.235
##
## Step:  AIC=63.81
## mpg ~ wt + cyl
##
##        Df Sum of Sq    RSS    AIC
## + hp    1   22.2810 160.78 61.657
## <none>              183.06 63.810
## + qsec  1   10.9487 172.11 63.837
## + vs    1    1.8416 181.22 65.487
## + disp  1    0.1096 182.95 65.791
## + am    1    0.0903 182.97 65.794
## + drat  1    0.0727 182.99 65.798
## + gear  2    6.6815 176.38 66.620
## + carb  5   10.9080 172.15 71.844
##
## Step:  AIC=61.66
## mpg ~ wt + cyl + hp
##
##        Df Sum of Sq    RSS    AIC
## + am    1    9.7520 151.03 61.655
## <none>              160.78 61.657
## + drat  1    2.4377 158.34 63.168
## + vs    1    0.6545 160.12 63.527
## + disp  1    0.6508 160.13 63.527
## + qsec  1    0.2294 160.55 63.611
## + gear  2    7.3662 153.41 64.156
## + carb  5    8.9383 151.84 69.827
##
## Step:  AIC=61.65
## mpg ~ wt + cyl + hp + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              151.03 61.655
## + vs    1    7.3459 143.68 62.059
## + qsec  1    7.0439 143.98 62.126
## + disp  1    0.6168 150.41 63.524
## + drat  1    0.2202 150.81 63.608
## + gear  2    1.3605 149.66 65.365
## + carb  5    5.6330 145.39 70.438
```

The selected model is as follows (the AIC is the smallest of all models):

```r
fit3<-lm(mpg ~ wt + cyl + hp + am, data=cars)

summary(fit3)
```

```
## 
## Call:
## lm(formula = mpg ~ wt + cyl + hp + am, data = cars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.9387 -1.2560 -0.4013  1.1253  5.0513 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## wt          -2.49683    0.88559  -2.819  0.00908 ** 
## cyl6        -3.03134    1.40728  -2.154  0.04068 *  
## cyl8        -2.16368    2.28425  -0.947  0.35225    
## hp          -0.03211    0.01369  -2.345  0.02693 *  
## am1          1.80921    1.39630   1.296  0.20646    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401 
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

We analyze heteroskedasticity and normality of residuals:

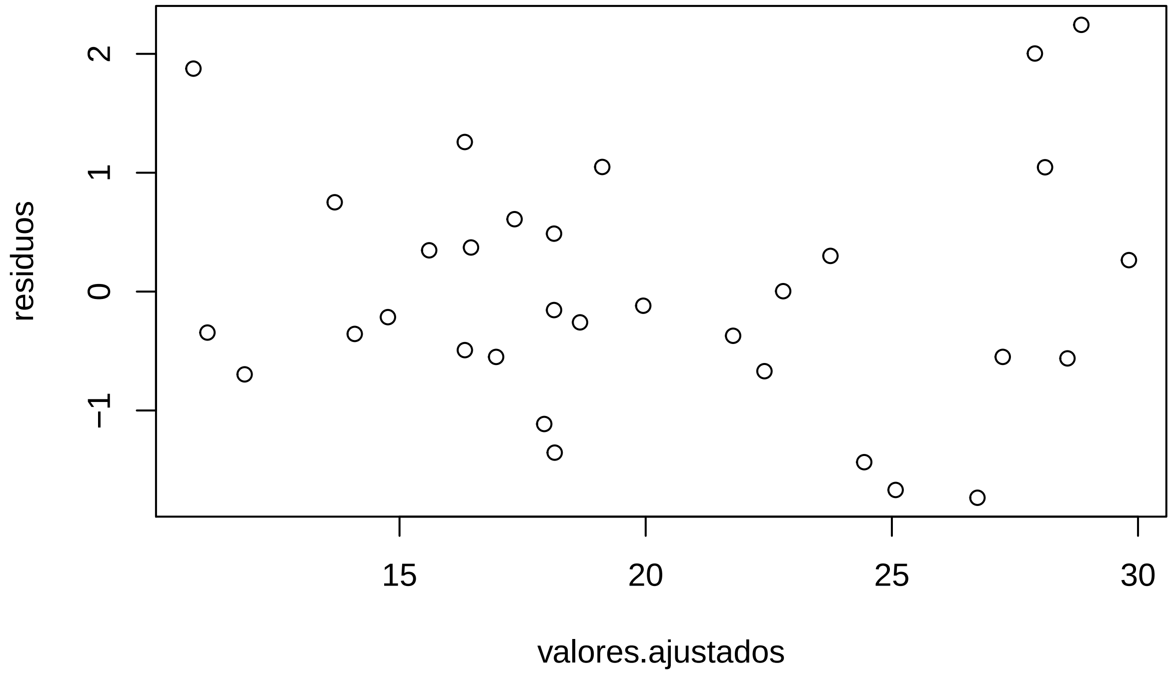
Test heteroskedasticity, Null is constant variance. Tested whether the disturbances variance is not constant throughout the observations.

```r
library(car)

ncvTest(fit3)
```

