Mtcars Regression Analysis

Herve Yu

Monday, January 19, 2015

Analysis variables over miles per gallon, compare automatic versus manual transmission

Executive Summary:

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). The overall MPG gain by manual transmission over automatic transmission is substantial around 30%.

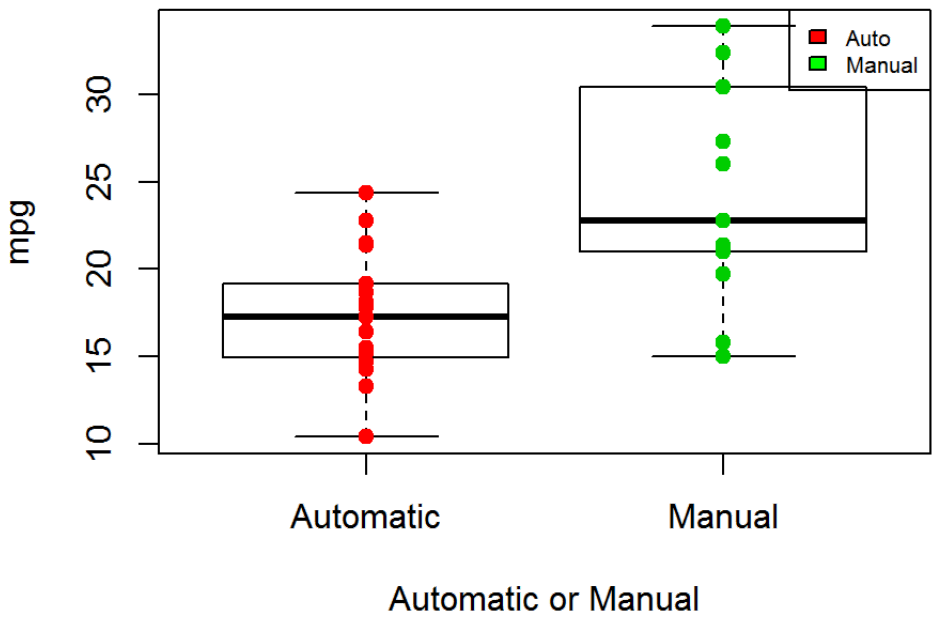
The linear model with MPG as outcome and transamission (Automatic or Manual), Weight, Quarter mile time as regressors gives an acceptable confidence and predictive intervals.

- For automatic transmission: MPG = 9.62 - 3.92 * Weight + 1.22 * Quarter miles time
- For manual transmission: MPG = 12.56 - 3.92 * Weight + 1.22 * Quarter miles time
- The model predict an actual MPG with a variation of + or - 4.9 MPG from the predicted MPG.
- Units of measure: Weight in 1000 lbs, Quarter Mile time in seconds.

Note: The data provided are insufficient to conform to the Central Limit Theorem, and the data also introduces biasness. For example 3 gears cars only include manual whereas 5 gears only include automatic.

Simple plot MPG over transmission type

```
# Load data
data(mtcars); mc <- mtcars
mpg <- mc$mpg; wt <- mc$wt; hp <- mc$hp; cyl <- mc$cyl; disp <- mc$disp
drat <- mc$drat; qsec <- mc$qsec; am <- mc$am; gear <- mc$gear; vs <- mc$vs
carb <- mc$carb
# Simple Plot of MPG over Automatic or Manual
par(mfrow=c(1,1))
f <- function(am){if (am==0) "Automatic" else "Manual"}
amn <- sapply(am,f)
plot(mpg~factor(amn),pch=19, xlab="Automatic or Manual")
points(mpg~factor(amn), pch=19,col=((am==1)*1+2))
legend("topright", legend = c("Auto","Manual"),fill =c("red","green"), ncol = 1, cex = 0.65); fit1 <-
 lm(mpg~factor(am)); summary(fit1)$coefficients
```




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

The plot mpg over type transmission (Automatic or Manual) shows that Manual transmission return better MPG than Automatic transmission. Manual transmission return **7.2449393** mpg better than automatic transmission.

