2. MULTICOLLINEARITY AND MISSING OBS

[1] MULTICOLLINEARITY

(1) Perfect Collinearity:

- When regressors are perfectly linearly related: The X matrix is less than full rank (Rank(X) < k).

Example 1:

$GDP_t = \beta_1 + \beta_2 G_t + \beta_3 T_t + \beta_4 DEF_t + \varepsilon_t$.

$\rightarrow DEF_t = G_t - T_t$ for any t.

$\rightarrow$ Rank(X) = 3.

Example 2: Dummy variable trap

$\log(w_t) = \beta_1 + \beta_2 age_t + \beta_3 d_{t1} + \beta_4 d_{t2} + \beta_5 d_{t3} + \varepsilon_t$,

where $d_{t1} = 1$ iff t's education level is lower than high school graduation ($d_{t1} = 0$, otherwise); $d_{t2} = 1$ iff person t is a high school graduate but not college graduate ($d_{t2} = 0$, otherwise); $d_{t3} = 1$ if person t is a college graduate ($d_{t3} = 0$, otherwise).

$\rightarrow d_{t1} + d_{t2} + d_{t3} = 1$.

- Consequence of perfect multicollinearity:
  - Cannot compute OLS estimates.

(2) Near Multicollinearity

- When regressors are highly (not perfectly) correlated.

1) Consequences of near multicollinearity on OLS estimators:
   - Under SIC, OLS estimators are unbiased, consistent and efficient.
   - Under WIC, OLS estimators are consistent and asymptotically normal and efficient.
   - Then, what is the problem?

     $\rightarrow$ Individual estimates are unreliable.

2) Symptoms of near multicollinearity:
   - Small changes in sample lead to large changes in estimates.
   - High R^2, but low t statistics.
   - High R_j^2 's (R_j^2 is R^2 from OLS of x_{tj} on $x_{t1}, \ldots, x_{t,j-1}, x_{t,j+1}, \ldots, x_{tk}$).
   - High value for $[\lambda_{max}/\lambda_{min}]^{1/2}$, where λ's are eigenvalues of $X'X$.

     (See Greene)
   - Estimates may be wildly different from those suggested by theory.
     - Think about a regression of consumption on one, income and wealth.

Question: Why does multicollinearity make values of t-statistic low?

Theorem:

x_j = j'th column of X;

X_j^* = X with j'th column deleted.

SSE_j = SSE from a regression of x_j on X_j^*.

$SST_j = \Sigma_t(x_{tj} - \bar{x}_j)^2$.

Then, the j'th diagonal of $(X'X)^{-1} = \frac{1}{SSE_j} = \frac{1}{SST_j(1-R_j^2)}$.

Implication:

- $\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}$ → As R_j^2↑, $\text{var}(\hat{\beta}_j)$↑.
- $\text{se}(\hat{\beta}_j) = \sqrt{\frac{s^2}{SST_j(1-R_j^2)}}$ → As R_j^2↑, $\text{se}(\hat{\beta}_j)$↑.
- (t statistic for H_o: $\beta_j = 0$) $= \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$ → As R_j^2↑, |t|↓.

3) Remedies

1. Drop some regressors highly correlated with others (?)
2. Collect a richer data set.
3. Use alternative estimators.

(3) Alternative Estimators

1) Ridge regression estimator:

$$\hat{\beta}_r = (X'X + rI_k)^{-1}X'y, \ r > 0.$$

- Biased but smaller MSE than OLS variances.

$$Cov(\hat{\beta}_r) = \sigma^2[X'X + rI_k]^{-1}X'X[X'X + rI_k]^{-1}.$$

- This estimator solves multicollinearity problem?
- No clear meaning to statistical inferences.
- What is optimal choice of r?

2) Principal Component Estimator.

- Procedure:
  - Compute eigenvalues of $X'X$ and sort them as $\lambda_{1.}\ \lambda_{2.}\ \ldots.\ \lambda_k$.
    (and $c_{k\times1}$ such that $X'Xc = \lambda c$.)
  - Choose normalized eigenvectors, i.e., $c_j'c_j = 1$.
  - Choose L largest λ's ($\lambda_1,\lambda_2,\ldots,\lambda_L$) and corresponding $c_1, \ldots, c_L$.
  - Define $C_L = [c_1, \ldots , c_L]$.
  - Let $Z = XC_L$, and do OLS on $y = Z\gamma + error$: $\hat{\gamma} = (Z'Z)^{-1}Z'y$.
  - Principal Component Estimator: $\hat{\beta}_{pc} = C_L\hat{\gamma}$.

- Facts:
  - $\hat{\beta}_{pc} = C_L C_L{}'\hat{\beta}$ (See Greene).
  - Biased unless L = k. (In fact, $\hat{\beta}_{pc} = \hat{\beta}$ if L = k.)
  - Sensitive to the scale of measurement on regressors.
  - No clear meaning to statistical inferences.

3) Suggestion:

Ridge regression or PC estimators could be useful for prediction, but may not be much helpful for statistical inferences.

(4) An ad hoc alternative (M is conquered?)

1) Situation: Consider the following regression model:

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \varepsilon_t .$$

Suppose that x_{t3} and x_{t4} are so highly correlated that the t-tests for β_2 and β_3 indicate insignificance of the two parameters.

<Example>

Dependent Variable: LWAGE
Sample: 1 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.517432	0.124819	44.20360	0.0000
EDUC	0.077987	0.006624	11.77291	0.0000
EXPER	0.016256	0.013540	1.200595	0.2302
EXPER^2	0.000152	0.000567	0.268133	0.7887
R-squared	0.130926	Mean dependent var		6.779004

2) Alternative regression:

Step 1: Regress x_{t3} on one and x_{t4}, and residuals u_t.

Step 2: Regress y_t on one, x_{t2}, u_t and x_{t4}.

<Example>

Dependent Variable: LWAGE
Included observations: 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.604536	0.101658	55.13115	0.0000
EDUC	0.077987	0.006624	11.77291	0.0000
U	0.016256	0.013540	1.200595	0.2302
EXPER^2	0.000812	0.000138	5.869204	0.0000
R-squared	0.130926	Mean dependent var		6.779004

Observe that $EXPER^2$ is now significant!!!

3) Logic of this treatment:

- Let $x_{t3} = \delta_1 + \delta_2 x_{t4} + u_t$. Substitute this into the original regression model:

$$\begin{aligned} y_t &= \beta_1 + \beta_2 x_{t2} + \beta_3(\delta_1 + \delta_2 x_{t4} + u_t) + \beta_4 x_{t4} + \varepsilon_t \\ &= (\beta_1 + \beta_3\delta_1) + \beta_2 x_{t2} + \beta_3 u_t + (\beta_4 + \delta_2\beta_3)x_{t4} + \varepsilon_t. \end{aligned}$$

  Since u_t and x_{t4} are uncorrelated, this alternative model does not suffer from M.

- Is it really?

3) Problems in this approach.

- Your estimate of β_4 is an estimate of $(\beta_4+\delta_2\beta_3)$, not of β_4!
- You can't use real u_t. Instead, you use the estimated u_t. When you estimated u_t, the OLS covariance matrix is no longer of the form $\sigma^2(X'X)^{-1}$. This problem is called "generated regressor" problem.

4) Conclusion: No magic! No solution!

[2] MISSING OBSERVATIONS

- Suppose that some values of y_t, x_{t1}, ..., x_{tk} are missing for some people.
- Is there any good way to fill up the missing values?
  - There might be. But may better not to do so.
  - See Greene Chapter 4.9.2.