应用统计学II 第2讲 相关分析与一元线性回归分析

Instructor: 郝壮

haozhuang@buaa.edu.cn
School of Economics and Management
Beihang University

May 9, 2022

相关分析与一元线性回归模型

对应教材第十章.

- 相关系数 (correlation coefficient)
- 一元线性回归模型(simple linear regression model)

第十章　相关分析与一元线性回归模型

- §10.1 相关系数的概念
- §10.2 线性回归模型
- §10.3 最小二乘估计方法
- §10.4 模型效果分析
- §10.5 显著性检验
- §10.6 变量筛选 (本节跳过, 并在多元回归中介绍)
- §10.7 残差分析

本章推荐额外参考教材: 伍德里奇. 计量经济学导论: 现代观点 (第六版). 人民大学出版社.

或

英文原版 Wooldridge, J. M. (2016). Introductory econometrics: A modern approach. Nelson Education.

有关本章内容的大量证明均可参考此教材.

第一节　相关系数(correlation coefficient)

常见的两类变量关系

- **确定性关系** (deterministic relationship): 每一个 X 值都唯一地对应一个Y值, 如$Y = f(X)$.
- **随机关系** (stochastic relationship): 当 X的值给定时, Y的取值服从一个分布.

问题： 如果X与Y为随机关系, 那么给定X的取值, Y的分布可以用什么函数描述?

第一节　相关系数(correlation coefficient)

常见的两类变量关系

- **确定性关系** (deterministic relationship): 每一个 X 值都唯一地对应一个Y值, 如$Y = f(X)$.
- **随机关系** (stochastic relationship): 当 X的值给定时, Y的取值服从一个分布.

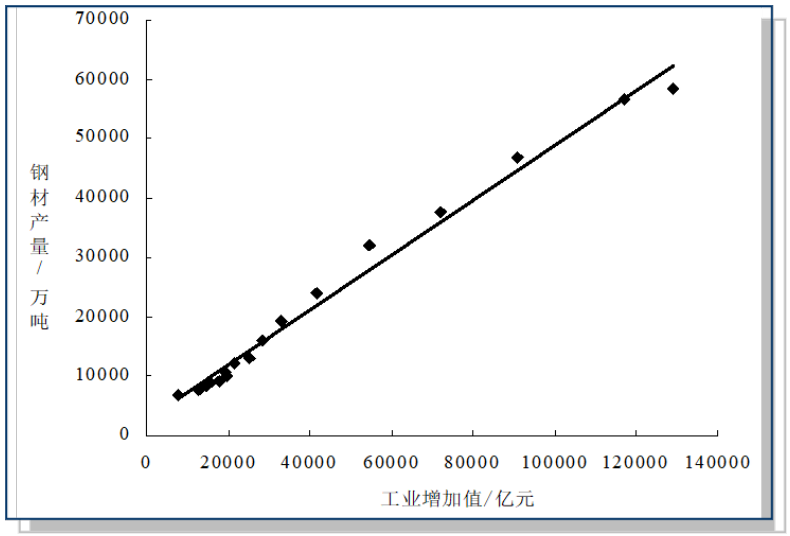
问题： 如果X与Y为随机关系, 那么给定X的取值, Y的分布可以用什么函数描述?
答：条件分布函数.
*所以教材中pp267 所谓"函数关系"实际是对应于"随机关系"的"确定性关系", 用"确定性关系"描述更准确.

随机关系例1: 1992-2009年我国钢材消费量与工业增加值

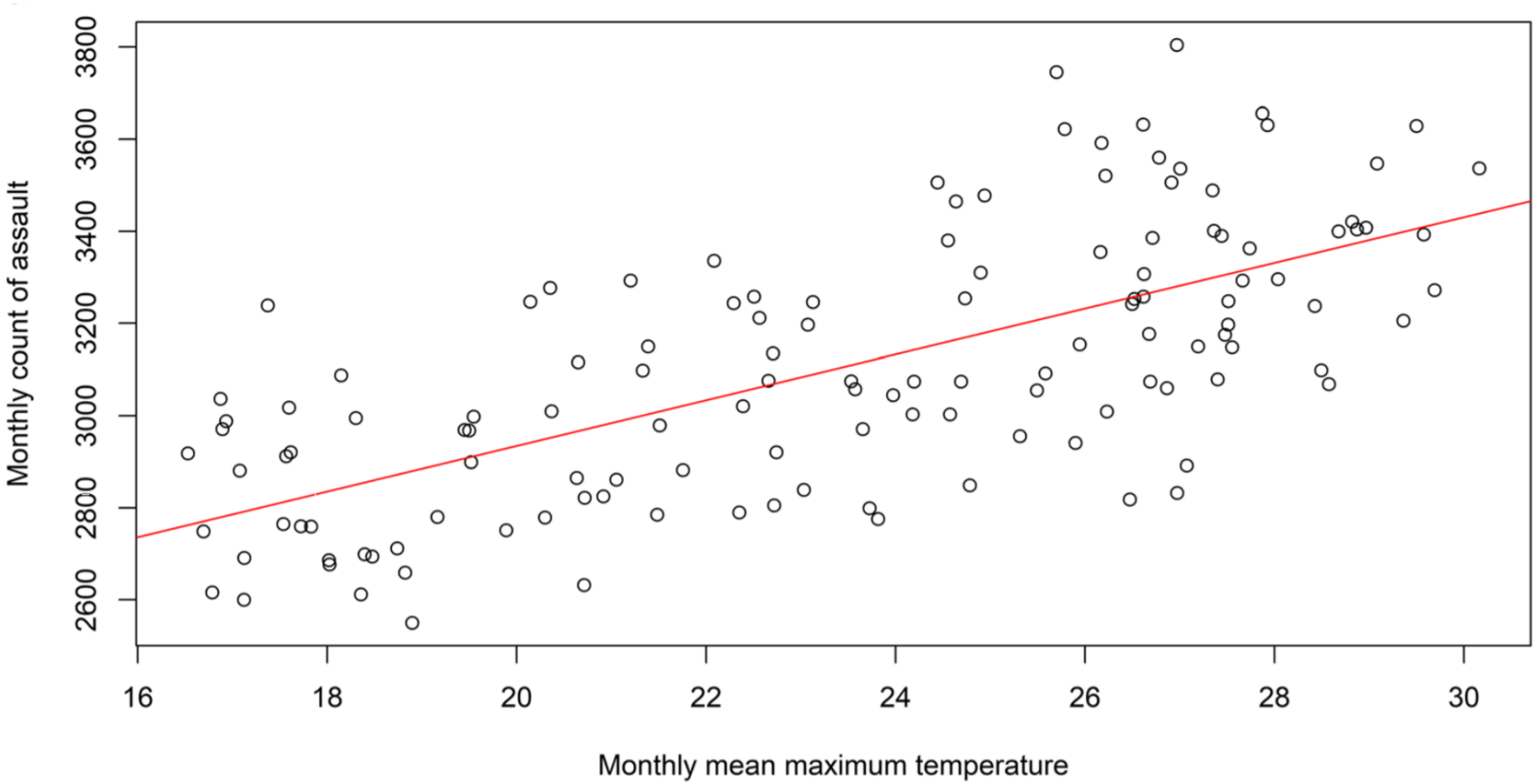
- Y —钢材消费量(万吨)
- X —工业增加值(亿元)

年份	钢材产量 (万吨)	工业增加值 (亿元)
1992	6697.0	7665.5
1993	7716.0	12842.6
1994	8428.0	14700.1
1995	8979.8	15446.1
1996	9338.0	18026.1
1997	9978.9	19835.2
1998	10737.8	19421.9
1999	12109.8	21564.7
2000	13146.0	25394.9
2001	16067.6	28329.4
2002	19251.6	32994.8
2003	24108.0	41990.2
2004	31975.7	54805.1
2005	37771.1	72187.0
2006	46893.4	91075.7
2007	56560.9	117048.4
2008	58488.1	129112.0

数据来源： 1993-2009 中国统计年鉴

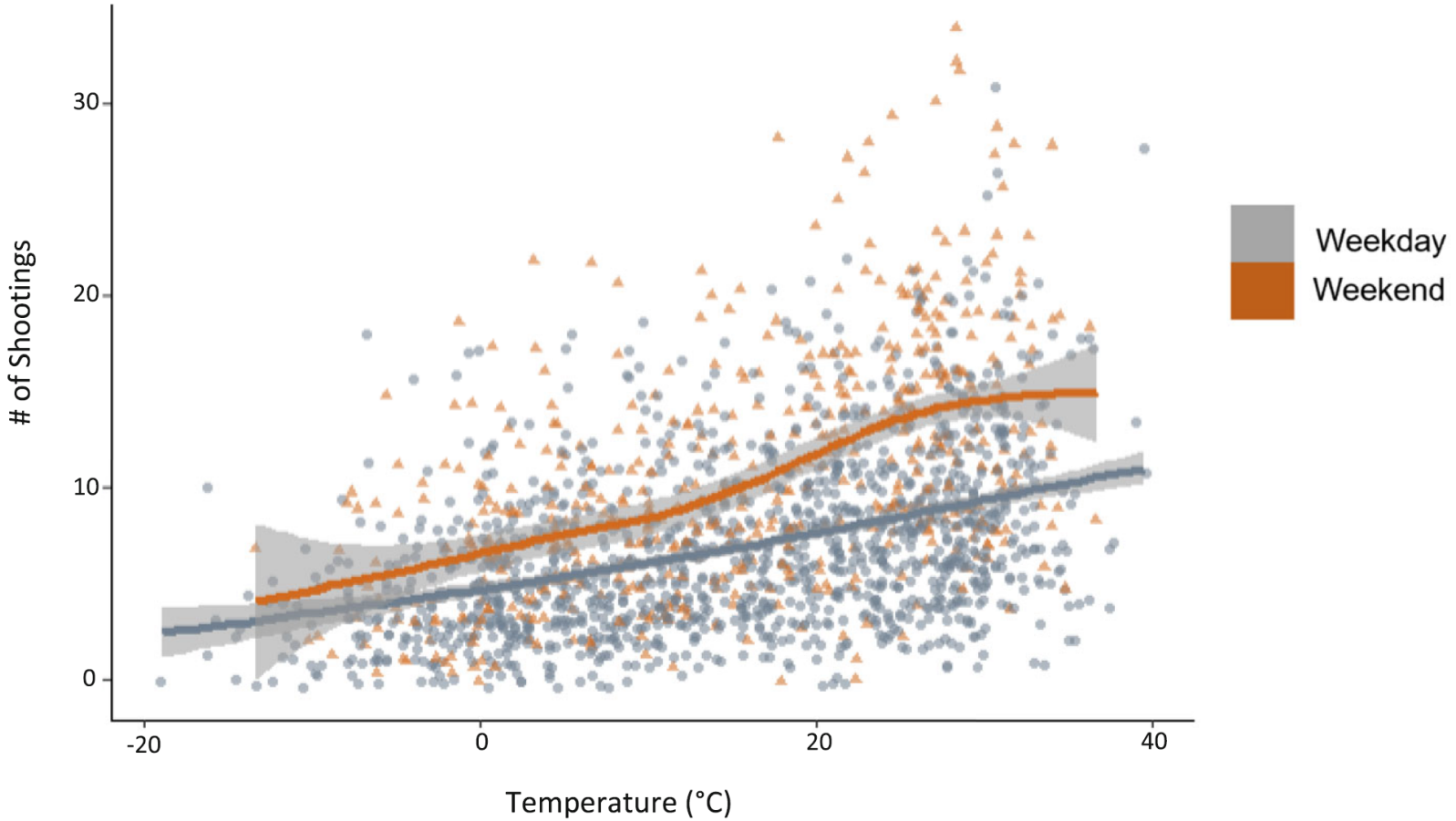


随机关系例2：气温与犯罪-悉尼2006-2016年气温与袭击案发数量关系



来源： 国际生物气象学杂志 Stevens et al. (2019). Hot and bothered? Associations between temperature and crime in Australia. International journal of biometeorology.

