计量经济学

Econometrics

作者：张丹丹, LaTeXed by VectorPikachu

目录

第一章 Introduction

计量经济学 - Econometrics.

经济学分为两大阵营: Theoretical analysis v.s. Empirical analysis. 理论分析观察人的行为, 把行为高度的抽象, 建立模型, 寻求关系. 实证分析搜集数据, 定量的描述变量的关系, 验证理论的正确性.

计量经济学的应用:

1. forecast macroeconomic variables (长期的分析, 时间序列的数据, 不讲);
2. evaluate the impact of an intervention (一个政策干预的影响的评估, 因果关系);
3. study economic relationships.

1.1 Forecasting Macroeconomic Variables

在时间维度上的变化, 时间序列数据. 根据过去的数据来预测未来的.

1.2 Impact Evaluation

做了一个干预, 到底会有什么样的效果.

因果关系 - Causality.

$$X \to Y$$

这里面 X 是因 (Cause), Y 是果 (Effect).

Causality v.s. Correlation? 鸡叫天就亮?

The effect of medical treatment.

$$X \to Y$$

$$\text{go to hospital} \to \text{healthier}$$

去医院的人是不健康的, 不去医院的人是健康的, 那么是否可以说去医院导致人变得不健康? 问题在于去医院的人和不去医院的人是不可比的, 因为去医院的人本身就是不健康的, 这里 Y 影响了 X, 也就是健康状况影响了人是不是去医院.

我们要研究的应该是单向的关系, 也就是除去 Y 对 X 的影响的部分.

D_i 表示是否去医院. $E(Y_{1i}|D_i = 1)$ 表示去医院的人的健康状况. $E(Y_{0i}|D_i = 0)$ 表示不去医院的人的健康状况. 但是我们不应该比较这两个. 应该是同样的人, 去和不去两种状况之间的差. 应该比较的是该去医院也去了的人与该去医院但没去的人之间的关系.

经济发展程度 (X_2) 影响空气污染 (X_1), 经济发展程度影响预期寿命 (Y), 空间污染影响预期寿命. 那么探究 X_1 和 Y 之间的关系就会受到其他因素, 如经济发展程度的干扰.

吸烟量 (X) 和健康 (Y), 但是观察数据发现吸烟越多的人越健康. 这是因为 Y 会反过来影响 X, 因为健康状况不好的人不吸烟了.

为什么因果关系很重要? 要想改变世界, 比如让人更加健康, 那么我们就可以改变 X 来改变 Y. 我们可以借此做干预.

如何识别因果关系? 最理想的方法, **理想实验** (Ideal Experiments). 例如临床试验, X 为是否吃药, Y 为某种健康指标. 将病人随机分成两个组, $D = 1$ 为吃药的实验组 (Treatment), $D = 0$ 为不吃药的控制组 (Control). 那么影响为 $\Delta = E(Y_0) - E(Y_1)$. 这里面最重要的就是 **随机分组 (Random Assignment)**. 要求的是每一个个体被分到 Treatment 和 Control 的概率是完全一样的, 无差异的. 否则会存在**系统性的差异**, 两者是不可比的.

在统计学意义上无差异, 但是实际上是有差异的. 这两个均值的值可能是不一样的, 但是做某种检验发现没有显著的差异.

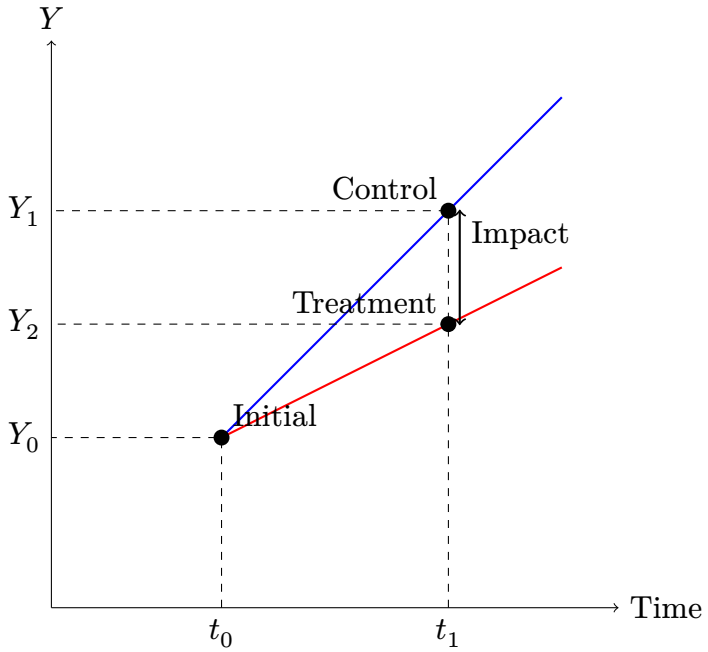

图 1.1: Ideal Experiment

在一张均值的图上, 起点是一样的, 这意味着两组是可比的. 在这个实验中所有其他的因素必须都控制成为一模一样的.

这种因果判断不能说明每一个个体的情况. 平均趋势不能反映到个案上. 只能判断整体的趋势.

理想实验的问题:

1. Random assignment is often considered as unethical.(伦理问题, 随机分成上大学的和不上大学的是不伦理的)
2. Nonrandom assignment for practical reasons.(有些政策已经实施了)
3. Placebo effects: the outcome may change if people know that they are being treated. (安慰剂效应)

大部分问题是非实验的. 例: The impact of minimum wage adjustment on migrants' employment.

Question: How does the intervention (raising MW) affect the outcome measure (employment)?

Identification problem: We do not know what would have happened if the program had not existed. 如果调整的城市不调整的情况是怎样的. 必须要知道一个和事实相反的状态. 这就好比我们必须知道应该去医院的人如果不去医院会怎么样.

Counterfactual Situation: 反事实状态.

Identification Assumption: 假设 Control 的变化程度是和实际的不参与的人是一样的, 也就是 non-treatment 平移得到 Control. 这样就得到了一个 Counterfact. 我们需要向前收集数据证明变化程度是一样的.

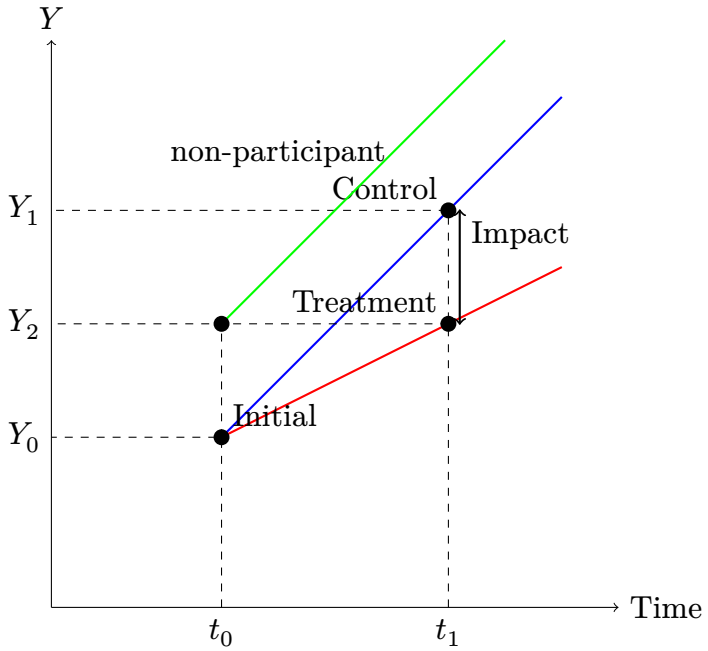

图 1.2: Counterfactual Situation

Causal (Ceteris Paribus) Effect: The ultimate goal of an evaluation is the isolation of the causal effect of an intervention.

1.3 Studying Economic Relationships

Becker 的犯罪模型:

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$$

y 是 hours spent in criminal activities, x_1 是 "wage" for an hour spent in criminal activity, x_2 是 hourly wage in legal employment, x_3 是 income other than from crime or employment,

x_4 是 probability of getting caught, x_5 是 probability of being convicted if caught, x_6 是 expected sentence if convicted, x_7 是 age.

现在我们来建立一个实证的模型, 我们把 f 一般选为线性回归, 一些无法被观测到的因素记为误差项 u.(unobserved factors, error term).

$$\text{crime} = \beta_0 + \sum_{i=1}^{6} \beta_i x_i + u$$

1.4 The Structure of Economic Data

1. Cross-sectional data, 截面数据;
2. Time series data, 时序数据;
3. Pooled cross-sectional data, 合并的截面数据, 里面把若干年份的样本合并在一起, 第二期需要重新抽样;
4. Panel data, 面板数据, 追踪数据, 第一年是一个截面数据, 第二期开始还找这五十个人不重新抽样.

截面数据, 一个时间点上对不同的个体的观察. 第一列通常是 id, 后面的每一列是这个人的特征.

第二章 Probability Theory & Estimation and Hypothesis Testing

probability density function(pdf), $f(x)$.

cumulative distribution function(cdf), $F(x)$.

joint probability density function.

marginal probability density function.

conditional probanility density function.

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}.$$

集中的趋势, Measures of central tendency: 期望和中位数.

变动的趋势, 方差.

相关的关系, 协方差.

一阶中心距: 期望.

二阶中心矩: 方差.

三阶中心距: skewness, 偏度.

四阶中心矩: 峰度, kurtosis.

独立, 线性独立 (Linear independet, $\mathrm{Cov}(X,Y) = 0$), 均值独立 (Mean independent, $E(Y|X) = E(Y)$).

最严格的是独立, 其次是均值独立, 最不严格的是线性独立.

计量经济学使用的是均值独立.

独立 $\Rightarrow$ 均值独立:

$$E(Y|X) = \int_{-\infty}^{+\infty} yf(y|x)\mathrm{d}y = \int_{-\infty}^{+\infty} y\frac{f(x,y)}{f(x)}\mathrm{d}y.$$

当 X, Y 独立的时候, 上式继续化简为: $E(Y|X) = \int_{-\infty}^{+\infty} yf(y)\mathrm{d}y = E(Y)$.

均值独立 $\Rightarrow$ 线性独立:

$$E(XY) = E_X[E(XY|X)] = E_X[XE(Y|X)].$$

如果均值独立, 则 $E(XY) = E_X[XE(Y)] = E(X)E(Y)$, 从而 X, Y 线性独立.

2.1 The Law of Iterated Expectations

定理 2.1 (The Law of Iterated Expectations)

$E(E(Y|X)) = E(Y)$.

♡

证明

$$\begin{aligned} E(E(Y|X)) &= \sum_x E(Y|X=x)P(X=x) \\ &= \sum_x \sum_y yP(Y=y|X=x)P(X=x) \\ &= \sum_y \sum_x yP(Y=y|X=x)P(X=x) \\ &= \sum_y yP(Y=y) = E(Y). \end{aligned}$$

2.2 Estimation and Hypothesis Testing

2.2.1 Statistical Inference and Estimation

我们从总体 (population) 中抽取一些样本 (sample), 利用样本做出一些对总体的关系的推断. 总体中的特征我们称为 parameters of interest.

对于一个随机的样本 $\{Y_1, \cdots, Y_n\}$, 我们对于一个位置的参数 θ, 我们可以取定一个 estimator(估计量) W, 它对于每一个可能的结果一个 θ 的取值.

$$W = h(Y_1, Y_2, \cdots, Y_n).$$

例如, 对均值的估计量可以是: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

2.2.2 Unbiasedness, Efficiency, Consistency

Unbiasedness: 如果 $E(W) = \theta$, 那么我们说它是无偏的估计量.

Efficiency: 方差小的好. 抽样分布就是估计量的分布, 因为估计量只和抽样有关. 还有就是把估计量的方差叫做抽样方差. (v.s. 样本方差.)

Consisitency: 对于任意的 $\varepsilon > 0$, $P(|W_n - \theta| > \varepsilon) \to 0$ as $n \to 0$.

2.2.3 Interval Estimation and Confidence Intervals

区间估计最终要估计出未知参数所在的区间，这个区间就是经常听到的置信区间.

定义 2.1 (置信区间)

设总体 X 的分布函数 $F(X;\theta)$ 中有未知参数 θ, 对于给定值 α, 若根据 X 的样本 $X_1, X_2, \ldots, X_n$ 确定的两个统计量 θ_s, θ_e, 满足: $P(\theta_s < \theta < \theta_e) \geqslant 1-\alpha$, 则称随机区间 (θ_s, θ_e) 是 θ 的置信水平为 $1-\alpha$ 的置信区间, $1-\alpha$ 称为置信水平, θ_s 和 θ_e 则分别称为置信下限和置信上限. ♣

如果我们设 population 符合正态分布 $N(\mu, \sigma^2)$, 同时 $Y_1, Y_2, \cdots, Y_n$ 是随机抽样的, 那么:

$$\bar{Z} = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

那么:

$$P(-d_{\alpha/2} < \bar{Z} < d_{\alpha/2}) = 1-\alpha.$$

最后化简得到: $[\bar{Y} - d_{\alpha/2}\sigma/\sqrt{n}, \bar{Y} + d_{\alpha/2}\sigma/\sqrt{n}]$. 这就是置信区间 (confidence intervals).

如果 σ 未知, 使用 t 分布.

$$\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t(n-1).$$

2.2.4 Hypothesis Testing

null hypothesis: $H_0 : \theta = \mu_0$.

alternative hypothesis: $H_1 : \theta \neq \mu_0$.

significance level: α.

p-value: $1-\Phi(t)$ 或者 $2(1-\Phi(t))$.

下面摘录部分来自于《概率论与数理统计教程》[1] 的假设检验的章节.

假设检验的基本步骤.

1. 建立假设.
   双侧假设: null hypothesis: $H_0 : \theta = \theta_0$, alternative hypothesis: $H_1 : \theta \neq \theta_0$.
   单侧假设: null hypothesis: $H_0 : \theta = \theta_0$, alternative hypothesis: $H_1 : \theta < (>)\theta_0$.
2. 选择检验统计量, 给出拒绝域形式.
   W 是一个拒绝域, 当样本属于 W 时, 拒绝 H_0, 否则接受 H_0. W 是一个样本的集合, 从而可以依据样本的一些统计量来确定.
3. 选择显著性水平.
   犯第一类错误的概率为: $\alpha(\theta) = P_\theta\{X \in W | H_0\text{为真}\}$. 也就是 H_0 为真, 我们却拒绝了 H_0 的概率.
   犯第二类错误的概率为: $\beta(\theta) = P_\theta\{X \in \bar{W} | H_1\text{为真}\}$. 也就是 H_1 为真, 我们却接受了 H_0 的概率.
   但是注意到我们不可能同时控制两类错误都小, 所以一般控制犯第一类错误的概率.

[1] 茆诗松, 程依明, 濮晓龙. 概率论与数理统计教程 [M]. 3. 高等教育出版社, 2019.

这就是 Fisher 的显著性检验.

显著性水平为 α 的检验: $\alpha(\theta) \leqslant \alpha$.

4. 给出拒绝域.

   确定显著性水平后, 我们就可以给出拒绝域了.

5. 做出判断.

现在可能出现另一个问题, 比如在一个较大的显著性水平 $\alpha = 0.05$ 的情况下, 我们得到拒绝原假设的结论; 而在另一个显著性水平 $\alpha = 0.01$ 的情况下, 我们却得到了接受原假设的结论.

现在就给出 p 值的定义: 在一个假设检验问题中, 利用样本观测值能够做出的拒绝原假设的最小的显著性水平称为 p 值. 也就是 p-value $= \min \alpha(\theta)$.

那么, $p \leqslant \alpha$, 则在显著性水平 α 下拒绝 H_0; $p > \alpha$, 则在显著性水平 α 下接受 H_0.

正态总体参数假设检验

我们只摘录单个正态总体的均值和方差的假设检验.

单个正态总体均值的检验:

1. $\sigma = \sigma_0$ 已知. 进行 u 检验.

$$u = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0,1). \tag{2.1}$$

I. 对于检验 $H_0 : \mu \leqslant \mu_0, H_1 : \mu > \mu_0$, 拒绝域形式为: $W = \{(x_1, x_2, \cdots, x_n) | u \geqslant c\}$. 这里我们得到: $c = u_{1-\alpha}$.

如果使用 p 值的话, $p = P(u \geqslant u_0) = 1 - \Phi(u_0)$. 这里 u_0 是根据观测值计算出来的.

II. 对于检验 $H_0 : \mu \geqslant \mu_0, H_1 : \mu < \mu_0$, 拒绝域形式为: $W = \{(x_1, x_2, \cdots, x_n) | u \leqslant c\}$. 此时 $W_{\mathrm{II}} = \{u \leqslant u_\alpha\}$. $p = P(u \leqslant u_0) = \Phi(u_0)$. 这里 u_0 是根据观测值计算出来的.

III. 对于检验 $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$, 拒绝域形式为: $W = \{(x_1, x_2, \cdots, x_n) | |u| \geqslant c\}$. 此时 $W_{\mathrm{III}} = \{|u| \geqslant u_{1-\alpha/2}\}$. $p = P(|u| \geqslant |u_0|) = 2(1 - \Phi(u_0))$. 这里 u_0 是根据观测值计算出来的.

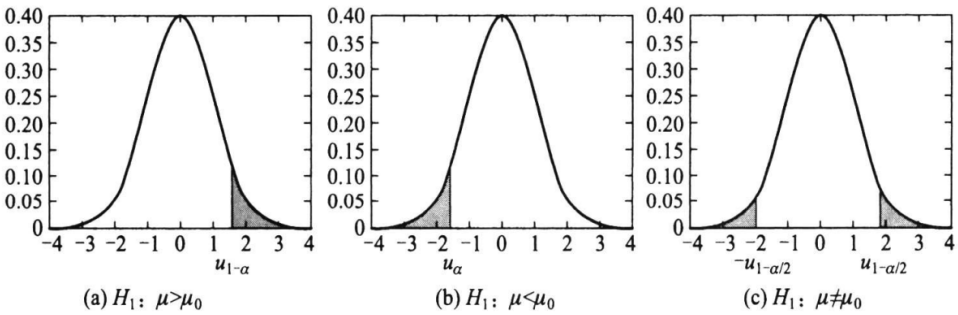

图 2.1: u 检验的拒绝域

2. σ 未知时的 t 检验.

样本的标准差: s. t 检验的统计量:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(n-1). \tag{2.2}$$

I. $W_{\mathrm{I}}=\{t\geqslant t_{1-\alpha}(n-1)\}$. $p_{\mathrm{I}}=P(t\geqslant t_0)=1-\Phi(t_0)$.

II. $W_{\mathrm{II}}=\{t\leqslant t_{\alpha}(n-1)\}$. $p_{\mathrm{II}}=P(t\leqslant t_0)=\Phi(t_0)$.

III. $W_{\mathrm{III}}=\{|t|\geqslant t_{1-\alpha/2}(n-1)\}$. $p_{\mathrm{III}}=P(|t|\geqslant |t_0|)=2(1-\Phi(|t_0|))$.

下面给出一个例题来说明假设检验的问题:

设随机变量 X 的分布密度 $f(x)$ 可取以下两个密度函数:

$$f_0(x)=2x, 0<x<1; f_1(x)=2(1-x), 0<x<1.$$

基于 X 的一个观测值, 对于假设检验问题

$$H_0: f(x)=f_0(x) \leftrightarrow H_1: f(x)=f_1(x),$$

利用上面介绍的原则求检验水平为 0.01 的检验法, 并求此时第二类错误的概率.

提示: 注意到 $f_0(x)$ 和 $f_1(x)$ 在 $(0,1)$ 内分别为增/减函数.

解答:

检验规则: 当 $x\leqslant C$ 时, 拒绝 H_0(给出了拒绝域的形式). 因为 $f_0(x)$ 表明观测到的 x 越小, 在原假设成立的条件下出现的概率越小, 在一件小概率事件出现的时候, 我们就可以怀疑我们的原假设了.

第 I 类错误:

$$\alpha(C)=P\{\text{拒绝}H_0|H_0\text{为真}\}=P\{x\leqslant C|f(x)=f_0(x)\}=\int_0^C 2x\mathrm{d}x=C^2\leqslant\alpha.$$

那么我们取 $C=\sqrt{\alpha}=0.1$ 即可, 此时第 II 类错误:

$$\beta=P\{\text{接受}H_0|H_1\text{为真}\}=P\{x>0.1|f(x)=f_1(x)\}=\int_{0.1}^1 2(1-x)\mathrm{d}x=0.81.$$

第三章 The Simple Regression Model

3.1 The Basic Concept

定义 3.1 (简单线性回归模型)

我们选定 x (Independent Variable, 解释变量) 和 y (Dependent Variable, 被解释变量) 代表两个经济学里的总体, 并且我们关心 $x \to y$ 的关系. 我们用误差项 (error term)u 来代表其他因素的影响, 我们用下面的式子来表示 y 和 x 的关系:

$$y = \beta_0 + \beta_1 x + u \tag{3.1}$$

这里我们又可以称之为双变量线性回归模型. ♣

等式3.1中回归系数 (Regeression Coefficients) 是 β_0, β_1. β_0 是 intercept parameter, β_1 是 slope parameter, 同时也是 parameter of interest. y 和 x 之间的关系不是线性的. 这里的线性指的是参数之间的线性的关系. 变量有三个.

