Abbie Pearson

Jessica Warren

June Suh

STA4203

Oct. 12, 2016

Homework 7

1. Using the divusa data:

a) Fit a regression model with divorce as the response and unemployed, femlab, marriage, birth and military as predictors. Compute the condition indexes and interpret their meaning.

Code:

*proc import out=divusa*

*datafile="/home/jes13j0/Homework/STA 4203/divusa.csv"*

*dbms = csv replace;*

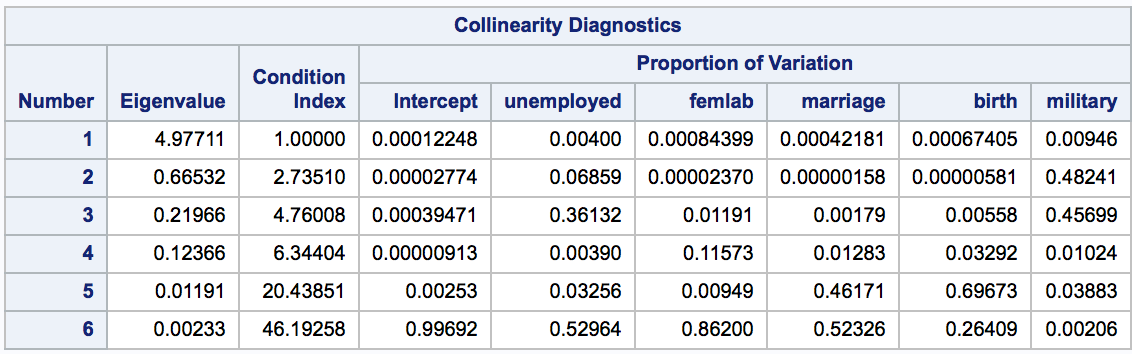
*run;*

*proc reg data=divusa;*

*model divorce = unemployed femlab marriage birth military / collin;*

*run;*

*quit;*



Since condition number is 46.19258 (κ = 46.19258 > 30) it indicates collinearity, when divorce is the dependent variable.

b) For the same model, compute the VIFs.

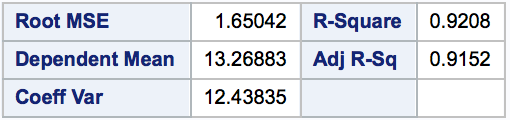
Code:

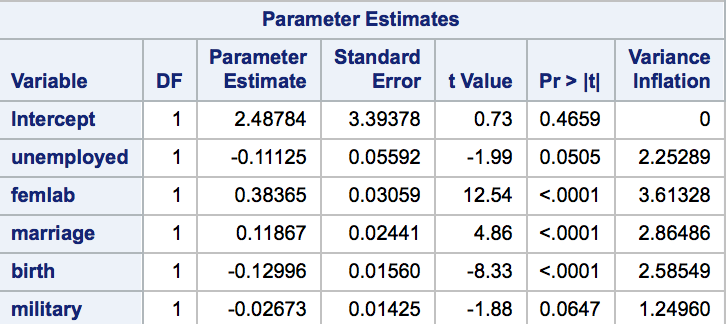
*proc reg data=divusa;*

*model divorce = unemployed femlab marriage birth military / vif;*

*run;*

*Quit;*





c) Is there any evidence of collinearity?

VIF for divorce is 12.626, according to the equation VIF(divorce) = (1/(1-R^2(divorce)) = 12.626.

Since VIF is larger than 1, we conclude that there is evidence of collinearity.

d) Does the removal of insignificant predictors from the model reduce the collinearity?

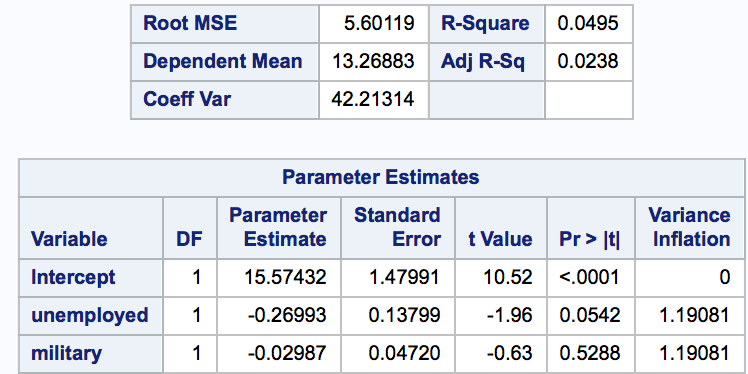
Code:

*proc reg data=divusa;*

*model divorce = unemployed military / vif;*

*run;*

*quit;*



Yes, R^2 drops significantly and VIF for the predictors are almost 1.

