Collinearity of independent variables

Collinearity is a condition in which some of the independent variables are highly correlated.

Why is this a problem?

Collinearity tends to inflate the variance of at least one estimated regression coefficient, $\widehat{\beta}_j$.

This can cause at least some regression coefficients to have the wrong sign.

The standard error of $\widehat{\beta}_j$ has the form

$$\sigma S_j,$$

where S_j depends only upon the values of the independent variables. If collinearity is present, at least one S_j will be large.

Ways of dealing with collinearity:

1) Ignore it. If *prediction* of y values is the object of your study, then collinearity is not a problem.

2) Use an estimator of the regression coefficients other than the least squares estimators. An alternative is to use *ridge regression* estimators; Draper and Smith (1981), *Applied Regression Analysis*, 2nd edition, pp. 313-324.

3) Get rid of the "redundant" variables by using a variable selection technique.

How is collinearity identified?

1) Examine the correlation coefficient for each pair of independent variables. A value of the correlation near ± 1 indicates that the two variables are highly correlated. Use *Analyze* $\rightarrow$ *Correlate* $\rightarrow$ *Bivariate* in SPSS to obtain the correlation coefficients.

2) The *variance inflation factors* are also very useful. $VIF(j)$ is the factor by which the variance of $\widehat{\beta}_j$ is increased over what it would be if x_j was uncorrelated with the other independent variables. If all values of $VIF(j)$ are near 1, then collinearity is not a problem. $VIF(j) > 10$ indicates serious collinearity.

   The variance inflation factors are obtained via *Regression* $\rightarrow$ *Linear* $\rightarrow$ *Statistics* $\rightarrow$ *Collinearity diagnostics*.

Variable selection

We'd like to select the smallest subset of independent variables that explains almost as much of the variation in the response as do *all* the independent variables.

Several methods can be used for variable selection, including R^2, adjusted R^2, Mallow's C_p, forward selection, and backward elimination.

$\underline{R^2}$

Generally speaking, models with high R^2 values are good. However, one must be careful not to carry this idea too far.

Model 1: contains $x_1, x_2, \ldots, x_k$

Model 2: contains $x_1, x_2, \ldots, x_k$ *plus* other independent variables

R^2 for Model 2 must be at least as big as R^2 for Model 1.

If the increase in R^2 is small in comparison to the number of extra independent variables, it's usually not a good idea to use the more complicated model.

> *Parsimony*, or Occam's razor applied to statistical modeling: Use the simplest model that's consistent with the data.

Adjusted R^2

For a model with m independent variables,

$$R^2_{\text{adj}} = 1 - (1 - R^2)\left(\frac{n-1}{n-m-1}\right).$$

Choose the model that maximizes R^2_{adj}.

Mallow's C_p

For a model with p regression coefficients (including the intercept),

$$C_p = \frac{SSE_p}{MSE} - (n - 2p),$$

where SSE_p is the SSE for the model with p independent variables and MSE is the mean squared error for the full model.

Good models are ones with small C_p values and/or values of C_p close to p.

Forward selection

- Variables are chosen sequentially in a series of m steps. (Total number of independent variables is m.)

- At each step, one variable is added to ones that were chosen on previous steps.

- A variable is added at step j if

  i) it maximizes R^2 among all models containing a new variable and variables added at steps $1, \ldots, j-1$, and

  ii) its P-value (using a Type II F-test) is smaller than a prespecified threshold.

Note: Take the threshold to be larger than the usual 0.05 if you want to ensure that all variables end up being entered.

Consider a case with 8 independent variables: $x_1, x_2, \ldots, x_8$.

After step 2, suppose the variables in the model are x_1 and x_8.

Let $R^2(x_1, x_8, x_j)$ be the R^2 for a model that contains x_1, x_8 and x_j.

- Compute $R^2(x_1, x_8, x_j)$ for $j = 2, 3, 4, 5, 6, 7$.

- Let's suppose $R^2(x_1, x_8, x_5)$ is the largest of these six numbers.

- Variable x_5 is added at step 3 if its Type II SS P-value in the model containing x_1, x_5, x_8 is smaller than the threshold.

- If the P-value is larger than the threshold, no further steps are performed.

The forward selection method is obtained in SPSS via *Analyze* → *Regression* → *Linear*.

- From the drop down menu next to *Method:* select *Forward*.

- By default the entry threshold is 0.05. If you want a different one, click *Options* and change the number next to Entry: to the desired threshold. (Note: The number next to *Removal:* must be bigger than the entry threshold.)

Soil Evaporation Data

Data recorded from June 16 to July 21 at a location in west central Texas.

Response y is daily amount of evaporation from the soil.

Ten independent variables:

- x_1: maximum soil temperature
- x_2: minimum soil temperature
- x_3: average soil temperature

- x_4: maximum air temperature
- x_5: minimum air temperature
- x_6: average air temperature
- x_7: maximum relative humidity
- x_8: minimum relative humidity
- x_9: average relative humidity
- x_{10}: total wind (miles per day)

Summary of Soil Evaporation Analysis

1. There are strong correlations between some of the variables, and hence the possibility of collinearity problems.

2. Collinearity verified by large variance inflation factors in the model with all ten independent variables.

3. Several insignificant t-tests indicate the likelihood that not all ten variables are needed in the model. So, we proceeded to variable selection.

4. Forward selection and use of adjusted-R^2 suggest using model 6, the one with variables $x_1, x_3, x_6, x_8, x_9, x_{10}$.

5. Personally, I prefer model 4 that has the four independent variables x_3, x_6, x_9, x_{10}. Why?

   - Model 4 is simpler but still has an R^2 that is only a bit smaller than that of model 6.

   - All the variance inflation factors for model 4 are smaller than 10, while three variance inflation factors for model 6 are larger than 20.

   - No P-value for a model 4 t-test is larger than 0.120, while model 6 has three t-tests with P-values larger than 0.20.

6. Before definitely choosing a model, residual plots should be examined to investigate possible nonlinear relationships.