随机关系例3：气温与犯罪-芝加哥2012-2016年气温与枪击案发数量关系



来源： Reeping and Hemenway (2020). The association between weather and the number of daily shootings in Chicago (2012–2016). Injury epidemiology.

相关系数 (correlation coefficient)

概率论中我们已经学过从随机变量角度衡量相关关系的协方差和相关系数. 本节课给出三个从样本角度衡量相关关系的相关系数.
测量两个随机变量之间相关关系强弱的工具：

- **Pearson 相关系数**(Pearson correlation coefficient/ Pearson's r): $r(x, y)$测度线性相关
- **Spearman 秩相关系数** (Spearman's rank-order correlation) 测度非线性相关
- **Kendall** τ (Kendall rank correlation): 测度排序一致性

1.2 Pearson 相关系数$r(x,y)$

定义： 给定观测值 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，**Pearson 相关系数为：**

$$r_{xy}=\frac{s_{xy}}{s_x s_y}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

其中s_{xy}是样本协方差(sample covariance)

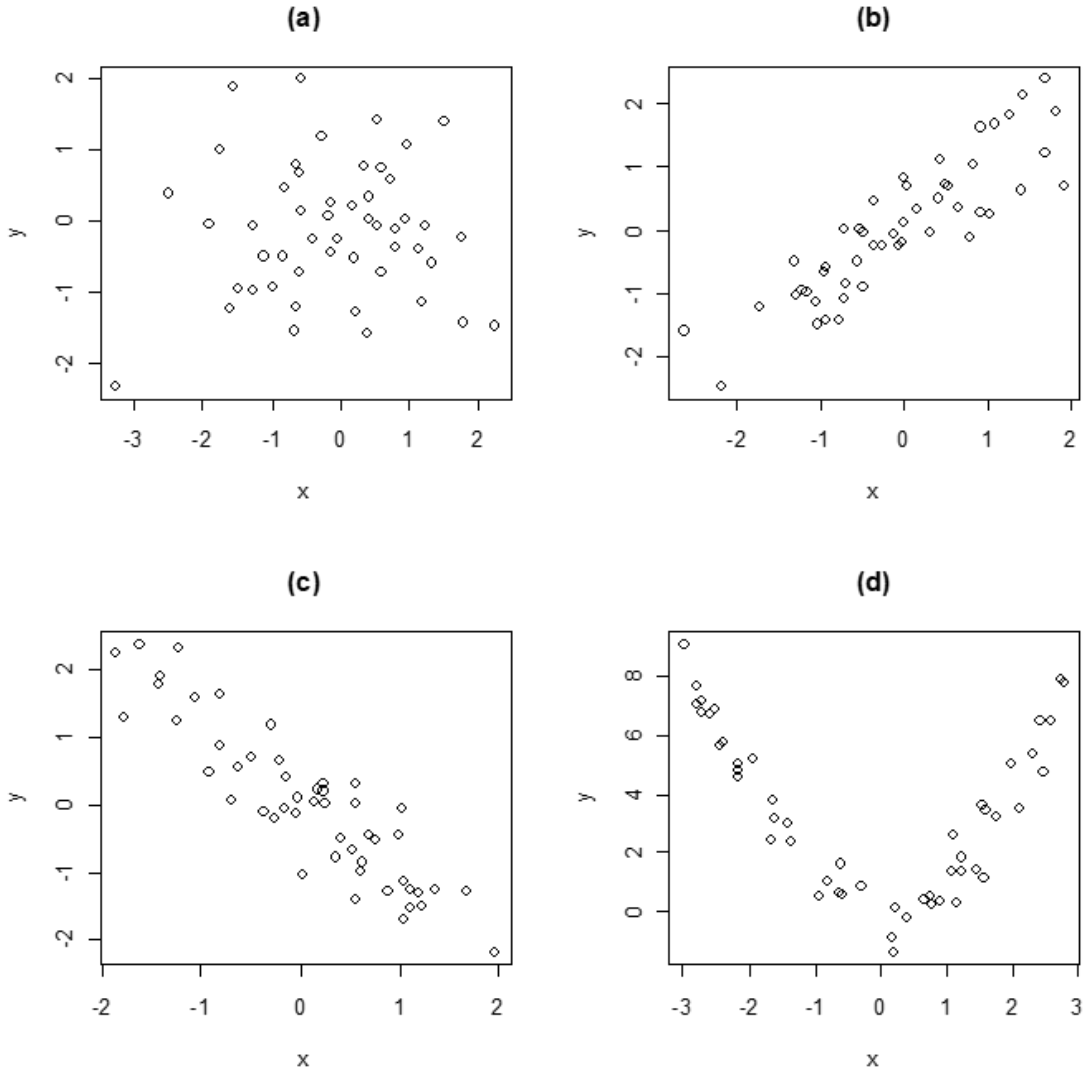
$$s_{xy}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

s_x,s_y是样本标准差(sample standard deviation)

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

易知, $-1\le r_{xy}\le 1$.

(a)$r(x,y)=0$ (b)$r(x,y)>0$ (c)$r(x,y)<0$



注意： Pearson相关系数用于测量2个变量之间的线性关系, 不能较好反映(d)图显示的非线性关系.

$r(x,y)$ 的性质

r	相关性
$r>0$	正线性相关
$r<0$	负线性相关
$r=0$	线性无关
$r=1$	完全正线性相关
$r=-1$	完全负线性相关

经验判断

r	相关性
$\lvert r_{xy}\rvert \geq 0.8$	强线性相关
$0.5 \leq \lvert r_{xy}\rvert < 0.8$	中度线性相关
$0.3 \leq \lvert r_{xy}\rvert < 0.5$	弱线性相关
$\lvert r_{xy}\rvert < 0.3$	不相关

- **注意：** $|r_{xy}| < 0.3$ 可近似认为"不相关"并不是总成立! (思考: 取决于哪些因素?)
- **问题:** 如何严谨地确定x与y是否相关? 如$r_{xy} = 0.25$, x与y是否相关或不相关?

经验判断

r	相关性
$\lvert r_{xy}\rvert \geq 0.8$	强线性相关
$0.5 \leq \lvert r_{xy}\rvert < 0.8$	中度线性相关
$0.3 \leq \lvert r_{xy}\rvert < 0.5$	弱线性相关
$\lvert r_{xy}\rvert < 0.3$	不相关

- **注意：** $|r_{xy}| < 0.3$ 可近似认为"不相关"并不是总成立! (思考: 取决于哪些因素?)
- **问题:** 如何严谨地确定x与y是否相关? 如$r_{xy} = 0.25$, x与y是否相关或不相关?
- 答: 利用假设检验, 检验两个总体的相关系数是否为0.

回忆总体相关系数定义

1. 总体方差

$$\sigma_X^2 = E\left[(X-\mu_X)^2\right], \quad \sigma_Y^2 = E\left[(Y-\mu_Y)^2\right]$$

2. 总体协方差

$$\sigma_{XY} = \mathrm{Cov}(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

3. 总体相关系数

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

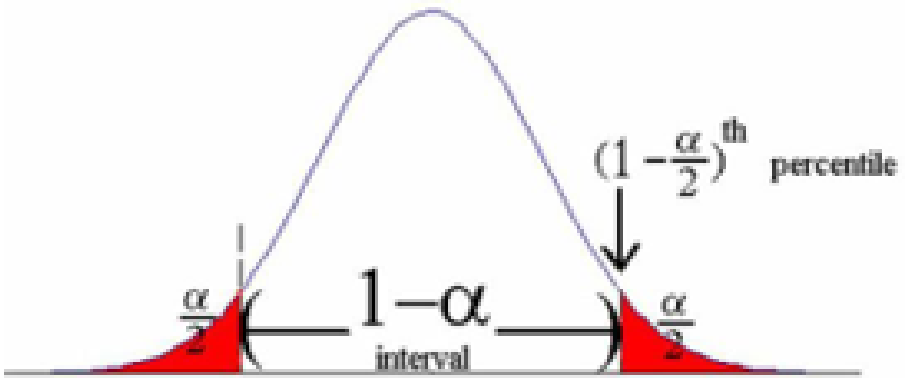
关于总体相关系数ρ的假设检验

$H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$ 检验统计量为：

$$T = \frac{r\sqrt{n-2}}{1-r^2}$$

其中r为 Pearson 相关系数(通过样本可求出)
在 H_0 下, $T \sim t(n-2)$

取 $\alpha = 0.05 \Rightarrow \mathrm{P}\left\{|T| > t_{\alpha/2}\right\} = 0.05$
如果 $|T| > t_{\alpha/2}$, 或 (P 值 < 0.05), 拒绝 H_0
如果 $|T| \leq t_{\alpha/2}$, 或 (P 值 ≥ 0.05), 不拒绝 H_0



Stata 计算Pearson 相关系数(Pearson's correlation coefficient)与相关性假设检验

```
1. 仅看相关系数
corr x y

2. 既想知道相关系数, 又想知道显著水平(significance level)p-value
pwcorr x y, sig
```

Stata 计算Pearson 相关系数(Pearson's correlation coefficient)与相关性假设检验

```
// Setup
sysuse auto

. list price weight mpg headroom in 1/10
```

	price	weight	mpg	headroom
1.	4,099	2,930	22	2.5
2.	4,749	3,350	17	3.0
3.	3,799	2,640	22	3.0
4.	4,816	3,250	20	4.5
5.	7,827	4,080	15	4.0
6.	5,788	3,670	18	4.0
7.	4,453	2,230	26	3.0
8.	5,189	3,280	20	2.0
9.	10,372	3,880	16	3.5
10.	4,082	3,400	19	3.5

Stata 计算Pearson 相关系数(Pearson's correlation coefficient)与相关性假设检验

```
// Estimate all pairwise correlations
pwcorr price weight mpg headroom

//  Add significance level to each entry
pwcorr price weight mpg headroom, sig

//  Add stars to correlations significant at the 1% level
 pwcorr price weight mpg headroom, star(.01) sig
```

```
             |    price   weight      mpg headroom
-------------+------------------------------------
       price |   1.0000
             |
      weight |   0.5386*  1.0000
             |   0.0000
             |
         mpg |  -0.4686* -0.8072*  1.0000
             |   0.0000   0.0000
             |
    headroom |   0.1145   0.4835* -0.4138*  1.0000
             |   0.3313   0.0000   0.0002
```

2、Spearman 秩相关系数(Spearman's rank-order correlation)

Pearson样本相关系数只能测量两个随机变量之间是否存在线性相关关系. 非线性相关关系可用Spearman 秩相关系数测量.

