- **接下来，我们将讨论横截面数据的回归分析**
- **第2章中，我们介绍用一个变量去解释另一个变量的简单线性回归模型**

简单线性回归在应用计量经济学中使用并不广泛，但其涉及的代数知识及其解释都相对简单明了

- **第3~4章中，我们将讨论多元回归分析：容许一个以上的变量影响我们所要解释的变量。**

第3章介绍了模型估计方法，第4章的重点则是讨论统计推断。

① $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ $\leftarrow$ $u \sim x$ $\mathrm{cov}(u, x) \neq 0$

截距常数

$E[u|x] \neq 0$.

② $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{u}$ $\leftarrow$ ✓ $E[u|x] = 0$.

抽样，随机，AGE，书附录。

$$\begin{cases} \frac{1}{n}\sum \hat{u}_i = 0 \\ \frac{1}{n}\sum x_i \hat{u}_i = 0 \end{cases}$$

$\frac{1}{n}\sum \hat{u}_i^2 = \hat{\sigma}^2 \to \sigma^2$

u

$y = \beta x + \varepsilon$, ε 均值为0，σ^2.

$\beta^* = \frac{\sum y}{\sum x}$ 估计 β. 问估计好坏?

$= \frac{\sum(\beta x + \varepsilon)}{\sum x} = \beta + \boxed{\frac{\sum \varepsilon}{\sum x}}$

ε 大部 > 0,

< 0,

Chapter 2 简单回归模型

章节框架

- 介绍简单回归模型的定义
- 推导模型的估计方法：普通最小二乘法（OLS）
- 讨论OLS估计值的代数性质
- 介绍度量单位与函数形式并讨论他们对OLS的影响
- 推导OLS的统计性质

简单线性回归模型的定义

- **简单线性回归模型的定义**

“用变量 x 来解释变量 y ”或者“研究y 如何随 x 变化而变化”

截距项（ Intercept ）

斜率参数（ Slope parameter ）

features

$$\frac{\partial y}{\partial x} = \beta_1 + \cancel{\frac{\partial u}{\partial x}} \qquad y = \beta_0 + \boxed{\beta_1} x + u \qquad u(x)$$

显著的?

显著不为0?

因变量 Dependent variable,
被解释变量 explained variable,
响应变量 response variable,…

regressand

自变量 Independent variable,
解释变量 explanatory variable,
回归元 regressor,…

误差项 Error term,
干扰项 disturbance,
观测不到的因素 unobservables,…

简单线性回归模型的定义

- **简单线性回归模型的解释**

“研究y如何随着x的变化而变化:”

$$\frac{\Delta y}{\Delta x} = \beta_1 \quad \text{as long as} \quad \boxed{\frac{\Delta u}{\Delta x} = 0}$$

如果自变量增加一个单位，
因变量会变化多少？

只有当自变量增加一个单位时，
所有其他条件保持相等时，解释
才正确

- **简单线性回归模型在实践中很少适用，但出于教学方面的原因，它的讨论很有用**

简单线性回归模型的定义

- **例子：大豆收成与施肥量**

$$yield = \beta_0 + \beta_1 fertilizer + u$$

降雨量、土地质量、寄生虫的存在…

在保持所有其他因素不变的情况下，测量肥料对产量的影响

Griliches (1976)

- **例子：一个简单的工资方程**

$$wage = \beta_0 + \beta_1 educ + u$$

劳动力经验、在现任雇主任职期间、职业道德、智慧…

在所有其他因素不变的情况下，衡量额外一年受教育后小时工资的变化

简单线性回归模型的定义

- 什么时候有因果解释？ 4. $E(E(u|x)) = E(u) = 0$
- 当x变化时，需要其他条件不变：保持所有其他因素（u中）不变
- 需要对无法观测的u与解释变量x之间的关系加以约束
- 首先，可以假设 $\mathrm{E}(u) = 0$
- 此外，还需要u的平均值与x无关
- 条件均值独立性假设

1. 核心假设
2. 不能检验
3. $E(u|x) = c$

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

$y = \underbrace{(\beta_0 + c)}_{\beta_0} + \beta_1 x + \underbrace{(u - c)}_{\tilde{u}}$

