RegressionModelsProject

2024-07-01

Executive Summary

in this analysis we try to answer the following questions in mtcars data set:
-Is an automatic or manual transmission better for MPG?

-Quantify the MPG difference between automatic and manual transmissions
I used model selection skills to fit the best model and also used diagnostics tools to analyse my results.

Explotory Data Analysis of mtcars dataset

A data frame with 32 observations on 11 (numeric) variables [, 1] mpg Miles/(US) gallon
[, 2] cyl Number of cylinders
[, 3] disp Displacement (cu.in.)
[, 4] hp Gross horsepower
[, 5] drat Rear axle ratio
[, 6] wt Weight (1000 lbs)
[, 7] qsec 1/4 mile time
[, 8] vs Engine (0 = V-shaped, 1 = straight)
[, 9] am Transmission (0 = automatic, 1 = manual)
[,10] gear Number of forward gears
[,11] carb Number of carburetors

```
 carsdf <- data.frame(mtcars)
head(carsdf)
```

```
##                   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
```

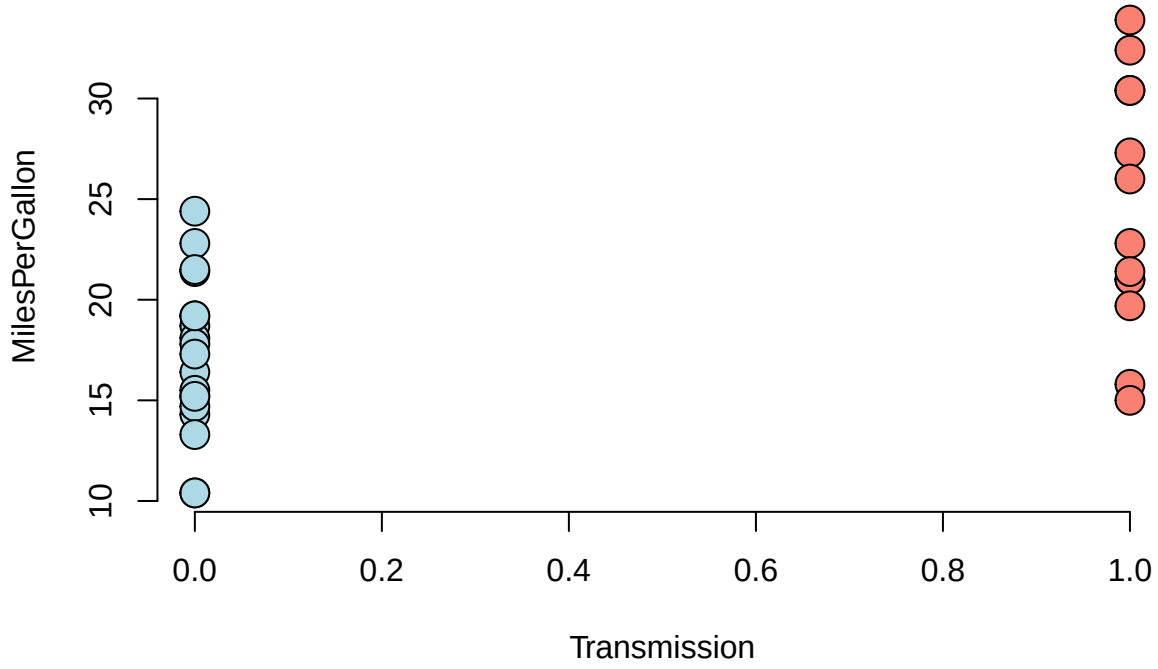
```
summary(carsdf)
```

```
##       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
```

```
y<-carsdf$mpg
x<-carsdf$am
idx<-x==0

plot(x, y,xlab="Transmission",ylab="MilesPerGallon" ,type = "n", frame = FALSE)
points(x[idx], y[idx], pch = 21, col = "black", bg = "lightblue", cex = 2)
points(x[!idx], y[!idx], pch = 21, col = "black", bg = "salmon", cex = 2)
```



