Multivariate Regression Models Project (Johns Hopkins Data Science Certification)

MD Alamgir, Ph.D.

December 4, 2016

Executive Summary

This multivariate linear regression data analysis project uses the “mtcars” dataset to address whether automatic or manual transimission results in higher mpg (miles per gallon), and to quantify the mpg difference between automatic and manual transmissions.The key steps include data loading, exploratory analysis to identify varaiables in addition to the transmission type(automatic or manual) that may have noticeable impact on the mpg. Four multiple linear regression models are fitted after identifying predictors that may have more significant effect on the outcome(mpg). Overall, cars with manual transmission have on average about 42% more mpg than automatic transmission cars. The fit3 linear model with outcome “mpg” and predictors “wt” and “qsec” controlled by transmission type “am” provides the highest R-squared value and explains nearly 90% of the variance in data. The coefficients of the regression models are discussed as to their relative importance. Plot of residuals indicate good normality (gaussian distribution).

1. Loading and Processing the Data

```
setwd("C:/Users/MD/Documents/Week4_Regression_Project")
library("ggplot2")
# Read mtcars data.
data(mtcars)
```

2. Exploratory Analysis

Appendix section A-1 contains structure of the mtcars dataset showing variables and data types, followed by check for missing data.

- Convert variable “am’ to factor type with levels”automatic and “manual”

```
mtcars$am<-as.factor(mtcars$am)
levels(mtcars$am)<- c("automatic", "manual")
```

- Explore correlation coefficients for preliminary insight: We see that the mpg has negative correlation with weight(wt), and positive correlation with rear axle ratio drat, and the acceleration capability (qsec).

```
cor_mpg_wt=cor(mtcars$mpg, mtcars$wt)
cor_mpg_drat=cor(mtcars$mpg, mtcars$drat)
cor_mpg_qsec=cor(mtcars$mpg, mtcars$qsec)
cor_coefs<- c(cor_mpg_wt, cor_mpg_drat, cor_mpg_qsec)
round(cor_coefs, 2)
```

```
## [1] -0.87  0.68  0.42
```

3. Linear Regression Modeling and Inferences

Exploratory analyses above, and in Appendix section A-1, have provided general insights on likely dependencies of mpg with other variables. Below, four linear regression models are fitted and key predictive statistics such as coefficiens, and R-squared values are obtained to pick the appropriate fit from this set. The significance of the coefficients for these models are discussed in Appendix Section A-2. High R-squared value is used as metric for assessing the fits. Residuals of the chosen fit (fit3) are briefly discussed via plots in Section A-3 of Appendix.

- “fitall” model: “mpg”" vs. all other remaining variables are fitted. The coefficients show that the variables, am(manual transmission), wt, drat, and qsec as having notiecable impact on mpg. Fitting all variables may lead to overfitting, and may not be efficient when used in an algorithm.
- “fit1” model: We use a step function analysis to narrow the predictor variable list suggested by the function
- “fit2” model: This linear model uses predictors “wt” and “drat” controlled by “am” as predictors of the outcome “mpg”.
- “fit3” model: This linear model uses predictors “wt” and “qsec” suggested by the “fit1” model, and additionally controlled by “am”.

```
fitall=lm(mpg ~., data=mtcars)
```

The fit0 model coefficients indicate that the mean mpg of manual transmission is about 7.2 mpg more than the mpg for automatic transmission of 17.15. However, the simplistic model fit0 has a an undesirable very low R-squared value (0.36), which means only 36% of the variance is explained by this fit.

```
fit0=lm(mpg~ factor(am), data=mtcars)
summary(fit0)$coef
```

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

```
fit1=step(lm(mpg~., data=mtcars), steps=3000, trace=0)
fit2=lm(mpg ~ factor(am):wt + factor(am):drat, data=mtcars)
fit3=lm(mpg ~ factor(am):wt + factor(am):qsec, data=mtcars)
```

