

Chapter 01. 经典模型

1.0 阅读材料

1.1 基本概念

1.2 估计方法及其精度

1.3 CLRM和N-CLRM假设以及OLS估
计量性质

1.4 变异分解与拟合优度

1.5 置信区间和假设检验

1.6 t检验和F检验

1.7 回归预测

1.8 一个数值案例

1.9 报告回归分析结果

1.0 测试和准备



测试互动：回归报告

Equation: EQ_M0 Workfile: LONGLEY::employee\				
View Proc Object Print Name Freeze Estimate Forecast Stats Resids				
Dependent Variable: Y				
Method: Least Squares				
Date: Time:				
Sample: 1 16				
Included observations: 16				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	77270.12	22506.71	3.433204	0.0075
X1	1.506187	8.491493	0.177376	0.8631
X2	-0.035819	0.033491	-1.069516	0.3127
X3	-2.020230	0.488400	-4.136427	0.0025
X4	-1.033227	0.214274	-4.821985	0.0009
X5	-0.051104	0.226073	-0.226051	0.8262
X6	1829.151	455.4785	4.015890	0.0030
R-squared	0.995479	Mean dependent var	65317.00	
Adjusted R-squared	0.992465	S.D. dependent var	3511.968	
S.E. of regression	304.8541	Akaike info criterion	14.57718	
Sum squared resid	836424.1	Schwarz criterion	14.91519	
Log likelihood	-109.6174	Hannan-Quinn criter.	14.59449	
F-statistic	330.2853	Durbin-Watson stat	2.559488	
Prob(F-statistic)	0.000000			

郎利数据的OLS回归结果



测试互动2：计量检验

Equation: EQ_M0 Workfile: CAR::mileage\

Heteroskedasticity Test: White

F-statistic	7.506997	Prob. F(14,66)	0.0000
Obs*R-squared	49.75474	Prob. Chi-Square(14)	0.0000
Scaled explained SS	85.28402	Prob. Chi-Square(14)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 04/12/18 Time: 22:28
Sample: 1 81
Included observations: 81

Variable	Coefficient
C	12784.67
X2^2	0.988388
X2*X3	-0.615805
X2*X4	0.203134
X2*X5	1.619221
X2	-226.3550
X3^2	0.097839
X3*X4	-0.061219
X3*X5	-0.573088
X3	72.41228
X4^2	0.000744
X4*X5	0.145326
X4	-20.63295
X5^2	0.901007
X5	-191.1448

R-squared	0.614256
Adjusted R-squared	0.532432
S.E. of regression	15.81967
Sum squared resid	16517.30
Log likelihood	-330.3015
F-statistic	7.506997
Prob(F-statistic)	0.000000

Table: TAB_WHITE Workfile: CAR::mileage\

Heteroskedasticity Test: White

A	B	C	D	E
1	7.506997	Prob. F(14,66)	0.0000	
3	49.75474	Prob. Chi-Square(14)	0.0000	
5	85.28402	Prob. Chi-Square(14)	0.0000	

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 16393 Time: 0.105
Sample: 1 81
Included observations: 81

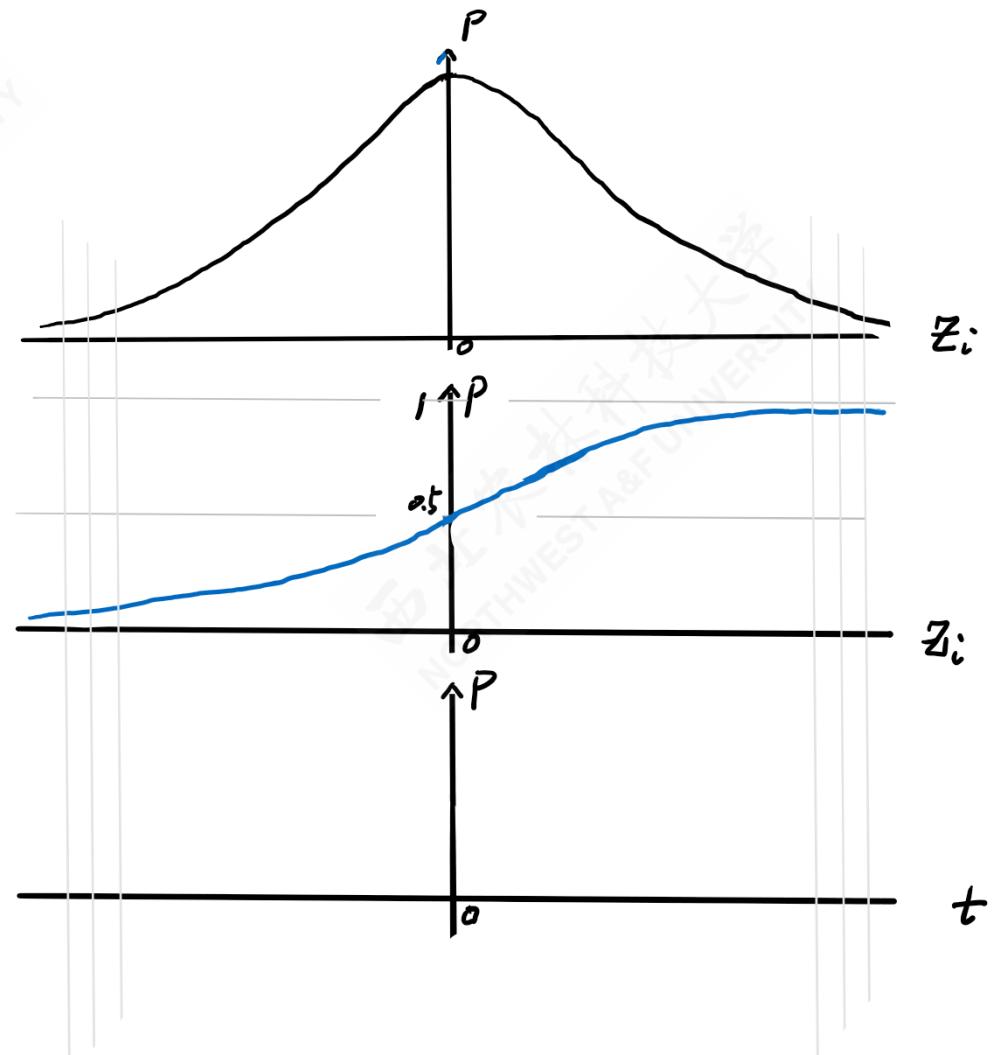
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	12784.67	5865.570	2.179612	0.0329
X2^2	0.988388	0.629866	1.569204	0.1214
X2*X3	-0.615805	0.418737	-1.470625	0.1461
X2*X4	0.203134	0.056687	3.583441	0.0006
X2*X5	1.619221	0.987719	1.639355	0.1059
X2	-226.3550	121.6560	-1.860616	0.0673
X3^2	0.097839	0.069894	1.399826	0.1662
X3*X4	-0.061219	0.018407	-3.325759	0.0014
X3*X5	-0.573088	0.339424	-1.688414	0.0961
X3	72.41228	40.95267	1.768194	0.0816
X4^2	0.000744	0.004538	0.163926	0.8703
X4*X5	0.145326	0.053942	2.694116	0.0089
X4	-20.63295	5.507167	-3.746564	0.0004
X5^2	0.901007	0.441951	2.038704	0.0455
X5	-191.1448	95.74529	-1.996389	0.0500

R-squared	0.614256	Mean dependent var	11.65130
Adjusted R-squared	0.532432	S.D. dependent var	23.13529
S.E. of regression	15.81967	Akaike info criterion	8.525962
Sum squared resid	16517.30	Schwarz criterion	8.969378
Log likelihood	-330.3015	Hannan-Quinn criter.	8.703866

Path = d:\econometrics\reviews DB = none WFE = car



测试互动3：概率问题





测试互动4：数据问题

海关总署 贸易数据

← → ⌂ ▲ 不安全 | customs.gov.cn/customs/302249/302274/302277/index.html

★ Bookmarks 业余学习库 农业专业库 社科专业库 经济思想史 课题网站 每日文献 搜索引擎库 大学 计量经济学 R relative 西北农林科技大学 经济管理学院 杨凌高新区小学 Go

年份 2020年 2019年 2018年 2017年 2016年 2015年 2014年

统计制度

统计快讯

统计月报 →

数据公布特殊标准

进出口监测预警

关区统计

统计服务指南

数据在线查询

人民币 美元

表名 (美元)	月份
(1)2020年进出口商品总值表 A:年度表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(1)2020年进出口商品总值表 B:月度表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(2)2020年进出口商品国别(地区)总值表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(3)2020年进出口商品构成表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(4)2020年进出口商品类章总值表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(5)2020年进出口商品贸易方式总值表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(6)2020年出口商品贸易方式企业性质总值表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(7)2020年进口商品贸易方式企业性质总值表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(8)2020年进出口商品收发货人所在地总值表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(9)2020年进出口商品境内目的地/货源地总值表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(10)2020年进出口商品关别总值表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(11)2020年特定地区进出口总值表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(12)2020年外商投资企业进出口总值表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(13)2020年出口主要商品量值表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(14)2020年进口主要商品量值表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(15)2020年对部分国家(地区)出口商品类章金额表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月
(16)2020年自部分国家(地区)进口商品类章金额表	1-2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月



测试互动4：数据问题

海关总署 贸易数据

(15) 2020年4月对部分国家(地区)出口商品类章金额表(人民币值)

发布时间: 2020-05-23 11:07

文章来源: 海关总署

【字体: 大 中 小】

分享到:

类章	缅甸		中国香港		印度		4,
	4月	1至4月	4月	1至4月	4月	1至4月	
总值	631,233	2,537,207	15,235,179	50,171,440	2,252,300	13,034,862	2,560,
第1类 活动物；动物产品	3,407	12,146	129,381	425,670	89	662	10,
01章 活动物	-	1	31,507	84,197	-	6	
02章 肉及食用杂碎	-	9	29,750	104,328	-	-	
03章 鱼及其他水生无脊椎动物	-	-	50,004	176,116	67	232	7,
04章 乳；蛋；蜂蜜；其他食用动物产品	7	381	11,613	35,878	-	7	
05章 其他动物产品	3,400	11,756	6,506	25,150	22	417	3,
第2类 植物产品	11,693	96,024	160,303	481,269	4,998	30,721	79,
06章 活植物；茎、根；插花、簇叶	423	2,401	547	6,870	-	775	
07章 食用蔬菜、根及块茎	1,805	7,599	83,105	246,565	2,849	13,645	45,
08章 食用水果及坚果；甜瓜等水果的果皮	3,975	70,610	18,809	57,007	3	2,032	14,
09章 咖啡、茶、马黛茶及调味香料	4,936	8,124	33,838	98,012	138	1,291	3,
10章 谷物	-	-	1,057	3,645	-	-	
11章 制粉工业产品；麦芽；淀粉等；	366	3,612	4,641	12,565	247	1,895	9,
12章 油籽；子仁；工业或药用植物；饲料	158	3,488	13,251	37,427	222	3,086	2,
13章 虫胶；树胶、树脂及其他植物液、汁	31	190	4,337	17,213	1,524	7,949	3,



测试互动4：数据问题

月度出口贸易数据形态 |

year	month	currency	cat	country	period	value	files	source	data_
2016	8	千美元	08 章	菲律宾	A	10873	2016- 08- D.html	custom	expd
2015	3	千美元	20 章	罗马尼 亚	B	3416	2015- 03- D.html	custom	expd
2015	6	千美元	10 章	缅甸	A		2015- 06- D.html	custom	expd
2020	4	千美元	19 章	西班牙	B	2170	2020- 04- D.html	custom	expd

Showing 1 to 4 of 1,000 entries

Previous

1 | 2 | 3 | 4 | 5 | ... | 250 | Next



测试互动4：数据问题

月度出口贸易数据形态2

year	month	currency	cat	country	files	source	data_set	A_dollar	B_dollar
2018	4	千美元	03 章	意大利	2018-04-D.html	custom	export	9797	100752
2020	3	千美元	03 章	泰国	2020-03-D.html	custom	export	100752	8350
2015	12	千美元	03 章	比利时	2015-12-D.html	custom	export	8350	3
2016	10	千美元	03 章	瑞士	2016-10-D.html	custom	export	3	100752

Showing 1 to 4 of 1,000 entries

Previous



阅读材料

材料1：计量经济学的语言 [链接1](#)

材料2：计量经济学的那点调性和套路 [链接1](#)



自学材料

西北农林科技大学
NORTHWEST A&F UNIVERSITY

1.1 基本概念

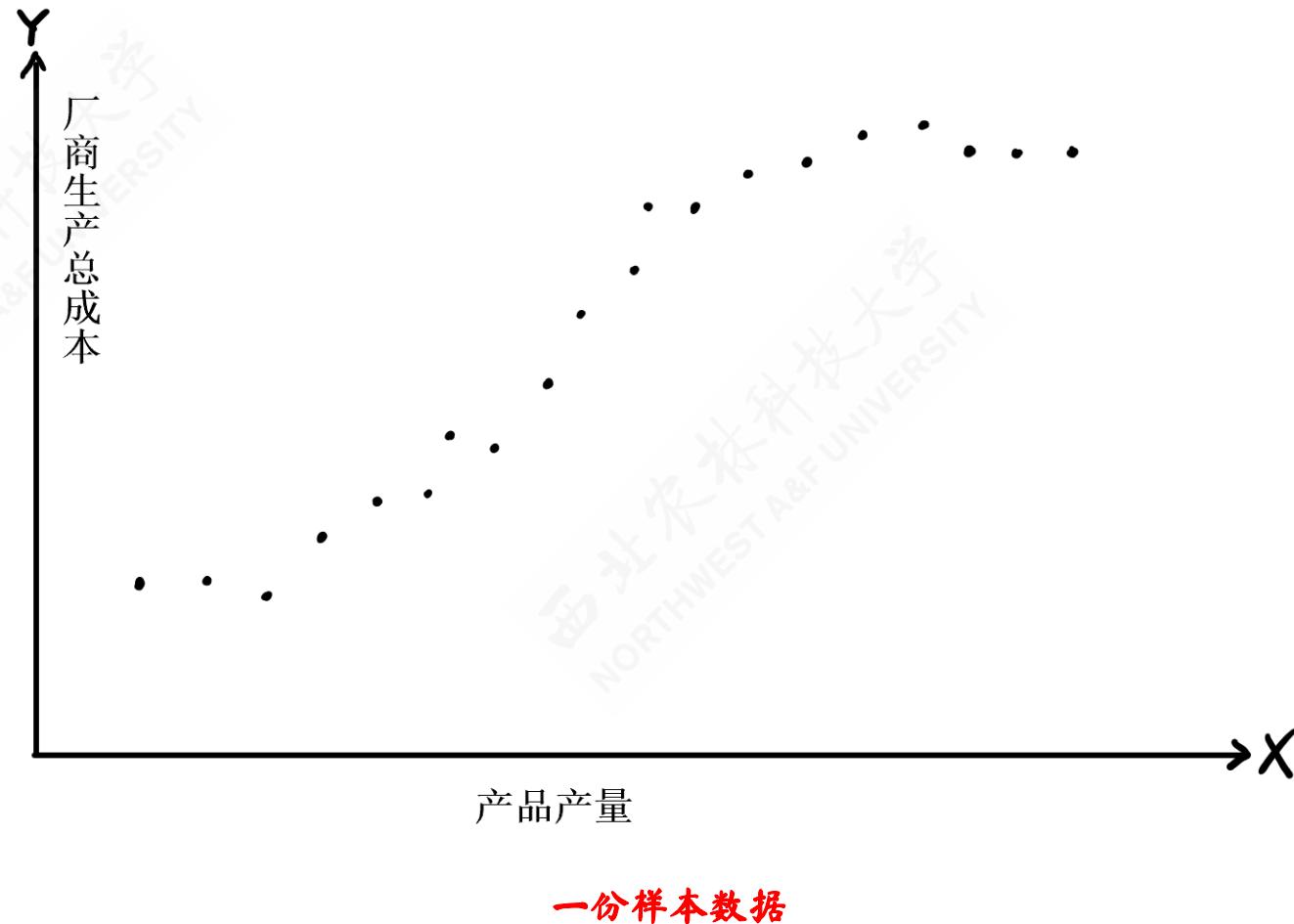


计量经济分析的基本过程（概览）



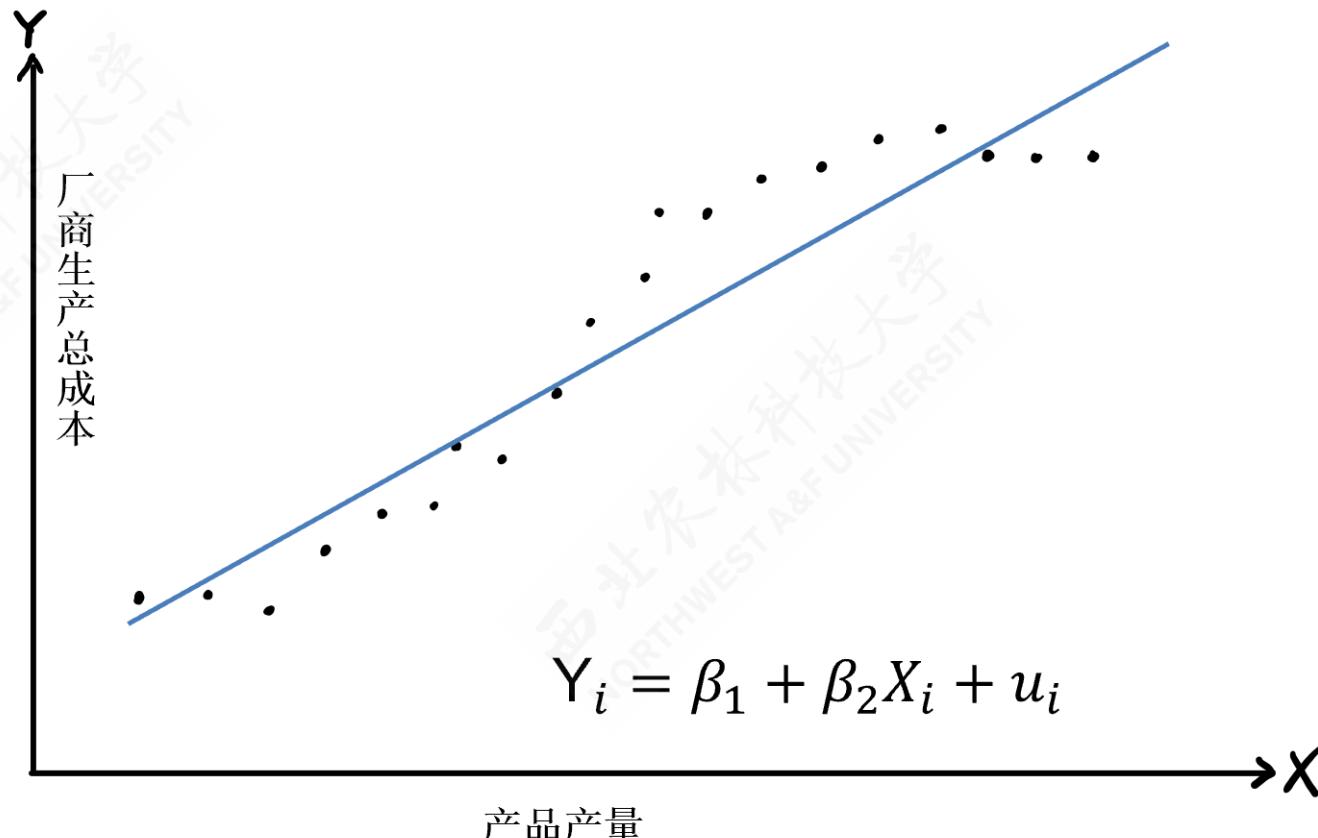


计量模型的数学形式[图片演示]





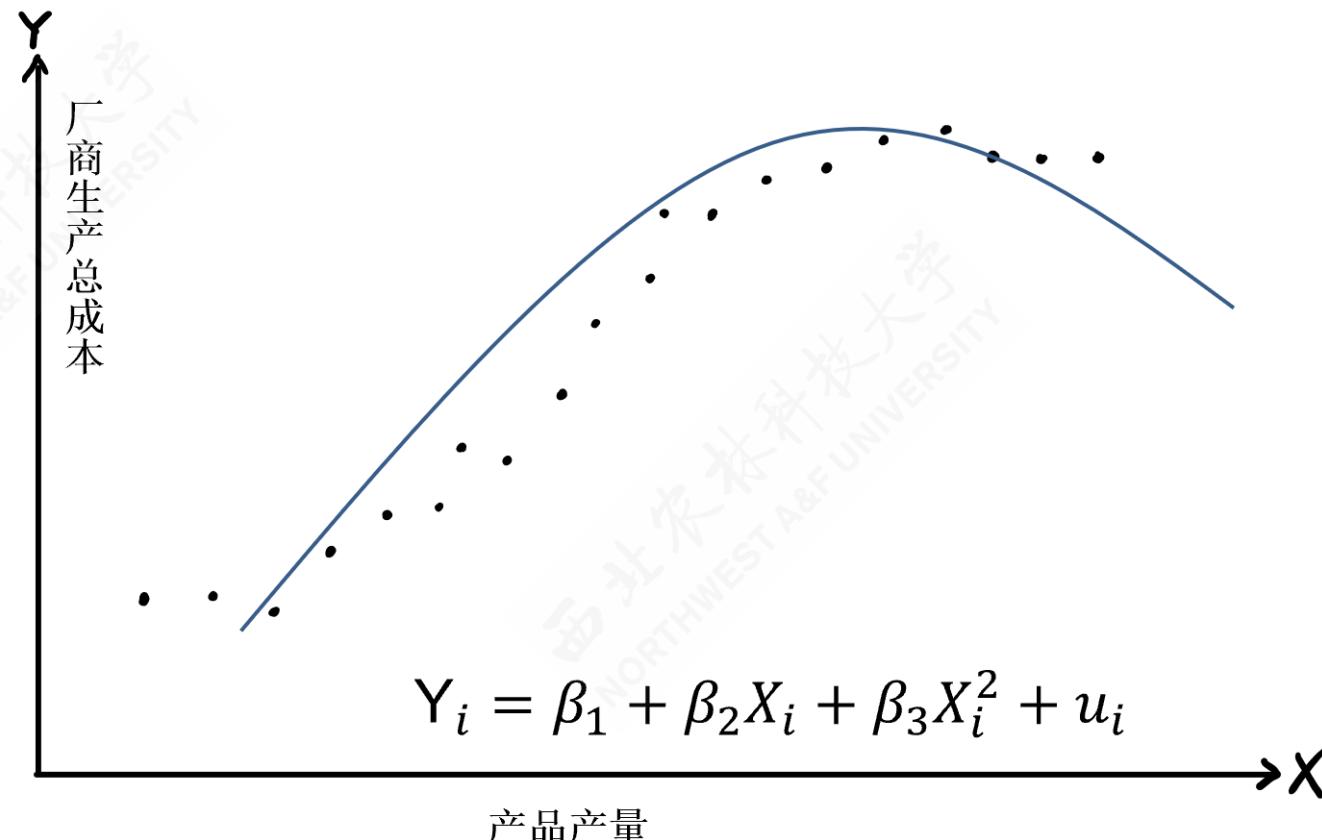
计量模型的数学形式[图片演示]



