Econometrics Note 1

2024年7月11日 14:54

Note 1

2021年12月13日 16:00

L_1

1) Formal Econometric model

需求的经济模型：

$$quantity = f(price,\ income)$$

计量模型：

$$quantity_{it} = \beta_0 + \beta_1 price_t + \beta_2 income_{it} + u_{it}$$

未知参数

误差项(无明显关系的其它项)

2) Hypothesis

e.g. Returns to Education

$$wages_i = \beta_0 + \beta_1 educ_i + \beta_2 IQ_i + u_i$$

如：educ 与 wages 无关：$\beta_1 = 0$

IQ 对 wage 有正的影响：$\beta_2 > 0$

L_2 一元线性回归

1) $y_i = \beta_0 + \beta_1 x_i + u_i$

$i = 1, \ldots, n$；每一个(y_i, x_i)与其它的相互独立

β_1: slope parameter；β_0: intercept parameter

Assume: · on average the error term is zero: $E(u_i) = 0$

(意为u与x之间无任何关系)(误差项不会对函数造成影响)

① $\Delta y = \beta_1 \Delta x$ if $\Delta u = 0$

一但①成立，$E(u_i) = 0$

· more importantly, $E(u_i | x_i) = 0$

2) OLS Estimator

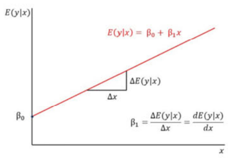
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \ ; \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$$

fitted values from LR (线性回归)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

residuals (残差)

$$\hat{u}_i = y_i - \hat{y}_i$$
$$= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$



3) Gauss-Markov Assumption (高斯-马尔科夫假设)

SLR.1 (模型的变量是线性的)

$$y = \beta_0 + \beta_1 x + u$$

SLR.2 (自变量为随机数据)

$$(x_i, y_i): i = 1, \ldots, n$$

SLR.3 (解释变量有样本差异) (解释变量间无完全共线性)

SLR.4 (零均值) (ZCM)

$E(u_i | x_i) = 0 \longrightarrow E(y|x) = \beta_0 + \beta_1 x$

Population Regression Function (PRF) (总体回归函数)

· 反映y的均值随x的变化，非在所有单位下y等于$\beta_0 + \beta_1 x$

(for all units in the population)

(是线性的意味着x改变1个单位，y的期望值更改β_1)

SLR.5 (误差项u与其它任何x的值有相同的方差)

(OLS估计量的线性性)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{\bar{y} \sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \sum \left[\frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] y_i$$

令 $\lambda = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$

$\therefore \hat{\beta}_1 = \sum \lambda_i y_i$

$\therefore \hat{\beta}_1$是解释变量y_i的线性组合

$$\sum_{i=1}^n \lambda_i = \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0$$

$$\sum_{i=1}^n \lambda_i (x_i - \bar{x}) = \sum_{i=1}^n \lambda_i x_i - \bar{x} \sum_{i=1}^n \lambda_i = \sum_{i=1}^n \lambda_i x_i = 1$$

$$\sum_{i=1}^n \lambda_i^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

4) Unbiasedness of OLS (无偏性)

· $\hat{\beta}_1$ is estimated from data

($\hat{\beta}_1$是对数据的一个预测；不同数据有不同的$\hat{\beta}_1$)

· $\hat{\beta}_1$ is a random variable

($\hat{\beta}_1$是随机变量；它是随机误差项的函数)

$$E[\hat{\beta}_1] = \beta_1 ?$$

用预测的β_1所求的期望

（β_1 是随机变量；它是随机误差项的函数）

$E[\hat{\beta}_1] = \beta_1$?

用预测的 β_1 所求的期望

5) OLS estimator

如果均符合高斯-马尔科夫假设（SLR.1 - SLR.4），则OLS预测量为自变量系数的无偏估计量）

$$E(\hat{\beta}_0) = \beta_0$$
$$E(\hat{\beta}_1) = \beta_1$$

（OLS估计量的无偏性）

$$\hat{\beta}_1 = \sum_{i=1}^n \lambda_i y_i$$
$$= \sum_{i=1}^n \lambda_i (\beta_0 + \beta_1 x_i + u_i)$$
$$= \beta_0 \sum_{i=1}^n \lambda_i + \beta_1 \sum_{i=1}^n \lambda_i x_i + \sum_{i=1}^n \lambda_i u_i$$
$$= \beta_1 + \sum_{i=1}^n \lambda_i u_i$$
$$E(\hat{\beta}_1) = E(\beta_1 + \sum_{i=1}^n \lambda_i u_i)$$
$$= \beta_1 + \sum_{i=1}^n \lambda_i E(u_i)$$
$$= \beta_1 + 0$$
$$= \beta_1$$

Assume Homoskedasticity（同方差性）

·SLR.5 u 和其它 x 有相同的方差

$$Var(u|x) = \sigma^2$$

（SLR.1 ~ SLR.5）

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}$$

e.g. $y = \beta_0 + \beta_1 x + u$（$x$ 1~n的数减 x 的平均值）（y 1~n的数减 y 的平均值）

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad ; \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{Cov(x, y)}{Var(x)}$$

（x 1~n的数减 x 的平均值的平方）

Fitted model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$（OLS regression line / Sample regression function）

Interpetation: $\Delta \hat{y} = \hat{\beta}_1 \Delta x$

Residuals: $\hat{u}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

Properties:

Sample counterparts

$$\frac{1}{n}\sum_{i=1}^n \hat{u}_i = 0$$ （误差平均值为0；$E(u) = 0$）

$$\frac{1}{n}\sum_{i=1}^n x_i \hat{u}_i = 0$$ （零均值；$E(u|x) = 0$）

OLS将 y 分解为拟合值和残差：

$$SST = SSE + SSR$$

Total Sum of Squares (SST)（总偏差平方和（反映全部数据误差大小的平方和））

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

Explained Sum of Squares (SSE)（回归平方和（因变量回归值 $\hat{y}$，因变量平均值 y 的离差平方和））

$$SSE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Residual Sum of Squares（残差平方和（衡量变量拟合程度））

$$SSR = \sum_{i=1}^n \hat{u}_i^2$$

（如果预测结果相近，SSE与SST几近相同）

6) 拟合程度（Coefficient of determination / R - squard）

$$R^2 \equiv \underline{\frac{SSE}{SST}} = 1 - \frac{SSR}{SST}$$

$$\frac{SST - SSR}{SST}$$

（拟合，SST=>SSE，SSR=>0，所以 R^2 =>1）

7) OLS中用虚拟变量（dummy variable）

（虚拟变量用0或1代表可量化的观测值）

$$y_i = \beta_0 + \beta_1 x_i + u_i$$ （✓，$x=1$；X，$x=0$）

Interpretation:

$$\hat{\beta}_1 = E(y_i | x_i = 1) - E(y_i | x_i = 0)$$

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad (\checkmark, x=1;\ \times, x=0)$$