Exploratory data analysis

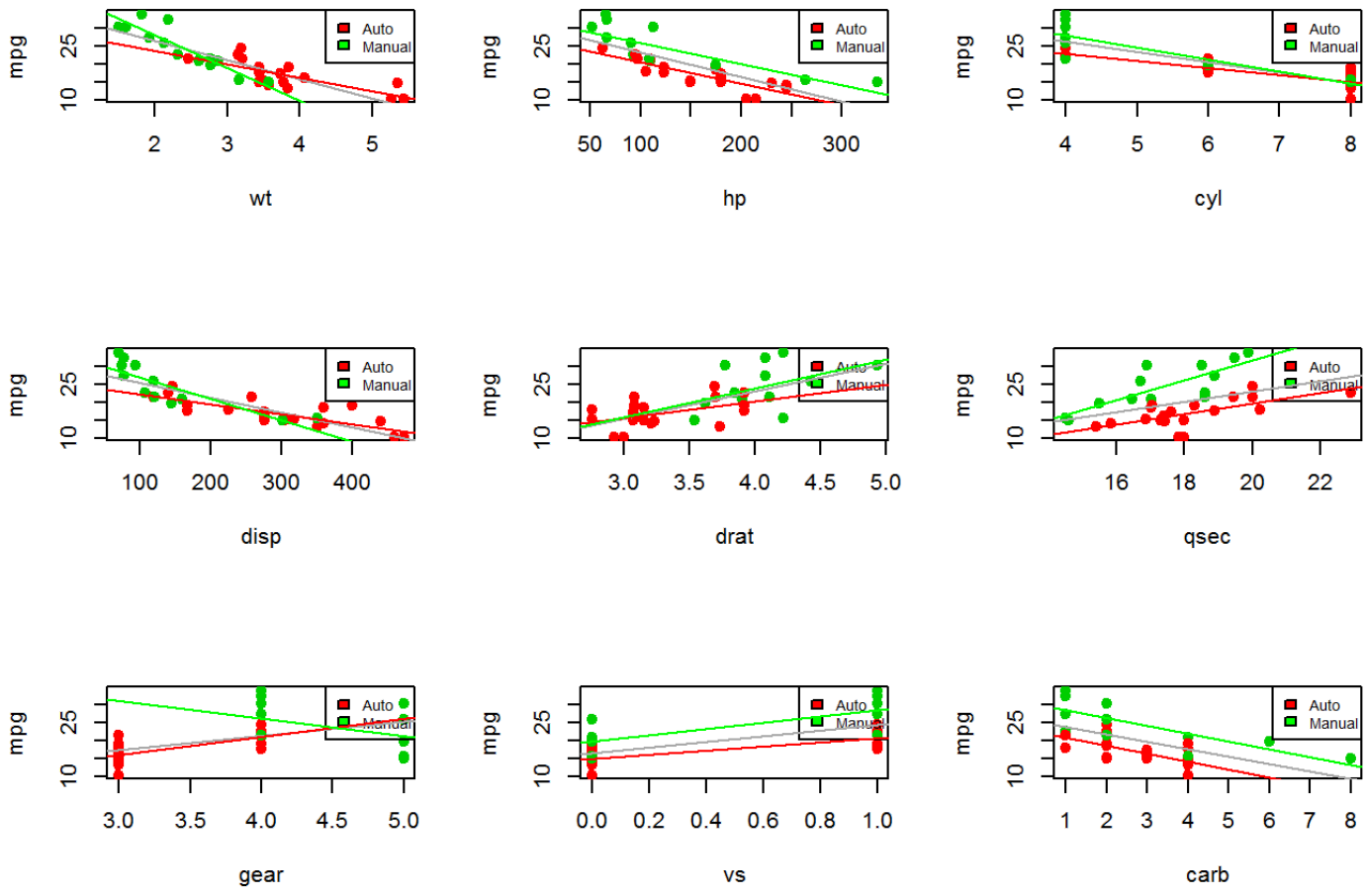
Let us find the linear regression model for the mpg over factor of transmission and other confounders.

