L7: Multicollinearity

Feng Li
feng.li@cufe.edu.cn

School of Statistics and Mathematics
Central University of Finance and Economics

Example – Whats wrong with it?

- Assume we have this data

```
Y     X2      X3
--------------
 2     1       3
 5     4      12
 7     8      24
13    16      48
---------------
```

- We want to make a simple regression model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

- By applying the formula in Chapter 7,

$$\hat{\beta}_2 = \frac{\sum y_i x_{2i} \sum x_{3i}^2 - \sum y_i x_{3i} \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i}\right)^2} = \frac{110068 - 110068}{154476 - 154476} = \frac{0}{0}$$

- Something went wrong?

Multicollinearity

- Perfect multicollinearity: the **covariates** are exactly linear combined

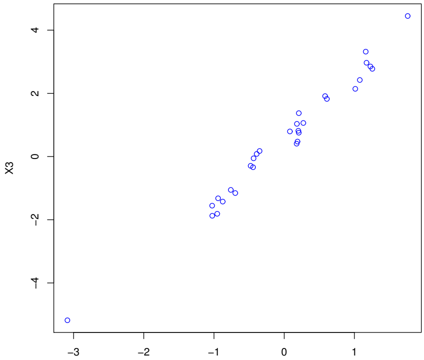
$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k = 0$$

e.g. $X_3 = 2X_2$.

- Less perfect multicollinearity (common in practice):

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k + v_i = 0$$

where v_i is some random values. E.g.



Estimation problems if multicollinearity in the covariates

- A perfect multicollinearity:
  - coefficients are indeterminate and
  - infinite large of standard errors for the coefficients.
- A less perfect multicollinearity:
  - coefficients are determinate but could not be estimated preciously and
  - very large of standard errors for the coefficients.

Sources of multicollinearity

- If we want to estimate how much electricity used in a family (Y) and we observe some variables might be used
  - X_1 : How big is the house
  - X_2 : How many people in this house
  - X_3 : How many rooms in the house
- Discussion with these models
  - $Y_i = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + u_i$
  - $Y_i = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_3 + u_i$

Practical consequences of high multicollinearity

- The OLS estimate is still BLUE (?) but with big variance.
- The OLS estimate can be very sensitive to a small change of data.
- Much bigger confidence interval $\Rightarrow$ easy to accept the null hypothesis.
- R^2 **very high but not significant** t **statistic**.

How to detect high multicollinearity (1)

- R^2 **very high but not significant** t **statistic**.
- Use **VIF** in the two-variable model
  - Assume r_{ij} is the coefficient of correlation between X_i and X_j. If X_i and X_j have collinearity problem, then $r_{ij} \to 1$.
  - Define **variance-inflating factor** (VIF) as

$$\text{VIF} = \frac{1}{1 - r_{ij}^2}$$

  - If there is no collinearity between X_i and X_j, $VIF = 1$.
  - If there is high collinearity between X_i and X_j, VIF is usually bigger than 10 and tends to ∞.
  - Recall the variance of estimate $\hat{\beta}_2$ in our first example

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} = \frac{\sigma^2}{\sum x_{2i}^2} \text{VIF}.$$

  - The inverse of VIF is called **tolerance** (TOL)

$$\text{TOL} = \frac{1}{VIF_j} = 1 - r_{ij}^2$$

  why does the last equality hold?

Use VIF in the k-variable model

- The variance of a coefficient in the model

$$var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2(1-R_j^2)} = \frac{\sigma^2}{\sum x_j^2}\mathrm{VIF}_j.$$

where R_j^2 is the R^2 for the **auxiliary regression** X_j with the remaining $k-1$ regressors

$$X_j = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{j-1}X_{j-1} + \beta_{j+1}X_{j+1} + ... + \beta_k X_k$$

- You can calculate VIF in two ways

$$\mathrm{VIF}_j = \frac{1}{1-R_j}$$

$$\mathrm{VIF}_j = \frac{var(\hat{\beta}_j)\sum x_j^2}{\hat{\sigma}^2}$$

- The $\mathrm{TOL}_j = 1/\mathrm{VIF}_j$
- Rule of thumb:
  if $VIF_j > 10$, indicating $R_j^2 > 0.9 \Rightarrow$ highly collinear of X_j.

how to detect high multicollinearity (2)

- high pair-wise correlations among regressors.
- auxiliary regression.
- the scatter plot.
- eigenvalues and conditional numbers of $X'X$
  - the basic idea: $X'X$ is invertible if there is not strong collinearity (all eigenvalues of $X'X$ are positive and in a reasonable range).
  - leading to the conditional number k

$$k = \frac{\text{Max eigenvalue}}{\text{Minimal eigenvalue}}$$

    and the conditional index $\sqrt{k}$

  - Rule of thumb:
    if $100 < k < 1000$, moderate to strong collinearity;
    if $k \geqslant 1000$, severe collinearity;
    if $0 < k < 100$, good

How to remedy multicollinearity problem?

- Drop a variable, usually firstly drop the most nonsignificant variable.
- Transform the variable.
- Do nothing if your purpose is prediction(see next slide).

Is multicollinearity always bad?

- The higher $\bar{R}^2$ the better prediction.
- So multicollinearity is not really a problem, if your purpose is prediction only.

Take home questions

10.10, 10.12, 10.19, 10.21