1-2. More Issues in Linear Model

- Interpretation of coefficient
- Let $y_i = \alpha + \beta x_i + u_i, \quad \triangle log(x) = \triangle x / x$

Y	X	β	Interpretation
y	x	$\triangle y / \triangle x$	1 more $x \Longrightarrow \beta$ more y
$\log y$	x	$(\triangle y / y) / \triangle x$	1 more $x \Longrightarrow y$ increases by $(\beta \times 100)\%$
y	x	$\triangle y / (\triangle x / x)$	1% increase of $x \Longrightarrow \beta/100$ more y
$\log y$	$\log x$	$(\triangle y / y) / (\triangle x / x)$	1% increase of $x \Longrightarrow y$ increases by $\beta\%$

p-value and One-Sided Test

- p-value : the largest significance level at which we could carry out the test and still fail to reject the null hypothesis. (If p-value is small enough, we reject the null hypothesis. For example, at 5% significance level, if p-value is less than 0.05, we reject the null hypothesis.)

  $H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$ (two-sided)

  If $\widehat{\beta}_1$ is large or small enough, we reject H_0.

  $H_0 : \beta_1 = 0, \quad H_1 : \beta_1 > 0$ (one-sided)

  If $\widehat{\beta}_1$ is large enough, we reject H_0. If $\widehat{\beta}_1$ is small, doen not reject H_0.

- p-value (two-sided) $= 2\times$ p-value (one-sided): easier to reject H_0? : You need a proper reason to use an one-sided test.

Dummy Variable

- $\text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta 2\text{female}_i + u_i$
  where

  $\text{female}_i = 1$ if the person i is female

  $\text{female}_i = 0$ if the person i is male.

  $\text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{female}_i + \beta_3 \text{male}_i + u_i$

  $\Rightarrow \text{female}_i + \text{male}_i = 1$

  Perfect Collinearity

- For data with seasonality, one can use dummy variables (quaterly/monthly) if seasonally adjusted data are unavailable.

- If your dependent variable is a dummy variable? (see probit model)

How Good Is the Fitting?

- Standard error of the regression
- Coefficient of determination R^2 : a measure of 'Goodness of Fit'
  $Y = X\beta + U$

  Let $\widehat{Y} = X\widehat{\beta}$ and $\widehat{U} = Y - \widehat{Y}$
  Then $Y'Y = (\widehat{Y} + \widehat{U})'(\widehat{Y} + \widehat{U}) = \widehat{Y}'\widehat{Y} + \widehat{U}'\widehat{U}$
  because $\widehat{Y}'\widehat{U} = \widehat{U}'\widehat{Y} = 0$.
  $Y'Y$: total sum of squares (TSS)
  $\widehat{Y}'\widehat{Y}$: explained sum of squares (ESS)
  $\widehat{U}'\widehat{U}$: residual sum of squares (RSS)
  $R^2 = ESS/TSS = \frac{\widehat{Y}'\widehat{Y}}{Y'Y} = 1 - \frac{\widehat{U}'\widehat{U}}{Y'Y}$
  $\overline{R}^2 = 1 - \frac{\widehat{U}'\widehat{U}/(n-k-1)}{Y'Y/(n-1)}$

Projection

- V vector space,

  $P : V \longrightarrow V$ (linear transfromation)
  P is a projection if $P(P_X) = P_X$.

  If P is a projection, so is $I - P$.

  Orthogonal Projection : null space and range of P are orthogonal.

  Projection of Y on (the space spanned by) X is
  $P_X Y = X(X'X)^{-1}X'Y$

Two step regression

- Partitions of X

$$Y = X_1\beta_1 + X_2\beta_2 + u$$

$$\widehat{\beta}_2 = (X_2'(I - P_1)X_2)^{-1}X_2'(I - P_1)Y$$

$$\widehat{\beta}_1 = (X_1'X_1)^{-1}X_1'(Y - X_2\widehat{\beta}_2)$$

where $P_1 = X_1(X_1'X_1)^{-1}X_1'$

If $X_1'X_2 = 0$,

$$\widehat{\beta}_2 = (X_2'X_2)^{-1}X_2'Y$$

$$\widehat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y$$

Omitted Variable

- Exclusion of a relevant variable W

  True Model : $Y = W\alpha + X\beta + U$

  Estimate $Y = X\beta + U$

  $$\widetilde{\beta}_{OLS} = (X'X)^{-1}X'Y$$

  $$= \beta + (X'X)^{-1}X'(W\alpha + U)$$

  Unless $X'W = 0$, $E(\widetilde{\beta}_{OLS}) \neq \beta$

  $\Rightarrow$ Omitted variable bias

Redundant Variable

- Inclusion of an irrelevant variable W

  True Model : $Y = X\beta + U$

  Estimate $Y = W\alpha + X\beta + U$, $E(U|W) = 0$

  $$\overline{\beta}_{OLS} = (X'(I - P_W)X)^{-1}X'(I - P_W)Y$$

  $$= \beta + (X'(I - P_W)X)^{-1}X'(I - P_W)U$$

  where $P_W = W(W'W)^{-1}W'$

  $\Rightarrow$ Unbiased but Not Efficient

Multicollinearity

- Multicollinearity is different from perfect collinearity or perfect multicollinearity. Consider

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

- Let $\frac{1}{n}X'X = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, i.e., we assume $var(x_{1i}) = var(x_{1i}) = 1$ and $\rho = corr(x_{1i}, x_{2i})$. In this case,

$$var\left(\widehat{\beta}_{OLS}\right) = \frac{\sigma^2}{n}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} = \frac{\sigma^2}{n(1-\rho^2)}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$$

- If $|\rho|$ is closer to one $\Rightarrow$ large standard error
- If sample size n is very small, we have the same problem. (better to have a large n).
- If two explanatory variables are highly correlated, they contain similar information and, therefore, an easy solution is to remove one variable.