- mpg : miles per gallon US
- wt : weight in 1000 of lbs
- hp : Gross horse power
- cyl : Number of Cylinders
- disp : Displacement (cu.in.)
- drat : Rear axle ratio
- qsec : 1/4 mile time (in seconds)
- am : Transmission (0 = automatic, 1 = manual)
- gear : Number of forward gears
- vs : 0 V Engine /1 Straight Engine
- carb : Number of carburetors

Let us individually the MPG over confounders(wt,hp,cyl,disp,drat,gear,vs,carb), then color the transamissions data points and linear model

- automatic in red
- manual in green
- add the regressin linear for all data points in dark grey

```
#plotting function using base plot
mplot <- function(vart){ if     (vart=="cyl") var <- cyl
else if  (vart=="hp")  var <- hp
else if  (vart=="wt")  var <- wt
else if  (vart=="disp")  var <- disp
else if  (vart=="drat")  var <- drat
else if  (vart=="qsec")  var <- qsec
else if  (vart=="gear")  var <- gear
else if  (vart=="carb")  var <- carb
else if  (vart=="vs")  var <- vs
plot(mpg~var,pch=19, xlab=vart)
# am 0 automatic col 2 (red) and manual col 3 (green)
points(mpg~var, pch=19,col=((am==1)*1+2))
# add small legend
legend("topright", legend = c("Auto","Manual"),  fill =c("red","green"), ncol = 1,cex = 0.65)
lmvar  <- lm(mpg~var); lmvarA <- lm(mpg[am==0]~var[am==0]);lmvarM <- lm(mpg[am==1]~var[am==1])
abline(lmvar,col="darkgrey",lwd=1); abline(lmvarA,col="red");abline(lmvarM,col="green")}
par(mfrow=c(3,3));mplot("wt");mplot("hp");mplot("cyl");mplot("disp");mplot("drat");mplot("qsec");mplo
t("gear");mplot("vs");mplot("carb")
```



Discussion:

- The plots: hp, carb, cyl, qsec, drat, vs show the manual transmission (green datapoints) returning better mpg than automatic (red datapoints)
- The regression lines of manual transmission (green) also shows better mpg than the automatic transmission (red line)
- The wt plot shows below 3000 lbs manual transmission has better mpg, beyond 3000 lbs the automatic transmission start to return better MPG. But we can see very few manual vehicles have higher than 3000 lbs of weight and beyond 3600 lbs, only automatic vehicles data have been collected. Whereas below 3000 lbs there is only 1 automatic vehicle.
- The disp plot shows below 300 cu.in manual transmission has better mpg, beyond 300 cu.in the automatic transmission start to return better MPG.
- The gear plot shows at 3 gears only automatic vehicles data have been collected, and at 5 gears only manual data collected. At 4 gears there is a mix between manual and automatic vehicles and again manual cars return better MPG than automatic cars.

From this serie of plots, all variables give significant impacts on the outcome MPG analysis between manual and automatic transmissions. One critic though, the data is not enough distributed across the range for both types of transmission, this might introduce biasness. For example 3 gears cars only include manual whereas 5 gears only include automatic.

Since all variables show impact on mpg between automatic and manual transmission, we will start by including all variables in the analysis.

Regresson Model with all variables inclusive

```
fit <- lm(mpg~factor(am)+wt+hp+cyl+disp+drat+qsec+carb+vs)
sf <- summary(fit); sf$coefficients
```

```
##                Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) 15.64180646 16.78606059  0.93183307 0.36153694
## factor(am)1  2.80344698  1.91663440  1.46269261 0.15769184
## wt          -3.86609068  1.82849991 -2.11435104 0.04605085
## hp          -0.02062744  0.02127976 -0.96934533 0.34290994
## cyl         -0.27315026  0.95981077 -0.28458762 0.77862172
## disp         0.01395085  0.01747263  0.79844026 0.43315023
## drat         0.84088668  1.60057435  0.52536559 0.60458429
## qsec         0.79507208  0.71495481  1.11205920 0.27811774
## carb        -0.04506069  0.73653480 -0.06117932 0.95176880
## vs           0.35800211  2.06357126  0.17348667 0.86385526
```

Only weight has a pValue 0.0460509 lower than .05. In this current model, only wt variable has sufficient significance according T-statistics for inference. The other variables'slopes beta coefficients in 1st column of the coefficients table can be considered as flat ie 0 for inferential regression.