Interpretation:

$$\hat{\beta}_1 = E(y_i \mid x_i = 1) - E(y_i \mid x_i = 0)$$

8) 回归计算中用对数（logarithms）

（当变量中有大量变异时使用）

$$\log(y_i) = \beta_0 + \beta_1 x_i + u_i$$

TABLE 2.3 Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

L3 多元线性回归

1) $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$

（从 $i=1 \sim n$，每个观测值 $(y_i, x_{i1}, \ldots, x_{ki})$ 彼此独立）

依旧假设：1、$E(u_i) = 0$

2. $E(u_i \mid x_{1i}, \ldots, x_{ki}) = 0$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

$\hat{\beta}_0$，当 $x_1 = x_2 = 0$ 时 y 的预测值

$\hat{\beta}_1$，x_1 对 y 的部分影响（partial effect），是 x_2 的条件影响

（x_1 对 y 的影响是在 x_2 不变下，ceteris paribus effect）

$\hat{\beta}_2$，x_2 对 y 的部分影响（partial effect）

2) 多元回归下的OLS预测值

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

当 $\hat{r}_i$ 是估计残差时，x_2 与 x_1 之间无任何影响（x_1 无法解释 x_2）

$$x_{1i} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{2i} + \hat{r}_{1i}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{1i} y_i}{\sum_{i=1}^{n} \hat{r}_{1i}^2}$$

（当 x_1，x_2 和 y 不变下的 $\hat{\beta}_1$）

3) 矩阵中的OLS预测值

$$Y = X\beta + u$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

（OLS估计量的无偏性）

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta+u) \\
&= (X'X)^{-1}X'(X\beta+u) \\
&= \beta + (X'X)^{-1}X'u
\end{aligned}$$

$$\begin{aligned}
E(\hat{\beta} \mid X) &= E[\beta + (X'X)^{-1}X'u \mid X] \\
&= \beta + (X'X)^{-1}X'E(u \mid X) \\
&= \beta + 0 \\
&= \beta
\end{aligned}$$

$$E(\hat{\beta}) = E_X E(\hat{\beta} \mid X) = E_X(\beta) = \beta$$

4) 高斯马尔科夫假设

$$\hat{\beta} = \underbrace{(X'X)^{-1}X'}_{\text{看作为矩阵系数}}y$$

则 $\hat{\beta}$ 与 y 为线性组合

MRL.1 总体模型是线性的

$$Y = X\beta + u$$

MRL.2 数据是总体中的随机值 $(y_i, x_{1i}, \ldots x_{ki})$

MRL.3 X无共线性

MRL.4 Zero Condition Mean (ZCM)

$$E(u \mid x_1, x_2, \ldots, x_k) = 0$$

从MRL.1～4得出：$E(\hat{\beta}) = \beta$

MRL.5 Homoskedasticity（u 和其它任意 $(x_1, x_2, \ldots x_k)$ 有相同方差）

$$Var(u \mid x_1, x_2, \ldots x_k) = \sigma^2$$

5) Variance of OLS estimator

$$Var(\hat{\beta} \mid X) = \sigma^2 (X'X)^{-1}$$

6) Estimating the Standard Error of the regression

$\sigma^2 = E[u^2]$

(不观测 u)

$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} = \frac{SSR}{n-k-1}$

(自由度, n = 观测数量, k+1 是预测变量)

(!!! 自由度有时为 n-k, 取决于k的定义)

7) ZCM

$E(u|x_1, x_2, \ldots, x_k) = 0$

当满足ZCM时,我们说我们有 exogenous explanatory variables

如果:

$E[u|x_j] \neq 0$

我们说 x_j 是 endogenous explanatory variable

(这是因果推断的关键, 如有 omitted variables 则易被违反)

8) 矩阵

$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + u_i$

(有N个关于变量的观测值)

$y = X\beta + \epsilon$

(y是N×1, β是(p+1)×1 变量的向量, X是N×(p+1)观测值的矩阵, ε是 N×1误差的矩阵)

9) 假设:

$E(\epsilon) = 0_N$

$E(\epsilon\epsilon') = \sigma^2 I_N$

X is fixed in repeated samples

X has rank less than N

(所以我们假设误差项的期望为0, 同方差, X不变, 观测值比变量多)

10) 向量 β, $\hat{\beta} = [\hat{\beta}_0, \ldots, \hat{\beta}_p]$

a residual column vector of e:

(残差向量) $e = y - X\hat{\beta}$ (矩阵X) (向量β的预测值)

(再将 β̂ 残差的平方和最小化,便可定义为数字 e'e (缩小为1×1矩阵))

如下:

(表示是e的转置) $e'e = (y - X\hat{\beta})'(y - X\hat{\beta})$

$= (y' - \hat{\beta}'X')(y - X\hat{\beta})$

$= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$

(化简)

(皆可转置) $\hat{\beta}'X'y = y'X\hat{\beta}$

$\therefore\ e'e = y'y - \hat{\beta}'X'y - \hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$

$e'e = y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$ ← (我们所需最小化的关于变量的矩阵的表达式)

(再找其最小)

