Intermediate Econometrics

Professor: Xiaojun Song

Timekeeper: Zircon

Spring 2023

目录

1 简单回归模型 **3**
1.1 OLS . . . 3
1.2 拟合优度 . . . 4
1.3 非线性回归模型 . . . 4
1.4 OLS 估计量的期望和方差 . . . 4

2 多元线性回归：估计 **7**
2.1 OLS . . . 7
2.2 OLS 估计量的期望和方差 . . . 7
2.3 拟合优度 . . . 8
2.4 误设模型问题 . . . 8
2.5 OLS 估计量的有效性 . . . 10

3 多元回归分析：推断 **11**
3.1 正态抽样分布 . . . 11
3.2 单个参数的 t 检验 . . . 11
3.3 参数线性组合的检验 . . . 12
3.4 多个线性约束的检验 . . . 12
3.5 经济显著性 . . . 13

4 多元回归分析：OLS 的渐进性 **14**
4.1 一致性 . . . 14
4.2 渐进正态和大样本推断 . . . 15

5 MLR Analysis: Further Issues **16**
5.1 More on Functional Form . . . 16
5.2 Goodness of Fit . . . 17

6 MLR Analysis with Dummy **19**
6.1 Different Intercepts: Dummy as Regressors . . . 19
6.2 Different Slopes: Interaction with Dummy . . . 21
6.3 Binary Regressand: Linear Probability Model . . . 22
6.4 Problem of Self-Selection . . . 23

7 Heteroskedasticity **25**
7.1 Properties of OLS under Heteroskedasticity . . . 25
7.2 Testing for Heteroskedasticity . . . 27
7.3 Weighted Least Squares Estimation . . . 28

8 More on Specification and Data Issues **32**
8.1 Test for Functional Form Misspecification . . . 32
8.2 Proxy Variables . . . 34
8.3 Random Coefficient Model . . . 35
8.4 Measurement Error . . . 36
8.5 Missing Data . . . 38
8.6 Outliers & Influential Observations . . . 39

9 Panel Data **40**
9.1 DiD . . . 40
9.2 First Difference . . . 42
9.3 Fixed Effects Estimation . . . 43
9.4 Random Effects Estimation . . . 46
9.5 Correlated Random Effects Estimation . . . 47
9.6 General Policy Analysis with Panel Data . . . 48

10 IV Estimation & 2SLS **50**
10.1 Instrumental Variable . . . 50
10.2 IV Estimation in SLR . . . 51
10.3 IV Estimation in MLR . . . 52
10.4 Two Stage Least Squares Estimation . . . 54
10.5 Weak Instrument . . . 56
10.6 IV in Measurement Error . . . 57
10.7 Testing for Endogeneity of Explanatory Variables . . . 57
10.8 Over-Identifying Restrictions . . . 58

1 简单回归模型

简单线性回归模型 (SLR Model)：$y = \beta_0 + \beta_1 x + u$

- y：因变量、被解释变量、响应变量、回归子
- x：自变量、解释变量、控制变量、回归元
- u：误差项、干扰项

定义截距：$\mathrm{E}(u) = 0$

零条件均值 (zero conditional mean) 假定：$\mathrm{E}(u|x) = 0$

总体回归函数 (PRF)：$\mathrm{E}(y|x) = \beta_0 + \beta_1 x$

1.1 OLS

1.1.1 OLS 的必要条件

1. $\mathrm{E}(u) = 0$
2. $\mathrm{E}(xu) = 0 \Longleftarrow \mathrm{Cov}(x, u) = 0$

1.1.2 OLS 估计量

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

样本回归方程 (SRF)：$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \Longrightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (\hat{u}_i := y_i - \hat{y}_i)$

1.1.3 OLS 估计量的代数性质

- $\sum_{i=1}^{n} \hat{u}_i = 0$（残差和为 0，残差均值为 0）
- $\sum_{i=1}^{n} x_i \hat{u}_i = 0$ （回归元和残差的样本协方差为 0）
- $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$（均值点在 OLS 回归线上）

注：

- $\beta_1 = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} = \hat{\rho}_{xy} \cdot \frac{\hat{\sigma}_y}{\hat{\sigma}_x}$
- 妙用 $\sum_{i=1}^{n} c(x_i - \bar{x}) = 0$，其中 c 是任意常数

1.2 拟合优度

$$
\begin{aligned}
SST &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
SSE &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
SSR &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
R^2 &= \frac{SSE}{SST} = 1 - \frac{SSR}{SST}
\end{aligned}
$$

拟合优度 R^2 衡量的是模型可以解释的数据变异性占总变异性的比例。

注：

- 特别地，R^2 实际上是因变量的预测值和真实值的相关系数的平方。
- R^2 的进一步含义为，回归模型与虚无模型对比能够多解释的变异性。

1.3 非线性回归模型

- $y \sim x \to \Delta y = \beta_1 \Delta x$
- $y \sim \log x \to \Delta y = \frac{\beta_1}{100}\%\Delta x$
- $\log y \sim x \to \%\Delta y = (100\beta_1)\Delta x$
- $y \sim x \to \%\Delta y = \beta_1 \%\Delta x$

1.4 OLS 估计量的期望和方差

1.4.1 基本假设

- SLR.1-线性于参数：总体回归函数为 $y = \beta_0 + \beta_1 x + u$
- SLR.2-随机抽样：$\{(x_i, y_i) : i = 1, 2, ..., n\} \to y_i = \beta_0 + \beta_1 x_i + u_i$
- SLR.3-自变量样本有波动：$SST_x > 0$
- SLR.4-零条件均值：$\mathrm{E}(u|x) = 0$
- SLR.5-同方差性：$\mathrm{Var}(u|x) = \sigma^2$

1.4.2 期望和无偏性

在 SLR.1~SLR.4 下，OLS 估计量是无偏的，$\mathbf{E}(\hat{\beta}_1)=\beta_1, \mathbf{E}(\hat{\beta}_0)=\beta_0$。

$$
\begin{aligned}
\because \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i-\bar{x})y_i}{SST_x} = \frac{\sum_{i=1}^n (x_i-\bar{x})(\beta_0+\beta_1 x_i+u_i)}{SST_x} \\
&= \frac{\sum_{i=1}^n (x_i-\bar{x})\beta_0 + \sum_{i=1}^n (x_i-\bar{x})\beta_1 x_i + \sum_{i=1}^n (x_i-\bar{x})u_i}{SST_x} \\
&= \frac{0+\beta_1 \sum_{i=1}^n (x_i-\bar{x})x_i + \sum_{i=1}^n (x_i-\bar{x})u_i}{SST_x} \\
&= \frac{0+\beta_1 \sum_{i=1}^n (x_i-\bar{x})(x_i-\bar{x}) + \sum_{i=1}^n (x_i-\bar{x})u_i}{SST_x} \\
&= \beta_1 + \frac{\sum_{i=1}^n (x_i-\bar{x})u_i}{SST_x} \\
\therefore \mathrm{E}(\hat{\beta}_1) &= \beta_1 + \frac{1}{SST_x}\cdot\sum_{i=1}^n (x_i-\bar{x})\mathrm{E}(u_i) = \beta_1 \\
\because \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \\
&= \frac{1}{n}\cdot\sum_{i=1}^n y_i - \hat{\beta}_1\bar{x} = \frac{1}{n}\cdot\sum_{i=1}^n (\beta_0+\beta_1 x_i+u_i) - \hat{\beta}_1\bar{x} \\
&= \beta_0 + \beta_1\bar{x} + \frac{1}{n}\cdot\sum_{i=1}^n u_i - \hat{\beta}_1\bar{x} = \beta_0 + (\beta_1-\hat{\beta}_1)\bar{x} + \frac{1}{n}\cdot\sum_{i=1}^n u_i \\
\therefore \mathrm{E}(\hat{\beta}_0) &= \beta_0 + \frac{1}{n}\cdot\sum_{i=1}^n \mathrm{E}(u_i) = \beta_0
\end{aligned}
$$

若 SLR.1~SLR.4 中有一个不成立，那么无偏性一般是不成立的；尤其是 SLR.4，误差项中可能存在与自变量相关的因素。

1.4.3 方差

在 SLR.1~SLR.5 下，$\mathrm{Var}(\hat{\beta}_1)=\frac{\sigma^2}{SST_x}, \mathrm{Var}(\hat{\beta}_0)=\frac{n^{-1}\sigma^2\sum_{i=1}^n x_i^2}{SST_x}$

$$
\begin{aligned}
\because \hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^n (x_i-\bar{x})u_i}{SST_x} \\
\therefore \mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}(\frac{\sum_{i=1}^n (x_i-\bar{x})u_i}{SST_x}) = \frac{1}{(SST_x)^2}\cdot\sum_{i=1}^n (x_i-\bar{x})^2\mathrm{Var}(u_i) = \frac{\sigma^2}{SST_x}
\end{aligned}
$$

误差方差通常未知，误差方差的无偏估计为

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

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

$$\mathrm{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{(SST_x)^{\frac{1}{2}}}$$

2 多元线性回归：估计

多元线性回归模型 (MLR Model)：$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$

2.1 OLS

2.1.1 OLS 的必要条件

- $\mathrm{E}(u) = 0$
- $\mathrm{E}(x_j u) = 0, \forall j$

2.1.2 OLS 的估计量

OLS 斜率系数为（以 $\hat{\beta}_1$ 为例，其余相同）

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \hat{r}_{i1} y_i}{\sum_{i=1}^n \hat{r}_{i1}^2}$$

其中 $\hat{r}_{i1}$ 是基于现有样本，将 x_1 对其他所有回归元（$x_2, ..., x_k$）回归得到的残差。

$\hat{\beta}_1$ 度量的是，排除其他所有回归元的影响之后，y 和 x_1 的样本关系。

偏效应指的是，**其他条件不变时**，x_j 对 y 的影响体现为 β_j。（注意，截距项可能不具有实际意义）

$\hat{\beta}_0$ 可以通过回归线通过样本中心点这一特性直接得出。

2.1.3 OLS 估计量的代数性质

- 残差的样本均值为 0 $\Leftrightarrow \sum_{i=1}^n \hat{u}_i = 0, \bar{y} = \bar{\hat{y}}$
- 每个自变量和残差之间的样本协方差为 0 $\Leftrightarrow \sum_{i=1}^n x_{ij}\hat{u}_i = 0, \forall j = 1, 2, ..., k \Leftrightarrow \sum_{i=1}^n \hat{y}_i \hat{u}_i = 0$，也即 OLS 拟合值和残差之间的样本协方差为 0
- 点 $(\bar{x}_i, \bar{x}_2, ..., \bar{x}_k, \bar{y})$ 总在回归线上 $\Leftrightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + ... + \hat{\beta}_k \bar{x}_k$

2.2 OLS 估计量的期望和方差

2.2.1 基本假设

- MLR.1-线性与参数：总体回归函数为 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$
- MLR.2-随机抽样：$\{(x_{i1}, x_{i2}, ..., x_{ik}, y_i) : i = 1, 2, ..., n\} \rightarrow y_i = \beta_0 + \beta_1 x_i + u_i$

- MLR.3-不存在完全共线性：样本中没有一个自变量是常数（共线于斜率），自变量之间没有严格的线性关系需要注意，可能存在等价转换之后的共线性，如 $\log w$ 和 $\log w^2$
- MLR.4-零条件均值：$\mathrm{E}(u|x_1, x_2, ..., x_k) = 0$
- MLR.5-同方差性：$\mathrm{Var}(u|x_1, x_2, ..., x_k) = \sigma^2$

2.2.2 OLS 估计量的期望和无偏性

在 MLR.1~MLR.4 下，$\mathrm{E}(\hat{\beta}_j) = \beta_j, \forall j = 0, 1, 2, ..., k$

2.2.3 OLS 估计量的方差

在 MLR.1~MLR.5 下，$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j \cdot (1-R_j^2)}, \forall j = 1, 2, ..., k$

其中，定义方差膨胀因子 $VIF_j = \frac{1}{1-R_j^2}$，$VIF < 10$ 被认为是底线。

2.3 拟合优度

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = \frac{(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y}))^2}{(\sum_{i=1}^n (y_i - \bar{y})^2) \cdot (\sum_{i=1}^n (\hat{y}_i - \bar{y})^2)}$$

注：

- 在回归中添加自变量时，R^2 至少不会减少
- 判断模型中是否应该加入变量，判定标准应该是**偏效应是否非零**，而不是 R^2 是否增大
- 不同类型、不同自变量的模型之间的比较应该看调整后的 R^2

2.4 误设模型问题

2.4.1 引入无关变量

引入**无关变量**不影响 OLS 估计量的无偏性，但会**影响方差**。

2.4.2 遗漏变量

假设真实的回归方程应为 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2$，但遗漏 x_2 后的误设回归模型为 $\tilde{y}=\tilde{\beta}_0+\tilde{\beta}_1x_1$。而 x_2 对 x_1 的回归有 $\tilde{x}_2=\tilde{\delta}_0+\tilde{\delta}_1x_1$，综合来有

$$\tilde{\beta}_1=\hat{\beta}_1+\hat{\beta}_2\tilde{\delta}_1$$
$$\mathrm{Bias}(\tilde{\beta}_1)=\mathrm{E}(\tilde{\beta}_1)-\beta_1=\beta_2\tilde{\delta}_1$$

若 $\mathrm{Bias}(\tilde{\beta}_1)=\mathrm{E}(\tilde{\beta}_1)-\beta_1=\beta_2\tilde{\delta}_1>0$，则说明有向上的偏误；若 $\mathrm{Bias}(\tilde{\beta}_1)=\mathrm{E}(\tilde{\beta}_1)-\beta_1=\beta_2\tilde{\delta}_1<0$，则说明有向下的偏误；若 $\mathrm{E}(\tilde{\beta}_1)$ 比 β_1 更接近 0，则说明有向零的偏误。

在更一般的情形下，**如果一个解释变量与误差相关，那么这通常会导致所有的 OLS 估计量都产生偏误**。只要被遗漏的变量与其他解释变量都有关，但因误设而被放入误差中。

如果 $\beta_2=0$，则 $\tilde{\beta}_1$ 和 $\hat{\beta}_1$ 都是无偏的。并且有 $\mathrm{Var}(\tilde{\beta}_1)=\frac{\sigma^2}{SST_1}$，$\mathrm{Var}(\hat{\beta}_1)=\frac{\sigma^2}{SST_1\cdot(1-R_1^2)}$，因此 $\mathrm{Var}(\tilde{\beta}_1)<\mathrm{Var}(\hat{\beta}_1)$。实际上，这说明了往模型中加入无关变量，会加剧多重共线性的问题，使得 $\hat{\beta}_1$ 的估计效率下降。