Spearman 相关系数定义： 给定观测值 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, Spearman 相关系数 ρ 定义为观测值的秩(Rank)之间的 Pearson 相关系数：

$$\rho_{xy} = \frac{\sum (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum (u_i - \bar{u})^2}\sqrt{\sum (v_i - \bar{v})^2}}$$

其中, u_i 和 v_i 代表 x_i 和 y_i 的秩.

(注意: 1. 教材9.5章定义为本定义的特例. 2. 教材中写样本1, 样本2,...样本n不准确, 应为观测1, 观测2,...)

例： Spearman 秩相关系数计算

设观测值为

i	1	2	3	4	5	6
x_i	1	2	3	4	5	6
y_i	1	4	9	16	25	36

求Spearman 秩相关系数.

例： Spearman 秩相关系数计算

首先计算秩

i		1	2	3	4	5	6
rank	(x_i)	1	2	3	4	5	6
rank	(y_i)	1	2	3	4	5	6

将秩带入公式, 得到Spearman 秩相关系数.

问题： Spearman 相关系数=?

例2： Spearman 秩相关系数计算

对某地区12个街道经济展水平与卫生条件按规定的标准打分. 评分及Spearman相关系数计算过程见下表

编号	经济水平	卫生水平	u (经济)	v(卫生)
1	82	86	6	9
2	87	78	9	6
3	60	65	1	2
4	98	88	12	10
5	75	64	3	1
6	89	90	10	11
7	84	80	7	7
8	78	77	4	5
9	80	76	5	4
10	94	96	11	12
11	85	85	8	8
N2	68	70	2	3

计算Spearman秩相关系数: $\rho = 0.8881$, 说明地区的经济水平与卫生水平存在正相关关系.

Spearman秩相关假设检验

检验统计量为：

$$T = \rho\sqrt{\frac{n-2}{1-\rho^2}}$$

在 H_0 下, $T \sim t(n-2)$.

Stata实现: Spearman秩相关系数及其假设检验

```
sysuse auto
// Two variables; output displayed in tabular form by default
spearman price  weight

// Two variables; output displayed in matrix form
spearman price  weight mpg headroom, matrix

// Use all nonmissing observations between a pair of variables
 spearman price  weight mpg headroom, pw

// Star all correlation coefficients significant at the 5% level or lower
spearman price  weight mpg headroom, pw star(.05)

             |    price   weight      mpg headroom
-------------+------------------------------------
       price |   1.0000
      weight |   0.4865*  1.0000
         mpg |  -0.5419* -0.8576*  1.0000
    headroom |   0.0969   0.5281* -0.4866*  1.0000
```

Kendall τ相关系数: 对于顺序相关性的测度

Kendall τ 相关系数给定观测值 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, 对于任意的$i < j$, 如果 $(x_i - x_j)(y_i - y_j) > 0$, (x_i, y_i) 和 (x_j, y_j) 叫作同向数对 (反之几位反向对).

全部数据所有可能的数对为 $C_n^2 = \frac{n(n-1)}{2}$, 其中同向数对记为 N_c, 反向数对记为 N_d, Kendall τ 相关系数:

$$\tau = \frac{N_c - N_d}{n(n-1)/2}$$

Kendall τ相关系数可用于分析两个定序变量的协同性 (例: 两个专家对n个事物的评分是否一致)

Kendall τ 相关系数

Kendall τ 相关系数:

$$\tau = \frac{N_c - N_d}{n(n-1)/2}$$

- 如果所有对数完全同向： $\boldsymbol{\tau = 1}$
- 如果所有对数完全反向： $\boldsymbol{\tau = -1}$
- 如果数对中的同向与反向的数量相同： $\boldsymbol{\tau = 0}$

可以证明, 当 $\boldsymbol{n \to \infty}$ 时：

$$\tau\sqrt{\frac{9n(n-1)}{2(2n+5)}} \underset{n\to\infty}{\to} N(0,1)$$

总结

- Pearson相关系数适用于两个连续变量间呈线性相关;
- Spearman秩相关系数适合定序变量、非线性等情形;
- Kendall τ 相关系数适用于分析两个定序变量的协同性(例如两个专家对n个事物的评价, 可能他们的评分松紧不同, 但关键是评价排序是否一致)

三种相关系数的Stata 实现

```
clear
set obs 100
gen x=_n
gen x2=x*x


corr x x2

spearman x x2

ktau x x2

//ktau will produce ktau_a and ktau_b which are both right,
// formula given in accompanied pdf file.
```

1.3 相关关系与因果关系讨论

案例： 18世纪在西班牙首次发现糙皮病. 糙皮病是非常贫困的居民中体弱多病、伤残、夭折的一个重要原因. 患病者家庭贫困, 环境条件恶劣, 到处有苍蝇. 而在欧洲, 一种吸血蝇与糙皮病有同样的地理分布范围; 而吸血蝇在春天最为活跃,恰恰是糙皮病发生病历最多的季节. 许多流行病专家认为这种疾病是传染性的——由昆虫传染.

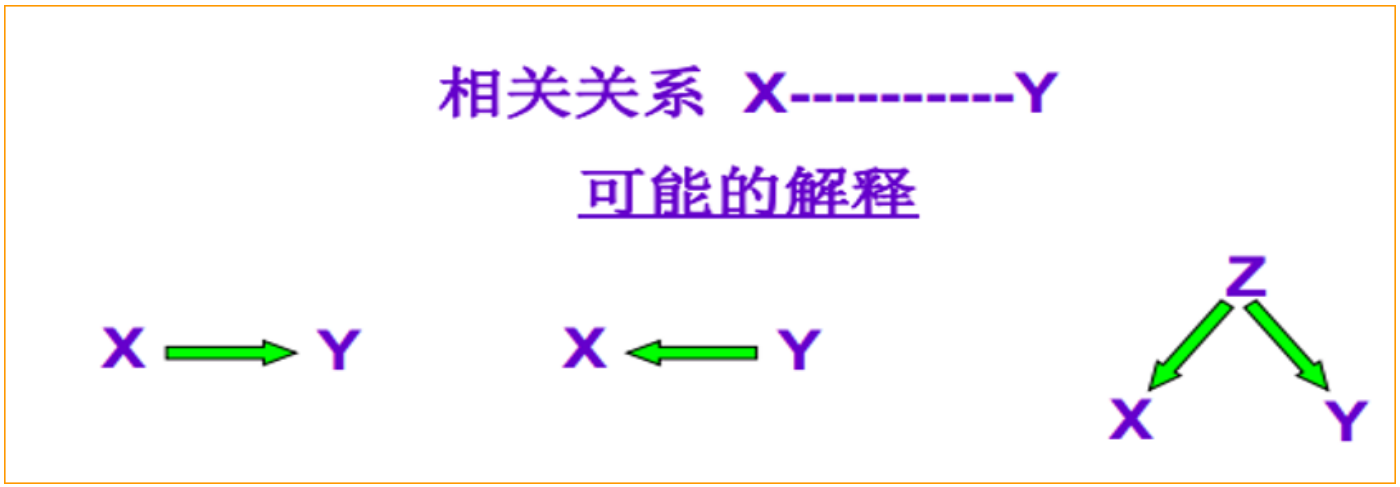
1914年初, 美国医生Joseph Goldberger通过实验研究证实, 糙皮病是由于不良饮食引起的, 可以通过食用含烟酸的食物而预防和治疗. 烟酸天然存在与肉、奶、蛋和一些蔬菜、谷物中. 发病地区的穷人主要以玉米为食物, 而玉米几乎不含烟酸.

苍蝇是贫穷的标志, 而不是糙皮病的起因!

还有一些例子

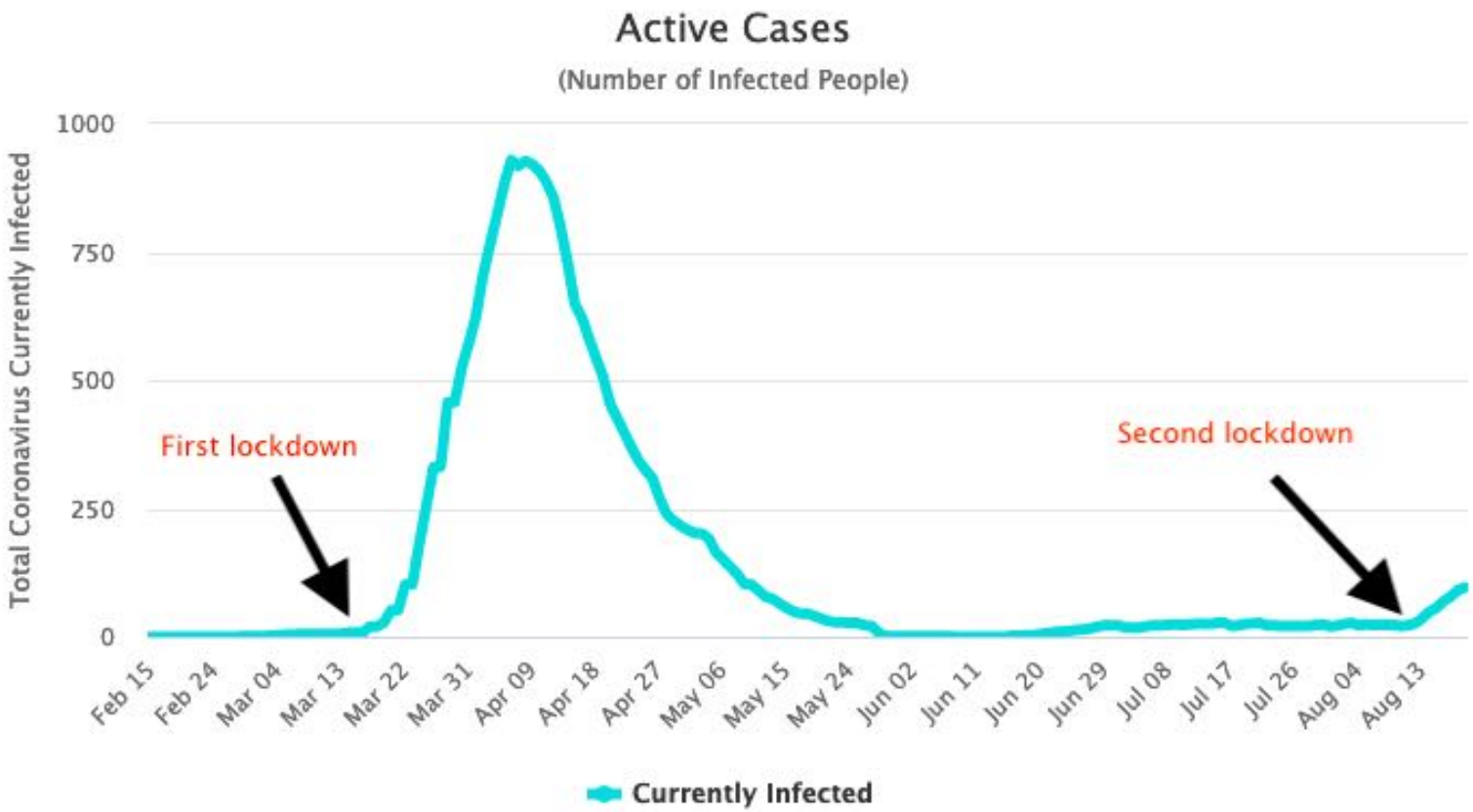
- (1)闪电是打雷的原因吗?
- (2)公鸡打鸣, 天就亮.
- (3)哲学家罗素(Bertrand Russell)讨论因果问题: 在一只鸡看来, 农妇到来, 就会有食物从天而降.
- (4)航空运量的增长是经济增长的线性趋势.