R-squared values for the various models (fits)

```
rs<-c(summary(fitall)$r.squared, summary(fit0)$r.squared, summary(fit1)$r.squared, summary(fit2)$r
squared, summary(fit3)$r.squared)
round(rs,2)
```

```
## [1] 0.87 0.36 0.85 0.83 0.89
```

4. Results / Conclusions

- Manual transmission has a 42% more mpg when all other variables are held constant. However, there are other variables like wt and qsec that influence this dependency.
- The mpg differences (automatic vs. manual) are further quantfied via coefficients of the secondary variables, and the desirable high R-squared values (Section3, and Appendix Section A-2.)
- The “fit3” multivariate linear regression model is recommended, and further optimization may be performed with additional interaction terms in the model and detail study of the residuals.

Appendix

A-1: Exploratory Analysis Details

```
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  : Factor w/ 2 levels "automatic","manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

- variable definitions can be obtained using the command ?mtcars in R. #mtcars is a data frame with 32 observations on 11 variables.
- mpg=Miles/(US) gallon, cyl=Number of cylinders, disp=Displacement (cu.in.), hp=Gross horsepower, drat=Rear axle ratio, wt=Weight (1000 lbs), qsec=1/4 mile time, vs=V/S, am=Transmission (0 = automatic, 1 = manual), gear=Number of forward gears, carb=Number of carburetors
- mtcars dataset has no missing data

```
colSums(is.na(mtcars))
```

```
##  mpg  cyl disp   hp drat   wt qsec   vs   am gear carb
##    0    0    0    0    0    0    0    0    0    0    0
```

A-2: Regression Coefficients of the various linear model fits

```
# "fitall"" model coefficients indicate that manual transmission improves mpg by about 2,52 mile
s vs. automatic, mpg decreases by 3.71 mpg per 1000 lb increase in weight, and increases by 0.82
 and 0.79 per unit increase in qsec and drat values. These variables appaer to affect mpg the mo
st.
summary(fitall)$coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## ammanual     2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

```
# "fit0" coefficients show that mpg increases by 7.24 (42%  vs. automatic transmission mpg of 1
7.15
summary(fit0)$coef
```

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

```
# The "fit1" coefficients generally confirm the "fitall" model conclusions above, narrowed down
 to wt, qsec, and manual transmission providing most impact on mpg.
summary(fit1)$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
## ammanual     2.935837  1.4109045  2.080819 4.671551e-02
```

```
# The "fit2" coefficients show that manual transmission mpg decreases more (-7.75 mpg) per unit
 weight increase than the automatic cars (-3.91 mpg) when drat predictor is included.
summary(fit2)$coef
```

```
##                            Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)              29.9268717  6.5522868  4.5673935 9.739025e-05
## factor(am)automatic:wt   -3.9124383  0.8170903 -4.7882571 5.380798e-05
## factor(am)manual:wt      -7.7496236  1.1730274 -6.6065152 4.352001e-07
## factor(am)automatic:drat  0.6098141  1.5303710  0.3984747 6.934140e-01
## factor(am)manual:drat     3.2331540  1.4324484  2.2570823 3.230082e-02
```

```
# The "fit2" coefficients show that manual transmission mpg decreases more (-6.1 mpg) per unit w
eight increase than the automatic cars (-3.18 mpg) when qsec predictor is included.
summary(fit3)$coef
```

```
##                            Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)              13.9692069  5.7756116  2.418654 2.259367e-02
## factor(am)automatic:wt   -3.1758862  0.6362299 -4.991727 3.114029e-05
## factor(am)manual:wt      -6.0991935  0.9685466 -6.297264 9.703599e-07
## factor(am)automatic:qsec  0.8337859  0.2601709  3.204762 3.458031e-03
## factor(am)manual:qsec     1.4463757  0.2692125  5.372616 1.120875e-05
```