(矩阵求导法则) 如 $f(x) = a'X$; a是 q×1向量, $\frac{\delta f(x)}{\delta a} = X$

如 $f(x) = a'Xa$; X是对称矩阵, $\frac{\delta f(x)}{\delta a} = 2Xa$

$\frac{\delta(e'e)}{\delta\hat{\beta}} = -2X'y + 2X'X\hat{\beta}$

(令其=0找β̂) $-2X'y + 2X'X\hat{\beta} = 0$

$-X'y + X'X\hat{\beta} = 0$

$X'X\hat{\beta} = X'y$

$\hat{\beta} = (X'X)^{-1}X'y$

(将 $y = X\beta + \epsilon$ 代入) $\hat{\beta} = (X'X)^{-1}X'(X\beta + \epsilon)$

$= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon$

$= \beta + (X'X)^{-1}X'\epsilon$

（两边互取期望） $E[\hat{\beta}] = E[\beta + (X'X)^{-1}X'\epsilon]$

$$= E[\beta] + E[(X'X)^{-1}X'\epsilon]$$
$$= \beta + (X'X)^{-1}X'E[\epsilon]$$
$$= \beta$$

结论：所找寻的预测变量的期望与变量为真实值（无偏 unbiased）

（如果为伪，则期望应为 $\beta+\delta$，我们对 δ 不够了解，不足以使用）

11）判断估测的准确性

（预测变量的期望与变量为真值，但实际值未必为真实值）

（解答 actual value = true value ？）

$$Var(\hat{\beta}) = E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))']$$

公式：$V(z) = E(z - E(z))^2$

$$= E[(\beta + (X'X)^{-1}X'\epsilon - \beta)(\beta + (X'X)^{-1}X'\epsilon - \beta)']$$

代入 $\begin{cases} \hat{\beta} = \beta + (X'X)^{-1}X'\epsilon \\ E[\hat{\beta}] = \beta \end{cases}$

$$= E[((X'X)^{-1}X'\epsilon)((X'X)^{-1}X'\epsilon)']$$
$$= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}]$$
$$= (X'X)^{-1}X'E[\epsilon\epsilon']X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}X'I_{p+1}X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}$$

（方差为 方差、协方差矩阵；我们没有所需的一切，因此要从数据中估计出 σ^2）

（所得出的 OLS 估计量为 BLUE（Best Linear Unbiased Estimator））

（这代表我们可以证明估计值在所有无偏的估计量中是差异最小的）

证明：

$$\hat{\beta} = (X'X)^{-1}X'y$$

$(X'X)^{-1}X'$ 是 $(p+1)\times N$ 的矩阵，每一个 $p+1$ 皆为观测值的平均权重

$\therefore\ W = (X'X)^{-1}X' + Q$

（非 $(p+1)\times N$ 矩阵但占有权重）

12）估测值的无偏性

新预测值： $\tilde{\beta} = Hy$

代入 $\begin{cases} y = X\beta + \epsilon \\ H = ((X'X)^{-1}X' + Q)(X\beta + \epsilon) \end{cases}$

$$\tilde{\beta} = ((X'X)^{-1}X' + Q)(X\beta + \epsilon)$$
$$\tilde{\beta} = (X'X)^{-1}X'X\beta + QX\beta + (X'X)^{-1}X'\epsilon + Q\epsilon$$
$$\tilde{\beta} = \beta + QX\beta + (X'X)^{-1}X'\epsilon + Q\epsilon$$

要知道 $E[\tilde{\beta}] = \beta$，需要 $QX = X'Q' = 0_{(p+1)\times N}$，（任何矩阵均满足线性及无偏）

13）估测值的方差

$$Var(\tilde{\beta}) = E[(\tilde{\beta} - E[\tilde{\beta}])(\tilde{\beta} - E[\tilde{\beta}])']$$
$$= E[(\beta + (X'X)^{-1}X'\epsilon + Q\epsilon - \beta)(\beta + (X'X)^{-1}X'\epsilon + Q\epsilon - \beta)']$$
$$= E[((X'X)^{-1}X'\epsilon + Q\epsilon)((X'X)^{-1}X'\epsilon + Q\epsilon)']$$
$$= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1} + (X'X)^{-1}X'\epsilon\epsilon'Q' + Q\epsilon\epsilon'X(X'X)^{-1} + Q\epsilon\epsilon'Q']$$
$$= \sigma^2((X'X)^{-1} + (X'X)^{-1}X'Q' + QX(X'X)^{-1} + QQ')$$
$$= \sigma^2((X'X)^{-1} + QQ')$$

$$QX = Q'X' = 0_{(p+1)\times N}$$

$$Var(\tilde{\beta}) = \sigma^2((X'X)^{-1} + QQ') = Var(\hat{\beta}) + \sigma^2 QQ'$$

$\because QQ'$ 的定义为 positive-semi definite matrix

$\therefore$ diagonal elements 是非负的（non-negative）

新估计量比 $(X'X)^{-1}X'y$ 效率低

14）小结：

1. 零误差期望假设下是无偏的
2. 在误差方差一致假设下是有效的
3. 当综合在一起时得出高斯-马尔科夫理论
4. 高斯-马尔科夫理论下说明变量向量 β 是 BLUE

15）Omitted Variable Bias（遗漏变量偏差）

（设定模型时遗漏了相关解释变量／忽略了某些对解释变量有趋势性影响的因素）

总体模型：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

遗漏 x_2：

如 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ （u 满足 $E(u)=0$）

忽略 x_3： $y = \beta_0' + \beta_1' x_1 + \beta_2' x_2 + u'$

误差项： $u' = (\beta_0 - \beta_0') + (\beta_1 - \beta_1')x_1 + (\beta_2 - \beta_2')x_2 + \beta_3 x_3 + u$

满足： $E(u') = (\beta_0 - \beta_0') + (\beta_1 - \beta_1')x_1 + (\beta_2 - \beta_2')x_2 + \beta_3 x_3$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

遗漏 x_2：

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

依模型得出解释变量关系：

$$x_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1$$

因此：

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$$

（有偏估计量）

$$E[\tilde{\beta}_1] = E[\hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1]$$
$$= E[\hat{\beta}_1] + E[\hat{\beta}_2]\tilde{\delta}_1 \quad \text{（假设}\tilde{\delta}_1\text{是固定的）}$$
$$= \beta_1 + \beta_2 \tilde{\delta}_1$$

$$\therefore \quad Bias(\tilde{\beta}_1) = E[\tilde{\beta}_1] - \beta_1 = \beta_2 \tilde{\delta}_1$$