如果 $\beta_2\neq 0$，则 $\tilde{\beta}_1$ 将会是有偏的，不过方差上的劣势可以通过样本量弥补，因此偏好将 x_2 纳入回归。

如果 $\delta_1=0$，由于**偏误是以样本为条件的**，所以特定样本中 $\tilde{\delta}_1$ 并不一定为零，可能仍然会有偏误。

2.4.3 多重共线性

$$VIF_j=\frac{1}{1-R_j^2}$$

如果 $VIF_j>10$，或者 $R_j^2\to 1$，那么意味着存在多重共线性。

形式上看，多重共线性会让系数的误差变大，不仅更难显著，也意味着估计难以精确。同时，存在多重共线性的回归元之间由于有高度的线性关系，可以相互表出，计算系数可能都不准确。此外，**多重共线性的影响可能不是通过扩大样本量能够直接消除的**。

不过，即便**多重共线性对单一系数的 t 检验能够产生影响，F 检验对多个变量之间的多重共线性免疫**。

2.4.4 OLS 估计量的标准误

方差是很难得到的，只能通过特定的样本估计标准误，在 MLR.1~MLR.5 下，无偏估计为

$$\hat{\sigma}^2=\frac{1}{n-k-1}\cdot\sum_{i=1}^{n}\hat{u}_i^2 \qquad \mathrm{E}(\hat{\sigma}^2)=\sigma^2$$

$\hat{\sigma}=\sqrt{\hat{\sigma}^2}$ 为回归标准误 (SER, standard error of regression)，且有

$$\mathrm{se}(\hat{\beta}_j)=\frac{\hat{\sigma}}{\sqrt{SST_j\cdot(1-R_j^2)}}=\frac{\hat{\sigma}}{\sqrt{n\cdot\mathrm{sd}(x_j)\cdot(1-R_j^2)}}$$

2.5 OLS 估计量的有效性

高斯-马尔科夫定理：在 MLR.1~MLR.5 下，$\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ 分别是 $\beta_0, \beta_1, ..., \beta_k$ 的最优线性无偏估计量。(BLUE, best linear unBiased estimator)

- 最优：具有最小的方差
- 线性：β_j 的一个估计量 $\hat{\beta}_j$ 是线性的充分必要条件是，$\hat{\beta}_j$ 能表示成因变量数据的一个线性函数，也即 $\hat{\beta}_j = \sum_{i=1}^{n} w_I y_i$，其中 w_i 可以是关于自变量样本值的一个函数
- 无偏：$\mathrm{E}(\hat{\beta}_j) = \beta_j$

注：称 MLR.1~MLR.5 为高斯-马尔科夫假定。

3 多元回归分析：推断

3.1 正态抽样分布

当以样本中自变量的值为条件时，OLS 估计量的抽样分布取决于其背后的误差分布。为使 $\hat{\beta}_j$ 的抽样分布易于处理，假定总体中不可观测的误差是正态分布的。

MLR.6-正态性：总体误差独立于解释变量 $x_1, x_2, ..., x_k$，并且 $u \sim N(0, \sigma^2)$。

MLR.6 是一个强假定，相当于在 MLR.4 & MLR.5 的基础上假设了 u 的分布；因此，MLR.6 也可以推出 MLR.4 & MLR.5。

CLM 假定下，总体可以表示为 $y|\vec{x} \sim N(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k, \sigma^2)$。

在 CLM 假定下，OLS 估计量是最小方差无偏估计量。这比高斯-马尔科夫假定下具有更强的效率，不需要把比较局限于对 y_i 为线性的估计量内。

正态抽样分布定理：在 CLM 假定下，以自变量的样本值为条件，有 $\hat{\beta}_j \sim N(\beta_j, \mathbf{Var}(\hat{\beta}_j))$

证明过程只要理解 OLS 估计量是样本误差的一个线性组合、并且样本误差是独立同分布的随机变量即可。在此基础上，该定理也可以加强，OLS 估计量的任何线性组合也都服从人态分布。

注：

- 称 MLR.1~MLR.6 为经典线性模型假定 (CLM assumptions, classical linear model)
- 误差服从正态分布：误差是影响因变量但又无法观测的诸多因素之和，可以借助中心极限定理断定误差具有近似正态分布。但近似的效果取决于误差中的因素、各自的分布差异，以及彼此之间的独立性。
- 大样本下 OLS 估计量的正态性近似成立

3.2 单个参数的 t 检验

标准化估计量的 t 分布：在 CLM 假定下，$\frac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\hat{\beta}_j)} \sim t_{n-k-1}$，其中 $k+1$ 是总体模型中未知参数的个数。

假设检验和构造置信区间是等价的，后者可以直接判断是否拒绝特定的原假设。

注：

- 原假设中对 β_j 的判断，不一定是 0
- 原假设和备择假设的方向，注意检验是单尾还是双尾，注意对应的 p 值和临界值
- 大样本的 t 检验可以用标准正态分布近似
- “不能拒绝原假设” 不等同于 “接受原假设”

3.3 参数线性组合的检验

假设有回归方程 $\hat{y}=\hat{\beta}_o+\hat{\beta}_1x_1+\hat{\beta}_2x_2$，如果要检验参数的线性组合，如 $\beta_1=p+q\beta_2$，那么可以**最一般性地表出** β_1 **为** $\beta_1=p+q\beta_2+\theta$，如果原假设成立，那么 $\theta=0$。将 $\beta_1=p+q\beta_2+\theta$ 代入回归方程，可以得到

$$\begin{aligned}
y&=\beta_0+\beta_1x_1+\beta_2x_2=\beta_0+(p+q\beta_2+\theta)x_1+\beta_2x_2\\
\Longrightarrow y-px_1&=\beta_0+\theta x_1+\beta_2(qx_1+x_2)
\end{aligned}$$

这时只要构造新的因变量为 $y-px_1$，新的自变量为 qx_1+x_2，并检验 x_1 的系数是否显著地异于 0 即可。同时，这也可以最简单地得到 $\mathrm{se}(\hat{\theta})=\mathrm{se}(\hat{\beta}_1-q\hat{\beta}_2)$，而后者是很难计算得出的。

要计算 $\mathrm{se}(\hat{\beta}_1-q\hat{\beta}_2)$，那么必须能够计算 $\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2)$。记住，OLS 估计量都是 y_i 的线性函数，而 y_i 之间是独立的。

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2)&=\mathrm{Cov}(\frac{\sum_{i=1}^n\hat{r}_{i1}y_i}{\sum_{i=1}^n\hat{r}_{i1}^2},\frac{\sum_{i=1}^n\hat{r}_{i2}y_i}{\sum_{i=1}^n\hat{r}_{i2}^2})\\
&=\frac{1}{\sum_{i=1}^n\hat{r}_{i1}^2\cdot\sum_{i=1}^n\hat{r}_{i2}^2}\cdot\mathrm{Cov}(\sum_{i=1}^n\hat{r}_{i1}y_i,\sum_{i=1}^n\hat{r}_{i2}y_i)\\
&=\frac{1}{\sum_{i=1}^n\hat{r}_{i1}^2\cdot\sum_{i=1}^n\hat{r}_{i2}^2}\cdot\sum_{i=1}^n\hat{r}_{i1}\hat{r}_{i2}\cdot\sigma^2\\
\Longrightarrow\mathrm{Var}(w_1\hat{\beta}_1+w_2\hat{\beta}_2)&=\sigma^2\cdot\frac{\sum_{i=1}^n(w_2\hat{r}_{i1}+w_1\hat{r}_{i2})^2}{\sum_{i=1}^n\hat{r}_{i1}^2\cdot\sum_{i=1}^n\hat{r}_{i2}^2}\\
\Longrightarrow\mathrm{Var}(\sum_{i=1}^k w_i\hat{\beta}_i)&=\frac{1}{2}\sum_{m\neq n}\mathrm{Var}(w_n\hat{\beta}_m+w_m\hat{\beta}_n)+\sum_{m\neq n}w_mw_n\cdot\mathrm{Cov}(\hat{\beta}_m,\hat{\beta}_n)
\end{aligned}$$

3.4 多个线性约束的检验

假设总体参数为 0 的原假设为排除性约束，同时包含多个参数的检验的原假设为多重约束，对多重约束进行的检验被称为多重假设检验或联合假设检验。

在排除性约束下的模型为受约束模型，它嵌套于不受约束模型之中，两个模型的自由度之差即为受约束的（即待检验的）自变量个数。

联合假设检验看的是不受约束模型和受约束模型之间的 SSR 相对变化，定义 F 统计量

$$F=\frac{(SSR_r-SSR_ur)/q}{SSR_ur/(n-k-1)}\sim F_{q,n-k-1}$$

其中 q 即为两个模型的自由度之差，是受约束的（即待检验的）自变量个数；F 统计量的分母是不受约束模型的误差的方差无偏估计。F 统计量的 SSR 表出形式是普适的，无论因变量在约束和不受约束模型是否一致。

如果拒绝原假设，即说受约束的自变量在对应显著性水平上是**联合统计显著**的，否则是联合不显著的。

联合假设检验可以穿透多重共线性，而多重共线性往往会干扰 t 检验的结果。实际上，单变量的双边 t 检验等价于对应的 F 检验，但 t 检验可以应对**单边**检验，且 t 统计量更**容易获得**。

当受约束模型和不受约束模型的因变量相同时，F 统计量还可以表示成 R^2 型。

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1-R_{ur})^2/(n-k-1)} \sim F_{q,n-k-1}$$

特别地，检验回归的整体显著性时，相当于对所有自变量施加约束，F 统计量的 R^2 型可以表示为

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

对于一般的线性约束，可以拆解为排除性约束和非排除性约束。对于后者，如 $\beta_1 = f(\vec{\beta})$，其中 $f(\vec{\beta})$ 是 $\vec{\beta}$ 的线性函数，仍然用 $\beta_1 = f(\vec{\beta}) + \theta$ 的方式最一般化地表出。一般来说，有常数会对应创建新的因变量，有其他回归系数的线性关系会对应创建新的自变量。

3.5 经济显著性

变量的经济显著性（或实际显著性）完全由系数的大小和符号决定；而统计显著性完全由 t 统计量的大小决定。

- 在样本容量扩大时有理由使用更小的显著性水平，经济上和统计上的显著性更可能达成一致，抵偿逐渐减小的标准误。
- 优先检查统计显著性，对于经济显著性应当注意变量在方程中出现的方式（比如单位、对数等）
- 不符合预期但不显著的变量可以忽略，但可能折射了模型变量的问题

4 多元回归分析：OLS 的渐进性

4.1 一致性

OLS 的一致性：在假定 MLR.1~MLR.4 下，对所有的 $j = 0, 1, ..., k$，OLS 估计量 $\hat{\beta}_j$ 都是 β_j 的一致估计。

实际上，一致性只需要假定误差项和单一自变量的零相关即可（相当于放款了 MLR.4），因此可以写出假定

MLR.4'-零均值和零相关：对所有的 $j = 1, 2, ..., k$，都有 $\mathrm{E}(u) = 0$ 和 $\mathrm{Cov}(x_{ji}, u) = 0$。

关于 MLR.4' 和 MLR.4 的讨论：

- MLR.4' 显得直接，因为这即为 OLS 推导过程的条件；
- 推导一致性时，满足 MLR.4' 即可，但这样的 OLS 估计值可能会有偏（但一致的）；
- MLR.4 更强，因为在有限样本下更希望考虑 OLS 估计量的精确性质；
- MLR.4 意味着模型正确地设定了总体回归函数，即 $\mathrm{E}(y|\vec{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$，于是可以得到解释变量对 y 的期望值的偏效应。如果只假设 MLR.4'，$\beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ 就不一定代表了总体回归函数，有可能面临自变量的非线性函数与误差相关的可能性。

如果 u 和 $x_1, x_2, ...x_k$ 中的任何一个相关，那么通常也会导致 OLS 估计量失去一致性；同时，估计量还失去了无偏性。

对于估计量 $\hat{\theta}$，不一致性可以表述为 $\mathrm{plim}\ \hat{\theta} - \theta$，这也可以称为渐进偏误。

在简单回归模型中，$\hat{\beta}_1$ 的不一致性为

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

可以进一步将其拓展，推导遗漏变量偏误的渐进类似情况。那么有 $\mathrm{plim}\ \hat{\beta}_1 = \beta_1 + \beta_2 \delta_1$，其中 $\delta_1 = \frac{\mathrm{Cov}(x_1, x_2)}{\mathrm{Var}(x_1)}$。因此，实际上可以把这种不一致性看作偏误，但区别在于，不一致性是用 x_1 的总体方差与 x_1 和 x_2 之间的总体协方差表示的，而偏误则基于其对应样本量（因为推导偏误时以特定样本为条件）。

不一致性和遗漏变量偏误既有联系也有区别，比如思考如下例子。给定总体回归方程为 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$，其中 x_1 和 x_2 是独立的。对于一个给定的样本，这时如果分别进行如下三个线性回归

$$\begin{aligned}
&(1) \qquad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\
&(2) \qquad \hat{y} = \tilde{\alpha}_0 + \tilde{\alpha}_1 x_1 \\
&(3) \qquad \hat{y} = \tilde{\gamma}_0 + \tilde{\gamma}_2 x_2
\end{aligned}$$

假设 x_2 对 x_1 的回归斜率系数为 $\tilde{\delta}_1$，若 x_1 和 x_2 在总体中不相关，那么 $\delta_1 = 0$，$\tilde{\alpha}_1$ 仍然是 β_1 的一致估计量，$\tilde{\gamma}_2$ 仍然是 β_2 的一致估计量。但遗漏变量偏误是基于特定样本的，在特定样本中通常 $\hat{\delta}_1 \neq 0$，因此 $\tilde{\alpha}_1$ 不一定是 β_1 的无偏估计量，$\tilde{\gamma}_2$ 也不一定是 β_2 的无偏估计量。

4.2 渐进正态和大样本推断

注意，**正态性对于 OLS 的无偏性不起任何作用，也不影响 OLS 在 MLR.1~MLR.5 之下成为最优线性无偏估计的结论**。

尽管 y_i 不是来自一个正态分布，但仍然可以用到中心极限定理证明 OLS 的估计量满足渐进正态性——在大样本容量的情况下，OLS 估计量是近似正态分布的。

OLS 的渐进正态性：在 MLR.1~MLR.5 下，

- $\sqrt{n}(\hat{\beta}_j - \beta_j) \overset{a}{\sim} N(0, \sigma^2/a_j^2)$，其中 σ^2/σ_j^2 是渐进方差，斜率系数 $a_j^2 = \text{plim}(\frac{\sum_{i=1}^n \hat{r}_{ij}^2}{n})$，$\hat{r}_{ij}$ 是 x_j 对其余自变量进行回归所得到的残差。称 $\hat{\beta}_j$ 是渐近正态分布的
- $\hat{\sigma}^2$ 是 $\sigma^2 = \text{Var}(u)$ 的一个一致估计量
- $(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j) \overset{a}{\sim} N(0,1), (\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \overset{a}{\sim} N(0,1)$

注：从渐近的观点来看，$\hat{\sigma}$ 和 σ 是 “等价” 的，因此无论是考虑 $\text{sd}(\hat{\beta}_j)$ 还是 $\text{se}(\hat{\beta}_j)$，标准化的 $\hat{\beta}_j$ 都服从于渐近标准正态分布

在大样本之下，$\hat{\beta}_j$ 的估计方差的形式也和此前的相似。

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

随着样本容量的扩大，$\hat{\sigma}^2$ 概率收敛为 σ^2，$SST_j \to n\sigma_j^2$，因此 $\widehat{\text{Var}(\hat{\beta}_j)}$ 以 $\frac{1}{n}$ 的速度收敛至零。

5 MLR Analysis: Further Issues

5.1 More on Functional Form

5.1.1 Variables Included

Adding regressors may reduce the error variance, but may also exacerbate multicollinearity problems. Variables that are uncorrelated with other regressors should be added because they reduce error variance without increasing multicollinearity. However, such uncorrelated variables may be hard to find. If excessive numbers of variables are held fixed, this may betray your research purpose!

