Practice 1.3 (Level 1): Dealing with Multicollinearity in a Simple Polynomial Regression Model

Task

In this practice, you diagnose and remediate multicollinearity in a simple polynomial regression model.

In the previous practice, you fit a simple polynomial regression model to the data set **mydata.cafeteria**. This model characterizes the relationship between the number of self-service coffee dispensers in a cafeteria line and coffee sales in a sample consisting of fourteen cafeterias.

Note: Before you complete this practice, you must run the Practice 1.1 code in the same SAS session.

Reminder: Make sure you've defined the **mydata** library.

1. Write a PROC CORR step to compute the Pearson correlation coefficient between **Dispensers** and **Dispensers^2**, as follows:
    - As the input data set, specify the data set that contains the design matrix, which was output from PROC GLMSELECT in the previous Level 1 practice.
    - In the VAR statement, reference the &_GLSMOD macro variable to specify the variables.
    - In the PROC CORR statement, use the NOSIMPLE and PLOTS( )=MATRIX options.

    Submit the code and examine the tabular and graphical output. What do you conclude?

    ```
    proc corr data=d_disp nosimple plots()=matrix;
       var &_GLSMOD;
    run;
    ```

    In the results, the Pearson Correlation Coefficients table shows that the variables **Dispensers** and **Dispensers^2** are highly correlated. The correlation coefficient is 0.96 and is significantly different from zero. The scatter plot between **Dispensers** and **Dispensers^2** shows a strong curvilinear relationship, as expected. **Note:** Depending on your SAS interface, the name of the second-degree polynomial term of **Dispensers** varies by one character. SAS Studio assigns the variable name **Dispensers^2** and Base SAS assigns the name **Dispensers_2**. However, in both interfaces, the label is **Dispensers^2**.

2. Write a PROC REG step with the appropriate options in the MODEL statement to compute the variance inflation factor (VIF) and the collinearity diagnostics statistics. Submit the code. Is there collinearity among the independent variables? If so, which ones?

    ```
    proc reg data=d_disp;
       model Sales=&_GLSMOD /
             vif collin collinoint;
    run;
    quit;
    ```

    In the results, the Parameter Estimates table shows that the the VIF values for **Dispensers** and **Dispensers^2** are both 13, which indicates strong collinearity. However, in the Collinearity Diagnostics table, the largest condition index value is 13.8, which suggests weak collinearity. It is not uncommon to reach inconsistent conclusions about collinearity based on different statistics. It might be a good idea to reduce the possible collinearity for more reliable inferences.

3. Use PROC GLMSELECT with an EFFECT statement to create a new, centered, quadratic effect for **Dispensers** and fit a model with the centered polynomial terms. Use the OUTDESIGN option to create a new model design data set.

Obtain collinearity diagnostics from PROC REG, using the design data set produced by PROC GLMSELECT. Does the centered model appear to have multicollinearity among the independent variables?

```
proc glmselect data=mydata.cafeteria outdesign=d_dispc;
        effect q_dispc=polynomial(Dispensers /
          degree=2 standardize(method=moments)=center);
        model Sales = q_dispc / selection=none;
run;

ods select ParameterEstimates CollinDiag
                CollinDiagNoInt;
proc reg data=d_dispc;
   model Sales = &_GLSMOD / vif collin collinoint;
   title 'Centered Quadratic Model';
run;
title;
quit;
```

As shown in the results from PROC REG, after you center the independent variable **Dispensers**, the VIF is reduced to 1 for both centered variables. The largest condition index value is only 2.68. There does not seem to be any collinearity among the (centered) independent variables.

4. After refitting the polynomial model using the centered variables, in the previous step, did both the Analysis of Variance table and the Parameter Estimates table change? Why or why not?

The Parameter Estimates table changed but the ANOVA table did not. Centering the data changes the actual data values, which causes the change in the parameter estimates. However, centering the data does not change the relationship between the variables. Centering performs only a location change, not a scale change, so the values in the ANOVA table do not change.

Hide Solution

Close