model selection comparing variance inflation factor for all variables, we can see that number of cylinders, displacement and weight which have higher standard deviations are highly related to each other. following a anova test shows that including weight and cylinder is necessary.

```r
#install.packages("car")
library(car)
```

```
## Loading required package: carData
```

```r
fit <- lm(y ~ .,data=carsdf)
sqrt(vif(fit))
```

```
## Warning in summary.lm(object, ...): essentially perfect fit: summary may be
## unreliable
```

```
##      mpg      cyl     disp       hp     drat       wt     qsec       vs
## 2.763061 3.922010 4.711089 3.207489 1.847118 4.235849 2.824955 2.229633
##       am     gear     carb
## 2.231788 2.325210 2.816124
```

```r
fitbase<- lm(y~factor(x))

fitcylwt<-lm(y~factor(x)+factor(carsdf$cyl)+carsdf$wt)

anova(fitbase,fitcylwt)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ factor(x)
## Model 2: y ~ factor(x) + factor(carsdf$cyl) + carsdf$wt
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     30 720.90
## 2     27 182.97  3    537.93 26.46 3.401e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear Regression

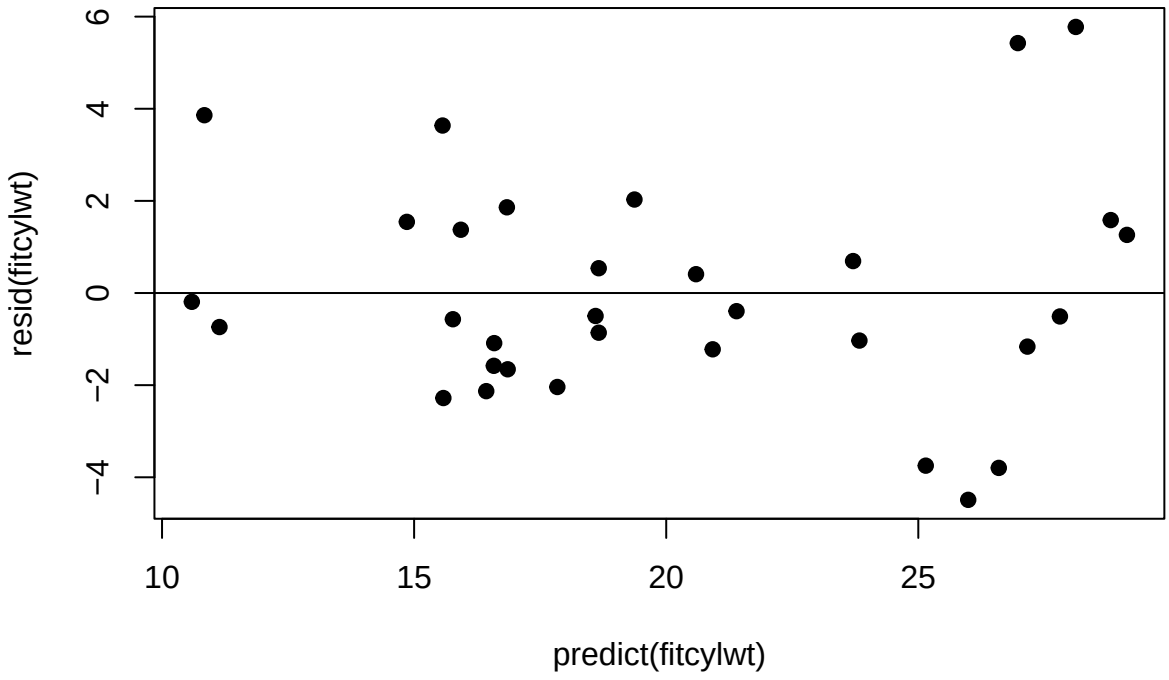
the linear regression shows that changing transmission type from automatic to manual (including wt and cyl) increases 0.15 miles per gallon.

```r
summary(fitcylwt)$coef
```

```
##                       Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)         33.7535920  2.8134831 11.9970836 2.495549e-12
## factor(x)1           0.1501031  1.3002231  0.1154441 9.089474e-01
## factor(carsdf$cyl)6 -4.2573185  1.4112394 -3.0167231 5.514697e-03
## factor(carsdf$cyl)8 -6.0791189  1.6837131 -3.6105432 1.227964e-03
## carsdf$wt           -3.1495978  0.9080495 -3.4685309 1.770987e-03
```

##plot residuals no specific pattern is shown

```r
plot(predict(fitcylwt),resid(fitcylwt),pch=19)
abline(h=0)
```



diagnostics and homoscedasticity

no data entry errors has seen based on hatvalues function. there is not specific outlier according to the dffits function.
the plot also shows the model fits the assumptions of homoscedasticity.
-seeing a bias in the residuals would indicate a bias in the error
-red lines representing the mean of the residuals are all basically horizontal and centered around zero so no outliers or bias
-QQ one-to-one line