But 1 confounder is added to factor(am) at a time, those confounders all have significant impact to the mpg outcome. Too many variables decrease significance of regression model.

For example:

```
fitdrat <- lm(mpg~factor(am)+drat)
sfdrat <- summary(fitdrat); sfdrat$coefficients
```

```
##              Estimate Std. Error    t value   Pr(>|t|)
## (Intercept) -1.949883   7.073285 -0.2756687 0.78475740
## factor(am)1  2.807061   2.282159  1.2300023 0.22858143
## drat         5.811143   2.129833  2.7284496 0.01069548
```

Coefficient slope for variable drat is changed from 0.84089 to 5.811. Mostly its significance increase, the pValue is changed from 60% to 1%.

We need to find a better model. Search in internet helps to point to AIC (Akaike Information Criterion), professor Akaike spent his study to finding the criterion to select the best fit in regression, by penalizing too many variables. The AIC criterion for model selection is implemented in R using step commend.

Regresson Model using Akaika Information Criterion

```
sfit <- step(fit,trace=0)
sfits <- summary(sfit); sfits
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt + qsec)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## factor(am)1   2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

The linear model selected is **mpg ~ factor(am) + wt + qsec**

Interpretation of model:

Estimate column:

- Intercept: 9.6177805 represents mpg for automatic transmission at 0 weight and 0 quarter miles time. Note this is a theoretical value, it does not exist in reality, since no vehicle would have 0 lbs of weight, but it does not matter, the model should be used within the real cases range.
- factor(am)1: 2.9358372 represents additional mpg from intercept at the same levels of regressors with manual transmission.
- wt: -3.9165037 represents the slope of weight on mpg, decrease of -3.92 mpg per 1000 lbs weight increase.
- qsec 1.225886 represents the slope of qsec on mpg, increase of 1.23 mpg per quarter time increase.

Confidence interval analysis:

- H0 null hypothesis of a variable is 0, that is no impact on the outcome.
- HA alternative hypothesis of a variable is different to 0, and has an impact on the outcome.
- The variable weight pValue 6.96e-06 (<.05), the H0 is rejected.
- The variable quarter mile time pValue .000216 (<.05), the H0 is rejected.
- Both variables have significant impacts on regression for mpg.