5.1.2 Log-Even Model

It's learned that logarithmic functional forms are more convenient for percentage and elasticity interpretation, and slope coefficients of logged variables are invariant to rescalings. Noteworthy that, taking logs often eliminates problems with outliers and helps to secure normality and homoskedasticity. However, variables measured in percentage points like education rate, or in units such as years should not be logged, and variables should be non-negative!

When $\log y$ is the dependent variable is helpful, this helps to predict y, say $\log y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$. Equivalently, rewrite this as $y = \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k) \cdot \exp(u)$. We will show that it is hard to reverse the log-operation when constructing predictions.

Under the additional assumption that u is independent of $x_1, ..., x_k$:

$$E(y|x) = \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k) \cdot E(\exp(u))$$

Then from the PRF, the SRF would be

$$\hat{y} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k) \cdot (\frac{1}{n} \sum_{i=1}^{n} \exp(\hat{u}_i))$$

From Jensen's inequality,

$$E(e^u) \geq e^{E(u)} = 1$$

For $\frac{1}{n} \sum_{i=1}^{n} exp(\hat{u}_i)$, such a nonlinear function, it is hard to guarantee it to be an unbiased estimator of $\exp(u)$. However, for cases with large sample, consistency still holds, then plim $\hat{y} = y$.

5.1.3 Quadratic Form

Quadratic form fits better when dependent variable and independent variable(s) are nonlinearly-correlated. However, pay attention to the pinnacle of the quadratic function and your interpretation on both sides!

Sometimes, the "unwanted" relationship between DV and IV on the "unwanted" side does not make sense. In this case, be careful about how many observations in the sample lie in the "unwanted" side (before or after the turnaround point). If indeed that "unwanted" side's relationship does not make sense, and the proportion of sample lying there is non-negligible, then the regression model has to be refined!

5.1.4 Interaction Terms

Interaction term is a representation of mediation effect, and interaction effects make our interpretation of parameters more complicated.

Start from a concrete example, if we are to construct the following regression with interaction term

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u \\ \implies \quad & \frac{\Delta y}{\Delta x_2} = \beta_2 + \beta_3 x_1 \end{aligned}$$

Note that slope coefficients may no longer make sense! Theoretically, β_2 means the effect of x_2 on y, only when $x_1 = 0$; when $x_1 \neq 0$, β_2 does not fully cover the effect of x_2 alone; in some cases, $x_1 = 0$ is even impossible!

Hence, we do the reparametrization and center the data:

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u$$

Note that μ_1, μ_2 stand for population means. They may be replaced by sample means when conducting the regression. Here, δ_2 means the effect of x_2 if all variables take on their mean values, which is more reasonable. If necessary, interaction may be centered at other interesting values. Another advantage of reparametrization is that, it will make standard errors for *partial* effects at the mean (as in the example above) values available!

More generally speaking, in models with quadratics, interactions, and other nonlinear functional forms, the partial effect depends on the values of one or more explanatory variables. Average Partial Effect (APE) is a summary measure to describe the relationship between dependent variable and each explanatory variable. The computation is clear, after getting the partial effect and plugging in the estimated parameters, average partial effects for each unit across the sample. For quadratics or interactions, centering the explanatory variable will ensure the coefficient is APE itself.

5.2 Goodness of Fit

General remarks on R^2:

- A high R^2 does not imply that there is a causal interpretation.
- A low R^2 does not preclude the precise estimation of partial effects.

Recap what we have learned for R^2:

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{SSR/n}{SST/n} \to 1 - \frac{\sigma_u^2}{\sigma_y^2}$$

which is an estimate for $1 - \frac{\sigma_u^2}{\sigma_y^2}$.

A better estimate should take into account degrees of freedom. The adjusted R^2 is then

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

Clearly, the adjusted R-squared imposes a penalty for adding new regressors. The adjusted R-squared increases if and only if, the t statistic of a newly-added regressor is greater than 1 in absolute value.

The relationship between R^2 and adjusted R^2 is

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \cdot (1 - R^2)$$

Note that the adjusted R^2 may even get negative.

One advantageous application of adjusted R^2 is that, $\bar{R}^2$ can be used to compare and choose between nonnested models. In contrast, a comparison between the ordinary R^2 of both nonnested models would be unfair, because they may differ in the number of parameters, forms of parameters. However, neither ordinary R^2 or adjusted R^2 can be used to compare models which differ in their definition of the dependent variable.

6 MLR Analysis with Dummy

A way to incorporate qualitative information is to use dummy variables. They may appear as the dependent (Linear Probability Model) or independent variables.

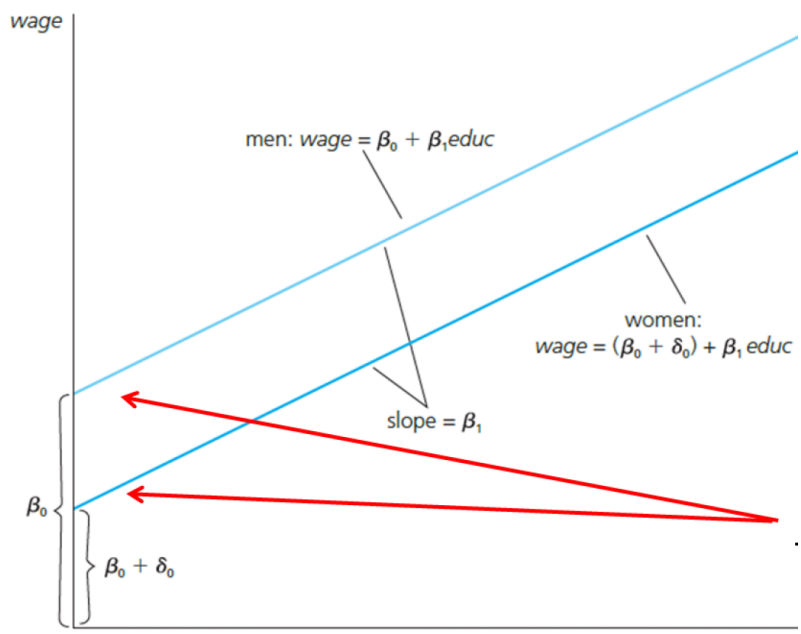
6.1 Different Intercepts: Dummy as Regressors

A Dummy is also called a 0-1 variable, or a binary variable. It does not matter whether to code the dummy variable under certain state as 0 or 1. Anyway, bear in mind what the base group (or, benchmark group) is!

Start with a simple example where $d_0 = 0$ implies an observation belongs to the base group, $d_0 = 1$ otherwise.

$$y = \beta_0 + \delta_0 d_0 + \beta_1 x_1 + u$$

For the base group, its intercept is β_0; for its counterpart, the intercept is $\beta_0 + \delta_0$. δ_0 represents an intercept shift on graphical illustration.



Generally speaking, the slope coefficient of the dummy variable indicates the difference in the mean of dependent variable across the two states, holding others factors constant.

6.1.1 Dummy Variable Trap

Mind yourself of Dummy Variable Trap! If the number of dummy variables describing one attribute equals or even exceeds its levels, then perfect collinearity creeps in!

Alternatively, one could include all dummies representing all levels and **omit the intercept term**,

$$y = \gamma_0 d_0 + \gamma_1 d_1 + \beta_1 x_1 + u$$

In this case, the estimated difference between groups is $|\gamma_1 - \gamma_0|$. However, in this method, we can only get the correct estimation for each coefficient, but the differences between the parameters (groups) are harder to be tested. Moreover, R^2 is invalid, since the intercept term is removed.

6.1.2 Incorporating Ordinal Information Using Dummies

To study ordinal variables, start with a case where we are interested in the relationship between city credit ratings (CR, integer ranging from 0 to 4) and municipal bond interest rates (MBR). One possible and straightforward way is to estimate

$$MBR = \beta_0 + \beta_1 CR + o.f.$$

which is the fixed partial effect model.

However, the differences between any two adjacent levels may not hold constant across all levels. They only stands for ordinal meanings, instead of interval meanings. Since most of the time ordinal variable only takes limited values, one better solution is to define a dummy for each level of the ordinal variable (Pay attention to dummy variable trap meanwhile!)

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_1 CR_2 + \delta_1 CR_3 + \delta_1 CR_4 + o.f.$$

Note that $CR = 0$ is the base group here. After introducing dummies to represent every possible levels of the ordinal variable, we allow for more possibilities of the effect of level change.

Fixed partial effect model is a special case. The constraint is $\delta_2 = 2\delta_1, \delta_3 = 3\delta_1, \delta_4 = 4\delta_1$. Plug in the constraint back to get the restricted model as

$$\begin{aligned} MBR &= \beta_0 + \delta_1(CR_1 + 2CR_2 + 3CR_3 + 4CR_4) + o.f. \\ &= \beta_o + \delta_1 CR + o.f. \\ &\Longleftarrow \quad H_0 : \delta_2 = 2\delta_1, \delta_3 = 3\delta_1, \delta_4 = 4\delta_1 \end{aligned}$$

Sometimes, the ordinal variable of interest takes too many values to include each level as a dummy. In this case, we can classifiy the ordinal variables into several categories by dividing them according to some ranges.

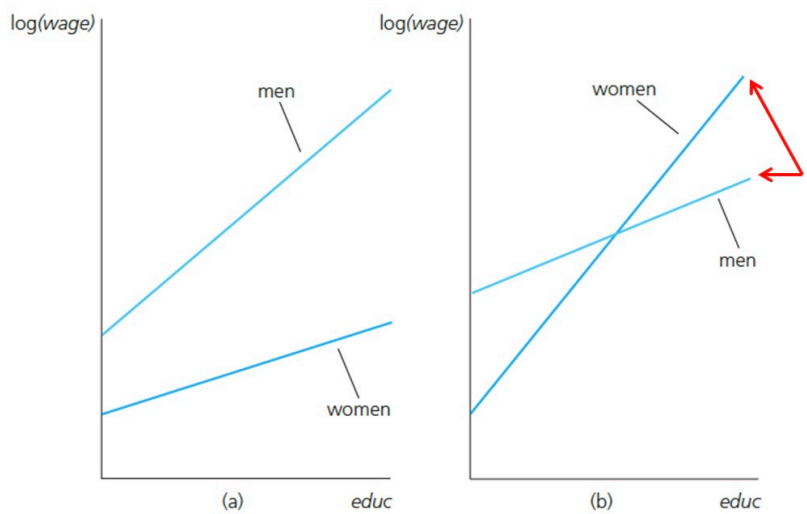
6.2 Different Slopes: Interaction with Dummy

6.2.1 Graphical Illustration

An interaction term of dummy variable(s) allows for different slopes between groups

$$y = \beta_0 + \beta_1 x_1 + \delta_0 d_0 + \delta_1 (d_0 \cdot x_1) + u$$

where $\beta_0 + \delta_0$. δ_0 represents an intercept shift on graphical illustration.



where δ_1 means the partial effect of x_1 on y in the base group and its counterpart differs for δ_1. If $\delta_1 = 0$, this means partial effect of x_1 is the same for both groups. If $\delta_0 = \delta_1 = 0$, this means there is no difference between groups, and the whole equation is the same! Also note that, if an interaction term is introduced into the model, be careful about the interpretation of β_1! β_1 equals to x_1's partial effect if, and only if $d_1 = 0$. Pay attention to the same issue with δ_0. In order to make δ_0 practically meaningful and reasonable, the common trick is to centralize x_1 to its mean.

6.2.2 Chow Test

From the simple case above, we have grasped the overall picture of detecting differences between groups using dummies and even interaction terms with dummies. In order to prove that the dummy classification has no effect (or difference between groups), we need to conduct a joint test about all slope coefficients for regressors with dummies. Take s step further, if we aim to test if the regression function differs for different groups, we need to introduce all interaction terms into the model (including interaction with intercept, i.e., dummy itself). This is practical if explanatory variables are not that much; however, if we have a lot of explanatory variables, we can turn to an alternative way to compute F-statistic. Chow-test is such a convenient way. Steps for Chow test follows as

1. Run separate regressions for the base group and its counterpart; the unrestricted SSR is given by the sum of the SSR of these two regressions. ($SSR_{ur} = SSR_1 + SSR_2$)

2. Run regression for the restricted model and store restricted SSR_p (**pooled** data this time).

3. Compute F statistics as

$$F = \frac{(SSR_p - SSR_{ur})/(k+1)}{SSR_{ur}/(n-2\cdot(k+1))} = \frac{[SSR_p - (SSR_1 + SSR_2)]}{SSR_1 + SSR_2} \cdot \frac{n-2(k+1)}{k+1}$$

   Note that the test assumes a constant error variances across groups.

The hypothesis for Chow Test is that, there is no difference between groups, including all the partial effects and the intercept. However, if we allow for the differences in intercepts across groups and then test the differences in slope coefficients, such test would be more meaningful. In this way, the number of constraints is k, instead of $k+1$.

Moreover, Chow Test can be applied to test structural change over time (including intercept change here). The principle is still the same as before, but this time we turn to dummies that represents time periods. First, for each period in the T periods of interest, conduct a regression and obtain its SSR. The unrestricted SSR_{ur} is then $SSR_{ur} = SSR_1 + SSR_2 + ... + SSR_T$. If there are k explanatory variables and T periods, we have to test for $(T-1)(k+1)$-many constraints. In the unrestricted model, there are $T(k+1)$ parameters to estimate. Suppose $n = n_1 + n_2 + ... + n_T$ is the total number of observations, then degrees of freedom for F statistic is $(T-1)(k_+1)$ and $n - T(k+1)$. Note again that Chow Test cannot tolerate with heteroskedasticity. If with heteroskadasticity, we have to actually construct all the interaction terms and conduct a regression (i.e., the direct method).

