RegressionModels_Project

MVS

September 24, 2015

Executive Summary

In this issue of Motor Trend magazine we examine the relationship between a set of car variables (mtcars) and miles per gallon (mpg). In particular, we look at whether an automatic or manual transmission is better for fuel consumption in terms of mpg. We quantify and visualize the mpg differences for the mtcars dataset. We begin with an EDA of the mtcars dataset: We check the relationship between each independent variable and dependent variable, as well as among independent variables, using scatterplots and correlations. Variables that are not related to the dependent variable, or those that are redundant because they are highly correlated with other independent variables, are removed from the analysis. Finally, we look at p-values to establish statistical significance of our final model coefficients.

Exploratory Data Analysis

The mtcars dataset comes from the 1974 Motor Trend US magazine. It comprises fuel consumption in miles per gallon (mpg) and 10 car variables pertaining to automobile design and performance for 32 models from 1973-1974.

A sample of the first few lines of the dataset shows us the variables collected for each car:

```
head(mtcars, n=4)
```

```
##                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
```

Fitting mutliple linear regression models:

We first attempt a simple linear regression model using mpg and transmission type:

```
summary(lm(mpg~am, data=mtcars))$coefficients
```

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

The coefficient estimate indicates the estimated change in mpg for every unit change in transmission type. Since transmission type is a binary variable (automatic = 0, manual = 1) we would interpret a 7.244 mpg increase, on average, when switching from an automatic to a manual transmission car. However, we know that other variables are involved and without properly accounting for their effect on mpg we can't explain if mpg changes are entirely controlled by transmission type.

A multivariable regression model is needed to see the effect of other variables on mpg, and on eachother:

```
summary(lm(mpg~., data=mtcars))$coefficients
```

```
##               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
## am           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
```

Using all variables as part of the model, we would interpret that a 2.52 mpg increase occurs (on average) when switching to a manual transmission car, all other variables held constant.

Looking at the (sorted) correlation values between mpg and each of the other variables:

```
sort(abs(cor(mtcars)[,1]), decreasing=TRUE)
```

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

we note that variables having a weak relationship with mpg will not contribute to the linear model. Variables that are correlated to mpg should be included in the final multivariable regression model. However, independent variables that are multicollinear with each other should also be removed to simplify the model. We can use the correlation values and the scatterplots (appendix) to see those independent variables that 1) form a linear relationship with mpg (i.e. used in our final model) 2) have no linear relationship with mpg (i.e. not used in the final model) and 3) are multicollinear with other independent variables (i.e. not used in the model).

Results

Based on the correlation values and multicollinearity found among some independent values (e.g. cyl, displ, hp, wt, and carb are strongly negatively correlated, where as drat, qsec, vs, am, and gear are strongly positively correlated, thus only one of them in each case is needed in the final model) we have the two covariate model:

```
summary(lm(mpg~cyl+am, data=mtcars))$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 34.522443  2.6031842 13.261621 7.694408e-14
## cyl         -2.500958  0.3608282 -6.931159 1.284560e-07
## am           2.567035  1.2914280  1.987749 5.635445e-02
```

Here we see that every cylinder increase decreases mile-to-fuel consumption rates by approximately 5 mpg and that approximately 2.56 mpg gain is achieved by driving a manual car (from that period!). The p-value belo 0.05 indicates a statistically significant (95%) number.

An estimate for a new value would be given by # y_hat = beta_0 + beta1 * x1 + beta_2 * x2 + error

```
beta_0 = 24.495
beta_1 = -5.002
beta_2 = 2.567
x1 = 10
x2 = 1
```

