STA3010 Regression Analysis

Feng YIN

The Chinese University of Hong Kong (Shenzhen)

yinfeng@cuhk.edu.cn

March 5, 2020

Multicollinearity

- Multicollinearity is said to exist, when there are near linear dependencies among the inputs/features/regressors.
- For a subset of inputs, there exists linear dependency $\sum_{i=1}^{p} t_i \mathbf{x}_i = \mathbf{0}$, $t_i \neq 0$. Even the above equality approximately holds, your computer may doubt about the existence of the multicollinearity problem.

Effects of Multicollinearity

- Consider a multiple linear regression model with two inputs, x_1 and x_2. Both the output y and the inputs x_1 and x_2 are scaled to unit length.
- The least-squares estimator, $\hat{\beta}_1$ and $\hat{\beta}_2$ are obtained as:

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{1 - r_{12}^2}, \qquad \hat{\beta}_2 = \frac{r_{2y} - r_{12}r_{1y}}{1 - r_{12}^2}. \tag{1}$$

- Strong multicollinearity results in:
  - the correlation coefficient r_{12} will be large.
  - $var(\hat{\beta}_j) = \sigma^2 \mathbf{C}_{jj} \to \infty$, as $|r_{12}| \to 1$, namely large variances and covariances for the least-squares estimators of the model parameters.

Multicollinearity Diagnosis

1. Examination of the correlation matrix: Diagnose multicollinearity through inspecting the off-diagonal elements. If input x_i and x_j are nearly linearly dependent, then $|r_{ij}|$ will be close to 1.

$$
\mathbf{X}'\mathbf{X} = \begin{bmatrix}
1.000 & 0.224 & -0.958 & -0.132 & 0.443 & 0.205 & -0.271 & 0.031 & -0.577 \\
 & 1.000 & -0.240 & 0.039 & 0.192 & -0.023 & -0.148 & 0.498 & -0.224 \\
 & & 1.000 & 0.194 & -0.661 & -0.274 & 0.501 & -0.018 & 0.765 \\
 & & & 1.000 & -0.265 & -0.975 & 0.246 & 0.398 & 0.274 \\
 & & & & 1.000 & 0.323 & -0.972 & 0.126 & -0.972 \\
 & & & & & 1.000 & -0.279 & -0.374 & 0.358 \\
 & & & & & & 1.000 & -0.124 & 0.874 \\
 & & & & & & & 1.000 & -0.158 \\
 & & & & & & & & 1.000 \\
\text{Symmetric} & & & & & & & &
\end{bmatrix}
$$

Source: textbook

This method is only able to detect linear or near-linear dependence between pairs of inputs/regressors !

Multicollinearity Diagnosis

2. Eigensystem analysis of $\mathbf{X}^T\mathbf{X}$: The eigenvalues of $\mathbf{X}^T\mathbf{X}$ can be used to measure the extent of multicollinearity in the data.

The procedure is the following:

- Step 1: perform eigenvalue decomposition of $\mathbf{X}^T\mathbf{X}$ and obtain the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$.
- Step 2: rank the eigenvalues and compute the condition number of $\mathbf{X}^T\mathbf{X}$ as $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$.
- Step 3: If the condition number κ exceeds 1000, severe multicollinearity is indicated. Otherwise, it is numerically safe.

Multicollinearity Diagnosis

For the acetylene data (refer our textbook), the resulting condition indices, $\kappa_j = \frac{\lambda_{max}}{\lambda_j}$, $j = 1, 2, ..., p$, are obtained as

$$\kappa_1 = \frac{4.2048}{4.2048} = 1, \quad \kappa_2 = \frac{4.2048}{2.1626} = 1.94, \quad \kappa_3 = \frac{4.2048}{1.1384} = 3.69$$

$$\kappa_4 = \frac{4.2048}{1.0413} = 4.04, \quad \kappa_5 = \frac{4.2048}{0.3845} = 10.94, \quad \kappa_6 = \frac{4.2048}{0.0495} = 84$$

$$\kappa_7 = \frac{4.2048}{0.0136} = 309.18, \quad \kappa_8 = \frac{4.2048}{0.0051} = 824.47, \quad \boxed{\kappa_9 = \frac{4.2048}{0.0001} = 42{,}048}$$

Source: textbook

There is at least one strong near-linear dependence in the acetylene data. Eigensystem analysis can also be used to identify the nature of the near-linear dependencies in data.