Chapter 10: Multicollinearity.

Intro:

The quality of estimates, as measured by their variances, can be seriously and adversely affected if the independent variables are closely related to each other. This situations is call multicollinearity.

Multicollinearity and Its effect:

If the columns of X are linearly dependent, then $X'X$ is singular and the estimate of β, which depends on $(X'X)^{-1}$, cannot be unique.

A much more troublesome situation arises when the columns of X are nearly linearly dependent.

Since singularity may be defined in terms of the existence of a unit vector of c ($c'c = 1$) such that

$Xc = 0$ or $c'X'Xc = 0$, we may characterize near singularity in terms of the existence of a unit vector c such that $\|Xc\|^2 = c'X'Xc = \delta$ is small.

(i.e. for some $c = (c_0, \cdots c_K)'$ of unit length,

the length of $\sum_{j=0}^{K} c_j x_{[j]}$ is small where

$X = (x_{[0]}, \cdots, x_{[K]})$ ($x_{[K]}$: the column of X.

x_i: the row of X)

when near singularity exists, the variance of estimates can be adversely affected (large)

$$\|c\|^2 = 1 = c'c = [c'c]^2 = [c'(X'X)^{-\frac{1}{2}}(X'X)^{\frac{1}{2}}c]^2.$$

Cauchy-Schwartz inequality,

$$(u'v)^2 \le \|u\|^2\|v\|^2 = (u'u)(v'v) \quad u' = c'(X'X)^{-\frac{1}{2}}$$

$$v = (X'X)^{\frac{1}{2}}c.$$

$$\Rightarrow 1 \le [c'(X'X)^{-1}c][c'(X'X)c] = \delta c'(X'X)^{-1}c$$

$$\Rightarrow \quad c'(X'X)^{-1}c \geq \frac{1}{\delta}.$$

$$\Rightarrow \quad var(c'\hat{\beta}) = \sigma^2 c'(X'X)^{-1}c \geq \frac{\sigma^2}{\delta}.$$

$var(c'\hat{\beta})$ will be large if δ is small.

($\Leftrightarrow$ near singularity of $X'X$)

Moreover, near singularity can magnify effects of inaccuricies in the elements of X.

Since $\sum_{j=0}^{k} c_j x_{[j]}$ is affected by the units in which the variables are measured, when assessing smallness it is desirable to scale X. i.e instead of $y = X\beta + \varepsilon$.

Consider the equivalent model

$$y = X_{(s)}\beta_{(s)} + \varepsilon.$$

where $X_{(s)} = XD_{(s)}^{-1}$, $\beta_{(s)} = D_{(s)}\beta$

and $D_{(s)} = diag(\|x_{[0]}\|, \cdots, \|x_{[k]}\|)$.

$$\hat{\beta}_{(s)} = \left(X'_{(s)} X_{(s)} \right)^{-1} X'_{(s)} \cdot y$$

$$= \left(D_{(s)}^{-1} X' X D_{(s)}^{-1} \right)^{-1} D^{-1}_{(s)} X' y$$

$$= D_s (X'X)^{-1} D_{(s)} D^{-1}_{(s)} X' y$$

$$= D_s (X'x)^{-1} X' y.$$

$$\Rightarrow \hat{\beta}_{(s)} = D_{(s)} \hat{\beta}, \quad cov(\hat{\beta}_{(s)}) = D_{(s)} \, Cov(\hat{\beta}) \, D_{(s)}$$

A consequence of this scaling is that it removes from consideration near singularity caused by a single $X_{[j]}$ being small length.

Detecting Muticollinearity.

1. Tolerances and variance inflation factors

One obvious method of assessing the degree to which each independent variable is related to all other independent variables is to examine R_j^2, which is the value of R^2 between the variable x_j and all other independent variables.

The tolerance $TOL_j = 1 - R_j^2$

TOL_j is close to one if x_j is not closely related to other predictors.

The variance inflation factor $VIF_j = TOL_j^{-1}$

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

A value of VIF_j close to one indicates no relationship, while larger values indicate presence of multicollinearity.

Eigenvalues and condition numbers.

$$X_{(s)} = X \cdot D_{(s)}^{-1}$$

$$D_{(s)} = \mathrm{diag}\left(\|x_{[0]}\|, \ldots, \|x_{[k]}\|\right)$$

Since the sum of eigenvalues is equal to the trace, and each diagonal element of $X'_{(s)}X_{(s)}$ is 1.

$$\sum_{j=0}^{k} \lambda_j = tr\left(X'_{(s)} X_{(s)}\right) = k+1.$$

where λ_j's are the eigenvalues of $X'_{(s)}X_{(s)}$.

A method of judging the size of one eigenvalue in relation to the others is through the use of the condition number η_j.

$$\eta_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}}$$

An eigenvalue with $\eta_j > 30$ be flagged for further examination.

Variance Components.