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

有 $$E(u|x) = 0$$

解释变量不得包含有关未观察因素平均值的信息

$\Rightarrow$ 5. $cov(u, x) = 0$ $\nRightarrow$

$cov(u,x) = E(ux) - E(u)E(x)$, $E(ux) = 0$, $E(u) = 0$

$= E(E(ux|x))$

$= E(x \cdot E(u|x))$

- 还称为u的均值独立于x

简单线性回归模型的定义

- **例子：大豆收成与施肥量**

$$yield = \beta_0 + \beta_1 fertilizer + u$$

Rainfall,
land quality,
presence of parasites, …

如果该地区的其他条件跟施肥量没有关系，条件均值独立性假设就可以成立

- **例子：工资等式**

$$wage = \beta_0 + \beta_1 educ + u$$

例如：智慧…

条件均值独立性假设不太可能成立，因为受过更多教育的人平均也会更聪明。

$E(u|x)=0$ X

简单线性回归模型的定义

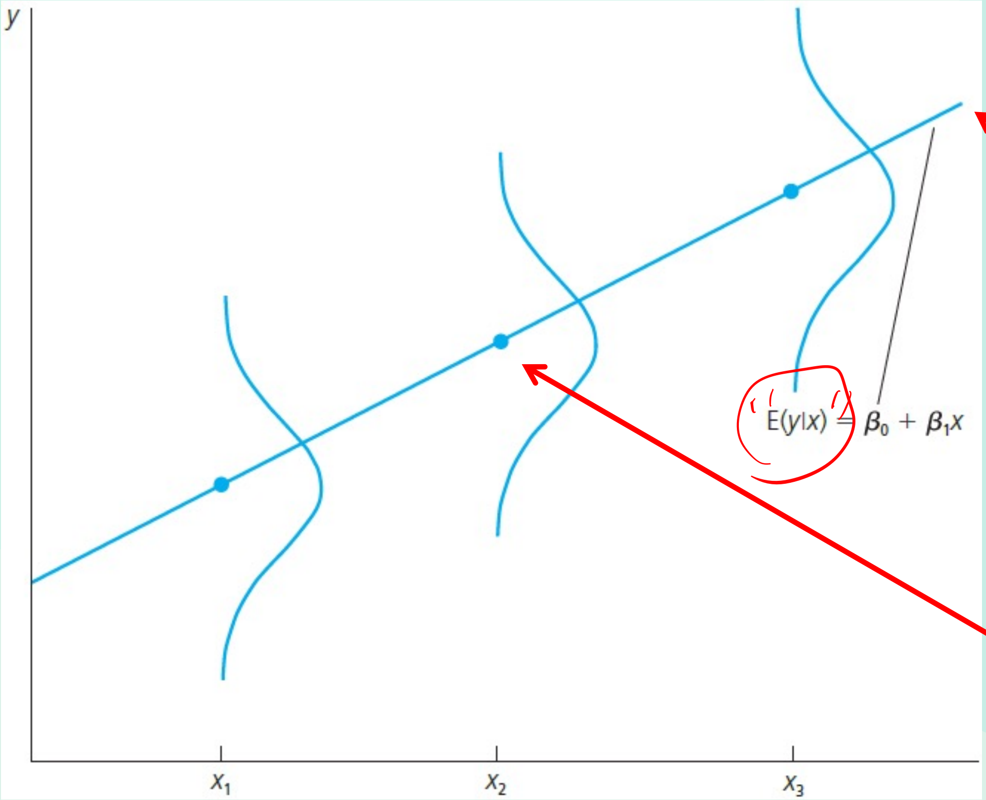
- **总体回归函数(Population regression function, PFR)**
  - 条件均值独立性假设意味着

$$
\begin{aligned}
E(y|x) &= E(\beta_0 + \beta_1 x + u|x) \\
&= \beta_0 + \beta_1 x + E(u|x) \\
&= \beta_0 + \beta_1 x
\end{aligned}
$$

=0

  - 这意味着$E(y|x)$可以被表达成一个自变量的线性函数
  - 对于任意给定的x， y的分布都以$E(y|x)$为中心

简单线性回归模型的定义



总体回归函数

对于个体 $x = x_2$, y 的平均值是 $\beta_0 + \beta_1 x_2$

普通最小二乘法的推导

- 推导普通最小二乘估计
- 需要数据来进行估计
- 假设一个随机样本有 n 个观测（observations）

$(x_1, y_1) \longleftarrow$ First observation

$(x_2, y_2) \longleftarrow$ Second observation

$(x_3, y_3) \longleftarrow$ Third observation

$\vdots$

$(x_n, y_n) \longleftarrow$ n-th observation

$$\{(x_i, y_i) : i = 1, \dots, n\}$$

第i个观测下的自变量取值

第i个观测下的因变量取值

$x=3,\ y=10$

$\overset{10}{y} = \beta_0 + \beta_1 \overset{3}{x}$

① $\beta_0 + \beta_1 3 = 10$ 无穷解

② $\begin{cases} x_1=3,\ y_1=10 \\ x_2=4,\ y_2=8 \end{cases} \to \beta_0, \beta_1$ 唯一解

③ 3个以上？

$$\begin{cases} \beta_0 + \beta_1 x_1 = y_1 \\ \vdots \\ \beta_0 + \beta_1 x_n = y_n \end{cases}$$

普通最小二乘法的推导

- “尽可能好(as good as possible)”是什么意思？
- 回归残差(residuals)

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

- 最小化回归残差的平方和

$$\min \sum_{i=1}^{n} \hat{u}_i^2 \quad \rightarrow \quad \hat{\beta}_0, \hat{\beta}_1$$