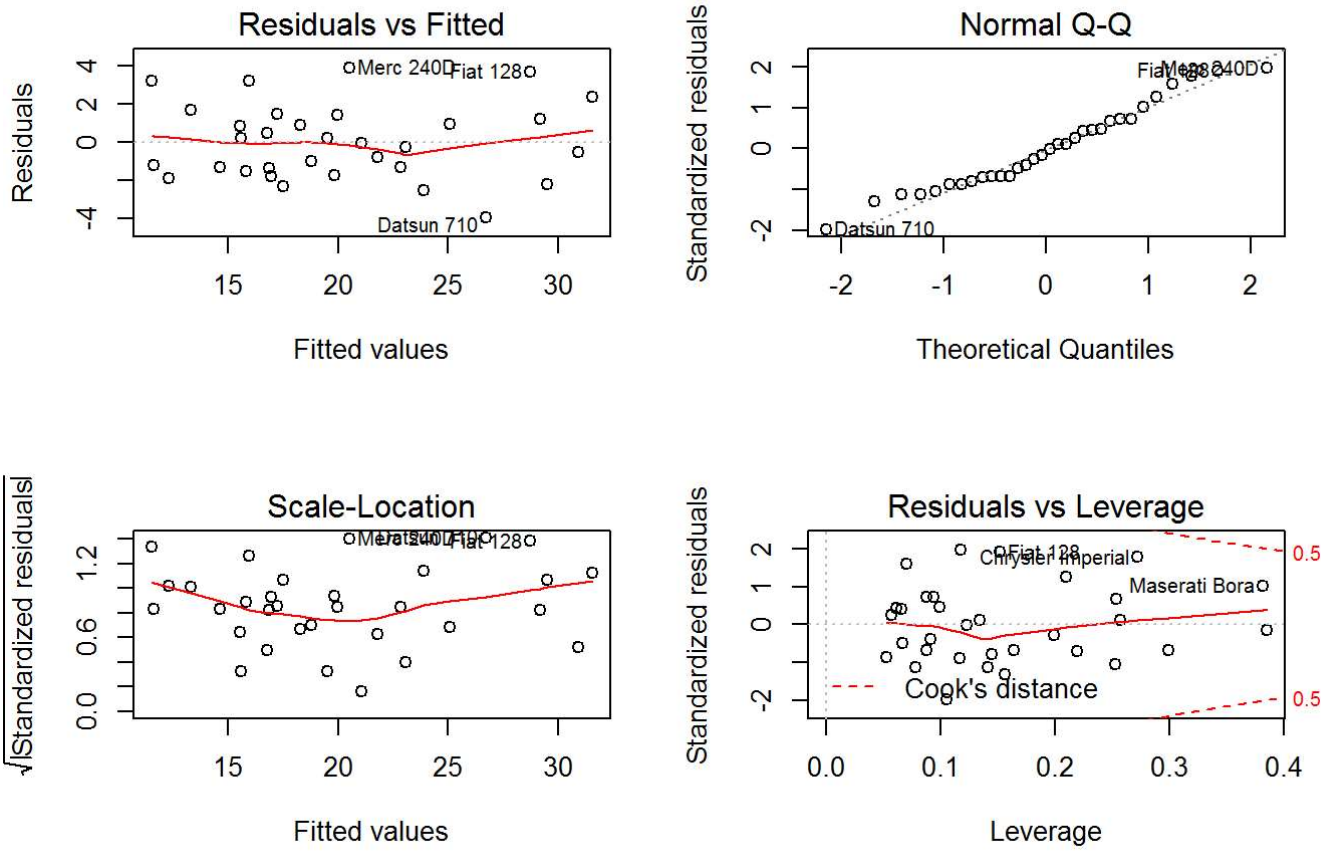
A-3: Plot of Residuals

- Patterns in residual plots generally indicate some poor aspect of model fit such as heteroskedasticity (non constant variance), and/or Missing model terms. Residual QQ plots shown indicate normality of the errors. Leverage measures (hat values) can be useful for diagnosing data entry errors and points that have a high potential for influence.

```
# Residual variance estimate (in mpg units) for the fit3 model
summary(fit3)$sigma
```

```
## [1] 2.096725
```

```
par(mfrow = c(2,2))
plot(fit3)
```



[End of Report]