例题 3.1 给出下面的式子:

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + u$$

u 中可以有 experience, gender, industry, family background, location, ability, occupation, tenure ...

一个最简单的假设 (not restricted assumption, 只要有 β_0, 我们也可以做出这个假设):

$$E(u) = 0 \tag{3.2}$$

但是我们想要做出因果判断, 必须在 u 和 x 之间做出一个限制性的假设. 其中一个假设就是 Zero Conditional Mean Assumption:

$$E(u|x) = E(u) = 0 \tag{3.3}$$

这就意味着在 x 的条件下, u 的平均水平不会改变, 这是均值独立. 这个假设可以论证因果关系.

如果 u 是 ability (这是一个无法衡量的量, 被称为 unobserved variable), 假设3.3就意味着教育和能力没有任何的线性和非线性的关系, 这是不对的, 实际上是有关系的. 我们应当讨论3.3是否成立, 成立的时候得到因果关系, 否则还要继续讨论.

在给定假设3.3的前提下, 我们有:

$$E(y|x) = \beta_0 + \beta_1 x \tag{3.4}$$

3.2 Ordinary Least Squares

如果有 n 个样本 (observation), 来估计 β_0, β_1.

3.2.1 Method of Moments(MoM)

在给定假设3.3的情况下:
我们有:

$$\mathrm{Cov}(x,u) = E(xu) - E(x)E(u) = 0 \tag{3.5}$$

现在有 $u = y - \beta_0 - \beta_1 x$, 而 $E(u) = 0, E(xu) = 0$, 那么:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3.6}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3.7}$$

根据3.6可以得到:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3.8}$$

带入3.7得到:

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}) \tag{3.9}$$

同时 $\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, $\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2$.
那么:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\widehat{\mathrm{Cov}(x,y)}}{\widehat{\mathrm{Var}(x)}} \tag{3.10}$$

这里 x 应该是非常数的.

3.2.2 Ordinary Least Squares(OLS)

我们考虑残差 $\hat{u}_i$:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \tag{3.11}$$

$\hat{y}_i$ 是 fitted value, predicted value.
我们的优化目标:

$$\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \hat{u}_i^2 = \min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \tag{3.12}$$

我们设上面的式子为 S. 则:

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3.13}$$

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3.14}$$

现在根据3.13我们可以得:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3.15}$$

继续带入3.14, 我们得到:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2} \tag{3.16}$$

会很大程度上被 outlier 影响. 总体回归线是确定的, 但是估计的线可以有无数条. 没有 hat 的是误差项, 有 hat 的是残差. $y_i = \hat{y}_i + \hat{u}_i$, $u_i = y_i - E(y|x)$.

3.3 Algebraic Properties of OLS Statistics

1. OLS 残差的样本均值为 0:

$$\sum_{i=1}^{n}\hat{u}_i = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3.17}$$

但是 $\sum_{i=1}^{n} u_i \neq 0$.

2. x_i 和 $\hat{u}_i$ 的协方差为 0:

$$\sum_{i=1}^{n}x_i\hat{u}_i = \sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3.18}$$

3. $(\bar{x}, \bar{y})$ 总是在 regression line 上的.

一些推论:

1. $\bar{\hat{y}} = \bar{y}$.
2. $\hat{y}$ 和 $\hat{u}$ 之间的 Cov $= 0$, 这是因为 x 和 $\hat{u}$ 没有线性关系, 而 x 和 $\hat{y}$ 是线性关系.

Total sum of squares(SST):

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.19}$$

Explained sum of squares(SSE) = 概率统计中的 Square Sum of Regeression(SSR):

$$SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{3.20}$$

Residual sum of squares(SSR) = 概率统计中的 Square Sum of Errors(SSE):

$$SSR = \sum_{i=1}^{n}\hat{u}_i^2 \tag{3.21}$$

那么我们有 $SST = SSR + SSE$.

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

拟合优度: 用来判断 x 多好的可以解释 y

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \tag{3.22}$$

3.4 Units of Measurement and Functional Form

改变单位不会改变对结果的解读, R^2 不改变. 但是会改变 estimator.

在线性回归里引入非线性:

$$f(y) = \beta_0 + \beta_1 g(x) + u$$

我们可以计算弹性 (Elasticity).

1. Model without logarithm: x 增加一个单位, y 增加 β_1 个单位.
2. Model with logarithm of y (semi-log model): x 增加一个单位, y 增加 $100\beta_1$ 个百分点. 推导过程:
$$\log(y + \Delta y) - \log y = \beta_1 \Delta x \Rightarrow \frac{\Delta y}{y} \approx \log\left(1 + \frac{\Delta y}{y}\right) = \beta_1 \Delta x.$$
3. Model with logarithm of x (semi-log model): x 增加一个百分点, y 增加 $\beta_1/100$ 个单位. 推导过程:
$$\Delta y = \beta_1(\log(x + \Delta x) - \log x) = \beta_1 \log\left(1 + \frac{\Delta x}{x}\right) \approx \frac{\Delta x}{x}.$$
4. Model with logarithm of x and y (log-log model): x 增加一个百分点, y 增加 β_1 个百分点. 推导过程:
$$\begin{aligned}\frac{\Delta y}{y} &\approx \log\left(1 + \frac{\Delta y}{y}\right) = \log(y + \Delta y) - \log y \\ &= \beta_1(\log(x + \Delta x) - \log x) = \beta_1 \log\left(1 + \frac{\Delta x}{x}\right) \approx \frac{\Delta x}{x}.\end{aligned}$$

或者是加入平方项. 也就是 $y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$. 它的拐点是导数等于 0 的地方. 我们可以得到边际影响 (Marginal effect):

$$\Delta y = (\beta_1 + 2\beta_2 x)\Delta x \Rightarrow \frac{\Delta y}{\Delta x} = \beta_1 + 2\beta_2 x.$$

3.5 Unbiasedness of OLS

The Gauss-Markov Assumptions for Simple Regression.

1. Linear in Parameters.
$$y = \beta_0 + \beta_1 x + u \tag{3.23}$$
2. Random Sampling. 随机抽样. 选 n 个样本.
$$y_i = \beta_0 + \beta_1 x_i + u_i, i = 1, 2, \cdots, n \tag{3.24}$$
3. Sample Variation in the Explanantory Variable. 也就是 x 不能都相同.
4. Zero Conditional Mean.
$$E(u|x) = 0 \tag{3.25}$$
5. Homoskedasticity.(同方差性)
$$\mathrm{Var}(u|x) = \sigma^2 \tag{3.26}$$

我们现在用 Assumption 1-4 来推出 OLS 的无偏性 (unbiasedness):

$$E(\hat{\beta}_0) = \beta_0 \tag{3.27}$$

$$E(\hat{\beta}_1) = \beta_1 \tag{3.28}$$

首先我们有:

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

注意到: $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})y_i = \sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)$.
那么我们有:

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 \bar{x} + u_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\ &= \beta_1 + E\left(\frac{\sum_{i=1}^n (x_i - \bar{x})u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\ &= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} E(u_i) = \beta_1. \end{aligned}$$

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{y} - \hat{\beta}_1 \bar{x}) = E\left(\beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}\right) \\ &= \beta_0 + E(\bar{u}) + E(\beta_1 - \hat{\beta}_1)\bar{x} = \beta_0. \end{aligned}$$

3.6 Variances of the OLS Estimators

$\mathrm{Var}(u|x) = \sigma^2 \Rightarrow \mathrm{Var}(y|x) = \sigma^2$.

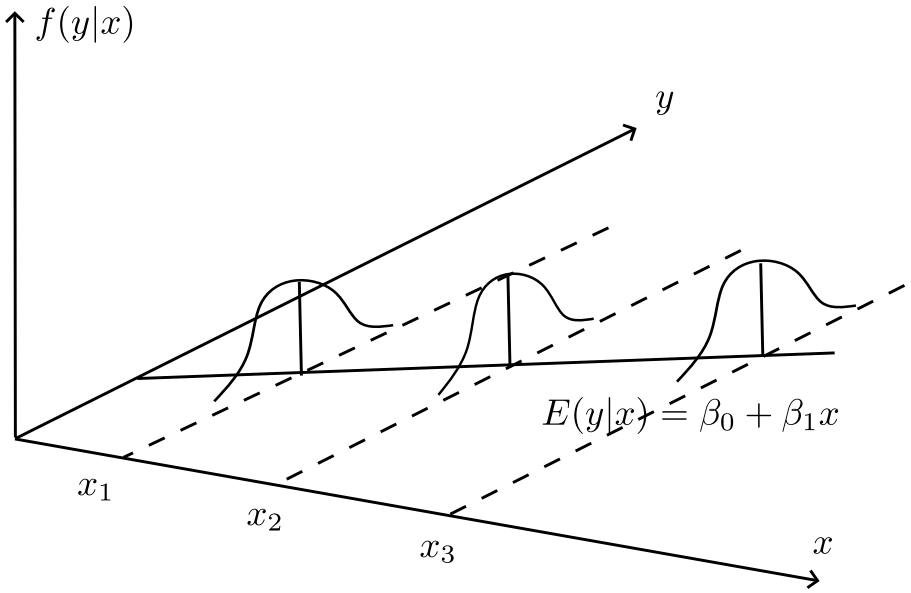

图 3.1: The Simple Regression Model under Homoskedaticity

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\mathrm{Var}(u_i)}{(\sum_{i=1}^{n}(x_i-\bar{x})^2)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{3.29}$$

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 n^{-1}\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{3.30}$$

我们来推导一下式3.30:

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_0) &= \mathrm{Var}(\bar{y}-\hat{\beta}_1\bar{x}) \\
&= \mathrm{Var}\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})(\bar{y}(x_i-\bar{x})-y_i\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \\
&= \frac{\sum_{i=1}^{n}\mathrm{Var}(\bar{y}(x_i-\bar{x})-y_i\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}\mathrm{Var}(\frac{1}{n}\sum_{j\neq i}(x_i-\bar{x})y_j+(\frac{x_i-\bar{x}}{n}-\bar{x})y_i)}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
&= \frac{\sigma^2\sum_{i=1}^{n}(\frac{n-1}{n^2}(x_i-\bar{x})^2+\frac{1}{n^2}(x_i-\bar{x})^2+n\bar{x}^2)}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
&= \frac{\sigma^2 n^{-1}\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}.
\end{aligned}$$

Standard Deviation(标准差):

$$sd(\hat{\beta}_1) = \sqrt{\mathrm{Var}(\hat{\beta}_1)}$$

$$sd(\hat{\beta}_0) = \sqrt{\mathrm{Var}(\hat{\beta}_0)}.$$

我们来估计 σ^2, 但是之前有两个等式3.17和3.18限制了, 我们损失了两个自由度. 那么应该估计为:

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

Standard Error(标准误):

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_x}}.$$

同样的, 我们把 $sd(\beta_0)$ 的 σ 替换成为 $\hat{\sigma}$ 可以得到 $se(\hat{\beta}_0)$.

第四章 The Multivariate Linear Regression Model

4.1 Motivation of Multiple Regression

零均值假设非常的不现实. 很难获取因果关系.

1. Allows us to explicitly control for many other factors that simultaneously affect the dependent variable;
2. Can accommodate many explanatory variables that may be correlated;
3. We can hope to **infer causality** in cases where simple regression analysis would be misleading (可以放松条件期望零值假设, $E(u|x_1, x_2, \cdots, x_k) = 0$);
4. If we add more regressors to our model, then more of the variation in y can be explained;
5. Can **have a better prediction** of the dependent variable (R^2 很大, 解释的力度很强);
6. Can incorporate **more general functional form relationships**.

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{expr} + u$$

给定经验不变 (holding experience fixed), 判断教育对工资的影响.

consumption vs. income:

$$\text{cons} = \beta_0 + \beta_1 \text{inc} + \beta_2 \text{inc}^2 + u$$

现在我们不能说 holding inc^2 fixed, 判断 inc 对 cons 的边际影响. 反而我们要写成:

$$\frac{\Delta \text{cons}}{\Delta \text{inc}} = \beta_1 + 2\beta_2 \text{inc}$$

对于 k 个解释变量:

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

现在的条件期望零值假设:

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

4.2 Mechanics and Interpretation of OLS

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \tag{4.1}$$

我们现在的残差为:

$$\sum_{i=1}^{n} \hat{u_i}^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik})^2 \tag{4.2}$$

我们写成矩阵的形式:

$$Y = X\beta + U \tag{4.3}$$

那么:

$$\hat{U}^\top \hat{U} = (Y - X\hat{\beta})^\top (Y - X\hat{\beta}) = Y^\top Y - Y^\top X\hat{\beta} - \hat{\beta}^\top X^\top Y - \hat{\beta}^\top X^\top X\hat{\beta} \tag{4.4}$$

那么:

$$\frac{\partial \hat{U}^\top \hat{U}}{\partial \hat{\beta}} = -2X^\top Y + 2(X^\top X)\hat{\beta} \tag{4.5}$$

我们有下面的几个性质:

$$\sum_{i=1}^{n} \hat{u}_i = 0 \tag{4.6}$$

$$\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0, j = 1, 2, \cdots, k \tag{4.7}$$

$$\mathrm{Cov}(x_{ij}, \hat{u}_i) = 0 \tag{4.8}$$

$$(\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_k, \bar{Y}) \text{ lies on the regression line} \tag{4.9}$$

4.2.1 Interpretation

Partial effects:

$$\Delta\hat{y} = \hat{\beta}_1\Delta x_1 + \hat{\beta}_2\Delta x_2 + \cdots + \hat{\beta}_k\Delta x_k \tag{4.10}$$

Partialling out interpretation (排除其它影响):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2} \tag{4.11}$$

这里面的 $\hat{r}_{i1}$ 是:

$$x_{i1} = \hat{\gamma}_1 + \hat{\gamma}_2 x_{i2} + \cdots + \hat{\gamma}_k x_{ik} + \hat{r}_{i1} = \hat{x}_{i1} + \hat{r}_{i1} \tag{4.12}$$

这样其他因素的影响都被包含在了 $\hat{x}_{i1}$ 之中.

$$\frac{\partial \sum_{i=1}^{n} \hat{u}_i^2}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} (\hat{x}_{i1} + \hat{r}_{i1})(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

现在由于 $\sum_{i=1}^{n} \hat{x}_{i1}\hat{u}_i = 0$, 那么:

$$\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

现在继续可以发现 $\sum_{j=2}^{k} x_{ij}\hat{r}_{i1} = 0$, 那么:

$$\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_1 x_{i1}) = 0 \tag{4.13}$$

4.2.2 Comparison of Bivariate and Multivariate Regression Estimates

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 \tag{4.14}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \tag{4.15}$$

那么有:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1 \tag{4.16}$$

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

$\tilde{\beta}_1 = \hat{\beta}_1$:

1. $\hat{\beta}_2 = 0$, 也就是 x_2 不影响 y, 应该放在 u 里;
2. $\tilde{\delta}_1 = 0$, 也就是 x_1 和 x_2 不相关, 那么也没有必要把 x_2 从 u 里拿出来, 当然拿出来也可以. 拿出来可以让 R^2 变大, 对 y 的预测做得更好.

Let $\hat{\beta}_j, j = 0, 1, \cdots, k$ be the OLS estimators from the regression using full set of explanatory variables.

Let $\tilde{\beta}_j, j = 0, 1, \cdots, k-1$ be the OLS estimators from the regression that leaves out x_k.

Let $\tilde{\delta}_j$ be the slope coefficient on x_j in the regression of x_k on $x_1, \cdots, x_{k-1}$.

也就是: $x_k = \tilde{\delta}_0 + \tilde{\delta}_1 x_1 + \cdots + \tilde{\delta}_{k-1} x_{k-1}$.

那么:

$$\tilde{\beta}_j = \hat{\beta}_j + \hat{\beta}_k \tilde{\delta}_j. \tag{4.17}$$

这里 $\hat{\beta}_j$ 是 x_j **直接**对 y 的影响, 而 $\hat{\beta}_k \tilde{\delta}_j$ 是 x_j 通过 x_k 对 y 的**间接**影响.

$$R^2 = \frac{SSE}{SST} = \left(\frac{\mathrm{Cov}(y_i, \hat{y}_i)}{\sqrt{\mathrm{Var}(y_i)}\sqrt{\mathrm{Var}(\hat{y}_i)}} \right)^2. \tag{4.18}$$

这个式子的证明, 我们可以让分子分母同时乘以 $\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2$, 再考虑到 $\bar{\hat{y}} = \bar{y}$, 就可以得证. 这里的 y_i 指的是样本. 还要注意到: $\sum_{i=1}^n \hat{u}_i (\hat{y}_i - \bar{\hat{y}}) = 0$, 因为 $\hat{y}_i, \bar{\hat{y}}$ 中含有的都是 x_i 的线性组合, 而 x_i 和 $\hat{u}_i$ 无关.

4.3 The Expected Value of the OLS Estimators

4.3.1 Assumptions for Multiple Linear Regeression

我们有下面的假设.

1. Linear in Parameters. 参数是线性的. 没有要求 x 和 y 之间是线性的.

2. Random Sampling. 随机抽样.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i. \tag{4.19}$$

3. No Perfect Collinearity. 不存在完全的共线性. 线性相关的: 存在不全为 0 的 $a_1, \cdots, a_k$, s.t.

$$a_0 + a_1 x_{i1} + \cdots + a_k x_{ik} = 0, \forall i = 1, 2, \cdots, n. \tag{4.20}$$

也就是: $[\vec{e}, \vec{x}_1, \cdots, \vec{x}_k]\vec{a} = \vec{0}$ 有非零解 $\vec{a}$.
如果存在线性相关, 去掉一个 x. 或者保留 x 而去掉截距.
例如对于下面的 regressiong model:

$$\ln(wage_i) = \beta_1 educ_i + \beta_2 male_i + \beta_3 female_i + u_i$$

是可以的, 这是因为 $a_1 educ + (a_2 - a_3) male + a_3 = 0$ 只有零解, 从而 $a_1 = a_3 = 0 \Rightarrow a_2 = 0$, 但是如果有常数项的话, 就变成了 $a_1 educ + (a_2 - a_3) male + a_0 + a_3 = 0$ 只有零解, 这样就会导致 $a_0 + a_3 = 0$, 这个式子是可以有非零解的.

4. Zero Conditional Mean. $E(u|x_1, x_2, \cdots, x_k) = 0$.

导致4.失败的可能性:

1. The functional form is misspecified. 方程形式误设. log, 二次项, 交叉项. 这样的话一些非线性的关系会进入 Error term 中去, 从而导致了 u 和 x 有关系.
2. An important variable that is correlated with any model regressor is omitted. 重要变量遗失. Omitted Variable Problem. 如果没有办法去 Measure 一个重要的变量, 要用一些高级的方法来处理.
3. The explanatory variable is measured with error. 测量误差的问题. 强调的是 x 的测量误差.
4. One or more regressors are determined jointly with y. 反向因果. y 对 x 有影响.

只要 x 和 u 发生了关系, 我们就是说 x 是**内生的 (Endogenous)**. 但是 x 和模型里的 u 不相关, 就说它是**外生的 (Exogenous)**, 实际上也就是 MLR.4.成立. When Assumption MLR.4 holds, we call the explanatory variables exogenous. If x_j is correlated with u, then x_j is said to be an endogenous explanatory variable.

4.3.2 Unbiasedness of the OLS Estimators

定理 4.1 (Unbiasedness of OLS)

在上面的四条假设下, OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ are unbiased estimators of the population parameters $\beta_0, \beta_1, \cdots, \beta_k$:

$$E(\hat{\beta}_j) = \beta_j, j = 0, 1, \cdots, k. \tag{4.21}$$

♡

证明 首先, 有:

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

同时: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$.