同学的视界：一条直线



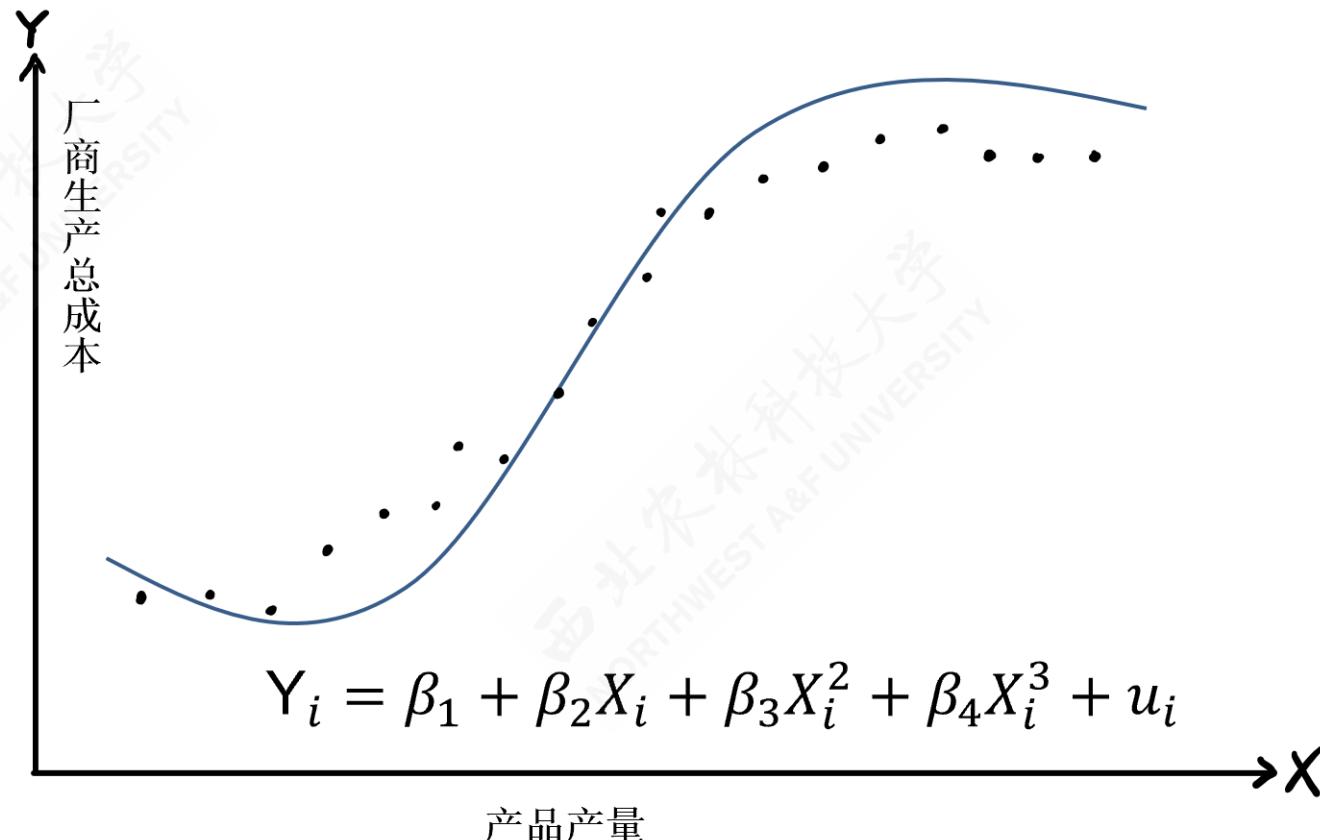
计量模型的数学形式[图片演示]



B同学的视界：一条抛物线



计量模型的数学形式[图片演示]



0同学的视界：一条S型曲线



术语符号

x 和 y 的各种术语约定

Y

- 应变量(Dependent variable)
- 被解释变量(Explained variable)
- 预测子(Predictand)
- 回归子(Regressand)
- 响应变量(Response variable)
- 内生(Endogenous)
- 结果变量(Outcome)
- 被控变量(Controlled variable)

X

- 解释变量(Explanatory variable)
- 自变量(Independent variable)
- 预测元(Predictor)
- 回归元(Regressor)
- 刺激变量 (Stimulus variable)
- 外生(Exogenous)
- 协变量(Covariate)
- 控制变量(Control variable)



k变量回归 vs k元回归

双变量回归分析(two-variables regression analysis):

| 双变量回归实际上就是一元回归。

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$$

多元回归分析(multiple regression analysis):

| 二元回归，实际上就是3变量回归

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i$$



两套符号表达体系

李子奈的k元回归：

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i$$

古扎拉蒂的k变量回归：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$



重要概念：关于总体 (Population)

总体回归曲线(Population Regression Curve, PRC)

条件期望值的轨迹表现为一条曲线(Curve), 或者一条直线 (Line)。

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

它是对总体回归曲线(PRC)的数学函数表现形式。

总体回归模型 (Population Regression model, PRM)

把总体回归函数表达成随机设定形式。

随机干扰项也被称为随机误差项(stochastic error term)

总体回归函数中忽略掉的但又影响着Y的全部变量的替代物。



重要概念：关于样本 (Sample)

样本回归线(Sample Regression Line, SRL)

是通过拟合样本数据得到的一条曲线（或直线）。

样本回归函数(Sample Regression Function, SRF)

是样本回归曲线的数学函数形式，可是是线性的或非线性。

样本回归模型 (Sample Regression Model, SRM)

把样本回归函数表现为“随机”形式。

残差 (Residual)

样本回归函数上Y的拟合值 (fitted value) 与Y的样本观测值 (observed value) 之间的离差。



样本回归与总体回归的比较

总体回归函数PRF:

$$E(Y|X_i) = \beta_1 + \beta_2 X_i \quad (\text{PRF})$$

样本回归函数SRF:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \quad (\text{SRF})$$

总体回归模型PRM:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (\text{PRM})$$

样本回归模型SRM:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \quad (\text{SRM})$$

思考：

- PRF无法直接观测，只能用SRF近似替代
- 估计值与观测值之间存在偏差
- SRF又是怎样决定的呢？

1.2 估计方法及其精度



经典估计方法

经典估计方法主要包括三类：

- 普通最小二乘法(Ordinary least squares, OLS)：朴素的认识论
- 极大似然法 (Maximum likelihood, ML) : 概率的神奇魔力
- 矩估计方法 (Moment method, MM) : 理工男的世界观



极大似然估计法(ML)

极大似然估计法(maximum likelihood, ML):

- 是由Fisher提出的一种参数估计方法基本思想：设总体分布的函数形式已知，但有未知参数 Θ ， Θ 可以取很多值，在 Θ 的一切可能取值中选一个使样本观察值出现的概率为最大的 $\hat{\Theta}$ 值作为 Θ 的估计值，并称估计值 $\hat{\Theta}$ 为参数 Θ 的极大似然估计值。这种求估计量的方法称为极大似然估计法。

似然函数表达式：

- 设总体 X_i 的概率密度函数 $f(X_i; \Theta)$ 为, 其中 Θ 为待估计参数。对于从总体中取得的样本观测值 $(X_1; X_2, \dots, X_n)$, 其联合密度函数为 $\prod f(X_i; \Theta)$, 它是参数 Θ 的函数, 称之为的似然函数, 记为 $L(\Theta)$:

$$L(\Theta) = \prod f(X_i; \Theta)$$



ML估计法与OLS估计法的关系

极大似然估计法(ML)比较复杂，我们仅需知道。在随机干扰项正态性假设下(N-CLRM)：

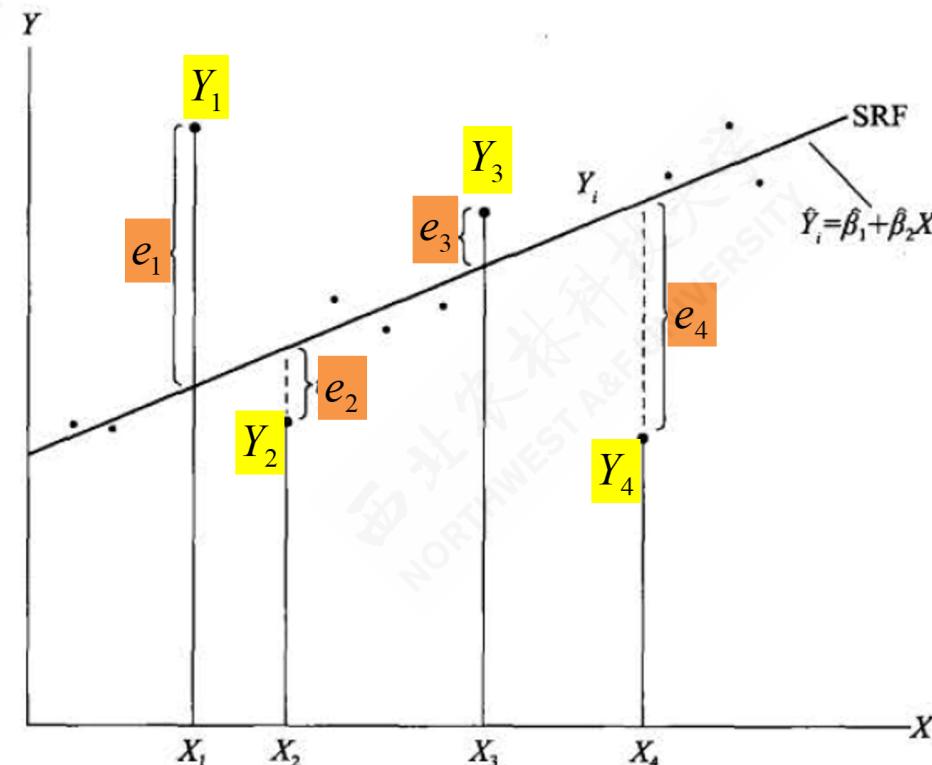
- 回归系数 β_i 的ML估计量和OLS估计量是相同的——无论是一元回归还是多元回归！
- 对于 σ^2 的估计，其ML估计量为 $\sum e_i^2/n$ ，是有偏的；其OLS估计量是 $\sum e_i^2/(n - 2)$ ，是无偏的。
- 关于 σ^2 的两种估计量，随着样本容量n的增大，两者将趋于相等！

启示：OLS方法真好！



普通最小二乘法原理

认识普通最小二乘法的原理：一个图示



最小二乘法的原理



普通最小二乘法原理

总体回归模型PRM:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

样本回归模型SRM:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

OLS的基本原理：用样本推断总体，使残差平方和最小化。

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)$$

$$\begin{aligned} Q &= \sum e_i^2 \\ &= \sum (Y_i - \hat{Y}_i)^2 \\ &= \sum \left(Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i) \right)^2 \\ &\equiv f(\hat{\beta}_1, \hat{\beta}_2) \end{aligned}$$

$$\text{Min}(Q) = \text{Min} \left(f(\hat{\beta}_1, \hat{\beta}_2) \right)$$



回归参数的OLS点估计值

OLS方法下，回归系数的计算公式1（Favorite Five, FF）：

$$\begin{cases} \hat{\beta}_2 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} \\ \hat{\beta}_1 = \frac{n \sum X_i^2 Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2} \end{cases} \quad (\text{FF solution})$$



回归参数的OLS点估计值

OLS方法下，我们也可以得到如下的离差公式(favorite five, ff)

$$\begin{cases} \hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} & (\text{ff solution}) \\ \hat{\beta}_1 = \bar{Y}_i - \hat{\beta}_2 \bar{X}_i \end{cases}$$

其中小写代表离差计算 $x_i = X_i - \bar{X}$; $y_i = Y_i - \bar{Y}$ 。



随机干扰项参数的OLS点估计值

求解残差平方和：

$$\sum e_i^2 = (\hat{\beta}_2 - \beta_2)^2 \sum x_i^2 + \sum (u - \bar{u})^2 - 2(\hat{\beta}_2 - \beta_2) \sum x_i(u - \bar{u})$$

求残差平方和的期望：

$$\begin{aligned} E(\sum e_i^2) &= \sum x_i^2 E[(\hat{\beta}_2 - \beta_2)^2] + E[\sum (u - \bar{u})^2] \\ &\quad + 2E[(\hat{\beta}_2 - \beta_2) \sum x_i(u - \bar{u})] \\ &\equiv A + B + C \\ &= \sigma^2 + (n - 1)\sigma^2 - 2\sigma^2 \\ &= (n - 2)\sigma^2 \end{aligned}$$



随机干扰项参数的OLS点估计值

回归误差方差 (Deviation of Regression Error) :

- 采用OLS方法下，总体回归模型PRM中随机干扰项 u_i 的总体方差的无偏估计量，记为 $E(\sigma^2) \equiv \hat{\sigma}^2$ ，简单地记为 $\hat{\sigma}^2$ 。

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n - 2}$$

回归误差标准差 (Standard Deviation of Regression Error) : 有时候也记为se。

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n - 2}}$$



OLS方法讨论：“估计值”与“估计量”

理解OLS方法下的“估计值”与“估计量”

回归系数的计算公式1 (Favorite Five, FF) :

$$\begin{cases} \hat{\beta}_2 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} \\ \hat{\beta}_1 = \frac{n \sum X_i^2 Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2} \end{cases} \quad (\text{FF solution})$$

- 如果给出的参数估计结果是由一个具体样本资料计算出来的，它是一个“估计值”，或者“点估计”，是参数估计量的一个具体数值；
- 如果把上式看成参数估计的一个表达式，那么，则它是 (X_i, Y_i) 的函数，而 $\$Y_i\$$ 是随机变量，所以参数估计也是随机变量，在这个角度上，称之为“估计量”。



OLS方法下 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的特征

OLS估计量是纯粹由可观测的(即样本)量(指X和Y)表达的，因此它们很容易计算。

它们是点估计量(point estimators)，即对于给定样本，每个估计量仅提供有关总体参数的一个(点)值。[我们以后还将考虑区间估计量(interval Estimators)]

一旦从样本数据得到OLS估计值，便容易画出样本回归线。



OLS方法下SR₂和SR_M的特征

- 特征1：样本回归线一定会经过样本均值点 (\bar{X}, \bar{Y}) :

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$$

- 特征2： Y_i 的估计值(\hat{Y}_i)的均值($\bar{\hat{Y}}_i$)等于Y的样本均值(\bar{Y})

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \\ &= (\bar{Y} - \hat{\beta}_2 \bar{X}) + \hat{\beta}_2 X_i \\ &= \bar{Y} - \hat{\beta}_2 (X_i - \bar{X})\end{aligned}$$

$$\begin{aligned}\Rightarrow 1/n \sum \hat{Y}_i &= 1/n \sum \bar{Y} - \hat{\beta}_2 (X_i - \bar{X}) \\ \Rightarrow \bar{\hat{Y}}_i &= \bar{Y}\end{aligned}$$



OLS方法下SR₂和SR_M的特征

- 特征3：残差的均值(\bar{e}_i)为零：

$$\sum \left[\hat{\beta}_1 - (Y_i - \hat{\beta}_2 X_i) \right] = 0$$

$$\sum \left[Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \right] = 0$$

$$\sum (Y_i - \hat{Y}_i) = 0$$

$$\sum e_i = 0$$

$$\bar{e}_i = 0$$



OLS方法下SRM和SRF的特征

- 特征4：SRM和SRF可以写成离差形式：

$$\left. \begin{array}{l} Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \\ \bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \end{array} \right\} \Rightarrow$$
$$Y_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) + e_i \Rightarrow$$
$$y_i = \hat{\beta}_2 x_i + e_i \quad (\text{SRM-dev})$$

$$\left. \begin{array}{l} \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \\ \bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \end{array} \right\} \Rightarrow$$
$$\hat{Y}_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) \Rightarrow$$
$$\hat{y}_i = \hat{\beta}_2 x_i \quad (\text{SRF-dev})$$



OLS方法下SR₂和SR_M的特征

- 特征5：残差(e_i)和 Y_i 的拟合值(\hat{Y}_i)不相关

$$\begin{aligned} Cov(e_i, \hat{Y}_i) &= E \left[(e_i - E(e_i)) \cdot (\hat{Y}_i - E(\hat{Y}_i)) \right] = E(e_i \cdot \hat{y}_i) \\ &= \sum (e_i \cdot \hat{\beta}_2 x_i) \\ &= \sum \left[(y_i - \hat{\beta}_2 x_i) \cdot \hat{\beta}_2 x_i \right] \\ &= \hat{\beta}_2 \sum \left[(y_i - \hat{\beta}_2 x_i) \cdot x_i \right] \\ &= \hat{\beta}_2 \sum \left[(y_i x_i - \hat{\beta}_2 x_i^2) \right] \\ &= \hat{\beta}_2 \sum x_i y_i - \hat{\beta}_2^2 \sum x_i^2 \\ &= \hat{\beta}_2^2 \sum x_i^2 - \hat{\beta}_2^2 \sum x_i^2 = 0 \end{aligned} \qquad \Leftrightarrow \hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

- 特征6：残差(e_i)和自变量(X_i)不相关



OLS方法下的离差公式总结

- 离差定义与符号：

$$x_i = X_i - \bar{X}$$

$$y_i = Y_i - \bar{Y}$$

$$\hat{y}_i = \hat{Y}_i - \bar{\hat{Y}}_i = \hat{Y}_i - \bar{Y}$$

- PRM及其离差形式：

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_i + u_i \\ \bar{Y} &= \beta_1 + \beta_2 \bar{X} + \bar{u} \end{aligned} \quad \Rightarrow$$
$$Y_i - \bar{Y} = \beta_2 x_i + (u_i - \bar{u}) \Rightarrow$$
$$y_i = \hat{\beta}_2 x_i + (u_i - \bar{u}) \quad (\text{PRM-dev})$$



如何知道OLS方法估计量是否可靠？

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

我们已经使用OLS方法分别得到总体回归模型(PRM)的3个重要参数（实际不止3个）的点估计值：

$$\begin{cases} \hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \\ \hat{\beta}_1 = \bar{Y}_i - \hat{\beta}_2 \bar{X}_i \\ \hat{\sigma}^2 = \frac{\sum e_i^2}{n - 2} \end{cases}$$



如何知道OLS方法估计量是否可靠？

问题是：

- OLS方法的点估计值是否稳定？
- OLS方法的点估计值是否可信？

因此，我们需要找到一种表达OLS方法估计稳定性或估计精度的指标！

在多次抽样下，点估计值会不断变化，成为具有随机分布特征的估计量，其方差（variance）和标准差（standard deviation）就是衡量估计稳定性或估计精度的一类重要指标！



斜率参数的方差和样本方差

斜率系数的总体方差 $\sigma_{\hat{\beta}_2}^2$ 和总体标准差 $\sigma_{\hat{\beta}_2}$:

$$Var(\hat{\beta}_2) \equiv \sigma_{\hat{\beta}_2}^2 = \frac{\sigma^2}{\sum x_i^2}$$
$$\sigma_{\hat{\beta}_2} = \sqrt{\frac{\sigma^2}{\sum x_i^2}}$$

其中， $Var(u_i) \equiv \sigma^2$ 表示随机干扰项 u_i 的总体方差。

斜率系数的样本方差 $S_{\hat{\beta}_2}^2$ 和样本标准差 $S_{\hat{\beta}_2}$:

$$S_{\hat{\beta}_2}^2 = \frac{\hat{\sigma}^2}{\sum x_i^2}$$
$$S_{\hat{\beta}_2} = \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2}}$$

其中， $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$ 表示对随机干扰项 (u_i) 的总体方差 σ^2 的无偏估计量。



截距参数的方差和样本方差

截距系数 ($\hat{\beta}_1$) 的总体方差 ($\sigma_{\hat{\beta}_1}^2$)
和总体标准差 ($\sigma_{\hat{\beta}_1}$) :

$$Var(\hat{\beta}_1) \equiv \sigma_{\hat{\beta}_1}^2 = \frac{\sum X_i^2}{n} \cdot \frac{\sigma^2}{\sum x_i^2}$$
$$\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sum X_i^2}{n} \cdot \frac{\sigma^2}{\sum x_i^2}}$$

- 其中, $Var(u_i) \equiv \sigma^2$ 表示随机干扰项 u_i 的总体方差。

截距系数 ($\hat{\beta}_1$) 的样本方差 ($S_{\hat{\beta}_1}^2$)
和样本标准差 ($S_{\hat{\beta}_1}$) :

$$S_{\hat{\beta}_1}^2 = \frac{\sum X_i^2}{n} \cdot \frac{\hat{\sigma}^2}{\sum x_i^2}$$
$$S_{\hat{\beta}_1} = \sqrt{\frac{\sum X_i^2}{n} \cdot \frac{\hat{\sigma}^2}{\sum x_i^2}}$$

- 其中, $E(\sigma^2) = \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$ 表示对随机干扰项 (u_i) 的总体方差的无偏估计量。



OLS估计方法为何成为经典？

OLS方法小结：

- 普通最小二乘方法（OLS）采用“铅垂线距离平方和最小化”的思想，来拟合一条样本回归线，进而求解出模型参数估计量。
- 拟合样本回归线，并求解出模型参数估计量的方法有很多，还有极大似然估计法（MLE）、据估计方法（MM）等等。不同估计方法代表不同的思想理念，当然各种方法的优劣势差异都需要考虑到“模型环境”（或模型假设）。
- 大家需要很熟练地记住OLS参数估计量公式，以及它们的几大重要特征！



OLS估计方法为何成为经典？

思考与讨论：

- OLS采用的“铅垂线距离平方和最小化”这一方案，凭什么它被奉为计量分析的经典方法？你觉得还有其他可行替代方案么？可以是“垂线距离平方和最小化”么？如果是距离的3次方或4次方之和，又会怎样？距离的绝对值之和可以么？对于这些方案，你有什么想法？
- 回归标准误差 se 的现实含义是什么？回归参数估计与随机干扰项的方差估计有什么内在联系么？
- OLS方法的几个特征，是不是使它“天生丽质”、“娘胎里生下来就含着金钥匙”？为什么能这么说？

1.3 CLRM 和 η -CLRM 假设 以及 OLS 估计量性质



仅仅利用OLS估计方法就足够了么？

我们已经知道OLS方法的原理和基本特征。

问题是：

- OLS方法凭什么能在其他众多拟合估计方法中“脱颖而出”？
- 要跨越“从样本推断总体”的巨大“鸿沟”，仅仅使用OLS方法就足够了么？

答案是：

OLS估计方法，还需要经典线性回归模型（CLRM）假设的加持，二者“双剑合璧”才能真正完成“从样本推断总体”的逻辑证明过程。



CLRM：关于模型的假设

CLRM假设1（模型是正确设置的）：这里大有学问，也是一切计量分析问题的根本来源。

思考：

- 我们怎么知道自己设置的模型是“正确的”？
- 我们有可能知道“正确的”模型么？

CLRM假设2（模型是参数线性的）：模型应该是参数线性的，具体而言模型中参数和随机干扰项必须线性，变量可以不是线性。

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

思考：为什么需要模型是“线性的”？



课堂讨论

以下模型都是线性的：

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i \quad (\text{quadratic polynomial})$$

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i \quad (\text{cubic polynomial})$$

$$Y_i = \beta_1 + \beta_2 \ln(X_i) + u_i \quad (\text{linear-log})$$

$$\ln(Y_i) = \beta_1 + \beta_2 X_i + u_i \quad (\text{log-linear})$$

$$Y_i = \beta_1 + \beta_2 \frac{1}{X_i} + u_i \quad (\text{reciprocal})$$

$$Q_t = AK_t^\alpha L_t^\beta u_t \quad (\text{Cobb-douglas})$$



CLRM：关于自变量 X 的假设

CLRM假设3（自变量 X 是外生的）： X 是固定的（给定的）或独立于误差项。也即自变量 X 不是随机变量。

$$\begin{aligned} \text{Cov}(X_i, u_i) &= 0 \\ E(u_i | X_i) &= 0 \end{aligned}$$

自变量 X 是固定的（给定的）是什么含义？

- 其方差 $\text{Var}(X)$ 是有限的正数。
 - 如 X 取值不能全部相同。如果全部 X 取值都一样，也即 $\text{Var}(X) = 0$ ，则会形成什么样的散点图？
 - 又例如回归系数估计值公式中分母为0，无法求解！
- X 变量没有异常值(outlier)，即没有一个 X 值对于其他值过大或过小。