TABLE 3.2 Summary of Bias in $\tilde{\beta}_1$ when x_2 Is Omitted in Estimating Eqution (3.40)

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

6) Multicollinearity（多重共线性）

（当一些解释变量高度相关时）

OLS斜率估计量的样本方差：

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}$$

拟合值与残差为同种方法定义的.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}$$
$$\hat{u}_i = y_i - \hat{y}_i$$

残差的样本平均值为0（$\Rightarrow \bar{y} = \bar{\hat{y}}$）

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

每个解释变量和残差间的样本协方差为0

$$\sum_{i=1}^{n} \hat{u}_i x_{ji} = 0 \qquad \text{for } j = 1, \ldots, k$$

7) 拟合程度

R^2：由所有解释变量解释的y样本变化的分数

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

（当添加解释变量时，R^2总会增加，R^2并非比较嵌套模型(nested model)的好工具）

∴ 用调整的R^2（Adjusted R-squared）

$$\bar{R}^2 \equiv 1 - \frac{\frac{SSR}{(n-k-1)}}{\frac{SST}{(n-1)}}$$

L4 Inference

1) 高斯－马尔科夫理论

当MLR.1 － MLR.5 成立时，OLS估计量 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ 是 $\beta_0, \beta_1, \ldots \beta_k$ 的最优线性无偏估计量（BLUE）

但为了统计推断，还需要误差项 $\hat{\beta}_j$ 的总体布局

（目前以知 $\hat{\beta}_j$ 的平均值和方差）

MLR. 6 总体误差项 u 与 X 独立，且与均值为0和方差σ^2服从正态分布

$$u \sim N(0, \sigma^2)$$

（高斯－马尔科夫及正态假设统称为经典线性模型(CLM)假设）

2) 样本分布

（在CLM假设下，并以回归量的样本值为条件）

$$\hat{\beta}_j \sim Normal(\beta_j, Var(\hat{\beta}_j))$$

说明：

$$\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim Normal(0, 1)$$

当将 $sd(\hat{\beta}_j)$ 里的σ^2和 $se(\hat{\beta}_j)$ 里的σ^2替换：

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1} = t_{df}$$

（$k+1$ 是总体模型 $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ 中未知参数的个数

$n-k-1$ 是自由度（df））

误差项：$u' = (\beta_0 - \beta_0') + (\beta_1 - \beta_1')x_1 + (\beta_2 - \beta_2')x_2 + \beta_3 x_3 + u$

满足：$E(u') = (\beta_0 - \beta_0') + (\beta_1 - \beta_1')x_1 + (\beta_2 - \beta_2')x_2 + \beta_3 x_3$

· 因 x_1, x_2 和 x_3 间无线性关系，$\beta_0, \beta_1, \beta_3$ 和 $\beta_0', \beta_1', \beta_2'$ 取任何值 $E(u')$ 不可能都为0. 故必然违反ZCM. 所以OLS估计量常常有偏与非一致性

3）假设检验

如：$Wage_i = \beta_0 + \beta_1 female_i + \beta_2 exper_i + u_i$

OLS估计：

$\widehat{Wage}_i = \hat{\beta}_0 + \hat{\beta}_1 female_i + \hat{\beta}_2 exper_i$

（想测试 experience 对 Wage 有什么影响时，先控制
其它所有的回归量（如：female））

做零假设：$H_0: \beta_2 = 0$

（一但除 x_2 以外的独立变量被计算在内，exper（x_2）对 Wage（y）的期望没有影响）

（不能理解为 exper（x_2）会对 wage（y）所产生的部分影响，因为只有当 β_2 是其它非零值时）

t统计量（t-statistic）

为了检验零假设：$H_0: \beta_2 = 0$

公式：

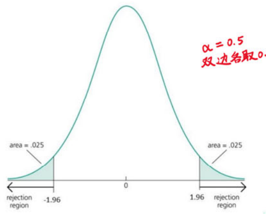
$$t \equiv \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)}$$

1. t和 $\hat{\beta}_2$ 同正负
2. $\hat{\beta}_2$ 越大，t-stat 越大
3. se 越大，t-stat 越小

假设检验：双边检验

零假设：$H_0: \beta_2 = 0$ （找证据拒绝零假设）

供选假设：$H_1: \beta_2 \neq 0$



$\alpha = 0.5$
双边各取0.25

先明确显著水平（significant level）：α

· 一类误差（拒绝 H_0 为真）的概率

（通常选取 5%，不过 1% 或 10% 也很常见）

（为了找临界值（critical value（c.v.）），需要当 H_0 为真时，样本的t统计分布，t

$$t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \sim t_{df}$$

（通过统计表找 c.v.）

（为了算t统计量，要从数据中估计出当零假设 $H_0: \beta_2 = 0$ 为真时的估计量 $\hat{\beta}_2$）

$$t = \frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = [一个数]$$

（$t > c.v.$，拒绝 H_0）

（c.v 在自由度为 n-k-1的t分布中占97.5%）

假设检验：单边检验

供选假设：$H_1: \beta_2 > 0$

如：OLS估计量：$\widehat{wage}_i = \hat{\beta}_0 + \hat{\beta}_1 female_i + \hat{\beta}_2 experience_i$

（实际上，零假设为 $H_0: \beta_2 \leq 0$，但如果我们证明
加教育经验，我们只关心 β_2 是非零的其它正数（
教育经验没有负的）也就是当拒绝 $\beta_2 = 0$，$\beta_2 < 0$。所以
$H_0: \beta_2 = 0$ VS. $H_1: \beta_2 > 0$）

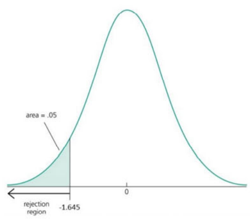
（探究女性平均工资是否比男性少$2.00

$H_0: \beta_1 \geq -2$ VS. $H_1: \beta_1 < -2$

（β_1 是性别变量的参数）

（探究女性平均人数是否少于男性）

$H_0: \beta_1 \geq 0$ VS. $H_1: \beta_1 < 0$



如：$\widehat{wage}_i = \underset{(0.191)}{9.71} - \underset{(0.175)}{2.32}\, female_i + \underset{(0.008)}{0.08}\, exper_i$

s.e（standard errors）

当 $H_0: \beta_1 \geq -2$ 为真，CLM假设吻合

$$\frac{\hat{\beta}_1 - (-2)}{se(\hat{\beta}_1)} \sim t_{df}$$

c.v = -1.645（当 Df 为 ∞，α 为 0.05，查表可知）

$$t = \frac{\hat{\beta}_1 - (-2)}{se(\hat{\beta}_1)} = \frac{-2.32-(-2)}{0.175} = -1.829$$