由于残差和解释变量无关, 那么:

$$\hat{\beta}_j = \frac{\sum_{i=1}^n \hat{r}_{ij}(\beta_j x_{ij} + u_i)}{\sum_{i=1}^n \hat{r}_{ij}^2}$$

同时 $x_{ij} = \hat{x}_{ij} + \hat{r}_{ij}$, 那么:

$$\hat{\beta}_j = \beta_j + \frac{\sum_{i=1}^n \hat{r}_{ij} u_i}{\sum_{i=1}^n \hat{r}_{ij}^2}$$

那么:

$$E(\hat{\beta}_j) = \beta_j + E_x \left[E\left(\frac{\sum_{i=1}^n \hat{r}_{ij} u_i}{\sum_{i=1}^n \hat{r}_{ij}^2} \middle| x_1, \cdots, x_k \right) \right]$$

那么:

$$E(\hat{\beta}_j) = \beta_j + E_x \left[\sum_{i=1}^n \frac{\hat{r}_{ij}}{\sum_{i=1}^n \hat{r}_{ij}^2} E\left(u_i | x_1, \cdots, x_k\right) \right] = \beta_j.$$

4.3.2.1 Overspecified Model

A model is *overspecified* when one or more of the independent variables is included in the model even though it has no partial effect on y in the population. 也就是对应的 $\beta_j = 0$.

我们设 true model 的 estimators 为 tilde, 同时只用 $k-1$ 个变量.

我们的现在的 model 的 estimators 为 hat, 用了 k 个变量.

x_k 对 $x_1, \cdots, x_{k-1}$ 的辅助回归为 $\tilde{\delta}_j$. 注意到 $E(\hat{\beta}_k) = \beta_k = 0$.

那么: $\beta_j = E(\tilde{\beta}_j) = E(\hat{\beta}_j) + E(\hat{\beta}_k)\tilde{\delta}_j = E(\hat{\beta}_j)$.

不影响估计量的无偏性, 但如果无关变量与解释变量存在相关性, 会丧失估计量的有效性.

4.3.2.2 Underspecified Model, Omitted Variable Bias

If a variable that actually belongs in the true model is omitted, we say the model is *underspecified*.

我们设现在的 model 的 estimators 为 tilde, 同时只用 $k-1$ 个变量.

我们设 true model 的 estimators 为 hat, 用了 k 个变量.

x_k 对 $x_1, \cdots, x_{k-1}$ 的辅助回归为 $\tilde{\delta}_j$.

$E(\tilde{\beta}_j) = E(\hat{\beta}_j) + E(\hat{\beta}_k)\tilde{\delta}_j = \beta_j + \beta_k \tilde{\delta}_j$.

所以现在是有偏的. Omitted Variable Bias.

$\tilde{\delta}_j$ 的符号就是: $\text{Corr}(x_j, x_k)$ 的符号.

即使 x_k 只和其中的一个 x_j 有关, 也会导致另一个 x_t 的 Estimator 有偏. 这是因为 x_j 和 x_t 有关的. 我们可以通过 partialling out 的等式来说明. 但是会比较接近于 0.

If an omitted variable has partial effects on y and it is correlated with at least one of the regressors, then the OLS estimators of all coefficients will be biased.

注 在考虑回归模型中是否应包含某个解释变量时, 要权衡它可能导致的偏误以及有效性丧失. 如果怀疑 x_2 与 x_1 和 y 同时相关, 研究中更倾向于包含 x_2, 即使 x_2 的加入会导致 x_1 的方差变大 (有效性丧失、显著性减弱). 因为, 有偏的估计是更严重的问题, 当样本足够大时, 多重共线性变的不是那么重要.

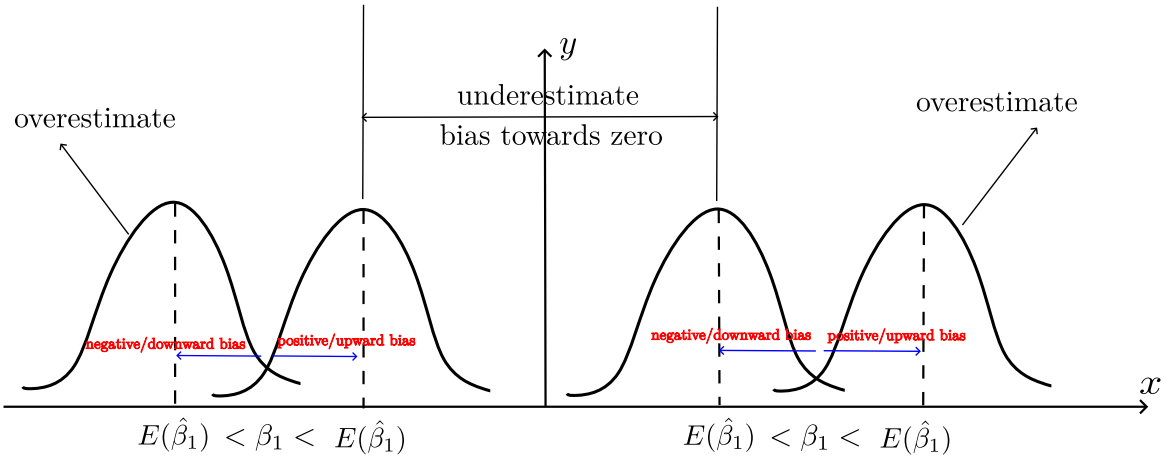

图 4.1: The Direction of the Bias

4.4 The Variance of the OLS Estimators

现在我们需要一个新的假设: Homoscedasticity

Assumption MLR.5 Homoscedasticity: The error u has the same variance given any values of the explanatory variables.

$$\text{Var}(u|x_1, \cdots, x_k) = \sigma^2. \tag{4.22}$$

同时有:

$$\text{Cov}(u_i, u_j) = 0, i \neq j. \tag{4.23}$$

Gauss-Markov Assumption: MLR.1 - MLR.5.

那么: $\text{Var}(y|\mathbf{x}) = \sigma^2$.

定理 4.2 (Sampling Variances of the OLS Slope Estimators)

Under Assumptions MLR.1-MLR.5, conditional on the sample values of the independent variables, the sampling variance of the OLS estimators is given by

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, j = 1, 2, \cdots, k. \tag{4.24}$$

这里的 $SST_j = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$ 表示 x_j 的总体变动. R_j^2 是辅助回归 reg x_j on all other independent variables 的 R-squared. ♡

注 因为多元线性回归是用 $\hat{r}_{ij}$ 来解释 y. 所以要用 R_j^2 来 partialling out.

证明 我们有:

$$\hat{\beta}_j = \beta_j + \frac{\sum_{i=1}^n \hat{r}_{ij} u_i}{\sum_{i=1}^n \hat{r}_{ij}^2}$$

那么:

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_j) &= \mathrm{Var}_{\mathbf{x}} \left[\sum_{i=1}^n \frac{\hat{r}_{ij}^2}{(\sum_{i=1}^n \hat{r}_{ij}^2)^2} \mathrm{Var}\left(u_i | \mathbf{x}\right) \right] \\ &= \mathrm{Var}_{\mathbf{x}} \left(\frac{\sigma^2}{\sum_{i=1}^n \hat{r}_{ij}^2} \right) = \frac{\sigma^2}{SSR_j} = \frac{\sigma^2}{SST_j(1-R_j^2)}. \end{aligned}$$

这里的 R_j^2 就代表了 Linear relationships among the independent variables, 如果 R_j^2 越大, 线性关系越好.

A high correlation between regressors is called *multicollinearity*.

$$\mathrm{Var}(\hat{\beta}_j) \to \infty, R_j^2 \to 1.$$

It could be reduced by appropriately dropping certain variables, or **collecting more data**. 所以样本的数量超极重要的.

为了无偏我们可以牺牲一些方差.

4.5 Estimating the Error Variance

我们来估计 $\sigma^2 = E(u^2)$.

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^n \hat{u}_i^2 = \frac{1}{n-k-1} SSR. \tag{4.25}$$

定理 4.3

在五条假设的前提下:

$$E(\hat{\sigma}^2) = \sigma^2. \tag{4.26}$$ ♡

我们有: Standard deviation of $\hat{\beta}_j$:

$$sd(\hat{\beta}_j) = \sqrt{\mathrm{Var}(\hat{\beta}_j)} = \sqrt{\frac{\sigma^2}{SST_j(1-R_j^2)}}. \tag{4.27}$$

standard error of $\hat{\beta}_j$:

$$se(\hat{\beta}_j) = \sqrt{\frac{\hat{\sigma}^2}{SST_j(1-R_j^2)}}. \tag{4.28}$$

4.6 Efficiency of OLS: The Gauss-Markov Theorem

定理 4.4 (Gauss-Markov Theorem)

Under assumptions MLR 1-5, OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ are **B**est **L**inear **U**nbiased **E**stimators ("BLUEs") of $\beta_0, \beta_1, \cdots, \beta_k$, respectively. ♡

注

1. **E**stimator: It is a rule that can be applied to any sample of data to produce an estimate.
2. **U**nbiased: an estimator β_j, 有: $E(\hat{\beta}_j) = \beta_j$.
3. **L**inear: An estimator $\tilde{\beta}_j$ is linear if and only if it can be expressed as a linear function (a "weighted average") of y_i:
$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i,$$
where each w_{ij} can be a function of the sample values of all the independent variables.
4. **B**est: Best is defined as smallest variance.

前面的三条都比较好证明, 我们来说明第 4 条. 也就是:

Let $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ denote the OLS estimators under Assumptions MLR.1-MLR.5. We need to show that, for any linear and unbiased estimator $\tilde{\beta}_j$, $\mathrm{Var}(\tilde{\beta}_j) \geqslant \mathrm{Var}(\hat{\beta}_j)$.

证明 现在我们有:

$$\begin{aligned}
\tilde{\beta}_j &= \sum_{i=1}^{n} w_{ij} y_i \\
&= \sum_{i=1}^{n} w_{ij} (\beta_0 + \beta_1 x_{ij} + \cdots + \beta_k x_{ij}) \\
&= \beta_0 \sum_{i=1}^{n} w_{ij} + \beta_1 \sum_{i=1}^{n} w_{ij} x_{ij} + \cdots + \beta_k \sum_{i=1}^{n} w_{ij} x_{ij}.
\end{aligned}$$

考虑到: $E(\tilde{\beta}_j) = \beta_j$, 那么:

$$\begin{aligned}
\sum_{i=1}^{n} w_{ij} x_{ij} &= 1, \\
\sum_{i=1}^{n} w_{il} x_{il} &= 0 (l \neq j),
\end{aligned}$$

这里 $x_{i0} = 1$.

那么:

$$\mathrm{Var}(\tilde{\beta}_j | x) = \mathrm{Var}\left(\sum_{i=1}^{n} w_{ij} u_i | x \right) = \sum_{i=1}^{n} w_{ij}^2 \mathrm{Var}(u_i | x) = \sigma^2 \sum_{i=1}^{n} w_{ij}^2,$$

这里用到了 MLR.5 的 $\mathrm{Cov}(u_i, u_j) = 0$.

现在我们考虑: $x_{ij} = \hat{x}_{ij} + \hat{r}_{ij}$, 也就是我们之前所讲的辅助回归. 这个 $\hat{x}_{ij}$ 是一堆 x 的线性的和 (不含有 x_j). 那么我们可以有: $\sum_{i=1}^{n} w_{ij} \hat{x}_{ij} = 0$.

既然如此, 那么我们有: $\sum_{i=1}^{n} w_{ij}\hat{r}_{ij} = 1$.

同时, $\mathrm{Var}(\hat{\beta}_j) = \sigma^2/(\sum_{i=1}^{n} \hat{r}_{ij}^2)$, 那么:

$$\begin{aligned}\mathrm{Var}(\tilde{\beta}_j|x) - \mathrm{Var}(\hat{\beta}_j|x) &= \sigma^2 \left(\sum_{i=1}^{n} w_{ij}^2 - \frac{(\sum_{i=1}^{n} w_{ij}\hat{r}_{ij})^2}{\sum_{i=1}^{n} \hat{r}_{ij}^2} \right) \\ &= \sigma^2 \sum_{i=1}^{n} (w_{ij} - \hat{\gamma}_j \hat{r}_{ij})^2,\end{aligned}$$

这里的 $\hat{\gamma}_j = \left(\sum_{i=1}^{n} w_{ij}\hat{r}_{ij}\right) / (\sum_{i=1}^{n} \hat{r}_{ij}^2)$.

从而这个式子一定非负, 也就是: $\mathrm{Var}(\tilde{\beta}_j) \geqslant \mathrm{Var}(\hat{\beta}_j)$.

注 必须是线性的 estimator, 如果是下面的题目的话:

If Gauss-Markov assumptions hold, let $\tilde{\beta}$ be an estimator satisfies $E(\tilde{\beta}) = \beta$, and $\hat{\beta}_{OLS}$ is the OLS estimator. Then we must have $\mathrm{Var}(\tilde{\beta}) < \mathrm{Var}(\hat{\beta_{OLS}})$ is for any sample.

这段话是错误的, 就是因为 $\tilde{\beta}$ 可能不是一个线性的 estimator.

第五章 Inference

5.1 Sampling Distributions of the OLS Estimators

现在我们加入新的假设, 正态性总体假设 (MLR.6 Normality Assumption):

The population error u is independent of the explanatory variables $x_1, x_2, \cdots, x_k$ and is normally distributed with zero mean and variance σ^2 : $u \sim N(0, \sigma^2)$.

Assumptions MLR.1-MLR.6 are called the **classical linear model (CLM) assumptions**.

这实际上是对 y 做了正态性的假设.

如果我们假设 u 是各种因素加在一起的, 那么根据中心极限定理, 这一堆东西加起来就很服从正态分布.

一个非常弱的假设: There are some examples where Assumption MLR.6 is clearly false (eg. y takes on just a few values).

定理 5.1 (Normal Sampling Distributions)

Under the CLM assumptions MLR.1-MLR.6, conditional on the sample values of the independent variables,

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

where $\mathrm{Var}(\hat{\beta}_j) = \sigma^2/[SST_j(1-R_j^2)]$. 从而,

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

♡

5.2 Testing Hypotheses about a Single Population Parameter: The t Test

我们可以假设 $\beta_j =$ a value, 然后再检验是不是对的.

定理 5.2 (t Distribution for Standardized Estimators)

Under the CLM assumptions MLR.1-MLR.6,

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

♡

证明 证明概要: 在 $se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2/[SST_j(1-R_j^2)]}$ 中, $\hat{\sigma}^2 = SSR/(n-k-1)$, 而 SSR 是一堆残差的平方和, 残差又服从正态分布, 从而上面的式子服从 t 分布.

注 当 t 分布的自由度 $\geqslant 120$ 的时候, t 分布都近似于标准正态分布.

5.2.1 Steps for Testing the Significancy of an Estimator

我们主要的 null hypothesis 是:

$$H_0 : \beta_j = 0.$$

我们使用 t statistic 或者 t ratio of $\hat{\beta}_j$:

$$t_{\hat{\beta}_j} \equiv \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j - a_j}{se(\hat{\beta}_j)}$$

一般 $a_j = 0$. $\hat{\beta}_j$ 大, 经济显著; $se(\hat{\beta}_j)$ 小, 更加准确.

当存在两个变量线性关系的时候, 对应的变量的 R_j^2 会增大, 因为更容易被另一个变量解释了, 同时方差也增加, $se(\hat{\beta}_j)$ 也增加, 导致 t 检验量变小, 使得 t 检验量更不容易落入拒绝域中, 从而使得 t 检验变得更加容易失败.

在 Stata 中默认的是双边检验且 $H_0 : \beta_j = 0$.

Significant value: α 表示弃真的概率.

Critical value: c, 根据 t 值和 c 对比来决定是否拒绝零假设.

1. One-sided alternative: $H_1 : \beta_j > a_j$. We reject H_0 in favor of H_1 when the parameter estimate $\hat{\beta}_j$ is "sufficiently greater than a_j":
   (a). Calculate the t-statistic: $t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - a_j}{se(\hat{\beta}_j)}$.
   (b). Compare $t_{\hat{\beta}_j}$ with the critical value c for a significance level α (example: $\alpha = 5\%$). 这里的 $c = t_{n-k-1}^{-1}(0.95)$.
   - If n is large, $c = 1.645$.
   - If $t_{\hat{\beta}_j} > c$, we reject the null hypothesis at a significance level of 5%.
   - If $t_{\hat{\beta}_j} \leqslant c$, we fail to reject the null hypothesis at a significance level of 5%.
2. One-sided alternative: $H_1 : \beta_j < a_j$. We reject H_0 in favor of H_1 when the parameter estimate $\hat{\beta}_j$ is "sufficiently smaller than a_j":
   (a). Calculate the t-statistic: $t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - a_j}{se(\hat{\beta}_j)}$.
   (b). Compare $t_{\hat{\beta}_j}$ with the critical value c for a significance level α (example: $\alpha = 5\%$). 这里的 $c = t_{n-k-1}^{-1}(0.95)$.
   - If n is large, $c = 1.645$.
   - If $t_{\hat{\beta}_j} < -c$, we reject the null hypothesis at a significance level of 5%.
   - If $t_{\hat{\beta}_j} \geqslant -c$, we fail to reject the null hypothesis at a significance level of 5%.
3. Two-sided alternative: $H_1 : \beta_j \neq a_j$. We reject H_0 in favor of H_1 when the parameter estimate $\hat{\beta}_j$ is "far from a_j in absolute value":
   (a). Calculate the t-statistic: $t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - a_j}{se(\hat{\beta}_j)}$.
   (b). Compare $t_{\hat{\beta}_j}$ with the critical value c for a significance level α (example: $\alpha = 5\%$). 这里的 $c = t_{n-k-1}^{-1}(0.975)$. 现在的临界值的计算要根据 $\alpha/2$.
   - If n is large, $c = 1.960$.

- If $|t_{\hat{\beta}_j}| > c$, we reject the null hypothesis at a significance level of 5%.
- If $|t_{\hat{\beta}_j}| \leqslant c$, we fail to reject the null hypothesis at a significance level of 5%.

如果我们拒绝了原假设, 我们说: x_j 对 y 的影响 is statistically significant at the α% level.

5.2.2 Computing p-values for t Tests

An alternative to the classical approach : If the calculated t statistic is used as critical value, what is the smallest significance level at which the null hypothesis would be rejected?

This level is known as the p-value. 就是最小的显著水平.

$$p\text{-value} = P(|T| > |t|).$$

我们可以标 $*$, $*$ 越多越显著.

5.2.3 Economic Significance VS Statistical Significance

Statistical significance is determined totally by how large the t statistic is.

Economic significance focuses on the magnitude of the estimated coefficients.

经济显著, 但是统计不一定显著的. 因为可能样本很小, 导致标准误特别大.

经济不显著, 但是统计显著. 这是可能是有一定影响, 但是影响没有那么大.

5.3 Confidence Intervals

对于这里的 t 检验, 我们有:

$$P(-c < t < c) = 95\%.$$

这里的 $t = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$, 从而我们可以解出括号里的 β_j 的取值, 这就是 β_j 的 95% 置信区间, 也就是: $[\hat{\beta}_j - c \cdot se(\hat{\beta}_j), \hat{\beta}_j + c \cdot se(\hat{\beta}_j)]$.

95% confidence interval: If random samples were obtained over and over again, then the unknown population value β_j would lie within the confidence interval for 95% of the samples.

5.4 Testing Hypotheses about a Single Linear Combination of the Parameters

现在我们要检验的原假设变为参数的一些线性组合, 比如: $H_0 : \beta_1 = \beta_2$.

那么我们的 t statistic 为:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)}.$$

而:

$$\text{Var}(\hat{\beta}_1 - \hat{\beta}_2) = \text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) - 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$$

获取 Covariance Matrix:

```
. matrix list e(V)


symmetric e(V)[4,4]
// 这里是一个协方差矩阵
```

但是我们有一个更好的办法.

我们可以设 $\theta_1 = \beta_1 - \beta_2$, 然后重新带入原方程. 对改造过后的方程作 t 检验即可.

5.5 Testing Multiple Linear Restrictions

Multiple null hypothesis:

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

Alternative hypothesis:

$$H_1 : H_0 \text{ is not true.}$$

The null hypothesis constitutes three **exclusion restrictions**: If H_0 is true, then the three variables have no joint effect on y and therefore should be excluded from the model.

Using separate t statistics to test a multiple hypothesis can be very misleading. 因为即使单个不显著, 可能在这三个模型总体作用下就显著了.

We estimate **a restricted model** without the three variables.

Now we can compare the restricted model to the unrestricted model (i.e. the model that includes all variables).

Since the OLS estimates are chosen to minimize the sum of squared residuals, the SSR always **increases** when variables are dropped from the model (this is also the reason why R^2 always increases when variables are added to the model). 也就是 x 越少, SSR 越大.

The SSR from the restricted model is greater than the SSR from the unrestricted model (and the R^2 from the restricted model is smaller than the R^2 from the unrestricted model).