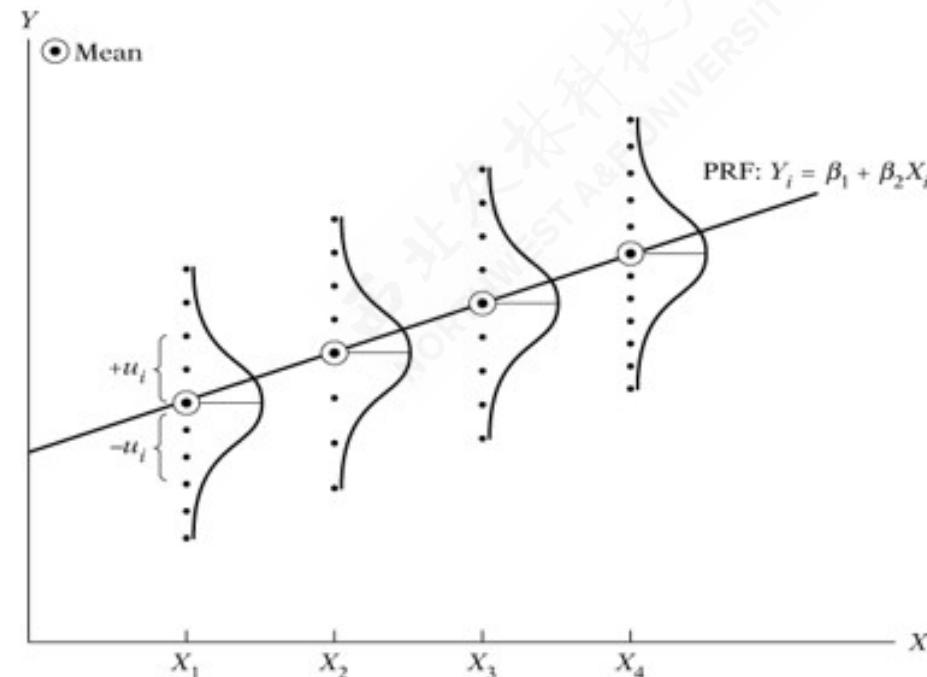
课堂思考



关于随机干扰项的假设1

CLRM假设4（随机干扰项条件期望值为零）：假设随机干扰项条件期望值为零。也即给定 X_i 的情形下，假定随机干扰项 u_i 的条件期望为零。

$$E(u|X_i) = 0$$





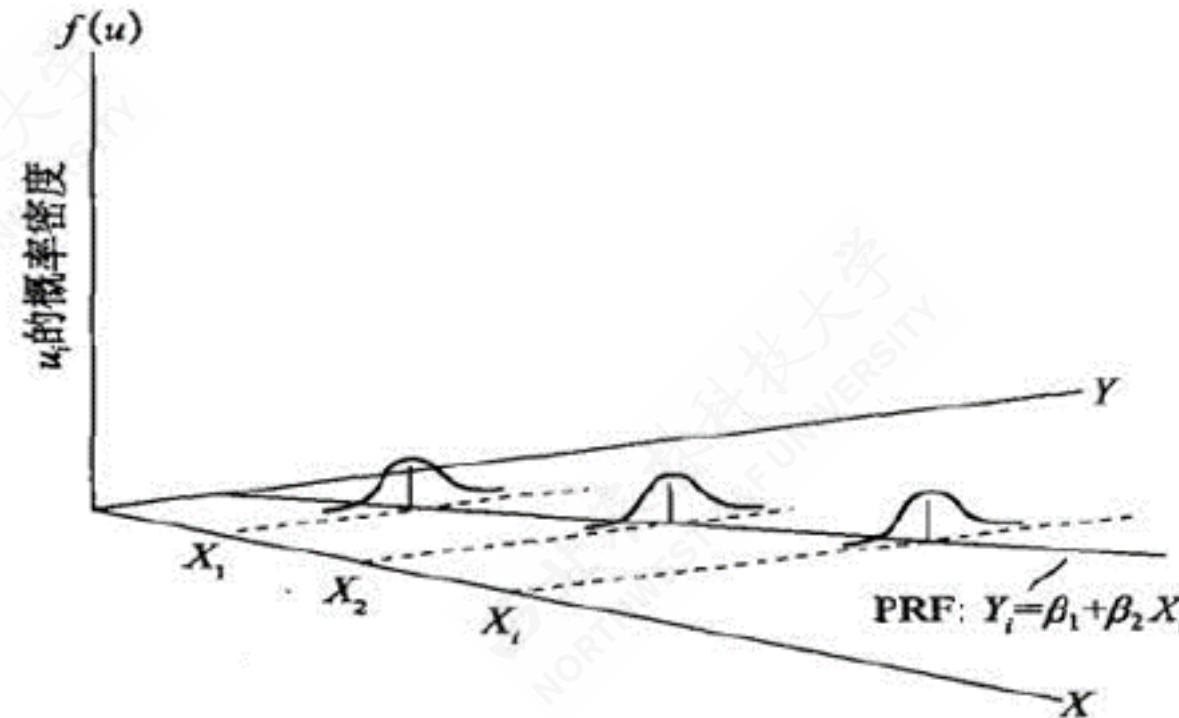
关于随机干扰项的假设2

CLRM假设5（随机干扰项的方差为同方差）：随机干扰项的方差为同方差。也即给定 X_i 的情形下，随机干扰项 u_i 的方差，处处都是相等的。记为：

$$\begin{aligned}Var(u_i|X_i) &= E\left[\left(u_i - E(u_i)\right)^2|X_i\right] \\&= E(u_i^2|X_i) \\&= E(u_i^2) \\&\equiv \sigma^2\end{aligned}$$



随机干扰项的方差为同方差

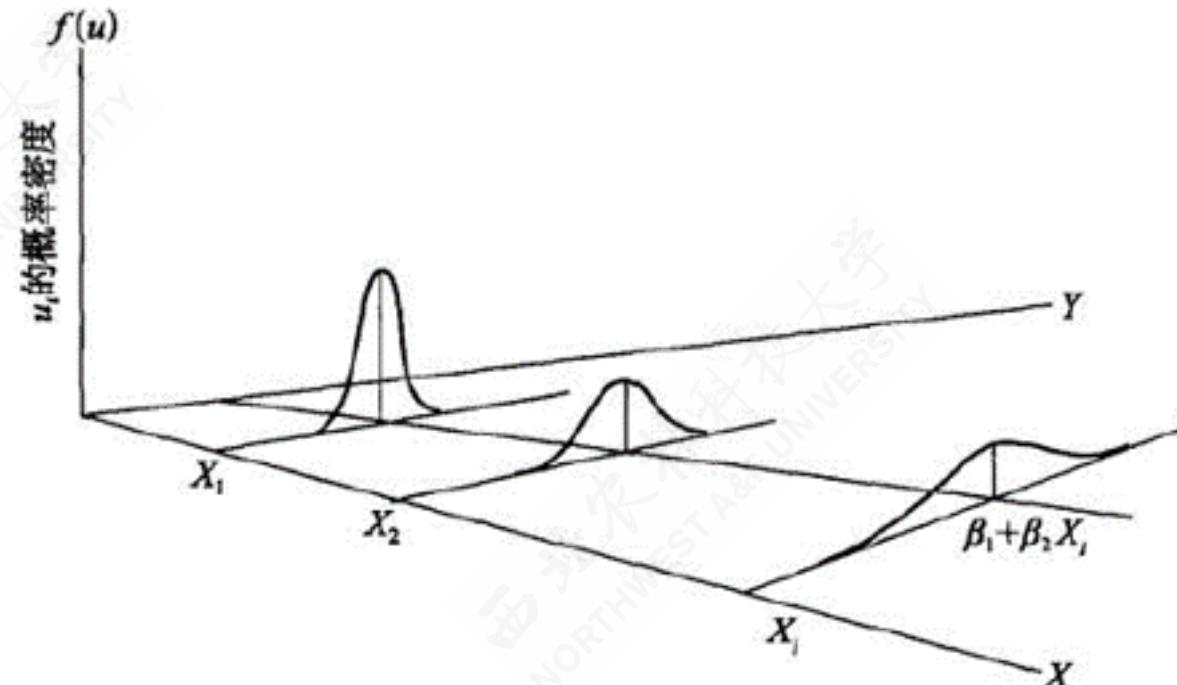


随机干扰项的方差处处相等

- 同方差性(homoscedasticity) : $Var(u_i|X_i) \equiv \sigma^2$



随机干扰项的方差为异方差



随机干扰项的方差随 X 取值不同而不同

- 异方差性(heteroscedasticity) : $Var(u_i|X_i) \equiv \sigma_i^2$



课堂讨论

讨论1: 如果 $Var(u_i|X_1) < Var(u_i|X_2)$, 是否意味着来自 $X = X_1$ 的总体, 相比来自 $X = X_2$ 的总体, 更靠近总体回归线PRL?

讨论2: 如何看待随机样本的质量? 或者, 那些离均值较近的Y总体的随机样本, 与远为分散的Y总体的随机样本, 前者是不是质量更好?

讨论3: 此时, Y_i 的条件方差 $Var(Y_i|X_i)$ 是多少? Y_i 的无条件方差 $Var(Y_i)$ 又是多少?

讨论4: 如果出现异方差, 会对OLS估计产生什么后果?



CLRM：关于随机干扰项的假设3

CLRM假设6（随机干扰项之间无自相关）：各个随机干扰之间无自相关。也即给定两个不同的自变量取值 ($X_i, X_j; i \neq j$) 情形下，随机干扰项 u_i, u_j 的相关系数为0。或者说 u_i, u_j 最好是相互独立的。

在 X_i 为给定情形下，且 $i, j \in (1, 2, \dots, n); i \neq j$ ，记为：

$$\begin{aligned} Cov(u_i, u_j | X_i, X_j) &= E[(u_i - E(u_i))(u_j - E(u_j))] \\ &= E(u_i u_j) \\ &\equiv 0 \end{aligned}$$



CLRM：关于随机干扰项的假设3

重要概念区别：

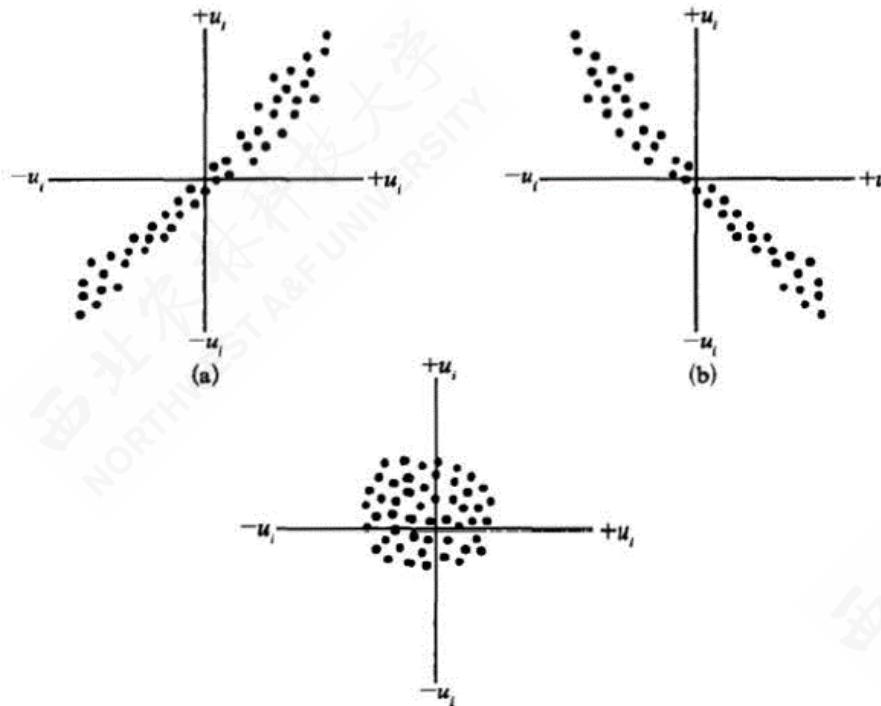
- 无序列相关(no serial correlation):
- 无自相关(no autocorrelation):

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$



课堂讨论



随机干扰项之间的相关情形

- 讨论1：该假设的目的和用处是什么？
- 讨论2：如果出现自相关，会对OLS估计产生什么影响？



CLRM：关于样本数的要求

CLRM假设7（观测样本数假设）： 观测次数n，要大于待估计参数个数。否则方程无法解出，参数不能估计出来。



关于CLRM假设体系的讨论

- 所有这些假设有多真实？

“假定无关紧要论”——弗里德曼

- 上述说有假设都是针对PRF，而不是SRF！

例如：PRF中随机干扰项有无自相关的假设 $Cov(u_i, u_j) = 0$ ；但是在SRF中，可能就会出现 $Cov(e_i, e_j) \neq 0$

- 前面提到的OLS方法正是试图“复制”CLRM的假设！

OLS方法中， $\sum e_i X_i = 0$ ，就类似于自变量X与随机干扰项不相关的假设（也即 $Cov(u_i, X_i) = 0$ ）。

OLS方法中， $\sum e_i = 0, (\bar{e}_i = 0)$ ，就类似于随机干扰项期望值为0的假设（也即 $E(u|X_i) = 0$ ）



关于CLRM假设体系的讨论

思考1：CLRM假设本质上是在讨论什么？

回答：数据是依据什么机制产生的？(data-generating process, DGP)

- 我们手头只有 n 个样本数据对 (Y_i, X_i) 。
- 但是我们希望能得到对总体参数集 Φ 的合理推断。
- 因此，如果不对总体回归模型（PRM）作任何假设的话，我们就没有更多的信息，来对总体参数集进行任何有价值的推断。



关于CLRM假设体系的讨论

思考2：CLRM假设既然有很多地方明显不符合现实，那么我们可以放宽这些假设么？如果现实根本就是违背了CLRM假设，OLS方法又将何去何从？对于参数估计量的性质会造成立致性的打击么？

回答：放宽CLRM假设和违背CLRM假设的后果是不同的。

- 如果只是放宽CLRM假设，则不会影响OLS方法的参数估计量的BLUE性质。
- 但是如果是违背了CLRM假设，则OLS方法的参数估计量的BLUE性质很可能无法保持！



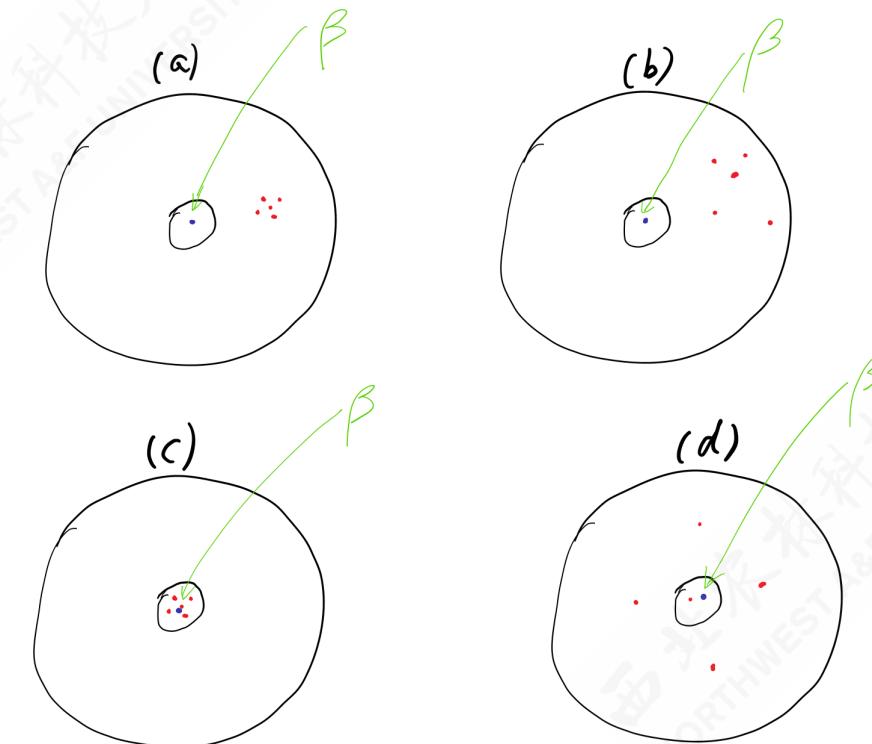
OLS方法怎么就“天生丽质”了？

我们已经知道了，OLS方法和CLRM假设“双剑合璧”下，参数估计量是最优线性无偏估计量（BLUE）。

问题是：

- OLS拟合估计方法很有“特点”，是不是意味着它就很“优秀”呢？
- 我们怎么知道，OLS方法和CLRM假设“双剑合璧”就是所向披靡呢？
- 同样在CLRM假设下，有没有一种不同于OLS的其他估计方法，也是同样那么优秀，甚至更好呢？

参数估计量的可能行为：





OLS方法怎么就“天生丽质”了？

某种参数估计方法（如OLS方法），得到的估计量（如 $\hat{\beta}_2, \hat{\beta}_1, \hat{\sigma}^2$ ）是总体参数（如 $\beta_2, \beta_1, \sigma^2$ ）的最优线性无偏估计量（Best Linear Unbiased Estimate, BLUE）需要满足如下三个条件：

- 线性的(Linear): 估计量是因变量 Y_i 的线性函数。
- 无偏的(Unbiased): 估计量的均值或期望值 ($E(\hat{\beta}_i)$) 等于参数的真值 (β_i)。
- 方差最小的 (Best) : 也即估计量是最有效的(Efficient), 是所有线性无偏估计量中有最小方差的那个估计量。

我们下面将证明：OLS方法在给定条件下就是那么“天生丽质”！

- 用记号表达为： $\hat{\Phi}$ of $\xrightarrow[CLRM]{OLS}$ Φ is BLUE
- 以上表达读作：在经典详细回归模型假设下 (CLRM)，采用普通最小二乘法 (OLS)，得到参数 Φ 的估计量 $\hat{\Phi}$ ，是最优线性无偏估计量 (BLUE)。



高斯-马尔可夫定理(Gauss-Markov Theorem)：

高斯-马尔可夫定理(Gauss-Markov Theorem): 在给定经典线性回归模型(CLRM)的假定下，最小二乘(OLS)估计量（如 $\hat{\beta}_2, \hat{\beta}_1, \hat{\sigma}^2$ ），在无偏线性估计量一类中，有最小方差，就是说它们是总体参数（如 $\beta_2, \beta_1, \sigma^2$ ）的最优线性无偏估计量(BLUE)。

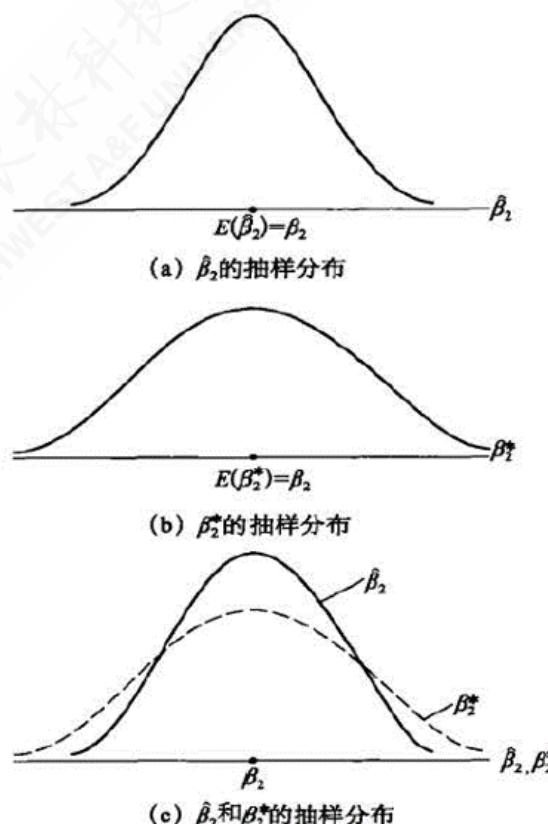
课堂思考与讨论：

- 讨论1：为什么最小二乘法(OLS)被计量学家奉为神明？还有其他选择吗？
- 讨论2：OLS得到的BLUE到底有什么值得你称赞？
- 讨论3：OLS得到BLUE还需要CLRM假设以外的更多假设吗？(正态性？？)



OLS方法最优线性无偏估计性质的证明

不同估计方法下两个估计量的抽样分布



- 图(a) OLS方法下估计量 $\hat{\beta}_2$ 是总体参数 β_2 的一个线性无偏估计量
- 图(b) 其他某种方法下估计量 $\hat{\beta}_2^*$ 也是总体参数 β_2 的一个线性无偏估计量
- 图(c) 那么哪一个估计量 ($\hat{\beta}_2$ 还是 $\hat{\beta}_2^*$) 更能为我们所接受呢?
- 讨论1：什么是抽样分布？
- 讨论2：怎样获得估计量分布？
- 讨论3：没有比OLS估计量更好的估计量吗？



CLRM假设下OLS估计量的性质1：线性性

线性性 (Linearity) : 是指 $\hat{\beta}_2$ 和 $\hat{\beta}_1$ 对 Y_i 是线性的。

具体证明过程如下：

步骤1：证明斜率系数估计量 $\hat{\beta}_2$ 对 Y_i 是线性的。

$$\hat{\beta}_2 = \sum k_i Y_i \quad \leftarrow \left[k_i = \frac{x_i}{\sum x_i^2} \right]$$

又因为 $k_i = \frac{x_i}{\sum x_i^2}$ 是不全为0的（为什么？），所以斜率系数估计量 $\hat{\beta}_2$ 对 Y_i 是线性的。



CLRM假设下OLS估计量的性质1：线性性

详细证明（反证法）：

- 假设 $H_0: k_i = \frac{x_i}{\sum x_i^2} = 0$, 也即全为零。
- 则有: $x_i = X_i - \bar{X} = 0$,
- 则有: X_i 处处等于 \bar{X} ,
- 也就意味着: x_i 是一个不变的量 (只有一个取值)
- 因此, 这是明显违背CLRM假设中关于自变量 X_i 的设定 (见前面)。
- 因此, H_0 是显然不成立的, 认为 k_i 不能全为零。[证明完毕]



CLRM假设下OLS估计量的性质1：线性性

步骤2：证明截距系数估计量 $\hat{\beta}_1$ 对 Y_i 是线性的。

$$\hat{\beta}_1 = \sum w_i Y_i \quad \leftarrow \left[w_i = \frac{1}{n} - k_i \bar{X} \right]$$

又因为 $w_i = \frac{1}{n} - k_i \bar{X}$ 是不全为0的（为什么？），所以斜率系数估计量 $\hat{\beta}_2$ 对 Y_i 是线性的。



CLRM假设下OLS估计量的性质1：线性性

详细证明（反证法）：

- 假设 $H_0: w_i = \frac{1}{n} - k_i \bar{X} = 0$, 也即全为零。
- 则有: $\sum w_i = \sum (\frac{1}{n} - k_i \bar{X}) = 0$,
- 则有: $1 - \bar{X} \sum k_i = 0$,
- 又因为: $\sum k_i = \sum \frac{x_i}{\sum x_i^2} = \frac{\sum x_i}{\sum x_i^2} = 0$,
- 因此有: $1 - 0 = 0$, 也即 $1 = 0$
- 因此, 这显然是错误的。
- 因此 H_0 是显然不成立的, 认为 w_i 不能全为零。[证明完毕]



CLRM假设下OLS估计量的性质2：无偏性

无偏性(Unbias): 估计量期望值 ($E(\hat{\beta}_i)$) 等于参数的真值 (β_i)。

步骤1: 证明斜率系数估计量 $\hat{\beta}_2$ 是无偏的, 也即 $E(\hat{\beta}_2) = \beta_2$ 。

容易有:

$$\begin{aligned}\hat{\beta}_2 &= \sum k_i Y_i && \leftarrow \left[k_i = \frac{x_i}{\sum x_i^2} \right] \\ &= \sum k_i (\beta_1 + \beta_2 X_i + u_i) \\ &= \beta_1 \sum k_i + \beta_2 \sum k_i X_i + \sum k_i u_i \\ &= 0 + \beta_2 + \sum k_i u_i\end{aligned}$$



CLRM假设下OLS估计量的性质2：无偏性

(续前) 因为有：

$$\begin{aligned}\sum k_i &= \sum \frac{x_i}{\sum x_i^2} = \frac{\sum x_i}{\sum x_i^2} = 0 \\ \sum k_i X_i &= \sum \left[\frac{x_i}{\sum x_i^2} \cdot X_i \right] = \frac{\sum x_i X_i}{\sum x_i^2} = \frac{\sum x_i(x_i + \bar{X})}{\sum x_i^2} \\ &= \frac{\sum x_i^2 + \sum x_i \bar{X}}{\sum x_i^2} = \frac{\sum x_i^2 + \bar{X} \sum x_i}{\sum x_i^2} = 1\end{aligned}$$

所以有：

$$\begin{aligned}\hat{\beta}_2 &= \beta_2 + \sum k_i u_i \\ E(\hat{\beta}_2) &= E(\beta_2 + \sum k_i u_i) = \beta_2 + E(\sum k_i u_i) \\ &= \beta_2 + \sum [k_i E(u_i)] = \beta_2\end{aligned}$$