有相关关系, 不一定有因果关系!

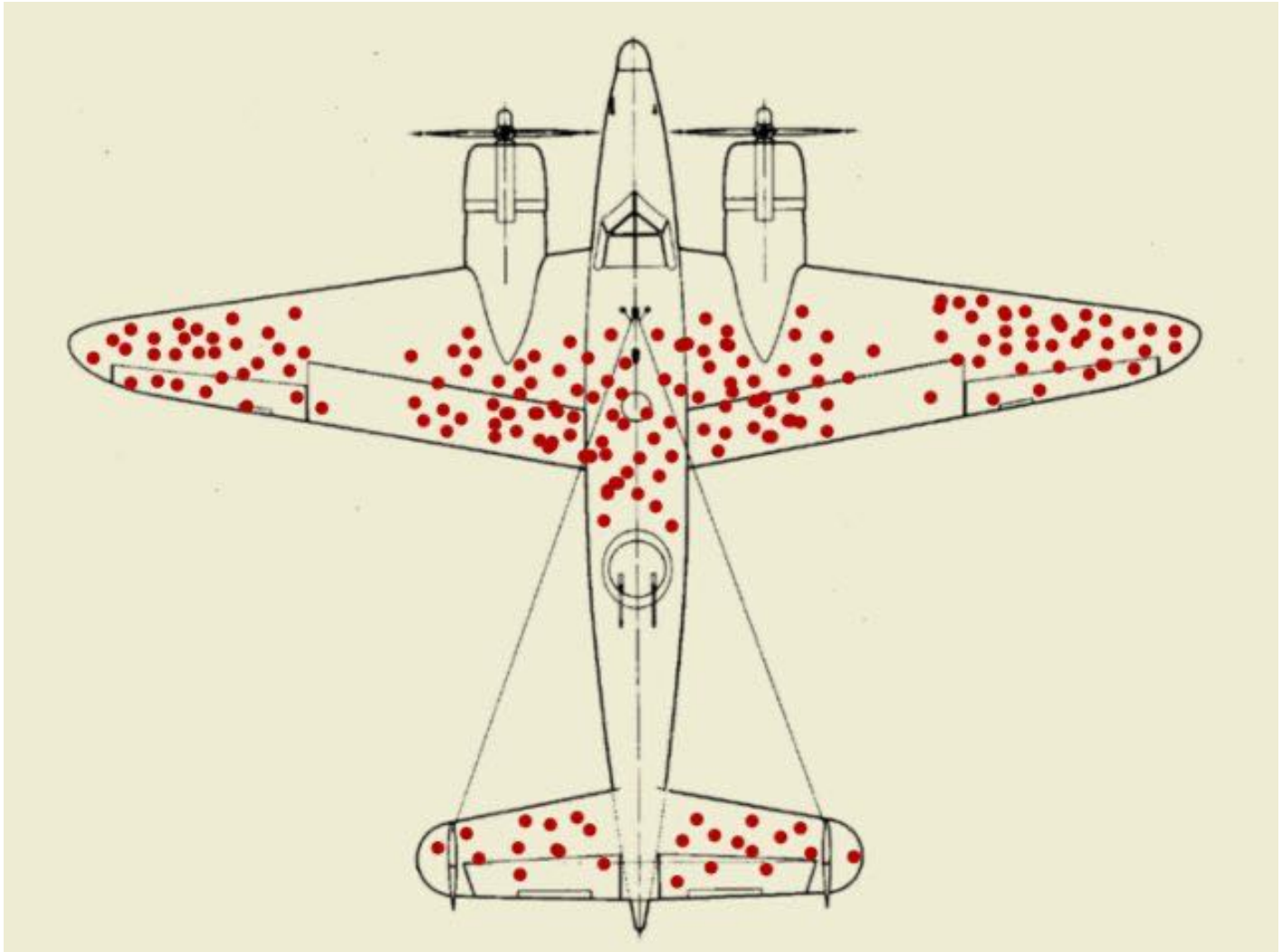


反向因果 reverse causality

Active Cases in New Zealand



样本选择偏误 sample selection bias



拓展内容：过去50年最重要的统计学思想

Gelman and Vehtari (2021). What are the most important statistical ideas of the past 50 years?

https://arxiv.org/pdf/2012.00174.pdf

中文翻译：

https://www.sohu.com/a/446294804_505915

排名第一的思想：反事实因果推断 (Counterfactual causal inference)

如何从相关分离因果(causal inference)？高级计量II.

再论:公鸡打鸣天就亮

虽然我们更想知道因果关系, 但即使在很多情况下分离不出因果, 相关关系也有其重要性和应用价值 (**相关关系也可以用作预测**).

案例1： 利用相关关系进行预测预警

- 利用电磁波和地震波的时间差来发布地震预警, 能抢出40秒时间. 中国"国家地震烈度速报与预警工程".
- 虽然电磁波不是地震的原因, 但我们能通过探测电磁波对地震进行预警.

案例2： 利用案发现场脚印长度预策犯罪嫌疑人身高.

问题： 其他应用相关关系进行社会经济分析的场景?

再论:公鸡打鸣天就亮

虽然我们更想知道因果关系, 但即使在很多情况下分离不出因果, 相关关系也有其重要性和应用价值 (**相关关系也可以用作预测**).

案例1： 利用相关关系进行预测预警

- 利用电磁波和地震波的时间差来发布地震预警, 能抢出40秒时间. 中国"国家地震烈度速报与预警工程".
- 虽然电磁波不是地震的原因, 但我们能通过探测电磁波对地震进行预警.

案例2： 利用案发现场脚印长度预策犯罪嫌疑人身高.

问题： 其他应用相关关系进行社会经济分析的场景？大数据推荐商品：您可能还喜欢.

第二节　线性回归模型(Linear Regression Models)

回归模型(Francis Galton 1889, Natural Inheritance): 研究因变量与自变量**(在一定假设条件下)**之间的“因果关系”
例:

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- 因变量 (dependent variable): Y
- 自变量 (independent variable): $X_1, X_2, \ldots$
- 总体参数: $\beta_0, \beta_1, \beta_2$
- 随机误差: ε

拓展: 一般化的线性回归模型的表达

设 y 与 x 间有相关关系, 称x为**自变量**, y为**因变量**, 在知道 x 的取值后 y 的取值是随机的, 因此有一个分布, 这个分布是知道x后的 Y 的条件密度函数

$$p(y \mid x).$$

我们关心的是 y 的均值 $E(y \mid x)$, 它是 x 的函数, 叫做雕件期望函数(conditional expectation function).
这个函数是确定性的:

$$f(x) = E(y \mid x) = \int_{-\infty}^{\infty} yp(y \mid x)\mathrm{d}y$$

这是y关于x的(总体)条件期望.

拓展: 一般化的线性回归模型的表达

无论y与x是否具有线性关系, 我们永远可以将y分解为

$$y = \mathrm{E}(y \mid \mathbf{x}) + \varepsilon$$
$$\mathrm{E}(\varepsilon \mid \mathbf{x}) = 0$$

(证明仅需要对第一个等式做对$\mathbf{x}$的条件期望即可.)
假设条件期望可以写为

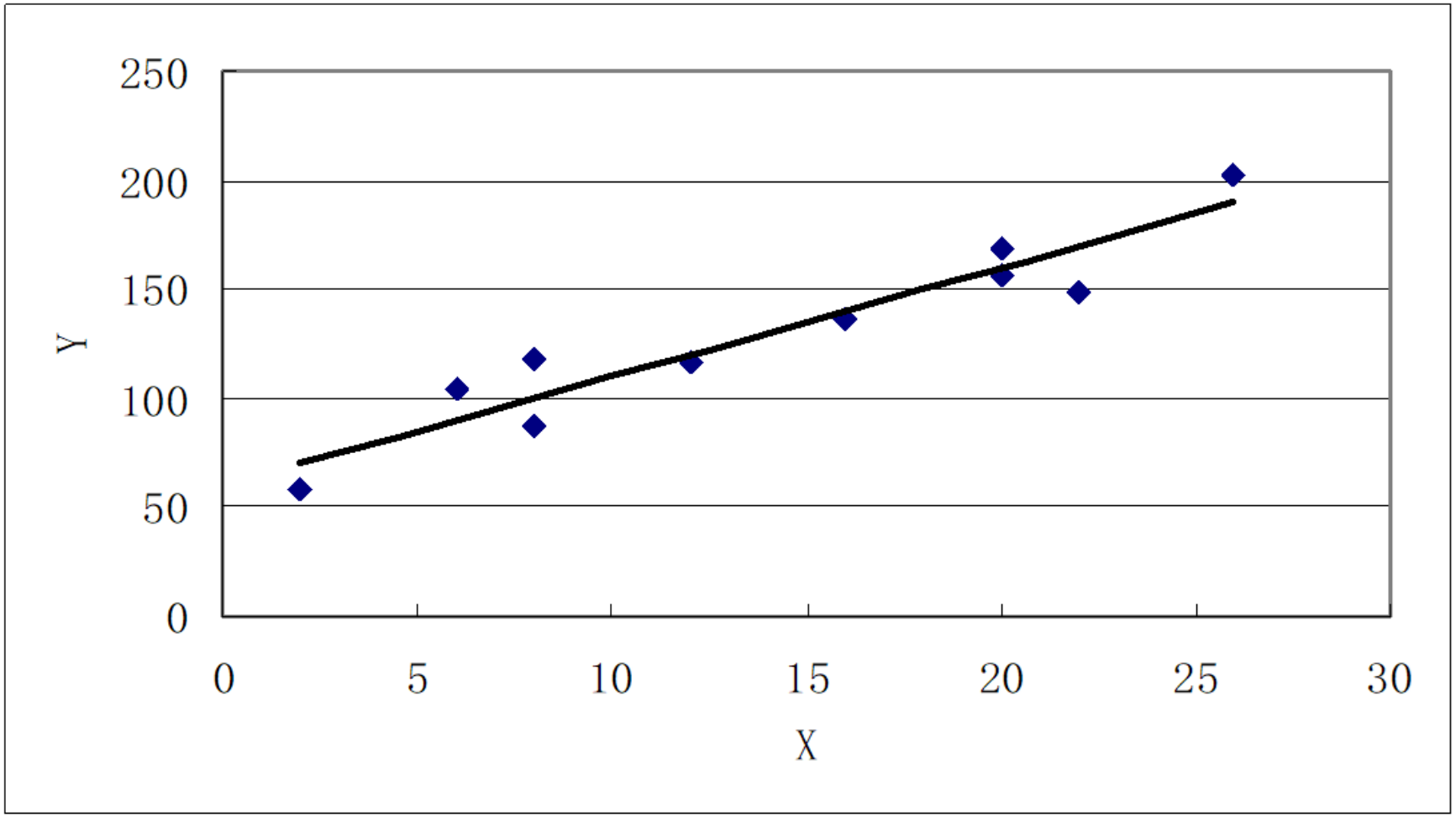
$$\mathrm{E}\left(y \mid x_1, x_2, \ldots\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

则上述分解构成一般化的线性回归模型.
(这个假设较弱, 因为x_2可以是x_1的高阶形式.)