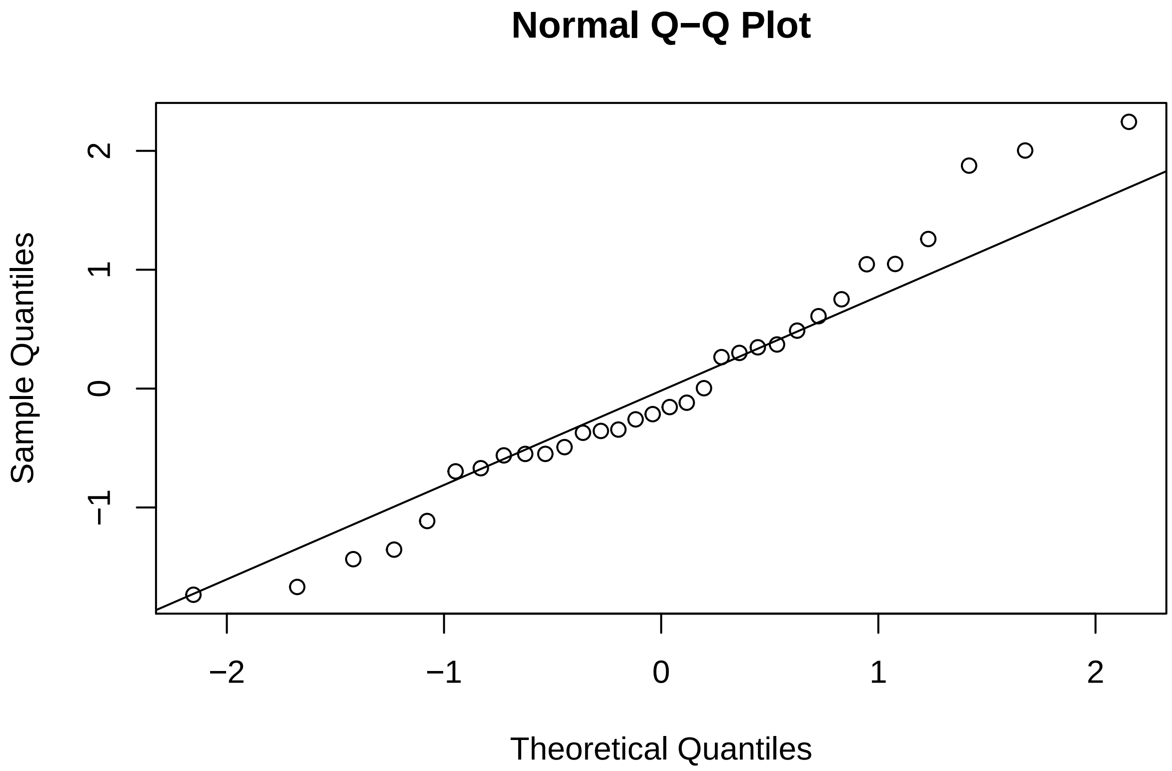
```
## Non-constant Variance Score Test 
## Variance formula: ~ fitted.values 
## Chisquare = 3.693268    Df = 1     p = 0.05463247
```

```r
residuos <- rstandard(fit3)
valores.ajustados <- fitted(fit3)
plot(valores.ajustados, residuos)
```