What we did wrong was that we LEFT the variables instead of TAKING THEM OUT and didn’t look at /collin condition number either

2. The seatPosEven and seatPosOdd datasets are the even and odd observations from the seatpos dataset. In this problem we will observe that removing collinearity obtains a more stable model that makes better predictions.

a) Fit a model on the seatPosEven data with hipcenter as the response and the other variables except id as predictors. We saw in class that this model has collinearity. Output the model into a dataset modelEven, and report its R square and the SSE (sum of squares of residuals).

Code:

*proc import out=seatPosEven*

*datafile="/home/jes13j0/Homework/STA 4203/seatPosEven(1).csv"*

*dbms = csv replace;*

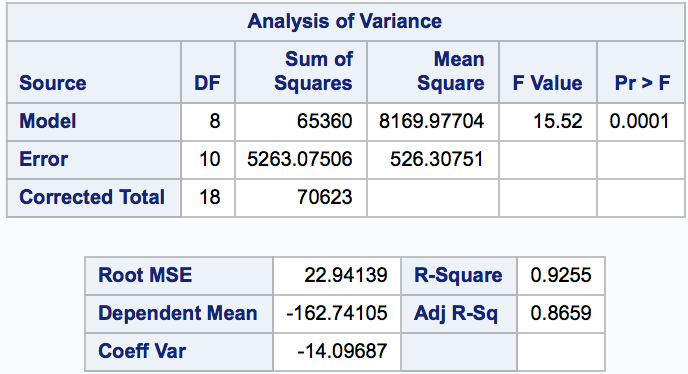
*run;*

*proc reg data=seatposeven outest=modelEven;*

*model hipcenter = age weight htshoes ht seated arm thigh leg;*

*run;*

*quit;*



R square value is 0.9255 and SSE is 5263.07506.

b) Use modelEven obtained above to obtain the predicted values and residuals on the seatPosOdd data using proc score. Report the sum of squares of the obtained residuals.

Code:

*proc import out=seatPosOdd*

*datafile="/home/jes13j0/Homework/STA 4203/seatPosOdd(1).csv"*

*dbms = csv replace;*

*run;*

*proc score data=seatposodd score=modeleven out=test1 residual type=parms;*

*var age weight htshoes ht seated arm thigh leg;*

*run;*

*proc print data=test1;*

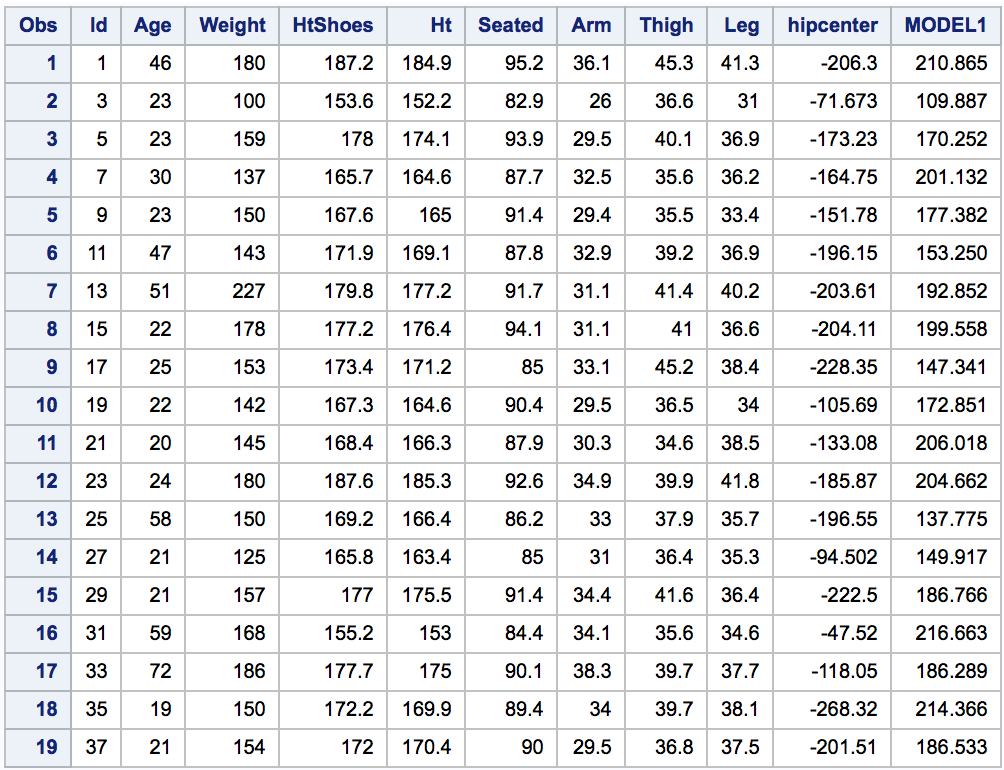
*run;*

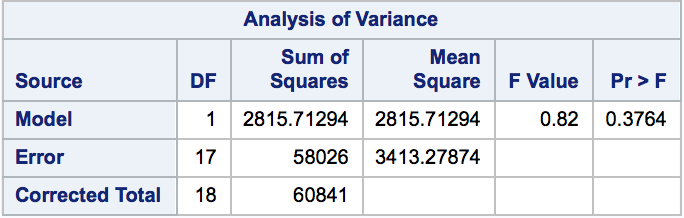
*proc reg data=test1;*

*model hipcenter = model1 /p;*

*run;*

*quit;*





SSE is 58026.

