Cars Course Project

LarionovaAnna

21 июня 2015 г.

Introduction

Looking at a data set of a collection of cars, we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). We will be using the mtcars dataset in R to do our analysis. We will be trying to answer the following questions:

*Is an automatic or manual transmission better for MPG

*Quantify the MPG difference between automatic and manual transmissions

Executive Summary

A simple model was first designed (mpg ~ factor(am)) to answer the question being asked. But since we found the R squared to be low and other variables in the data set having linear relationship with mpg we designed a multivariate regression model that ultimately increased the R squared value to 83.4%. The final verdict being manual transmission being better for MPG than automatic. 1.8 is the estimated expected increase in miles per gallon when compared manual transmission to automatic transmission.

Analysis

```
data("mtcars")
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.2.1

library(car)

## Warning: package 'car' was built under R version 3.2.1

head(mtcars)

##                    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
```

the am variable stores the data for transmission (0 = automatic, 1 = manual) creating boxplot to understand effect of transmission on mpg (see appendix)

```
g <- ggplot(mtcars, aes(x = factor(am), y = mpg, fill = factor(am))) +
  geom_boxplot() + ggtitle("Analyzing mpg ~ am data")
```

there is a positive effect on mpg in cars with manual transmission fit the simple linear model with am

```
fit <- lm(mpg ~ as.factor(am), data = mtcars)
coef(summary(fit))
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    17.147368   1.124603 15.247492 1.133983e-15
## as.factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

coefficient of manual transmission have value 7.24 as expected increase in miles per gallon when compared manual transmission to automatic transmission. So our basic model answers the first question positively. But if we look at the Adjuster R-squared value we will see that it is pretty low (34%), and the explanation of variance low as well. Plot2 in apendix shows that there are more variables that have linear relationship to mpg. Let's try to see which variable has big variance inflation factor (vif). First we add all variables to new model.

```
fit_all <- lm(mpg ~ ., data = mtcars)
sqrt(vif(fit_all))
```

```
##      cyl     disp       hp     drat       wt     qsec       vs       am
## 3.920948 4.649757 3.135608 1.837014 3.894212 2.743712 2.228424 2.156035
##     gear     carb
## 2.314617 2.812249
```

Here we find that am variable actually is not that very significant, however we leave it be in our model together with weight (wt), disp, cyl, hp.

```
fit2 <- lm(mpg ~ factor(cyl) + disp + hp + wt + factor(am), data = mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ factor(cyl) + disp + hp + wt + factor(am),
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.864276   2.695416  12.564 2.67e-12 ***
## factor(cyl)6 -3.136067   1.469090  -2.135   0.0428 *
## factor(cyl)8 -2.717781   2.898149  -0.938   0.3573
## disp          0.004088   0.012767   0.320   0.7515
## hp           -0.032480   0.013983  -2.323   0.0286 *
## wt           -2.738695   1.175978  -2.329   0.0282 *
## factor(am)1   1.806099   1.421079   1.271   0.2155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344 
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

Now adjuster R-squared is 83%, the model do better than the previous one after we added more variables in it. We can use anova to see whether our multivariable model (fit2) is better than the simple model (fit)

```
anova(fit, fit2)
```

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ as.factor(am)
## Model 2: mpg ~ factor(cyl) + disp + hp + wt + factor(am)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
## 1     30 720.90                                  
## 2     25 150.41  5    570.49 18.965 8.637e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

low p-value does signify that the final model is a good fit.We run a diagnostic to be doubly sure