$-1.829 < c.v(-1.645)$

（接受零假设，拒绝供选假设）

结论：女性平均工资确实比男性少$2（或更多）

统计显著性依靠：$t_{\hat{\beta}_j}$

经济显著性依靠：$\hat{\beta}_j$

(大的样本量用小的显著性水平(如1%)

小的样本量用大的显著性水平)

(可表示统计意义要么是 $\hat{\beta}_j$ 大；要么 $se(\hat{\beta}_j)$ 小

区分t统计很重要，不然导致错误结论(如变量对

解释y很重要，虽然其估计影响不大))

4) 区间估计(置信区间)

用定值估计(Point Estimates)和标准差来行成(1-α)%区间估计

$$P\left[-t_{(1-\frac{\alpha}{2},df)} < t < t_{(1-\frac{\alpha}{2},df)}\right] = 1-\alpha$$

(如：95%区间，β_1 $df \to \infty$)

$$\hat{\beta}_1 \pm 1.96 \times se(\hat{\beta}_1)$$

$$1.44 \pm 1.96 \times 0.0859$$

(如果分成100个区间，95个会包括真实参数 β_1)

5) P 值

(在t统计检验下，显著性水平最小足够使得我们拒绝零假设)

$$P(|T| > |t|)$$

决定法则：

当 p值 $< \alpha$，拒绝 H_0

6) 测试单一线性组合

$$\log(donation) = \beta_0 + \beta_1 \log(txhw) + \beta_2 \log(trhw) + \beta_3 sexhd + \beta_4 eduhd + \beta_5 eduwf + u$$

(探究户主和妻子的教育程度是否对慈善捐款有着一致的影响)

$$H_0: \beta_4 = \beta_5$$

测试统计量：

$$t = \frac{\hat{\beta}_4 - \hat{\beta}_5}{se(\hat{\beta}_4 - \hat{\beta}_5)}$$

当：$s.e.(\hat{\beta}_4 - \hat{\beta}_5) = \sqrt{Var(\hat{\beta}_4) + Var(\hat{\beta}_5) - 2cov(\hat{\beta}_4, \hat{\beta}_5)}$

(将问题重新参数化)，令 $\theta = \hat{\beta}_4 - \hat{\beta}_5$，则 $H_0: \theta = 0$)

$$H_0: \theta = 0 \quad VS. \quad H_1: \theta < 0$$

$$\therefore \log(donation) = \beta_0 + \beta_1 \log(txhw) + \beta_2 \log(trhw) + \beta_3 sexhd + (\theta + \beta_5) eduhd + \beta_5 eduwf + u$$

$$= \beta_0 + \beta_1 \log(txhw) + \beta_2 \log(trhw) + \beta_3 sexhd + \theta eduhd + \beta_5 (eduhd + eduwf) + u$$

(将新变量(eduhd + eduwf)取代 eduhwf 来估测 θ和观测标准差 $\hat{\theta}$)

令 $totedu = eduhd + eduwf$

$$\log(donation) = \beta_0 + \beta_1 \log(txhw) + \beta_2 \log(trhw) + \beta_3 sexhd + \theta eduhd + \beta_5 totedu + u$$

(这样使当估计新模型时，eduhd 上的系数为θ，且 $s.e.(\hat{\theta})$ 一同被估计)

7) Testing Multiple Linear Restrictions (F统计模型)

计量模型如：

$$\underbrace{\log(donation)}_{\text{慈善捐款}} = \beta_0 + \beta_1 \underbrace{\log(txhw)}_{\text{应纳税家庭收入}} + \beta_2 \underbrace{\log(trhw)}_{\text{其它转移收入}} + \beta_3 \underbrace{sexhd}_{\text{户主性别}} + \beta_4 \underbrace{eduhd}_{\text{户主教育水平}} + \beta_5 \underbrace{eduwf}_{\text{妻子教育水平}} + u$$

(探究一旦控制了收入，那么教育和性别是家主对捐款没有影响)

零假设：

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

供选假设：

$$H_1: \beta_k \neq 0 \quad for \quad k = 3, 4, 5$$

在零假设中用制约因素：

$$\log(donation) = \beta_0 + \beta_1 \log(txhw) + \beta_2 \log(trhw) + u$$

(比较有限制后和无限制模型)

$$F = \frac{\frac{SSR_r - SSR_u}{q}}{\frac{SSR_u}{df_u}}$$

(两个模型中都用SSR)

(受限制的个数)

(无制约模型的SSR；自由度)

(变化后运用R^2的)

$$F = \frac{\frac{(R_u^2 - R_r^2)}{q}}{\frac{(1 - R_u^2)}{df_u}}$$

在零假设，CLM假设成立情况下，F统计量是有着 q, df_u 的 F分布

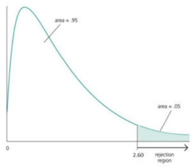
df_u

在零假设，CLM假设成立情况下，F统计量是有着q. df_u 的F分布

$$F \sim F_{q,\,df_u}$$

（假设步骤与t测试时一致）

1. 零假设和备选假设
2. 测试统计量（在零假设下的分布）
3. 显著性水平和临界值（从统计表中查）
4. 推演和结论



（模型中所有的斜率参数为0）

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_K = 0$$

限制后的模型为：

$$y = \beta_0 + u$$

F 统计量：

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

（通常被称为回归总显著性并包含于任何的回归结果中）

L5 方程形式

1) 对数

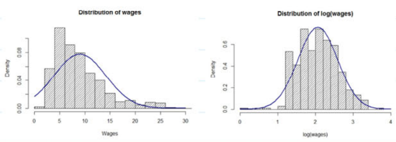
TABLE 2.3 Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