```r
hatvalues(fitcylwt)
```

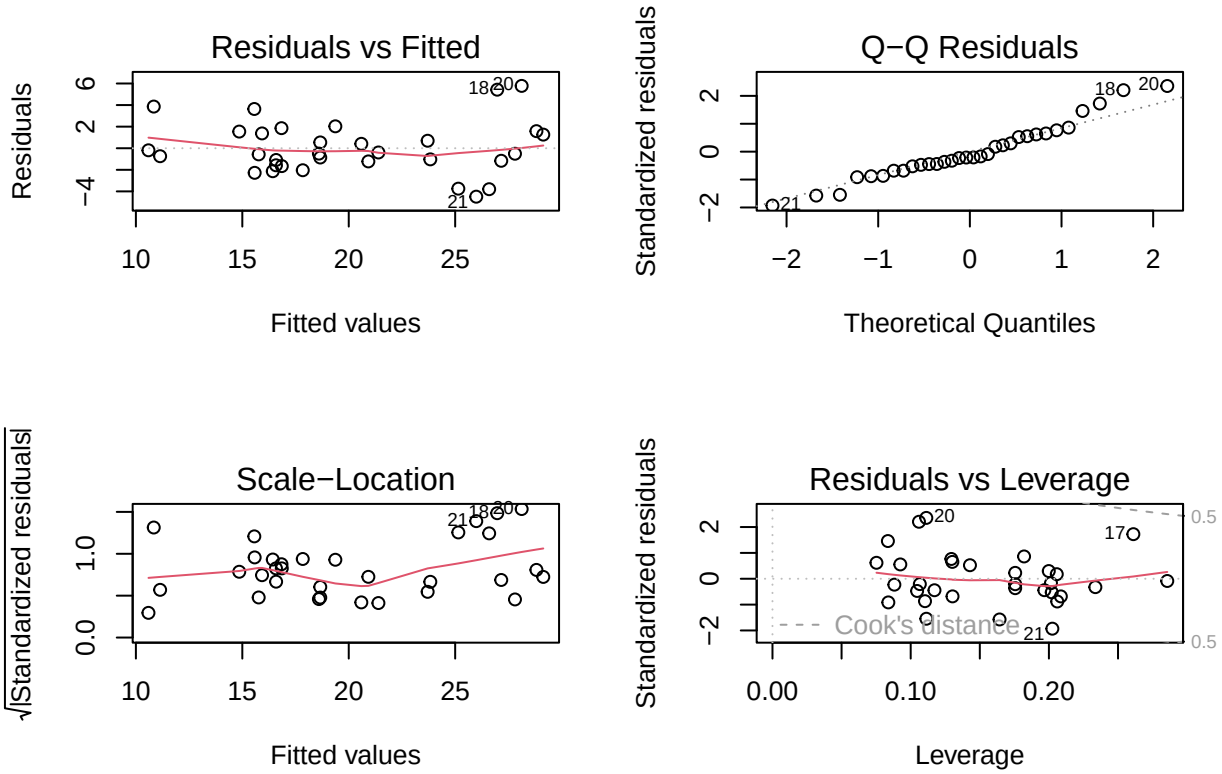
```
##          1          2          3          4          5          6          7
## 0.20149516 0.20568847 0.11134826 0.18203504 0.12944546 0.17562241 0.11035260
##          8          9         10         11         12         13         14
## 0.19990502 0.19671403 0.17559835 0.17559835 0.07524667 0.09249950 0.08819801
##         15         16         17         18         19         20         21
## 0.23360825 0.28562638 0.26109577 0.10600585 0.13014404 0.11129580 0.20249595
##         22         23         24         25         26         27         28
## 0.11720930 0.13026193 0.08383928 0.08351560 0.10662207 0.10464875 0.14287912
##         29         30         31         32
```