6.3 Binary Regressand: Linear Probability Model

When y is binary, our model turns out to be a linear probability model (LPM).

$$\begin{aligned}
& y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \\
\implies \quad & E(y|\vec{x}) = 1 \cdot P(y=1|\vec{x}) + 0 \cdot P(y=0|\vec{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \\
\implies \quad & P(y=1|\vec{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \\
\implies \quad & \beta_j = \frac{\Delta P(y=1|x)}{\Delta x_j}
\end{aligned}$$

In the linear probability model, the coefficients describe the effect of the explanatory variables on the probability that $y = 1$.

One of the most obvious drawback of the linear probability model is that, all observed factors in the model are considered of ***linear*** effect. However, actually most of the influence may be nonlinear but based on its level. Other disadvantages of the linear probability model include

- Predicted probabilities may be larger than one or smaller than zero.
- Marginal probability effects sometimes logically impossible.
- The linear probability model is necessarily heteroskedastic. Thus, heteroskedasticity consistent standard errors need to be computed.

$$Var(y|\vec{x}) = P(y = 1|\vec{x})[1 - P(y = 1|\vec{x})]$$

Clearly, the variance of y depends on x. However, in the classical linear model's assumptions (CLM), the variance should have nothing to do with $\vec{x}$.

6.4 Problem of Self-Selection

Consider a PRF as

$$E(y|w, x) = \alpha + \tau w + \gamma_1 x_1 + ... + \gamma_k x_k$$

where w is a treatment indicator. $w = 1$ when the treatment has been applied. Denote $y(0)$ as the outcome of y when $w = 0$, $y(1)$ when $w = 1$, then

$$y = (1 - w)y(0) + wy(1)$$

We include x_1 through x_k to account for the possibility that the treatment w is not randomly assigned and clean of self-selection problem (to get closer to random assignment into the treament and control group).

Now we need to make strong assumption that w is independent of $[y(0), y(1)]$ conditional upon $x_1, ..., x_k$. This is known as **regression adjustment** and allows us to adjust for differences across units in estimating the causal effect of the treatment.

We can relax the assumption of a constant treatment effect. We allow the treatment effect to vary across observations and estimate the average treatment effect (ATE).

$$y_i = \alpha + \tau w_i + \gamma_1 x_{i1} + ... + \gamma_k x_{ik} + \delta_1 w_i(x_{i1} - \bar{x}_1) + ...{}_+ \delta_k w_i(x_{ik} - \bar{x}_k) + u$$

where the estimated coefficient on w will be the ATE. The regression that allows individual treatment effects to vary is known as the unrestricted regression adjustment (URA). URA more is closer to the reality, since the individual effects won't be homogenious for all. By contrast, a restricted regression adjustment (RRA) forces the treatment effect to be identical across individuals.

Note that tt doesn't matter whether $\gamma_j x_{ij}$ is centralized, but it matters whether interaction terms are centralized, otherwise the interpretation of τ is affected.

Here is an alternative method for obtaining the URA ATE. For control group and treatment group, estimate the following equation respectively

$$y_i = \alpha + \gamma_1 x_{i1} + ... + \gamma_k x_{ik}$$

$$\Longrightarrow \quad \begin{cases} \text{Control: } \hat{y}_i^{(0)} = \hat{\alpha} + \hat{\gamma}_{0,1} x_{i1} + ... + \hat{\gamma}_{0,k} x_{ik}, \text{ using control observations} \\ \text{Treatment: } \hat{y}_i^{(1)} = \hat{\alpha} + \hat{\gamma}_{1,1} x_{i1} + ... + \hat{\gamma}_{1,k} x_{ik}, \text{ using treatment observations} \end{cases}$$

Now for every unit in the sample, predict $y_i(0)$ and $y_i(1)$ regardless of whether the unit is in the control of treatment groups. Then use these predicted values to compute the ATE as

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i^{(1)} - \hat{y}_i^{(0)}))$$

Though this yields the same ATE as running the regression with interaction terms, computing a standard error by hand can be tricky.

7 Heteroskedasticity

7.1 Properties of OLS under Heteroskedasticity

Assumptions for **Homoskedasticity**: $Var(u_i|x_i) = \sigma^2$.

7.1.1 Unbiasedness & Consistency

If we require OLS estimators to have some "good" properties,

- Unbiasedness: $E(u_i|x_i) = 0$ (zero conditional mean)
- Consistency: $E(u_i|x_i) = 0$ (zero conditional mean)

Based on those, even with heteroskedasticity,

- OLS still unbiased and consistent under heteroskedasticity;
- Interpretation of R^2 is not changed.

$$\text{plim}_{n\to\infty} \frac{1}{n} SSR = \sigma_u^2$$
$$\text{plim}_{n\to\infty} \frac{1}{n} SST = \sigma_y^2$$
$$R^2 = 1 - \frac{SSR}{SST} \xrightarrow{p} 1 - \frac{\sigma_u^2}{\sigma_y^2}$$

This is because SSR and SST measure the unconditional variance of error and DV respectively. Unconditional error variance is **unaffected** by heteroskedasticity (which refers to the conditional error variance). However, heteroskedasticity invalidates variance formulas for OLS estimators, and OLS is no longer the Best Linear Unbiased Estimator (BLUE).

7.1.2 Variance

Consider the simple linear regression $y_i = \beta_0 + \beta_1 x_i + u_i$, with $Var(u_i|x_i) = \sigma_i^2$ representing the general form for heteroskedasticity.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$Var(\hat{\beta}_1|\vec{x}) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 Var(u_i|x_i)}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2}$$

Note that $Var(\hat{\beta}_1|\vec{x}) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2}$ is valid regardless of heteroskedasticity or homoskedasticity, while $Var(\hat{\beta}_1|\vec{x}) = \frac{\sigma_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$ holds only for homoskedasticity.

However, from a given sample, we can only make inference based on unbiased estimators for σ_i^2.

$$\widehat{Var(\hat{\beta}_1)} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \hat{u}_i^2}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2}$$
$$se(\hat{\beta}_1) = \sqrt{\widehat{Var(\hat{\beta}_1)}} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2 \hat{u}_i^2}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2}}$$

Note that this version of $se(\hat{\beta}_1)$ here is heteroskedasticity-robust! And all formulas are only valid in large samples.

$$\text{plim}_{n\to\infty}\ nVar(\hat{\beta}_1|\vec{x}) = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2}{\left(\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} = \frac{E\left((x-\mu_x)^2\sigma_i^2\right)}{(\sigma_x^2)^2}$$
$$\text{plim}_{n\to\infty}\ n\widehat{Var(\hat{\beta}_1}|\vec{x}) = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \hat{u}_i^2}{\left(\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} = \frac{E\left((x-\mu_x)^2\sigma_i^2\right)}{(\sigma_x^2)^2}$$
$$\Longrightarrow \quad \text{plim}_{n\to\infty}\ n\widehat{Var(\hat{\beta}_1}|\vec{x}) = \text{plim}_{n\to\infty}\ nVar(\hat{\beta}_1|\vec{x})$$

The first two equations above can be proved via LLN and CLT.

$$\text{plim}_{n\to\infty}\ \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2 = \text{plim}_{n\to\infty}\ \frac{1}{n}\sum_{i=1}^n (x_i - \mu_x)^2 \sigma_i^2 \overset{LLN}{\Longrightarrow} E\left((x-\mu_x)^2\sigma_i^2\right)$$
$$\text{plim}_{n\to\infty}\ \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \hat{u}_i^2 = \text{plim}_{n\to\infty}\ \frac{1}{n}\sum_{i=1}^n (x_i - \mu_x)^2 \sigma_i^2 \overset{LLN}{\Longrightarrow} E\left((x-\mu_x)^2\sigma_i^2\right)$$
$$\begin{aligned} E\left((x-\mu_x)^2\sigma_i^2\right) =& E\left[E\left[(x-\mu_x)^2\sigma_i^2|x_i\right]\right] \\ =& E\left[(x-\mu_x)^2 E(\sigma_i^2|x_i)\right] \\ =& E\left[(x-\mu_x)^2\sigma_i^2\right] \end{aligned}$$

In the MLR case,

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + u_i \widehat{Var(\hat{\beta}_j}|\vec{x}) = \frac{\sum_{i=1}^n \hat{r}_{ij}\hat{u}_i^2}{SSR_j^2}$$

Its corresponding $se(\hat{\beta}_j|\vec{x})$ is called White/Huber/Elcker standard errors. They involve the squared residuals from the regression and from a regression of x_j on all other exlanatory variables.

7.1.3 Hypothesis Test

Under the heteroskedasticity-robust OLS standard error, the usual t test is valid *asymptotically* (If with heteroskedasticity, the t statistic constructed on usual standard error is no longer valid.). The heteroskedasticity-robust t statistic differs with the traditional one only in its version of standard error!

$$t = \frac{\text{estimate - hypothesized value}}{\text{standard error}}$$
$$t_{\text{robust}} = \frac{\text{estimate - hypothesized value}}{\text{heteroskedasticity-robust standard error}}$$

Generally speaking (NOT theoretically), heteroskedasticity-robust standard errors may be *slightly larger* than their non-robust counterparts. If they show great difference, strong heteroskedasticity is indicated.

However, the usual F statistic does not work under heteroskedasticity, but heteroskedasticity-robust versions are available in statistical softwares (Too complicated to master).

7.2 Testing for Heteroskedasticity

7.2.1 Breusch-Pagan Test

Under MLR.4,

$$H_0 : Var(u|\vec{x}) = \sigma^2$$
$$Var(u|\vec{x}) = E(u^2|\vec{x}) - [E(u|\vec{x})]^2 = E(u^2|\vec{x})$$
$$\implies \quad E(u^2|\vec{x}) = E(u^2) = \sigma^2$$

which means the mean of u^2 must not vary with $x_1, x_2, ..., x_k$.

Consider the following equation,

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + v$$

If the null hypothesis holds, then the joint hypothesis $\delta_1 = \delta_2 = ... = \delta_k = 0$ cannot be rejected. However, the "real" u^2 cannot be observed, so fine to substitute u^2 with $\hat{u}^2$.

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + error$$

The error term incorporates two parts, $error_i = v_i + \hat{u}_i^2 - u_i^2$. With sufficiently large samples, $\hat{u}_i^2 - u_i^2$ is thought to converge to 0.

Accordingly, conduct the F test. Construct the F statistic

$$F = \frac{R^2_{\hat{u}^2}/k}{(1 - R^2_{\hat{u}^2})/(n-k-1)} \sim F_{k,n-k-1}$$

whose alternative is the Lagrange Multiplier Test with LM statistic.

$$LM = n \cdot R^2_{\hat{u}^2} \sim \chi^2_k$$

Note that a small $R^2_{\hat{u}^2}$ may not correspond to a small LM statistic, since the latter is then multiplied by the sample size n, which is potentially large.

Remarks

- Taking logarithm form of variables can alleviate the issue of heteroskedasticity to some extent, and this is accomplished mainly through the dependent variable, instead of the explanatory variable.

- BP Test only considers the linear function of $\vec{x}$.
- More on F and LM.
  - $F_{k,n-k-1} \overset{n\to\infty}{\Longrightarrow} F_{k,\infty} \overset{a}{\sim} \chi_k^2/k$
  - And if $X_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, then $\frac{1}{n}\sum_{i=1}^n X_i^2 \sim \chi_n^2/n$. When n approaches infinity, $\frac{1}{n}\sum_{i=1} X_i^2 \overset{p}{\to} E(X_i^2) = Var(X_i) = 1$, then $\chi_n^2/n \overset{p}{\to} 1, n \to \infty$.
  - $LM \overset{a}{\sim} k \cdot F, n \to \infty$

7.2.2 White Test

Since BP Test only considers the linear function of $\vec{x}$, White Test sees to it and take more higher-level items of independent variables into account. Regress squared residuals on all explanatory variables, their squaresm and interactions. Take a 3-explanatory-variable case for example:

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + error$$

The main drawback is that, the items needed to taken into account will expand greatly with more explanatory variables. Then more estimators are needed to be estimated, and eventually the power of the test shrinks.

Alternative form of the White Test:

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + error$$

Note that it is the estimated $\hat{y}$ that is incorporated into the regression. And the equation makes sense only when $\hat{y}$, instead of y itself is used.

The predicted $\hat{y}$ and its square implicitly contain all of these terms. Thus ,the regression above **indirectly** tests the dependence of the squared residuals on the explanatory variables, their squares, and interactions. And this alternative form of White Test is also called the "Special Case" for White Test.

7.3 Weighted Least Squares Estimation

Suppose that, heteroskedasticity is known up to a multiplicative constant, i.e., the functional form of the heteroskedasticity is known.

$$Var(u_i|\vec{x}_i) = \sigma^2 \cdot h(\vec{x}_i), h(\vec{x}_i) = h_i > 0$$

Assumptions of the multiplicative constant are kind of arbitraty. Discussions about underlying beliefs are needed.

Note that $h(\vec{x}_i) > 0$, which is the primary condition to check! Meanwhile, this constraint limits WLS's application.

7.3.1 Model Transformation for WLS

Naturally, we can get

$$Var\left(\frac{u_i}{\sqrt{h_i}}\middle|\vec{x}_i\right) = \frac{1}{h_i} \cdot Var(u_i|\vec{x}_i) = \sigma^2$$

From the traditional MLR model, we can get a transformed model to tackle with heteroskedasticity, taking the funtional form into account.

$$\begin{aligned}
& y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + u_i \\
\implies \quad & \left[\frac{y_i}{\sqrt{h_i}}\right] = \beta_0 \left[\frac{1}{\sqrt{h_i}}\right] + \beta_1 \left[\frac{x_{i1}}{\sqrt{h_i}}\right] + ... + \beta_k \left[\frac{x_{ik}}{\sqrt{h_i}}\right] + \left[\frac{u_i}{\sqrt{h_i}}\right] \\
\iff \quad & y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + ... + \beta_k x_{ik}^* + u_i^*
\end{aligned}$$

Note that h_i is a function of $\vec{x}_i$. In such transformed model, here is no intercept term. Also, the transformation factor does not necessarily have to be exactly $\frac{1}{\sqrt{h_i}}$, fine to be any **multiple** of $\frac{1}{\sqrt{h_i}}$.

For the transformed model, conduct regression using OLS.

$$\begin{aligned}
& \min_{b_0, b_1, ..., b_k} \sum_{i=1}^{n} \left(\left[\frac{y_i}{\sqrt{h_i}}\right] - b_0 \left[\frac{1}{\sqrt{h_i}}\right] - b_1 \left[\frac{x_{i1}}{\sqrt{h_i}}\right] - ... - b_k \left[\frac{x_{ik}}{\sqrt{h_i}}\right] \right)^2 \\
\iff & \min_{b_0, b_1, ..., b_k} \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_{i1} - ... - b_k x_{ik}\right)^2 / h_i
\end{aligned}$$

If h_i is a constant, i.e., independent of $\vec{x}_i$, then this algorithm of WLS coincides with OLS. The "weighted" in WLS lies in the h_i for i. Observations with a larger variance are less informative than those with smaller variance and therefore should get less weight. WLS estimators sometimes have considerably smaller standard errors, which is line with the expectation that they are more efficient.