CLRM假设下OLS估计量的性质2：无偏性

步骤2：证明截距系数估计量 $\hat{\beta}_1$ 是无偏的，也即 $E(\hat{\beta}_1) = \beta_1$ 。

| 容易有：

$$\begin{aligned}\hat{\beta}_1 &= \sum w_i Y_i && \leftarrow \left[w_i = \frac{1}{n} - k_i \bar{X} \right] \\ &= \sum w_i (\beta_1 + \beta_2 X_i + u_i) \\ &= \beta_1 \sum w_i + \beta_2 \sum w_i X_i + \sum w_i u_i \\ &= \beta_1 + 0 + \sum k_i u_i\end{aligned}$$



CLRM假设下OLS估计量的性质2：无偏性

(续前) 因为有：

$$\sum w_i = \sum \left[\frac{1}{n} - k_i \bar{X} \right] = 1 - \bar{X} \sum k_i = 1$$

$$\sum w_i X_i = \sum \left[\left(\frac{1}{n} - k_i \bar{X} \right) \cdot X_i \right] = \sum \left(\frac{X_i}{n} - \bar{X} k_i X_i \right) = \bar{X} - \bar{X} \sum (k_i X_i) = 0$$

所以有：

$$\begin{aligned}\hat{\beta}_1 &= \beta_2 + \sum w_i u_i \\ E(\hat{\beta}_1) &= E(\beta_1 + \sum k_i u_i) \\ &= \beta_1 + E(\sum k_i u_i) \\ &= \beta_1 + \sum [k_i E(u_i)] \\ &= \beta_1\end{aligned}$$

[证明完毕]！



CLRM假设下OLS估计量的性质3：方差最小性

方差最小性 (Best) : 也即估计量是最有效的(Efficient), 是所有线性无偏估计量中, 方差为最小的那个估计量。

| 证明:

- 已知估计量 $\hat{\beta}_2$ 和 $\hat{\beta}_1$ 的总体方差分别是:

$$Var(\hat{\beta}_2) \equiv \sigma_{\hat{\beta}_2}^2 = \frac{\sigma^2}{\sum x_i^2}$$

$$Var(\hat{\beta}_1) \equiv \sigma_{\hat{\beta}_1}^2 = \frac{\sum X_i^2}{n} \cdot \frac{\sigma^2}{\sum x_i^2}$$



CLRM假设下OLS估计量的性质3：方差最小性

假设存在用其他方法估计的线性无偏估计量 $\hat{\beta}_2^*$ 和 $\hat{\beta}_1^*$:

$$\begin{aligned}\hat{\beta}_2^* &= \sum [(k_i + d_i)Y_i] = \sum c_i Y_i \\ \hat{\beta}_1^* &= \sum [(w_i + g_i)Y_i] = \sum h_i Y_i\end{aligned}$$

其中， d_i 和 g_i 为不全为零的常数（证明略），则可以证明（此处略）：

$$\begin{aligned}Var(\hat{\beta}_2^*) &\geq Var(\hat{\beta}_2) \\ Var(\hat{\beta}_1^*) &\geq Var(\hat{\beta}_1)\end{aligned}$$

因此，方差最小性得以证明！



关于OLS估计量性质的小结

评价不同估计方法的参数估计量性质，一般是从线性的(Linear)、无偏性(Unbiased)和有效性（方差最小性，Best）三个维度来共同测量。

OLS估计方法，在CLRM假设下，其参数估计量很好地满足如上三个性质，因此我们称OLS方法估计的参数估计量是最优线性无偏估计量（BLUE）。



关于OLS估计量性质的思考

关于OLS估计量有效性（方差最小性，best）的证明过程你满意么？你能不能自己查阅资料证明一下？找到自己满意的证明过程！

参考答案：建议参阅Greene的《计量经济分析》，他采用的矩阵方法做了完美的证明！

还有没有其他维度来评价不同估计方法的参数估计量性质？

参考答案：还可以从“一致性”（consistency）维度来评价，主要考察参数估计量的渐进性质，也即在样本不断接近总体时估计量的表现。

使得参数估计量具备BLUE性质，仅有OLS方法么（独孤求败）？你能说出一个么？



经典正态线性回归模型假设 (N-CLRM)

经典正态线性回归模型(classical normal linear regression model , N-CLRM): 在经典线性回归模型(CLRM)假设中再增加干扰项 u_i 服从正态性的相关假设。

- 均值为0: $E(u|X_i) = 0$
- 同方差: $Var(u_i|X_i) \equiv \sigma^2$
- 无自相关: $E(u_i, u_j|X_i) = 0$
- 正态性分布: $u_i \sim N(0, \sigma^2)$

以上几条也可以统写为: $u_i \sim iid. N(0, \sigma^2)$

其中, iid表示独立同分布(Independent Identical Distribution, iid)。



N-CLRM假设下OLS估计量的性质

在N-CLRM假设下，OLS估计量有如下统计性质：

- 性质1：无偏性
- 性质2：有效性（方差最小）
- 性质3：一致性（收敛到它们的总体参数上）



N-CLRM假设下OLS估计量的性质

- 性质4：估计量 $\hat{\beta}_2$ 是正态分布的：

$$\hat{\beta}_2 \sim N(\mu_{\hat{\beta}_2}, \sigma_{\hat{\beta}_2}^2)$$

$$\mu_{\hat{\beta}_2} = E(\hat{\beta}_2) = \beta_2$$

$$\sigma_{\hat{\beta}_2}^2 = \frac{\sigma^2}{\sum x_i^2}$$

随机变量 Z_2 服从标准正态分布：

$$Z_2 = \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} \sim N(0, 1)$$

$$\mu(Z_2) = E(\hat{\beta}_2 - \beta_2) = 0$$

$$\sigma_{Z_2}^2 = Var\left(\frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}}\right) = \frac{Var(\hat{\beta}_2)}{\sigma_{\hat{\beta}_2}^2} = 1$$



N-CLRM假设下OLS估计量的性质

- 性质5：估计量 $\hat{\beta}_1$ 是正态分布的：

$$\hat{\beta}_1 \sim N(\mu_{\hat{\beta}_1}, \sigma_{\hat{\beta}_1}^2)$$

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

$$\sigma_{\hat{\beta}_1}^2 = \frac{\sum X_i^2}{n} \cdot \frac{\sigma^2}{\sum x_i^2}$$

随机变量 Z_1 服从标准正态分布：

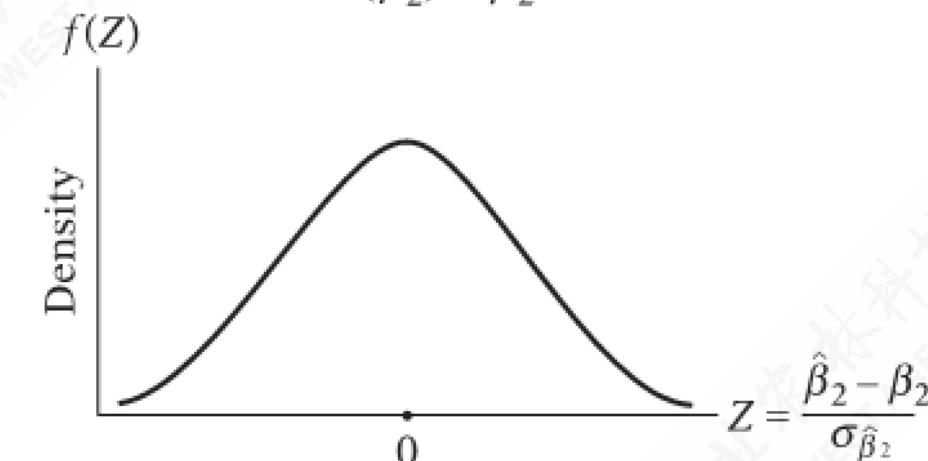
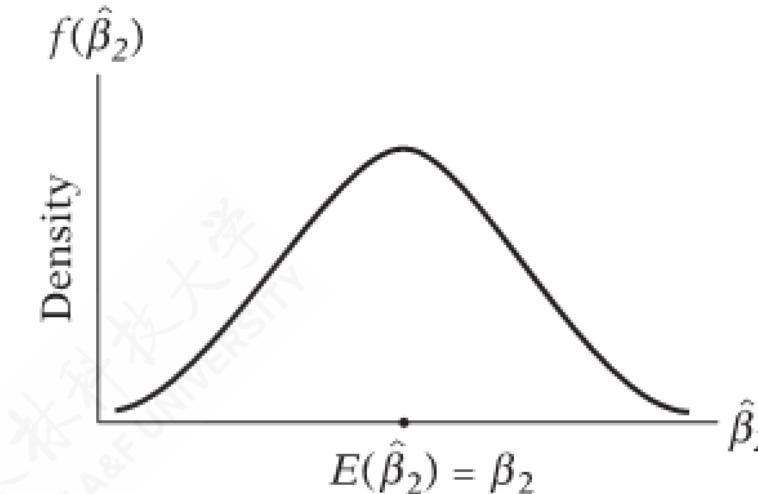
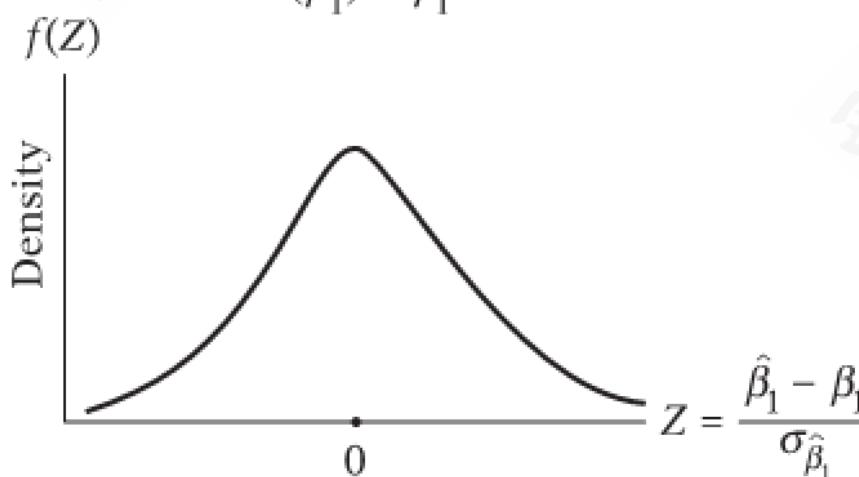
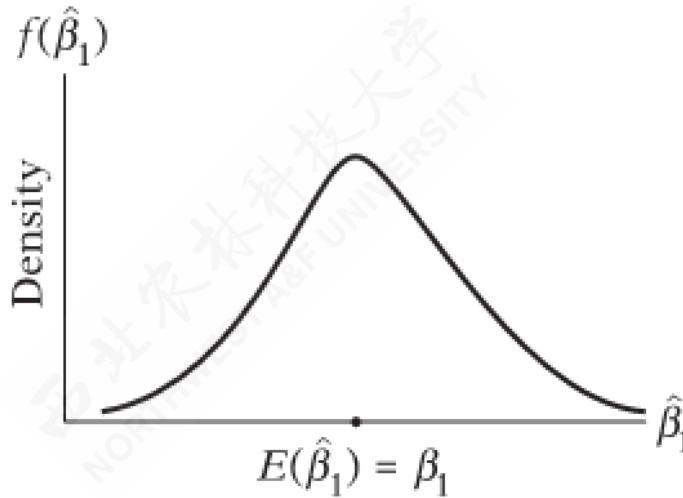
$$Z_1 = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0, 1)$$

$$\mu(Z_1) = E(\hat{\beta}_1 - \beta_1) = 0$$

$$\sigma_{Z_1}^2 = Var\left(\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}\right) = \frac{Var(\hat{\beta}_1)}{\sigma_{\hat{\beta}_1}^2} = 1$$



N-CLRM假设下OLS估计量的性质





N-CLRM假设下OLS估计量的性质

- 性质6: $X \equiv (n - 2)\hat{\sigma}^2/\sigma^2$ 服从自由度为 $(n - 2)$ 的卡方分布。

$$\begin{aligned} X &\equiv (n - 2)\hat{\sigma}^2/\sigma^2 \\ X &\sim \chi^2(n - 2) \end{aligned}$$

- 性质7: 随机变量 $(\hat{\beta}_2, \hat{\beta}_1)$ 的分布独立于随机变量 $\hat{\sigma}^2$
- 性质8: 估计量 $(\hat{\beta}_2, \hat{\beta}_1)$ 在所有无偏估计中, 无论是线性还是非线性, 都有最小的方差。也即, 它们是最有无偏估计量 (Best Unbiased Estimators, BUE) 。



关于N-CLRM的几点小结

核心观点：

除了关心参数估计量的性质（是否BLUE），我们还需要关注：参数估计量作为一个随机变量，会服从什么概率分布？这样才会为后面的假设检验提供基础！它将成为完成“由样本推断总体”逻辑分析的最后一个台阶，也是最重要的一个环节之一！

N-CLRM假设体系，是在CLRM假设基础之上，额外再增加了一条关于随机干扰项服从正态分布的假设。基于此，我们可以推断得到回归系数估计量也将服从正态分布，从而进一步可以构造出很多有用的样本统计量，例如后面要学的t统计量、F统计量等。



关于N-CLRM的思考与讨论

你能说出现实中，随机干扰项 u_i 服从其他概率分布的情形？

参考答案：显然，现实社会现象中有很多不服从正态分布的情形，比如t分布、二项分布等。

如果随机干扰项确实不服从正态分布，OLS方法+CLRM假设还能那么“天生丽质”、那么“无往不利”么？

参考答案：事实上，我们并不需要随机干扰项服从正态分布这条强假设。即时不服从正态分布，中心极限定理和大数定理也照样能保证OLS方法的有效性！

1.4 变异分解与拟合优度



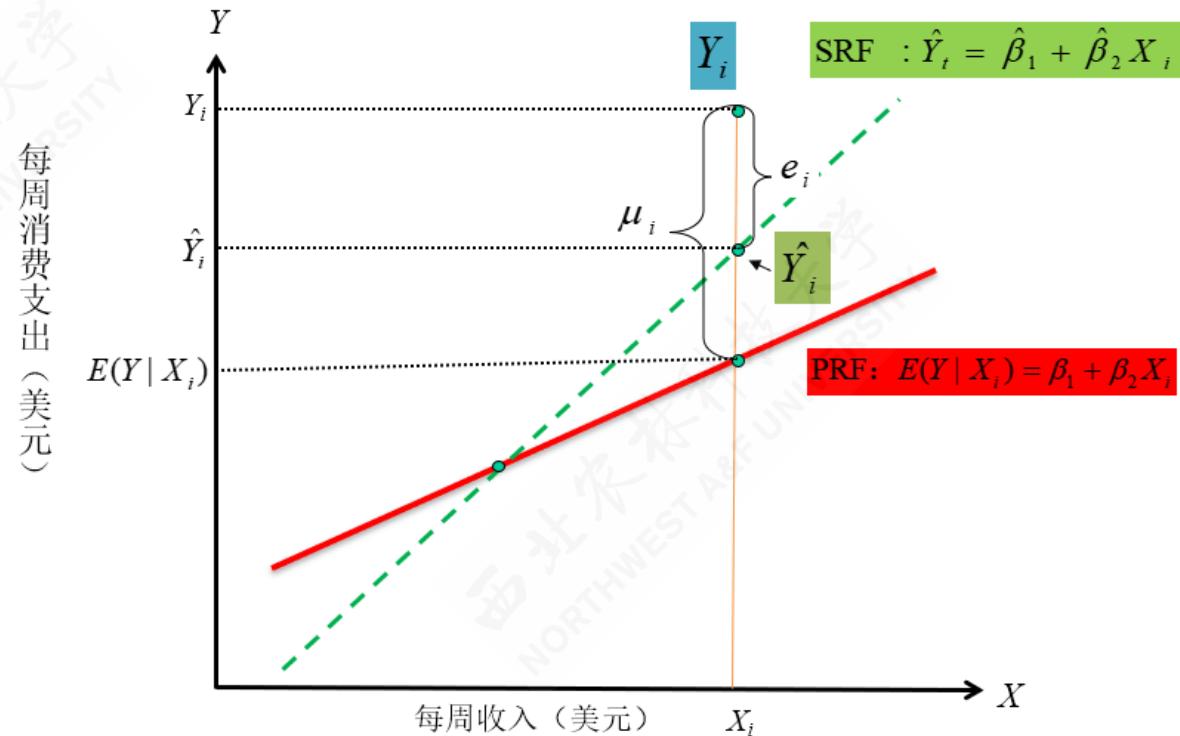
怎么来判定OLS方法对特定样本数据拟合的好坏？

请大家思考如下几个问题：

- 样本数据不完全落在拟合的直线（或曲线）上，是经常发生的么？
- 怎么来表达或测量这种对样本数据拟合的不完全性？
- 在OLS方法和CLRM假设“双剑合璧”下，对特定样本数据的拟合不是已经证明最好的么（BLUE）？为什么还要说“拟合”有“好坏之分”？



Y 变异的分解



$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$
$$y_i = \hat{y}_i + e_i$$



平方和分解

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

$$y_i = \hat{y}_i + e_i$$

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2$$

$$TSS = ESS + RSS$$

- TSS 表示总离差平方和; ESS 表示回归平方和; RSS 表示残差差平方和

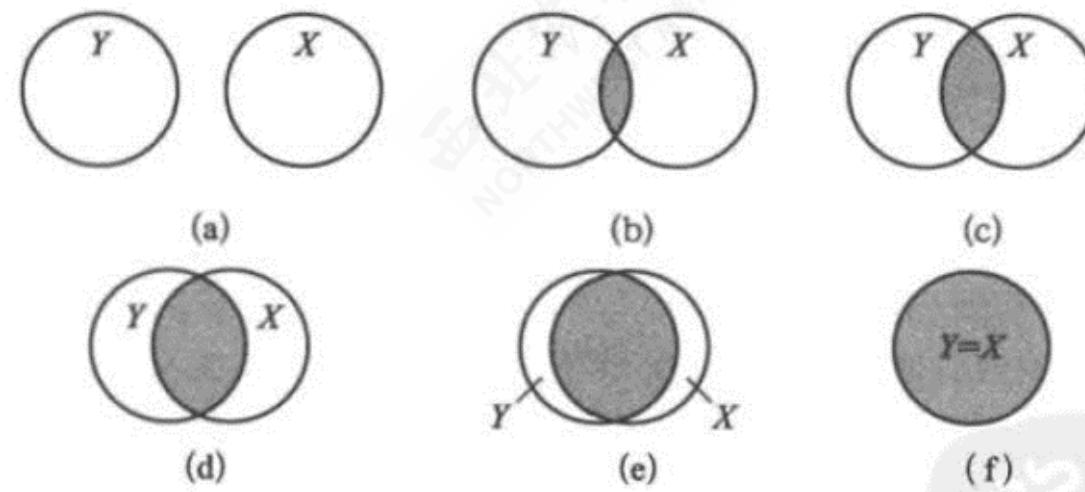
$$\begin{aligned}\sum y_i^2 &= \sum (\hat{y}_i e_i)^2 \\&= \sum (\hat{y}_i^2 + 2\hat{y}_i e_i + e_i^2) \\&= \sum \hat{y}_i^2 + 2 \sum \hat{y}_i e_i + \sum e_i^2 \\&= \sum \hat{y}_i^2 + 2 \sum ((\hat{\beta}_2 x_i) e_i) + \sum e_i^2 \\&= \sum \hat{y}_i^2 + 2\hat{\beta}_2 \sum (x_i e_i) + \sum e_i^2 \quad \leftarrow [\sum x_i e_i = 0] \\&= \sum \hat{y}_i^2 + \sum e_i^2\end{aligned}$$



拟合优度的度量

拟合优度 (Goodness of fit) : 判断样本回归线对一组数据拟合优劣水平的度量。

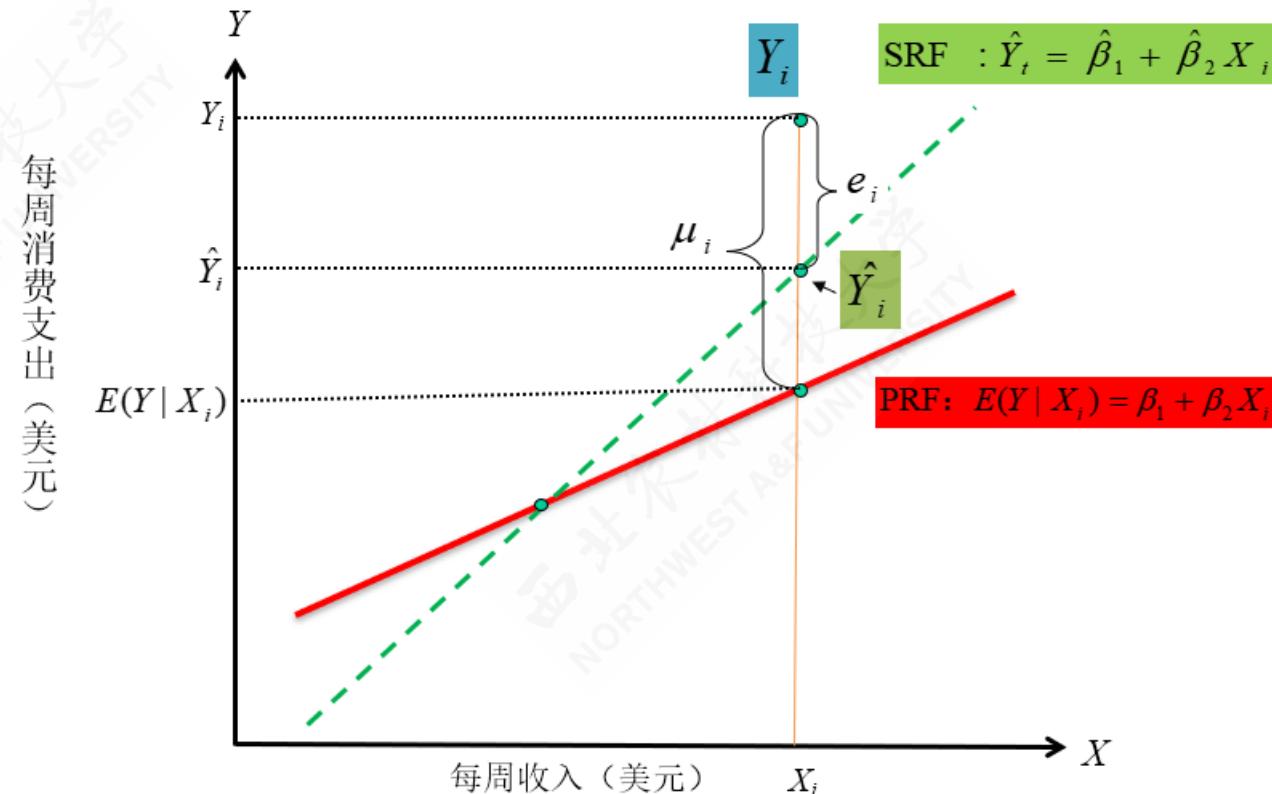
判定系数 (coefficient of determination) : 一种利用平方和分解, 考察样本回归线对数据拟合效果的总度量。一元回归中, 一般记为 r^2 ; 多元回归中, 一般记为 R^2 。



维恩图看拟合优度



拟合优度的度量



平方和分解看拟合优度



拟合优度的度量

判定系数 r^2 计算公式1:

$$r^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

判定系数 r^2 计算公式2:

$$r^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$



拟合优度的度量

判定系数 r^2 计算公式3:

$$r^2 = \frac{ESS}{TSS} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\sum (\hat{\beta}_2 x_i)^2}{\sum y_i^2} = \hat{\beta}_2^2 \frac{\sum x_i^2}{\sum y_i^2} = \hat{\beta}_2^2 \frac{S_{X_i}^2}{S_{Y_i}^2}$$

判定系数 r^2 计算公式4:

$$r^2 = \hat{\beta}_2^2 \cdot \frac{\sum x_i^2}{\sum y_i^2} = \left(\frac{\sum x_i y_i}{\sum x_i^2} \right)^2 \cdot \left(\frac{\sum x_i^2}{\sum y_i^2} \right) = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2}$$

课堂讨论:

- 讨论1: r^2 是一个非负量。为什么?
- 讨论2: $0 \leq r^2 \leq 1$, 两个端值分别意味什么?