（半弹性演绎（semi-elasticity interpretation）只在小百分比数值中有效
但仍可以让CLM假设更真实）

正态假设

CLM 假设 $\Longrightarrow y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$



· 当有正的货币性数值时（wage, firm sales, house prices...）使用对数
· 变量为时间性的（education, experience, age...）时通常不用
· 注意使用百分比的数据报告（interest rates, unemployment rates...）

2) $\widehat{\log(y)}$ 中 y 的预测

如：

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

由此得出 $\widehat{\log(y)}$ 预测为：

$$\widehat{\log(y)} = \hat\beta_0 + \hat\beta_1 x_1 + \hat\beta_2 x_2 + \cdots + \hat\beta_k x_k$$

y 的朴素估测：

$$\hat y = \exp(\hat\beta_0 + \hat\beta_1 x_1 + \hat\beta_2 x_2 + \cdots + \hat\beta_k x_k)$$

但在CLM假设下：

$$\hat y = \exp\left(\frac{\hat\sigma^2}{2}\right)\exp\left(\widehat{\log(y)}\right)$$

3) 二次型 （Quadratic）

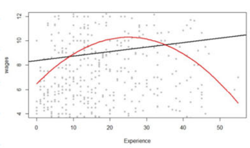
如：$wage_i = \beta_0 + \beta_1 experience_i + u_i$

（Quadratic form 可以使 experience 具有收益递减性）

Quadratic form：

$$\widehat{wage} = \hat\beta_0 + \hat\beta_1 exper + \hat\beta_2 exper^2$$

推演：$\frac{\Delta wage}{\Delta exper} = (\hat\beta_1 + 2\hat\beta_2 exper)$

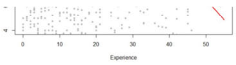


如果 $\hat\beta_2 > 0$，U的抛物线
如果 $\hat\beta_2 < 0$，∩的抛物线

如：$\widehat{wage} = 6.477 + 0.302\ exper - 0.00603\ exper^2$

如果 $\hat{\beta}_2>0$，U的抛物线

如果 $\hat{\beta}_2<0$，∩的抛物线



如：$\widehat{wage}=6.477+0.302\,exper-0.00603\,exper^2$

$1^{st}.(exper=0)$ $\quad 0.302-2\times0.00603\times0=\$0.302/h$

$2^{nd}.(exper=1)$ $\quad 0.302-2\times0.00603\times1=\$0.290/h$

$$\begin{cases}\hat{\beta}_1+2\hat{\beta}_2\,exper\\ \hat{\beta}_1=0.302\ ;\ \hat{\beta}_2=-0.00603\end{cases}$$

4）虚拟变量

变量如 educ 和 exper 是有量化（Quantitative）意义的（数的大小都有着信息）

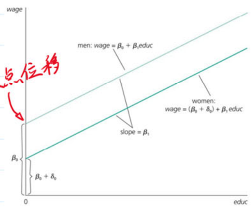
经验中还有些起重要因素的定性（Qualitative）因素

如：性别，（有无）宗教信仰等（通常是二元答案（如有或无））

可以用 1 或 0 分别代表

如：female = 1 为女性；female = 0 为男性

交点位移



a）$\widehat{wage}=\hat{\beta}_0+\hat{\delta}_0 female+\hat{\beta}_1 educ$

推演 $\hat{\delta}_0$：（假设 ZCM 假设成立（$E[u|female,educ]=0$））

$\hat{\delta}_0=E[wage|female=1,\ educ]-E[wage|female=0,\ educ]$

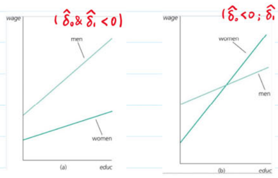
（使用虚拟变量使交点位移）

b）$\widehat{Wage}=\hat{\beta}_0+\hat{\delta}_0 female+\hat{\beta}_1 educ+\hat{\delta}_1 female\times educ$

当 $\hat{\delta}_0$ 为男或女时会导致交点位移

当 $\hat{\delta}_1$ 为男或女时会导致斜率不同

（$\hat{\delta}_0$&$\hat{\delta}_1<0$） （$\hat{\delta}_0<0$；$\hat{\delta}_1>0$）



$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_1x_2$$

$$\frac{\Delta\hat{y}}{\Delta x_1}=\hat{\beta}_1+\hat{\beta}_3x_2$$

当 $x_2=0$ 时，$\hat{\beta}_1$ 是与 Δx_1 关联的 $\Delta\hat{y}$

如：$\hat{\beta}_3>0$：当 x_2 为 1 时，x_1 增加对 $\hat{y}$ 增加的影响大于 x_2 为 0 时的 x_1 增加

$\hat{\beta}_3<0$：当 x_2 为 1 时，x_1 增加对 $\hat{y}$ 增加的影响小于 x_2 为 0 时的 x_1 增加

进行 t 检测：$H_0:\delta_1=0$

（检测同性别下教育的回报是否相同）

（如斜率的不同）

进行 F 检测：$H_0:\delta_0=\delta_1=0$

（如性别是否对工资有任何影响）

（将斜率和交点合并检验）

$Wage=\beta_0+\delta_0 female+\beta_1 married+\delta_2 female\times married+u$

- 当 female = 0 & married = 0 ⟹ 单身男性（β_0）
- 当 female = 0 & married = 1 ⟹ 已婚男性（$\beta_0+\delta_1$）
- 当 female = 1 & married = 0 ⟹ 单身女性（$\beta_0+\delta_0$）
- 当 female = 1 & married = 1 ⟹ 已婚女性（$\beta_0+\delta_0+\delta_1+\delta_2$）

5）分类变量

可取多种值的定性因素

如：学位类型；信仰

1 = 一等　　1 = 佛教

2 = 二等一　2 = 基督教

3 = 二等二　3 = 伊斯兰教

（顺序无所谓；需要转化为多个虚拟变量）

一些分类变量有着自然顺序

如：学位类型；学位等级

1 = GCSE

2 = A-Level

1 = GCSE
2 = A-Level
3 = BSc
4 = MSc
5 = PhD

$Wage = \beta_0 + \beta_1 degree + u$

(或者用虚拟变量列出每一个)