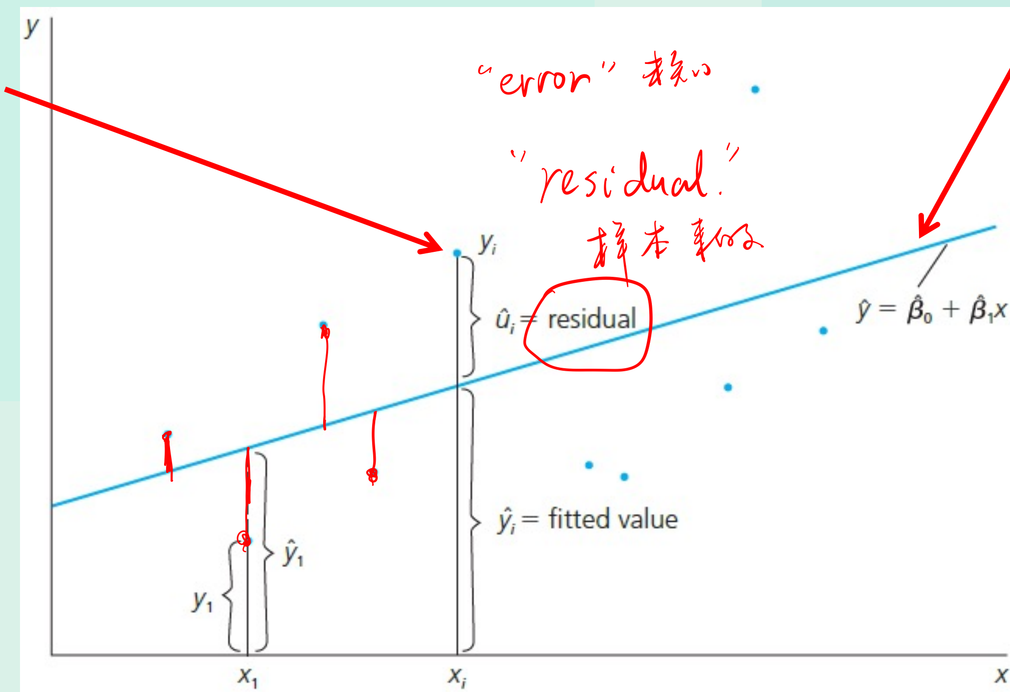
- 为什么不是最小化残差的其他某个函数，比如残差绝对值?

普通最小二乘法的推导

- **尽可能通过数据点拟合回归线：**

例如，第i个数据点
(x_i, y_i)

拟合回归线



普通最小二乘法的推导

$x^{(1)}$ $x^{(2)}$

- **普通最小二乘(Ordinary Least Squares, OLS)估计**

u σ^2 $\hat{u}_i^2$

$$min \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$\hat{\beta}_0$和$\hat{\beta}_1$满足:

$E(u) \overset{?}{=} 0$ $E(xu) = 0$

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

$x^{(2)} = 10 \cdot x^{(1)}$

$\frac{1}{n}\sum \hat{u}_i = 0 \Rightarrow E(u)=0$ $\frac{1}{n}\sum x_i \hat{u}_i = 0 \Rightarrow E(xu)=0$

“没有多重共线性”

因此:只要有$\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$

$\sum x_i(x_i-\bar{x}) = \sum(x_i-\bar{x})(x_i-\bar{x})$

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

$\sum x_i(y_i-\bar{y}) = \sum(x_i-\bar{x})(y_i-\bar{y})$

Frisch–Waugh Theorem

普通最小二乘法的推导

- **等价的，我们可以从下列两个条件得到$\hat{\beta}_0$和$\hat{\beta}_1$：**

$$E(u)=0\,,\quad E(xu)=0$$

$$\Longrightarrow E(y-\beta_0-\beta_1 x)=0\,,\qquad E(x(y-\beta_0-\beta_1 x))=0$$

"矩条件"

样本对应条件：

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

广义矩

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

Generalized Methods of Moments "GMM"

Hayashi (2000)

普通最小二乘法的推导

- **CEO 工资与股本回报率**

$\sum \hat{u} \neq 0$ X

$$salary = \beta_0 + \beta_1 roe + u$$

$\boxed{\text{单位：1千美元}}$

CEO公司的平均股本回报率

- **拟合回归**

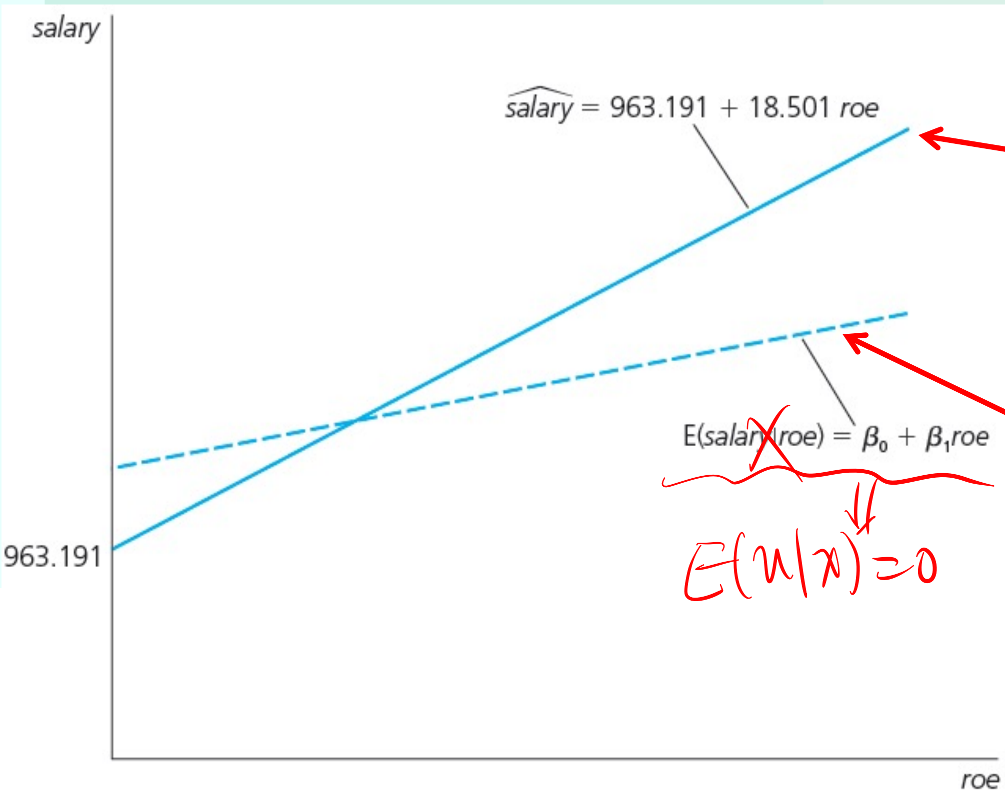
$$\widehat{salary} = 963.191 + 18.501\ roe$$

