Print

Demo: Calculating Collinearity Diagnostics Using PROC REG

Filename: **st105d03.sas**

In this demonstration, we use PROC CORR to investigate the correlations between the variable score and the other interval variables. Then we'll use PROC REG and the VIF option to assess the magnitude of the collinearity problem.

PROC CORR DATA=*SAS-data-set* <*options*>;
    MODEL *variables*;
    WITH *variables*;
    ID *variables*;
RUN;

PROC REG DATA=*SAS-data-set* <*options*>;
    MODEL *dependents* = <*regressors*> < / *options*>;
RUN;

1. Open program st105d03.sas.

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st105d03.sas*/  /* Part A*/
proc sort data=STAT1.ameshousing3 out=STAT1.ames_sorted;
    by PID;
run;
proc sort data=STAT1.amesaltuse;
    by PID;
run;

data amescombined;
    merge STAT1.ames_sorted STAT1.amesaltuse;
    by PID;
run;

title;
proc corr data=amescombined nosimple;
    var &interval;
    with score;
run;

/*st105d03.sas*/  /*Part B*/
proc reg data=amescombined;
    model SalePrice = &interval score / vif;
```

```
    title 'Collinearity Diagnostics';
run;
quit;

proc reg data=amescombined;
    NOSCORE: model SalePrice = &interval / vif;
    title2 'Removing Score';
run;
quit;
```

In Part A, we'll first combine the score data from the other research group with the data we already have. The PROC CORR step produces Pearson correlation statistics and corresponding p-values. We specify the new data set, amescombined, with the nosimple option to suppress the descriptive statistics. In the VAR statement, we specify the continuous variables listed in the interval macro variable. By default, SAS produces correlations for each pair of variables in the VAR statement, but we'll use the WITH statement to correlate each continuous variable with score.

2. Submit the code in Part A.

3. Review the output.

   In the Pearson Correlation table, the new variable, score, appears to be significantly correlated with all the interval variables, but focus your attention on the actual correlations in the first row. Recall that closer to 1 or -1 implies a stronger correlation between two variables. Score is highly correlated with Basement_Area, and moderately correlated with Above Ground Living Area and Total_Bathroom. The correlation with Basement_Area is large enough that they'll clearly provide redundant information about sale price and should not both be included in the same model. Let's check for additional sources of collinearity that might not be detected with correlation coefficients.

4. Let's go back to the code and look at Part B. We're going to use PROC REG with the VIF option to further assess the collinearity problem and identify the predictors involved in the problem. In the PROC REG statement, we specify the amescombined data set. The MODEL statement specifies SalePrice as the response variable and all of the variables in the macro variable and the score variable as predictors, followed by a forward slash, and then the VIF, or variance inflation factor, option. SAS calculates the VIF for each predictor term in the model. The VIF_i is the ratio of $VIF_i = \frac{1}{1-R_i^2}$, where R_i^2 is the R-square value when regressing the i^{th} predictor, X_i, on all the other predictors in the model. But the important thing to remember is the approximate cutoff value. If the VIF is greater than 10 for any predictors in the model, those predictors are likely involved in collinearity.

5. Submit the first PROC REG step in Part B.

6. Review the output.

   In the Parameter Estimates table, VIF values are displayed in the Variance Inflation column. The VIFs for Above Ground Living Area, Basement_Area, and score are much larger than 10, so a severe collinearity problem is present. At this point there are many ways to proceed. You might use some subject-matter expertise if available. Another option is to systematically remove variables starting with the highest VIF and re-run the analysis. Much like p-values, VIF values will need to be updated with each successive variable removal.

   We decided to contact the researchers that provided the score variable, and determined that score is a composite variable. The researchers, on the basis of prior literature, created a composite variable, which is a weighted function of the two variables, Above Ground Living Area and Basement_Area. Score is equal to 10,000 minus twice Above Ground Living Area plus 5 times Basement_Area, and rounded.

   This isn't an uncommon occurrence and illustrates an important point. If a composite variable is included in a model along with some or all of its component measures, there's bound to be collinearity. If the composite variable has meaning, it can be used as a substitute measure for both components, and you can remove the variables Above Ground Living Area and Basement_Area from the analysis. However, composite measures

have the disadvantage of losing some information about the individual variables. If this a concern, then you can remove score from the analysis.

We'll remove score from the analysis in order to maintain the information about the two variables, Above Ground Living Area and Basement_Area. Then we'll check the variance inflation factors again to see whether collinearity remains a problem.

7. Let's go back to the code. In the last PROC REG step, we've removed score from the MODEL statement, and added a label of NOSCORE. Let's run this new model and calculate the VIFs.

8. Submit the last PROC REG step.

9. Review the output.

   Scroll to the Parameter Estimates table. As you can see, all the VIF values are smaller than 2 now. Because collinearity can have a substantial effect on the outcome of a stepwise model selection procedure, it's advisable to deal with collinearity before using any automated model selection tool. The eight variables in question no longer exhibit a high degree of collinearity, and could now be safely passed into a stepwise selection approach.

Close