引例: 阿蒙德比萨饼屋的销售预测

连锁店阿蒙德比萨饼屋位于学生人数较多校园旁边的店比位于学生人数较少的校园边上的店有更大的销售额.
为了研究学生人数X(千人)与季度销售额Y(千美元)之间的关系, 采集了10家饭店的数据.

餐厅	Y	X
1	58	2
2	105	6
3	88	8
4	118	8
5	117	12
6	137	16
7	157	20
8	169	20
9	149	22
10	202	26


250
200
150
100
50
0
Y
0
5
10
15
20
25
30
X

本问题的回归模型为$y = a + bx$ 通过数据拟合, 可得到

$$\hat{y} = 60 + 5x.$$

该模型有如下结论

- $b = 5$ (为正), 说明学生人数增加时, 季度销售额便会增加
- 学生人数每增加$1,000$人, 预计销售额会增加$5,000$美元
- 如果要预测一个位于拥有$16,000$个学生的校园边的饭店的季度销售额, 可以预测季度销售额约为$140,000$美元.

2.2 一元线性回归模型

一元线性模型总体模型：

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

对应样本模型：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \cdots, n$$

基本假设(高斯马尔可夫假定)：

- (零均值, 同方差, 正态分布)对于任意的 i:

$$\varepsilon_i \sim N\left(0, \sigma^2\right) \begin{cases} E\left(\varepsilon_i\right) = 0 \\ Var\left(\varepsilon_i\right) = \sigma^2 \end{cases}$$

- (误差项不相关): $\operatorname{Cov}\left(\varepsilon_i, \varepsilon_j\right) = 0, \forall i \neq j$
- x_i**是非随机变量** (便于计算)

拓展知识: 有关假设的进一步说明

上文的假设可被放松: 可假设x_i是**随机变量**.

此时, 对应的简单线性回归其他假设可以调整为 (伍德里奇(第六版))

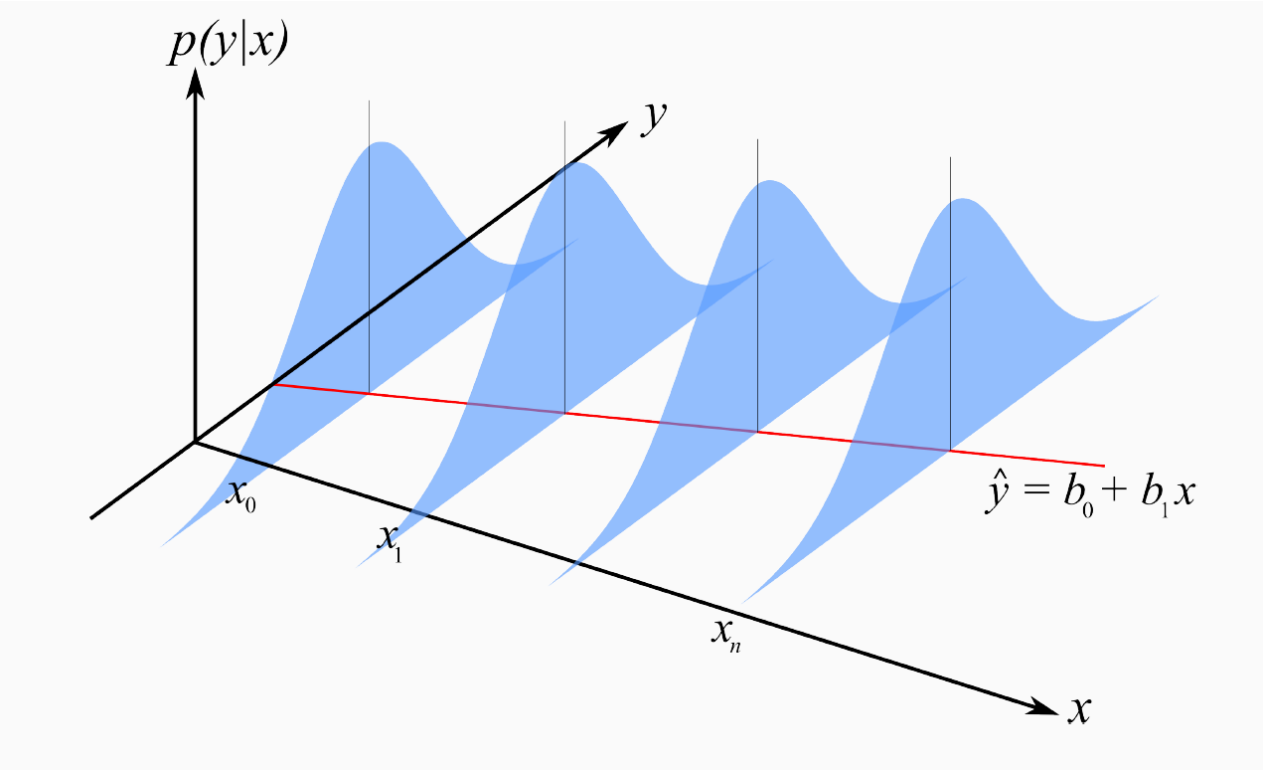
- $E(\varepsilon|x) = 0$ (可推出 x_i与ε_i 不相关: $\mathrm{Cov}\left(\varepsilon_i, x_i\right) = 0$)
- $Var(\varepsilon|x) = \sigma^2$

计量经济学中还将学习各假设被违反时, 如何得出参数无偏或一致估计的方法.

y_i的分布？

如果假设 x_i是**非随机变量**, 则很容易得出

- $E\left(y_i \mid X = x_i\right) = \beta_0 + \beta_1 x_i$
- $\mathrm{Var}\left(y_i \mid X = x_i\right) = \sigma^2$



拓展: 如果假设 x_i是**随机变量**, y_i的分布也可推出同样表达, 伍德里奇(第六版)

一元线性回归分析的问题

一元线性回归分析的问题：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \cdots, n$$

模型中需要估计的参数为 β_0, β_1, σ.

一元线性回归分析的问题

第一： 从样本数据中做出有关总体参数的推断.
从总体中抽取容量为 n 的样本:$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$,
则对应样本模型:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \cdots, n$$

希望估计总体参数: β_0, β_1 估计量:

$$\hat{\beta}_0 \xrightarrow{\text{estimates}} \beta_0, \quad \hat{\beta}_1 \xrightarrow{\text{estimates}} \beta_1$$

第二： 根据参数估计做出关于y的拟合与预策.
一元回归线:

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

$\hat{y}_i$: y_i 的拟合值 (y_i 的预测值).

第三节　最小二乘法 (Ordinary Least Squares, OLS)

如何估计$\hat{\beta}_0 \overset{\text{estimates}}{\rightarrow} \beta_0$, $\hat{\beta}_1 \overset{\text{estimates}}{\longrightarrow} \beta_1$?

定义**残差** (Residuals): $\hat{e}_i = (y_i - \hat{y}_i)$

则残差平方和: SSE (sum of squared error)

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 \rightarrow \min$$

第三节　最小二乘法 (Ordinary Least Squares, OLS)

OLS 估计是找出使上式最小的$\hat{\beta}_0, \hat{\beta}_1$:

$$\arg\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

(注意：教材中使用b_0, b_1 表示 $\hat{\beta}_0, \hat{\beta}_1$, 均是$\beta_0, \beta_1$估计的意思. $\hat{\beta}_0, \hat{\beta}_1$更常用.)

OLS求解-最小化残差平方和

第一步: 求一阶条件 (正则方程组)

$$\frac{\partial\left(\sum \hat{e}_i^2\right)}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\frac{\partial\left(\sum \hat{e}_i^2\right)}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

化简得

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

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

最小二乘估计, OLS estimator

求解得 (推导过程作为课程练习):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{XY}}{S_X^2} = \frac{\text{sample covariance}}{\text{sample variance}}$$

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

斜率为$\hat{\beta}_1$, 截距为$\hat{\beta}_0$.

拓展知识: 最大似然估计

我们的模型为：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \cdots, n$$

其中 $\varepsilon_i \sim N\left(0, \sigma^2\right)$.

我们知道： $y_i \mid X = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.
似然函数?

拓展知识: 最大似然估计

我们的模型为:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \cdots, n$$

其中 $\varepsilon_i \sim N\left(0, \sigma^2\right)$.

我们知道: $y_i \mid X = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.
似然函数? (即y_i的联合密度函数)

拓展知识: 最大似然估计

似然函数

$$
\begin{aligned}
& L\left(\beta_0, \beta_1, \sigma^2\right) \\
= & \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y_i-\beta_0-\beta_1 x_i}{\sigma}\right)^2\right] \\
= & \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)^2\right]
\end{aligned}
$$

拓展知识: 最大似然估计

上述似然函数取对数：对数似然

$$\begin{aligned}&\log L\left(\beta_0, \beta_1, \sigma^2\right)\\ =&-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)^2\end{aligned}$$

通过求导可以最大化对数似然：

$$\begin{aligned}\frac{\partial \log L}{\partial \beta_0} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)\\ \frac{\partial \log L}{\partial \beta_1} &= \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i\left(y_i-\beta_0-\beta_1 x_i\right)\\ \frac{\partial \log L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)^2\end{aligned}$$

拓展知识: 最大似然估计

MLE为

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

与OLS估计相同.

Gauss-Markov 定理

Gauss-Markov 定理: 如果Gauss-Markov 假设成立, 则OLS估计量 $\hat{\beta}_0, \hat{\beta}_1$ 是总体参数 β_0, β_1 的 **线性最小方差无偏估计量** (LMVUE, 或者BLUE).

拓展: 在更高级的课程中, 会学习在Gauss-Markov 假设不成立的条件下, 如何得到无偏或一致的估计.

在Stata 中建立一元线性回归模型

```
reg y x
```

第四节　回归模型评价

问题提出： 样本数据的回归模型总是可以求到的, 但是它是否确实是总体回归模型的正确估计呢?

- 该模型能否较好地解释 y_i 的取值变化规律?

  回归方程的质量如何? 误差多大?
- 关于一元线性回归模型的几个基本假设条件是否得到满足?

  拓展思考： 遗漏变量会违反什么假设?
- 自变量X真的可以解释Y吗?

1. 估计标准误差(standard error): s_e (即残差ε的标准差, 可用于评价模型精度; 可用于计算估计量的方差)

2. 判断拟合优度 (Goodness of fit): 可决系数 R^2

3. 在 X 和Y 之间是否存在线性关系? (F-test, 模型检验)

4. X 在解释Y时, 是否有作用? (t-test, 变量检验)

5. 残差分析: 检验基本假设是否被违反.