截距项

如果ROE增长1%，那么工资预计增长$18,501

- **因果关系解释？** 没有！

普通最小二乘法的推导



拟合回归线(取决于样本)

未知的总体回归线

普通最小二乘法的推导

- **工资与教育**

$$wage = \beta_0 + \beta_1 educ + u$$

$\boxed{\text{每小时工资}}$

受教育年限

- **拟合回归**

$$\widehat{wage} = -0.90 + 0.54\ educ$$

截距项

在该样本下，多受一年教育
伴随着每小时工资增加$0.54

- **“因果关系？”** 没有，

普通最小二乘法的推导

- **投票结果和竞选支出（两党）**

$$voteA = \beta_0 + \beta_1 shareA + u$$

候选人A的得票率

候选人A竞选支出的比例

- **拟合回归**

$$\widehat{voteA} = 26.81 + 0.464\ shareA$$

截距项

如果候选人A的支付比例增长1%，则他或她将多接受0.464%的投票比例

- 在有些情形中，回归分析不是用来确定因果关系，仅用于判断两个变量的相关关系。

OLS对任一样本数据的性质

大样本 $n \to \infty$

- **OLS的性质**
- **拟合值与残差**

$\min \sum (y_i - \beta_1 x_i)^2$

$\sum 2(y_i - \beta_1 x_i)(-x_i) = 0$

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

拟合或预测值　　$E(u)$?　　较回归线的偏离值(= 残差)

$\sum (x_i \hat{u}_i) = 0$

- **OLS回归的代数性质**

$$\boxed{\sum_{i=1}^{n} \hat{u}_i = 0} \qquad \boxed{\sum_{i=1}^{n} x_i \hat{u}_i = 0} \qquad \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$\frac{1}{n}\sum y_i = \frac{1}{n}\hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n}\sum x_i$

残差的均值为0（思考：如果没有截距项呢？）　没有截距

残差与自变量的协方差为0

y的样本均值落在回归线上（思考：如果没有截距项呢？）

$+\frac{1}{n}\sum \hat{u}_i$

$=0$

OLS对任一样本数据的性质

x_i

TABLE 2.1 Fitted Values and Residuals for the First 15 CEOs

obsno	roe	salary	salaryhat	uhat
1	14.1	1095	1224.058	–129.0581
2	10.9	1001	1164.854	–163.8542
3	23.5	1122	1397.969	–275.9692
4	5.9	578	1072.348	–494.3484
5	13.8	1368	1218.508	149.4923
6	20.0	1145	1333.215	–188.2151
7	16.4	1078	1266.611	–188.6108
8	16.3	1094	1264.761	–170.7606
9	10.5	1237	1157.454	79.54626
10	26.3	833	1449.773	–616.7726
11	25.9	567	1442.372	–875.3721
12	26.8	933	1459.023	–526.0231
13	14.8	1339	1237.009	101.9911
14	22.3	937	1375.768	–438.7678
15	56.3	2011	2004.808	6.191895

y_i

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

预测

$$\hat{u}_i = y_i - \hat{y}_i$$

例如，12号CEO的工资比预测值低了$526,023

$\sum \hat{u}_i = 0$, $\hat{u}_i$?

OLS对任一样本数据的性质

- **拟合优度(Goodness-of-Fit)** R^2

"自变量对因变量的解释程度如何？"

- **测量波动(Measures of Variation)**

$\hat{\beta}_0$, $\hat{\beta}_1$

$$SST \equiv \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SSE \equiv \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad SSR \equiv \sum_{i=1}^{n}\hat{u}_i^{\,2}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix} \qquad \hat{y}_i$$

总体平方和（Total sum of squares），代表因变量的总体变化

解释平方和（Explained sum of squares），代表被回归解释的y的变化

残差平方和（ Residual sum of squares ），代表没有被回归解释的y的变化

$$\underbrace{(y-\bar{y})'(y-\bar{y})} = (\hat{y}-\bar{y})'(\hat{y}-\bar{y}) + \hat{u}'\hat{u}$$

OLS对任一样本数据的性质

- **分解总体波动**

$$SST = SSE + SSR$$

总体波动　　被解释部分　　未被解释部分

思考：如果没有截距项呢？

$\min \sum (y-\beta x)^2$ vs.

