Math 3080 § 1. Treibergs

Car Example: Multicollinearity in Multiple Regression

Name: Example
March 30, 2014

This **R**© program explores multicollinearity. When the input variables are highly correlated, then the effects of the variable may be confounded. The data comes from Rosenkrantz, *Probability and Statistics for Science, Engineering and Finance,* CRC Press, Boca Raton, 2009. Twelve 1992 cars were measured for fuel efficiency. Two response variables used were miles per gallon (MPG) and gallons per 100 miles (GPM).

Fitting MPG to the other variables yields an estimated coefficient $\hat{\beta}_1$ that is negative and $\hat{\beta}_2$ which is positive. It suggests that the mileage of the car decreases with an increase in the weight of the car, holding the other variables constant as we expect. SHowever, an increase in displacement which measures the size of an engine will yield an INCREASE in the miles per gallon, contrary to our expectation that a larger engine will have poorer mileage. When variables are correlated, it may not be possible to increase one of the variable while holding the others fixed.

This confounding of variables occurs because the independent variables are highly correlated. In this example, the correlation coefficient is 0.9534827. Severe multicollinearity is indicated because $R^2 > .9$

Devore suggests regressing each independent variable x_i on the others and computing

$$V(\widehat{\beta}_i) = \frac{MSE}{\sum_j (x_{ji} - \hat{x}_{ji})^2}$$

for each variable, where x_{ji} is the jth observation of the ith variable and $\hat{x}_{ji}$ is the jth fitted value when a least squares fit is run for x_i depending on the other variables $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k$. In other words, the square of the standard error of the estimated coefficient is related to the how well each variable is predicted by the others. We compute both sides of this equation and check this formula.

Let R_i^2 be the coefficient of determination in the fit of x_i on the other independent variables. The ratios

$$VIF_i = \frac{1}{1 - R_i^2}$$

are called variance inflationary factors. VIF_i is small if there is little correlation. But VIF_i's above ten are considered large and indicate that variable x_i is highly dependent. In this example, $VIF = 11.00463$. One remedy is to use alternatives to least squares regression. Another is to drop the collinear variable, and rely on the other independent variables to be its proxy..

Data Set Used in this Analysis :

```
# Math 3080-1                Car Data             March 29, 2014
# Treibergs
#
# From Rosenkrantz, "Probability and Statistics for Science, Engineering
# and Finance," CRC Press, Boca raton, 2009. Table 10.2.
# 12 1992 cars were measured for fuel efficiency. Two response variables
# used were miles per gallon (MPG) and gallons per 100 miles (GPM).
#  Variables
#      car      Make of car
#      weight   Weight in 1000 lbs
#      mpg      Miles per gallon
#      disp     Engine displacement in liters
#      gpm      Gallons per 100 miles
"car"  "weight"  "mpg"  "disp"  "gpm"
Saturn   2.495  32  1.9  3.12
Escort   2.53   30  1.8  3.34
Elantra  2.62   29  1.6  3.44
CamryV6  3.395  25  3    4
Camry4   3.03   27  2.2  3.7
Taurus   3.345  28  3.8  3.58
Accord   3.04   29  2.2  3.44
LeBaron  3.085  27  3    3.7
Pontiac  3.495  28  3.8  3.58
Ford     3.95   25  4.6  4
Olds88   3.47   28  3.8  3.58
Buick    4.105  25  5.7  4
```

R Session:

```
R version 2.13.1 (2011-07-08)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-apple-darwin9.8.0/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.41 (5874) i386-apple-darwin9.8.0]

[History restored from /Users/andrejstreibergs/.Rapp.history]
```