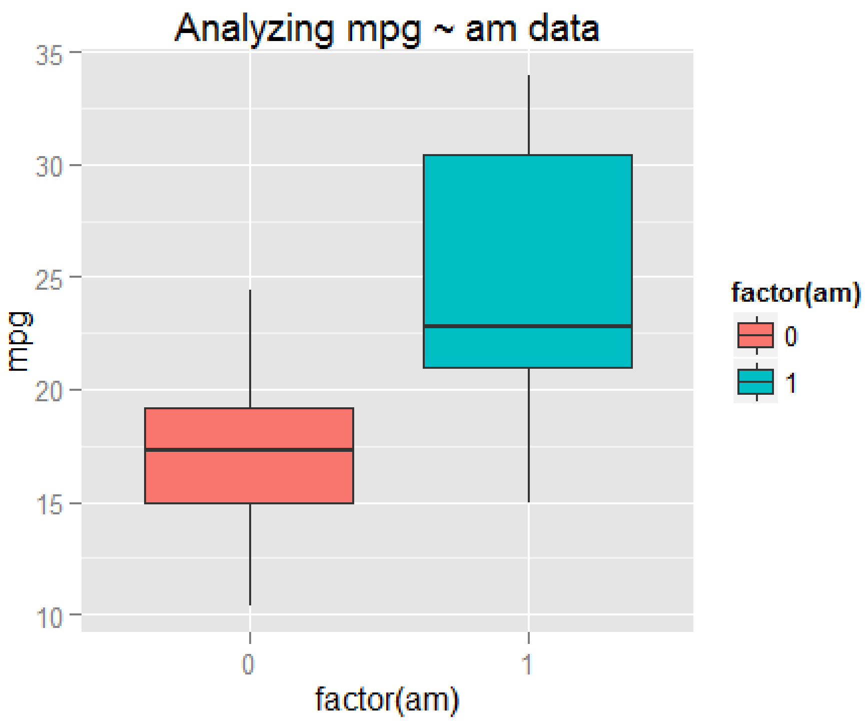
```
influence.measures(fit2)
```

```
## Influence measures of
##   lm(formula = mpg ~ factor(cyl) + disp + hp + wt + factor(am),      data = 
mtcars) :
## 
##                       dfb.1_  dfb.f..6  dfb.f..8 dfb.disp   dfb.hp
## Mazda RX4            0.03301 -0.268733 -0.108345 -0.02678  0.13847
## Mazda RX4 Wag        0.05482 -0.139273 -0.068909  0.01716  0.08146
## Datsun 710           0.04969  0.278493  0.129741  0.00306  0.04694
## Hornet 4 Drive       0.22579  0.156012 -0.301082  0.43636  0.01697
## Hornet Sportabout    0.15690  0.029907  0.030697  0.13202 -0.06234
## Valiant             -0.02214 -0.057548  0.034377 -0.03966  0.00719
## Duster 360          -0.06462  0.026131  0.043519 -0.04543 -0.08417
## Merc 240D            0.01724 -0.084385 -0.052359 -0.01705 -0.00140
## Merc 230             0.00220 -0.011501 -0.009104 -0.00318  0.00724
## Merc 280             0.01254  0.128074  0.009872 -0.11099  0.06612
## Merc 280C           -0.00204 -0.020868 -0.001608  0.01808 -0.01077
## Merc 450SE          -0.02799  0.079733  0.228068 -0.24916 -0.05948
## Merc 450SL           0.04344  0.073293  0.180948 -0.16534 -0.03651
## Merc 450SLC         -0.02703 -0.059078 -0.149167  0.14062  0.03152
## Cadillac Fleetwood   0.28713  0.152155  0.160390 -0.22566  0.13907
## Lincoln Continental  0.17194  0.076188  0.063873 -0.05268  0.04516
## Chrysler Imperial   -0.82503 -0.386979 -0.299315  0.06433  0.00496
## Fiat 128            -0.01916 -0.139366  0.210032 -0.24375 -0.36287
## Honda Civic          0.05274 -0.005569  0.018135  0.01702 -0.04632
## Toyota Corolla       0.31199 -0.115628  0.187992 -0.08076 -0.32499
## Toyota Corona       -0.49874  0.466668  0.449917 -0.03201 -0.46343
## Dodge Challenger    -0.24034 -0.177972 -0.324375  0.00796  0.28697
## AMC Javelin         -0.32179 -0.243668 -0.446732  0.07884  0.33155
## Camaro Z28          -0.04038  0.030010  0.031003 -0.00630 -0.09930
## Pontiac Firebird     0.20835 -0.000416 -0.030949  0.41576 -0.21534
## Fiat X1-9           -0.05193  0.032647 -0.032690  0.01221  0.07571
## Porsche 914-2       -0.03201  0.095840  0.075453 -0.07297  0.01566
## Lotus Europa         0.28728 -0.179079 -0.207120  0.18807  0.18436
## Ford Pantera L       0.00484  0.005010  0.045183 -0.15943 -0.10155
## Ferrari Dino         0.00965 -0.020165 -0.000776  0.00994 -0.01505
## Maserati Bora       -0.25148 -0.113120 -0.090201 -0.24735  0.64667
## Volvo 142E           0.34145  0.367371  0.188210  0.15301 -0.03583
##                       dfb.wt dfb.f..1   dffit cov.r   cook.d   hat inf
## Mazda RX4           -0.00419 -0.19523 -0.3681 1.532 1.98e-02 0.235
## Mazda RX4 Wag       -0.05071 -0.12446 -0.1990 1.718 5.86e-03 0.251
## Datsun 710          -0.12953 -0.25563 -0.6304 0.632 5.23e-02 0.112
## Hornet 4 Drive      -0.36272 -0.23119  0.6722 1.479 6.46e-02 0.318
## Hornet Sportabout   -0.17199 -0.06364  0.2544 1.511 9.52e-03 0.188
## Valiant              0.02329  0.04062 -0.1428 1.603 3.02e-03 0.190
## Duster 360           0.09134  0.08407 -0.1405 1.672 2.93e-03 0.220
## Merc 240D            0.04203 -0.05626  0.1549 1.623 3.56e-03 0.202
## Merc 230             0.00333 -0.01016  0.0185 1.745 5.12e-05 0.238
## Merc 280             0.05674 -0.11005  0.2871 1.558 1.21e-02 0.217
## Merc 280C           -0.00924  0.01793 -0.0468 1.695 3.26e-04 0.217
## Merc 450SE           0.17347  0.01195  0.3104 1.663 1.42e-02 0.261
## Merc 450SL           0.06628 -0.03013  0.2388 1.542 8.41e-03 0.194
## Merc 450SLC         -0.06418  0.01954 -0.1966 1.592 5.71e-03 0.201
## Cadillac Fleetwood  -0.16086 -0.19289 -0.5279 1.596 4.04e-02 0.305
## Lincoln Continental -0.12832 -0.09676 -0.2528 1.841 9.45e-03 0.307   *
## Chrysler Imperial    0.66151  0.34213  1.1503 0.656 1.70e-01 0.262
## Fiat 128             0.32200  0.43838  0.8858 0.449 9.78e-02 0.144
## Honda Civic         -0.02999  0.02518  0.1039 1.570 1.60e-03 0.166
## Toyota Corolla      -0.02743  0.28603  0.9313 0.330 1.04e-01 0.129
## Toyota Corona        0.29639  0.72186 -0.8989 1.026 1.11e-01 0.278
## Dodge Challenger     0.10563  0.02089 -0.5041 1.145 3.60e-02 0.174
## AMC Javelin          0.12051  0.05270 -0.6375 0.990 5.63e-02 0.186
## Camaro Z28           0.05407  0.08389 -0.1489 1.517 3.28e-03 0.154
## Pontiac Firebird    -0.31351 -0.02722  0.5980 1.096 5.02e-02 0.197
## Fiat X1-9           -0.00568 -0.07416 -0.2031 1.399 6.06e-03 0.125
## Porsche 914-2        0.03473 -0.06215 -0.2119 1.372 6.59e-03 0.119
## Lotus Europa        -0.32720 -0.11038  0.4998 1.269 3.57e-02 0.205
## Ford Pantera L       0.16276 -0.11530 -0.3896 1.640 2.22e-02 0.276
## Ferrari Dino        -0.00485 -0.00864 -0.0412 1.762 2.53e-04 0.246
## Maserati Bora        0.07802  0.05121  0.8772 2.209 1.11e-01 0.512   *
## Volvo 142E          -0.42089 -0.35143 -0.7616 0.741 7.73e-02 0.171
```