$\min \sum (y-\beta x-\tilde{\beta}\tilde{x})^2$

$\min \sum (y-\beta x-\tilde{\beta}\tilde{x})^2$ s.t. $\tilde{\beta}=0$

有截距

R^2 很不重要

- **拟合优度测量（R²）**

$\min \sum (y-x\beta)^2 \equiv \max R^2$

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

R2测量了被解释变化占总体变化的比例

推论：只要变量数目增加，R^2不会减小

$$y - \hat{y} = \hat{u}$$

$$(y - \hat{y} + \hat{y} - \bar{y})'(y - \hat{y} + \hat{y} - \bar{y})$$

$$= (\hat{u} + \hat{y} - \bar{y})'(\hat{u} + \hat{y} - \bar{y})$$

$$= \hat{u}'\hat{u} + \underbrace{\boxed{2(\hat{y} - \bar{y})'\hat{u}}}_{=0} + (\hat{y} - \bar{y})'(\hat{y} - \bar{y})$$

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1 x_1 \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 x_n \end{bmatrix} = \hat{\beta}_0 \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \hat{\beta}_1 \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

$$\hat{y}'\hat{u} = \hat{\beta}_0 \sum \hat{u}_i + \hat{\beta}_1 \sum x_i u_i$$

$$= 0$$

$$\bar{y}'\hat{u} = \begin{bmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix}' \hat{u} = \bar{y} \cdot \sum \hat{u}_i = 0$$

OLS对任一样本数据的性质

- CEO 工资与股本回报率

$$\widehat{salary} = 963.191 + 18.501\ roe$$

$$n = 209, \quad R^2 = 0.0132$$

The regression explains only 1.3% of the total variation in salaries

- 投票结果和竞选支出

$$\widehat{voteA} = 26.81 + 0.464\ shareA$$

$$n = 173, \quad R^2 = 0.856$$

The regression explains 85.6% of the total variation in election outcomes

- 注意：高的R2不意味着回归具有因果解释！

度量单位和函数形式

- **CEO 工资与股本回报率**

$$salary = \beta_0 + \beta_1 roe + u$$

单位：1千美元

CEO公司的平均股本回报率

- **拟合回归**

$$\widehat{salary} = 963.191 + 18.501\ roe$$

$$salarydol = 1000 \cdot salary = 963191 + 18501\, roe$$

- **假设不用千美元而用美元做单位，OLS估计值会如何变化？**

- **如果用roe/1000替代roe，OLS估计值会如何变化？**

$$roedec = \frac{roe}{1000}, \quad salary = 963.191 + \underbrace{18.501 \cdot 1000}\cdot roedec$$

度量单位和函数形式

- **非线性：半对数形式**

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

- **教育对工资的自然对数回归**

$y = \beta_0 + \beta_1 x + \beta_2 x^2$

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

工资的自然对数

$\exp(\beta_0)\exp(\beta_1 educ) + u$

线性

- **这改变了系数的含义：**

$wage = \exp(\beta_0)\cdot\exp(\beta_1 educ)\cdot\exp(u)$

$\frac{dwage}{wage}$ %

$$\beta_1 = \frac{\Delta\log(wage)}{\Delta educ} = \frac{1}{wage}\cdot\frac{\Delta wage}{\Delta educ} = \frac{\frac{\Delta wage}{wage}}{\Delta educ}$$

工资变化的百分比

$deduc$

... 如果教育年份增加一年

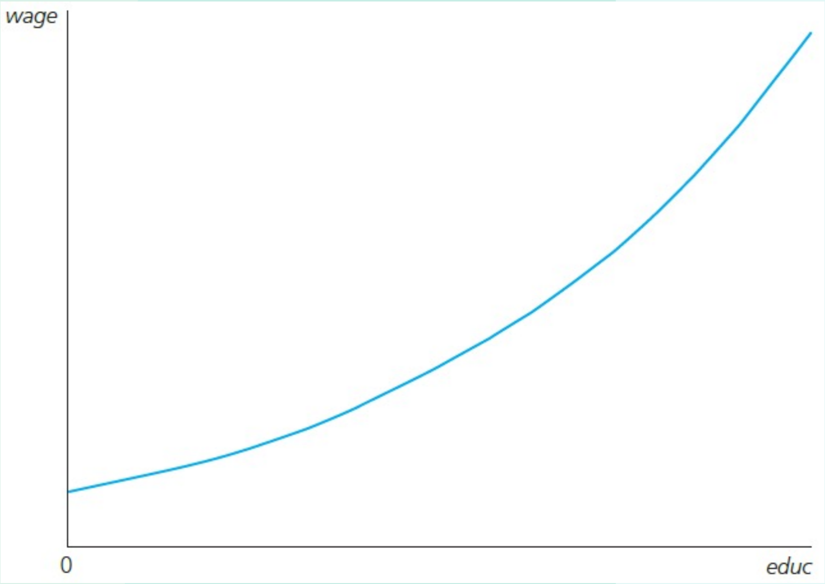
$\ln y$ x

度量单位和函数形式

- **拟合回归**

$$\widehat{\log}(wage) = 0.584 + 0.083\ educ$$

教育多一年，工资增长8.3%（=多教育一年的回报）



例如:

$$\frac{\frac{\Delta wage}{wage}}{\Delta educ} = \frac{\frac{+0.83\$}{10\$}}{+1\ year} = 0.083 = +8.3\%$$

度量单位和函数形式

- **非线性：对数形式**

$salary = \exp(\beta_0) \cdot sales^{\beta_1} \cdot \exp(u)$

- **CEO工资与公司销售**

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u$$

Natural logarithm of CEO salary

Natural logarithm of his/her firm`s sales

- **这改变了系数的含义：**

$$\beta_1 = \frac{\Delta \log(salary)}{\Delta \log(sales)} = \boxed{\frac{\frac{\Delta salary}{salary}}{\frac{\Delta sales}{sales}}}$$