二. 估计标准误差 (Standard Error of the Estimate)

估计标准误: 即残差ε的标准差

第一步: 计算残差的均值 (为0)

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

(由OLS一阶条件得出)

第二步: 计算标准误差

$$s_e = \sqrt{s_e^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(\hat{e}_i - 0)^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{SSE}{n-2}}$$

为什么自由度(Degree of Freedom)= n-2

因为在n个残差估计$\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$中, 只有$n-2$个可自由移动. 由正规方程组

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

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

我们有

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

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

标准误差的性质

可以证明, 如果基本假设成立: $\varepsilon_i \sim N\left(0, \sigma^2\right)$, 则

- s_e^2 是总体随机误差方差 σ^2 的无偏估计量: $\mathrm{E}\left(s_e^2\right) = \sigma^2$
- s_e 是总体标准差 σ 的一致估计量 (但一般是有偏估计量): $plim(s_e) = \sigma$

证明过程详见伍德里奇教材. 以下给出证明思路.

拓展知识：证明$\mathrm{E}\left(s_e^2\right)=\sigma^2$

证明： 利用OLS一阶条件, 我们知道:

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

或

$$\sum_{i=1}^{n}\left(\beta_0+\beta_1 x_i+\varepsilon_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)=0.$$

取平均

$$\bar{\varepsilon}-(\hat{\beta}_0-\beta_0)-(\hat{\beta}_1-\beta_1)\bar{x}=0.$$

将该平均值
从$\hat{\varepsilon}_i=y_i-\hat{\beta}_0-\hat{\beta}_1 x_i=\varepsilon_i-(\hat{\beta}_0-\beta_0)-(\hat{\beta}_1-\beta_1)x_i$减去可得

$$\hat{\varepsilon}_i=(\varepsilon_i-\bar{\varepsilon})-\left(\hat{\beta}_1-\beta_1\right)(x_i-\bar{x}).$$

拓展知识：证明$\mathrm{E}\left(s_e^2\right)=\sigma^2$

所以, $\hat{\varepsilon}_i^{\,2}=(\varepsilon_i-\bar{\varepsilon})^2+\left(\hat{\beta}_1-\beta_1\right)^2(x_i-\bar{x})^2-2\left(\varepsilon_i-\bar{\varepsilon}\right)\left(\hat{\beta}_1-\beta_1\right)(x_i-\bar{x})$.

对所有i加总可得

$\sum_{i=1}^n\hat{\varepsilon}_i^{\,2}=\sum_{i=1}^n(\varepsilon_i-\bar{\varepsilon})^2+\left(\hat{\beta}_1-\beta_1\right)^2\sum_{i=1}^n(x_i-\bar{x})^2$

$-2\left(\hat{\beta}_1-\beta_1\right)\sum_{i=1}^n\varepsilon_i(x_i-\bar{x})$

- 上式第一项期望等于 $(n-1)\sigma^2$
- 上式第二项期望等于 σ^2. 因为 $E\left[\left(\hat{\beta}_1-\beta_1\right)^2\right]=\mathrm{Var}\left(\hat{\beta}_1\right)=\sigma^2/s_x^2$.
- 上式第三项可写为 $2\left(\hat{\beta}_1-\beta_1\right)^2s^2$, 期望是$2\sigma^2$.

所以 $E\left(\sum_{i=1}^n\hat{\varepsilon}_i^{\,2}\right)=(n-1)\sigma^2+\sigma^2-2\sigma^2=(n-2)\sigma^2$, 所以$E[SSE/(n-2)]=\sigma^2$.

例题2: y_i-股票价格(\$); x_i-股息(\$)

股票	股息 (\$)	股价 (\$)
i	x_i	y_i
1	13	115
2	4	45
3	12	100
4	5	50
5	6	55
6	8	85
7	3	40
8	4	50
9	5	45
10	7	70

例：股价与股息之间的关系

解： 计算结果为

$$\hat{\beta}_0 = 15.2017, \quad \hat{\beta}_1 = 7.5072$$

$$\hat{y} = 15.2017 + 7.5072x$$

当 $x = \$13 \quad (y = 115.0)$,

$\hat{y} = 15.2017 + 7.5072 \times 13 = 112.80$

$\hat{e} = 115.00 - 112.80 = 2.20$

例：股价与股息之间的关系

估计标准误差

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 255.6196$$
$$s_e^2 = \frac{SSE}{n-2} = \frac{255.6196}{10-2} = 31.9525$$
$$s_e = \sqrt{s_e^2} = \sqrt{31.9525} = 5.6526$$

比较不同的模型时, 估计标准误差越小的模型, 精度越高

标准误差

可以推导出(证明见Woodridge Chapter 2)：

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right)\right)$$
$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right)$$

但是 σ 未知, 所以需要用 s_e 代替:

$$\mathrm{SE}\left[\hat{\beta}_0\right] = s_e\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}}$$
$$\mathrm{SE}\left[\hat{\beta}_1\right] = \frac{s_e}{\sqrt{s_{xx}}}$$

标准误差

从而有：

$$\frac{\hat{\beta}_0 - \beta_0}{\mathrm{SE}\left[\hat{\beta}_0\right]} \sim t_{n-2}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}\left[\hat{\beta}_1\right]} \sim t_{n-2}$$

β_0 和 β_1 的置信区间

β_0 的置信区间：

$$\hat{\beta}_0 \pm t_{1-\alpha/2}(n-2) \cdot \mathrm{SE}\left[\hat{\beta}_0\right]$$

β_1 的置信区间：

$$\hat{\beta}_1 \pm t_{1-\alpha/2}(n-2) \cdot \mathrm{SE}\left[\hat{\beta}_1\right]$$

三. 拟合优度 (Goodness of Fit)

评价模型的拟合优度就是评价y的变异(变化)中多大的部分而可以用x解释.

可用**可决系数/测定系数**(Coefficient of Determination)R^2评价.

三. 拟合优度 (Goodness of Fit)

回忆方差分析课程内容, 可以证明离差平方和SST有如下分解式$SST = SSR + SSE$ (平方和分解公式) :

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

- $\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$, 原始数据 y_i 的**总变异平方和**, 其自由度为 $\mathrm{df_T} = n - 1$;
- $\mathrm{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$, **拟合方程可解释的变异平方和**, 其自由度为 $\mathrm{df_R} = p$;
- $\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, **残差平方和**, 自由度为 $\mathrm{df_E} = n - p - 1$
- $\mathrm{df_T} = \mathrm{df_R} + \mathrm{df_E}$

(对应Stata “reg”命令左上角输出)

可决系数/测定系数

SSR越大：用回归方程解释 y_i 变异的部分越多;
SSE越小：观测值y_i 绕回归线越紧密, 拟合越好.
故可定义：

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

1) $0 \leq \mathrm{R}^2 \leq 1$
2) 当 $\mathrm{R}^2 = 1$: $\mathrm{SSR} = \mathrm{SST}$, or $\mathrm{SSE} = 0$
3) 当 $\mathrm{R}^2 = 0$: $\mathrm{SSR} = 0$, or $\mathrm{SST} = \mathrm{SSE}$

R^2性质

4) 在一元回归里, $\mathrm{R}^2 = \hat{\beta}_1^2 \frac{S_X^2}{s_Y^2} \Rightarrow$ if $\hat{\beta}_1 = 0$, then $\mathrm{R}^2 = 0$.

证明:

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

($\sum_{i=1}^{n}(\varepsilon_i)$和$\sum_{i=1}^{n}(\hat{\varepsilon}_i)$ 因为假设和一阶条件均为零)

$$\therefore R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \hat{\beta}_1^2 \frac{s_X^2}{s_Y^2}$$

R^2性质

5) (相关系数和可决系数关系) 在一元回归里,

$$|r(X,Y)| = \sqrt{R^2},$$

$r(\mathrm{X},\mathrm{Y})$ 的 $(\pm)$ 号与 $\hat{\beta}_1$ 相同

证明:

$$\because R^2 = \hat{\beta}_1^2 \frac{S_X^2}{S_Y^2}, \quad \hat{\beta}_1 = \frac{S_{XY}}{S_X^2} \therefore R^2 = \left(\frac{S_{XY}}{S_X^2}\right)^2 \cdot \frac{S_X^2}{S_Y^2} = \frac{S_{XY}^2}{S_X^2 S_Y^2} = [r(X,Y)]^2$$

例题2：股价与股利之间的关系

回归方程:

$$\hat{y}_i = 15.2017 + 7.5072 x_i$$

$$\mathrm{SST} = 6122.5, \quad \mathrm{SSE} = 255.6196$$

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{255.6196}{6122.5} = 0.9582$$

$$r(x, y) = 0.9789$$

课堂练习

1. $R^2 = 0.64$, X 与 Y 的相关系数等于
(a)0.64
(b)0.80
(c)0.32
(d)0.40

课堂练习

1. $R^2 = 0.64$, X 与 Y 的相关系数等于
(a)0.64
(b)0.80
(c)0.32
(d)0.40
答案： (b)

2. 指出下面哪一个方程一定是错误的
(a)$\hat{y} = 500 + 0.01x \quad r = 0.7$
(b)$\hat{y} = -100 + 0.9x \quad r = 0.86$
(c)$\hat{y} = -8 + 3x \quad r = -0.95$

课堂练习

1. $R^2 = 0.64$, X 与 Y 的相关系数等于
(a)0.64
(b)0.80
(c)0.32
(d)0.40
答案： (b)

2. 指出下面哪一个方程一定是错误的
(a)$\hat{y} = 500 + 0.01x \quad r = 0.7$
(b)$\hat{y} = -100 + 0.9x \quad r = 0.86$
(c)$\hat{y} = -8 + 3x \quad r = -0.95$
答案： (c)

第五节　显著性检验

- F-检验 (检验回归模型的线性关系)
- t-检验 (参数显著性检验)

注: 在一元模型中 F-检验和t检验二者等价.

四. F-检验 (检验回归模型的线性关系)

一元线性回归模型:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \cdots, n$$

F-test - 在 X 和 Y之间是否存在线性关系?

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$$

F-test :在 X 和 Y之间是否存在线性关系?

检验统计量

$$F = \frac{SSR/1}{SSE/n-2}$$

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

当H_0成立, $F \sim F(1, n-2)$.

可选显著性水平$\alpha = 0.05$

若 $F > F_{1-\alpha}(1, n-2)$, 拒绝 H_0 （不能否定线性模型）
若 $F \leq F_{1-\alpha}(1, n-2)$, 不拒绝$H_0$ (非线性模型或换变量)

教材错误

教材错误: pp282, 三处$F_{\alpha}(p, n-p-1)$均应为$F_{1-\alpha}(p, n-p-1)$, 表示自由度为$(p, n-p-1)$的F分布的$1-\alpha$分位数, 而非α分位数!