$Wage = \delta_0 + \delta_1 d_1 + \delta_2 d_2 + \cdots + \delta_5 d_5 + u$

L_6 一致性和内生性

1) 内生性的由来

ZCM 假设

$E(u|X_1, X_2, \dots, X_k) = 0$

然而，当：

$E[u|x_j] \neq 0$

我们称 x_j 是内生性解释变量

内生性的来源

- 遗漏变量
- 同步性
- 误差测量

内生性：模型中一个或多个解释变量与随机扰动相关

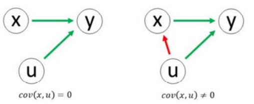
思路：比较 OLS 与 IV 估计量，若都是外生的，则 OLS 与 IV 皆一致；不同则为解释变量有内生性

如：$y_i = \alpha + \beta x_i + u_i$

变化 x 会导致 β 改变 y

以上三种均导致 x 是有内生性

Causal Paths



2) 一致性

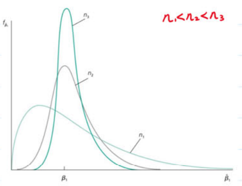
在高斯—马尔科夫假设下 OLS 是 BLUE（最优线性无偏估计量）

但如果高—马假设不成立，则没有具有无偏估计量的可能性

在该情况下，我们便能满足一致的估计量

（当估计量的分布折叠为 $n \rightarrow \infty$ 的真实值，则估计量是一致的）

（具有一致性是估计量的最低要求）



· 令 $\hat{\beta}_j$ 是 β_j 的 OLS 估计量，依高—马 MLR.1～4，$\hat{\beta}_j$ 是无偏的，该分布有均值 β_j。如果估计量是一致的，$\hat{\beta}_j$ 的分布将随着 β_j 的增加而愈发分布在 β_j 周围，当 n 取无限时，$\hat{\beta}_j$ 将会折叠到 β_j 点上。

· 高—马假设下 OLS 估计量是一致的（且无偏）

如果 $\text{plim}(\hat{\beta}_j) = \beta_j$，则 $\hat{\beta}_j$ 是 β_j 的一致估计量

（概率极限）

· 对于无偏性，假设 ZCM（零条件均值）

$E[u|x_1, \dots, x_k] = 0$

· 对于一致性，应用平均值和协方差为 0 这更少限制的假设

$E[u] = 0$ & $cov(x_j, u) = 0$ for $j = 1, \dots, k$

（如果都不满足，OLS 有偏且不一致）

3) 测量误差（Chapter 9）

如：$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$

但观测 y 而非 y^*

测量误差：$e_0 = y - y^*$

$\therefore y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + (u + e_0)$

假设 e_0 与 x_j' 无关，则 OLS 是无偏的

如：$y = \beta_0 + \beta_1 x_1^* + u$

（满足 MLR.1～4），但观测 x_1 而非 x_1^*

令 $x_1 = x_1^* + e_1$，当 $E[e_1] = 0$（均值为 0）

· 因为假设 u 和 e_1 均值为 0 且与 x_1 无关；

$u - \beta_1 e_1$ 均值为 0 且与 x_1 无关，那么 x_1 取代 x_1^*

考于 OLS 估计时会产生一致的估测量 β_1（和 β_0）

因 u 和 e_1 无关，误差的方差是 $Var(u - \beta_1 e_1)$

$= \sigma_u^2 + \beta_1^2 \sigma_{e_1}^2$。故当 $\beta_1 = 0$，测量误差增大了误差

如: $y=\beta_0+\beta_1 x_1^*+u$

(满足 MLR.1~4),但观测 x_1 而非 x_1^*

令 $x_1=x_1^*+e_1$,当 $E[e_1]=0$ (均值为0)

则: $y=\beta_0+\beta_1 x_1+(u-\beta_1 e_1)$

为了OLS估计时会产生一致的估测量 β_1(和 β_0)

因 u 和 e_1 无关,误差的方差是 $Var(u-\beta_1 e_1)$

$=\sigma_u^2+\beta_1^2\sigma_{e_1}^2$。故当 $\beta_1=0$,测量误差增大了误差

方差,但对OLS的性质无影响

经典误差值假设 (Classical errors-variables (CES))

$$Cov(x_1^*,e_1)=0$$

当满足 CEV 时,x_1 和 e_1 必有以下关系

$$cov(x_1,e_1)=E[x_1 e_1]=E[x_1^* e_1]+E[e_1^2]=0+\sigma_{e_1}^2$$

用在估测模型

$$y=\beta_0+\beta_1 x_1+(u-\beta_1 e_1)$$

x_1 与复合误差 $u-\beta_1 e_1$ 有关:

$$cov(x_1,u-\beta_1 e_1)=-\beta_1 Cov(x_1,e_1)=-\beta_1\sigma_{e_1}^2$$

OLS的有偏和不一致的量: (当 $Var(x_1)=Var(x_1^*)+Var(e)$)

$$plim(\hat{\beta}_1)=\beta_1+\frac{Cov(x_1,u-\beta_1 e_1)}{Var(x_1)}$$

$$=\beta_1-\frac{\beta_1\sigma_{e_1}^2}{\sigma_{x_1^*}^2+\sigma_{e_1}^2}$$

$$=\beta_1\left(1-\frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2+\sigma_{e_1}^2}\right)$$

$$=\beta_1\left(\frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2+\sigma_{e_1}^2}\right)$$

· β_1 的乘积,即 $\frac{Var(x_1^*)}{Var(x_1)}$ 的比例总是小于1,

因为 $Cov(x_1^*,e_1)=0$,因此,$plim(\hat{\beta}_1)$ 总是比

β_1 更靠近0 (OLS偏的衰减)

4) 联立方程模型 (simultaneous equation model)

如:

$$crime=\alpha_1\ police+\beta_1\ avginc+u_1 \quad ①$$

$$police=\alpha_2\ crime+\beta_2\ tax+u_2 \quad ②$$

将②代入①: $crime=\alpha_1(\alpha_2\ crime+\beta_2\ tax+u_2)+\beta_1\ avginc+u_1$

重新组合: $(1-\alpha_1\alpha_2)crime=\alpha_1\beta_2\ tax+\beta_1\ avginc+\alpha_1 u_2+u_1$

$$crime=\frac{\alpha_1\beta_2}{1-\alpha_1\alpha_2}tax+\frac{\beta_1}{1-\alpha_1\alpha_2}avginc+\frac{\alpha_1 u_2+u_1}{1-\alpha_1\alpha_2}$$