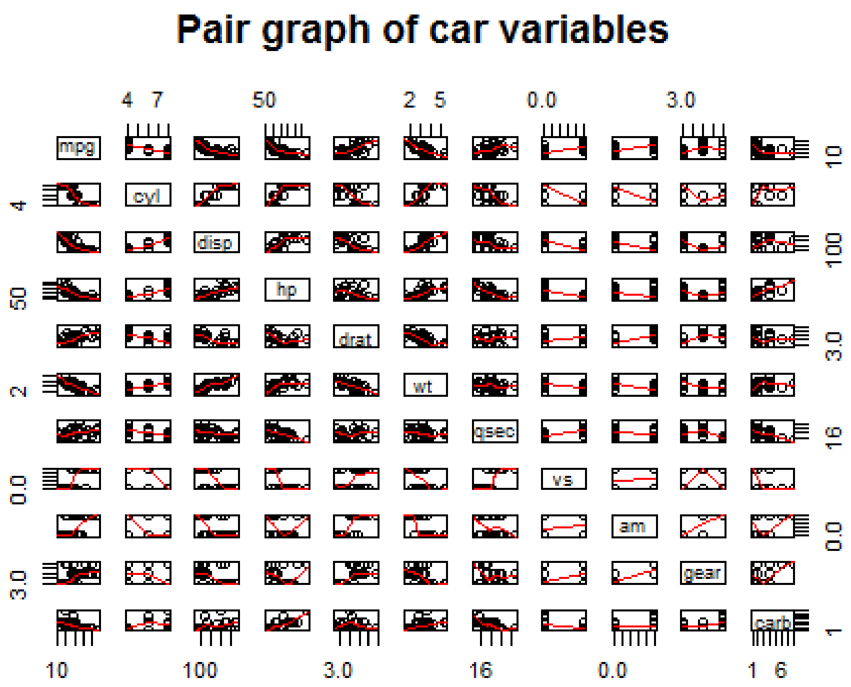
Residual plots are in Appendix. A pattern less residual plot signifies a good fit.