现在我们不妨假设我们检验最后的 q 个变量是否有零系数. 我们有 F statistic (or F ratio) 为:

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F(q, n-k-1).$$

如果 F 检验显著, 那么我们说它们是联合显著 (jointly significant) 的.

```
. test bavg=hrunsyr=rbisyr=0
F( 3, 347) = 9.55
  Prob > F = 0.0000
```

Using the fact that $SSR_r = SST(1-R_r^2)$ and $SSR_{ur} = SST(1-R_{ur}^2)$; we can derive an alternative formula for the F statistic:

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

overall significance: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$, 也就是 Stata 右上角的 F statistic.

如果我们检验的有一个系数不是 0 的话, 比如 $\beta_1 = 1, \cdots, \beta_3 = 0$, 那么我们实际上就变成了检验: $y - x_1 = \beta_0 + \beta_1 x_1 + \cdots + \beta_3 x_3$.

这个时候我们不能用 R^2 版本的, 因为 y 变了, Restricted model 的 y 变成了 $y - x_1$. 但是 SSR 是和 y 无关的, 因为 $u_i' = (\hat{y}_i - x_{i1}) - (y_i - x_{i1}) = \hat{y}_i - y_i = u_i$.

5.6 Reporting Regression Result

我们有下面的一张表:

Testing the Salary-Benefits Tradeoff			
Dependent Variable: log(*salary*)			
Independent Variables	(1)	(2)	(3)
b/s	-.825 (.200)	-.605 (.165)	-.589 (.165)
log(*enroll*)	———	.0874 (.0073)	.0881 (.0073)
log(*staff*)	———	-.222 (.050)	-.218 (.050)
droprate	———	———	-.00028 (.00161)
gradrate	———	———	.00097 (.00066)
intercept	10.523 (0.042)	10.884 (0.252)	10.738 (0.258)
Observations *R*-squared	408 .040	408 .353	408 .361

表 5.1: Reporting Regeression Result

上面的是系数, 下面的是标准误, 这样直接就可以作 t 检验. 同时又可以根据这 3 个模型之间的 R^2 的数据来作 F 检验.

第六章 Multiple Regression Analysis: Further Issues

6.1 Effects of Data Scaling on OLS Statistics

改变 the units of measurement 不会改变 R^2.

例题 6.1 对于下面的回归:

$$\widehat{bwght} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc,$$

我们作回归 `regress bwght cigs faminc`. 都是显著的. 如果我们设: $bwghtlbs = bwght/16$, 那么就有:

$$\widehat{bwght}/16 = \hat{\beta}_0/16 + \hat{\beta}_1/16 cigs + \hat{\beta}_2/16 faminc,$$

注 如果 $y^* = cy$, 那么 $se(\hat{\beta}_j^{\ *}) = c\ se(\hat{\beta}_j)$(因为根号里分子上的 SSR 放大了 c 倍, 分母不变(都是 x 的部分)), t 不变, p-value 不变, 置信区间上下界都变成 c 倍. 所有的 F 值不变.

如果我们设某个 $x_j^* = cx_j$, 那么 $se(\hat{\beta}_j^{\ *})$ 中的分子不变, 分母变成 c 倍, 那么总体变成 $1/c$, t 不变, p-value 不变, 置信区间上下界都变成 $1/c$. 所有的 F 值不变. 别的系数的标准误不变.

6.2 Standardized Coefficients

Instead of looking at the change in the test score by one unit, we ask what happens when the test score is one standard deviation higher.

我们要探究 IQ 在分布中增加了一个标准差会怎么样. 去看标准差的变化.

Standardized Regeression.

我们设:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} + \hat{u}_i.$$

1. 减去均值 (demean).

$$y_i - \bar{y} = \hat{\beta}_1(x_{i1} - \bar{x_1}) + \hat{\beta}_2(x_{i2} - \hat{x_2}) + \cdots + \hat{\beta}_k(x_{ik} - \hat{x_k}) + \hat{u}_i.$$

2. 除以标准差.

$$(y_i - \bar{y})/\hat{\sigma_y} = (\hat{\sigma_1}/\hat{\sigma_y})\hat{\beta}_1[(x_{i1} - \bar{x_1})/\hat{\sigma_1}] + \cdots + (\hat{\sigma_k}/\hat{\sigma_y})\hat{\beta}_k[(x_{ik} - \bar{x_k})/\hat{\sigma_k}] + \hat{u}_i/\hat{\sigma_y}.$$

3. 重新写成:

$$z_y = \hat{b_1} z_1 + \hat{b_2} z_2 + \cdots + \hat{b_k} z_k + error.$$

这里标准化的系数 (Standardized coefficients) 是

$$\hat{b_j} = \hat{\sigma_j}/\hat{\sigma_k}\hat{\beta}_j, 1 \leqslant j \leqslant k.$$

Interpretation: If x_1 increases by one standard deviation, then $\hat{y}$ changes by $\hat{b_1}$ standard deviations.

好处有:

1. unit-free.
2. 当 x 之间的单位不可比而想比较对 y 影响大小的时候. 我们可以得到 x 和 y 的 z-score.

这是一个没有截距项的回归方程. 不会影响 R^2 和 t-test 等.

如果我们在 Stata 中使用 `reg y x,beta` 即可得到标准化的系数.

如果遇见 score, 身高什么的, 要做标准化.

6.3 Functional Forms

只考虑下面的三种:

1. 自然对数. 我们在3.4节中讨论了.
2. 二次项. quadratic forms of x.
3. x 的交叉项. interactions of x variables.

时间单位不取 log.

1. Log form: logs are typically applied to positive dollar amounts (wages, expenditure, firm sales, etc.) and large integer values (population, school enrollment, etc.).
2. Level form: Variables that are measured in years (education, age, etc.) usually appear in original form.
3. Either in log or in level: Proportions/shares/rates (unemployment rate, Gini coefficient, etc.) can appear in both forms. Interpretation differs: percentage change vs. percentage point change.

如果有一些 0, 可以 $\log(1+y)$.

如果是二次项. x 对 y 的影响是:

$$\frac{\Delta \hat{y}}{\Delta x} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x.$$

在 $x = -\hat{\beta}_1/(2\hat{\beta}_2)$ 取得最大值, 观察左右两边的样本数量是否足够支持这一个最大值.

交互项. 我们考虑下面的模型:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u.$$

这里面的 β_2 就是当 $x_1 = 0$ 的时候 x_2 对 y 的影响. 这个参数的经济学含义非常的弱, 因为一般不会有 0 的情况.

我们重新写成:

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

那么:

$$\delta_1 = \beta_1 + \beta_3\mu_2,$$
$$\delta_2 = \beta_2 + \beta_3\mu_1.$$

这样做的好处在于: 使得 δ_2 表示 $x_1 = \mu_1$ 的时候 x_2 对 y 的影响, 这在经济学里还是很常见的, **使得参数的经济学含义更强了.**

6.4 More on Goodness-of-Fit

R^2 是用来检测 Goodness of fit 的. 注意到 $R^2 = SSE/SST$. 如果我们增加一个解释变量, R^2 总是会增加. 但是增加一个变量有时候并不意味着我们模型 Goodness of fit 增加了, 所以我们要选择一个更加合适的指标.

The Adjusted R^2: 添加了自由度.

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

对单纯的由于解释变量的增加导致的 R^2 的增加一个 punishment. 减小了单纯由于 k 的增加而导致的 R^2 的增加.

两者之间的关系:

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

甚至有可能是负的, 说明我们的方程非常的差.

R^2 和 $\bar{R}^2$ 不会告诉我们:

1. whether an included variable is statistically significant;
2. whether the X are a true cause of Y;
3. whether there is omitted variable bias, or you have chosen the most appropriate and complete set of regressors.

Adjusted R^2 的应用: 选择一个最好的模型.

Non-Nested Models: two equations are considered non-nested, if neither equation is a special case of the other.

嵌套模型: restricted model 和 unrestricted model.

例题 6.2 对于下面的两个模型:

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + u,$$
$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 rbisyr + u.$$

选择一个最好的.

F-statistic 可以帮助我们在嵌套模型里选择一个最好的. 它的作用是决定一个组里是否至少有一个变量对 dependent variables 有影响. 但是没有办法确定哪个变量有影响.

那么又如何对非嵌套模型来选择呢? 一个办法就是根据 Adj R^2 来选择, 选择 $\bar{R}^2$ 更大的

那一个.

$\bar{R}^2$ 可以选择不同形式的自变量的 functional forms. 比如:

例题 6.3

$$rdintens = \beta_0 + \beta_1 \log(sales) + u,$$
$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u.$$

我们之所以不使用 R^2 是因为这两个模型参数的数量不一样, 使用 R^2 不公平.

但是要注意的是: R^2 和 $\bar{R}^2$ 无法选择不同的因变量的 functional forms.

因为我们的 R^2 度量的是因变量总变化中可以被解释的比例, 现在这两个模型的因变量总变化的比例是不同的, 那么 R^2 就不能告诉我们哪个模型拟合得更好, 因为它们拟合的是 **两个完全不同的因变量**!

6.5 Prediction and Residual Analysis

假设我们想要预测一个参数 θ_0:

$$\begin{aligned}\theta_0 &= \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k \\ &= E(y|x_1 = c_1, x_2 = c_2, \cdots, x_k = c_k).\end{aligned}$$

那么 θ_0 的估计量就是:

$$\hat{\theta}_0 = \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k.$$

现在我们想获取这个预测 $\hat{\theta}_0$ 的置信区间. 那么我们就应该先得到 standard error of $\hat{\theta}_0$.

首先重新写成: $\beta_0 = \theta_0 - \beta_1 c_1 - \beta_2 c_2 - \cdots - \beta_k c_k$, 接着把这个式子插入原来的 OLS 方程中就会得到:

$$y = \theta_0 + \beta_1(x_1 - c_1) + \beta_2(x_2 - c_2) + \cdots + \beta_k(x_k - c_k) + u.$$

那么我们只要计算出这个方程的截距的 se 即可. 这样置信区间就是:

$$[\hat{\theta}_0 - c \cdot se(\hat{\theta}_0), \hat{\theta}_0 + c \cdot se(\hat{\theta}_0)],$$

这里 c 就是我们需要的显著水平计算出来的临界值. (定理5.2告诉我们这些参数服从的都是一个 t 分布.)

我们设我们现在收集到了一堆新的自变量的值, $x_1^0, x_2^0, \cdots, x_k^0$, 我们用 y^0 表示这个观测对应的因变量, 我们想估计它. 我们就通过回归方程计算得到了 $\hat{y}^0$, 现在我们想借此求得这个 y^0 的置信区间.

现在我们设:

$$\hat{e}^0 = y^0 - \hat{y}^0 = (\beta_0 + \beta_1 x_1^0 + \beta_2 x_2^0 + \cdots + \beta_k x_k^0) + u^0 - \hat{y}^0.$$

那么 $E(\hat{e}^0) = 0, \mathrm{Var}(\hat{e}^0) = \mathrm{Var}(\hat{y}^0) + \sigma^2$.

而 $se(\hat{y}^0)$ 根据上面的关于如何计算 $\hat{\theta}_0$ 的 se 可以得到.

所以最后一个 $100(1-\alpha)\%$ 的置信区间就是:

$$\hat{y}^0 \pm c_{\alpha/2} se(\hat{e}^0).$$

我们还可以给出一个从 $\log y$ 的式子得到 y 的估计:

如果 $u \sim N(0, \sigma^2)$, 那么 $E(e^u) = e^{\frac{1}{2}\sigma^2}$.

那么:

$$\begin{aligned} E(y|X) &= E[e^{\log(y)}|X] \\ &= e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k} E(e^u|X) \\ &= \exp\left(\frac{\sigma^2}{2} + (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)\right). \end{aligned}$$

所以最后我们就得到一个 y 的预测:

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

第七章 Dummy Variables

7.1 A Single Dummy Independent Variable

虚拟变量 (Dummy Variables): 值不重要, 但是分到哪个组非常重要.

Qualitative factors: binary information.

如果我们考虑下面的模型:

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u.$$

那么我们有:

$$\delta_0 = E(wage|female = 1, educ) - E(wage|female = 0, educ).$$

那么 δ_0 就代表在给定教育水平下, 女性和男性的**平均**薪资之间的差值. δ_0 经常被视为 "discrimination", 也就是女人挣得比男人少.

并且 u 一定不能和性别有关, 如果有关我们只能说有**差异**, 不能说有歧视.

基准组, 对照组, 取 0 的值, base, benchmark, comparison group.

如果我们换了一个虚拟变量为 $male$:

$$wage = \beta_0' + \delta_0' male + \beta_1' educ + u.$$

那么 $\beta_0' + \delta_0' = \beta_0$, $\delta_0' = -\delta_0$, $\beta_1' = \beta_1$.

7.2 Dummy Variables and Logged Dependent Variables

计算近似的组间的百分比的差异:

$$\Delta y/y \approx \hat{\beta}_1.$$

计算准确的组间的百分比的差异:

$$\Delta y/y = e^{\hat{\beta}_1} - 1.$$

7.3 Using Dummy Variables for Multiple Categories

General principle: if we want to consider group-specific intercepts for g groups, we need to include $g - 1$ dummy variables in the model along with an intercept. 否则有可能产生 dummy variable trap. 总体的截距就代表基准组的斜率.

例如, 两个组 male 和 female, 就丢掉 male 组, 防止多重共线性的问题. e.g. 我们有东,

西, 南, 北, 中五个组, 就要抛弃掉中这个组, 剩下:

$$y = \beta_0 + \beta_1 \text{east} + \beta_2 \text{west} + \beta_3 \text{south} + \beta_4 \text{north}.$$

我们还可以把两个 dummy variables 相乘得到四个 dummy variables.

7.3.1 Using Ordinal Information

例题 7.1 Effect of credit ratings (CR) on the municipal bond interest rate (MBR).

$$MBR = \beta_0 + \beta_1 CR + otherfactors.$$

但是如果说 CR 从 0 到 1 的变化更大呢? 我们可以用虚拟变量: CR_1, CR_2, CR_3, CR_4.

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + otherfactors.$$

所以一个虚拟变量和连续变量之间是可以互相转化的. 同时我们有:

$$\delta_1 = \beta_1, \delta_2 = 2\beta_1, \delta_3 = 3\beta_1, \delta_4 = 4\beta_1.$$

所以连续变量是一个非常受限制的模型.

7.4 Interactions Involving Dummy Variables

讨论虚拟变量的交互项.

我们首先考虑一个虚拟变量和另一个虚拟变量相乘的交互项.

7.4.1 Interaction of a Dummy with Another Dummy

例题 7.2 Consider the wage model with female-married interaction with the base group still be single men. 我们用下面的模型:

$$\log(wage) = \beta_0 + \beta_1 female + \beta_2 married + \beta_3 female \cdot married + otherfactors.$$

例如, 单身女性的截距项就会变成 $\beta_0 + \beta_1$. 已婚女性的截距就会变成 $\beta_0 + \beta_1 + \beta_2 + \beta_3$.

这里的 β_1 获取的是单身男性和单身女性之间的平均工资差异. β_2 是已婚男和单身男之间的平均工资差异. β_3 没有办法被单独解释.

7.4.2 Interacting Dummies with Continuous Variables

例题 7.3 我们允许男女的教育回报是不一样的. 我们考虑下面的模型:

$$\begin{aligned}\log(wage) &= (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female)educ + u \\ &= \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \cdot educ + u.\end{aligned}$$

下面是两种可能的情况:

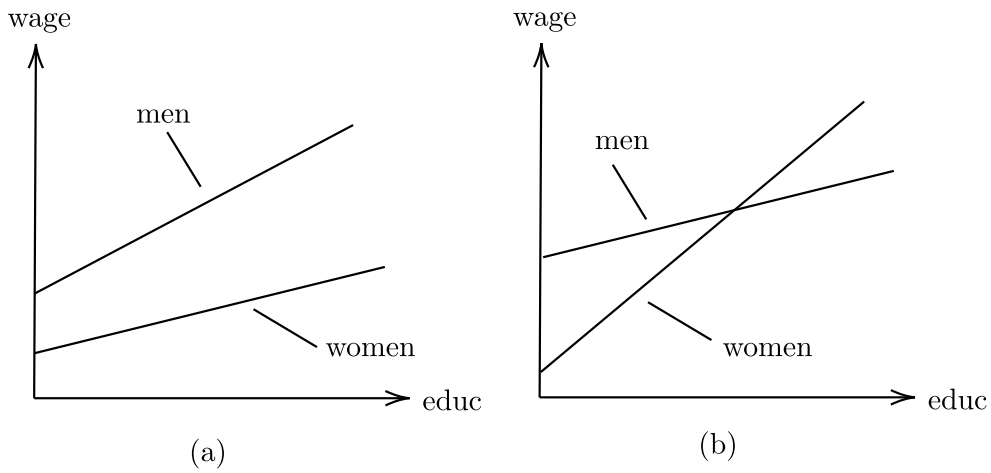

图 7.1: Are the returns to education different for women and men?

An important hypothesis is that the return to education is the same for women and men ($H_0 : \delta_1 = 0$).

We can also test H_0 : Average wages are identical for men and women who have the same levels of education. This is to test $H_0 : \delta_1 = 0, \delta_0 = 0$.

但是有可能 female 前面的系数并不显著, 这是因为我们加入了 $female \cdot educ$ 导致了不精确的系数的估计. 我们可以使用 $female \cdot (educ - \overline{educ})$ 来改进这个模型.

7.4.3 Testing for Differences in Regression Functions across Groups

检测组间是不是存在系统性的差异.

首先假设男性的回归方程是:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + u.$$

接着假设女性的回归方程是:

$$log(wage) = (\beta_0 + \delta_0) + (\beta_1 + \delta_1)educ + (\beta_2 + \delta_2)exper + u.$$

接着我们做检验: $H_0 : \delta_0 = \delta_1 = \delta_2 = 0$.

还可以做 estimate a fully interacted model to compare differences in intercept and slopes between two groups.

也就是我们令:

$$\begin{aligned} log(wage) = {} & \beta_0 + \beta_1 educ + \beta_2 exper \\ & + (\delta_0 + \delta_1 educ + \delta_2 exper) \cdot female + u. \end{aligned}$$

然后做 F 检验: $\delta_0 = \delta_1 = \delta_2 = 0$.

7.4.3.1 The Chow Test

周氏检验.

We estimate the following **unrestricted model** separately for the 2 groups:

$$y = \beta_{g,0} + \beta_{g,1}x_1 + \cdots + \beta_{g,k}x_k + u.$$

We compare the estimates to those of a **restricted model** in which all coefficients are the same:

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

The sum of squared residuals from the unrestricted model can be obtained from two separate regressions, one for each group:

$$SSR_{ur} = \sum_g SSR_g.$$

unrestricted model 的自由度为 $n - 2(k+1)$. 因为有 $k+1$ 个参数, 2 组.

The restricted sum of squared residuals is the SSR from pooling the two groups and estimating a single equation:

$$SSR_r = SSR_P.$$

我们就会得到 F statistic (**Chow statistic**):

$$F = \frac{[SSR_P - (SSR_1 + SSR_2)]}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k+1)]}{k+1}.$$

我们的 Stata 代码如下:

```
regress lwage educ exper if female==1
regress lwage educ exper if female==0
regress lwage educ exper
```

然后计算 F 检验量.

7.5 The Linear Probability Model

如果我们把解释变量 y 改造成一个 dummy variable.

线性可能性模型.

$$E(y|\mathbf{x}) = P(y=1|\mathbf{x})E(y=1) + P(y=0|\mathbf{x})E(y=0) = P(y=1|\mathbf{x})$$
$$\Rightarrow P(y=1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

In the LPM, β_j measures the change in the probability of $y = 1$ when x_j changes, holding other factors fixed:

$$\Delta P(y=1|\mathbf{x}) = \beta_j \Delta x_j.$$

LPM 一定具有异方差 (Heteroskedasticity).

首先我们有:

$$E(y|\mathbf{x}) = P(y = 1|\mathbf{x}) = p.$$

$$\text{Var}(y|\mathbf{x}) = E[(y - \bar{y})^2|\mathbf{x}] = (1-p)^2 \cdot p + (0-p)^2 \cdot (1-p) = p(1-p).$$

也就是:

$$\text{Var}(u|\mathbf{x}) = \text{Var}(y|\mathbf{x}) = P(y = 1|\mathbf{x})[1 - P(y = 1|\mathbf{x})].$$

Heteroskedasticity does not cause bias in OLS estimators, but the t and F statistics needs the homoskedasticity assumption. Hence the standard errors in LPM are wrong, but usually not far off.