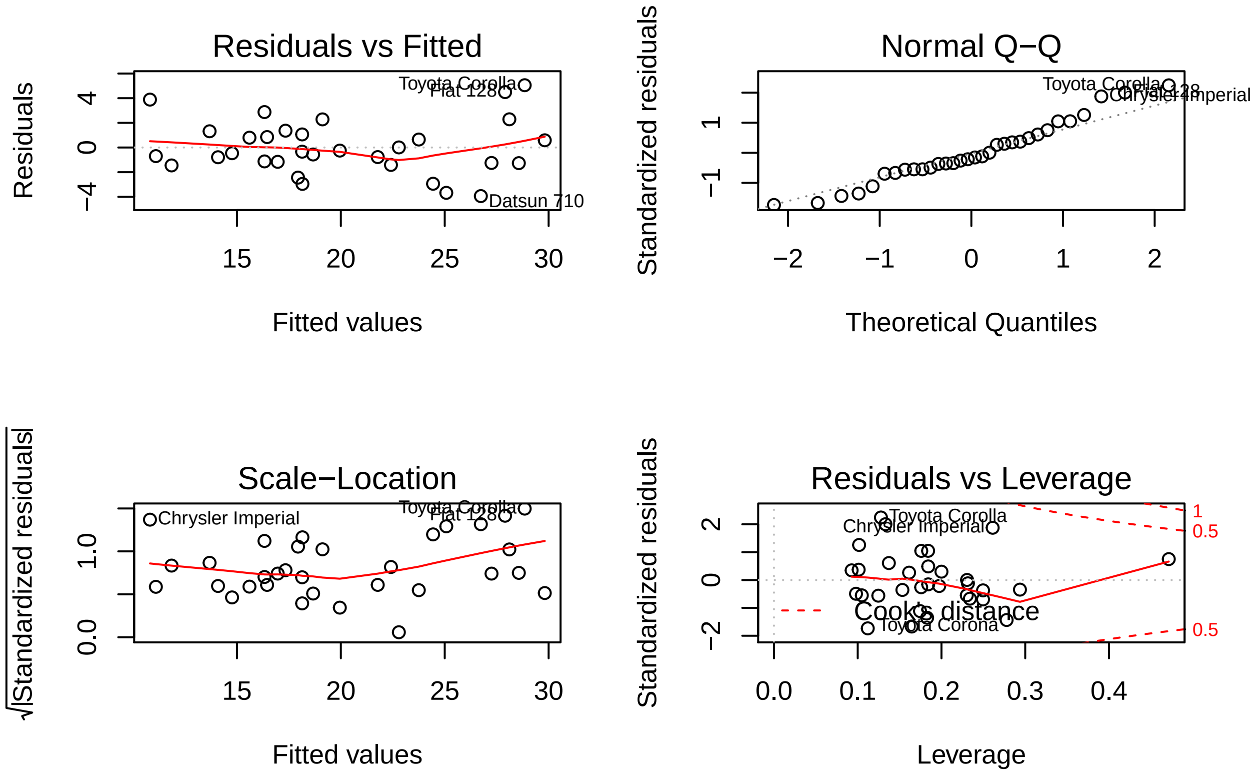
No special pattern is observed, so as homoscedasticity as both linearity are reasonable hypothesis.

```r
qqnorm(residuos)
qqline(residuos)
```



```
par(mfrow = c(2,2))

plot(fit3)
```



Conclusion.

The results from the multivariate regression reveal that, on average, manual transmission cars get 1.809 miles per gallon more than automatic transmission cars. The variables wt, cyl6 and hp influence in the variable mpg.render('test1.Rmd',pdf_document())

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.