WLS is a special case of generalized least squares. Under WLS, the corresponding deterministic coefficient is weighted R^2, $R^2 = 1 - \frac{wSSR}{wSST}$. (Not for you to master)

Moreover, if the other Gauss-Markov assumptions hold as well, OLS applied to the transformed model is the BLUE (Best Linear Unbiased Estimator).

Lastly, set a powerful reminder to yourself that, the interpretation of coefficients should come back to the **traditional** model; our transformation only aims to improve the effectiveness of estimation.

Also, note that both heteroskedasticity-robust standard error and WLS are ways to cope with issues of heteroskedasticity and propose a correction. If the estimated slope coefficients are close to each other under two methods (both in magnitude and error), just be confident about the cross-checked validation.

7.3.2 Important Special Case for WLS

If the observations are reported as average at the grand level (such as city, country, etc.), they should be **weighted** by the **size of the unit**. Suppose for firm i with number of employees to be m_i, and the indicator $y_{i,e}$ locates individual e in firm i.

$$\begin{aligned} & y_{i,e} = \beta_0 + \beta_1 x_{1,i,e} + \beta_2 x_{2,i,e} + \beta_3 x_{3,i} + u_{i,e} \\ \iff \quad & \frac{1}{m_i} \cdot \sum_{e=1}^{m_i} y_{i,e} = \beta_0 + \beta_1 \cdot \frac{1}{m_i} \cdot \sum_{e=1}^{m_i} x_{1,i,e} + \beta_2 \cdot \frac{1}{m_i} \cdot \sum_{e=1}^{m_i} x_{2,i,e} + \beta_3 x_{3,i} + \frac{1}{m_i} \cdot \sum_{e=1}^{m_i} u_{i,e} \end{aligned}$$

Thus for an entity i,

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{1,i} + \beta_2 \bar{x}_{2,i} + \beta_3 x_{3,i} + \bar{u}_i$$

It's intriguing that here $\bar{u}_i$ correlates with the sum-up level, i.e., size of the unit m_i.

$$Var(\bar{u}_i) = Var(\frac{1}{m_i} \cdot \sum_{e=1}^{m_i} u_{i,e}) = \frac{\sigma^2}{m_i}$$

where $u_{i,e}$ is i.i.d. with variance σ^2, i.e., errors are homoskedastic at the individual-level.

In sum, if errors are homoskedastic at the individual-level, WLS with weights equal to entity size m_i should be used. If assumption of homoskedasticity at the individual-level is not exactly satisfied, one can calculate robust standard errors after WLS (i.e., for the transformed model).

7.3.3 Unknown Heteroskedasticity Function in WLS

What we've discussed before comes with a strong assumption of known $h(\vec{x}_i)$. If $h(\vec{x})$ is unknown, given that $h(\vec{x}) > 0$ must hold, we can generally assume the following functional form:

$$Var(u|\vec{x}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + ... + \delta_k x_k) = \sigma^2 h(\vec{x})$$

where exp-function is used to ensure positivity.

Since $E(u^2|\vec{x}) = Var(u|\vec{x})$, we can write it as:

$$u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + ... + \delta_k x_k) = \sigma^2 h(\vec{x}) \cdot v$$

where v is the multiplicative error, assume to be independent of the explanatory variables. Then,

$$\implies \log(u^2) = \alpha_0 + \delta_1 x_1 + ... + \delta_k x_k + e$$

Substitute u^2 with sample-version $\hat{u}^2$.

$$\begin{aligned} & \log(\hat{u}^2) = \hat{\alpha}_0 + \hat{\delta}_1 x_1 + ... + \hat{\delta}_k x_k + error \\ \implies \quad & \hat{h}_i = \exp(\hat{\alpha}_0 + \hat{\delta}_1 x_1 + ... + \hat{\delta}_k x_k) \end{aligned}$$

And then $\hat{h}_i$ can be further used as weights in WLS. (Note that $\hat{h}_i$ may be a multiple of h_i, instead of right just the h_i.)

7.3.4 Misspecified Heteroskedasticity Function

If misspecified, WLS is still consistent under MLR.1~MLR.4. Convince yourself again that the heteroskedasticity has nothing to do with unbiasedness and consistency. However, if OLS and WLS produce very different estimates, this typically indicates that some other assumptions (e.g. MLR.4) are wrong.

If with strong heteroskedasticity, still often better to use a wrong form of heteroskedasticity in order to increase efficiency.

7.3.5 WLS in LPM

We have discussed heteroskedasticity in Linear Probability Model (LPM). Now we revisit it and see how WLS is summoned to cope with it.

$$\begin{aligned} & \Pr(y=1|\vec{x}) = p(x) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \\ \Longrightarrow \quad & Var(y|\vec{x}) = p(x)[1-(p(x)] \end{aligned}$$

In the LPM, the exact form of heteroskedasticity is known! Thus,

$$\Longrightarrow \hat{h}_i = \hat{y}_i(1-\hat{y}_i)$$

Remarks

- Infeasible if LPM predictions are below zero or greater than one. Otherwise, $\hat{h}_i$ would be non-positive.
  - If such cases are rare, they may be adjusted to values such as .01/.99 to stick with WLS.
  - Otherwise, probably better to use OLS with robust standard errors.

8 More on Specification and Data Issues

8.1 Test for Functional Form Misspecification

First revisit MLR.4 of Zero Conditional Mean:

$$E[u|\vec{x}] = 0$$

The most usual case that violates zero-conditional-mean assumption is, **misspecification of functional form**. (Leave out some important explanatory variables)

If MLR.4 doesn't hold, then the slope coefficients may **not** be unbiased and consistent. Even though there are only a few variables correlating with u, since other variables would inevitably correlate with them, the slope coefficients would be biased and inconsistent.

8.1.1 RESET (Regression Specification Error Test)

The idea of RESET is to include squares and possibly higher-order fitted values in the regression. (which is similar to the reduced White test.) Note that squared terms and interaction terms are included since $\hat{y}^2, \hat{y}^3$ have been induced.

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + error$$

We test for the exclusion of the $\hat{y}$ terms (both $\hat{y}^2, \hat{y}^3$). If they cannot be excluded, this is evidence for omitted higher order terms and interactions, which indicates the misspecification of functional form. Note that it's accepted and enough to test only $\hat{y}^2$ and $\hat{y}^2$.

We conduct a F test for $\hat{y}$ terms.

$$H_0 : \delta_1 = \delta_2 = 0$$

The constructed F statistic will asymptotically follows

$$F \sim F_{2,n-k-1-2}$$

The number of restricted coefficients is 2, and the degree of freedom for model's residuals is $n-1-(k+2) = n-k-3$.

Remarks

- If $\hat{y}$ itself is included, then multicollinearity problem arises.
- RESET provides little guidance as to **where** misspecification comes from. Even if significant, some higher order terms are omitted, but no further specific information can be implied.

8.1.2 Testing against Nontested Alternatives

Suppose we postulate two candidate model for y:

$$\text{Model 1: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$\text{Model 2: } y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + u$$

Neither Model 1 nor Model 2 is a special case under another. They are **nonnested** alternatives of each other. We can specify **either** Model 1 or Model 2 to be the null model. For example, if take Model 1 as H_0^1:

$$H_0^1 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$H_1^1 : y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + u$$

If take model 2 as H_0^2

$$H_0^2 : y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + u$$
$$H_1^2 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

In either cases, define a general and comprehensive model that contains both models as subcases, and test:

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \log x_1 + \gamma \log x_2 + u$$

Then the corresponding null hypothesis to H_0^1 is:

$$H_0^1 : \gamma_3 = \gamma_4 = 0$$

If H_0^1 cannot be rejected, we say the model under H_0^1 is **preferred**, but **not** necessarily *true* model. It's possible that neither Model 1 nor Model 2 is true model. True model stands, if and only if zero conditional mean holds. Similarly, we can treat model 2 as the null model and test against $H_2^0 : \gamma_1 = \gamma_2 = 0$. But note that which model is specified as null model really matters!

However, if $E(u|x_1, x_2) = 0$, which means the model under H_0^1 is true model, this will imply that $E(u|f(x_1), g(x_2)) = 0$ whenever $f(x_1), g(x_2)$ is different from x_1, x_2. That is true because starting from the true model, no more term should correlate with the error. Therefore, for any alternative hypothesis, theoretically it proves to be not preferred, since error in the model under alternative hypothesis is not considered practically. Jointly speaking, the nonnested model test has no real effect. An alternative way that enforces the consideration of the error in alternative hypothesis will make sense regardless of H_0's model's correctness:

1. Regress y on $\log x_1, \log x_2$, and obtain fitted value $\tilde{y}$.
2. Regress y on x_1, x_2 and $\tilde{y}$, and then check t statistic of coefficient of $\tilde{y}$.

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 \tilde{y} + u \Longrightarrow H_0 : \alpha_3 = 0$$

Remarks

- A **clear winner** need not emerge. In one pair of hypothesis, one model may be preferred; while in another pair of hypothesis, its counterpart may be preferred.
- The preferred model may still differ from the true one.
- This cannot be used if the models differ in **dependent** variable, i.e., the same y.

8.2 Proxy Variables

Some variables, though indispensable in the model, hard to interpret numerically. For those unobservable explanatory variables, the interpretations of them are rather opaque, so are their coefficients. And such tricky issue will further influence our estimation for rest of the slope coefficients. If this is the case, a **proxy** variable should be introduced for the unbiasedness of other variables of our interest.

8.2.1 General Approach

Assume that the population regression function is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

where x_3^* is unobservable. However, we have found a proxy x_3 for x_3^*.

$$x_3^* = \delta_0 + \delta_3 x_3 + v_3$$

Plug-in solution to omitted variable problem:

$$\text{Regress } y \text{ on } x_1, x_2, x_3$$

However, a "good" proxy variable must satisfy the following assumptions.

8.2.2 Assumptions

1. u is uncorrelated with x_1, x_2, x_3^*.
   - $E(u|x_1, x_2, x_3^*) = 0 \Longrightarrow Corr(x_3, u) = 0$. Note that x_3 differs from x_3^*!
   - The proxy is "just a proxy" for the omitted variable, and does **not** belong into the population regression. The proxy won't enter the true model, contrary to an explanatory variable.
2. v_3 is uncorrelated with x_1, x_2, x_3^*.
   - $E(x_3^*|x_1, x_2, x_3) = E(x_3^*|x_3) \Longrightarrow Corr(x_1, v_3) = Corr(x_2, v_3) = 0$.
   - The proxy variable is a good proxy for the omitted variable.

- i.e., using other variable in addition will not help predict the omitted variable.

Note that proxy variable is **not equivalent** to its corresponding unobserved variable.

8.2.3 Overall Model

Under these assumption,

$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + (\beta_3\delta_3)x_3 + (u + \beta_3 v_3)$$

In this regression model, the error term in uncorrelated with all the explanatory variables. Consequently, all slope coefficients will be correctly estimated using OLS, unbiased and consistent (i.e., correctly identified). Meanwhile, the coefficient for the proxy variable may also be of interest, and it is a multiple of the coefficient of the omitted variable.

8.2.4 Lagged DV as Proxy

Omitted unobserved factors may be proxied by the value of the dependent variable from an earlier time period.

For example, city crime rate has an inertia that when considering this year's crime rate, crime rate in the past year matters. Moreover, including the past crime rate will at least partly **control for** the many *omitted* factors that also determine the crime rate in a given year.

$$crime = \beta_0 + \beta_1 crime_{-1} + \beta_2 unem + u$$

Another way to interpret this equation is that, one compares cities which had the "same" crime rate last year. This avoids comparing cities that differ very much in unobserved crime factors.

8.3 Random Coefficient Model

Models with random slopes, are also called random coefficient models. Back to the simple linear regression model, however, in this time parameters for intercept and slope coefficients vary among individuals.

$$y_i = a_i + b_i x_i$$

It is pitiful that, with only dataset $\{y_i, x_i\}_{i=1}^n$, a_i, b_i cannot be estimated. Luckily, we can take apart each random coefficient into two part, the fixed (average) and the zero-mean random part.

$$y_i = (\alpha + c_i) + (\beta + d_i)x_i$$

where, α as the average intercept, and c_i as random component for intercept; β as the average slope, and d_i as random component for slope.

If we group all the deterministic parts together and attribute the random parts to the error term, we can obtain the following equation:

$$\Leftrightarrow y_i = (\alpha + \beta x_i) + (c_i + d_i x_i)$$

Assumptions are:

$$E(c_i|x_i) = E(d_i|x_i) = 0 \Leftrightarrow E(a_i|x_i) = E(b_i|x_i) = 0$$

which means the individual random components are independent of the explanatory variable.

Under the assumptions above, we can get

$$E(c_i + d_i x_i|x_i) = 0$$

then, for the random coefficient model, the slope coefficients are both **unbiased** and **consistent**.

However, the error term is inate with heteroskedasticity, since that will depend on x_i obviously.

$$\Longrightarrow Var(c_i + d_i x_i|x_i) = \sigma_c^2 + \sigma_d^2 x_i^2$$

Fortunately, though with heteroskedasticity, the specific functional form is known. WLS or OLS with standard errors will consistently estimate the **average** intercept and average slope in the population.

Finally note that, relatively precise as random coefficient model is, it is **incapable** of making predictions, since individuals have random parts and differ from each other.

8.4 Measurement Error

Consider two kinds of measurement error:

- on dependent variable
- on explanatory variable(s)

8.4.1 Dependent Variable

Suppose the regression function for the population is as follows:

$$y^* = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$$

where the starred y^* cannot be observed directly. In fact, you can only observed y with error e_0:

$$e_0 = y - y^*$$

Take the measurement error on dependent variable into account:

$$\begin{aligned} & y^* = y - e_0 = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \\ \Longleftrightarrow \quad & y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + (u + e_0) \end{aligned}$$

If e_0 is supposed to be uncorrelated with $x_1, ..., x_k$, i.e., $E(u+e_0|x_1, ..., x_k) = 0$, then all slope coefficients will be unbiased and consistent. Pratically, regress y on $x_1, ..., x_k$. This kind of measurement error imposes no influence on the estimation of slope coefficients.

However, the estimated slope coefficients will be less precise, since their variance will be larger because of measurement error e_0.

$$Var(u + e_0) = \sigma_u^2 + \sigma_e^2 > \sigma_u^2$$

8.4.2 Explanatory Variable

Suppose the regression function for the population is as follows:

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

where the starred x^* cannot be observed directly. The observed x deviates from true x^* systematically at e.

$$x = x^* + e$$

Taking measurement error into account, then the estimated regression is

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

Typically, there are two ways to treat the measurement error on explanatory variable.