In many applications, the usual OLS statistics are not far off, and it is still acceptable in applied work to present a standard OLS analysis of a linear probability model.

7.6 More on Policy Analysis and Program Evaluation

例题 7.4 The effect of the job training grants(拨款) on worker productivity (Holzer et al., 1993)

$$\log(scrap) = \beta_0 + \beta_1 grant + \beta_2 \log(sales) + \beta_3 \log(employ) + u.$$

销售的规模和就业的规模会影响企业是否申请 grant. (grant 是企业是否申请拨款)

Critics: Some unobserved factors may affect worker productivity and at same time correlated with whether the firm receives a grant.

The authors point out that grants were given on a first-come, first-served basis. This is definitely not the same as giving out grants randomly.

例题 7.5 The impact of microcredit program(小额贷款) on the re-employment of laid-off women(下岗女工)

$$reemployed = \beta_0 + \beta_1 participate + u.$$

participate 是是否参与小额贷款项目.

Self-selection problem(自选择问题): Individuals self-select into certain programs; participation is not randomly determined. (不是随机分派的, 是自己选择的.)

If that is the case, the binary indicator of participation might be systematically related to unobserved factors, i.e.,

$E(u|participate = 1) \neq E(u|participate = 0) \Rightarrow$ participate is endogenous, related to u.

内生性的问题.

例题 7.6 Testing for discrimination in loan approvals. (贷款申请的歧视检验)

A linear probability model is used to test for discrimination:

$$approved = \beta_0 + \beta_1 nonwhite + \beta_2 income + \beta_3 wealth + \beta_4 credrate + u.$$

Race would appear to be the perfect example of an exogenous explanatory variable, given that it is determined at birth. But there might be systematic differences in some social-economic factors across race.

Therefore, it is necessary to **control for factors that affect approval and are systematically different across race.**

raw wage gap: 什么都不控制, 硬比较. `reg wage gender`.

gender wage differentials.

discrimination: 什么都控制.

第八章 Heteroscedasticity

开始 relax 先前的假设. 首先 relax MLR.5.

8.1 Estimating Robust Standard Errors

MLR.5 不影响无偏性. 但是会对方差的估计有影响, 以及影响 standard error, 同时也不再 BLUE 了. 所有的假设检验也都不 reliable 了.

在异方差存在的情况下, 我们试图找到一个稳健的标准误.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i.$$

我们假设 MLR.1-4 成立. 我们设:

$$\mathrm{Var}(u_i | x_{i1}, x_{i2}, \cdots, x_{ik}) = \sigma_i^2.$$

$$\hat{\beta}_j = \frac{\sum_{i=1}^n \hat{r}_{ij} y_i}{\sum_{i=1}^n \hat{r}_{ij}^2} = \beta_j + \frac{\sum_{i=1}^n \hat{r}_{ij} u_i}{\sum_{i=1}^n \hat{r}_{ij}^2}.$$

那么:

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^n \hat{r}_{ij}^2 \sigma_i^2}{SSR_j^2}.$$

White(1980) 证明了下面是对 $\mathrm{Var}(\hat{\beta}_j)$ 的一个有效的估计, 当 n 很大的时候:

$$\widehat{\mathrm{Var}}(\hat{\beta}_j) = \frac{\sum_{i=1}^n \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}.$$

那么 the heteroscedastic-robust standard errors 就是:

$$se(\hat{\beta}_j) = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)} = \sqrt{\frac{\sum_{i=1}^n \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}}.$$

MacKinnon and White (1985): The formula has to be corrected by the degrees of freedom,

$$se(\hat{\beta}_j) = \sqrt{\frac{n}{n-k-1} \frac{\sum_{i=1}^n \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}}.$$

例题 8.1 Model with Usual OLS Standard Errors.

Model with Robust OLS Standard Errors:

```
regress lwage marrmale marrfem singfem educ exper expersq tenure tenursq, robust
   Linear regression
```

注 Robust standard errors can be either larger or smaller than the usual standard errors.

Empirically, the robust standard errors are often found to be larger than the usual standard errors.

一般来说, Robust OLS Standard Errors 更大.

With small sample sizes, robust t statistics may have **wrong** distributions. 我们不知道异方差性是否存在的, 所以就需要无论异方差性是否成立都可以有效的这个 robust se.

那既然如此为什么我们还需要一个 OLS standard errors 呢? 因为如果一旦同方差性成立了, 那么我们就会得到真实的 t 分布. In practice, we typically use robust standard errors in large samples, especially in using cross-sectional data.

大样本, 截面数据 $\Rightarrow$ 稳健的标准误.

8.2 Testing for Heteroscedasticity

根据图3.1, 我们知道如果是同方差性的话, 残差不会随着解释变量而变化. 所以我们可以画出残差和解释变量的这样一个散点图, 来判断是否是一个异方差性.

最好的检验方法就是散点图.

```
predict e,residuals
gen r2 = e*e
scatter r2 X
```

8.2.1 Breusch-Pagan Test for Heteroscedasticity

必须要 MLR.1-4 成立.

$$H_0 : \mathrm{Var}(u|x_1, x_2, \cdots, x_k) = \sigma^2.$$

接着我们转换成:

$$H_0 : \mathrm{Var}(u|x_1, x_2, \cdots, x_k) = E(u^2|x_1, x_2, \cdots, x_k) = E(u^2) = \sigma^2.$$

那么我们就要检验:

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

那么我们现在就检验:

$$H_0 : \delta_0 = \delta_1 = \cdots = \delta_k = 0.$$

F statistic 就是:

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

检验的步骤:

1. Estimate the model by OLS, as usual. Obtain $\hat{u}_i^2$ for each observation i.

2. Run the regression $\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + error$. Keep the R-squared from this regression, $R^2_{\hat{u}^2}$.
3. Form the F statistic and compute the p-value. If the p-value is sufficiently small, then we reject the null hypothesis of homoscedasticity.

```
regress price lotsize sqrft bdrms
predict uhat, residual
generate uhatsq=uhat*uhat
regress uhatsq lotsize sqrft bdrms
```

用 log-form 之后貌似同方差就没有被违反了. (取了 log-form 可能是解决了方程形式误设的问题, 解决内生性的问题.)

8.2.2 White Test for Heteroscedasticity

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

一个更简单的版本:

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

White Test 的步骤:

1. Estimate OLS regression. Obtain the OLS residuals $\hat{u}$ and the fitted values $\hat{y}$. Compute $\hat{u}^2$ and $\hat{y}^2$.
2. Run the regression $\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta \hat{y}^2 + error$. Keep the R-squared from this regression, $R^2_{\hat{u}^2}$.
3. Form the F statistic and compute the p-value. If the p-value is sufficiently small, then we reject the null hypothesis of homoscedasticity.

```
reg y x1 x2 x3
predict yhat, xb
predict uhat, residual
gen uhatsq = uhat * uhat
gen yhatsq = yhat * yhat
reg uhatsq yhat yhatsq
```

It is better to use explicit tests for functional form first, since functional form misspecification is more important than heteroskedasticity.

8.3 Weighted Least Squares (WLS) Estimation

这是一种新的减少异方差性的方法. 或者这个章节不如叫 GLS 刚好和下一章节也能联系起来.

我们假设:

$$\mathrm{Var}(u_i|x_{i1},x_{i2},\cdots,x_{ik})=\sigma^2 h(x_{i1},x_{i2},\cdots,x_{ik}).$$

For now, we assume that h_i is known.

我们考虑原来的方程;

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+u_i.$$

那么有:

$$y_i/\sqrt{h_i}=\beta_0/\sqrt{h_i}+\beta_1(x_{i1}/\sqrt{h_i})+\beta_2(x_{i2}/\sqrt{h_i})+\cdots+\beta_k(x_{ik}/\sqrt{h_i})+(u_i/\sqrt{h_i}).$$

接着变成:

$$y_i^*=\beta_0^*+\beta_1 x_{i1}^*+\beta_2 x_{i2}^*+\cdots+\beta_k x_{ik}^*+u_i^*.$$

同时可以计算出:

$$E(u_i^*)=E(u_i/\sqrt{h_i})=E(u_i)/\sqrt{h_i}=0.$$

$$\mathrm{Var}(u_i^*)=E((u_i^*)^2)=E(u_i^2)/h_i=\sigma^2.$$

```
generate ystar=nettfa/sqrt(inc)
generate x1star=inc/sqrt(inc)
generate x2star=agesq25/sqrt(inc)
generate x3star=male/sqrt(inc)
generate x4star=e401k/sqrt(inc)
gen constant=1/sqrt(inc)
regress ystar x1star x2star x3star x4star constant if fsize==1, noconstant
```

一定要 noconstant, 因为我们自己有一个 constant 了.

OLS → GLS → WLS.

The OLS estimators of the transformed equation, $\beta_0^*,\beta_1^*,\cdots,\beta_k^*$ are examples of **generalized least squares (GLS) estimators** (in this case, they are used to account for heteroscedasticity).

The GLS estimators for correcting heteroscedasticity are called WLS estimators because the β_j^* minimize the weighted sum of squared residuals. (每一个 squared residual 的权重是 $1/h_i$.)

也就是我们最小化下面的式子:

$$\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_kx_{ik})^2/h_i=\sum_{i=1}^{n}u_i^2/h_i$$
$$=\sum_{i=1}^{n}(y_i^*-\beta_0-\beta_1x_{i1}^*-\beta_2x_{i2}^*-\cdots-\beta_kx_{ik}^*)^2=\sum_{i=1}^{n}(u_i^*)^2.$$

这个 GLS model 会满足所有的理想的假设, 所以是 BLUE 的.

例题 8.2 Financial wealth equation

$$nettfa=\beta_0+\beta_1inc+\beta_2(age-25)^2+\beta_3male+\beta_4e401k+u.$$

现在我们假设 $\text{Var}(u|inc)=\sigma^2inc$.

WLS 就可以被得到, 并且比 OLS 更加的精确.

8.4 Feasible Generalized Least Squares (FGLS)

In most cases, the functional form of $h_i=h(x_{i1},\cdots,x_{ik})$ is unknown.

我们来估计: $\hat{h}_i$. 这样的使用就会产生 **feasible GLS (FGLS) estimator**.

我们假设:

$$\text{Var}(u_i|x_{i1},\cdots,x_{ik})=\sigma^2h_i=\sigma^2\exp(\delta_0+\delta_1x_{i1}+\delta_2x_{i2}+\cdots+\delta_kx_{ik}).$$

$$u^2=\sigma^2\exp(\delta_0+\delta_1x_1+\delta_2x_2+\cdots+\delta_kx_k)\nu, E(\nu|x_1,\cdots,x_k)=1.$$

这个后面的式子是根据 $E(u^2)=\text{Var}(u)$ 得到的.

$$\log(u^2)=\alpha_0+\delta_1x_1+\delta_2x_2+\cdots+\delta_kx_k+e, E(e|x_1,\cdots,x_k)=0.$$

接下来我们对 $\log(\hat{u}^2)$ 在 $x_1,x_2,\cdots,x_k$ 上面进行回归, 然后得到:

$$g_i\equiv\log(\hat{u}_i^2), \hat{g}_i\equiv\widehat{\log(\hat{u}_i^2)}.$$

那么我们对于 h_i 的估计 $\hat{h}_i=\exp(\hat{g}_i)$. 随后我们把 WLS 中的 $1/h_i$ 换成 $1/\hat{h}_i$.

A Feasible GLS Procedure to Correct for Heteroscedasticity:

1. Run the regression of y on $x_1,x_2,\cdots,x_k$ and obtain the residuals, $\hat{u}$.
2. Create $\log(\hat{u}^2)$ by first squaring the OLS residuals and then taking the natural log.
3. Run the regression of $\log(\hat{u}^2)$ on $x_1,x_2,\cdots,x_k$ and obtain the fitted values, $\hat{g}$.
4. Exponentiate the fitted values from the above regression: $\hat{h}=\exp(\hat{g})$.
5. Estimate the equation

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_kx_{ik}+u_i.$$

by WLS, using weights $1/\hat{h}$.

第二种估计 h_i 的方法就是: reg $\log(\hat{u}^2)$ on $\hat{y}$, $\hat{y}^2$. 然后得到 $\hat{h}_i=\exp(\hat{g}_i)$.

OLS vs WLS: 一般来说, OLS 和 WLS 之间的显著的差距表明 MLR.4(x 和 u 有关) 失效了.

问题: 如果说方差并不是 $\text{Var}(u_i|\mathbf{x}_i) \neq \sigma^2 h_i$ 怎么办?

WLS estimators 是无偏的. 但是 WLS 的 Standard errors 和检验的统计量都不再 valid 了. 可以用 robust se. 位于8.1.

8.5 Heteroscedasticity in the Linear Probability Model (LPM)

Estimating the LPM by FGLS:

$$\text{Var}(y|\mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})].$$

那么我们就可以估计 $\text{Var}(y|\mathbf{x})$ 为:

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

那么我们就可以运用 FGLS.

Estimating the LPM by FGLS:

1. Estimate the model by OLS and obtain the fitted values, $\hat{y}$.
2. Determine whether all of the fitted values are inside the unit interval. If so, proceed to step (3). If not, some adjustment is needed to bring all fitted values into the unit interval.
3. Construct the estimated variances in $\hat{h}_i = \hat{y}_i(1 - \hat{y}_i)$.
4. Estimate the equation $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ by WLS, using weights $1/\hat{h}_i$.

第九章 More on Specification and Data Issues

9.1 Functional Form Misspecification

这是不满足假设 4 的一种情况. 这是 Omitted variable bias 的一个特例, 只不过遗失的变量恰好是 x 和 y 之间的一个关系.

导致有偏和不一致的问题. 我们只需要考虑三种情形, 就可以解决这个问题:

1. 加上平方项.
2. 加上交互项.
3. 使用 the level of a variable rather than its log form.

9.1.1 How to detect Misspecified Functional Form

加入一些平方项, 对数项等, 如果这个时候我们做一个 F-test 的联合检验, 此时显著则可以加入, 之前的方程形式误设了.

9.1.2 RESET as a General Test for Functional Form Misspecification

一个常用的方法是: Ramsey's (1969) regression specification error test (RESET).

考虑下面的模型:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u.$$

一旦 MLR. 4 被满足了我们就不应该向里面加入任何的平方项. 但是只加入平方项存在很大的问题:

1. 失去了很多自由度. We lose many degrees of freedom by adding quadratics for everything.
2. 解释更加困难. The interpretation is more difficult.
3. Adding quadratic terms does not solve all problems of nonlinearities.

RESET 的步骤如下:

1. Estimate $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$ to obtain $\hat{y}$.
2. Estimate expanded equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 u.$$

3. 检验下面的假设:

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

这个的 F 检验是 2 和 $n-k-1-2 = n-k-3$. 如果 $\hat{\delta}_1$ 和 $\hat{\delta}_2$ 联合不显著, 则原来的模型不存在方程形式误设的问题.

Limitation of RESET

1. The RESET test provides no real direction on how to proceed if the model is rejected.
2. The RESET has no power for detecting omitted variables or heteroscedasticity whenever they have expectations that are linear in the included independent variables.
3. If the functional form is properly specified, RESET has no power for detecting heteroskedasticity.

9.1.3 Tests against Nonnested Alternatives: Davidson-MacKinnon test

The Davidson-MacKinnon (1981) test may be used to decide whether an independent variable should appear in **level or logarithmic** form:

我们来检测下面两个模型:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$
$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u.$$

这是 nonnested models, 所以不能简单的使用 F 检验.

Davidson-MacKinnon (1981) test:

1. Estimate the first model to obtain the predicted values $\hat{y}$.
2. Estimate the second model to obtain the predicted values $\hat{\hat{y}}$.
3. Estimate the models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \hat{\hat{y}},$$
$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \theta_2 \hat{y} + error.$$

现在如果 $\hat{\theta}_1$ 显著, 那么 level equation 就被拒绝了.

现在如果 $\hat{\theta}_2$ 显著, 那么 log equation 就被拒绝了.

Limitation of Davidson-MacKinnon Test

1. The test cannot be applied if the sets of independent variables are different.
2. The test is not helpful if both models are rejected. (这种情况出现时, 选择 (adj) R^2 大的那个.)
3. If the test rejects level equation using, it does not mean that the log equation is the correct model. The level model can be rejected for a variety of functional form misspecification.

4. An even more difficult problem is obtaining nonnested tests when the leading case is y vs $\log(y)$.

9.2 Proxy Variables

但是比较重要的一个问题是, 我们的模型要处理的一个变量是不可获得的. 例如对于下面的模型, *ability* 就是 unavailable 的:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 expr + \beta_3 ability + u.$$

A proxy variable is a variable that is related to the unobserved variable that we would like to include in our model. 例如, 我们用 IQ 作为 *ability* 的代理变量.

我们设未使用代理变量的模型为:

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

这里面的 x_3^* 就是 unobserved variable, x_3 是可观测的代理变量. 我们设:

$$x_3^* = \delta_0 + \delta_3 x_3 + \nu_3.$$

下面是几个假设来保证使用 x_3 得到对于 β_1, β_2 的无偏的估计:

1. u 必须和 x_1, x_2, x_3, x_3^* 无关. (因为 x_3 只能通过 x_3^* 影响 y, 如果 x_3 直接影响 y, 就会和 u 有关了.) (多余性假设, x_3 不能带有 x_3^* 之外的额外信息.)
2. ν_3 必须和 x_1, x_2, x_3 无关. $\Rightarrow E(x_3^*|x_1, x_2, x_3) = E(x_3^*|x_3) = \delta_0 + \delta_3 x_3$. (好的代理变量假设, x_3 又可以充分表示 x_3^*, 也就是剩下的信息没有多少了.)

接着我们得到:

$$\begin{aligned} y &= (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + u + \beta_3\nu_3 \\ \Rightarrow y &= \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e. \end{aligned}$$

现在只要考虑 e 和 x_1, x_2, x_3 无关就可以了, 也就是 ν_3 和 u 与 x_1, x_2, x_3 无关. 根据上面两个假设就得到了.

Suppose that, the unobserved variable, x_3^*, is related to all of the observed variables:

$$x_3^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \nu_3.$$

那么就会得到:

$$y = (\beta_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)x_1 + (\beta_2 + \beta_3\delta_2)x_2 + \beta_3\delta_3 x_3 + u + \beta_3\nu_3$$

这样不满足好的变量假设就会导致有偏.

但是我们无法检验我们选取的代理变量是否满足我们的两个假设.

9.2.1 Using Lagged Dependent Variables as Proxy Variables

Lagged dependent variable: 滞后变量. 现实经济生活中, 许多经济变量不仅受同期因素的影响, 而且还与它自身的前期值有关.

例题 9.1 We want to estimate the effects of unemployment ($unem$) and law enforcement expenditures ($expend$) on crime:

$$crime = \beta_0 + \beta_1 unem + \beta_2 expand + \beta_3 crime_{-1} + u.$$

为什么要引入 $crime_{-1}$? 因为有着更高历史犯罪率的城市一般会在犯罪预防上花费更多, 这样就存在着一个 unobserved variable 在影响 $expand$, 所以必须通过一个 proxy variable 来代表这个变量, 这里使用 $crime_{-1}$ 就比较合适.

9.3 Measurement Error

测量误差. 和 Omitted Variable 的区别: 一个是测不准, 还有一个是没放进来.

Sometimes we have the variable we want, but we think it is measured with error. e.g. Reported annual income as a measure of actual annual income.

两种错误: 解释变量和被解释变量都有可能有测量误差.

数据来自于 Survey - 这些数据的准确性是未知的.

9.3.1 Measurement Error in the Dependent Variable

我们设真实模型为:

$$y^* = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u,$$

满足 Gauss-Markov 假设.

我们设:

$$e_0 = y - y^*.$$

那么:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u + e_0.$$

1. $E(e_0) \neq 0$, 可以把 e_0 的一部分加到 β_0 上面.
2. 如果 e_0 和所有的 x 无关, 那么就是无偏且一致的.
3. 如果 e_0 和 u 无关, 那么 $\mathrm{Var}(e_0 + u) = \sigma_u^2 + \sigma_{e_0}^2 > \sigma_u^2$. 造成了更大的方差.

如果是系统性的误差, 和一些解释变量有关系的话, 就会造成有偏的 OLS 估计. 相反, 如果是一些随机的误差, 那么 OLS 就是合适的.

9.3.2 Measurement Error in an Explanatory Variable

Consider the regression model

$$y = \beta_0 + \beta_1 x_1^* + u,$$

现在 x_1^* 是观测不到的, x_1 是观测值. 我们设 $e_1 = x_1 - x_1^*$. 不失一般性的, 我们可以假设 $E(e_1) = 0$. (可以都扔到常数项上.)