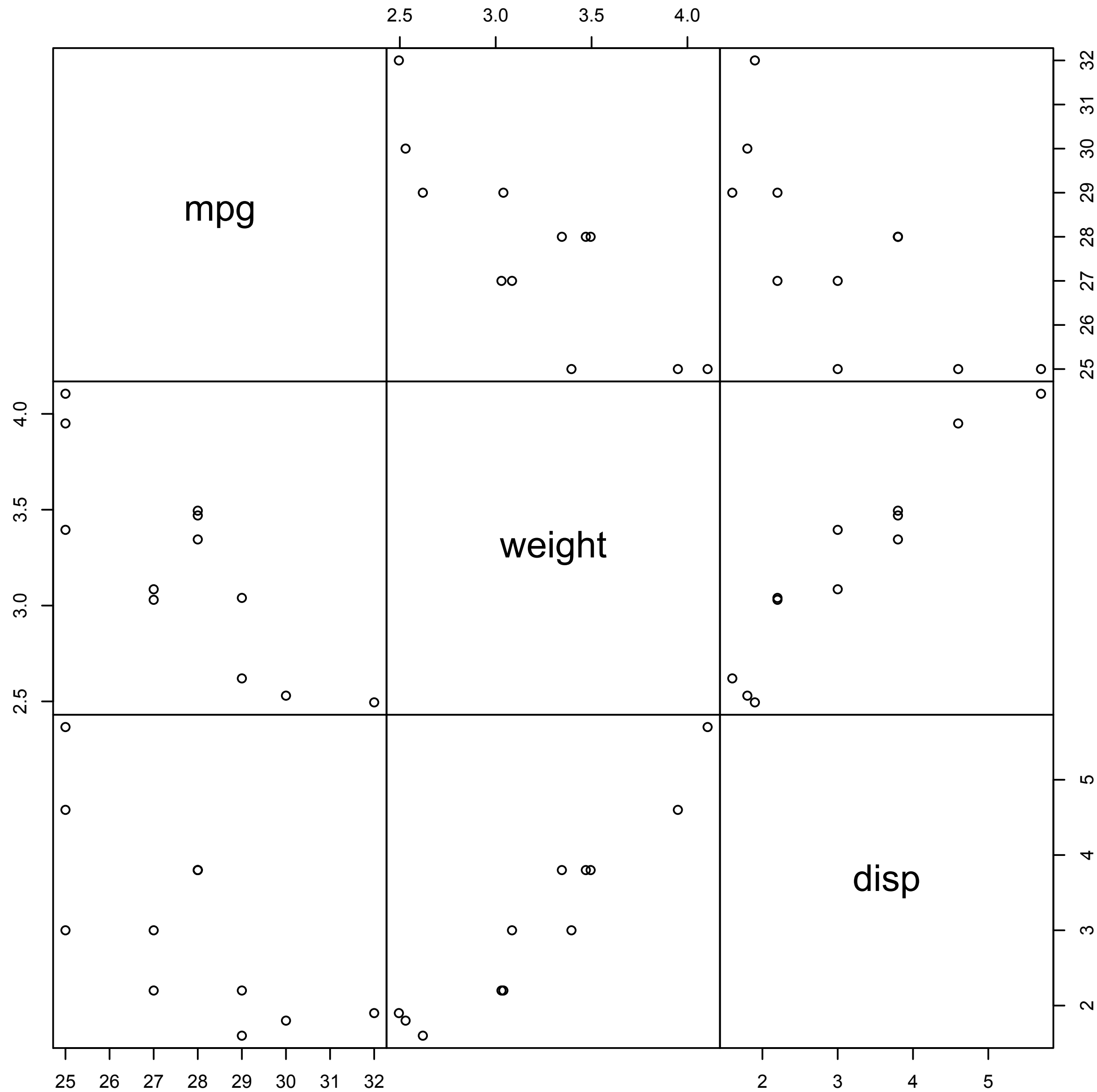
```
> tt=read.table("M3082DataCar.txt",header=T)
> attach(tt)
> tt
       car weight mpg disp  gpm
1   Saturn  2.495  32  1.9 3.12
2   Escort  2.530  30  1.8 3.34
3  Elantra  2.620  29  1.6 3.44
4  CamryV6  3.395  25  3.0 4.00
5   Camry4  3.030  27  2.2 3.70
6   Taurus  3.345  28  3.8 3.58
7   Accord  3.040  29  2.2 3.44
8  LeBaron  3.085  27  3.0 3.70
9  Pontiac  3.495  28  3.8 3.58
10    Ford  3.950  25  4.6 4.00
11  Olds88  3.470  28  3.8 3.58
12   Buick  4.105  25  5.7 4.00
> pairs(mpg~weight+disp, gap=0)
>
> ###########  FIT A LINEAR MODEL ON TWO VARIABLES  ###############
> f1=lm(mpg~weight+disp);  summary(f1);  anova(f1)

Call:
lm(formula = mpg ~ weight + disp)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5065 -0.5705 -0.2401  0.9839  1.5496

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  46.3513     4.2812  10.827 1.84e-06 ***
weight       -7.4770     2.1029  -3.556  0.00616 **
disp          1.7406     0.8632   2.016  0.07455 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.088 on 9 degrees of freedom
Multiple R-squared: 0.7878,Adjusted R-squared: 0.7406
F-statistic: 16.71 on 2 and 9 DF,  p-value: 0.000934

Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value    Pr(>F)
weight     1 34.770  34.770 29.3468 0.0004233 ***
disp       1  4.817   4.817  4.0659 0.0745539 .
Residuals  9 10.663   1.185
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> #########  NOTE THE UNEXPECTED SIGN ON COEFFICIENT beta2  ######
```

```
> #########  COMPUTE THE CORRELATION BETWEEN INDEP. VARS. ########

> cor(weight,disp)
[1] 0.9534827

> ######### VARIANCE INFLATIONARY FACTOR  ########################
> VIF = 1/(1-cor(weight,disp)^2); VIF
[1] 11.00463


> ######### GET SSE AND MSE FOR FULL MODEL  ######################
> SSE=sum(residuals(f1)^2); SSE
[1] 10.66309
> n=length(mpg);n;k=2
[1] 12
> MSE=s2/(n-k-1);MSE
[1] 1.184787



> ##########  STUDY THE DEPENDENCE OF ONE VARIABLE ON THE OTHER ####
> ########## FIT weight AS A FUNCTION OF disp  ####################
> f2=lm(weight~disp); summary(f2); anova(f2)

Call:
lm(formula = weight ~ disp)

Residuals:
      Min        1Q    Median        3Q       Max
-0.242120 -0.123552 -0.005247  0.160923  0.227331

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.99345    0.13079   15.24 3.00e-08 ***
disp         0.39141    0.03913   10.00 1.59e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1637 on 10 degrees of freedom
Multiple R-squared: 0.9091,Adjusted R-squared:   0.9
F-statistic:   100 on 1 and 10 DF,  p-value: 1.586e-06

Analysis of Variance Table

Response: weight
          Df  Sum Sq Mean Sq F value    Pr(>F)
disp       1 2.68049 2.68049  100.05 1.586e-06 ***
Residuals 10 0.26792 0.02679
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
```

```
> ###########  GET  SSEweight IN FIT ON disp  #####################
> s2=sum(residuals(f1)^2); s2
[1] 10.66309