```
## 0.20603900 0.20204540 0.20862936 0.16429081
```

```
dffits(fitcylwt)
```

```
##           1           2           3           4           5           6
## -0.08362867  0.08802080 -0.56294348  0.40469780  0.29299594 -0.09563785
##           7           8           9          10          11          12
## -0.30414152  0.14635533 -0.21565388  0.10325157 -0.16551067  0.17392266
##          13          14          15          16          17          18
##  0.17451788 -0.06992890 -0.17604909 -0.05388787  1.06676664  0.82251269
##          19          20          21          22          23          24
##  0.24943890  0.91675767 -1.02869925 -0.15962595 -0.26124531 -0.27594963
##          25          26          27          28          29          30
##  0.45027602 -0.07021666 -0.15913348  0.21081130 -0.44616239 -0.26082265
##          31          32
## -0.34688319 -0.71903764
```

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



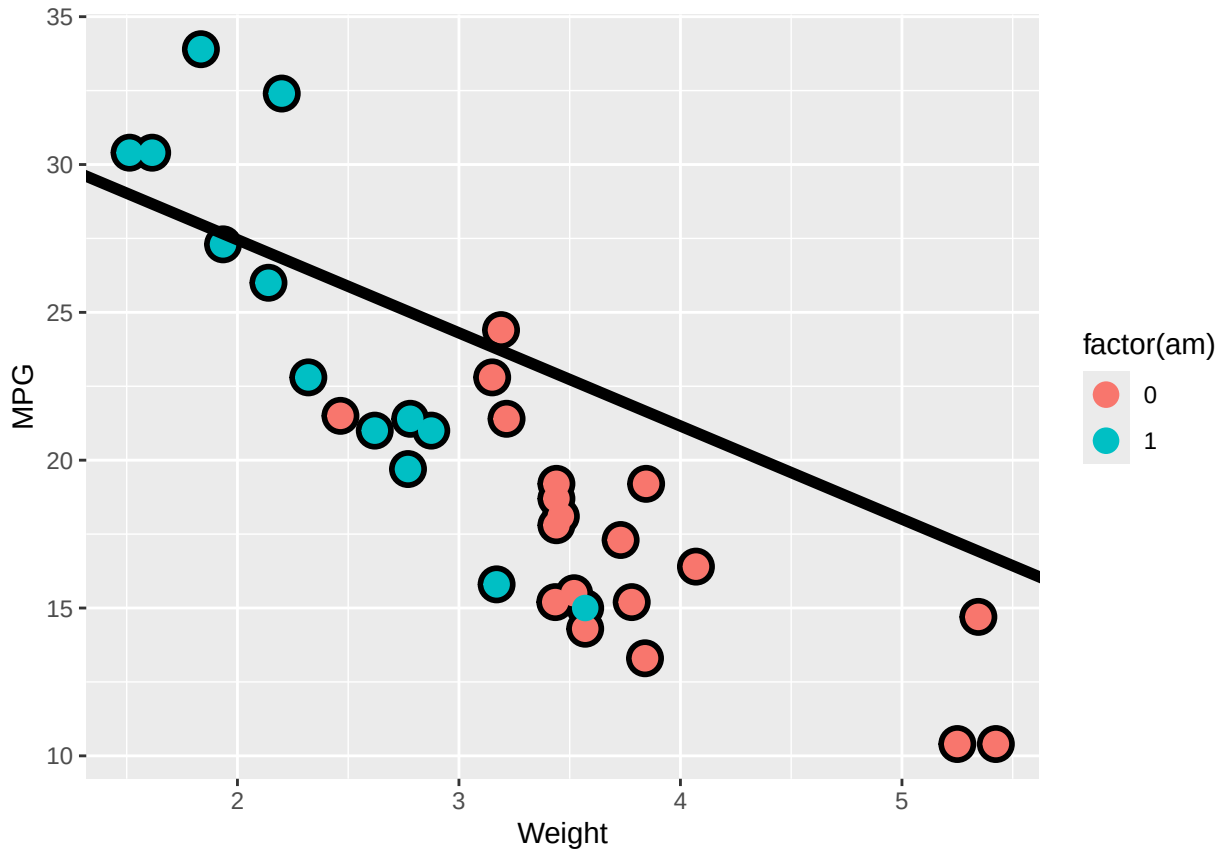
visualise the result

in this diagram I try to visualize the fitted line and its relation to cars weight and transmission type. as we can see manual type(1) contributes to higher mpg and it is also obvious that as weight increases the outcome decreases.

```
library(ggplot2)
coeffit<-summary(fitcylwt)$coef[,1]
g = ggplot(carsdf, aes(x = wt, y = mpg, colour = factor(am)))
g = g + geom_point(size = 6, colour = "black") + geom_point(size = 4)
g = g + xlab("Weight") + ylab("MPG")
g1=g
coeffit
```

```
##          (Intercept)          factor(x)1 factor(carsdf$cyl)6 factor(carsdf$cyl)8
##           33.7535920           0.1501031          -4.2573185          -6.0791189
##            carsdf$wt
##           -3.1495978
```

```
g1=g1+geom_abline(intercept=coeffit[1],slope=coeffit[5],linewidth=2)
g1
```



conclusion

for conclusion we cay tell that although there is a positive relationship between type of transmission and mpg, in other words type manual produces higher miles per gallon, the other variables such as weight and cylinders have negative effect and decrease this slope. and it is also notable that cars with higher weight in this data set appears to have automatic transmission mainly.