residual plot

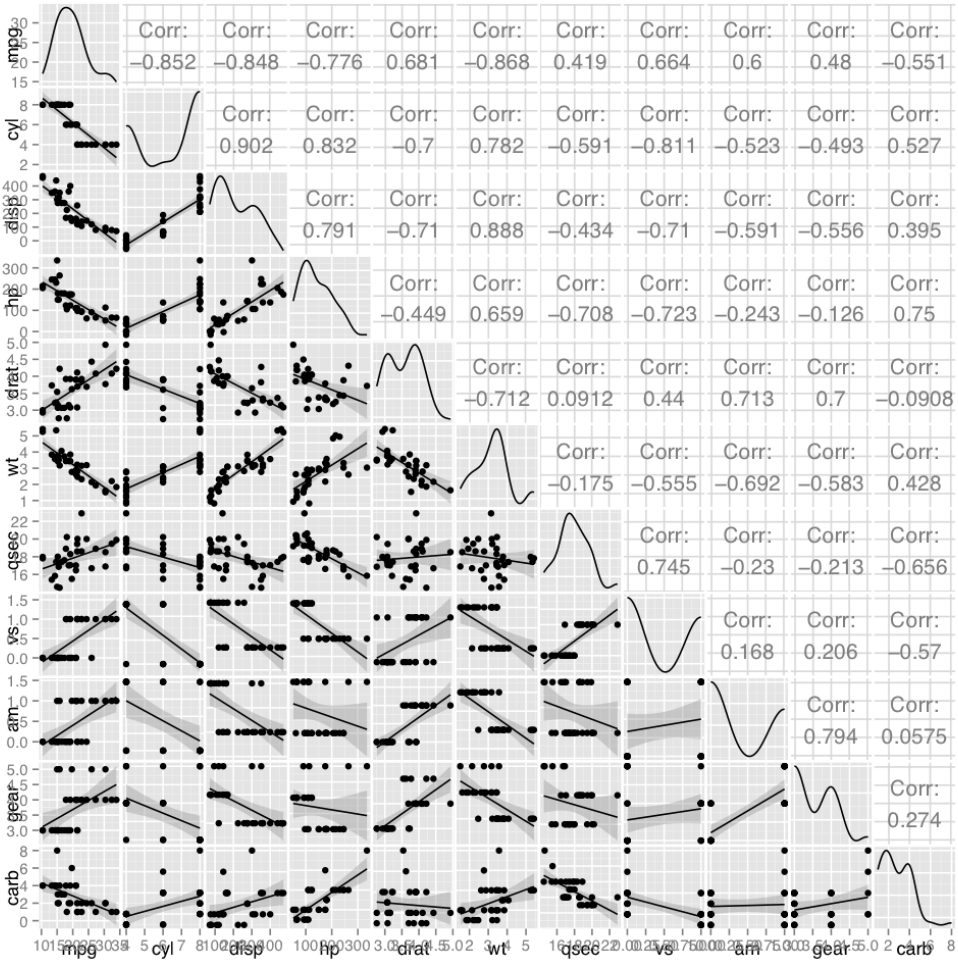
Appendix:

The scatterplots from all variables is useful for visually revealing those variables that are multicollinear with each other.

```
require(GGally)
```

```
## Loading required package: GGally
```

```
g = ggpairs(mtcars, lower = list(continuous="smooth"), params=c(method = "loess"))
g
```


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
Corr: −0.852
Corr: −0.848
Corr: −0.776
Corr: 0.681
Corr: −0.868
Corr: 0.419
Corr: 0.664
Corr: 0.6
Corr: 0.48
Corr: −0.551
Corr: 0.902
Corr: 0.832
Corr: −0.7
Corr: 0.782
Corr: −0.591
Corr: −0.811
Corr: −0.523
Corr: −0.493
Corr: 0.527
Corr: 0.791
Corr: −0.71
Corr: 0.888
Corr: −0.434
Corr: −0.71
Corr: −0.591
Corr: −0.556
Corr: 0.395
Corr: −0.449
Corr: 0.659
Corr: −0.708
Corr: −0.723
Corr: −0.243
Corr: −0.126
Corr: 0.75
Corr: −0.712
Corr: 0.0912
Corr: 0.44
Corr: 0.713
Corr: 0.7
Corr: −0.0908
Corr: −0.175
Corr: −0.555
Corr: −0.692
Corr: −0.583
Corr: 0.428
Corr: 0.745
Corr: −0.23
Corr: −0.213
Corr: −0.656
Corr: 0.168
Corr: 0.206
Corr: −0.57
Corr: 0.794
Corr: 0.0575
Corr: 0.274