If we wish to go further than the mere detection of multicollinearity to determine which linear combinations of columns of X are causing it, we can use the eigenvectors. A more frequently used approach involves the variance of the coefficients of x_j's. $(Var(\hat{\beta}_j^{(s)}))$

$$\hat{\beta}_{(s)} = D_{(s)} \hat{\beta}$$

$\beta_j^{(s)}$ is the j^{th} element of $\hat{\beta}_{(s)}$.

$$X'_{(s)} X_{(s)} = \Gamma D_\lambda \Gamma'$$

$$D_\lambda = diag(\lambda_0, \lambda_1, \dots \lambda_K)$$

λ_i : eigenvalues of $X'_{(s)} X_{(s)}$

$$\Gamma = \begin{bmatrix} \gamma_{00} & \cdots & \gamma_{0K} \\ \vdots & & \\ \gamma_{K0} & \cdots & \gamma_{KK} \end{bmatrix}$$ is an orthogonal matrix.

$$\text{cov}(\hat{\beta}_{(s)}) = \sigma^2 (X'_{(s)} X_{(s)}) = \sigma^2 \Gamma D_\lambda^{-1} \Gamma'$$

$$\text{var}(\hat{\beta}_j^{(s)}) = \sigma^2 \sum_{l=0}^{K} \lambda_l^{-1} \gamma_{jl}^2$$

$\lambda_l^{-1} \gamma_{jl}^2$ are called components of variance $\hat{\beta}_j^{(s)}$

$\phi_{lj} = \dfrac{\lambda_l^{-1} \gamma_{jl}^2}{\sum_{l=0}^{K} \lambda^{-1} \gamma_{jl}^2}$ is called the proportion of variance of jth coefficient $\hat{\beta}_j^{(s)}$ corresponding to the l^{th} eigenvector.

ϕ_{lj} and $\phi_{lK} > 0.5$ means a strong collinearity between variables x_j and x_K for l^{th} component of $X_{(s)}$

$$\sum_{l=0}^{K} \phi_{lj} = 1.$$

Interpretation of SAS Results.

Collinearity Diagnostics

	Eigenvalues (λ_j)	Condition index (η_j)	proportion of variation (ϕ_{lj}) x_1 x_2 $\cdots$ x_j
λ_{max}	λ_{max}	1	$\cdots$
$\vdots$			
λ_{min}	λ_j	$\sqrt{\dfrac{\lambda_{max}}{\lambda_j}}$	

> 30

$\Rightarrow$ problem of multicollinearity.

> 0.5

$\Rightarrow$ 1^{th} and j^{th} collinearity.

Remedies are proposed when multicollinearity is detected:

(1) Remove some independent variables.

(2) Ridge Regression

(3) Principal Components Regression (PCR)

(4) Partial Least Squares (PLS) Regression.

Ridge Regression.

Ridge regression is applied to the centered and scaled model. $X'X$ and $X'X+\Phi I$ have the same eigenvectors, but different eigenvalues (λ_j and $\lambda_j+\Phi$).

Estimate of β : $\hat{\beta}_\phi = (X'X+\Phi I)^{-1}X'y$

$$= (X'X+\Phi I)^{-1}X'(X\beta+\varepsilon)$$

$$= (X'X+\Phi I)^{-1}X'X\beta + (X'X+\Phi I)^{-1}X'\varepsilon.$$

$$\hat{\beta}_\phi = (Z'Z+\phi I)^{-1}Z'y \quad *$$

$$= (Z'Z+\phi I)^{-1}Z'Z\beta + (Z'Z+\Phi I)^{-1}Z'\varepsilon.$$

$$Z_j = \frac{x_j - \bar{x}_j \cdot \mathbf{1}}{S_{x_j}}$$

mean of x_j

standard error of x_j

Properties of estimates of ridge.

(1) $\hat{\beta}_\phi$ is biased.

$$\text{bias}(\hat{\beta}_\phi) = -\phi (Z'Z + \Phi I)^{-1} \beta$$

(2) $\text{Cov}(\hat{\beta}_\phi) = \sigma^2 (Z'Z + \Phi I)^{-1} Z'Z (Z'Z + \Phi I)^{-1}$.

(3) the sum of variances of components:

$$\text{tr}(\text{Cov}(\hat{\beta}_\phi)) = \sigma^2 \sum_{j=1}^{k} \lambda_j (\lambda_j + \Phi)^{-2}$$

(4) $\text{tr}[\text{Cov}(\tilde{\beta})] > \text{tr}[\text{Cov}(\hat{\beta}_\phi)]$

where $\tilde{\beta} = (Z'Z)^{-1} Z'y$.

(5) $$\text{tr}[\text{MSE}(\hat{\beta}_\phi)] = \sigma^2 \sum_{j=1}^{K} \lambda_j (\lambda_j + \Phi)^{-2} + \Phi^2 \beta' (Z'Z + \Phi I)^{-2} \beta$$

*: $\text{MSE}(\hat{\beta}_\phi) = \text{Cov}(\hat{\beta}_\phi) + \text{Bias}(\hat{\beta}_\phi)[\text{Bias}(\hat{\beta}_\phi)]'$

Determination of ϕ

(1) **Graphical method.** It consists of plotting a graph of the evolution of the coefficients of $\hat{\beta}_\phi$ as function of ϕ. The value $\hat{\phi}$ of ϕ is then chosen as (approximately) the smallest value at which the coefficients stabilize.

(2) **Analytical method.**

$$\tilde{\phi} = \frac{k\,\tilde{\sigma}^2}{\tilde{\beta}'\tilde{\beta}}$$

$\tilde{\beta} = (Z'Z)^{-1}Z'y$.

$\tilde{\sigma}^2$: estimate of σ^2 obtained by OLS over (y, Z)

k : number of independent variables.