判定系数和简单相关系数的区别与联系

总体相关系数：是变量 X_i 与变量 Y_i 总体相关关系的参数，一般记为 ρ 。

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X_i)Var(Y_i)}} = \frac{E(X_i - EX)(Y_i - EY)}{\sqrt{E(X_i - EX)^2}E(Y_i - EY)^2}$$

样本相关系数：是从总体中抽取随机样本，获得变量 X_i 与变量 Y_i 样本相关关系的统计量度量，一般记为 r 。

$$r = \frac{S_{XY}^2}{S_X * S_Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$



判定系数和相关系数的区别与联系

判定系数和简单相关系数的联系：

- 在一元回归中，判定系数 r^2 等于样本相关系数 r 的平方。

判定系数和简单相关系数的区别：

- 判定系数 r^2 表明因变量变异由解释变量所解释的比例，而相关系数 r 只能表明变量间的线性关联强度。
- 在多元回归中，这种区别会更加凸显！因为那时的相关系数 r 出现了偏相关的情形（交互关联）！



本节内容小结

使采用OLS方法，它对样本数据的拟合也是不完全的。意味着实际数据点在样本回归线附近，而不是在样本回归线上。我们可以把样本点行为的“变异”，划分为“回归”能解释的部分和“随机”的部分。并进一步获得变异平方和的分解。

判定系数 R^2 是对OLS拟合程度的测量，它使用了变异平方和分解的思想。在一元线性回归（含截距）中，判定系数与相关系数存在如下关系 $R^2 = r_{(X_i, Y_i)}^2$ 。注意，在多元回归中则不存在这种关系。



本节问题与思考

OLS方法的参数估计量，在CLRM假设满足情况下，就是最优线性无偏估计量（BLUE），为什么还要用判定系数来判断“拟合好还是不好？”对此，你的回答是什么？

还有没有其他指标，来反映估计方法对样本数据的拟合好坏程度？请说出一两个。

参考答案：还可以有均方误差和（MSE） $MSE = RSS/n = 1/n \sum (Y_i - \hat{Y}_i)^2$ ，以及均方误差根（RMSE）等。

1.5 置信区间和假设检验



重要概念1

- 显著性水平 α
- 置信度（或置信水平） $1 - \alpha$
- 置信区间
- 第I类错误：弃真错误 $\alpha = P(Z > Z_0 | H_0 = True)$
- 第II类错误：取伪错误 $\beta = P(Z \leq Z_0 | H_1 = True)$



重要概念2

- 区间估计

$$\Pr(\hat{\beta}_2 - \delta \leq \beta_2 \leq \hat{\beta}_2 + \delta) = 1 - \alpha$$

- 随机区间(random interval) : $(\hat{\beta}_2 - \delta, \hat{\beta}_2 + \delta)$
- 置信区间(confidence interval): $\hat{\beta}_2 - \delta \leq \beta_2 \leq \hat{\beta}_2 + \delta$
- 显著性水平(level of significance): α
- 置信度或置信系数(confidence coefficient): $1 - \alpha$
- 置信限 (confidence limits) 或临界值 (critical values)
- 置信上限 (lower confidence limit)
- 置信下限 (upper confidence limit)



区间估计

注意的几个问题（自己去巩固）：

- 陈述问题：
 - 落入给定界限内的概率是 $1 - \alpha$ 。 (X) ? ?
 - 使用我们的方法构造出来的区间包含 β 的概率为 $1 - \alpha$ 。
 - 抽样层面来理解：从重复多次抽样中来看，平均起来这些区间将有 $(1 - \alpha)$ 的可能包含着参数的真值。
- 我们构造的区间是只是随机区间！ (?)
 - 对于计算出的参数估计值而言，得到的区间中要么包含参数真值要么不包含。概率为0或1！
 - 例如：对于95%置信区间的 $0.4268 \leq \beta_2 \leq 0.5914$ 而言，不能说这个区间包含真实的 β_2 的概率是95%。这个概率不是1就是0。



区间估计

注意的几个问题（自己去巩固）：

- 两个游戏：
 - 掷硬币
 - 套圈游戏

请问：区间估计更象哪一个？

- 置信区间的两个特点：
 - 位置的随机性
 - 长度的随机性



斜率系数的置信区间

$$\hat{\beta}_2 \sim N(\mu_{\hat{\beta}_2}, \sigma_{\hat{\beta}_2}^2) \quad \leftarrow \left[\mu_{\hat{\beta}_2} = \beta_2; \quad \sigma_{\hat{\beta}_2}^2 = \frac{\sigma^2}{\sum x_i^2} \right]$$

$$Z = \frac{(\hat{\beta}_2 - \beta_2)}{\sqrt{\text{var}(\hat{\beta}_2)}} = \frac{(\hat{\beta}_2 - \beta_2)}{\sqrt{\sigma_{\hat{\beta}_2}^2}} = \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} = \frac{(\hat{\beta}_2 - \beta_2)}{\sqrt{\frac{\sigma^2}{\sum x_i^2}}} \quad \leftarrow Z \sim N(0, 1)$$

$$T = \frac{(\hat{\beta}_2 - \beta_2)}{\sqrt{S_{\hat{\beta}_2}^2}} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{S_{\hat{\beta}_2}^2}} = \frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}} \quad \leftarrow T \sim t(n-2)$$

$$S_{\hat{\beta}_2}^2 = \frac{\hat{\sigma}^2}{\sum x_i^2}; \quad \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

$$\Pr[-t_{\alpha/2, (n-2)} \leq T \leq t_{\alpha/2, (n-2)}] = 1 - \alpha$$



斜率系数的置信区间

$$\Pr \left[-t_{\alpha/2, (n-2)} \leq \frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}} \leq t_{\alpha/2, (n-2)} \right] = 1 - \alpha$$

$$\Pr \left[\hat{\beta}_2 - t_{\alpha/2, (n-2)} \cdot S_{\hat{\beta}_2} \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2, (n-2)} \cdot S_{\hat{\beta}_2} \right] = 1 - \alpha$$

因此， β_2 的 $100(1 - \alpha)\%$ 置信上限和下限分别为：

$$\hat{\beta}_2 \pm t_{\alpha/2} \cdot S_{\hat{\beta}_2}$$

β_2 的 $100(1 - \alpha)\%$ 置信区间为：

$$\left[\hat{\beta}_2 - t_{\alpha/2} \cdot S_{\hat{\beta}_2}, \quad \hat{\beta}_2 + t_{\alpha/2} \cdot S_{\hat{\beta}_2} \right]$$



截距系数的置信区间

$$\hat{\beta}_1 \sim N(\mu_{\hat{\beta}_1}, \sigma_{\hat{\beta}_1}^2) \quad \leftarrow \left[\mu_{\hat{\beta}_1} = \beta_1; \quad \sigma_{\hat{\beta}_1}^2 = \frac{\sum X_i^2}{n} \frac{\sigma^2}{\sum x_i^2} \right]$$

$$Z = \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\text{var}(\hat{\beta}_1)}} = \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\sigma_{\hat{\beta}_1}^2}} = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} = \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\frac{\sum X_i^2}{n} \cdot \frac{\sigma^2}{\sum x_i^2}}} \quad \leftarrow Z \sim N(0, 1)$$

$$T = \frac{(\hat{\beta}_1 - \beta_1)}{S_{\hat{\beta}_1}^2} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{S_{\hat{\beta}_1}^2}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \quad \leftarrow T \sim t(n-2)$$

$$S_{\hat{\beta}_1}^2 = \frac{\sum X_i^2}{n} \cdot \frac{\hat{\sigma}^2}{\sum x_i^2}; \quad \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

$$\Pr[-t_{\alpha/2, (n-2)} \leq T \leq t_{\alpha/2, (n-2)}] = 1 - \alpha$$



截距系数的置信区间

$$\Pr \left[-t_{\alpha/2, (n-2)} \leq \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \leq t_{\alpha/2, (n-2)} \right] = 1 - \alpha$$

$$\Pr \left[\hat{\beta}_1 - t_{\alpha/2, (n-2)} \cdot S_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, (n-2)} \cdot S_{\hat{\beta}_1} \right] = 1 - \alpha$$

因此， β_1 的 $100(1 - \alpha)\%$ 置信上限和下限分别为：

$$\hat{\beta}_1 \pm t_{\alpha/2} \cdot S_{\hat{\beta}_1}$$

β_1 的 $100(1 - \alpha)\%$ 置信区间为：

$$\left[\hat{\beta}_1 - t_{\alpha/2} \cdot S_{\hat{\beta}_1}, \quad \hat{\beta}_1 + t_{\alpha/2} \cdot S_{\hat{\beta}_1} \right]$$



随机干扰项的方差的置信区间

$$\chi^2 = (n - 2) \frac{\hat{\sigma}^2}{\sigma^2} \leftarrow \chi^2 \sim \chi^2(n - 2)$$

$$\Pr\left(\chi_{\alpha/2}^2 \leq \chi^2 \leq \chi_{1-\alpha/2}^2\right) = 1 - \alpha$$

$$\Pr\left(\chi_{\alpha/2}^2 \leq (n - 2) \frac{\hat{\sigma}^2}{\sigma^2} \leq \chi_{1-\alpha/2}^2\right) = 1 - \alpha$$

$$\Pr\left[(n - 2) \frac{\hat{\sigma}^2}{\chi_{1-\alpha/2}^2} \leq \sigma^2 \leq (n - 2) \frac{\hat{\sigma}^2}{\chi_{\alpha/2}^2}\right] = 1 - \alpha$$

因此， σ^2 的 $100(1 - \alpha)\%$ 为：

$$\left[(n - 2) \frac{\hat{\sigma}^2}{\chi_{1-\alpha/2}^2}, \quad (n - 2) \frac{\hat{\sigma}^2}{\chi_{\alpha/2}^2} \right]$$



假设检验的基本原理和思路

假设检验 (Hypothesis Testing) : 某一给定的观测或发现与某声称的假设是否相符? 进行统计假设检验, 就是要制定一套步骤和规则, 以使决定接受或拒绝一个虚拟假设 (原假设)。

虚拟假设(null hypothesis) —— H_0

- 指定或声称的假设, 如 $H_0 : \beta_2 = 0$
- 它是一个等待被挑战的“靶子”! “稻草人”!

备择假设(alternative hypothesis) ——

H_1

- 简单的 (simple) 备择假设, 如
 $H_1 : \beta_2 = 1.5$
- 复合的 (composite) 备择假设, 如
 $H_1 : \beta_2 \neq 1.5$

假设检验的具体方法:

- 置信区间检验 (confidence interval)
- 显著性检验 (test of significance)



置信区间检验法——双侧检验

双侧或双尾检验 (Two-sided or Two-Tail Test)

$$H_0 : \beta_2 = 0; \quad H_1 : \beta_2 \neq 0$$

- 假设检验目的：估计的是否与上述相容？
- 决策规则：
 - 构造一个 β_2 的 $100(1 - \alpha)\%$ 置信区间。
 - 如果 β_2 在 H_0 假设下落入此区间，就不拒绝 H_0 。
 - 如果它落在此区间之外，就要拒绝 H_0 。



显著性检验法

显著性检验方法(test-of-significance approach): 是一种用样本结果来证实 \$H_0\$ 真伪的检验程序。

关键思路:

- 找到一个适合的检验统计量(test statistic)。例如t统计量 χ^2 统计量、F统计量等。
- 知道该统计量在 H_0 下的抽样分布(pdf)。往往与待检验参数有关系。
- 计算样本统计量的值。也即能用样本数据快速计算出来，例如 $t_{\hat{\beta}_2}^* = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}}$ 。
- 查表找出给定显著性水平 α 下的理论统计量的临界值。例如

$$t_{1-\alpha/2}(n-2) = t_{0.975}(11) = 2.2010$$

- 比较样本统计量值和该临界值的大小。例如，比较 $t_{\hat{\beta}_2}^*$ 与 $t_{0.975}(11)$
- 做出拒绝还是接受 H_0 的判断。



假设检验：实际操作中的若干问题

关于显著性水平 α 和显著性概率值 p 。

选择显著性水平 α :

- 犯错误类型:
 - 第I类错误: 弃真错误 $\alpha = P(Z > Z_0 | H_0 = \text{True})$
 - 第II类错误: 取伪错误 $\beta = P(Z \leq Z_0 | H_1 = \text{True})$
 - [给定样本容量时]如果我们要减少犯第I类错误, 第II类错误就要增加; 反之亦然。
- 为什么 α 通常固定在 0.01、0.05、0.1 水平上?
 - 约定而已, 并非神圣不可改变!
 - 如何改变? ?



假设检验：实际操作中的若干问题

关于显著性水平 α 和显著性概率值 p

精确的显著性水平：p 值

- 对给定的样本算出一个检验统计量(如t统计量)，查到与之对应的概率：p值(p value)或概率值(probability value)
- 不约定 α ，而是直接求出犯错误概率p值，由读者自己去评判犯错误的可能性和代价！！因人而异！！



假设检验：实际操作中的若干问题

关于统计显著性与实际显著性。

- 不能一味追求统计显著性，有时候还需要考虑“实际显著性”的现实意义。
- 举例说明：
 - 边际消费倾向(MPC)是指GDP每增加1美元带来消费的增加数；宏观理论表明收入乘数为： $1/(1-MPC)$ 。
 - 若MPC的95%置信区间为(0.7129,0.7306)，当样本表明MPC的估计值为 $\widehat{MPC} = 0.74$ （此时，即乘数为3.84），你怎样抉择！！！



假设检验：实际操作中的若干问题

关于置信区间方法和显著性检验方法的选择。

- 一般来说，置信区间方法优于显著性检验方法！
- 例如：假设MPC $H_0 : \beta_2 = 0$ 显然荒谬的！

1.6 t检验和 χ^2 检验



回归系数的显著性检验：截距参数的t检验

对于截距参数 β_1 的显著性检验（t检验）。

- 步骤1：给出模型，并提出假设：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

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

- 步骤2：构造合适的检验统计量

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



回归系数的显著性检验：截距参数的t检验

- 步骤3：基于原假设 H_0 计算出样本统计量。

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

$$t^*_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \leftarrow H_0 : \beta_1 = 0$$

$$t^*_{\hat{\beta}_1} = \frac{-0.0145}{0.8746} = -0.0165$$

- 步骤4：给定显著性水平 $\alpha = 0.05$ 下，查出统计量的理论分布值。

$$t_{1-\alpha/2}(n-2) = t_{1-0.05/2}(13-2) = t_{0.975}(11) = 2.2010$$



回归系数的显著性检验：截距参数的t检验

- 步骤5：得到显著性检验的判断结论。

- 若 $|t_{\hat{\beta}_1}^*| > t_{1-\alpha/2}(n - 2)$, 则 β_1 的 t 检验结果显著。换言之，在显著性水平 $\alpha = 0.05$ 下，应显著地拒绝原假设 H_0 ，接受备择假设 H_1 ，认为截距参数 $\beta_1 \neq 0$ 。
- 若 $|t_{\hat{\beta}_1}^*| < t_{1-\alpha/2}(n - 2)$, 则 β_1 的 t 检验结果不显著。换言之，在显著性水平 $\alpha = 0.05$ 下，不能显著地拒绝原假设 H_0 ，只能暂时接受原假设 H_0 ，认为截距参数 $\beta_1 = 0$ 。



回归系数的显著性检验：斜率参数的t检验

对于斜率参数 β_2 的显著性检验（t检验）。

- 步骤1：给出模型，并提出假设：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$H_0 : \beta_2 = 0; \quad H_1 : \beta_2 \neq 0$$

- 步骤2：构造合适的检验统计量

$$T = \frac{\hat{\beta}_2 - \beta_2}{S_{\beta_2}} \leftarrow T \sim t(n-2)$$



回归系数的显著性检验：斜率参数的t检验

- 步骤3：基于原假设 H_0 计算出样本统计量。

$$T = \frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}} \leftarrow T \sim t(n-2)$$

$$t^*_{\hat{\beta}_2} = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}} \leftarrow H_0 : \beta_2 = 0$$

- 步骤4：给定显著性水平 $\alpha = 0.05$ 下，查出统计量的理论分布值。

$$t_{1-\alpha/2}(n-2) = t_{1-0.05/2}(13-2) = t_{0.975}(11) = 2.2010$$



回归系数的显著性检验：斜率参数的t检验

- 步骤5：得到显著性检验的判断结论。

- 若 $|t_{\hat{\beta}_2}^*| > t_{1-\alpha/2}(n - 2)$, 则 β_2 的 t 检验结果显著。换言之，在显著性水平 $\alpha = 0.05$ 下，应显著地拒绝原假设 H_0 ，接受备择假设 H_1 ，认为斜率参数 $\beta_2 \neq 0$ 。
 - 若 $|t_{\hat{\beta}_2}^*| < t_{1-\alpha/2}(n - 2)$, 则 β_2 的 t 检验结果不显著。换言之，在显著性水平 $\alpha = 0.05$ 下，不能显著地拒绝原假设 H_0 ，只能暂时接受原假设 H_0 ，认为斜率参数 $\beta_2 = 0$ 。



平方和分解

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

$$y_i = \hat{y}_i + e_i$$

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2$$

$$TSS = ESS + RSS$$

- 其中： TSS 表示总离差平方和； ESS 表示回归平方和； RSS 表示残差差平方和



双变量方差分析表

变异来源	平方和符号SS	平方和计算公式	自由度df	均方和符号MSS	均方和计算公式
回归平方和	ESS	$\sum (\hat{Y}_i - \bar{Y}_i)^2 = \sum \hat{y}_i^2$	1	MSS_{ESS}	$ESS/df_{ESS} = \hat{\beta}_2^2 \sum x_i^2$
残差平方和	RSS	$\sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2$	n-2	MSS_{RSS}	$RSS/df_{RSS} = \frac{\sum e_i^2}{n-2}$
总平方和	TSS	$\sum (Y_i - \bar{Y}_i)^2 = \sum y_i^2$	n-1	MSS_{TSS}	$TSS/df_{TSS} = \frac{\sum y_i^2}{n-1}$



模型整体显著性检验：F检验

- 步骤1：给出模型，并提出假设：

一元回归模型下：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$H_0 : \beta_2 = 0; \quad H_1 : \beta_2 \neq 0$$

多元回归模型下：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0; \quad H_1 : \text{not all } \beta_j = 0, \quad j \in 2, 3, \dots, k$$



模型整体显著性检验： χ^2 检验

- 步骤2：构造合适的检验统计量

$$\chi_1^2 = \left(\frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} \right)^2 = \left(\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\sigma^2 / \sum x_i^2}} \right)^2 = \frac{(\hat{\beta}_2 - \beta_2)^2 \sum x_i^2}{\sigma^2} \leftarrow \chi_1^2 \sim \chi^2(1)$$

$$\chi_2^2 = (n - 2) \frac{\hat{\sigma}^2}{\sigma^2} = \frac{\sum e_i^2}{\sigma^2} \leftarrow \chi_2^2 \sim \chi^2(n - 2)$$

$$F = \frac{\chi_1^2/1}{\chi_2^2/(n-2)} = \left(\frac{(\hat{\beta}_2 - \beta_2)^2 \sum x_i^2}{\sigma^2} \right) / \left(\frac{\sum e_i^2}{(n-2)\sigma^2} \right) = \frac{(\hat{\beta}_2 - \beta_2)^2 \sum x_i^2}{\sum e_i^2/(n-2)}$$

$$F \sim F(1, n - 2)$$



模型整体显著性检验：F检验

- 步骤3：基于原假设 H_0 计算出样本统计量。

$$\begin{aligned} F^* &= \frac{\left(\hat{\beta}_2 - \beta_2\right)^2 \sum x_i^2}{\sum e_i^2 / (n-2)} && \leftarrow H_0 : \beta_2 = 0 \\ &= \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum e_i^2 / (n-2)} \\ &= \frac{ESS/df_{ESS}}{RSS/df_{RSS}} = \frac{MSS_{ESS}}{MSS_{RSS}} = \frac{\hat{\beta}_2^2 \sum x_i^2}{\hat{\sigma}^2} \end{aligned}$$



模型整体显著性检验：F检验

- 步骤4：给定显著性水平 $\alpha = 0.05$ 下，查出统计量的理论分布值。 $F_{1-\alpha}(1, n - 2)$
- 步骤5：得到显著性检验的判断结论。
 - 若 $F^* > F_{1-\alpha}(1, n - 2)$ ，则 模型整体显著性的F检验结果显著。换言之，在显著性水平 $\alpha = 0.05$ 下，应显著地拒绝原假设 H_0 ，接受备择假设 H_1 ，认为斜率参数 $\beta_2 \neq 0$ 。
 - 若 $F^* < F_{1-\alpha}(1, n - 2)$ ，则 模型整体显著性的F检验结果不显著。换言之，在显著性水平 $\alpha = 0.05$ 下，不能显著地拒绝原假设 H_0 ，只能暂时接受原假设 H_0 ，认为斜率参数 $\beta_2 = 0$ 。



F检验和t检验的异同及联系

F检验与t检验的联系：

在一元回归模型中，t检验与F检验的结论总是一致的。

对于检验斜率参数 β_2 的显著性，两者可相互替代！在一元回归分析中，若假设 $H_0 : \beta_2 = 0$ ，则 $F^* \simeq (t^*)^2$



F检验和t检验的异同及联系

F检验与t检验的不同：

检验目的不同。F检验是检验模型的整体显著性；t检验是检验各个回归参数的显著性。

假设的提出不同：

- F检验：斜率系数联合假设 $H_0 : \beta_2 = 0; H_1 : \beta_2 \neq 0$
- t检验：回归系数分别假设 $H_0 : \beta_i = 0; H_1 : \beta_i \neq 0; i \in 1, 2$

检验原理的不同：F检验需要构造F统计量；t检验需要构造t统计量

1.7 回归预测



预测未来事件的一些惯常说法

- 算命术士:
 - “客官印堂发黑，明日必有凶象！”
- 天气预报播报词:
 - 预测西安明天是小雨，概率为95%。
 - 预测西安明天是小雨转阴，概率为95%。
 - 预测西安明天是天晴或阴天或雨天，概率为100%！
- 简要解析:
 - 人们在预测什么事件？
 - 预测多少个事件？它们发生的关系？
 - 预测如何令人信服？