Correlation analysis:

The Adjusted R-squared of .8336 suggests a strong correlation between MPG and the variables:

- Weight
- Quarter Mile time
- Transmission

The transmission, the difference of the intercept tells us that using automatic as base, the manual transmission returns 2.9358372 MPG better than automatic transmission.

Adjusted Correlation between mpg (dependent variable) and independent variables (am,wt,qsec) Adjusted R-squared 0.8336 is high, strong relationship. 84% variability of mpg is explained by the regressors (am, wt, qsec).

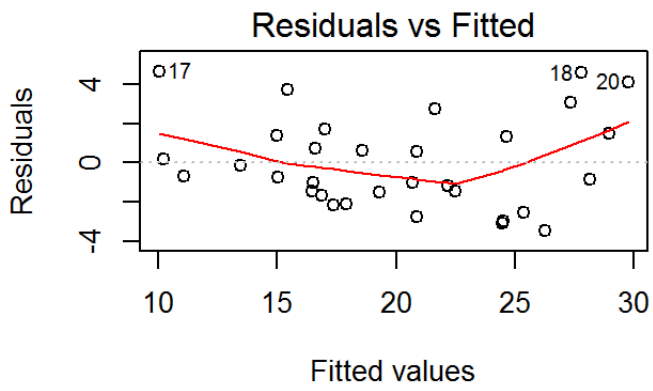
F-statitic pvalue 1.21e-11 suggests variation of mpg explained by the regressors more significant than the variation within the regressors. The number suggest a 95% confident interval the linear model reflects the change of mpg consumption.

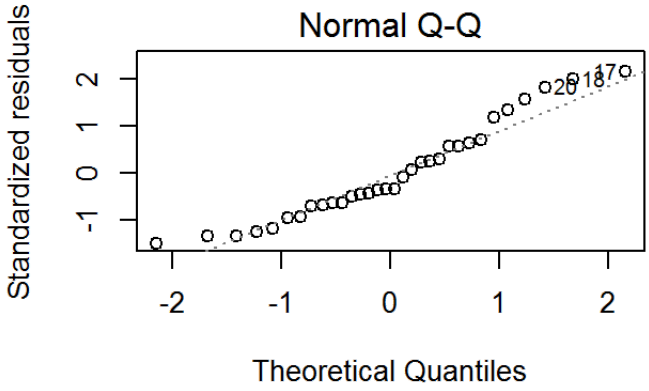
Residuals analysis:

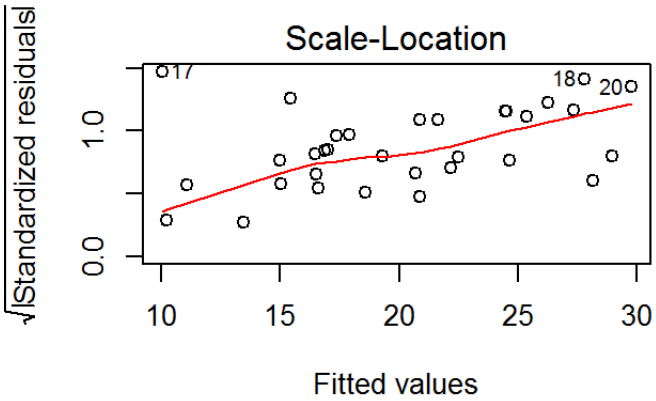
Residual standard error 2.4588465, coefficient of variation sigma/mean of outcome: 12.2387755%, “mpg hat”" compare to the “real mpg”" has a 12% variation. Based on the numbers, there is 95% confidence interval that for a given mpg output of combination of am, wt and qsec, the model provides an interval of [-4.917693, 4.917693] mpg around the real value.

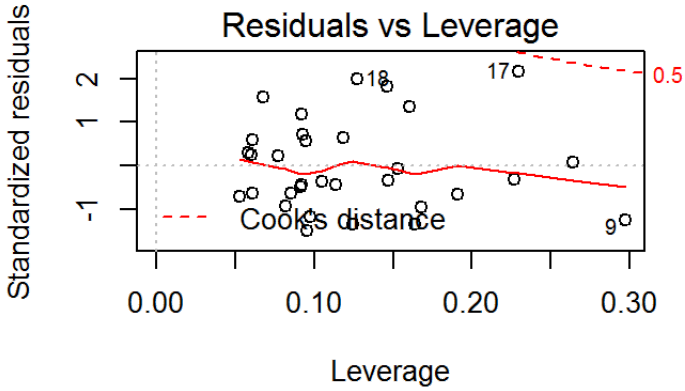
Residual Diagnostic plot:

```
par(mfrow=c(2,2)); plot(sfit)
```









Discussion:

- Residuals VS Fitted plot, the line is not perfectly around the value 0, suggests some pattern in residual versus the model, we also remember the data collected are limited and Central Limit Theorem tells only when size of the population grows with the convergence to population mean. Here we are limted data.
- Quantile-Quantile plot, suggest the residuals are approximately normally distributed.
- Scale-Location plot suggests some biasness.
- Residual VS leverage suggests some outliers may shift the model towards higher mpg, but overall coefficients change is not disturbed too much based on the cooks distance.

Conclusion:

Overall, the model looks fitted for confident and prediction analysis. It shows manual transmissions return higher MPG than automatic transmission. But the volume of data is low and there is some biasness in the data. For example, most of the heavy vehicles use automatic transmission versus most of the light vehicle are manually transamitted.