If $Cov(x, e) = 0$, then $E(x^*|x) = x$. x is regarded as a proxy variable of x^*, and β_1 is unbiased and consistent.

If $x = x^* + e$, the observed x is pratically decomposed into two parts, unobserved part x^* and measurement error e. Instead of assuming linear independence of observed value x and measurement error e, we assume unobserved part x^* to be uncorrelated with e:

$$Cov(x^*, e) = 0$$

This assumption is so important that it is called classical errors-in-variables assumption, **CEV** in short.

To consider potential issues of endogeneity, compute $Cov(x, u - \beta_1 e)$:

$$Cov(x, u - \beta_1 e) = Cov(x^* + e, u - \beta_1 e) = -\beta_1 \cdot Cov(x, e) + Cov(u, e) + \beta_1 \cdot Cov(x^*, e)$$

where $Cov(u, e)$ is safely supposed to be 0. Thus under CEV,

$$Cov(x, u - \beta_1 e) = -\beta_1 \cdot Cov(x, e) = -\beta_1 \cdot Cov(x^* + e, e) = -\beta_1 \sigma_e^2$$

Suppose $y = \beta_0 + \beta_1 x + v$, and regress y on x,

$$\text{plim } \hat{\beta}_1 = \beta_1 + \frac{Cov(v, x)}{Var(x)} = \beta_1(1 - \frac{\sigma_e^2}{\sigma_x^2}) = \beta_1 \cdot \frac{\sigma_{x^*}^2}{\sigma_x^2 + \sigma_e^2} = \beta_1 \cdot \frac{Var(x^*)}{Var(x)}$$

Notice that $\frac{\sigma^2_{x^*}}{\sigma^2_x+\sigma^2_e} \in (0,1)$ holds forever, then $\hat{\beta}_1$ has a **attenuation bias** (bias towards 0) compared with β_1. The magnitude of the effect will be attenuated towards zero.

In MLR, the problem will be more tricky.

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + ... + \beta_k x_k + u, \text{ with } x_1 = x_1^* + e$$
$$\Longrightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + (u - \beta_1 e)$$

Possible that $(u - \beta_1 e)$ correlates with x_1. In this case, even if $(u - \beta_1 e)$ is uncorrelated with all other variables, as long as any of the remaining variables correlates with x_1, it's likely that all those slope coefficients will be biased and not consistent. However, the deterministic equation of plim $\hat{\beta}_j$ is rather complicated and hence not covered here.

8.5 Missing Data

Missing data is viewed as a special case of sample selection (= nonrandom sampling), as the observations with missing information cannot be used.

If the sample selection is based on **independent** variables, there's no problem as a regression conditions on the independent variables (i.e., exogenous sample selection). In general, sample selection is no problem if it is uncorrelated with the error term of a regression.

If the sample selection is based on the **dependent** variable or the error term, this entails a problem (i.e., endogenous sample selection).

Missing Completely at Random (MCAR) means that missingness is unrelated to both u and $x_1, ..., x_k$.

$$\Pr(m_k = 1 | u, x_1, ..., x_k) = \Pr(m_k = 1)$$

Compared to MCAR, Missing at Random means that

$$\Pr(m_k = 1 | x_1, ..., x_k) = \Pr(m_k = 1)$$

where the missing mechanism is allowed to correlate with $x_1, ..., x_k$, but not allowed with u.

8.5.1 Missing Indicator Method (MIM)

Suppose we are missing some information on explanatory variable x_k, and with full information on $y, x_1, ..., x_{k-1}$.

One conservative but direct way is to work only with complete cases and get the complete-case estimator. However, the cost is hefty, though robust. Missing Indicator Method (MIM), will make better use of current data.

MIM creators two new variables:

$$z_{ik} = \begin{cases} x_{ik}, & \text{if } x_{ik} \text{ is observed} \\ 0, & \text{if } x_{ik} \text{ is missing} \end{cases}$$

$$m_{ik} = \begin{cases} 1, & \text{if } x_{ik} \text{ is observed} \\ 0, & \text{if } x_{ik} \text{ is missing} \end{cases}$$

where m_{ik} is a missing data indicator.

Then, regress y_i on $x_{i1}, ..., x_{i,k-1}, z_{ik}, m_{ik}$, using **all** observations.

Unfortunately, MIM is valid only under a strong assumption:

$$Cov(x_k, x_j) = 0, \forall j \neq k$$

The assumption is too strong to hold in practice.

Note that omitting m_{ik} is the same as assuming $x_{ik} = 0$ whenever it is missing.

8.6 Outliers & Influential Observations

If outliers come from mistakes that occurred when keying in the data, one should just discard the affected observations. But if outliers are the result of the data generating process, the decision on whether or not to discard the outliers is not so easy.

8.6.1 Least Absolute Deviations Estimation (LAD)

Algorithm in OLS will exemplify the outliers' influence, and causing pratical shock to the estimated result. Leat Absolute Deviations Estimation is a competitive alternative.

The least absolute deviations estimator minimizes the sum of absolute deviations, which may be more robust to outliers as deviations are not squared.

$$\min_{b_0, b_1, ..., b_k} \sum_{i=1}^{m} |y_i - b_0 - b_1 x_{i1} - ... - b_k x_{ik}|$$

The least absolute deviation estimator estimates the parameters of the **conditional median**, instead of the conditional mean with OLS. The conditional median and conditional mean coincide only when u in $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$ is **symmetric**. Furthermore, the least absolute deviations estimators is a special case of quantile regression.

9 Panel Data

In panel data, each individual is assumed i.i.d. Specifically speaking, indicators within individual is allowed with cross-correlation, but correlations across individuals are not allowed.

9.1 DiD

Most of the time, we are interested in an event that happened in the timeline (often supposed to be exogenous) and see its influence. Naturally, we classify our data into the control and treatment group. We can construct a dummy as a cutoff to represent the happening of this event, such as year, province, country, etc. Usually, we assume parallel trend behind the control and the treatment group, and reason that any difference between groups are caused by the event. The key for DiD is to understand the "trend", for "parallel trend" and "difference in trend". Thus, what we are interested in is not the trend itself, but the difference between trends. Graphically speaking, differences in slopes. Hence, it is crystal clear that an interaction term is needed.

In the example, we are interested in the effect of building a new garbage incinerator on housing prices (price before and after the garbage incinerator was built). Variable *rprice* means the real housing price, *nearinc* is a dummy meaning whether this house is near the incinerator. A naive rookie may estimate such a simple model using the data in the year when the incinerator was built like $rprice = \gamma_0 + \gamma_1 nearinc + u$, and claim that γ_1 is the effect. However, the fact is that, regardless of whether or not the incinerator was built, those houses near the incinerator was far from center of the city and had a relatively lower price. This is the underlying "parallel trend". The trend itself is not of interest; an interaction term representing differences in slopes is indeed of need. (In order to detect slope differences, you need to first control for parallel trends, *nearinc* and the time dummy are indispensable.) Under such logic, it is smooth to understand difference-in-differences (DiD) estimator and why it is called that way.

Define a dummy $y81$. $y81 = 1$ if the observation comes from year of 1981 when construction of the incinerator started; $y81 = 0$ otherwise. Use the following equation to estimate difference-in-difference estimator and its standard error:

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 (y81 \cdot nearinc) + u$$

where δ_1 is what we want.

Generally speaking, DiD estimator can be used to estimate the effect of natural experiment or quasi-natural experiment. In a (quasi-) natural experiment, observations can be divided into control group and treatment group. (Quasi-) natural experiment is different from true experiment. In a true experiment, subjects are randomly assigned to either control or treatment group. However in a natural experiment, control and treatment group all come from a certain change in "policies". So, we have to control for the systematic difference between the two groups. To achieve that, we need data from (at least) two years, one before the

policy change, one after the policy change. We estimate the equation

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 (d2 \cdot dT) + o.f.$$

where C stands for control group, T for treatment group; $dT = 1$ if from a treatment group. $d2 = 1$ indicates the second period after the policy change.

Careful interpretation of the equation will help a lot to our understanding of DiD. The term $\beta_1 dT$ controls for fixed effect of the two groups. $\delta_0 d2$ controls for the time trend. $\delta_1(d2 \cdot dT)$ detects any differences in slopes after controlling for fixed effect and time trend.

	$Before$	$After$	$After - Before$
$Control$	β_0	$\beta_0 + \delta_0$	δ_0
$Treatment$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \delta_0 + \delta_1$	$\delta_0 + \delta_1$
$Treatment - Control$	β_1	$\beta_1 + \delta_1$	δ_1

Therefore, δ_1 measures the effect of policy, which is also the average treatment effect (ATE as covered before). From the table above, we can see there are two ways to understand δ_1:

1. $\hat{\delta}_1 = (\bar{y}_{1,T} - \bar{y}_{1,C}) - (\bar{y}_{0,T} - \bar{y}_{0,C})$
2. $\hat{\delta}_1 = (\bar{y}_{1,T} - \bar{y}_{0,T}) - (\bar{y}_{1,C} - \bar{y}_{0,C})$

The pratical meaning of $\hat{\delta}_1$ can be interpreted as it that, compare the difference in outcomes of the units that are affected by the policy change (= treatment group) and those not affected (= control group), before and after the policy was enacted. $\hat{\delta}_1$ represents before/after comparisons in "(quasi-)natural experiments", so DiD can be used to evaluate policy changes or other exogenous events. The substraction hopefully *controls for* any changes in *external factors* that are common to both the treated and control groups, which will be the case when we have *random* assignment. In this case, the DiD estimator can be interpreted as the average treatment effect.

Note that DiD only works if the difference in outcomes between two groups is not changed by other factors than the policy change (e.g. there must be no differential trends). Parallel trends assumes that any trends in the outcome y would trend at the same rate towards the same direction in the absence of the intervention. If the parallel assumption is violated, we cannot be sure that the DiD estimator is identifying the effects of the policy or simply other unaccounted factor causing different trends between these groups. However in such cases, we can add flexibility by adding an additional control group. In the regression equation, more interaction terms are considered, in order to account for possibly different trends in different levels of some other factors (e.g., difference-in-difference-in-differences estimators).

9.2 First Difference

In the example, we are interested in the effect of x_{it} on y_{it}. Assume that no other explanatory variables are available, we wonder whether it would be possible to estimate the causal effect of x_{it} on y_{it}. Luckily, if the individuals are observed for at least two periods, and other factors affecting y_{it} stay approximately *constant* over those periods.

$$y_{it} = \beta_0 + \delta_0 d_t + \beta_1 x_{it} + v_{it}, \text{where } v_{it} = a_i + u_{it}$$

where

- d_t: time dummy for the second second
- a_i: unobserved time-constant factors (= fixed effect)
- u_{it}: other unobserved factors (= idiosyncratic error)

However as has been emphasized intensively before, this simple regression may encounter issues of omitting variables. One possible way out is to control more factors as possible, but some are unobservable, some hard to control. Another way is to categorize the composite error v_{it} into two parts: the time-invariant a_i and time-variant part u_{it}, as we did above. The most direct and simple way is to pool the data and use OLS to estimate coefficients. By doing this, we must assume that v_{it} is uncorrelated with x_{it}. Even if we assume that idiosyncratic (time-varying) error is uncorrelated with x_{it}, as long as a_i correlates with x_{it}, pooling OLS will give us biased and inconsistent estimators. The error caused by this is also called heterogeneity bias. In most applications, we collect panel data mainly to consider the arbitrary correlation of unobserved fixed effect a_i and explanatory variables. Since a_i is time-invariant, we can take the difference of two adjacent years. For two distinct t such that $d_t = 0, 1$ respectively, we can estimate that

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

which is called the first-differenced equation. As long as the assumption that Δu_i is uncorrelated with Δx_i , the regression comes back to the ordinary case. The key assumption holds if for any period t, u_i is uncorrelated with explanatory variables in the two periods (strict exogeneity).

Note again that there may be arbitrary correlation between the unoserved time-invariant characteristics and the included explanatory variables. If a_i correlates with x_{it}, the error term v_{it} will correlate with x_{it}, then both δ_0, β_1 will be biased. OLS in the original equation would therefore fail. The first-differenced panel estimator is thus a way to consistently estimate causal effects in the presence of time-invariant endogeneity. Because a_i will be eliminated, **any form of the correlation** of α_i and x_i is allowed, since all that will be offset in first difference. In other words, α_i is free to be general. However, "general" is no equivalent to "random", and α_i is fixed.

However, in order to eliminate a_i, first differencing may cut down the variation in explanatory variables greatly. Though x_{it} may vary a lot in each period, it is still possible that Δx_i does not vary much. Little variation in explanatory variables will lead to a relatively large standard error, and less precision of estimation. Another

issue with first differencing is that, it cannot be used to estimate time-invariant variable's effect, because all that will be eliminated once you do the first differencing (such as education level, gender, etc.). As was mentioned, strict exogeneity is a critical assumption. If such assumption is violated, first-differenced estimator will lose consistency.

One last note. First differenced equation will yield the same result as DiD if x_i here is a dummy variable. In DiD, we introduce the term $\beta_1 dT$ to control for fixed effect of the two groups, $\delta_0 d2$ to control for the time trend. $\delta_1(d2 \cdot dT)$ detects any differences in slopes after controlling for fixed effect and time trend. In first differenced equation, first differencing comes first to wipe out the fixed effect; the time trend is attributed to the intercept term, and the $\beta_1 \Delta x_i$ term by definition is equivalent to that interaction term.

First differencing can be applied to multi-period data. Take a concrete example of a panel data in which each person has three observations across three periods.

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{it} + u_{it}$$

where two time dummies are introduced to allow for different intercepts across periods. Take the first difference and we will get

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta x_{it} + \Delta u_{it}$$

where $\Delta d2_t = 1, \Delta d3_t = 0$ when $t = 2$; $\Delta d2_t = -1, \Delta d3_t = 1$ when $t = 3$. Note that the equation above does not include an intercept, which is not convenient for computing R^2. Better to estimate an equivalent equation like

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \beta_1 \Delta x_{it} + \Delta u_{it}$$

First differenced equation requires no serial correlation. It is easy to check this requirement. Let $r_{it} = \Delta u_{it}$, if r_{it} follows $r_{it} = \rho r_{i,t-1} + e_{it}$, the requirement demands $\rho = 0$. We can check it by testing against the null hypothesis $H_0 : \rho = 0$, and draw the conclusion based on t statistics.

9.3 Fixed Effects Estimation

Same as before, start with the simple equation of

$$y_{it} = \beta_1 x_{it} + \alpha_i + u_{it}$$

where α_i is the fixed effect for individual i, u_{it} is the idiosyncratic error.

9.3.1 Within Estimator

Use **time variation** with cross-sectional units to get **within estimators**.