两类预测

一元回归模型下：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

预测什么？

均值预测(mean prediction)：

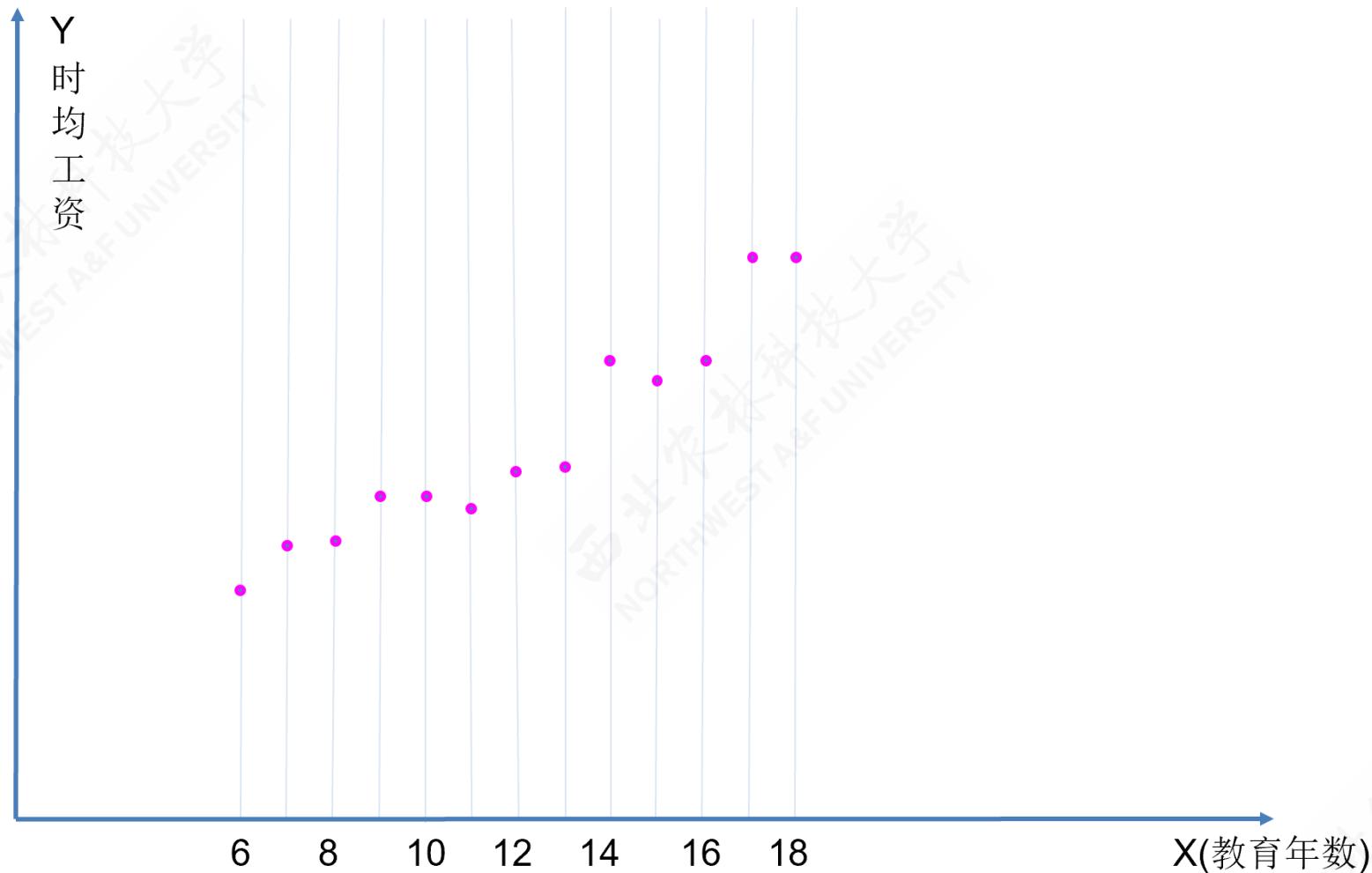
- 给定 X_0 , 预测Y的条件均值 $E(Y|X = X_0)$

个值预测(individual prediction)：

- 给定 X_0 , 预测对应于 X_0 的Y的个别值 $(Y_0|X_0)$

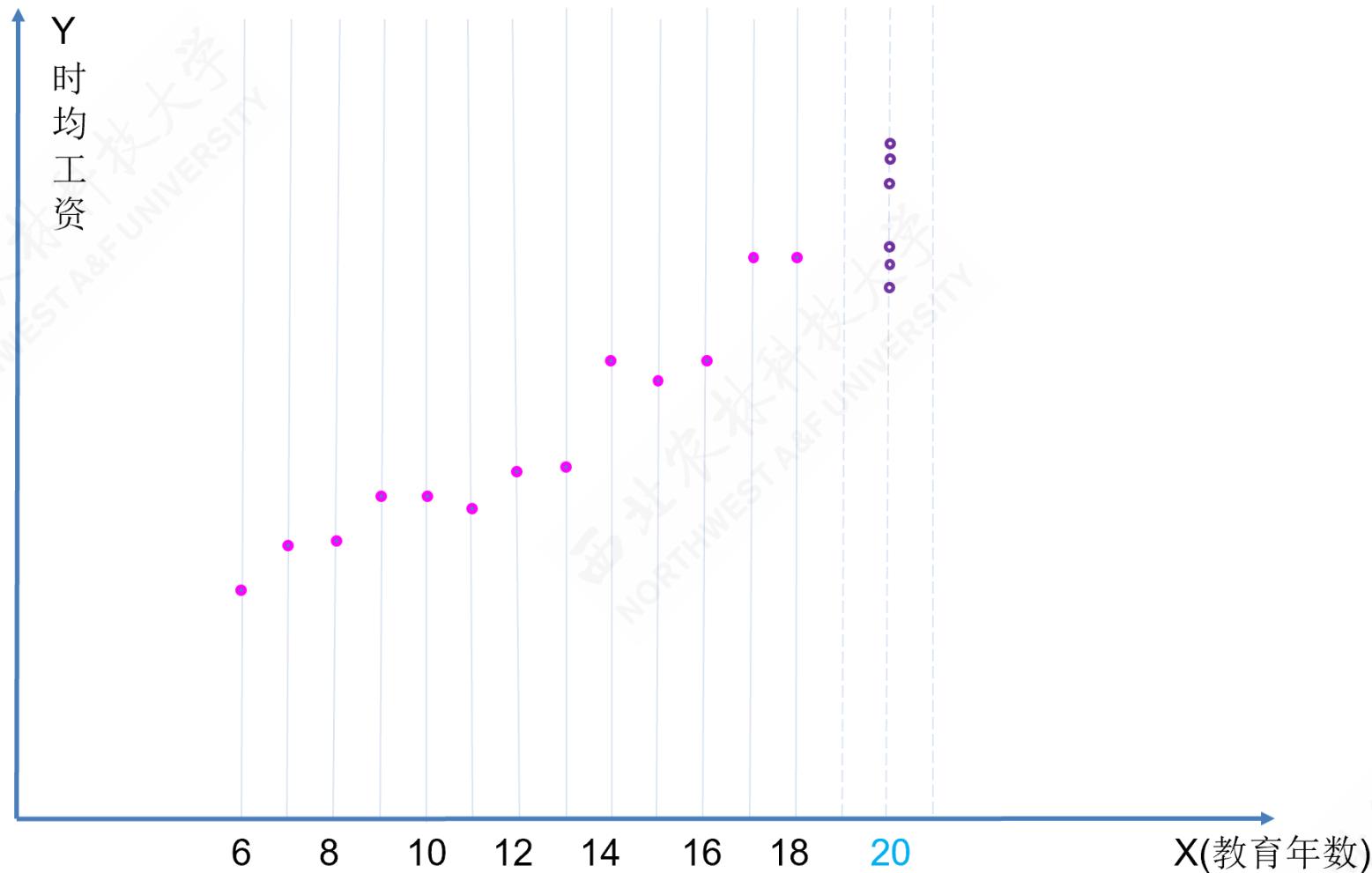


两类预测——图示（样本内）



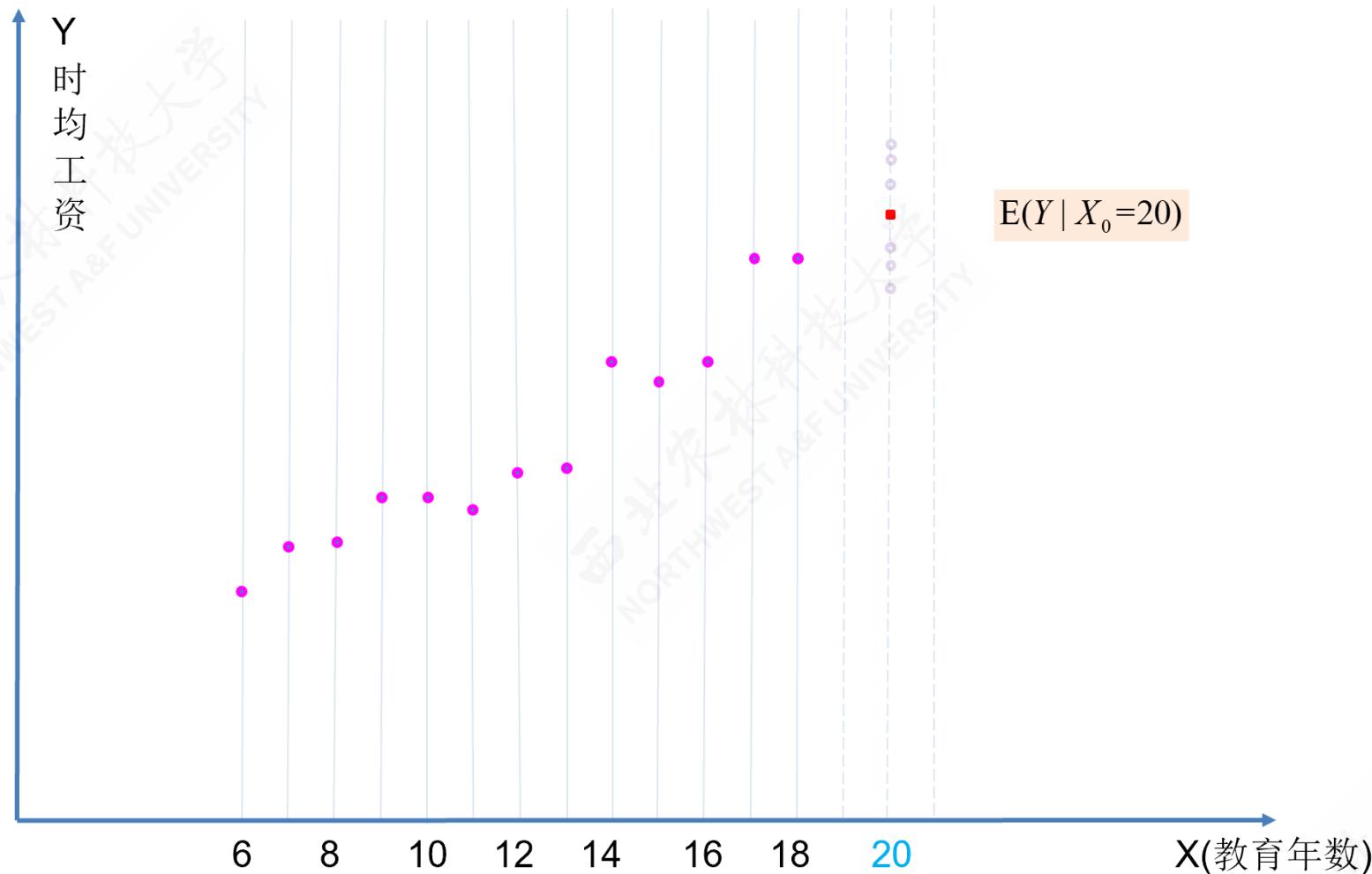


两类预测——图示（样本外）



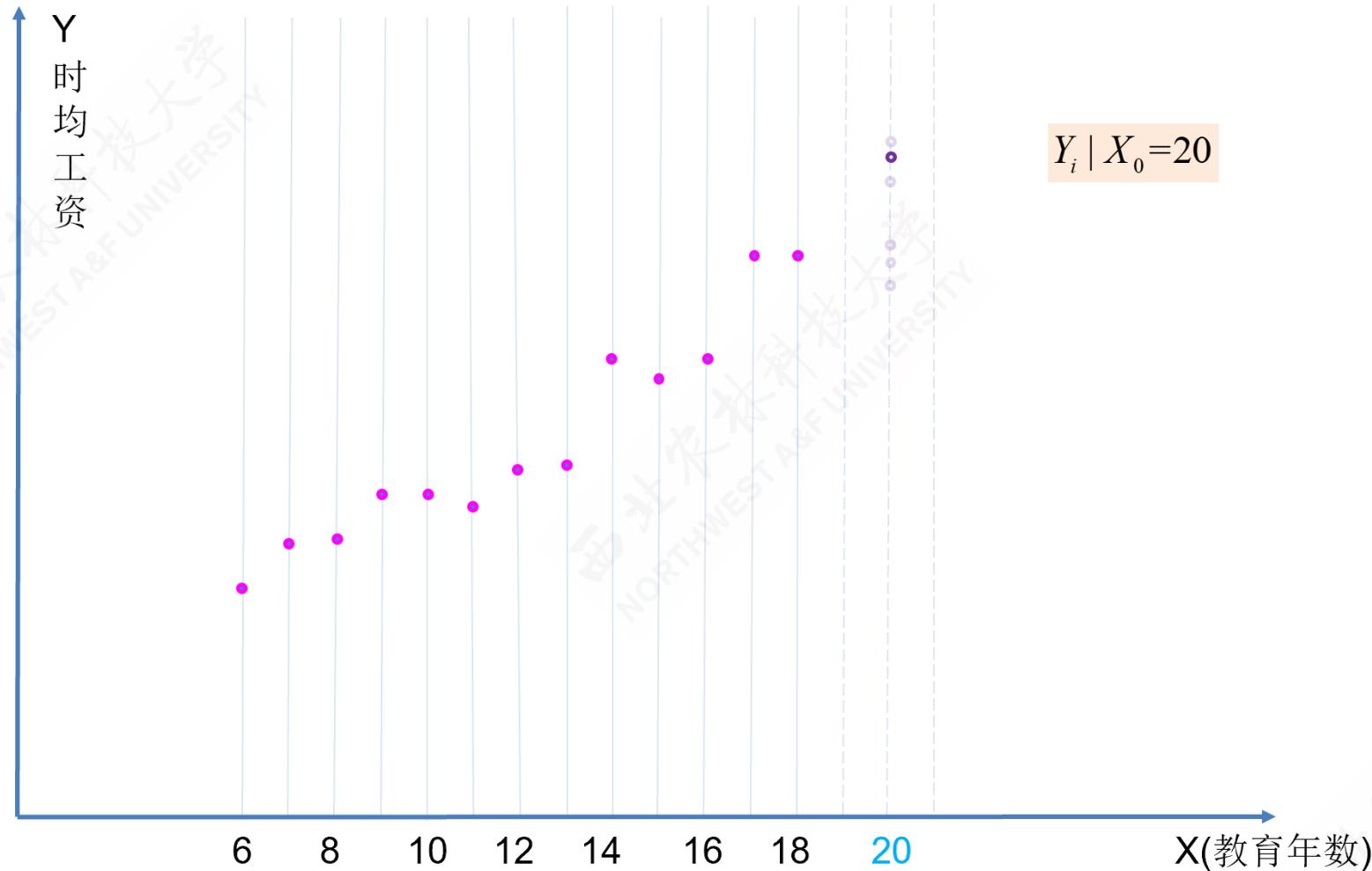


两类预测——图示（均值预测）





两类预测——图示（个值预测）





预测分析的关键

拿什么来预测？——样本数据？样本回归线？样本拟合值？

样本外拟合值 $\hat{Y}_0|X = X_0$:

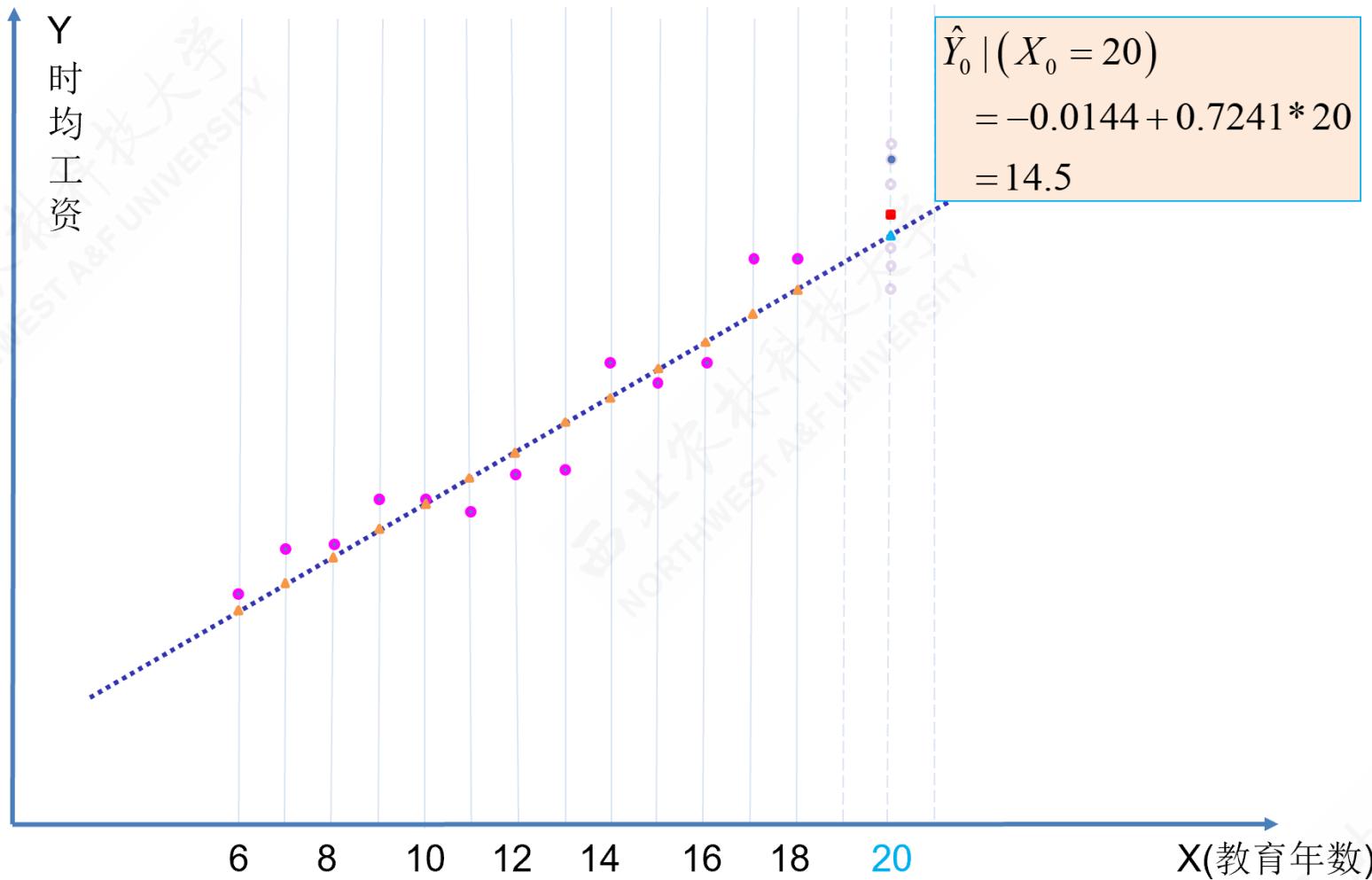
- 可以证明：样本外拟合值 $\hat{Y}_0|X = X_0$ 是均值 $E(Y|X = X_0)$ 的一个**BLUE**
- 也可以证明：样本外拟合值 $\hat{Y}_0|X = X_0$ 是个值 $(Y_0|X = X_0)$ 的一个**BLUE**

工资案例中，给定 $X_0 = 20$ ，则可以得到样本外拟合值：

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 = -0.0145 + 0.7241 * 20 = 14.4675$$



预测分析的关键





均值预测

在**N-CLRM**假设和**OLS**方法下，可以证明（证明过程略）给定 X_0 下的拟合值 \hat{Y}_0 服从如下正态分布：

$$\hat{Y}_0 \sim N\left(\mu_{\hat{Y}_0}, \sigma_{\hat{Y}_0}^2\right)$$

$$\mu_{\hat{Y}_0} = E\left(\hat{Y}_0\right) = E\left(\hat{\beta}_1 + \hat{\beta}_2 X_0\right) = \beta_1 + \beta_2 X_0 = E(Y|X_0)$$

$$\text{var}\left(\hat{Y}_0\right) = \sigma_{\hat{Y}_0}^2 = \sigma^2 \left[\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{\sum x_i^2} \right]$$

$$\hat{Y}_0 \sim N\left(E(Y|X_0), \sigma^2 \left[\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{\sum x_i^2} \right]\right)$$



均值预测

对 \hat{Y}_0 构造 t 统计量：

$$T = \frac{\hat{Y}_0 - E(Y|X_0)}{S_{\hat{Y}_0}} \sim t(n-2) \iff S_{\hat{Y}_0} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right]}$$

得到均值 $E(Y|X = X_0)$ 置信区间为：

$$\Pr\left[\hat{Y}_0 - t_{1-\alpha/2}(n-2) \cdot S_{\hat{Y}_0} \leq E(Y|X_0) \leq \hat{Y}_0 + t_{1-\alpha/2}(n-2) \cdot S_{\hat{Y}_0}\right] = 1 - \alpha$$

$$\Pr\left[\hat{\beta} + \hat{\beta}_2 X_0 - t_{1-\alpha/2}(n-2) \cdot S_{\hat{Y}_0} \leq E(Y|X_0) \leq \hat{\beta} + \hat{\beta}_2 X_0 + t_{1-\alpha/2}(n-2) \cdot S_{\hat{Y}_0}\right] = 1 - \alpha$$



个值预测

在**N-CLRM**假设和**OLS**方法下，可以证明（证明过程略）给定 X_0 下的个别值 $Y_0 = \beta_1 + \beta_2 X_0 + u_0$ 服从如下正态分布：

$$Y_0 \sim N(\mu_{Y_0}, \sigma_{Y_0}^2)$$

$$\mu_{Y_0} = E(Y_0) = E(\beta_1 + \beta_2 X_0) = \beta_1 + \beta_2 X_0$$

$$Var(Y_0) = Var(u_0) = \sigma^2$$

$$Y_0 \sim N(\beta_1 + \beta_2 X_0, \sigma^2)$$



个值预测

进一步可以构造新的随机变量 $(Y_0 - \hat{Y}_0)$, 其将服从如下正态分布:

$$Y_0 \sim N(\beta_1 + \beta_2 X_0, \sigma^2)$$

$$\hat{Y}_0 \sim N\left(\beta_1 + \beta_2 X_0, \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right]\right)$$

$$Y_0 - \hat{Y}_0 \sim N\left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right]\right)$$

$$Y_0 - \hat{Y}_0 \sim N\left(0, \sigma_{Y_0 - \hat{Y}_0}^2\right)$$



个值预测

对 $Y_0 - \hat{Y}_0$ 构造 t 统计量：

$$T = \frac{(Y_0 - \hat{Y}_0)}{S_{(Y_0 - \hat{Y}_0)}} \sim t(n-2) \quad \Leftarrow S_{(Y_0 - \hat{Y}_0)} = \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right]}$$

得到个值 Y_0 置信区间为：

$$\Pr \left[\hat{Y}_0 - t_{1-\alpha/2}(n-2) \cdot S_{(Y_0 - \hat{Y}_0)} \leq Y_0 \leq \hat{Y}_0 + t_{1-\alpha/2}(n-2) \cdot S_{(Y_0 - \hat{Y}_0)} \right] = 1 - \alpha$$

$$\Pr \left[\hat{\beta} + \hat{\beta}_2 X_0 - t_{1-\alpha/2}(n-2) \cdot S_{(Y_0 - \hat{Y}_0)} \leq Y_0 \leq \hat{\beta} + \hat{\beta}_2 X_0 + t_{1-\alpha/2}(n-2) \cdot S_{(Y_0 - \hat{Y}_0)} \right] = 1 - \alpha$$



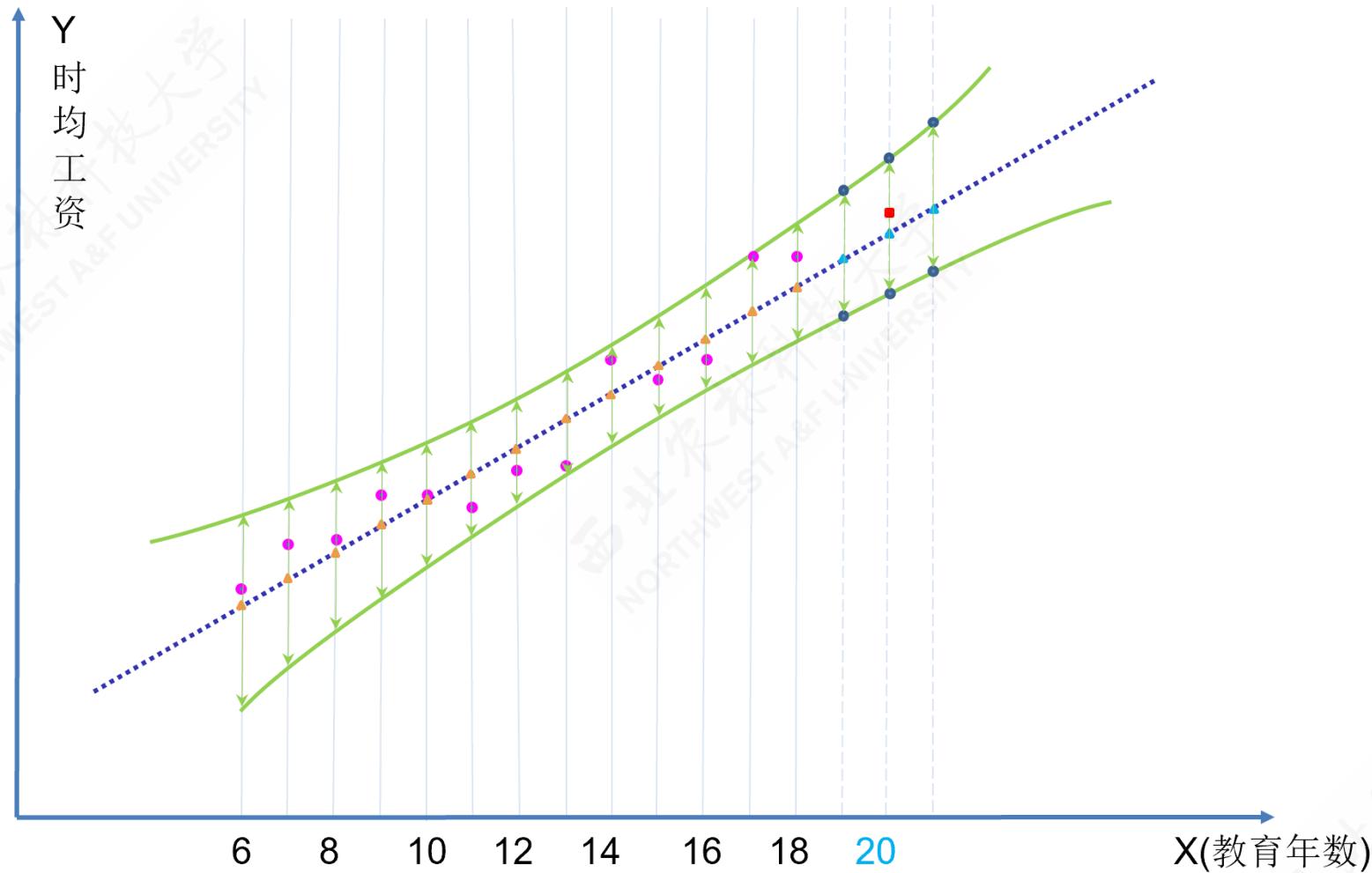
置信带

置信带(confidence interval): 对所有的X值，分别进行均值和个值分别进行预测，就能得到：

- 均值预测的置信带——总体回归函数的置信带
- 个值预测的置信带
- 预测如何可信？
 - 均值预测置信区间
 - 均值预测置信带
- 样本内置信带。——检验可靠性
- 样本外置信带。——预测未来值范围