那么:

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

我们设一个多余性假设:

$$E(u|x_1) = E(u|x_1^*) = 0.$$

现在只需要讨论 e_1 和 x_1 之间的关系:

1. $\text{Cov}(e_1, x_1) = 0$. 这可以推出: $E(u - \beta_1 e_1) = 0$, $\text{Cov}(x_1, u - \beta_1 e_1)$=0. 不改变无偏性. 但是 $\text{Var}(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_{e_1}^2$. 减弱有效性.
2. $\text{Cov}(e_1, x_1) \neq 0$. 我们可以写成: $x_1 = x_1^* + e_1$.
   Classical errors-in-variable (CEV) assumption: $\text{Cov}(x_i^*, e_1) = 0$.
   $\text{Cov}(x_1, e_1) = \text{Cov}(x_1^*, e_1) + E(e_1^2) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2$.
   那么:

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

   首先:

$$\begin{aligned}
\text{plim } \hat{\beta}_1 &= \beta_1 + \frac{\text{Cov}(x_1, error)}{\text{Var}(x_1)} \\
&= \beta_1 + \frac{\text{Cov}(x_1, u - \beta_1 e_1)}{\text{Var}(x_1)} \\
&= \beta_1 - \frac{\beta_1 \sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \\
&= \beta_1 \cdot \frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}.
\end{aligned}$$

   β_1 is biased towards zero: attenuation bias(衰减偏误)
   Due to correlations between regressors, measurement error causes inconsistency in **all estimators**.

9.4 Missing Data, Nonrandom Samples, and Outliers

9.4.1 Missing Data

If the data are missing at random, then the size of the random sample is simply reduced. Although this makes the estimates less precise, it does not introduce any bias.

A problem can arise if the data is missing systematically - say high income individuals refuse to provide income data.

⇒ Nonrandom samples.

9.4.2 Nonrandom Samples

外生的样本选择 (exogenous sample selection), 无偏的:

$$saving = \beta_0 + \beta_1 income + \beta_2 age + \beta_3 size + u.$$

Consider a nonrandom sample of people over 35 years of age. 这实际上相当于一条线一下子把 35 左边的部分都抛弃了. 总体回归线是 $E(saving|income, age, size)$, 它在任何的一个子集下实际上是不变的.

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 expr + \beta_3 expr^2 + \beta_4 IQ + u.$$

Observed 概率和解释变量有关. 如果 IQ 比较高的话, 更容易报告, 这就导致了 IQ 高的人被观察到的概率高, IQ 低的人被观察到的概率低. 这现在不是一个子集的问题了, 而是一个概率随着 IQ 增加而不断增加的关系. 这个样本选择是一个概率性的. 这还是一个外生的样本选择. 还有一种是 IQ 被观测到的概率和 educ 有关, 依然是外生的, unbiased. 如果 IQ 被观测的概率是和 IQ 测试点和家的距离相关, 这并不会影响到 $wage$. 这实际上就是 missing at random.

我们设一个样本选择方案通过概率密度 $f(x)$ 来体现. 总结上面的情况是:

1. $f(x_1) = 0, \forall x_1 \in S$, $f(x_1) = 1, \forall x_1 \in U - S$.
2. $f(x_1)$ 和 x_1 有关.
3. $f(x_1)$ 和 x_2 有关.
4. $f(x_1)$ 和 x^f 有关, x^f 不是我们的解释变量和被解释变量, 而且 x^f 不影响 y.

这四种情况都将是无偏的, exogenous samples selection.

If the sample selection based on the dependent variables, this is called endogenous sample selection. e.g. IQ 观测的可能性和 EQ 有关, 而 EQ 影响了解释变量 wage.

Tobit model - 把东西补齐, 重新做回归.

A common method of data collection is *stratified sampling*. 分层抽样.

9.4.3 Outliers

OLS 受到 outliers 的影响有时候会比较大.

如何处理奇异值:

1. Drop the outliers.
2. Transform the data into functional forms that are less sensitive to outliers. Log forms are often used.
3. Use method that is less sensitive to outliers than OLS.

   LAD (Least absolute deviations).

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

   可以在 Stata 里使用`qreg`.

第十章 Instrumental Variables

10.1 Omitted Variables in a Simple Regression Model

如果说出现了 Omitted Variable Problem:

1. 直接不处理. 因为这个影响还是在的.
2. 代理变量. (Proxy variable) 找一个变量代理看不见的变量.
3. 面板数据. Panel data. 在不同的时间点观测这个人, 在时间维度上差分, 解决不随时间变化的东西.
4. 工具变量方法. Instrumental variables approach (IV).

定义 10.1 (工具变量)

为了得到 β_0 和 β_1 的一致估计, 可以使用工具变量. 工具变量满足下面两个假设:

1. $\mathrm{Cov}(z,u)=0$. Instrument exogeneity(工具外生性): z should have no partial effect on y (after controlling for x) and z should be uncorrelated with the omitted variable.
2. $\mathrm{Cov}(z,x)\neq 0$. Instrument relevance(工具相关性): z must be correlated, either positively or negatively, to the endogenous explanatory variable x.

满足上面的假设, 就是一个 valid IV, 可以得到 x 和 y 之间的一致性的估计. ♣

Birth quarter of children: Birth quarter is exogenous to ability but may be correlated with education outcomes. 一个很奇怪的工具变量.

10.2 Indentification with Instrument Variables

10.2.1 Method of Moments

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

那么:

$$\mathrm{Cov}(z,y)=\beta_1\mathrm{Cov}(z,x)+\mathrm{Cov}(z,u)\Rightarrow\beta_1=\frac{\mathrm{Cov}(z,y)}{\mathrm{Cov}(z,x)}.$$

如果是在样本的情况下, 得到:

$$\hat{\beta}_1=\frac{\sum_{i=1}^n(z_i-\bar{z})(y_i-\bar{y})}{\sum_{i=1}^n(z_i-\bar{z})(x_i-\bar{x})}. \tag{10.1}$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}.$$

10.2.2 Two Stage Least Square (TSLS,2SLS)

First stage: reg x on $z \rightarrow$ predict $\hat{x}$.

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

$$\hat{x}_i = \hat{\pi}_0 + \hat{\pi}_1 z_i.$$

Second stage: reg y on $\hat{x}$.

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

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

而注意到:

$$\hat{x}_i - \bar{x} = (\hat{\pi}_0 + \hat{\pi}_1 z_i) - (\hat{\pi}_0 + \hat{\pi}_1 \bar{z}) = \hat{\pi}_1 (z_i - \bar{z}).$$

把分子和分母里的替换掉一个 x 就得到了 MoM 的结果.

Stata 代码: `ivreg2 y (x=z)`.

10.2.3 Consisitency and Biasedness

IV estimator is consistent, 也就是 $\text{plim}(\hat{\beta}_1) = \beta_1$.

但是 IV estimator 不是无偏的.

如何计算 IV estimator 的方差:

我们假设 $E(u^2|z) = \sigma^2 = \text{Var}(u)$.

那么根据大数定理, 因为10.1里的每一项都有方差有一个上限, 那么就有这个是一致的.

同时, $\hat{\beta}_1$ 的渐进方差为:

$$\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}. \tag{10.2}$$

在10.2式子里 σ_x^2 是 x 的总体方差, $\rho_{x,z}^2$ 是 x 与 z 的总体相关系数的平方.

如果我们设第一阶段的 R^2 为 $R_{x,z}^2$, 那么 $\hat{\beta}_1$ 的渐进标准误为:

$$\frac{\hat{\sigma}^2}{SST_x R_{x,z}^2}. \tag{10.3}$$

如果说 x 和 z 非常的不相关, 那么 $R_{x,z}^2$ 就会很小, 方差就会很大.

10.3 IV Estimation of the Multiple Regression Model

定义 10.2 (Structural Equation)

我们可以把变量分为两组, 内生的和外生的. 比如:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1.$$

y_1 and y_2 are endogenous; z_1 is exogenous. ♣

我们现在再找一个外生的工具变量 z_2, 满足下面的条件:

$$E(u_1) = 0, \mathrm{Cov}(z_1, u_1) = 0, \mathrm{Cov}(z_2, u_1) = 0.$$

我们就可以得到下面的方程组并且来求解 β_s:

$$\sum_{i=1}^{n}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} z_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

First stage 包含了所有的外生变量, 也就是不仅是工具变量 z_2, 还需要原先的外生变量 z_1:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \nu_2, \pi_2 \neq 0.$$

比如 z_1 和 z_2 很像, 如果不加 z_1, 那这个就没有意义了, 所以必须要把结构模型也放进去.

10.4 Two Stage Least Squares

有时候对于同一个内生变量 y_2 我们有不同的工具变量.

考虑下面的结构方程 (Structural equation):

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1.$$

并且我们有两个外生的变量 z_2 和 z_3. 并且它们满足排他性约束 (Exclusion restrictions):

1. z_3 和 z_3 不出现在结构方程中.
2. z_2 和 z_3 与误差 u_1 不相关.

由于 z_1, z_2, z_3 和 u_1 都不相关, z_2 最好的工具变量是它们的线性组合:

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

同时我们不能让 y_2^* 和 z_1 完全相关, 所以我们有一个关键的识别假设: $\pi_2 \neq 0$, $\pi_3 \neq 0$.

注 我们可以通过 F 检验来检验是否有: $H_0 : \pi_2 = 0$ and $\pi_3 = 0$.

如果给定了 z_j 的数据, 我们就可以计算出 y_2^*:

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

接下来计算原来的结构方程中的 OLS 只需要:

$$\sum_{i=1}^{n}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} \hat{y}_{i2}^*(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

With multiple instruments, the IV estimator is also called the **two stage least squares (2SLS)** estimator.

1. 第一阶段: run the regression for the reduced form equation to obtain the fitted values $\hat{y}_2$.
2. 第二阶段: the OLS regression of y_1 on $\hat{y}_2$ and z_1.

10.4.1 The (Asymptotic) Variance of the 2SLS Estimator

2SLS 中对于 β_1 的方差的渐进估计可以是:

$$\frac{\sigma^2}{\widehat{SST}_2(1 - \widehat{R}_2^2)}.$$

这里面的 $\sigma^2 = \text{Var}(u_1)$, $\widehat{SST}_2$ 是 $\hat{y}_2$ 的总体变动, $\widehat{R}_2^2$ 是 $\hat{y}_2$ 对结构方程中的其他所有外生变量做回归得到的 R^2.

对于为什么 2SLS 的方差大于 OLS 的方差的解释:

1. $\hat{y}_2$ 比 y_2 有更小的总体变动.
2. $\hat{y}_2$ 和外生变量的线性关系比 y_2 和外生变量的线性关系更强.

10.4.2 Multiple Endogenous Explanatory Variables

2SLS 也可以被应用在有更多内生变量的方程中.

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + u_1.$$

Order condition for identification of an equation: 就是说我们需要找的外生的变量至少要和我们要替换的内生变量一样多.

10.5 Testing for Endogeneity

2SLS 比 OLS 更不有效, 会有更大的标准差、标准误. 因此, 检测是否有变量内生, 进而决定是否应该使用 2SLS 就很必要了.

除了下面的结构方程之外, 假设我们又有两个外生变量 z_3 和 z_4.

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1.$$

我们来检测 y_2 是否是内生的, 也就是 y_2 和 u_1 是否相关.

我们把 y_2 写成:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2.$$

由于前面的 z_j 和 u_1 都不相关, 所以我们只要检测 v_2 和 u_1 是否相关就可以了.

我们写成: $u_1 = \delta_1 v_2 + e_1$.

那么接下来就是检测: $\delta_1 = 0$.

那么我们就来检测:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + error.$$

然后来检测 $H_0 : \delta_1 = 0$, 通过 t 检验来判断.

Testing for endogeneity of a single explanatory variable:

1. 计算 y_2 在所有的外生变量上做回归, 并且得到 $\hat{v}_2$. (including those in the structural equation and the additional IVs.)
2. 把 $\hat{v}_2$ 加入原先的结构方程中 (which includes y_2), 检测它前面的系数.

第十一章 Pooled Cross-Sectional and Panel Data

Pooled Cross-Sectional and Panel Data 是跨越时间的, 一个变量随时间的变化.

11.1 Policy Analysis with Pooled Cross-Sectional Data

如何判断一个政策的影响, 政策颁布后效果如下:

	$D_i = 1$	$D_i = 0$
$y_{ti} + \Delta_i$	observable	unobservale
y_{ti}	unobservale	observale

y: Outcome measure.

$x_1, \cdots, x_k$: Variables that affect the outcome measure and the participation in the intervention.

D: Dummy variables, 1 for participants, 0 for non-participants.

T: 是干预之后的 period $t = 1$, 干预前的 period $t' = 0$.

11.1.1 Cross-Section Comparison

识别假设: The average value of the outcome measure of participants would have changed in the same way as the outcome measure of non-participants if the participants had not participated in the intervention.

$$E(y_t|x_1, \cdots, x_k, D = 1) = E(y_t|x_1, \cdots, x_k, D = 0).$$

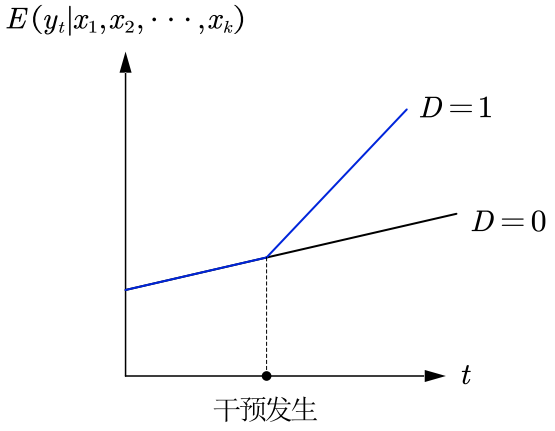

图 11.1: Identification Assumption in Cross-Section Comparison

Regression Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \alpha D_i + u_i.$$

Mean effect of treatment on the treated:

$$\hat{\alpha} = E(y_t + \Delta | x_1, \cdots, x_k, D = 1) - E(y_t | x_1, \cdots, x_k, D = 0).$$

只要 x 找的够准够多就可以识别.
但是会有自选择问题, 观察不到的异质性等.

11.1.2 Before-After Comparison

识别假设: The expected outcome measure of a participant with characteristics $x_1, x_2, \cdots, x_k$ at time t would have the same value as the expected outcome measure at time t' if the participant had not participated.

$$E(y_t | x_1, \cdots, x_k, D = 1) = E(y_{t'} | x_1, \cdots, x_k, D = 1).$$

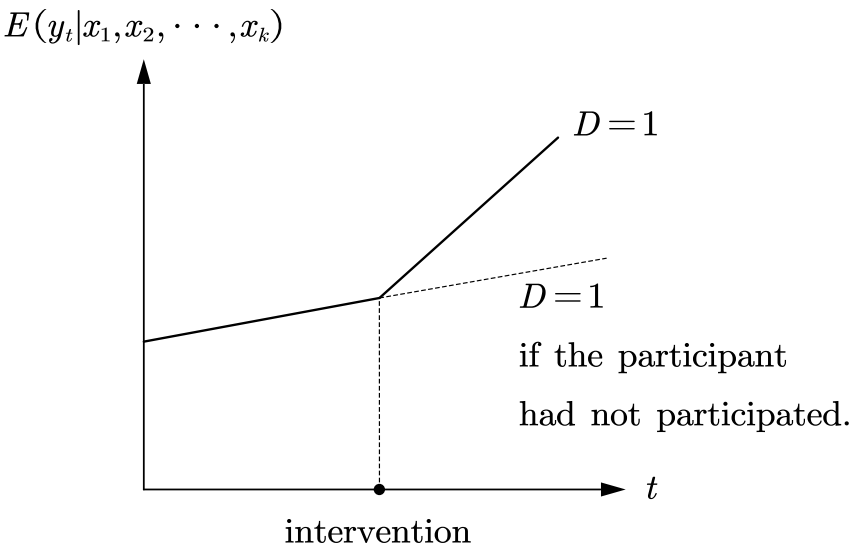

图 11.2: Identification Assumption in Before-After Comparison

Regression Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \gamma T_{it} + u_i.$$

Mean effect of treatment on the treated:

$$\hat{\gamma} = E(y_t + \Delta | x_1, \cdots, x_k, D = 1) - E(y_t | x_1, \cdots, x_k, D = 1).$$

Pooled cross-sectional or panel data are needed.
时间太长不行, 时间太短也不行. 因为有 Business cycle sensitivity.
Ashenfelter's Dip. 对 intervention 的预期会导致研究对象的行为的短期改变.

11.1.3 Difference-in-Differences (DiD) Estimation

识别假设: The difference in the outcome measure between the two groups would have remained constant over time if the participants had not participated in the intervention.

$$\begin{aligned}&E(y_t+\Delta|x_1,\cdots,x_k,D=1)-E(y_t|x_1,\cdots,x_k,D=1)\\=&E(y_t+\Delta-y_{t'}|x_1,\cdots,x_k,D=1)-E(y_t-y_{t'}|x_1,\cdots,x_k,D=0)\end{aligned}$$

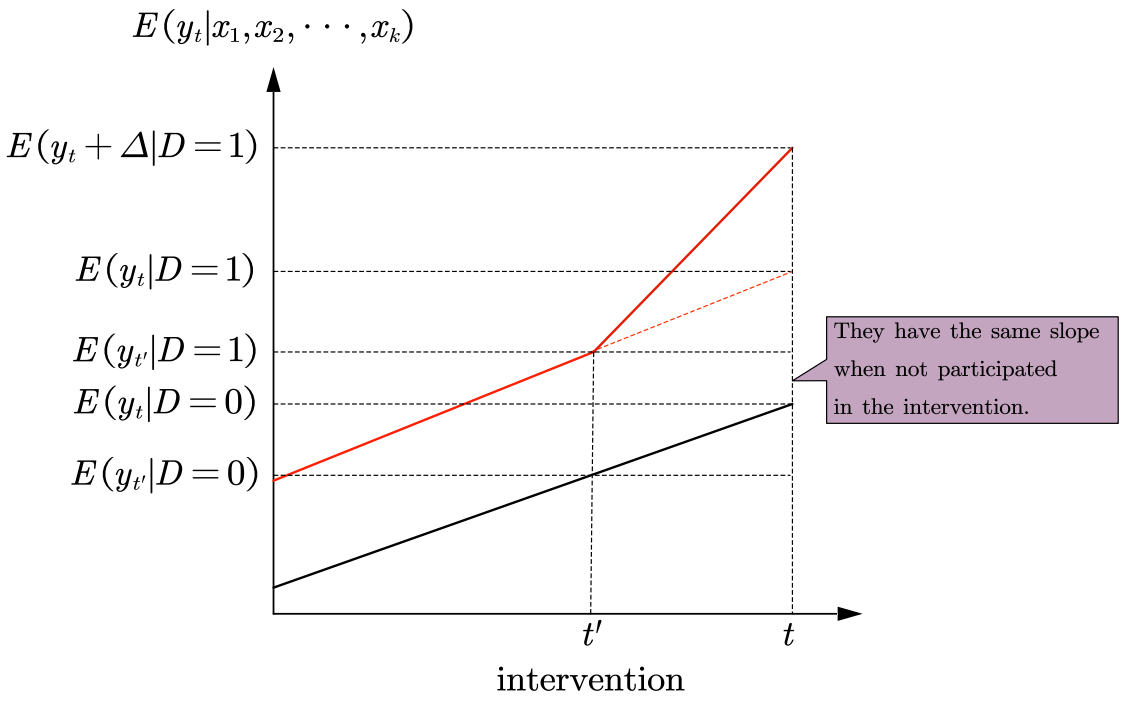

图 11.3: Identification Assumption in DiD Estimation

我们现在的 Regression Model 就是:

$$y_{it}=\beta_0+\beta_1x_{it1}+\beta_2x_{it2}+\cdots+\beta_kx_{itk}+\delta_1D_i+\delta_2T_{it}+\delta_3(D_i\cdot T_{it})+u_{it}.$$

Mean effect of treatment on the treated:

$$\begin{aligned}\hat{\delta}_3=&[E(y_t+\Delta|x_1,\cdots,x_k,D=1)-E(y_{t'}|x_1,\cdots,x_k,D=1)]\\&-[E(y_t|x_1,\cdots,x_k,D=0)-E(y_{t'}|x_1,\cdots,x_k,D=0)].\end{aligned}$$

如果说 $\hat{\delta}_3$ 显著的不等于 0, 那么说明 intervention 有影响.

11.1.4 Natural Experiment

Natural experiment (quasi-experiment) occurs when some exogenous event - often a change in government policy - changes the environment that individual, firm, city operates.