From the regression equation, first sum it up and then get the average.

$$\frac{1}{T}\sum_{i=1}^{T} y_{it} = \beta_1 \cdot \frac{1}{T}\sum_{i=1}^{T} x_{it} + \alpha_i + \frac{1}{T}\sum_{i=1}^{T} u_{it}$$

which for simplicity can be written as

$$\bar{y}_i = \beta_1 \bar{x}_i + \alpha_i + \bar{u}_i$$

Then, we can get the new regression equation under fixed effect.

$$y_{it} - \bar{y}_i = \beta_1(x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i)$$

Or equivalently,

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{u}_{it}$$

where $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$ (same for $\ddot{x}_{it}, \ddot{u}_{it}$, and the fixed effect α_i is by nature *removed*. $\ddot{y}_{it}, \ddot{x}_{it}$ are called **time-demeaned** data. Such transformation for fixed effect model is also called within transformation. The demeaned equation is without intercept term. If you wonder why I did not include an intercept term like β_0 in the original equation, you can do it yourself the demeaning process and see if β_0 will be eliminated. One comfortable justification is that, "intercept" is perfectly absorbed by individual fixed effect α_i.

However, the basic logic under within estimator happens to be its drawback. If some regressors of interest are **invariant** to time, only individual-dependent, they will be canceled out. Gender and education background are typical examples.

When it comes to the estimation of individual fixed effect, since

$$\bar{u}_i = \frac{1}{N}\sum_{i=1}^{N} u_{it} \xrightarrow{p} 0$$

We have

$$\hat{\alpha}_i = \bar{y}_i - \beta_1 \bar{x}_i$$

If the original regression equation has k-many explanatory variables, after within transformation, the demeaned equation for fixed effect will have k-many variables (But note that the equation does not have an intercept!). However, in order to estimate the average over time for each individual, we will lose 1 df for each average we estimated. Therefore, the degree of freedom for residual is $df = NT - N - k = N(T-1) - k$.

9.3.2 Between Estimator

On the other hand, **between estimator** does *not* consider time variation for each individual.

Use $\bar{y}_i = \beta_1 \bar{x}_i + \alpha_i + \bar{u}_i$ for $i = 1, 2, ..., N$. and regress $\bar{y}_i$ on $\bar{x}_i$ to get the regressed equation. The corresponding coefficients are called between estimator. Since α_i is allowed to have unknown correlation with x_i, slope

coefficients may be biased and inconsistent. If we think a_i is uncorrelated with x_i, using random effect estimator might be more powerful. Between estimator ignores the important information across time.

Note that, for unbiasedness and consistency, strict exogeneity is required under $\bar{y}_i = \beta_1 \bar{x}_i + \alpha_i + \bar{u}_i$, i.e., u_{it} is not correlated with x_{it} at any time period, including the past, present and future.

9.3.3 Discussions

Within Estimator

- The R^2 of the demeaned equation is inappropriate, since α_i is not considered here.
- The effect of time-invariant variables cannot be estimated.
  - However, the effect of interactions for time-variant variables with time-invariant ones can be estimated.
- If a **full** set of time dummies are included, the effect of variables whose change over time is constant cannot be estimated. (e.g., working experience by year)
  - Multicollinearity
- Degrees of freedom have to be adjusted, since the N time averages are estimated in addition.
  - Resulting degrees of freedom to be $NT - N - K$.

Alternative Interpretation of Fixed Effects

Another way to interpret fixed effects is to introduce a dummy for each individual in the original regression.

$$y_{it} = a_1 ind1_{it} + a_2 ind2_{it} + ... + a_N indN_{it} + \beta_1 x_{it} + u_{it}$$

where $indK_{it}$ indicates if the observation stems from individual K. If so, $indK_{it} = 1$; and 0 otherwise. The equation does not fall into dummy variable trap since intercept term is not included. However, in model with fixed effect α_i, the story would be reversed. This is because with the full set of N-many indicators and the intercept, multicollinearity problem will arise. When N is large, this method will be impratical.

Fixed Effect v.s. First Differencing

- In the case $T = 2$, fixed effects and first differencing are identical.
  - Note that time dummy is naturally required in first differencing model, but that is not the case for fixed effect model. For equivalence, we have to include a time dummy to represent the second period.
  - First differencing is more straightforward. Heteroskedasticity-robust standard errors are easier to compute.
- For $T > 2$, fixed effects is more efficient if classical assumptions hold.
- First differencing may work better in the case of severe serial correlation in the errors.

  - $u_{it} = \rho u_{i,t-1} + v_{it}$. The errors follow a random walk if $\rho = 1$. Then, $\Delta y_{it} = \beta_1 \Delta x_{it} + v_{it}$ makes you relieved.
  - If T is large relative to N, the panel has a pronounced time series character.
- In practice, better to compute both and check robustness.

Unbalanced Panel Data

A panel data is unbalanced when not all cross-sectional units has the same number of observations. If you are to pick between FE and FD to deal with unbalanced panel data,

- FE considers average value, and some missing values are tolerated.
- FD requires that each observation have available data for both t and $t-1$.

FE generally preserves more data than FD with unbalanced panels.

9.4 Random Effects Estimation

Consider a random effect model as

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + u_{it}$$

where α_i is assumed to be random, and $Cov(x_{it}, \alpha_i) = 0$, i.e., completely unrelated to explanatory variable(s).

Then, the composite error, $v_{it} = \alpha_i + u_{it}$, is uncorrelated with the explanatory variables. But v_{it} is serially correlated for observations coming from the same individual i.

$$\begin{aligned} Cov(v_{it}, v_{is}) &= Cov(\alpha_i + u_{it}, \alpha_i + u_{is}) \\ &= Cov(\alpha_i, \alpha_i) + Cov(\alpha_i, u_{is}) + Cov(\alpha_i, u_{it}) + Cov(u_{it}, u_{is}) \\ &= \sigma_\alpha^2 + 0 + 0 + 0 \\ &= \sigma_\alpha^2 \neq 0 \end{aligned}, \forall t \neq s$$

under the assumption that idiosyncratic errors are serially uncorrelated, i.e., $Cov(u_{it}, u_{is}) = 0$.

$$Corr(v_{it}, v_{is}) = \frac{Cov(v_{it}, v_{is})}{\sqrt{Var(v_{it})Var(v_{is})}} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2} > 0$$

If OLS is used, standard errors have to be adjusted for the fact that errors are correlated over time for given i, i.e., take the serial correlation structure into account. Because of the serial correlation, OLS is not efficient. We're to find a transformation to the model so that GM-assumptions would hold. As in fixed effect model, time-demeaned equation might be a candidate.

$$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + (v_{it} - \bar{v}_i)$$

However, $(v_{it} - \bar{v}_i)$ is still serially-correlated.

We plug in a λ to do the transformation

$$y_{it} - \lambda\bar{y}_i = \beta_1(x_{it} - \lambda\bar{x}_i) + (v_{it} - \lambda\bar{v}_i)$$

in the hope that $e_{it} = v_{it} - \lambda\bar{v}_i$ would be serially-uncorrelated, i.e.

$$Cov(e_{it}, e_{is}) \neq 0, \forall t \neq s$$

λ is then called the quasi-demeaning parameter, and the regression equation $y_{it} - \lambda\bar{y}_i = \beta_1(x_{it} - \lambda\bar{x}_i) + (v_{it} - \lambda\bar{v}_i)$ is called the quasi-demeaned equation. It can be proved that

$$\lambda = 1 - \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}}, \qquad \lambda \in [0, 1]$$

λ is theoretically unknown but can be estimated by a given sample.

- If the random effect is relatively unimportant compared to the idiosyncratic error, $\lambda \to 0$, FGLS will be close to pooled OLS.
- If the random effect is relatively important compared to the idiosyncratic error, $\lambda \to 1$, FGLS will be close to fixed effects.

Luckily, random effects estimation can be used to estimate the effect of time-invariant variables. (Compared to fixed-effect model's limitations) But in economics, unobserved individual effects are seldomly uncorrelated with explanatory variables so that fixed effects is more convincing.

Remarks

Leave yourself with proof of $Cov(e_{it}, e_{is}) \neq 0, \forall t \neq s$, with

1. $E(e_{it}) = 0$;
2. $Var(e_{it}) = \sigma_u^2$;
3. $Cov(e_{it}, e_{is}) = 0, \forall t \neq s$.

Meanwhile, you can prove that β_1 is **BLUE**.

9.5 Correlated Random Effects Estimation

When using CRE to choose between FE and RE, we must include any time-constant variables α_i that appear in RE estimation.

$$y_{it} = \beta_1 x_{it} + \alpha_i + u_{it}$$

We try to combine FE and RE. This time, α_i is not only allowed to correlate with x_{it}, but also assumed to follow a distribution (or a certain way of correlation with x_{it}).

$$\alpha_i = \gamma_0 + \gamma_1\bar{x}_i + \gamma_i$$

where γ_i is assumed to be random and uncorrelated with each x_{it}. Then, $Cov(\gamma_i, \bar{x}_i) = 0$.

The CRE equation is then

$$y_{it} = \gamma_0 + \beta_1 x_{it} + \gamma_1 \bar{x}_i + v_{it}$$

where $v_{it} = u_{it} + \gamma_i, E(v_{it}) = 0, Cov(v_{it}, x_{it}) = 0$. Compared to FE or RE, the differences in equations lies in $\gamma_1 \bar{x}_i$,

Estimating the equation by RE (or even just pooled OLS) yields

$$\hat{\beta}_{CRE,j} = \hat{\beta}_{FE,j}, \forall j = 1, ..., k$$
$$\hat{\alpha}_{CRE,j} = \hat{\alpha}_{FE,j}, \forall t = 1, ..., T$$

which indicates time-varying estimates will be the same as in FE.

This intriguing discovery tells us that, adding time average $\bar{x}_i$ and using RE estimates is equivalent to first demeaning time average and using pooled OLS. This provides a new way to interpret FE: when estimating partial effect of x_{it} on y_{it}, FE controls for time average $\bar{x}_i$. Moreover, CRE "visualizes" the difference between FE and RE. CRE itself will yield the same result as FE. When setting $\gamma_1 = 0$, the CRE equation is then reduced to RE equation. Obviously, when $\gamma_1 \neq 0$, since $\bar{x}_i$ and x_{it} appear in the same equation, problems of multicollinearity will arise, which will cause $\hat{\beta}_{FE}$ to have a larger standard error and less precision; especially when variation of x_{it} is small across time. Hence, generally speaking, FE estimators are less precise than those in RE. Another advantage of CRE is that it allows for estimation of the effects of time-constant explanatory variables, which is not possible using FE.

At the meantime, CRE provides a simple but formal way to choose between FE and RE. RE will set $\gamma_1 = 0$, while FE will estimate γ_1. We can conduct a hypothesis test where the null hypothesis is in support of RE.

$$H_0 : \gamma_1 = 0$$

Under H_0, RE is sufficient. If H_0 is rejected, FE is preferred.

9.6 General Policy Analysis with Panel Data

The two-period, before-after setting is a special case of a more general policy analysis framework when $T \geq 2$.

$$y_{it} = a_1 + a_2 d2_t + ... + a_T dT_t + \beta w_{it} + x_{it}\varphi + \alpha_i + u_{it}$$

where w_{it} is the binary policy variable and β estimates the average treatment effect (ATE) of the policy.

To allow w_{it} to be systematically related to the unobserved fixed effect a_i, such as self-selection problem, we estimate the regression with either FD or FE, using cluster-robust standard errors.

We need to be careful if the policy variable w_{it} reacts to past shocks, which is the feedback from the error term to the policy variable. In this case, we can introduce an extra $\delta w_{i,t+1}$ into the model and test for feedback, i.e., testing against $H_0 : \delta = 0$.

Moreover, if time trends are unique across individuals ($T \geq 3$), we can safely introduce a new term $g_i t$ into the model, which is a unit-specific time trend. This allows the policy intervention to not only be correlated with level differences among units (captured by α_i), but also by trend differences. The model is still estimated by FE, and $g_i t$ will be differenced to be the intercept term.

10 IV Estimation & 2SLS

Instrumental variables estimation and two stage least squares are equivalent under some conditions. However, two stage least squares are more general compared to instrumental variables estimation.

The endogeneity problem is **endemic** in social sciences and economics.

- In many cases, important personal variables cannot be observed.
- These are often correlated with observed explanatory information.
- In addition, measurement error may also lead to endogeneity.
- Solutions to endogeneity problems considered so far:
  - Proxy variables method for omitted regressors
  - Fixed effects methods if
    - panel data is available,
    - endogeneity is time-constant and
    - regressors are not time-constant.

Instrumental Variables Method (IV) is now the most well-known and favored method to address endogeneity problems.

To begin with, first consider the simple linear regression model,

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

where the "weak" requirement of $Cov(u, x) = 0$ makes sure that slope coefficient is consistent. If the requirement is met, OLS is the best method for estimation (i.e., BLUE); no need to find alternatives.

However, if $Cov(u, x) \neq 0$, endogeneity problem arises. Explanatory variable x is then called endogenous variable. (Note that, you do not need to struggle with the origin of endogeneity. You can simply think of it as omitting some important variables.) Moreover in MLR, $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$, as long as $Cov(x_j, u) \neq 0$ for some j, then x_j is called endogenous variable.

10.1 Instrumental Variable

Definition:

1. It does not appear in the regression.
   - It did not participate in explaining variation of y.
   - In mathematics, if an IV also appears in the structural equation, you will not have enough equation to estimate slope coefficients via method of moments; if you use 2SLS, multicollinearity problem will be quite severe.

2. It is highly correlated with the endogenous variable. (Instrument Relevance)

- This requirement ensures the induced instrument is a good "representative".
  - Otherwise, such IV is a *weak* one.
- This condition can be checked by regression.

3. It is uncorrelated with the error term. (Instrument Exogeneity)
   - However, the argument mostly comes down to a "belief" and *cannot* be checked.
   - Since x is endogenous, $\hat{\beta}_0, \hat{\beta}_1$ are not correctly estimated, so are residuals $\hat{u}_i$; impractical and impossible to check.

10.2 IV Estimation in SLR

10.2.1 Exogenous Case

Start from the simple linear regression,

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

Identification of β_1 can be obtained from *zero conditional mean*.

$$\begin{aligned}
&Cov(x_i, u_i) = 0 \\
\Longrightarrow &Cov(x_i, y_i - \beta_0 - \beta_1 x_i) = 0 \\
\Longrightarrow &Cov(x_i, y_i) - \beta_1 Var(x_i) = 0 \\
\Longrightarrow &\beta_1 = \frac{Cov(x_i, y_i)}{Var(x_i)}
\end{aligned}$$

Replace "population parameter" with its sample-version:

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

Also, $\hat{\beta}_1$ is consistent, i.e., $\hat{\beta}_1 \xrightarrow{p} \beta_1$, as long as exogeneity holds.