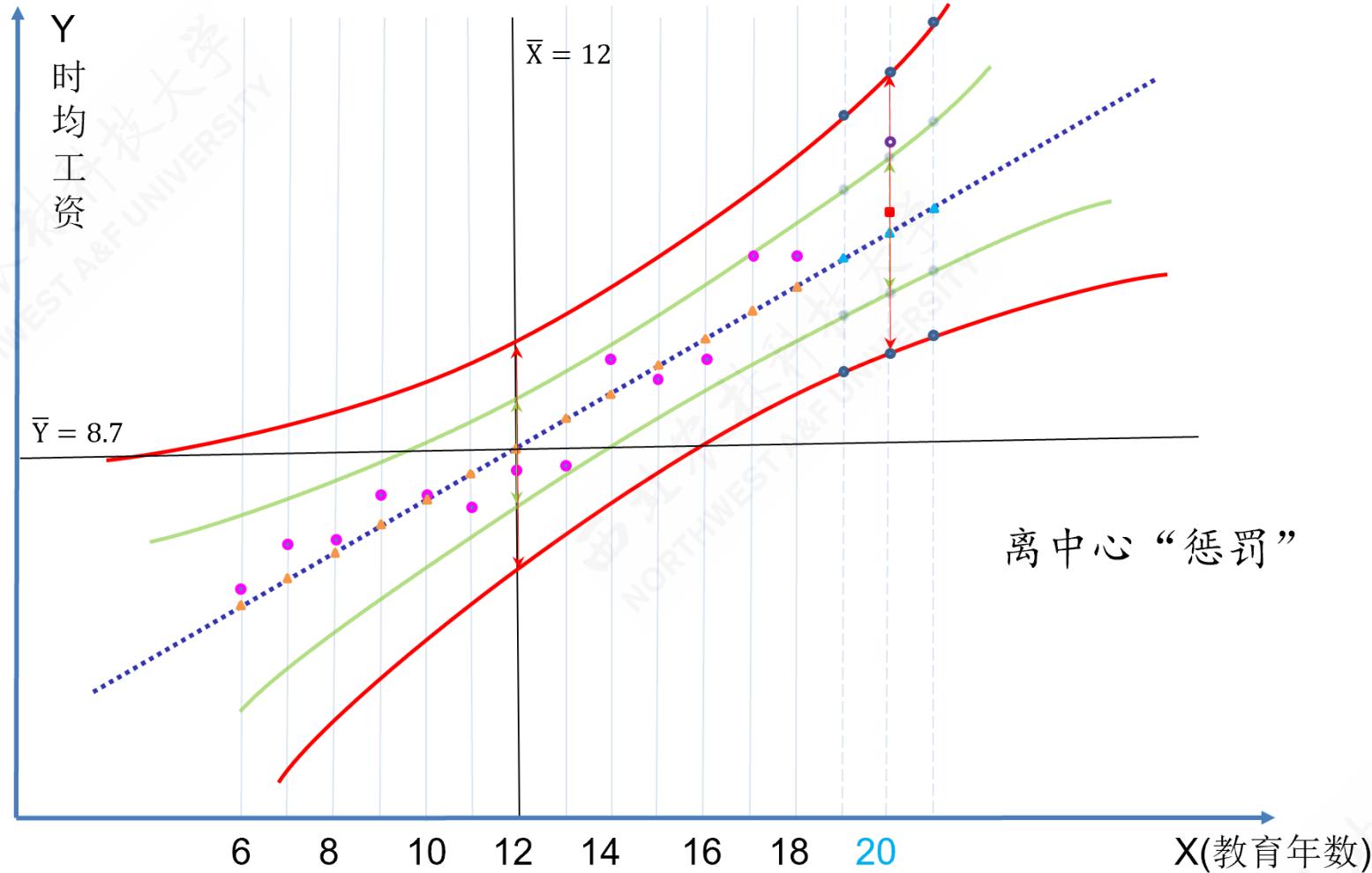


置信带





置信带





置信带

如何理解置信带？

- 谁更宽？——均值预测更准确
- 何处最窄？——中心点 $(\bar{X}, \bar{Y}) = (12, 8.67)$ 是历史信息的集中代表。



总结与思考

内容总结：

- 回归预测基于一套坚实严密的“底座”：OLS估计方法、CLRM假设、BLUE估计性质
- 均值预测置信带和个值预测置信带，是对预测可信度的形象表达。
- （同等条件下）均值预测比个值预测更准确（置信带宽窄）

课堂思考：

- 同样是95%置信度区间，两个人的认识是一样的么？

课后作业：工资与教育案例扩展

- 请计算置信度 $100(1 - \alpha) = 95\%$ 下， $X_0 = 20$ 时均值的置信区间。与 $100(1 - \alpha) = 90\%$ 时相比，有什么差异？
- 99%更值得可信么？

1.8 一个数值例子



如何把计量分析思想的进行软件验证？

说一千道一万，重要的是我们自己能不能动手，利用统计软件对前述的计量分析理论进行计算和验证！

下面，我们利用样本数据对教育和工资案例进行一次完整的计算和验证吧！

| 教育和工资案例的总体回归模型（PRM）如下：

$$\begin{aligned}Wage_i &= \beta_1 + \beta_2 Edu_i + u_i \\Y_i &= \beta_1 + \beta_2 X_i + u_i\end{aligned}$$



计算表的制作

obs	X_i	Y_i	$X_i Y_i$	X_i^2	Y_i^2	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	6.00	4.46	26.74	36.00	19.86	-6.00	-4.22	25.31	36.00	17.79
2	7.00	5.77	40.39	49.00	33.29	-5.00	-2.90	14.52	25.00	8.44
3	8.00	5.98	47.83	64.00	35.74	-4.00	-2.70	10.78	16.00	7.27
4	9.00	7.33	65.99	81.00	53.75	-3.00	-1.34	4.03	9.00	1.80
5	10.00	7.32	73.18	100.00	53.56	-2.00	-1.36	2.71	4.00	1.84
6	11.00	6.58	72.43	121.00	43.35	-1.00	-2.09	2.09	1.00	4.37
7	12.00	7.82	93.82	144.00	61.12	0.00	-0.86	-0.00	0.00	0.73
8	13.00	7.84	101.86	169.00	61.39	1.00	-0.84	-0.84	1.00	0.70
9	14.00	11.02	154.31	196.00	121.49	2.00	2.35	4.70	4.00	5.51
10	15.00	10.67	160.11	225.00	113.93	3.00	2.00	6.00	9.00	4.00
11	16.00	10.84	173.38	256.00	117.42	4.00	2.16	8.65	16.00	4.67
12	17.00	13.62	231.46	289.00	185.37	5.00	4.94	24.70	25.00	24.41
13	18.00	13.53	243.56	324.00	183.09	6.00	4.86	29.14	36.00	23.58
sum	156.00	112.77	1485.04	2054.00	1083.38	0.00	0.00	131.79	182.00	105.12



计算回归系数

公式1：(Favorite Five, FF形式)

$$\begin{aligned}\hat{\beta}_2 &= \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} \\ &= \frac{13 * 1485.04 - 156 * 112.771}{13 * 2054 - 156^2} = 0.7241\end{aligned}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 8.6747 - 0.7241 * 12 = -0.0145$$

公式2：(离差形式, favorite five, ff形式)

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{131.786}{182} = 0.7241$$

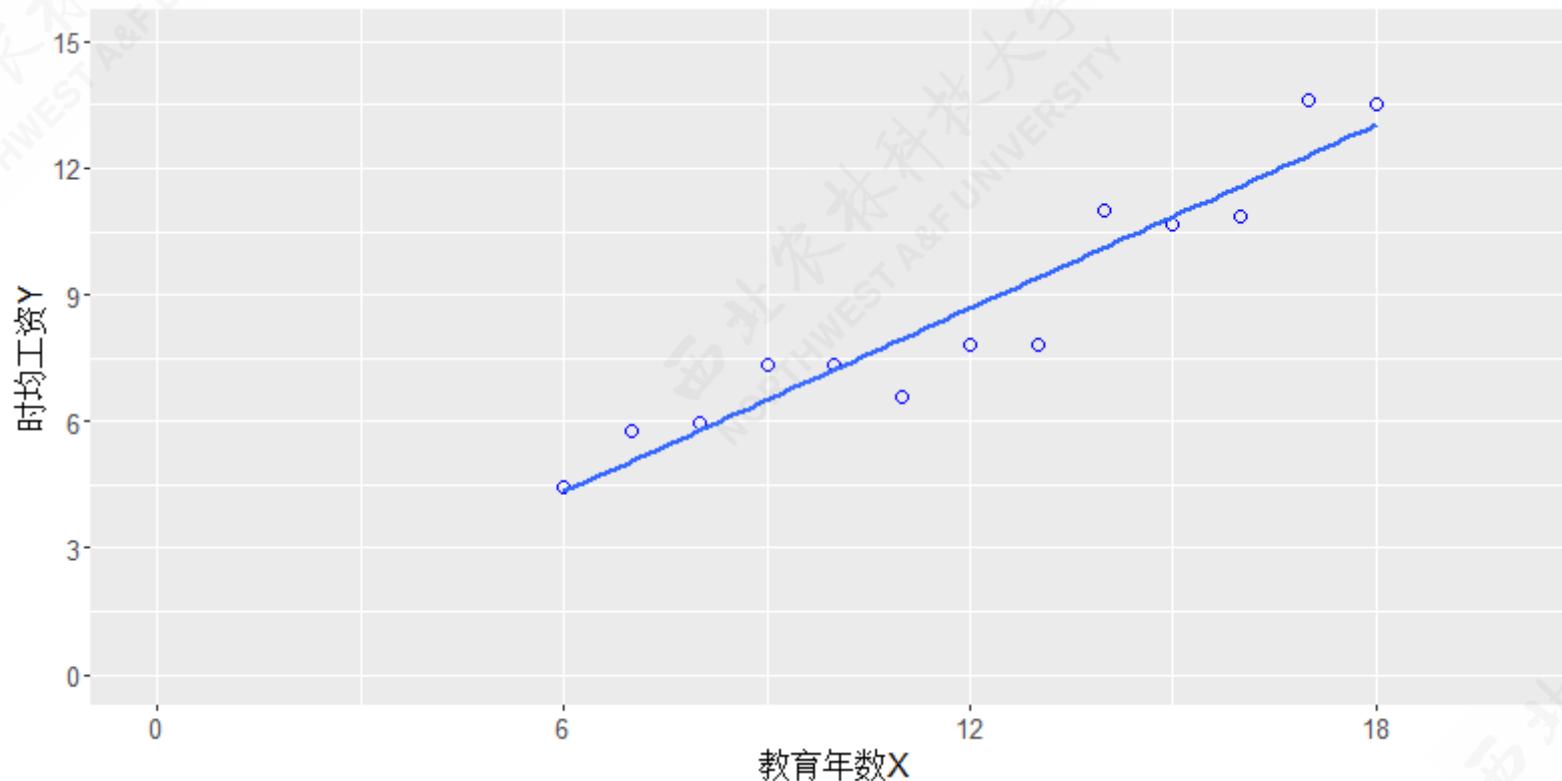
$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 8.6747 - 0.7241 * 12 = -0.0145$$



样本回归结果

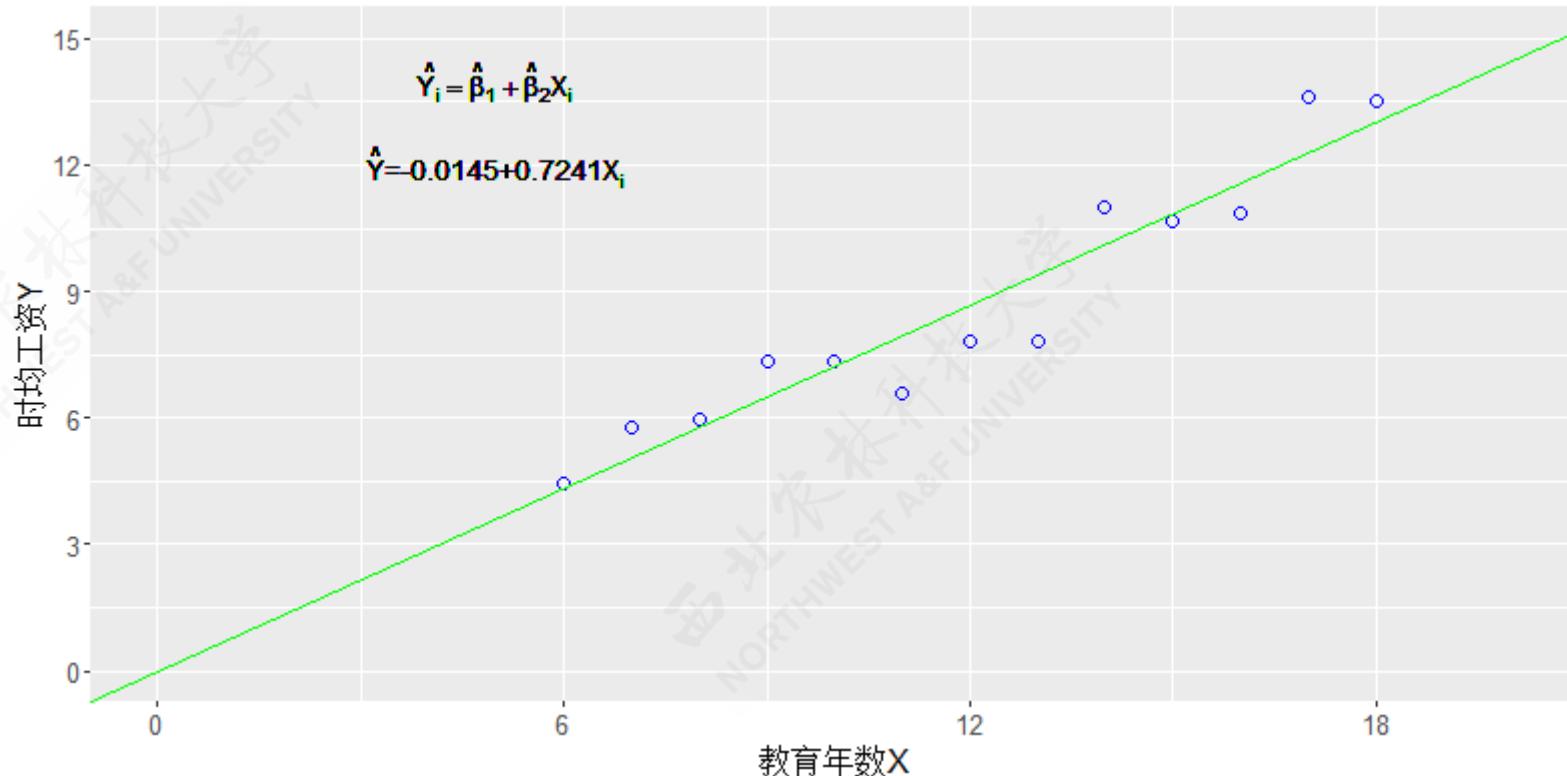
样本回归方程SRF:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i = -0.0145 + 0.7241 X_i$$





样本回归线SRL





计算得到拟合值和残差

obs	`X _i `	`Y _i `	`Ŷ _i `	`e _i `	`e _i ² `
1	6.0000	4.4567	4.3301	0.1266	0.0160
2	7.0000	5.7700	5.0542	0.7158	0.5123
3	8.0000	5.9787	5.7783	0.2004	0.0402
4	9.0000	7.3317	6.5024	0.8293	0.6877
5	10.0000	7.3182	7.2265	0.0917	0.0084
6	11.0000	6.5844	7.9506	-1.3662	1.8665
7	12.0000	7.8182	8.6747	-0.8565	0.7336
8	13.0000	7.8351	9.3988	-1.5637	2.4452
9	14.0000	11.0223	10.1229	0.8994	0.8089
10	15.0000	10.6738	10.8470	-0.1732	0.0300
11	16.0000	10.8361	11.5711	-0.7350	0.5402
12	17.0000	13.6150	12.2952	1.3198	1.7419
13	18.0000	13.5310	13.0193	0.5117	0.2618
sum	156.0000	112.7712	112.7712	0.0000	9.6928

根据以上样本回归方程，可以计算得到 Y_i 的回归拟合值 \hat{Y}_i ，以及回归残差 e_i 。

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

$$e_i = Y_i - \hat{Y}_i$$



回归误差方差和标准差

回归误差方差 $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{(n - 2)} = \frac{9.693}{11} = 0.8812$$

回归误差标准差 $\hat{\sigma}$:

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{(n - 2)}} = \sqrt{0.8812} = 0.9387$$



计算回归系数的样本方差

$$S_{\hat{\beta}_2}^2 = \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{0.8812}{182} = 0.0048$$

$$S_{\hat{\beta}_2} = \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2}} = \sqrt{0.0048} = 0.0696$$

$$S_{\hat{\beta}_1}^2 = \frac{\sum X_i^2}{n} \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{2054}{13} \frac{0.8812}{182} = 0.765$$

$$S_{\hat{\beta}_1} = \sqrt{\frac{\sum X_i^2}{n} \frac{\hat{\sigma}^2}{\sum x_i^2}} = \sqrt{0.765} = 0.8746$$



计算平方和分解

$$TSS = \sum (Y_i - \bar{Y})^2 = 105.1183$$

$$RSS = \sum (Y_i - \hat{Y}_i)^2 = 9.693$$

$$ESS = \sum (\hat{Y}_i - \bar{Y})^2 = 95.4253$$



相关系数和判定系数

样本相关系数 r :

$$r = \frac{S_{XY}^2}{S_X * S_Y} = \frac{10.9821}{3.8944 * 2.9597} = 0.9528$$

回归方程的判定系数 r^2 :

$$r^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{9.693}{105.1183} = 0.9078$$

二者关系



计算回归系数的置信区间

下面我们进一步计算回归系数的置信区间：

那么，截距参数 β_1 的 95% 置信区间为：

$$\begin{aligned}\hat{\beta}_1 - t_{1-\alpha/2} \cdot S_{\hat{\beta}_1} &\leq \beta_1 \leq \hat{\beta}_1 + t_{1-\alpha/2} \cdot S_{\hat{\beta}_1} \\ -0.0145 - 2.201 * 0.8746 &\leq \beta_1 \leq -0.0145 + 2.201 * 0.8746 \\ -1.9395 &\leq \beta_1 \leq 1.9106\end{aligned}$$

那么，斜率参数 β_2 的 95% 置信区间为：

$$\begin{aligned}\hat{\beta}_2 - t_{1-\alpha/2} \cdot S_{\hat{\beta}_2} &\leq \beta_2 \leq \hat{\beta}_2 + t_{1-\alpha/2} \cdot S_{\hat{\beta}_2} \\ 0.7241 - 2.201 * 0.0696 &\leq \beta_2 \leq 0.7241 + 2.201 * 0.0696 \\ 0.5709 &\leq \beta_2 \leq 0.8772\end{aligned}$$



计算随机干扰项方差的置信区间

- 给定 $\alpha = 0.05$, $(1 - \alpha)100\% = 95\%$
- 查卡方分布表可知:
 - $\chi_{\alpha/2}^2(n - 2) = \chi_{0.05/2}^2(11) = \chi_{0.025}^2(11) = 3.8157$
 - $\chi_{1-\alpha/2}^2(n - 2) = \chi_{1-0.05/2}^2(11) = \chi_{0.975}^2(11) = 21.9200$

们之前已算出回归误差方差 $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = 0.8812$ 。因此可以算出 σ^2 的 95% 置信区间为:

$$\begin{aligned}(n - 2) \frac{\hat{\sigma}^2}{\chi_{\alpha/2}^2} &\leq \sigma^2 \leq (n - 2) \frac{\hat{\sigma}^2}{\chi_{1-\alpha/2}^2} \\ 11 * \frac{0.8812}{21.92} &\leq \sigma^2 \leq 11 * \frac{0.8812}{3.8157} \\ 0.4422 &\leq \sigma^2 \leq 2.5403\end{aligned}$$



置信区间检验法

对于斜率参数 β_2 的置信区间检验法。

- 步骤1：给出模型，并提出假设：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$H_0 : \beta_2 = 0.5; \quad H_1 : \beta_2 \neq 0.5$$

- 步骤2：给定 $\alpha = 0.05$, $(1 - \alpha)100\% = 95\%$
- 步骤3：根据前述计算结果，计算斜率参数 β_2 的 95% 置信区间为：

$$\begin{aligned} \hat{\beta}_2 - t_{1-\alpha/2} \cdot S_{\hat{\beta}_2} &\leq \beta_2 \leq \hat{\beta}_2 + t_{1-\alpha/2} \cdot S_{\hat{\beta}_2} \\ 0.5709 &\leq \beta_2 \leq 0.8772 \end{aligned}$$

- 步骤4：那么我们可以对斜率参数 β_2 做出如下检验判断：
 - 拒绝原假设 H_0 , 接受 H_1 。认为，长期来看很多个区间 $[0.5709, 0.8772]$ 有 95% 的可能性不包含 0.5 ($\beta_2 \neq 0.5$)。



置信区间检验法

对于截距参数 β_1 的置信区间检验法。

- 步骤1：给出模型，并提出假设：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

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

- 步骤2：给定 $\alpha = 0.05$, $(1 - \alpha)100\% = 95\%$
- 步骤3：根据前述计算结果，计算截距参数 β_1 的95%置信区间为：

$$\begin{aligned} \hat{\beta}_1 - t_{1-\alpha/2} \cdot S_{\hat{\beta}_1} &\leq \beta_1 \leq \hat{\beta}_1 + t_{1-\alpha/2} \cdot S_{\hat{\beta}_1} \\ -1.9395 &\leq \beta_1 \leq 1.9106 \end{aligned}$$

- 步骤4：那么我们可以对截距参数 β_1 做出如下检验判断：
 - 不能拒绝原假设 H_0 ，暂时接受 H_0 。认为，长期来看很多个区间 $[-1.9395, 1.9106]$ 有95%的可能性包含0 ($\beta_1 = 0$)。



进行t检验

我们之前已算出：

- 回归系数： $\hat{\beta}_1 = -0.0145$; $\hat{\beta}_2 = 0.7241$; $\hat{\sigma}^2 = 0.8812$ 。
- 回归误差方差： $\hat{\sigma}^2 = 0.8812$ 。
- 回归系数的样本方差： $S_{\hat{\beta}_1}^2 = \frac{\sum X_i^2}{n} \cdot \frac{\hat{\sigma}^2}{\sum x_i^2} = 0.7650$; $S_{\hat{\beta}_2}^2 = \frac{\hat{\sigma}^2}{\sum x_i^2} = 0.0048$;
- 回归系数的样本标准差： $S_{\hat{\beta}_1} = 0.8746$; $S_{\hat{\beta}_2} = 0.0696$ 。

