Multicollinearity

Exact Multicollinearity (1)

What if $[\mathbf{X}^T\mathbf{X}]^{-1}$ does not exist or $\det[\mathbf{X}^T\mathbf{X}]$ is zero? β cannot be estimated.

It is a data problem not the technique. Remember that **X** has been pre-selected.

If it really happens, get a new data set. Nothing else can be done.

Exact multicollinearity is a rare event. Easy to "fix".

Exact Multicollinearity (2)

How could Exact multicollinearity happen?

A column in **X** is exactly a multiple of one column, i.e., $X_{2i} = 2X_{3i}, \forall i = 1,..n$

or a column in **X** is equal to a linear combination of one or more other columns in **X**, i.e.,

$X_{2i} = 2X_{3i} - X_{4i} + 2.7X_{5i} - 1.2X_{6i}, \forall i = 1,..n$

==> Rank(**X**) <*K*.

Near Multicollinearity (1)

Real problem is the Near Multicollinearity.

How could it happen?

A column in **X** is almost a multiple of one column or almost equal to a linear combination of one or more other columns in **X**

=>Rank(X) is still *K* but $\det[\mathbf{X}^T\mathbf{X}]$ is almost zero.

Near Multicollinearity (2)

Results

Estimator of β's involved in the problem will have a very large standard error (bad accuracy) even though the hypothesis that these β's are zero has been rejected using an F-test with very high confidence (at very low significant level).

==> accept that the coeff. is zero but in fact it is not.

Near Multicollinearity (3)

For example, given that

$$X_{2i} \approx 2X_{3i} - X_{4i} + 2.7X_{5i} - 1.2X_{6i}, \forall i = 1,..n$$

it is likely that s.e.'s of $\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6$

are large. t-tests will accept that each parameter is zero but the F-test will reject

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

$$H_1 : \beta_2 \neq \beta_3 \neq \beta_4 \neq \beta_5 \neq \beta_6 \neq 0$$

Near Multicollinearity (4)

Symptoms

1) low t-stat for at least 2 β's but high R^2.

2) high correlation among X's

3) non-robust of t-stat when removing variables. t-stat changes in sign or magnitude or both.

Symptoms do not guarantee the existence or pattern of multicollinearity.

Near Multicollinearity (5)

Detection (identification)

1) Ratio of max. and min. eigenvalues of $\mathbf{X}^T\mathbf{X}$ is greater than 100

2) Auxiliary regression among X's has R^2 greater than 0.9

Near Multicollinearity (6)

Auxiliary Regressions

$$X_{1i} = \gamma_{12}X_{2i} + \gamma_{13}X_{3i} + \ldots + \gamma_{1K}X_{Ki} + \varepsilon_{1i}$$

$$X_{2i} = \gamma_{21}X_{1i} + \gamma_{23}X_{3i} + \ldots + \gamma_{2K}X_{Ki} + \varepsilon_{2i}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$X_{Ki} = \gamma_{K1}X_{1i} + \gamma_{K3}X_{3i} + \ldots + \gamma_{K,K-1}X_{K-1,i} + \varepsilon_{K-1,i}$$

High R^2 implies that the X on the left hand side is almost a linear combination of the remaining X's on the right-hand side.

Near Multicollinearity (7)

Remedies

1) neglect if no specific coefficient is of interest. Even with multicollinearity, no problem in prediction of Y. Why?

   Note that OLS is still BLUE.

2) get more data or get a new data set

3) change model, i.e., use log-log model

Near Multicollinearity (8)

Remedies (cont'd)

4) drop variables that cause the problem. Drop one-by-one the dependent X's that yield the highest auxiliary R^2 until no high auxiliary R^2. However, dropping a variable will create another problem. For example, the new estimator will be biased and inconsistent. See Model specification.

Near Multicollinearity (9)

Remedies (cont'd)

5) If all the variables have to be maintained in the model, no chance to improve on OLS. Try other methods, such as, Factor Analysis, Determinant Analysis, LISREL

Polynomial model (1)

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3$$
$$+ \ldots + \beta_P X_i^P + \varepsilon_{1i}$$

It is very easy to have a multicollinearity in this model due to high correlation between the independent variables because they are all related to the same variables

Polynomial model (2)

To alleviate the problem, the polynomial model should be re-written

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$$
$$+ \ldots + \beta_P x_i^P + \varepsilon_{1i}$$

where $x_i = X_i - \overline{X}$