10.2.2 Endogenous Case

However, if $Cov(x_i, u_i) \neq 0$, x_i is endogenous. Luckily, z_i is found with $Cov(z_i, u_i) = 0$.

$$\begin{aligned}
&Cov(z_i, u_i) = 0 \\
\Longrightarrow &Cov(z_i, y_i - \beta_0 - \beta_1 x_i) = 0 \\
\Longrightarrow &Cov(z_i, y_i) - \beta_1 Var(z_i) = 0 \\
\Longrightarrow &\beta_1 = \frac{Cov(z_i, y_i)}{Var(z_i)}
\end{aligned}$$

The instrument z_i has to satisfy two important properties:

- Instrument Exogeneity: $Cov(z_i, u_i) = 0$
- Instrument Relevance: $Cov(z_i, x_i) \neq 0$
  - However, when $Cov(z_i, x_i) \to 0$, z_i is a weak instrument, and this will have a pratical influence on slope coefficients.

With sample-version statistics replacing population ones:

$$\hat{\beta}_{IV} = \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Var}(z_i, x_i)} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

Note that even with endogeneity, $\bar{y} = \beta_0 + \beta_1 \bar{x}$ holds (as long as $E(u) = 0$). Intercept coefficient $\hat{\beta}_0$ can be obtained in this way.

If x_i is an exogenous variable, totally fine to set $z_i \equiv x_i$, and goes back to OLS estimation. Here we can get the sense that, any exogenous variable is the instrument for itself.

Note that **instrument must be biased, though consistent.** (Not covered here)

$$E(\hat{\beta}_{IV}) \neq \beta_1, \quad \hat{\beta}_{IV} \xrightarrow{p} \beta_1$$

10.2.3 Properties of IV with a POOR Instrumental Variable

IV may be much more inconsistent than OLS if the instrumental variable is not completely exogenous and only weakly related to x.

$$\begin{aligned}
\hat{\beta}_{1,IV} &= \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Var}(z_i, x_i)} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \\
&= \frac{\sum_{i=1}^{n}(z_i - \bar{z})(\beta_0 + \beta_1 x_i + u_i)}{\sum(z_i - \bar{z})(x_i - \bar{x})} \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(z_i - \bar{z})u_i}{\sum(z_i - \bar{z})(x_i - \bar{x})} \\
\Longrightarrow &\begin{cases} \text{plim } \hat{\beta}_{1,OLS} & = \beta_1 + Corr(x, u) \cdot \frac{\sigma_u}{\sigma_x} \\ \text{plim } \hat{\beta}_{1,IV} & = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)} \cdot \frac{\sigma_u}{\sigma_x} \end{cases}
\end{aligned}$$

IV is worse than OLS iff

$$\frac{Corr(z, u)}{Corr(z, x)} > Corr(x, u)$$

10.3 IV Estimation in MLR

Consider the *structural equation*

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + ... + \beta_k z_{k-1} + u_1$$

where y_2 is an endogenous variable, and all $z_j, \forall j = 1, 2, ..., k-1$ are exogenous variables. And note that such equation is called a structural equation because we are interested in β_j, but it does not mean the equation is a representation of causality.

We are to find an instrumental variable z_k satisfies that, z_j

1. Does not appear in regression equation above;
2. Is *UNcorrelated* with error term;
3. Is ***partially*** correlated with endogenous explanatory variable.

For the third requirement of partial correlation, consider the following equation

$$y_2 = \pi_0 + \pi_1 z_1 + ... + \pi_k z_{k-1} + \pi_k z_k + v_2$$

This is the so-called "*reduced form regression*". $z_1, ..., z_{k-1}$ are induced in case of any bias from omitting variables. The significance of π_k measures strength of correlation of y_2 and z_k. We are in the hope that null hypothesis $\pi_k = 0$ is rejected.

Traditionally, IV estimation can be realized by either method of moments, or Two Stage Least Squares.

10.3.1 Method of Moments

Consider the simple regression

$$y_{1i} = \beta_0 + \beta_1 y_{2i} + \beta_2 z_{1i} + u_i$$

where y_2 is endogenous, z_1 is exogenous, and we have found an IV z_2 for y_2. That is, with $E(u_i) = 0, Cov(z_{1i}, u_i) = 0$,and $Cov(z_{2i}, u_i) = 0$.

$$\begin{cases} E(u) = 0 \\ \quad \Longrightarrow \frac{1}{n}\sum_{i=1}^{n}(y_{1i} - \beta_0 - \beta_1 y_{2i} - \beta_2 z_{1i}) = 0 \\ Cov(z_{1i}, u_i) = 0 \Longrightarrow E(z_{1i}, u_i) = 0 \\ \quad \Longrightarrow \frac{1}{n}\sum_{i=1}^{n} z_{1i}(y_{1i} - \beta_0 - \beta_1 y_{2i} - \beta_2 z_{1i}) = 0 \\ Cov(z_{2i}, u_i) = 0 \Longrightarrow E(z_{2i}, u_i) = 0 \\ \quad \Longrightarrow \frac{1}{n}\sum_{i=1}^{n} z_{2i}(y_{1i} - \beta_0 - \beta_1 y_{2i} - \beta_2 z_{1i}) = 0 \end{cases}$$

We can find the solutions $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ for the three equations.

$$\begin{cases} \sum_{i=1}^{n}(y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 y_{2i} - \hat{\beta}_2 z_{1i}) = 0 \\ \sum_{i=1}^{n} z_{1i}(y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 y_{2i} - \hat{\beta}_2 z_{1i}) = 0 \\ \sum_{i=1}^{n} z_{2i}(y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 y_{2i} - \hat{\beta}_2 z_{1i}) = 0 \end{cases}$$

If with more endogenous variables, more IV are needed. For instance, if you have two endogenous variables, you have to find at least two IVs. However, if you are lucky enough to find more than 2 IVs, say 3, method of moments would fail, since number of solutions binds one-to-one with number of equations. Therefore, if the number of IVs equates that of endogenous variables, this is a just-identified case. If more, over-identified; if less, under-identified. If with more IV than number of endogenous variables, it is really a pity that some potentially valuable cannot be utilized in method of moments.

10.4 Two Stage Least Squares Estimation

10.4.1 Approach

Consider a general case of multiple linear regression,

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + ... + \beta_k x_{k-1} + u$$

where y_2 is endogenous, and all other explanatory variables are exogenous. We have found an IV z_k for y_2.

- First Stage (Reduced Form Regression)
  - The endogenous explanatory variable y_2 is predicted using only exogenous information:

    $$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + ... + \hat{\pi}_{k-1} z_{k-1} + \hat{\pi}_k z_k$$

    In essence, you partial out endogenous part in y_2, and only its exdogenous part is kept. (Also, you should prove it to yourself that $\hat{\pi}_k$ is not zero. If there are more than one IV for y_2, prove it that the joint hypothesis that all slope coefficients for IVs are zero can be rejected!)
- Second Stage
  - OLS with y_2 replaced by its prediction $\hat{y}_2$ from the first stage.

$$y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 z_1 + ... + \beta_k z_{k-1} + error$$

10.4.2 Principle of TSLS

10.4.2.1 From intuition

- All variables in the *second* stage regression are ***exogenous***, because y_2 was replaced by a prediction based on only exogenous information, and has been purged of its endogenous part (which is related to the error term).

- If there is one endogenous variable and one instrument, then 2SLS is equivalent to IV. Moreover, 2SLS estimation can also be used if there is more than one endogenous variables and at least as many instruments. (2SLS is more general than method of momenets.)

10.4.2.2 From Regression Back to the simplest case where the structural equation is $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u$. And we have found two exogenous variables z_2, z_3. z_2, z_3 did not appear in structural equation, and is each uncorrelated with u. Such assumption for z_2, z_3 is called exclusion restrictions. (equivalent to requirements for IV except instrument relevance)

Since exogenous in the structural equation is uncorrelated with u, and IV is also uncorrelated with u, any combination of them must also be uncorrelated with u. Therefore, any arbitrary linear combination of exogenous variables would construct a effective IV. To find the one that correlates with y_2 the most, we resort to the reduced form equation of y_2,

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v$$

where $E(v) = 0, Cov(z_1, v) = Cov(z_2, v) = Cov(z_3, v) = 0$.

The "best" IV for y_2 turns out to be the linear combination in the reduced form equation above. Denoted it as y_2^*,

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$$

In order to make sure that such IV y_2^* is not completely correlated with z_1, or say instrument relevance should hold, we test the joint hypothesis that $z_2 = z_3 = 0$. If the joint hypothesis cannot be rejected, then the structural equation cannot be identified.

The reduced form of y_2 successfully take apart y_2 into its endogenous part with correlation to u, and the exogenous part uncorrelated with u. The former is the reason why y_2 may be endogenous.

Luckily, although we do not know the true value of π_j, we can estimate the reduced form equation via OLS.

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3$$

Once we have got $\hat{y}_2$, we can use it as IV for y_2. Note that regardless of how many IV you originally induced for y_2, after estimating the reduced form equation and get $\hat{y}_2$ (in essence, find a best linear combination of exogenous variables), $\hat{y}_2$ is the single IV for y_2. Using method of moments,

$$\begin{cases} \sum_{i=1}^n (y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 y_{2i} - \hat{\beta}_2 z_{1i}) = 0 \\ \sum_{i=1}^n z_{1i}(y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 y_{2i} - \hat{\beta}_2 z_{1i}) = 0 \\ \sum_{i=1}^n \hat{y}_{2i}(y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 y_{2i} - \hat{\beta}_2 z_{1i}) = 0 \end{cases}$$

From such equations under method of moments, an intriguing discovery is that, this turns out to be equivalent to carry a Two Stage Least Squares. Especially in the second stage of regression, we can see ($y_2 = y_2^* + v$),

$$y_1 = \beta_0 + \beta_1 y_2^* + \beta_2 z_1 + (u + \beta_1 v)$$

where $E(u+\beta_1 v)=0$, and $u+\beta_1 v$ is uncorrelated with both y_2^* and z_1. That is why the regression of y_1 on $\hat{y}_2, z_1$ is effective.

Along with this logic, we will in the meantime feel at ease to grasp the principle for endogeneity test. (Covered later)

10.4.3 More Endogenous Variables

Consider the case with 2 endogenous variables:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + u$$

And suppose we have found two IVs, z_2, z_3.

- In the first stage
  - Regress y_2 on z_1, z_2, z_3, and get $\hat{y}_2$.
  - Regress y_3 on z_1, z_2, z_3, and get $\hat{y}_3$.
- In the second stage
  - Regress y_1 on $\hat{y}_2, \hat{y}_3, z_1$.

10.4.4 Remarks for IV

- There will be the problem of multicollinearity if with more than one endogenous variables, since $\hat{y}_2, \hat{y}_3$ are both linear combinations of z_1, z_2, z_3. Nevertheless, if not perfect multicollinearity, totally fine to conduct TSLS.
- The standard errors from the OLS second stage regression are wrong, since $\hat{y}_2$ is itself not y_2. If you are to get the "right" standard errors, you have to introduce "uncertainty" from the first to the second stage. Note that the **magnitude** of slope coefficients are *correct*.

10.5 Weak Instrument

$$\hat{\beta}_{1,IV} = \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Var}(z_i, x_i)} = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})}$$

If z is a weak IV, denominator $\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})$ is close to 0, the distribution of corresponding t statistic is heavily-tailed. If such problem is neglected, you would take more **risk** committing type I error before you even know it.

10.6 IV in Measurement Error

IV can not only help cope with issues of omitting variables, but also can address measurenment error. For example, consider the simple linear regression

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

where x^* is unobservable, $Cov(x^*, u) = 0$. x is a feasible measurement for x^* with measurement error $x = x^* + e$. Under CEV assumption, $Cov(x^*, e) = 0 \Longrightarrow Cov(x, e) = \sigma_e^2 \neq 0$. The equation can be equivalently rewritten as

$$y = \beta_0 + \beta_1 x + (u - \beta e)$$

Since $(u - \beta e)$ is then correlated with x, x is an endogenous variable due to measurement error.

What we need is an IV for x_1, which should highly correlate with x_1, and uncorrelate with measurement error. One possibility is to get a second measurement of x^*, say z; and z has another version of measurement error, $z = x^* + \eta$. We must assume that e is uncorrelated with η. By nature, z must have high correlation with x (Correlate through their common correlation with x^*), and z is uncorrelated with u. Just take z as an instrument for x.

10.7 Testing for Endogeneity of Explanatory Variables

In the general case, consider structural equation as

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + ... + \beta_k z_{k-1} + u_1$$

where y_2 is suspected to be endogenous, i.e., $Cov(y_2, u) \neq 0$. All other explanatory variables, $z_1, ..., z_{k-1}$, are exogenous.

One simple and intuitive way to think about endogeneity is to compare the estimated coefficients under OLS and 2SLS. If statistically salient difference is detected, then it is safe to allege that y_2 Is endogenous. However, using regression whenever possible is more convenient. Consider the reduced form regression,

$$y_2 = \pi_0 + \pi_1 z_1 + ... + \pi_k z_k + v_2$$

where z_k satisfies exclusion restrictions.

As before, $\hat{y}_2$ is the cleaned version without endogeneity, with residual v_2 of such reduced form equation absorbs all potential endogeneity. Then, y_2 is exogenous if and only if v_2 is uncorrelated with u_1. Rewrite u_1 as linear function of v_2

$$u_1 = \delta_1 v_2 + e_1$$

That is, $\delta_1 = 0$ in such equation.

Use our sample to estimate the equations above. Use OLS to estimate reduced form equation of y_2, and get $\hat{v}_2$. Then, conduct a regression for the equation

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + ... + \beta_k z_{k-1} + \delta_1 \hat{v}_2 + error$$

Use t statistics to test the null hypothesis $H_0 : \delta_1 = 0$. If we successfully reject H_0, we can deduce from the correlation of v_2 and v_1 that y_2 is endogenous.

To sum up, the general approach for endogeneity test for a set of variable(s) is

1. Regress each variable in the set on all exogenous (including those in the structural equation and induced IV) to get the estimated reduced form equation, and keep the residual;
2. Plug in all residuals(s) to the original structural equation and conduct a OLS regression;
3. Test against the joint hypothesis that all slope coefficient(s) of residual(s) be 0.

10.8 Over-Identifying Restrictions

When the instruments we have are more than our need, we will be able to test whether part of them are correlated with the structural error term. If this is the case, then the estimation using each single IV would be statistically different. If all IVs are exogenous, apart from sampling error, the error of 2SLS should be uncorrelated with all IVs. The general approach works like as follows.

1. Use 2SLS to estimate the structural equation, and get residual $\hat{u}_1$;
2. Regress $\hat{u}_1$ on all exogenous variables, and get R^2;
3. Under the null hypothesis that all IVs are uncorrelated with u_1, $nR^2 \overset{a}{\sim} \chi_1^q$, where q is the difference between the number of IVs and that of endogenous variables. If nR^2 exceeds the critical value, then H_0 can be rejected, and we can say that at least a set of all the IVs is not exogenous.