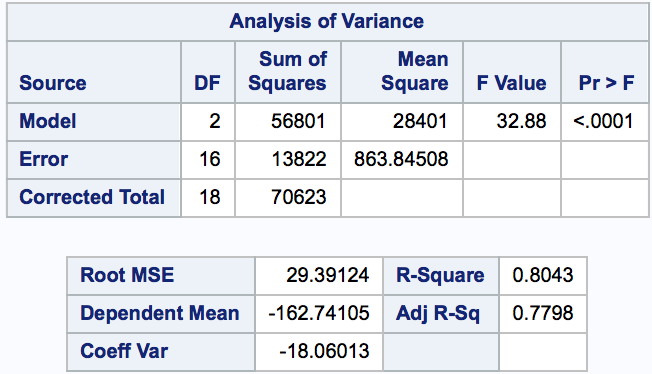
c) Fit a model on the seatPosEven data with hipcenter as the response and age and ht as predictors. Output the model into a dataset modelEven2 , and report its R square and the SSE.

Code:

*proc reg data=seatposeven outest=modelEven2;*

*model hipcenter = age ht;*

*run;*



R2 is 0.8043 and SSE is 13822.

d) Use modelEven2 obtained above to obtain the predicted values and residuals on the seatPosOdd data. Report the sum of squares of the obtained residuals.

Code:

*proc score data=seatposodd score=modeleven2 out=test2 residual type=parms;*

*var age ht;*

*run;*

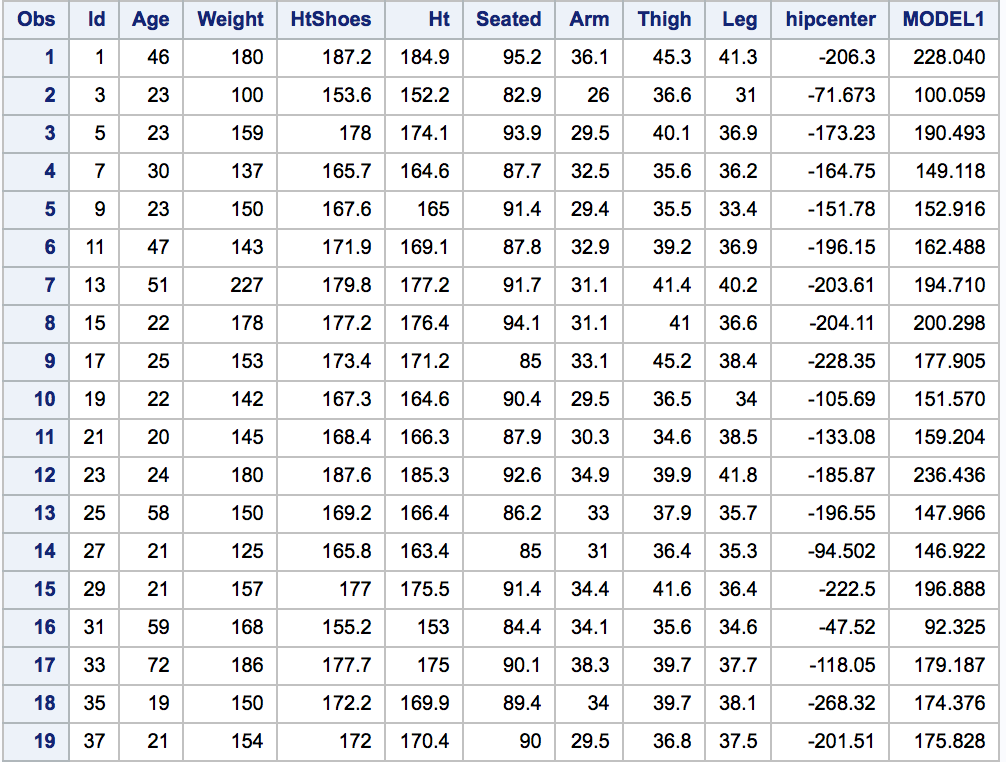
*proc print data=test2;*

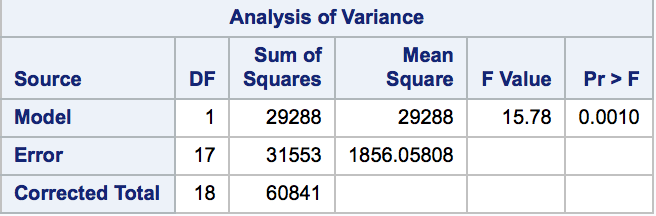
*run;*

*proc reg data=test2;*

*model hipcenter = model1 /p;*

*run;*





SSE is 31553.