工资变化的增长率

... 如果销售增长<u>1%</u>

度量单位和函数形式

- **CEO工资与公司销售：拟合回归**

$$\widehat{\log}(salary) = 4.822 + 0.257 \log(sales)$$

+ 1% sales; + 0.257% salary

- **例如：**

$$\frac{\frac{\Delta salary}{salary}}{\frac{\Delta sales}{sales}} = \frac{\frac{+2{,}570\$}{1{,}000{,}000\$}}{\frac{+10{,}000{,}000\$}{1{,}000{,}000{,}000\$}} = \frac{+0.257\%\ salary}{+1\%\ sales} = 0.257$$

- **对数形式假设一个恒定的弹性模型，而半对数形式假设一个半弹性模型**

度量单位和函数形式

- **回归模型"线性"的含义**
- **简单回归允许解释变量以非线性形式出现:**
- 例:

$$cons = \beta_0 + \beta_1 \sqrt{inc} + c$$

- **简单回归不能处理参数的非线性:**

例:

$$cons = 1/(\boxed{\beta_0} + \boxed{\beta_1}\sqrt{inc}) + c$$

OLS估计量的期望值和方差

- OLS估计值的期望与方差
- 估计的回归系数是随机变量，因为它们是由一个随机样本估计得到的（跨样本随机变量）

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

Data is random and depends on particular sample that has been drawn

- 估计值的平均值是多少，以及它们的可变性有多大?

$$E(\hat{\beta}_0) = ?,\ E(\hat{\beta}_1) = ? \qquad Var(\hat{\beta}_0) = ?,\ Var(\hat{\beta}_1) = ?$$

（注意：本节中讨论的其实为OLS估计在给定x1,…,xn 条件下的均值与方差）

OLS估计量的期望值和方差

- **线性回归的标准假设（高斯-马尔科夫假定）**
- **假设 SLR. 1 （线性于参数）**

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

在总体模型中，y、x和u是线性的

- **假设 SLR. 2 （随机抽样）**

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

数据由总体样本中随机抽取

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

每一个数据点都遵循总体等式

OLS估计量的期望值和方差

- **讨论随机抽样：工资与教育**
  - 例如，一个国家的人口由所有工人组成
  - 在总体等式中，工资（或log工资）与受教育年限之间存在线性关系
  - 从人群中随机抽取一名工人
  - 被抽取的工人的工资和受教育年限是随机的，因为人们事先不知道哪个工人被抽取
  - 将工人返回人群中，重复随机抽取n次
  - 抽样工人的工资和受教育年限用于估计工资和教育之间的线性关系

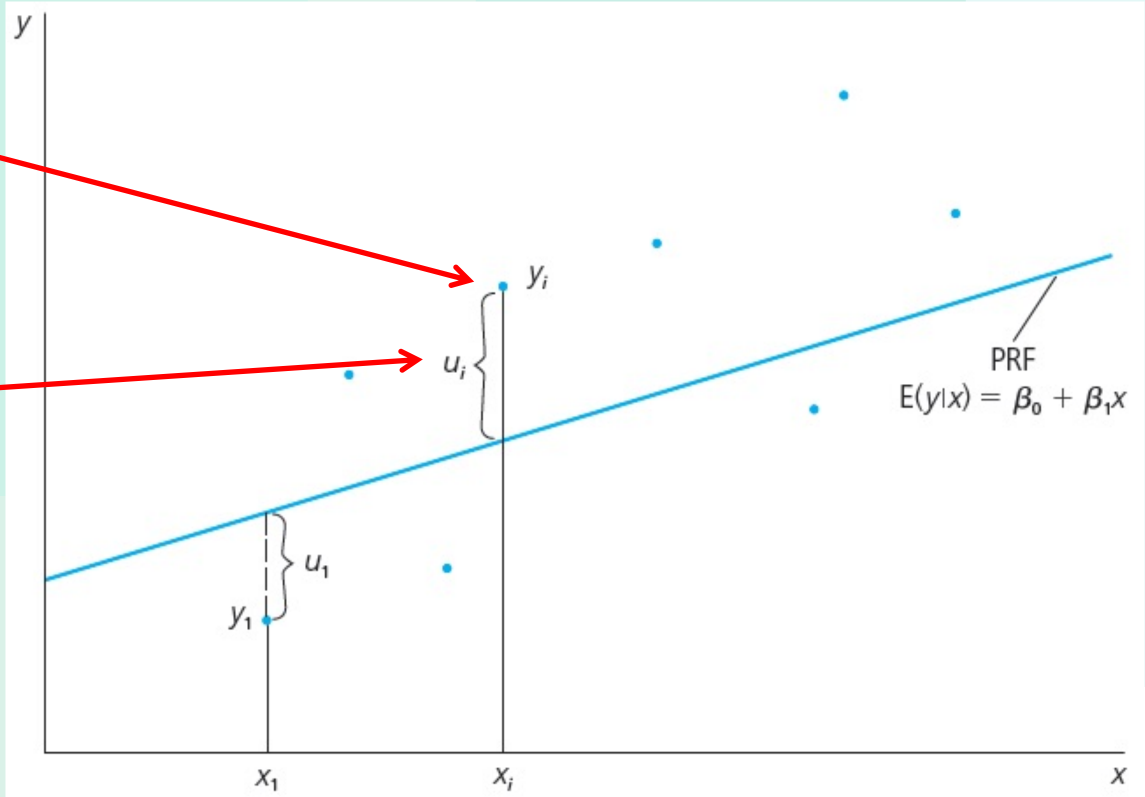
OLS估计量的期望值和方差

第i个工人的值

(x_i, y_i)