例题2：股价与股利之间的关系

$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$

$n = 10$

$\hat{y}_i = 15.2017 + 7.5072 x_i$

$\text{SST} = 6122.5, \text{SSE} = 255.6, \text{SSR} = 5866.9$

$F = \frac{5866.9/1}{255.6/(10-2)} = 183.61$

$\alpha = 0.05 \quad \Rightarrow F_{1-0.05}(1, n-2) = 5.32$

$F = 183.61 > F_{1-\alpha} = 5.32$

拒绝 H_0. (通过F检验)

Stata实现

```
clear
cd "D:\OneDrive - Washington State University (email.wsu.edu)\teaching\applied statistics\202
用统计学课件\chapter B1-线性回归分析"

use "股价-股息.dta"

reg y x
```

```
      Source |       SS           df       MS      Number of obs   =        10
-------------+----------------------------------   F(1, 8)         =    183.61
       Model |   5866.8804         1   5866.8804   Prob > F        =    0.0000
    Residual |  255.619597         8  31.9524496   R-squared       =    0.9582
-------------+----------------------------------   Adj R-squared   =    0.9530
       Total |      6122.5         9  680.277778   Root MSE        =    5.6526

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   7.507205   .5540216    13.55   0.000     6.229628    8.784781
       _cons |   15.20173   4.119925     3.69   0.006     5.701166    24.70229
------------------------------------------------------------------------------
```

五. t检验(回归系数的检验)

t检验-X 对Y 是否有解释作用? $H_0: \beta_1 = 0, \quad H_1: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t(n-2)$$

如果 H_0 为真, 则有检验统计量为:

$$t = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} \sim t(n-2)$$

其中, $s_{\hat{\beta}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$.

可选显著性水平$\alpha = 0.05$

若 $|t| > t_{1-\alpha/2}$, 拒绝 H_0, X 有解释作用

若 $|t| \leq t_{1-\alpha/2}$, 不拒绝 H_0, X 没有解释作用

p值为$p = P(|t(n-2)| \geq |t|)$

教材错误

教材错误: pp283, 三处$t_{\alpha/2}(n-p-1)$均应为$t_{1-\alpha/2}(n-p-1)$, 表示自由度为$(n-p-1)$的t分布的$1-\alpha/2$分位数, 而非$\alpha/2$分位数!

Stata实现

```
clear
cd "D:\OneDrive - Washington State University (email.wsu.edu)\teaching\applied statistics\202
用统计学课件\chapter B1-线性回归分析"

use "股价-股息.dta"

reg y x
```

```
      Source |       SS           df       MS      Number of obs   =        10
-------------+----------------------------------   F(1, 8)         =    183.61
       Model |   5866.8804         1   5866.8804   Prob > F        =    0.0000
    Residual |  255.619597         8  31.9524496   R-squared       =    0.9582
-------------+----------------------------------   Adj R-squared   =    0.9530
       Total |      6122.5         9  680.277778   Root MSE        =    5.6526

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   7.507205   .5540216    13.55   0.000     6.229628    8.784781
       _cons |   15.20173   4.119925     3.69   0.006     5.701166    24.70229
------------------------------------------------------------------------------
```

模型诊断

一元线性回归模型满足假设吗?

如果不满足假设, 回归结果可能出现多种问题.

我们学习几类基础的模型诊断方法: 残差分析; 异方差检验; 正态性; 序列相关

在更高级的课程中, 我们会学习在不满足假设条件时如何得出一致和无偏的估计, 以及估计对应的正确的标准误差.

第六节　残差分析 (residual analysis)

在满足假设的条件下：

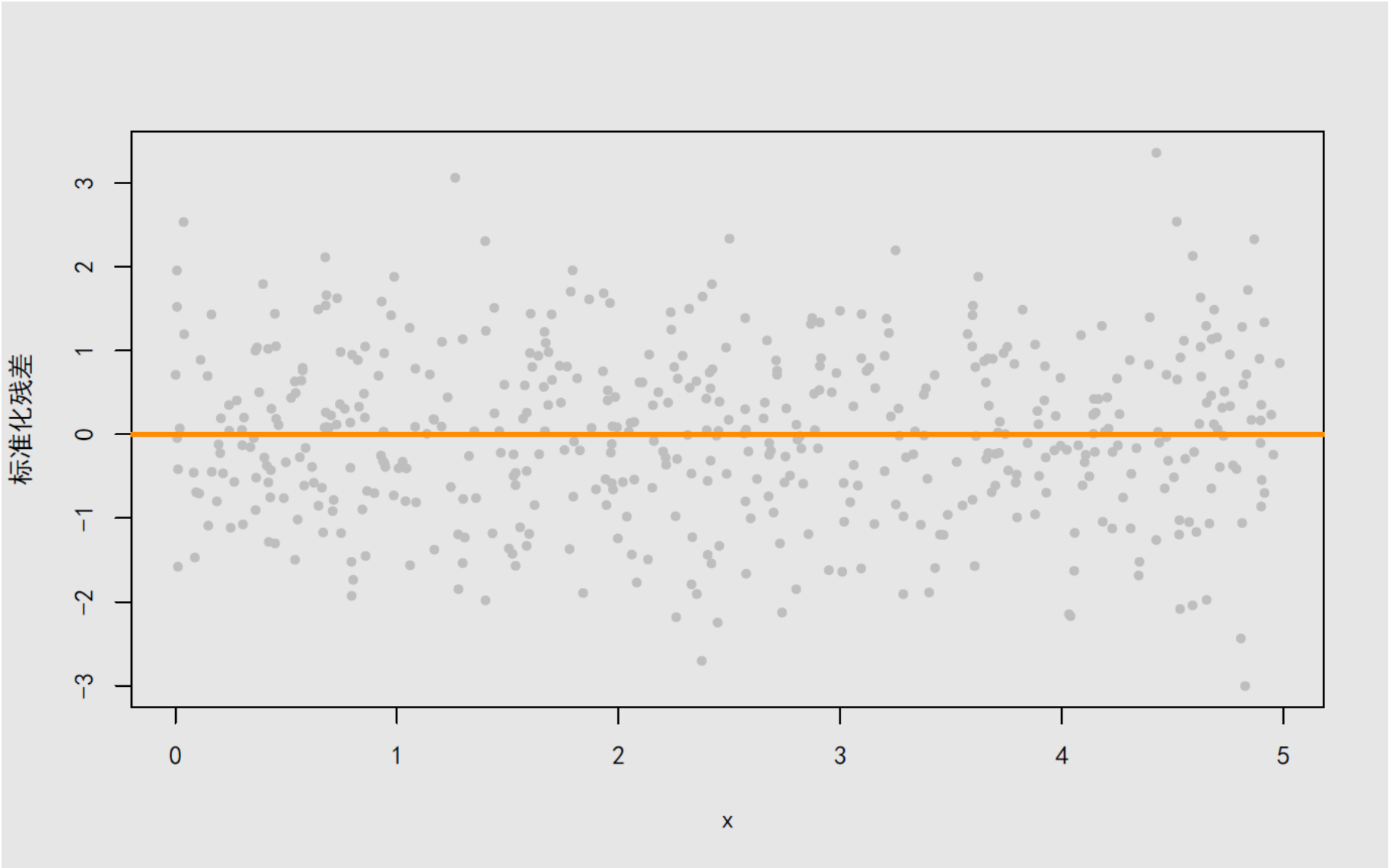
$$\bar{e} = 0 \quad \mathrm{Var}\,(e_i) = s_e^2$$

定义标准化残差 $e_i^* = \frac{e_i - 0}{s_e}$. 当 $n \to +\infty$ 时 , $e_i^* \sim N(0,1)$

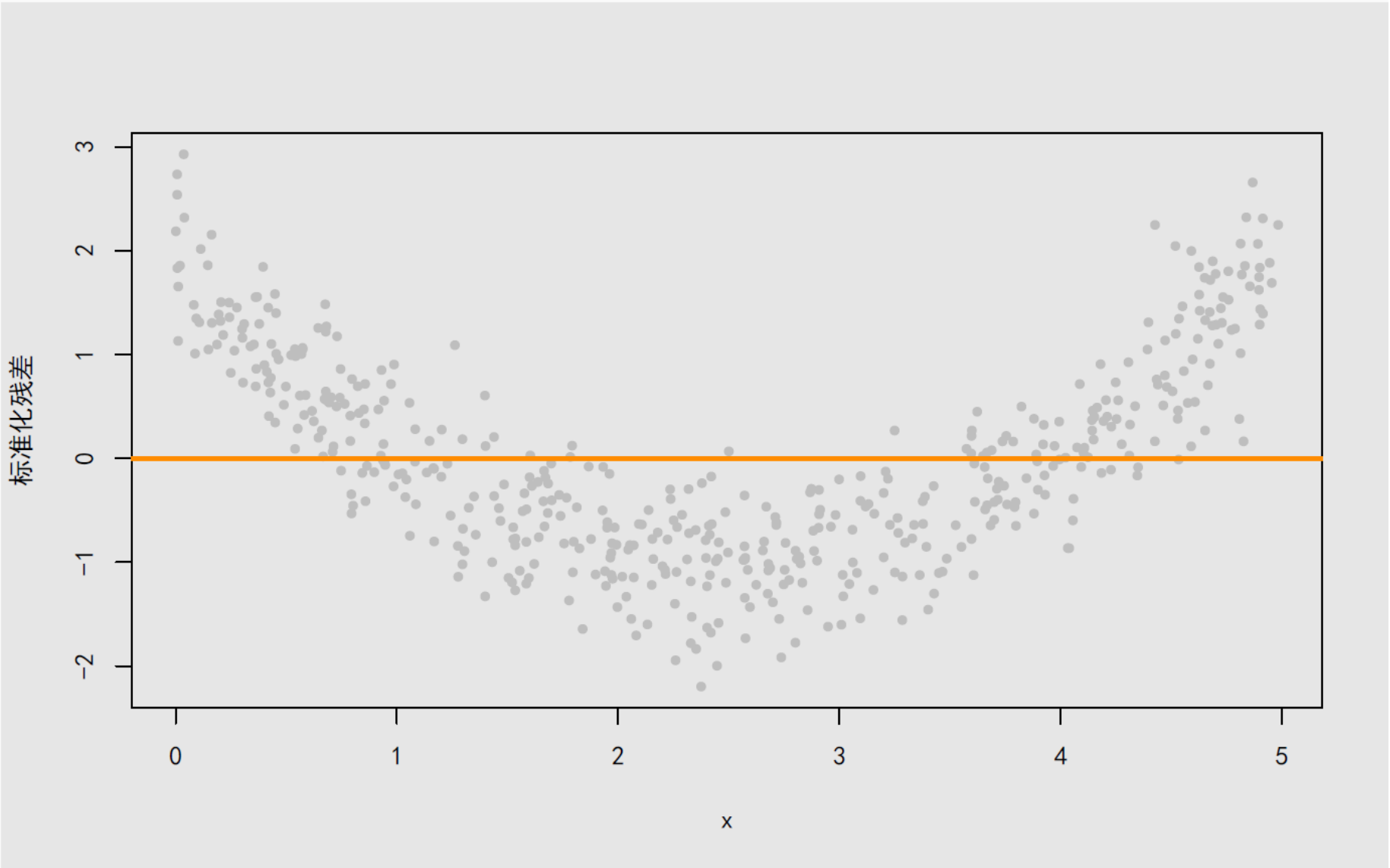
标准化残差图： 以x_i 横坐标, 以e_i^*为纵坐标, 将数据(x_i, e_i^*)标在平面图上.