> SSEweight=sum(residuals(f2)^2);SSEweight
[1] 0.267925

> ########## FIT disp AS A FUNCTION OF weight  ####################
> f3=lm(disp~weight); summary(f3); anova(f3)

Call:
lm(formula = disp ~ weight)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53863 -0.29351  0.05814  0.29727  0.51225

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.3470     0.7550  -5.757 0.000183 ***
weight        2.3227     0.2322  10.002 1.59e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3987 on 10 degrees of freedom
Multiple R-squared: 0.9091,Adjusted R-squared:   0.9
F-statistic:   100 on 1 and 10 DF,  p-value: 1.586e-06

Analysis of Variance Table

Response: disp
          Df  Sum Sq Mean Sq F value    Pr(>F)
weight     1 15.9067  15.907  100.05 1.586e-06 ***
Residuals 10  1.5899   0.159
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> ###########  GET  SSEdisp IN FIT ON weight  #####################
> SSEdisp=sum(residuals(f3)^2);SSEdisp
[1] 1.589936

> ####  RATIOS OF MSE TO SSEweight, SSEdisp ARE VAR. HAT-BETAS  ####
> ####  THEIR ROOTS ARE ST. ERRORS OF BETAS !!!  ###################

> c(MSE/SSEweight, MSE/SSEdisp)
[1] 4.4220858 0.7451792

> sqrt(c(MSE/SSEweight, MSE/SSEdisp))
[1] 2.1028756 0.8632376
>
```

```
> ############  disp HAS SMALLER t-VALUE SO TRY DROPPING THIS VAR. ####
> f4=lm(mpg~weight); summary(f4); anova(f4)

Call:
lm(formula = mpg ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1261 -0.8883  0.1077  0.8095  1.7832 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  38.7847     2.3559  16.463 1.43e-08 ***
weight       -3.4340     0.7246  -4.739 0.000793 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 1.244 on 10 degrees of freedom
Multiple R-squared: 0.6919,Adjusted R-squared: 0.6611 
F-statistic: 22.46 on 1 and 10 DF,  p-value: 0.000793 

Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value   Pr(>F)    
weight     1  34.77  34.770  22.461 0.000793 ***
Residuals 10  15.48   1.548                     
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

> ############  TRY DROPPING weight INSTEAD. ##############
> f5=lm(mpg~disp); summary(f5); anova(f5)

Call:
lm(formula = mpg ~ disp)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8884 -0.9140  0.2383  1.0604  2.8071 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  31.4462     1.2795  24.576 2.84e-10 ***
disp         -1.1859     0.3828  -3.098   0.0113 *  
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 1.601 on 10 degrees of freedom
Multiple R-squared: 0.4897,Adjusted R-squared: 0.4387 
F-statistic: 9.597 on 1 and 10 DF,  p-value: 0.01129 
```

```
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value  Pr(>F)
disp       1 24.608 24.6083   9.597 0.01129 *
Residuals 10 25.642  2.5642
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> ################ SO  mpg ~ weight  IS THE SUPERIOR MODEL  ########
```


mpg
weight
disp
2.5
3.0
3.5
4.0
25
26
27
28
29
30
31
32
2
3
4
5