第i个工人推导出的与总体等式的偏差

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



OLS估计量的期望值和方差

- **假设 SLR.3（解释变量的样本有波动）**

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

解释变量的值并不完全相同（否则就不可能研究解释变量的不同值如何导致因变量的不同值）

- **假设 SLR.4（零条件均值）**

$$E(u_i|x_i) = 0$$

解释变量的值不得包含有关未观察因素平均值的信息

OLS估计量的期望值和方差

- **定理2.1（OLS的无偏性）**

$$SLR.1 - SLR.4 \quad \Rightarrow \quad E(\hat{\beta}_0) = \beta_0,\ E(\hat{\beta}_1) = \beta_1$$

- **对无偏性的解释**
  - 估计系数可能更小或更大，这取决于随机抽取的结果样本
  - 然而，平均而言，它们将等于代表总体样本中y和x之间真实关系的值
  - “平均”是指重复抽样，即抽取随机样本并进行多次估计
  - 在给定的样本中，估计值可能与真实值存在很大差异

OLS估计量的期望值和方差

- **OLS估计值的方差**
  - 根据样本，估计值可能更接近或更远离真实总体值
  - 我们预计的估计值与真实总体值（=抽样可变性）相差多远？
  - 抽样可变性由估计值的方差来衡量

$$Var(\hat{\beta}_0),\ Var(\hat{\beta}_1)$$

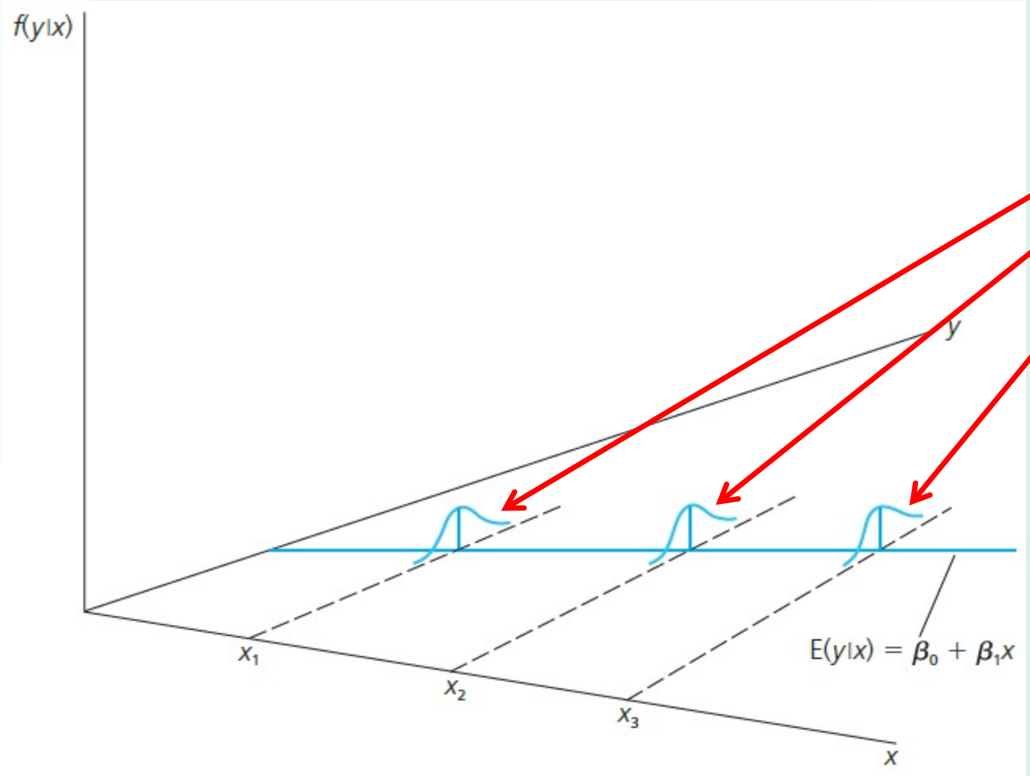
- **假设 SLR.5（同方差性）**

$$Var(u_i|x_i) = \sigma^2$$

解释变量的值不得包含有关
未观察因素可变性的信息

OLS估计量的期望值和方差

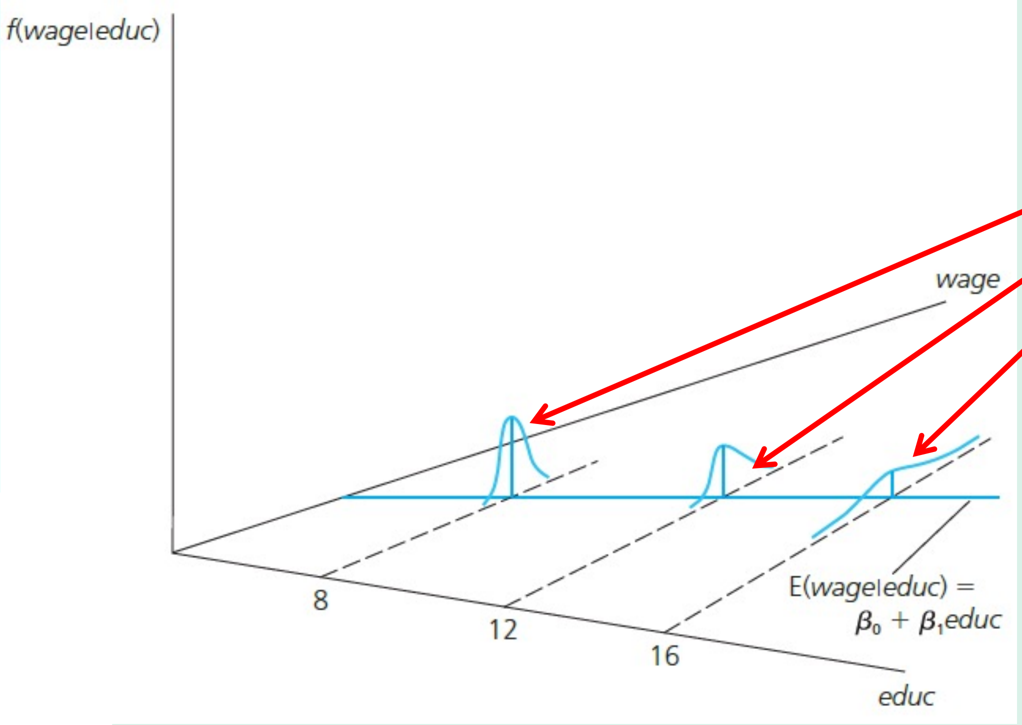
- **同方差的几何表述**



未观察到因素的可变性并
不取决于解释变量的值

OLS估计量的期望值和方差

- **一个异方差的例子：工资与教育**



未观测到的因素的变化随着教育水平的增长而增长

OLS估计量的期望值和方差

- **定理2.2（OLS估计量的抽样方差）**

  根据假设 SLR.1 – SLR.5:

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

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

- **结论**:
  - 估计回归系数的抽样可变性越高，未观测因素的波动性越大，解释变量的波动性越高

OLS估计量的期望值和方差

- **在讨论方差的估计之前，我们首先强调误差u_i与残差$\hat{u}_i$的区别**
- 误差$\mathrm{u_i}$出现在简单回归模型中，满足某些条件，但无法被观测

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

- $\hat{u}_i$代表拟合的残差，可以用样本（y_i和x_i）和OLS估计值（$\hat{\beta}_0$与$\hat{\beta}_1$）求出

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