Appendix

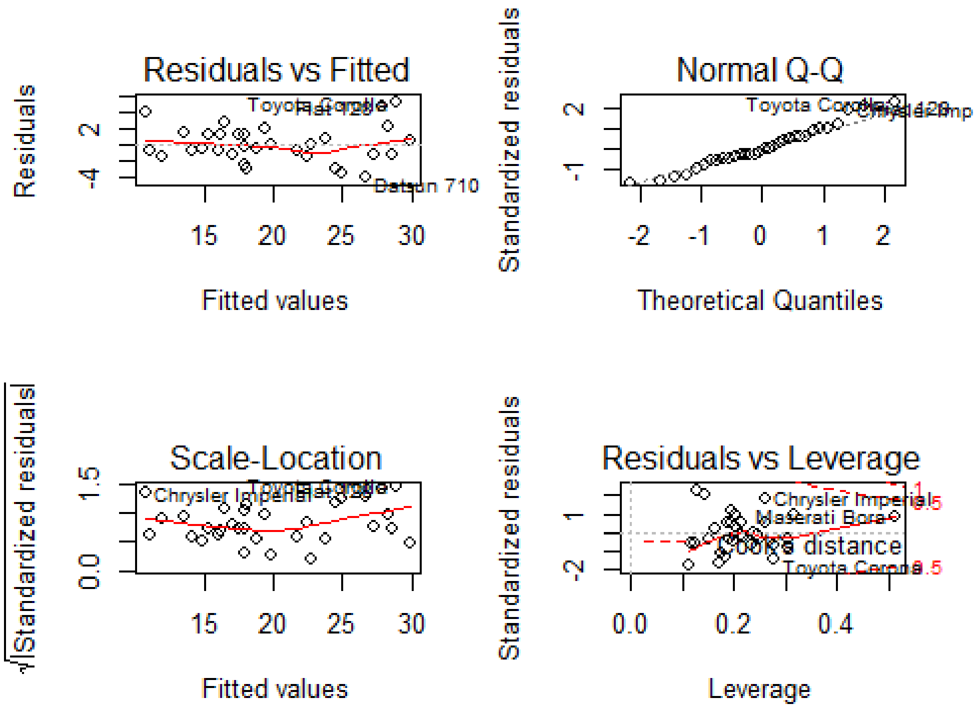
```
g
```



```r
pairs(mtcars, panel = panel.smooth, main = "Pair graph of car variables")
```



```r
par(mfrow = c(2,2))
plot(fit2)
```


Residuals vs Fitted
Residuals
2
-4
15
20
25
30
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q-Q
Standardized residuals
2
-1
-2
-1
0
1
2
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imp
Scale-Location
√|Standardized residuals|
1.5
0.0
15
20
25
30
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
1
-2
0.0
0.2
0.4
Leverage
Chrysler Imperial
Maserati Bora
Toyota Corona
Cook's distance
1
0.5
0.5