为了得到这个事件的影响, 我们需要:

1. Two years of data: one before change ($t=1$) and one after ($t=2$).

2. Two groups of people: one affected (T) by the change and one unaffected (C).

The average treatment effect can be estimated by:

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

11.2 Panel Data: Introduction

Panel Data(面板数据), or Longitudinal Data(纵向数据), 记录了一个个体的不同时间点的情况.

如果 n 个人在 T 个时间点的数据都被记录了, 总共有 nT 个数据, 这就叫做 balanced panel.

Two Period Panel Data Analysis:

假设: 只有两个时期而且每个人都在每个时期被记录了.

例题 11.1 Does unemployment cause crime? 我们有一个特别简单的:

$$\widehat{crmrte} = 128.38 - 4.16unem$$
$$(20.76)\,(3.42)$$
$$n = 46, R^2 = .033$$

这就很可能有遗失变量的问题, 但是不一定可以得到这些变量.

那么我们考虑下面的模型:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \delta_0 d2_t + \alpha_i + u_{it}.$$

$d2_t$ is the dummy variables for time period 2, which indicates things that vary over time, but not across individuals.

现在:

1. α_i 中蕴含着 things that vary across individuals, but not over time. $\alpha_i \to$ fixed effect, unobserved heterogeneity, unobserved individual effect
2. u_{it} 中蕴含着 things that vary across individuals and over time.

我们的这个模型被称为 fixed effects model/unobserved effect model.

panel data analysis 最重要的作用就是可以移除 α_i, 通过下面两个技术:

1. Differencing (First-differenced model)
2. Demeaning (Fixed effect model)

11.2.1 Differencing

For a cross-sectional observation i, write the two years as

$$y_{i2} = \beta_0 + \beta_1 x_{i2} + \delta_0 + \alpha_i + u_{i2}, (t = 2)$$
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + \alpha_i + u_{i1}, (t = 1)$$

把这两个式子相减:

$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1(x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

或者是

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

This is the first-differenced (FD) equation. The unobserved effect α_i 就被消除了.

接下来我们要对 FD equation 进行 OLS, 那么关键的假设就是:

1. $\text{Cov}(\Delta x_i, \Delta u_i) = 0$, 也就是 u_{it} 和解释变量在两个时期都无关.
2. Δx_i must have variation.

对 β_1 的解释不发生变化: the change in y for a one-unit change in x, holding the other variables in the model constant.

11.2.2 Demeaning

Fixed Effects (FE) Estimation:

我们依然从下面的模型出发:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \delta_0 d2_t + \alpha_i + u_{it}.$$

然后我们求出关于 t 的均值, 也就是 $\bar{y}_i = E_t(y_{it})$, 以此类推:

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + \delta_0 + \overline{d2} + \alpha_i + \bar{u}_i.$$

接着用原来的式子减去这个式子就得到:

$$(y_{it} - \bar{y}_i) = \beta_1(x_{it} - \bar{x}_i) + \delta_0(d2_t - \overline{d2}) + (u_{it} - \bar{u}_i).$$

If $T = 2$, first differencing and fixed effects produce **identical** coefficient estimates and standard errors.

Can not estimate any thing that is **constant over time** and any thing that **changes at a constant rate over time**.

If $T = 3$, FE is more efficient if the u_{it} are serially uncorrelated.

附录 A Stata Basics

A.1 A Hint of Wooldridge Datasets

Boston College Department of Economics 中列出了所有可能用到的数据集.

你可以直接在 Stata 中使用到这些数据集, 通过输入下面的命令:

```
bcuse affairs // 使用名为affairs的数据集
```

如果 Stata 提示你: `command bcuse is unrecognized`, 请先键入以下命令:

```
ssc install bcuse // 安装bcuse
```

A.2 Exercise Lesson 9.16

```
clean all // 清除
set more off // 自动显示一步程序所有结果
capture log close // 关掉之前日志文件
cd "Sth./YourDirectory" // 文件读取路径
. log using ass1.log, replace
-----------------------------------------------------------------------
      name: <unnamed>
       log: sth.\习题课9.16\ass1\ass1.log
  log type: text
 opened on: 24 Oct 2024, 09:45:38

. disp(exp(3)-20) // 显示当前的值
.08553692

. disp((exp(3)-20)>0)
1
```

```
use "sth\score.dta" // 打开一个dta文件
import excel ass2.xlsx, sheet("Sheet1") firstrow //导入表格数据, 第一行作变量名
save score.dta,replace
bysort 性别: sum 成绩 //按性别分组统计成绩
gen x=(成绩>=90)
log close
shellout ass2.log //导出日志
```

这里的sum x命令就是要求我们 summarize x这个变量的信息.

我们可以得到:

```
-> 性别 = 女

    Variable |      Obs       Mean    Std. dev.       Min        Max
-------------+--------------------------------------------------------
        成绩 |        3         88    10.14889        77         97
-> 性别 = 男

    Variable |      Obs       Mean    Std. dev.       Min        Max
-------------+--------------------------------------------------------
        成绩 |        4       85.5    5.322906        80         91
```

```
sysuse auto.dta // 使用系统自带的数据
hist price,density by (foreign) xtick(0(1000)15000) bin(8) note("数据来源于美国汽车协
    会") title("price直方图")
graph export ass3.png, replace //导出图片
```

最后画出的图像如下:

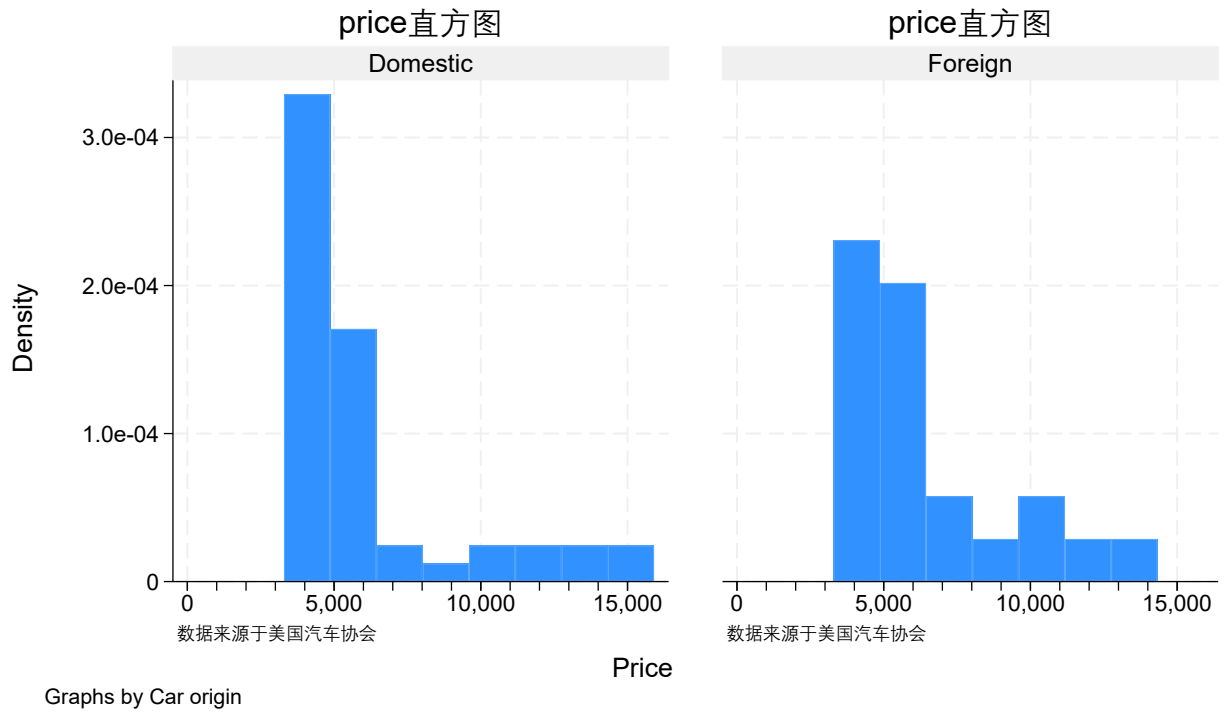

图 A.1: Stata 绘图实践

A.3 Assignment 1

在 Assignment 1 中的 Stata Exercises 部分要求我们进行一些 Stata 的实践, 结果如下.

A.3.1 Inverse regression

(1) Run the following codes in STATA.

```
set obs 100
gen z = rnormal ()
gen u1 = rnormal ()
gen u2 = rnormal ()
gen x = z + 0.4 * u1
gen y = z + 0.4 * u2
scatter y x
```

这份代码产生了标准正态分布 $z \sim N(0,1), u_1 \sim N(0,1), u_2 \sim N(0,1)$. x, y 是在 z 之上分别加上了一点东西, 所以不会相差很远, 应该都在 $y = x$ 附近. 在运行完代码后, 我们产生的图形如下:

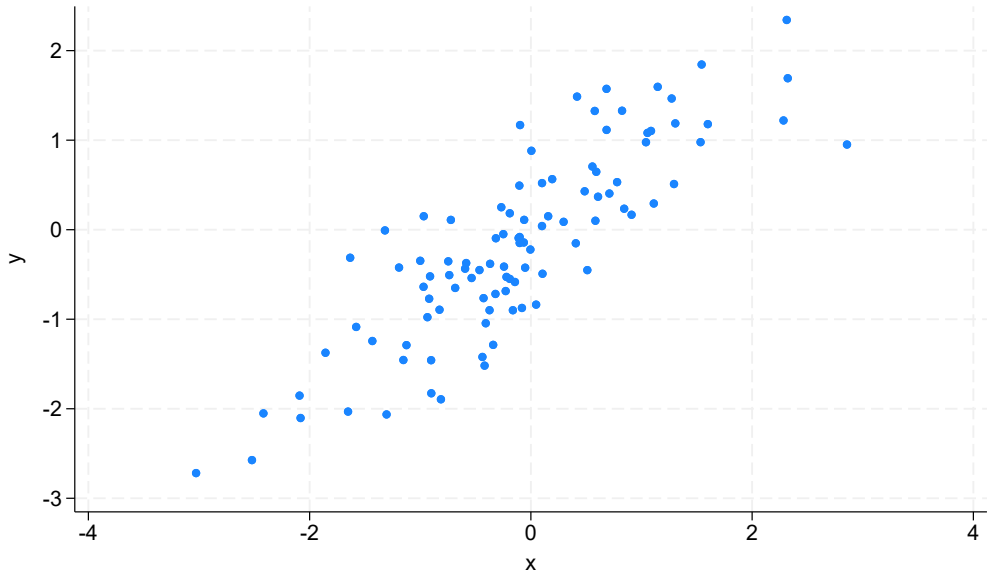

图 A.2: Result of `scatter` Command

(2) 在`reg y x`之后, 我们得到如图A.3的结果.
其中 F 检验的 p 值为 0.0000, 所以是显著的. 对于 β_1, t 检验的 p 值为 0.000, 是显著的; 对于 β_0, t 检验的 p 值为 0.177, 是不显著的.

(3) 在`reg x y`之后, 我们得到如图A.4的结果.
其中 F 检验的 p 值为 0.0000, 所以是显著的. 对于 β_1, t 检验的 p 值为 0.000, 是显著的; 对于 β_0, t 检验的 p 值为 0.576, 是不显著的.

```
9 . reg y x
```

Source	SS	df	MS	Number of obs	=	100
				F(1, 98)	=	244.03
Model	77.2880571	1	77.2880571	Prob > F	=	0.0000
Residual	31.0385033	98	.316719422	R-squared	=	0.7135
				Adj R-squared	=	0.7105
Total	108.32656	99	1.09420768	Root MSE	=	.56278

y	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
x	.8294396	.0530965	15.62	0.000	.7240713	.9348079
_cons	-.0769299	.0566214	-1.36	0.177	-.1892932	.0354334

图 A.3: Result of Regressing y on x

```
10 . reg x y
```

Source	SS	df	MS	Number of obs	=	100
				F(1, 98)	=	244.03
Model	80.1530931	1	80.1530931	Prob > F	=	0.0000
Residual	32.1890877	98	.328460078	R-squared	=	0.7135
				Adj R-squared	=	0.7105
Total	112.342181	99	1.1347695	Root MSE	=	.57311

x	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
y	.8601866	.0550648	15.62	0.000	.7509123	.9694608
_cons	.0325668	.0581088	0.56	0.576	-.0827482	.1478817

图 A.4: Result of Regressing x on y

(4) 这两个斜率是相近的. 这是因为 x 和 y 是两个相同的分布, 那么他们的样本两者调换顺序进行回归也应该差不多. 但是这两个斜率不是 consistent(一致) 的. 这是因为我们有 $y = x - 0.4u_1 + 0.4u_2$, 然后:

$$E(-0.4u_1 + 0.4u_2|x) = E(-0.4u_1|x) \neq 0,$$

不满足 $E(u|x) = 0$ 的假设.

A.3.2 The quality of a linear model

(1) 这份代码就是产生一个 $u \sim N(0,1)$ 附近的分布 y 和 $x \sim U(5,20)$.

```
set obs 500
gen x = 5+15*runiform()
gen u =rnormal()
gen y=100/x+u
```

(2) 在`reg y x`之后, 结果如图A.5.
这里的结果都是显著的, 同时 $R^2 = 0.7135$, 说明模型拟合的比较好, 数据的波动比较小.

(3) 不是的, 比如这题的常数项是显著的, 而前面的题目常数项是不显著的.

(4) Use x and u to generate z: $z_i = x_i^2 - 12x_i + u_i$.

```
5 . reg y x
```

Source	SS	df	MS	Number of obs	=	500
				F(1, 498)	=	2527.00
Model	6755.30479	1	6755.30479	Prob > F	=	0.0000
Residual	1331.28084	498	2.67325471	R-squared	=	0.8354
				Adj R-squared	=	0.8350
Total	8086.58563	499	16.2055824	Root MSE	=	1.635

y	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
x	-.8387882	.0166859	-50.27	0.000	-.8715717	-.8060047
_cons	19.77836	.2228421	88.76	0.000	19.34054	20.21619

图 A.5: Result of Regressing y on x

(5) 在`reg z x`之后, 结果为如图A.6.

```
7 . reg z x
```

Source	SS	df	MS	Number of obs	=	500
				F(1, 498)	=	5761.05
Model	1595653.96	1	1595653.96	Prob > F	=	0.0000
Residual	137932.386	498	276.972663	R-squared	=	0.9204
				Adj R-squared	=	0.9203
Total	1733586.34	499	3474.12093	Root MSE	=	16.642

z	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
x	12.89137	.1698432	75.90	0.000	12.55768	13.22507
_cons	-135.6691	2.268274	-59.81	0.000	-140.1257	-131.2125

图 A.6: Result of Regressing z on x

这题的结果说明 y 和 x 大体上是正相关的, 而 (2) 中的基本上是负相关的.

Now R^2 is larger than R^2 in (2)(which is 0.83), so the linear model explains more variation of y. However, neither the 2 models is accurate of unbiased, because z is not a linear function of x, neither.

A.4 Assignment 2

Are rent rates influenced by the student population in a college town? Let **rent** be the average monthly rent paid on rental units in a college town in the United States. Let **pop** denote the total city population, **avginc** the average city income, and **pctstu** the student population as a percentage of the total population. One model to test for a relationship is

$$\log(\text{rent}) = \beta_0 + \beta_1 \log(\text{pop}) + \beta_2 \log(\text{avginc}) + \beta_3 \text{pctstu} + u.$$

```
use RENTAL.dta
keep if year = 90 // 选出其中90年代的数据
sum pctstu, detail // summarize percent of student
```

```
reg lrent lpop lavginc pctstu // 进行多元回归
test pctstu = 0 // 检测pctstu前面的系数是否为0


. test pctstu = 0
( 1) pctstu = 0
  F( 1, 60) = 10.44
    Prob > F = 0.0020
```

A.5 常用的 Stata 命令

1. 生成新变量.

```
gen z = x1 + x2
```

2. 删除或者保留某一变量.

```
drop x1
keep x2
```

3. 删除或保留某些观测.

```
drop if x1 >= 2
keep if x2 != 1
```

4. 按某一变量分类作描述性统计.

```
by x2, sort:summarize x1
```

5. 进行简单运算.

```
display -(2+3)/sqrt(2*3)
dis 1/10
```

6. y 对 x 作回归, 并生成拟合值与残差.

```
reg y x
predict yhat,xb // xb的含义是线性预测拟合值
predict uhat,residual // residual的含义是残差
```

7. 两步法估计多元回归中 x_1 的系数.

```
reg x1 x2 x3
predict rhat,residual
reg y rhat
```

8. 查看回归系数及其协方差.

```
reg y x1 x2 x3
matrix list e(b) // 查看回归系数
matrix list e(V) // 查看协方差
```

9. 对多个回归系数作联合检验.

```
reg y x1 x2 x3 x4 x5
test x3=x4=x5
```

10. 求回归方程的标准化系数.

```
reg y x1 x2,beta
```

11. 画含拟合线的散点图.

```
twoway (scatter y x) (lfit y x)
```

附录 B 24-25 秋季期中考试试题

B.1 True or False

(Total mark 30', 3' for each question) Judge if the following statements are **True or False**, more importantly, **provide a brief explanation** for your answer in each case.

1. There is a trade-off (权衡) between unbiasedness and efficiency, for getting more efficient Ols estimator, we can sacrifice unbiasedness.
   Soln: False. 无偏性是计量估计中最基本的要求.
2. For establishing the unbiasedness of Ordinary Least Squares (Ols) estimators, we assume that the error term has the same variance given any value of the explanatory variable.
   Soln: False. 保证系数无偏只需要 MLR.1-4.
3. In an Ols regression model, adding an additional control variable will always decrease the variance of the Ols estimator for other coefficients.
   Soln: False. 当增加的控制量与现有变量高度相关时, 会引发多重共线性, 使得系数方差反而增大.
4. In the presence of heteroscedasticity (异方差), the Ols estimator is still the Best Linear Unbiased Estimator (BLUE).
   Soln: False. 异方差不会影响 OLS 估计量的无偏性, 但使得 OLS 估计量不再是最小方差估计量, 因此不再是 BLUE.
5. In an Ols regression model, if $Cov(u, X) = 0$, this condition alone ensures that the Ols estimator is unbiased.
   Soln: False. 虽然 $Cov(u, X) = 0$ 是无偏性的必要条件, 但它不足以单独保证无偏性. 无偏性要求零条件均值假设成立, 这意味着在任何给定的解释变量值下, 误差项的期望为 0.
6. In an Ols regression model, if an explanatory variable X is scaled by a constant factor of 10 (e.g. X is multiplied by 10), the new Ols estimate of the coefficient for X will be 10 times the original coefficient.
   Soln: False. 当解释变量 X 被乘以 10 时, 新的 OLS 估计系数会变为原系数的 1/10, 而不是 10 倍. 这是因为系数会反向调整, 以保持 $X \cdot \beta$ 项对因变量的实际影响不变. 因此, 缩放解释变量会导致系数的大小按相反的比例变化.
7. In an Ols regression model, removing the intercept term will not affect the Ols estimates of the slope coefficients.
   Soln: False. 移除截距项会影响 OLS 估计量, 因为模型会被强制通过原点.
8. In an Ols regression model, the residuals are always uncorrelated with each of the

explanatory variables.

Soln: True. 在 OLS 回归模型中, 残差与每个解释变量的样本值的协方差为零.

9. Taking the logarithm of an explanatory variable will change the R^2 of the regression model.

Soln: True. 对解释变量取对数会改变该变量对因变量的影响形式, 从而可能改变型的拟合优度 R^2.

10. Both the t-test and the F-test can be used to test the statistical significance of a single parameter in an Ols regression.

Soln: True. F 检验中待检验约束个数为 1 的特定形式可以检验单个参数的显著性.

B.2 Theoretical Deduction

(Total mark 25') Answer the following questions and write the procedure in detail.

(1) **Consider the saving function** (the relationship between household savings and income):

$$sav = \beta_0 + \beta_1 inc + u \text{ where } u = \sqrt{inc} \cdot e$$

e is a random variable with $E(e) = 0$ and $Var(e) = \sigma^2$. Assume e is independent of inc.

(a) (4') **Show** that the key zero conditional mean assumption (Assumption of Simple Linear Regression (SLR).4) is satisfied.

(b) (3') Is the **homoscedasticity** (同方差性) assumption (Assumption of SLR.5) violated? **Explain** your result.

(c) (4') Provide your economic intuition that supports the conclusion in (b).