- 从而，可以建立u_i与$\hat{u}_i$的关系：

$$\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i$$

OLS估计量的期望值和方差

- **误差方差的估计**

$$Var(u_i|x_i) = \sigma^2 = Var(u_i)$$

u的方差不取决于x，而等于无条件方差

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

计算样本残差的方差来估计误差方差；
不幸的是，这个估计是有偏的

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

误差方差的无偏估计可以通过从观测值的数量中减去估计回归系数的数量来获得

OLS估计量的期望值和方差

- **定理 2.3（误差方差的无偏估计）**

$$SLR.1 - SLR.5 \quad \Rightarrow \quad E(\hat{\sigma}^2) = \sigma^2$$

- **计算回归系数标准误差**

$$se(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)} = \sqrt{\hat{\sigma}^2 / SST_x}$$

用 $\hat{\sigma}^2$ 来替代未知的 σ^2

$$se(\hat{\beta}_0) = \sqrt{\widehat{Var}(\hat{\beta}_0)} = \sqrt{\hat{\sigma}^2 n^{-1} \sum_{i=1}^{n} x_i^2 / SST_x}$$

回归系数的估计标准偏差称为“标准误差”。它们测量回归系数的估计精度。

过原点回归及对常数回归

- **在少数情形中，我们希望施加如下约束：当x=0时，y的期望值为零**

  例如：若收入（x）为零，那收入税所得（y）也必须为零

- **我们得到一个没有截距项的回归模型（过原点回归）：**

$$y = \beta_1 x + u$$

- **通过最小化** $min \sum_{i=1}^{n}(y_i - \tilde{\beta}_1 x)^2$ **可得：**

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} {x_i}^2}$$

- **拟合优度测量 (R²)**:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{\beta}_1 x)^2}{\sum_{i=1}^{n} {y_i}^2}$$

（注意与有截距项情况的区别）

过原点回归及对常数回归

- 如果只对一个常数进行回归（对常数回归）：

$$y = \beta_0 + u$$

- 通过最小化$min \sum_{i=1}^{n}(y_i - \tilde{\beta}_0)^2$可得：

$$\tilde{\beta}_0 = \bar{y}$$

小节

- 我们在本章介绍了简单线性回归模型，并讨论了它的基本性质
- 普通最小二乘法被用于估计总体模型中的斜率和截距参数
- 我们说明了 OLS 的代数性质，包括拟合值和残差的计算，以及拟合优度测量（R^2）
- 我们讨论了度量单位和函数形式对OLS估计的影响
- 我们证明了在SLR.1至SLR.4四个假定之下，OLS估计量是无偏的
- 当我们加入同方差假定SLR.5后，便得到OLS估计量方差的简单表达式