Universidade do Estado do Rio de Janeiro
Instituto Politécnico
Departamento de Modelagem Computacional

Disciplina de Modelos Lineares 2012-2

Professora Ariane Ferreira

Análise de Colinearidade e Multicolinearidade

Quando trabalhamos com mais de uma variável regressora, é muito importante verificar se essas variáveis explicativas são correlacionadas. Desta forma, se não houver nenhum relacionamento entre elas, dizemos que são ortogonais.

Na prática, é muito difícil que as variáveis de entrada sejam ortogonais e, felizmente, a falta de ortogonalidade não é séria. Mas se as variáveis forem muito correlacionadas, as inferências baseadas no modelo de regressão podem ser errôneas ou pouco confiáveis.

Por isso, é necessário verificar se as variáveis são altamente correlacionadas. Na literatura, os termos **Colinearidade (Multicolinearidade)** são utilizados para indicar a existência forte de correlação entre duas (ou mais) variáveis independentes. Entretanto, alguns autores designam de **Colinearidade** a existência de relação linear entre duas variável explicativa (matriz de correlação) e de **Multicolinearidade** a existência de relação linear entre uma variável explicativa e as demais.

Se a matriz X'X é singular, isto é, algumas variáveis explicativas são combinações lineares de outras, temos **Multicolinearidade** e não há Estimadores de Mínimos Quadrados único para os parâmetros. Se X'X é aproximadamente singular, temos **Multicolinearidade** aproximada.

Colinearidade

Algumas características no ajuste do modelo podem indicar problema de colinearidade. São elas:

- $\hat{\beta}_i$ com sinal oposto ao esperado;
- grandes mudanças em $\hat{\beta}_i$ quando adicionamos ou excluimos variáveis ou observações;
- $\hat{\beta}_i$ não significante para um X_i teoricamente importante.

Para diagnosticar colinearidade, temos a seguinte opção:

- verificar se a matriz de correlações das variáveis explicativas apresenta altas correlações. Se a correlação de duas variáveis for próxima de 1, indica problema.

Matriz de Correlação

A matriz de correlação apresenta a correlação das variáveis com todas as outras em consideração. Vamos supor que temos três variáveis: X_1, X_2e X_3. A matriz de correlação é dada por

$$r = \begin{bmatrix} r_{(x_1,x_1)} & r_{(x_1,x_2)} & r_{(x_1,x_2)} \\ r_{(x_2,x_1)} & r_{(x_2,x_2)} & r_{(x_2,x_3)} \\ r_{(x_3,x_1)} & r_{(x_3,x_2)} & r_{(x_3,x_3)} \end{bmatrix},$$

em que

$$r_{x_j,x_k} = \frac{\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)}{(n-1)s_{x_j}s_{x_k}}.$$

s_{x_j} e s_{x_k} são os desvios padrão de X_j e X_k, respectivamente.

Se a correlação de duas variáveis é maior de 0,9, é indicativo de correlação forte entre elas.

Multicolinearidade

A multicolinearidade é um problema no ajuste do modelo que pode causar impactos na estimativa dos parâmetros. Podemos diagnosticar Multicolinearidade por meio do VIF (Variance Inflation Factor).

VIF

Os elementos da diagonal principal de $(X'X)^{-1}$são também úteis para detectar multicolinearidade. O j-ésimo elemento da diagonal principal $(X'X)^{-1}$, C_{jj} pode ser escrito como

$$C_{jj} = (1 - R_j^2)^{-1}, \qquad j = 1, ..., p.$$

em que R_j^2 é o R^2 da regressão de X_j sobre as outras variáveis explicativas.

C_{jj} é chamado de fator de inflação da variância e outra notação usada é VIF_j. Assim, o VIF_j é dado por

$$VIF_j = \frac{1}{1 - R_j^2}.$$

Verificamos que VIF_j mede o quanto a variância do coeficiente $\hat{\beta}_j$ é inflacionada por sua colinearidade.

Geralmente, o VIF é indicativo de problemas de multicolinearidade se VIF>10.

Bibliografia:

Neter ,J.; Wasserman, William; Kutner, M.H., Applied linear statistical models;
Draper,N.R.; Smith,H., Applied Regression Analysis.
Montgomery and Peck, Introduction to Linear Regression Analysis;
Seber, G.A.F., Linear Regression Analysis.
Myers and Montgomery, Generalized Linear Models.