Soln:

(a) $E(e|inc) = E(e) = 0, E(u|inc) = E(\sqrt{inc} \cdot e|inc) = \sqrt{inc}E(e|inc) = \sqrt{inc} \cdot 0$ (因为 $E(e|inc) = E(e) = 0$)

(b) Yes. $Var(e|inc) = \sigma_e^2$, $Var(u|inc) = Var(\sqrt{inc} \cdot e|inc) = inc \cdot Var(e|inc) = \sigma_e^2 inc$.

(c) 低收入家庭在储蓄方面没有太多自由裁量权, 他们的支出主要由必要的生活支出构成, 而高收入家庭在储蓄方面有自由裁量权 (既可能花很多钱消费也可能将收入的大部分用于储蓄), 因此储蓄变化的幅度会更大.

(2) **Consider a multiple linear regression model**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i.$$

(a) (4') Write down "**Partialling out**" interpretation of *Multiple Regerssion* and the derivation process.

(b) (4') **Prove** that: $E(\hat{\beta}_1) = \beta_1$.

(c) (6') Suppose that:

$$\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)=0.$$

Show that:

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y})}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}.$$

Soln:

(a) $\hat{\beta}_1=\frac{\sum_{i=1}^{n}\hat{r}_{i1}y_i}{\sum_{i=1}^{n}\hat{r}_{i1}^2}$. 证明过程如下:

$$x_{i1}=\hat{r}_1+\hat{r}_2x_{i2}+\hat{r}_{i1}=\hat{x}_{i1}+\hat{r}_{i1}.$$

代入主回归中, $\frac{\partial\sum_{i=1}^{n}\hat{u}_i^2}{\partial\hat{\beta}_1}=-2\sum_{i=1}^{n}(\hat{x}_{i1}+\hat{r}_{i1})(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2})=0.$
由于主回归残差与解释变量不相关, 而 $\hat{x}_{i1}$ 是 x_{i2} 的线性方程, 因此 $\sum_{i=1}^{n}\hat{x}_{i1}\hat{u}_i=0.$
得到 $\sum_{i=1}^{n}\hat{r}_{i1}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2})=0.$
辅助回归的残差和辅助回归的解释变量 x_{i2} 不相关, 同时 $\hat{\beta}_0$ 为常数, 残差和为 0, 得到
$\sum_{i=1}^{n}\hat{r}_{i1}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2})=\sum_{i=1}^{n}\hat{r}_{i1}(y_i-\hat{\beta}_1x_{i1})=0.$
代入 $x_{i1}=\hat{x}_{i1}+\hat{r}_{i1}$, 由于 $\sum_{i=1}^{n}\hat{x}_{i1}\hat{r}_{i1}=0$, 则: $\sum_{i=1}^{n}\hat{r}_{i1}(y_i-\hat{\beta}_1\hat{r}_{i1})=0.$
得到: $\hat{\beta}_1=\frac{\sum_{i=1}^{n}\hat{r}_{i1}y_i}{\sum_{i=1}^{n}\hat{r}_{i1}^2}$.

(b) $\hat{\beta}_1=\frac{\sum_{i=1}^{n}\hat{r}_{i1}y_i}{\sum_{i=1}^{n}\hat{r}_{i1}^2}=\frac{\sum_{i=1}^{n}\hat{r}_{i1}(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+u_i)}{\sum_{i=1}^{n}\hat{r}_{i1}^2}$.
由于残差 u_i 的期望是 0, 所以 $E(\hat{\beta}_1)=\frac{\sum_{i=1}^{n}\hat{r}_{i1}(\beta_0+\beta_1x_{i1}+\beta_2x_{i2})}{\sum_{i=1}^{n}\hat{r}_{i1}^2}$.
由于辅助回归的残差和辅助回归的解释变量 x_{i2} 不相关, 同时 $\hat{\beta}_0$ 为常数, 残差和为 0,
所以 $\sum_{i=1}^{n}\hat{r}_{i1}=\sum_{i=1}^{n}\hat{r}_{i1}x_{i2}=0.$
又由于 $x_{i1}=\hat{r}_{i1}+\hat{x}_{i1}$, 而 $\sum_{i=1}^{n}\hat{x}_{i1}\hat{u}_i=0,$
所以 $\sum_{i=1}^{n}\hat{r}_{i1}x_{i1}=\sum_{i=1}^{n}\hat{r}_{i1}(x_{i1}+\hat{r}_{i1})=\sum_{i=1}^{n}\hat{r}_{i1}^2.$
因此 $E(\hat{\beta}_1)=\frac{\sum_{i=1}^{n}\hat{r}_{i1}(\beta_0+\beta_1x_{i1}+\beta_2x_{i2})}{\sum_{i=1}^{n}\hat{r}_{i1}^2}=\frac{\beta_1\sum_{i=1}^{n}\hat{r}_{i1}^2}{\sum_{i=1}^{n}\hat{r}_{i1}^2}=\beta_1.$

(c)

$$\begin{aligned}
&-2\sum_{i=1}^{n}x_{i1}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2})\\
&=-2\sum_{i=1}^{n}x_{i1}[(y_i-\bar{y})-\hat{\beta}_1(x_{i1}-\bar{x}_1)-\hat{\beta}_2(x_{i2}-\bar{x}_2)]\\
&=-2\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)[(y_i-\bar{y})-\hat{\beta}_1(x_{i1}-\bar{x}_1)-\hat{\beta}_2(x_{i2}-\bar{x}_2)]=0.
\end{aligned}$$

整理得 $\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y})-\hat{\beta}_1\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2-\hat{\beta}_2\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_{i2})=0.$
结合条件, 有 $\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y})-\hat{\beta}_1\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2=0.$
因此 $\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y})}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$.

B.3 Application 1

(Total mark 15') In recent years, the role of family size has drawn attention regarding its potential effects on individual's economic outcomes (家庭规模对个体经济结果的影响). A key perspective suggests that reducing the number of children may enhance the quality of resources available to each child, thus influencing individual socioeconomic success. *Lu* aims to investigate **the relationship between individual income and sibling count** (兄弟姐妹数), utilizing survey data from the China Family Panel Studies (CFPS).

The variables included in her analysis are as follows:

income: the annual personal income of individual i (measured in yuan).

sibling: the number of siblings of individual i.

```
      Source |       SS         df         MS        Number of obs   =    24,783
-------------+------------------------------        F(  2,   497)   =    255.97
       Model |  669.884797         1   669.884797     Prob > F        =    0.0000
    Residual |  64852.0969    24,781   2.61700887     R-squared       =    0.0102
-------------+------------------------------        Adj R-squared   =    0.0102
       Total |  65521.9817    24,782   2.64393437     Root MSE        =    1.6177

------------------------------------------------------------------------------
     lnincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
      sibling |  -.0860677   .0053795     [    ]   0.000    -.0966119   -.0755236
        _cons |   8.870594   .1861844    476.44   0.000     8.834101    8.907087
------------------------------------------------------------------------------
```

图 B.1: **Figure 1**

(1) (2') Write the Stata codes to **generate the variable "lnincome"** and **run the regression** to obtain the estimated results for **Figure 1**, where "*lnincome*" represents the natural logarithm of individual income.

Soln: `gen lnincome = log(income)`

`reg lnincome sibling`

(2) (4') Using the information provided in **Figure 1**, **calculate** the concealed (被遮住) ***t*-value** in the figure, rounding to two decimal places (保留两位小数) and justify its statistical and economic (magnitude of the coefficient) significance.

Soln: 由公式 $t = \frac{\hat{\beta}}{se(\hat{\beta})}$, 计算得 $t = \frac{-0.0860677}{0.0053795} = -16.00$, 通过表格中的 p 值可以发现统计意义显著, 但是由于系数不是很大, 我们可以认为它在经济意义上未必显著.

(3) (1') Now, consider a multivariable regression where "*gender*" represents individual gender (male = 1, female = 0) and "*age*" represents the individual's age in years. Please **write a one-line regression code** to obtain the results shown in **Figure 2**.

Soln: `reg lnincome sibling gender age`

(4) (2') Based on **Figure 2**, please **specify** the **mean** of *lnincome* for the group of 30-year-old women with no siblings, rounding to two decimal places.

Soln: 当 *gender* 和 *sibling* 为 0, *age* 是 30 时, *lnincome* 的均值是 $9.331 - 30 \times 0.022 =$

8.671 ≈ 8.67.

```
      Source |       SS       df       MS              Number of obs   =    24,783
-------------+------------------------------           F(  2,   497)   =    936.08
       Model |   669.81946         3   2223.27315       Prob > F        =    0.0000
    Residual |  58852.1622    24,779   2.37508222       R-squared       =    0.1018
-------------+------------------------------           Adj R-squared   =    0.1817
       Total |  65521.9817    24,782   2.64393437       Root MSE        =    1.5411

------------------------------------------------------------------------------
     lnincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
      sibling |    -.23926   .0053763    -4.45   0.000    -.034464   -.0133881
       gender |   .7585894   .0196741    38.56   0.000     .7200269    .7971519
          age |  -.0229479   .0006786   -33.82   0.000     -.24278   -.0216178
        _cons |   9.331861   .0333758   279.60   0.000    9.266442     9.29728
------------------------------------------------------------------------------
```

图 B.2: **Figure 2**

(5) (6') Lu employs the following code to generate the new variable *lnincomek*:

```
gen incomek = income/1000
gen lnincomek = log(incomek)
```

He regresses *lnincomek* on *sibling*, *gender*, and *age*. Can you identify the coefficients for these variables and explain whether they can or cannot be interpreted meaningfully (providing your reasoning)?

Soln: 相关系数与 **Figure 2** 中模型相同.

lnincomek 的值等于 *lnincome* 减去一个常数 ($log(1000000)$), 新的模型衡量其他条件不变时, *sibling*, *gender*, *age* 三个变量分别变动 1 个单位时, *lnincomek* 的变化量和原模型中的 *lnincome* 的变化量是相同的, 因此模型的相关系数与 **Figure 2** 中的模型相同.

B.4 Application 2

题干未看清.

ln($PM2.5$): 未看清.

exp: Professional experience (years).

*exp*2: Square of professional experience (years).

practice: Average daily training duration (hours) in the week preceding matches.

temp: Outdoor temperature at the training facility (°C).

prestige: Team prestige score (0-100 points, based on historical performance, fan base size, etc., with higher scores indicating greater prestige).

Estimation results are presented in the table below:

VARIABLES	(1)	(2)	(3)
ln(PM2.5)	-3.186	-3.842	-3.526
	(0.421)	(0.389)	(0.424)
exp	2.400	2.356	2.383
	(0.450)	(0.448)	(0.445)
exp2	-0.200	-0.196	-0.198
	(0.068)	(0.066)	(0.065)
temp	-0.240	-0.285	-0.282
	(0.012)	(0.092)	(0.091)
prestige		0.224	0.226
		(0.043)	(0.043)
practice			2.845
			(0.518)
Constant	82.643	75.924	73.156
	(2.152)	(2.363)	(2.425)
Observation	2000	2000	2000
R2	0.315	0.392	0.428

表 B.1: Standard errors are in parentheses.

(1) (5') Using the results from **Column (1)**, derive the *marginal effect* of experience on player performance. Next, calculate the *optimal years* of experience (最优经验年) based on these results in **Column (1)**. Please provide an *intuitive explanation* for why an optimal level may exist in the context of professional esports.

Soln: exp 的边际效应为 $2 \times (-0.2) \times exp + 2.4 = 2.4 - 0.4exp$, 当边际效应为 0 的时候, 选手成绩最佳, 此时最优的 exp 为 6 年. 适度的训练可以让选手在职业比赛中更熟练, 然而过度的训练会使选手思维僵化, 因此成绩降低.

(2) (5') Based on results from **Column (1)**, *interpret* the coefficient of In(PM2.5). Below are three potential scatter plots showing the relationship between performance and the absolute value of PM2.5. *Determine* which scatterplot below (Figures A-C) best aligns with the model specification in this study. Provide your *explanation*.

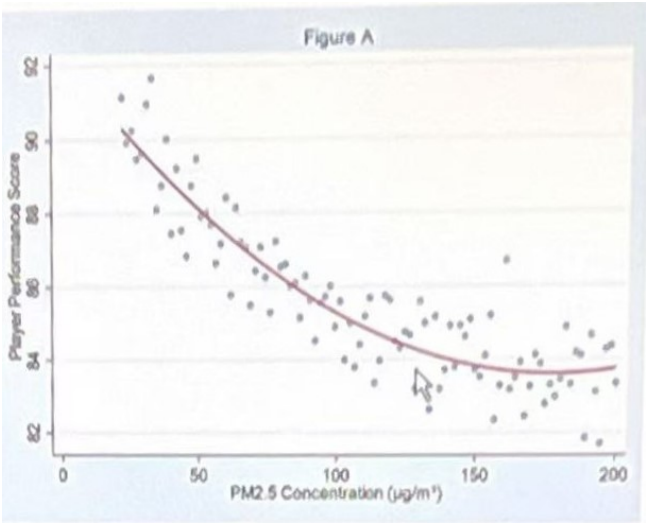

图 B.3: Figure A

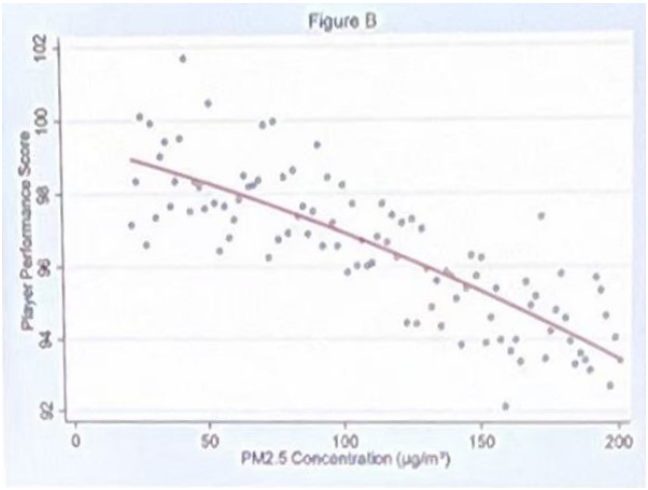

图 B.4: Figure B

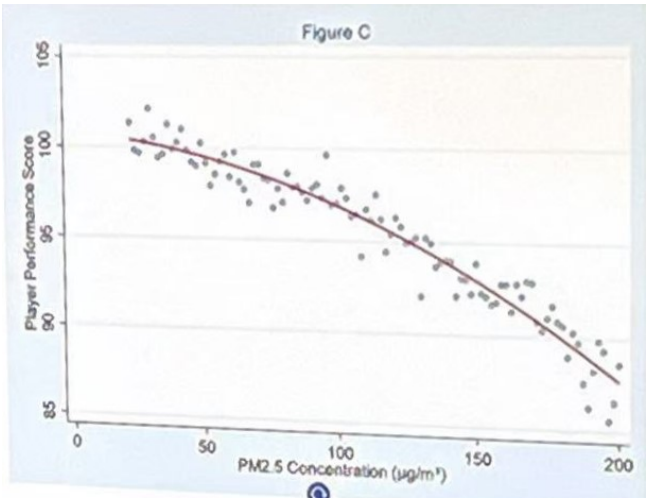

图 B.5: Figure C

Soln: $\ln(PM2.5)$ 的系数-3.186 表明训练地城市的 PM2.5 水平每上升 1%, 选手的成绩平均下降 0.03186 分. 图 A, 选手的成绩和 $\ln(PM2.5)$ 呈线性关系, 因此是关于 PM2.5 水平的凸函数, 只有 A 图符合要求.

(3) In **Column (2)**, we observe that after including the variable for team prestige, the coefficient of $\ln(PM2.5)$ becomes *more negative* compared to that in Column (1).

(a) (3') Are higher-prestige esports clubs **more likely to be located in areas with higher or lower pollution levels**? Provide an *econometric explanation* for your inference.

(b) (2') If an omitted variable, **such as sponsorship advertising revenue**(赞助商广告收入), is **only related to esports event performance and not related to pollution**, how would the *standard error* of the estimated cocfficient of $\ln(PM2.5)$ change, and *why*?

(c) (3') Beyond team prestige, what *other variables* that Song may have omitted in her study could bias the estimated coefficient of $\ln(PM2.5)$? Provide an *example* and explain the *potential direction* of the bias.

Soln:

(a). 假设遗漏变量 prestige 时 $\ln(PM2.5)$ 的相关系数为 β_1, 加入 prestige 后 $\ln(PM2.5)$ 的相关系数为 β_2, 成绩与 prestige 的相关系数为 δ, $\ln(PM2.5)$ 和 prestige 的相关系数为 α, 那么 $\beta_1 = \beta_2 + \delta\alpha$, 从题目中信息中可以得出 $\beta_1 > \beta_2$, $\delta > 0$, 代入可得 $\alpha > 0$, 即声望更高的队伍坐落于污染更严重的城市.

(b). 当变量被遗漏时, 系数的标准误将增加, 这是因为遗漏的变量被归为误差项中, 增大了误差项的方差.

(c). 队伍所在城市的 GDP. GDP 越高的城市能给选手提供更好地训练条件, 因此 GDP 与成绩呈正相关关系, 而 GDP 越高的城市工业越发达, 污染水平越高, 因此 GDP 和 $\ln(PM2.5)$ 呈正相关关系. 假设遗漏变量 GDP 时 $\ln(PM2.5)$ 的相关系数为 β_1, 加入 GDP 后 $\ln(PM2.5)$ 的相关系数为 β_2, 成绩与 GDP 的相关系数为 δ, $\ln(PM2.5)$ 和 GDP 的相关系数次 α, 那么 $\beta_1 = \beta_2 + \delta\alpha$, 由于 δ 和 α 大于 0, β_1 大于 β_2, 忽略 GDP 会高估 $\ln(PM2.5)$ 的相关系数.

(4) Based on results from Column (2)-(3), Song wants to perform *a joint test of the effect of prestige and training hours:*

(a). (2') State the *null hypothesis* for this joint test.

(b). (5') Using the information provided in the table, explain how to *construct the F-statistic* to test the joint effect. Write down the *formula* with all required components.

(c). (5') *Calculate* the F-value, *identify the appropriate critical value* for testing joint significance, and conduct the test *at 5% significance level* given the critical value you identified from *the following options*:

Critical F - values:

$F(2, 1000, 0.05) = 3.02$ $F(2, 1993, 0.05) = 3.00$ $F(2, 2000, 0.05) = 2.99$

$F(3, 2000, 0.05) = 2.61$ $F(3, 2000, 0.025) = 3.01$ $F(3, 1994, 0.025) = 3.02$

$F(3, 1993, 0.025) = 3.02$

Soln:

(a). H_0: prestige 和 practice 的相关系数均为 0.

(b). $F = \frac{(SSR_1 - SSR_2)/q}{SSR_2/(n-k-1)}$, F 统计量服从分布 $F(q, n-k-1)$. 其中 $q = 2, n = 2000, k = 6$, SSR_1 是 H_0 成立时的多元回归模型 (即 Column (1)) 的残差平方和, SSR_2 是 H_0 不成立时的多元回归模型 (即 Column(3)) 的残差平方和. 给定总样本 SST 不变, 我们只需求出 SSR_1 和 SSR_2 的比值便可求得 F.

(c). 给定总样本 SST 不变, 由 Column (2) 和 Column (3) 所示, SSR_1 为 SST 的 $1-R^2$ 倍即 0.685 倍, SSR_2 为 SST 的 0.572 倍, 因此 $SSR_1/SSR_2 = 0.685/0.572 = 1.198$, 因此 $F = \frac{0.198/2}{1/1993}$, 即 197.3. 由于 F 服从 $F(2, 1993)$ 分布, 所以选择的临界值 $F(2, 1993, 0.05) = 3.00$, 由于 197.3 大于 3.00, 因此我们拒绝 H_0.

(5) (5') Propose an *ideal experiment* (理想实验) design to identify the causal impact of air pollution on player performance. [Hint: Your proposal should include the following elements: (a) list other factors that could bias the estimation of pollution's effect and explain how your experimental design mitigates these biases; (b) detail the experimental framework and randomization method; and (c) discuss the feasibility of your experiment, addressing potential implementation challenges.]

Soln: 空气污染质量和经济发展水平、工业产值之间有很强的关联, 而在富裕和工业发达的城市, 电竞队能够得到更多比赛资源, 获得更好成绩. 因此我们需要通过理想实验剔除其他变量影响, 确认污染与成绩之间的相关性. 我们可以将抽样范围缩小至 GDP 和工业产值相似的城市群, 在这些城市中以成绩为因变量, 以污染水平为自变量作 OLS 回归, 探讨空气污染质量与选手成绩的因果关系. 如果要进行这个实验, 我们需要在一组 GDP、工业产值相似的城市中找到足够多的竞赛队样本以进行回归分析.