- (1)**拟合良好：** 若数据点$(x_i, e_i^*), i = 1, 2, \ldots, n,$ 在$(-2, 2)$区间内随机分布, 则说明对总体模型的假设是正确的, 因而推断回归方程的拟合是良好的.
- (2)**拟合不充分：** 若数据点$(x_i, e_i^*), i = 1, 2, \ldots, n$ 排列有规律, 或其中有许多点落在$(-2, 2)$区间之外, 则说明回归方程对数据的拟合不充分,这时随机误差项不再服从正态分布$N(0, \sigma^2)$.
- 原因例如：回归方程的形式选择不当(非线性); 缺少重要的解释变量.

标准化残差图：拟合良好



标准化残差图：拟合不充分



异方差(heteroskedasticity)

"同方差"假设：$\varepsilon_i \sim N\left(0, \sigma^2\right), \forall i = 1, \cdots, n.$

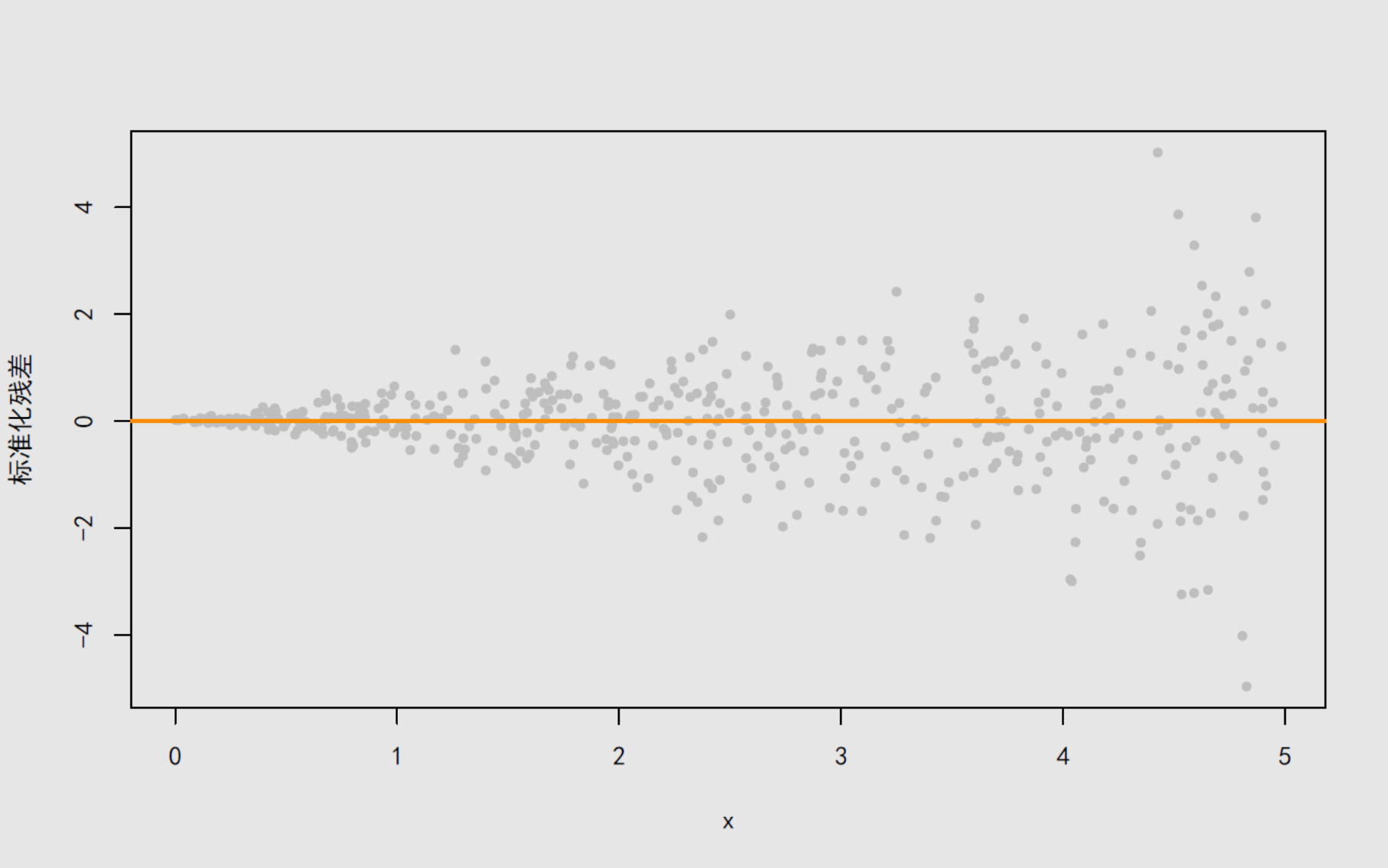
截面数据中较容易出现异方差现象$\varepsilon_i \sim N\left(0, \sigma_i^2\right)$. 例如：储蓄行为的差异随着收入水平而变化.

后果：

- 参数估计方差扩大(仍然无偏, 但不再有效);
- F检验和 t检验失效 (会低估估计量的方差, 得到的t值很高, 对t检验产生误导)

拓展知识： 解决办法: OLS estimation with "robust standard errors"

标准化残差图：异方差



用Stata画残差图和标准残差图

```
clear

use "股价-股息.dta"
reg y x
predict e, residuals
scatter e x
predict standardized_e, rstandard
scatter standardized_e x
```

回归模型残差的正态性检验

- 残差的直方图(Histogram)
- P-P 图(累计概率分布图)(Percentile-Percentile); Q-Q 图(Quantile-Quantile)
- 正态性检验：如 Shapiro-Wilk normality test

序列相关

总体模型中无序列相关假设：$\mathrm{Cov}\left(\varepsilon_i, \varepsilon_j\right) = 0, \forall i \neq j$

1. 序列相关的测量

$\hat{e}_t$	$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_3$	$\hat{e}_4$	$\cdots$	$\hat{e}_n$
$\hat{e}_{t-1}$	—	$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_3$	$\cdots$	$\hat{e}_{n-1}$

2. 序列相关现象产生的原因

- 重要的解释变量被遗漏, 模型函数形式错误
- 时间序列自变量： $Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t$
- 蛛网现象: 当期价格和下期供应量间的关系

3. 序列相关的后果

- (1)估计量的误差范围扩大(不再是有效估计量);
- (2)t检验和F检验不再有效;
- (3)稳健性差：最小二乘估计量对抽样波动变得十分敏感.

4.检查：序列相关现象——残差图 (e_i, e_{i-1})在序列不相关成立的条件下应该表现为随机的点.

有关序列相关在时间序列课程和面板数据课程中会有更多的学习.

拓展内容: 第七节应用回归模型进行预测

当 $X = x_p$, 预测 y_p 的数值.
Point Estimation:

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

Interval Estimation:

$$\hat{y}_p \pm t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

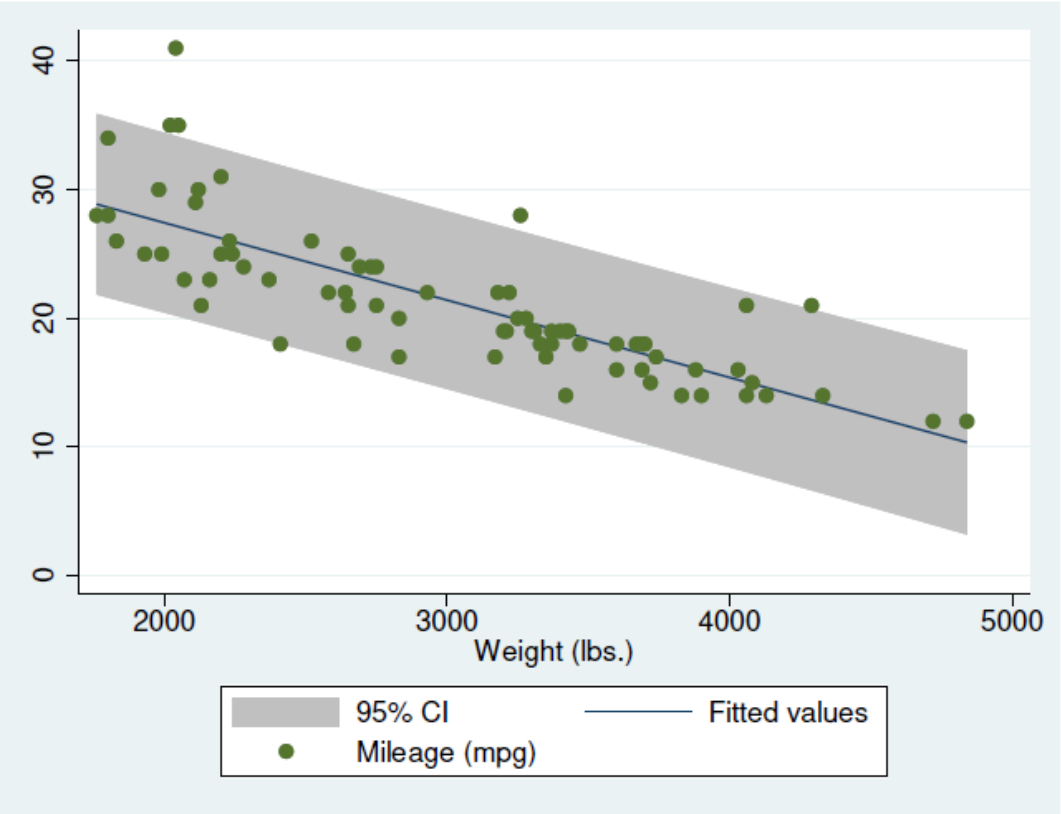
(1)$(x_p - \bar{x}) \uparrow$, C.I. $\uparrow$ (自变量距离平均值越远, 预测越不准确)
(2)$n \uparrow$, C.I. $\downarrow$ (样本容量越大, 预测越准确)
推导过程: Woodridge (Intro)Chapter 6

Stata 中画y的预测值与置信区间

```
help graph twoway lfitci

sysuse auto, clear

twoway lfitci mpg weight, stdf || scatter mpg weight
```



应用统计学II 作业2

题目1. 推导一元线性回归模型$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$的OLS估计量$\hat{\beta}_0, \hat{\beta}_1$. 请写出具体步骤, 无具体步骤不得分.

题目2. 证明上述OLS估计量$\hat{\beta}_0, \hat{\beta}_1$是$\beta_0, \beta_1$的无偏估计量. 请写出具体步骤, 无具体步骤不得分.

上机实验6

题目3. 运用Stata自带数据(课程中心已同步上传)
sysuse nlsw88.dta, U.S. National Longitudinal Study of Young Women (NLSW, 1988 extract)

- 报告样本中受教育年限(grade, 近似代表受教育年限)的分布, 工资分布, 平均受教育年限, 平均工资
- 计算教育年限和工资的Pearson、Spearman、 Kendal相关系数; 请给出最终分析报告, 并说明你将采用哪一种相关系数作为结论, 为什么?
- 建立简单线性模型, 用OLS估计教育对小时工资的影响, 解释模型, 分析经典假设是否被满足.
- 请阐述： F检验统计量的构造原理; t检验统计量的构造原理