L7 工具变量 (Instrumental Variables) (Chapter 15)

1) OLS 的渐近性 (Asymptotic)

(描述 $n\to\infty$ 的大型样本中的估计量的性质)

1、无偏:

$E[\hat{\beta}]=\beta$,估计量 $\hat{\beta}$ 的性质

(意为估计量的样本分布与真实值相等)

2、一致:

当样本量扩大,估计量的覆盖率达到真实的总值

(意为当样本量扩大时 估计量的样本方差怎样变的问题)

2) 工具变量

如: $Wage_i=\beta_0+\beta_1 educ_i+\beta_2 ability_i+u_i$

但 ability 不可测量

$\therefore$ $wage_i=\beta_0+\beta_1 educ_i+u_i$

如果 educ 和 ability 是有相关性的,那 $E[u|x]\neq 0$; $cov(x,u)\neq 0$

$\Longrightarrow$ OLS 既有偏又不一致

且 $\hat{\beta}_1$ 不再能因果推演 (causal interpretation)

$$y=\beta_0+\beta_1 x+u$$

($Cov(x,u)\neq 0$)

假设有观测值 Z,且满足:

1) $Cov(z,u)=0$

2) $Cov(z,x)\neq 0$

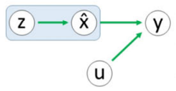
Z 便是工具变量

1) 为工具外生性

2) 为工具相关

3) Z 不应包括在模型中

Casual Paths



（为了估测设定所需的因果关系，z 对 y 有影响，
但只在内生性变量 x 下有影响）

工具变量的估测量

在 $E[u|x] \neq 0$ 下：

$$y = \beta_0 + \beta_1 x + u$$

一：假设 $cov(z, x) \neq 0$

$$x = \pi_0 + \pi_1 z + v$$

$$\frac{\Delta x}{\Delta z} = \pi_1$$

二：

$$y = \beta_0 + \beta_1 [\pi_0 + \pi_1 z + v] + u$$

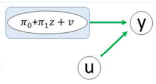
$$= [\beta_0 + \beta_1 \pi_0] + \beta_1 \pi_1 z + [\beta_1 v + u]$$

$$\frac{\Delta y}{\Delta z} = \beta_1 \pi_1$$

故：

$$\frac{\Delta y}{\Delta x} = \frac{\frac{\Delta y}{\Delta z}}{\frac{\Delta x}{\Delta z}} = \frac{\beta_1 \pi_1}{\pi_1} = \beta_1$$

假设：$cov(z, x) \neq 0$ 且 $cov(z, u) = 0$



估测步骤一：

$$\hat{x} = \hat{\pi}_0 + \hat{\pi}_1 z$$

将拟合值代入二：

$$y = \beta_0 + \beta_1 [\pi_0 + \pi_1 z + v] + u$$

$$= \beta_0 + \beta_1 \hat{x} + u$$

用 $cov(z, x) \neq 0$ 和 $cov(z, u) = 0$ 来确认 β_1

Identification of β_1：把 β_1 写成可被样本数据估测的总体矩阵条件（Population moment condition）

用 $y = \beta_0 + \beta_1 x + u$ 去获得 z 和 y 的协方差

$$cov(z, y) = \beta_1 cov(z, x) + cov(z, u)$$

重新排列：

$$\beta_1 = \frac{cov(z, y)}{cov(z, x)}$$

样本副本：

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

∴ 工具变量（IV）的估计量：

$$\hat{\beta}_1^{IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \quad \& \quad \hat{\beta}_0^{IV} = \bar{y} - \hat{\beta}_1^{IV} \bar{x}$$

（一致但有偏）因此 IV 估计量只用于判断'大'变量

当 $z = x$ 时，（如 x 是实际外生性且可被认为是自身的工具）

则 $\hat{\beta}_1^{IV} = \hat{\beta}_1^{OLS}$ （比较 OLS 与 IV 的估计量）

好的工具变量的来源：

1、相关性 $cov(z, x) \neq 0$

2、外生性 $cov(z, x) = 0$

3. z 不应包含在模型中

好的 IV 难以找寻：

如：自然灾害，政策变化

L8 面板数据方法估计量（Panel Data Methods Estimators）

1) 邹氏检验（CHOW test）（F检测）

截面数据（Cross Section Data）

（进行跨单元观测，以查看它们的差异）

独立面板样本

如：$\log(wage)_{it} = \beta_0 + \delta_0 wave2_{it} + \beta_1 educ_{it} + \beta_2 exper_{it} + \beta_3 female_{it} + \beta_4 female_{it} wave2_{it} + \epsilon_{it}$

（i 指代人；t 为时间段；$wave_{it}$ 指观测指为2阶段：随时间的推移，性别工资差距）

（使用邹氏检测（主要为两个时间段的F检测）；虚拟变量的另一应用便是发现时间段下的模型差异）

如：$g = 1, 2$ 两个组下的模型

$$y = \beta_{0,g} + \beta_{1,g} x_1 + \cdots \beta_{p,g} x_p + \epsilon$$

一：$\beta_{i,g}$ 为每一组下的参数；因此可用联合假设（joint hypotheses）

$$H_0 : \beta_{i,1} = \beta_{i,2} \ \forall i$$

（全称量词，对所有的 i；每一个 i）

（应用F检测，其两组中所有参数的受限模型相等）

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

如：$\log(wage)_{it} = \beta_0 + \delta_0 wave2_{it} + \beta_1 educ_{it} + \beta_2 exper_{it} + \beta_3 female_{it} + \beta_4 female_{it} wave2_{it} + a_i + \epsilon_{it}$

（a_{it} 是随时间推移而不变的所有未被观测的异质性）（如与x的方差有关，则有因遗漏恒定变量时间造成的有偏）

二：∴要移出异致性偏差（虚拟变量需要照顾每一单（R^2 同理会变高）因此不好）