给定 $\alpha = 0.05$, $(1 - \alpha)100\% = 95\%$, 我们可以查t分布表得到理论参照值：

$$t_{1-\alpha/2}(n - 2) = t_{1-0.05/2}(11) = 2.2010$$



截距参数的t检验

- 步骤1：给出模型，并提出假设：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

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

- 步骤2：构造合适的检验统计量

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

- 步骤3：基于原假设 H_0 计算出样本统计量。

$$t^*_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{-0.0145}{0.8746} = -0.0165 \quad \leftarrow H_0 : \beta_1 = 0$$



截距参数的t检验

- 步骤4：给定显著性水平 $\alpha = 0.05$ 下，查出统计量的理论分布值。

$$t_{1-\alpha/2}(n-2) = t_{1-0.05/2}(13-2) = t_{0.975}(11) = 2.2010$$

- 步骤5：得到显著性检验的判断结论。因为 $|t_{\hat{\beta}_1}^*| = 0.0165$ 小于 $t_{0.975}(11) = 2.2010$ 。因此，认为 β_1 的 t 检验结果不显著。换言之，在显著性水平 $\alpha = 0.05$ 下，不能显著地拒绝原假设 H_0 ，只能暂时接受原假设 H_0 ，认为截距参数 $\beta_1 = 0$ 。



斜率参数的t检验

- 步骤1：给出模型，并提出假设：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$H_0 : \beta_2 = 0; \quad H_1 : \beta_2 \neq 0$$

- 步骤2：构造合适的检验统计量

$$T = \frac{\hat{\beta}_2 - \beta_2}{S_{\beta_2}} \leftarrow T \sim t(n-2)$$

- 步骤3：基于原假设 H_0 计算出样本统计量。

$$t^*_{\hat{\beta}_2} = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}} = \frac{0.7241}{0.0696} = 10.4064 \quad \leftarrow H_0 : \beta_2 = 0$$



斜率参数的t检验

- 步骤4：给定显著性水平 $\alpha = 0.05$ 下，查出统计量的理论分布值。

$$t_{1-\alpha/2}(n-2) = t_{1-0.05/2}(13-2) = t_{0.975}(11) = 2.2010$$

- 步骤5：得到显著性检验的判断结论。 $|t^*_{\hat{\beta}_2}| = 10.4064$ 大于 $t_{0.975}(11) = 2.2010$ 。因此，认为 β_2 的 t 检验结果显著。换言之，在显著性水平 $\alpha = 0.05$ 下，应显著地拒绝原假设 H_0 ，接受备择假设 H_1 ，认为斜率参数 $\beta_2 \neq 0$ 。



计算方差分析(ANOVA)表

教育程度与时均工资案例的ANOVA分析表

变异来源	平方和SS	自由度df	均方和MSS
回归平方和ESS	95.425	1	95.425
残差平方和RSS	9.693	11	0.881
总平方和TSS	105.118	12	7.086



模型整体显著性检验

- 步骤1：给出模型 $Y_i = \beta_1 + \beta_2 X_i + u_i$, 提出假设: $H_0 : \beta_2 = 0$; $H_1 : \beta_2 \neq 0$
- 步骤2：构造合适检验的分布：

$$F = \frac{\left(\hat{\beta}_2 - \beta_2\right)^2 \sum x_i^2}{\sum e_i^2 / (n - 2)} \leftarrow F \sim F(1, n - 2)$$

- 步骤3：基于原假设 $H_0 : \beta_2 = 0$, 可以计算出样本统计量。

$$F^* = \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum e_i^2 / (n - 2)} = \frac{ESS/df_{ESS}}{RSS/df_{RSS}} = \frac{MSS_{ESS}}{MSS_{RSS}} = \frac{95.4253}{0.8812} = 108.2924$$



模型整体显著性检验

- 步骤4：给定 $\alpha = 0.05$ 下，查出 F 理论值 $F_{1-\alpha}(1, n - 2) = F_{0.95}(1, 11) = 4.8443$
- 步骤5：得到显著性检验的判断结论。因为 $F^* = 108.2924$ 大于 $F_{0.95}(1, 11) = 4.8443$ ，所以模型整体显著性的 F 检验结果显著。换言之，在显著性水平 $\alpha = 0.05$ 下，应显著地拒绝原假设 H_0 ，接受备择假设 H_1 ，认为斜率参数 $\beta_2 \neq 0$ 。



回归预测：均值预测

给定 $X_0 = 20$ 时，根据早前计算结果： $\hat{\sigma}^2 = 0.8812$; $\bar{X} = 12.0000$; $\sum x_i^2 = 182.0000$ 。因此可以得到：

$$S_{\hat{Y}_0}^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right] = 0.8812 \left(\frac{1}{13} + \frac{(20 - 12)^2}{182} \right) = 0.3776; \quad S_{\hat{Y}_0} = \sqrt{S_{\hat{Y}_0}^2} = 0.6145$$

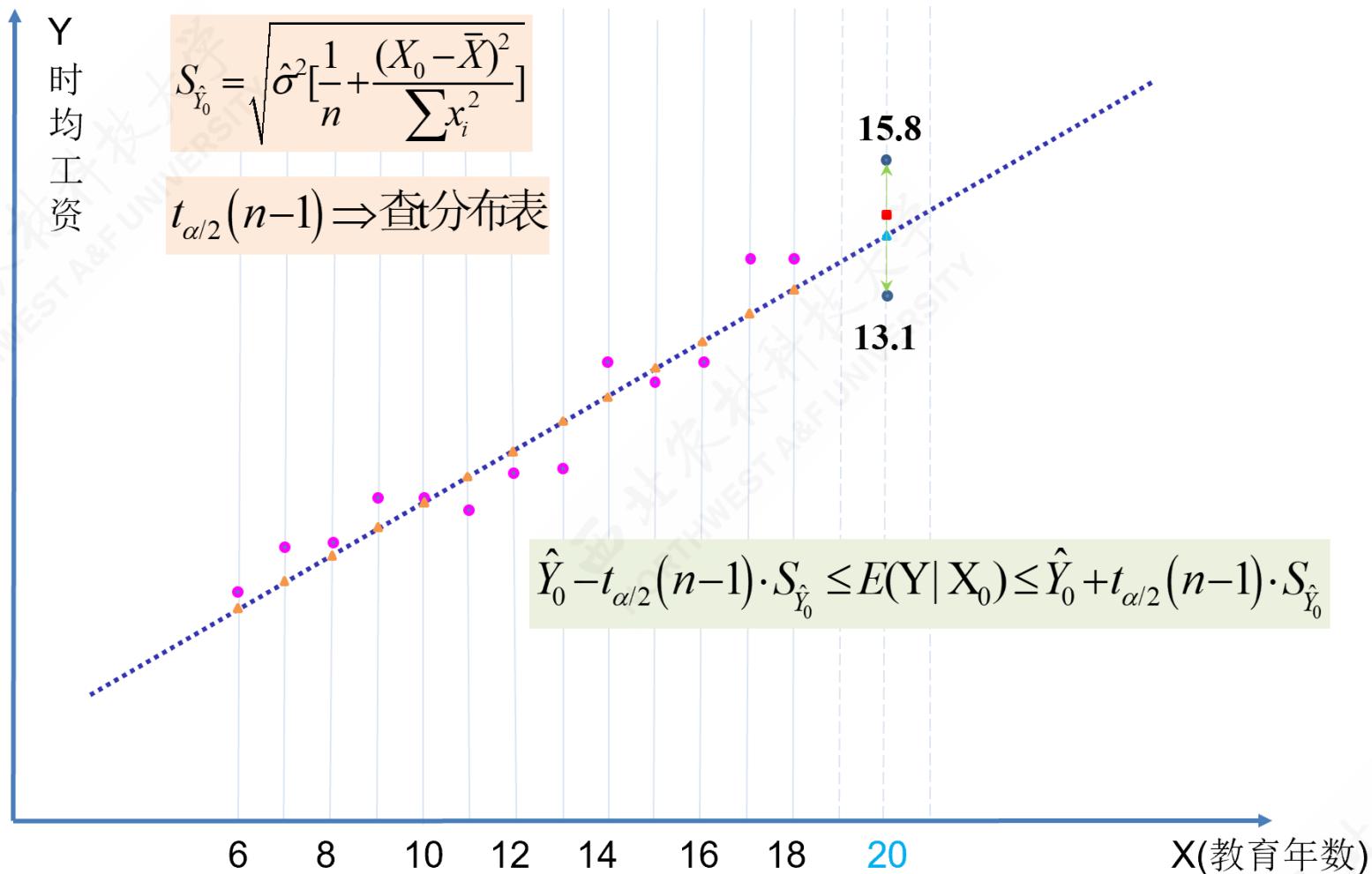
$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 = -0.0145 + 0.7241 * 20 = 14.4675$$

因此，可以计算得到均值 $E(Y|X=20)$ 置信区间为：

$$\begin{aligned} \hat{\beta} + \hat{\beta}_2 X_0 - t_{1-\alpha/2}(n-2) \cdot S_{\hat{Y}_0} &\leq E(Y|X_0) \leq \hat{\beta} + \hat{\beta}_2 X_0 + t_{1-\alpha/2}(n-2) \cdot S_{\hat{Y}_0} \\ 14.4675 - 1.7959 * 0.6145 &\leq E(Y|X_0 = 20) \leq 14.4675 + 1.7959 * 0.6145 \\ 13.3639 &\leq E(Y|X_0 = 20) \leq 15.5711 \end{aligned}$$



回归预测：均值预测





回归预测：个值预测

给定 $X_0 = 20$ 时，根据早前计算结果： $\hat{\sigma}^2 = 0.8812$; $\bar{X} = 12.0000$; $\sum x_i^2 = 182.0000$ 。因此可以得到：

$$S_{(Y_0 - \hat{Y}_0)}^2 = \hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right] = 0.8812 \left(1 + \frac{1}{13} + \frac{(20 - 12)^2}{182} \right) = 1.2588$$
$$S_{\hat{Y}_0} = \sqrt{S_{(Y_0 - \hat{Y}_0)}^2} = 1.122$$

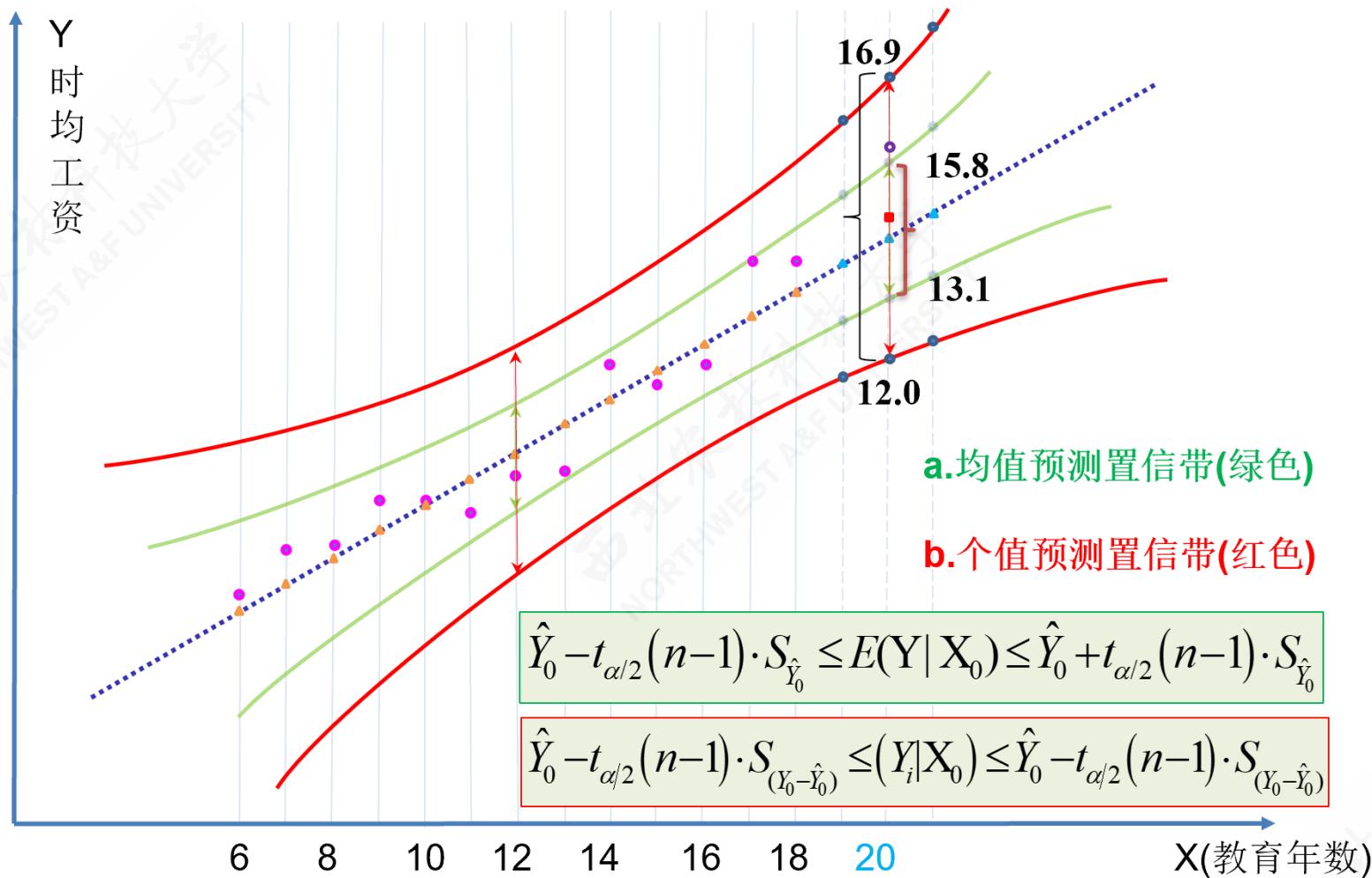
$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 = -0.0145 + 0.7241 * 20 = 14.4675$$

因此，可以计算得到个值 ($Y_0 | X = 20$) 置信区间为：

$$\hat{\beta} + \hat{\beta}_2 X_0 - t_{1-\alpha/2}(n-2) \cdot S_{(Y_0 - \hat{Y}_0)} \leq Y_0 | X = X_0 \leq \hat{\beta} + \hat{\beta}_2 X_0 + t_{1-\alpha/2}(n-2) \cdot S_{(Y_0 - \hat{Y}_0)}$$
$$14.4675 - 1.7959 * 1.122 \leq Y_0 | X_0 = 20 \leq 14.4675 + 1.7959 * 1.122$$
$$12.4525 \leq Y_0 | X_0 = 20 \leq 16.4824$$



回归预测：个值预测





本节内容小结

下面对本节内容做一个小结：

- 通过样本数值例子，我们可以全部自己“手工计算”OLS方法参数估计、估计的精度（估计量样本方差）、变异的平方和分解，以及判定系数等过程环节。
- 实际上任何一款统计软件都会很快帮我们完成这些计算过程，实证分析中我们不需要亲自去计算它们。
- 送上一句忠告，统计软件就像一个“技术黑箱”，如果你不理解“黑箱”里面的运作原理，那么你就永远只是被它“牵着鼻子走”！所以，大家起码要自己手工计算一遍！
- 而且，随着开源软件（如R或Python等）的普遍流行，“你”的介入和作用可能越来越重要！因为你可以更加灵活、更加自由地进行改造、重塑和创新！



本节内容思考

下面提出若干问题与思考：

- 请大家自己使用任何熟悉的统计软件，完成本节的所有环节的计算！
- 如果你和其他同学的随机样本数据，都来自同一个总体，你们的计算结果会是一样的么？全班同学全部计算结果，会给“从样本推断总体”带来什么启示？

1.9 报告回归分析结果



回归分析的形式

课程要求：会熟练、正确阅读统计软件给出的各类分析报告，理解其中的关键信息和内涵。这些分析报告包括：传统的多元回归分析报告；以及各种计量检验的辅助分析报告（如异方差white检验报告）等。

根据统计软件的不同（stata；Eview；R……），各种分析报告呈现形式略有差异，但基本要素和信息都大抵一致。

给定如下一元回归模型：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$



回归分析的形式（多行方程表达法）

形式1：多行方程表达法（整理好的精炼报告）：

根据统计软件的原始报告，往往是选取最关键的信息，经过整理并以多行样本回归方程（SRF）的形式呈现。



回归分析的形式（多行方程表达法）

例如，一种精炼报告的具体形式写为：

$$\begin{aligned}\hat{Y} &= -0.01 + 0.72X \\(t) &\quad (-0.0165) \quad (10.4065) \\(se) &\quad (0.8746) \quad (0.0696) \\(\text{fitness}) R^2 &= 0.9078; \bar{R}^2 = 0.8994 \\F^* &= 108.29; p = 0.0000\end{aligned}$$

- 第1行表示样本回归函数（回归系数）
- 第2行(t)表示回归系数对应的样本t统计量 ($t_{\hat{\beta}_i}^*, i \in 1, 2, \dots, k$)
- 第3行(se)表示回归系数对应的样本标准误差 ($S_{\hat{\beta}_i}, i \in 1, 2, \dots, k$)
- 第4行(fitness)表示回归模型拟合情况和统计检验的简要信息，其中 R^2 表示判定系数， \bar{R}^2 表示调整判定系数，F表示模型整体显著性检验中的样本F统计量值 (F^*)，p 表示样本F统计量值对应的概率值。



回归分析的形式（表格列示法）

形式2：表格列示法（整理好的精炼报告）：根据统计软件的原始报告，往往是选取最关键的信息，经过整理以表格形式呈现，表格列示法的形式呈现为：

term	estimate	std.error	statistic	p.value
(Intercept)	-0.0144527	0.8746239	-0.0165245	0.9871118
X	0.7240967	0.0695813	10.4064779	0.0000005

- 第1列：`term`表示回归模型中包含的变量，也即 $X_{2i}, X_{3i}, \dots, X_{ki}$ ，其中截距项默认为 (Intercept)。
- 第2列：`estimate`表示回归系数的估计值，也即 $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ 。
- 第3列：`std.error`表示回归系数对应的样本标准误差，也即 $S_{\hat{\beta}_i}, i \in 1, 2, \dots, k$ 。
- 第4列：`statistic`表示回归系数对应的样本t统计量，也即 $t^*_{\hat{\beta}_i}, i \in 1, 2, \dots, k$
- 第5列：`p.value`表示回归系数样本t统计量对应的概率值，也即 $Pr(t = t^*_{\hat{\beta}_i}) = p$



回归分析的形式（EViews软件原始报告）

形式3：原始报告：分析软件如EViews、R、STATA等直接自动生成的多元回归分析报告。EViews软件原始分析报告形式如下：抬头区域

Equation: EQ_WAGE Workfile: CHPT2::wage\				
View Proc Object Print Name Freeze Estimate Forecast Stats Resids				
Dependent Variable: Y				
Method: Least Squares				
Date: 03/09/19 Time: 10:55				
Sample: 1 13				
Included observations: 13				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.014453	0.874624	-0.016525	0.9871
X	0.724097	0.069581	10.40648	0.0000
R-squared	0.907791	Mean dependent var	8.674708	
Adjusted R-squared	0.899409	S.D. dependent var	2.959706	
S.E. of regression	0.938704	Akaike info criterion	2.852004	
Sum squared resid	9.692810	Schwarz criterion	2.938920	
Log likelihood	-16.53803	Hannan-Quinn criter.	2.834139	
F-statistic	108.2948	Durbin-Watson stat	1.737984	
Prob(F-statistic)	0.000000			

- Dependent Variable: Y: 因变量
- Method: Least Squares: 分析方法
- Date: 03/09/19 Time: 10:55: 分析的时间
- Sample: 1 13: 样本范围
- Included observations: 13: 样本数n



回归分析的形式 (EViews软件原始报告)

形式3：原始报告：分析软件如EViews、R、STATA等直接自动生成的多元回归分析报告。EViews软件原始分析报告形式如下：三线表区域

Equation: EQ_WAGE Workfile: CHPT2::wage\				
View	Proc	Object	Print	Name
Dependent Variable: Y				
Method: Least Squares				
Date: 03/09/19	Time: 10:55			
Sample: 1 13				
Included observations: 13				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.014453	0.874624	-0.016525	0.9871
X	0.724097	0.069581	10.40648	0.0000
R-squared	0.907791	Mean dependent var	8.674708	
Adjusted R-squared	0.899409	S.D. dependent var	2.959706	
S.E. of regression	0.938704	Akaike info criterion	2.852004	
Sum squared resid	9.692810	Schwarz criterion	2.938920	
Log likelihood	-16.53803	Hannan-Quinn criter.	2.834139	
F-statistic	108.2948	Durbin-Watson stat	1.737984	
Prob(F-statistic)	0.000000			

- 第1列：Variable表示模型包含的变量， $X_{2i}, X_{3i}, \dots, X_{ki}$ ，其中截距项默认为C。
- 第2列：Coefficient回归系数，也即 $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ ；
- 第3列：Std. Error回归系数的样本标准误差，也即 $S_{\hat{\beta}_i}, i \in 1, 2, \dots, k$ 。
- 第4列：t-Statistic表示回归系数对应的样本t统计量，也即 $t^*_{\hat{\beta}_i}, i \in 1, 2, \dots, k$ ；



回归分析的形式（EViews软件原始报告）

形式3：原始报告：分析软件如EViews、R、STATA等直接自动生成的多元回归分析报告。EViews软件原始分析报告形式如下：指标值区域（左）

Equation: EQ_WAGE Workfile: CHPT2::wage\				
View	Proc	Object	Print	Name
Dependent Variable: Y				
Method: Least Squares				
Date: 03/09/19				
Time: 10:55				
Sample: 1 13				
Included observations: 13				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.014453	0.874624	-0.016525	0.9871
X	0.724097	0.069581	10.40648	0.0000
R-squared	0.907791	Mean dependent var	8.674708	
Adjusted R-squared	0.899409	S.D. dependent var	2.959706	
S.E. of regression	0.938704	Akaike info criterion	2.852004	
Sum squared resid	9.692810	Schwarz criterion	2.938920	
Log likelihood	-16.53803	Hannan-Quinn criter.	2.834139	
F-statistic	108.2948	Durbin-Watson stat	1.737984	
Prob(F-statistic)	0.000000			

- R-squared: 回归判定系数 R^2 。
- Adjusted R-squared: 回归模型调整判定系数 \bar{R}^2 。
- S.E. of regression: 回归模型的回归误差标准差 $\hat{\sigma}$ 。
- Sum squared resid: 回归模型的残差平方和 RSS $RSS = \sum e_i^2$ 。
- Log likelihood: 回归模型的对数似然值。
- F-statistic: 回归模型整体显著性的样本F统计量 F^* 。



回归分析的形式（EViews软件原始报告）

形式3：原始报告：分析软件如EViews、R、STATA等直接自动生成的多元回归分析报告。EViews软件原始分析报告形式如下：指标值区域（右）

Equation: EQ_WAGE Workfile: CHPT2::wage\				
View	Proc	Object	Print	Name
Dependent Variable: Y				
Method: Least Squares				
Date: 03/09/19	Time: 10:55			
Sample: 1 13				
Included observations: 13				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.014453	0.874624	-0.016525	0.9871
X	0.724097	0.069581	10.40648	0.0000
R-squared	0.907791	Mean dependent var	8.674708	
Adjusted R-squared	0.899409	S.D. dependent var	2.959706	
S.E. of regression	0.938704	Akaike info criterion	2.852004	
Sum squared resid	9.692810	Schwarz criterion	2.938920	
Log likelihood	-16.53803	Hannan-Quinn criter.	2.834139	
F-statistic	108.2948	Durbin-Watson stat	1.737984	
Prob(F-statistic)	0.000000			

- Mean dependent var: Y均值 \bar{Y} 。
- S.D. dependent var: Y样本标准差 S_Y 。
- Akaike info criterion: AIC信息准则。
- Schwarz criterion: 回归的 Schwarz准则。
- Hannan-Quinn criter.: 回归的 Hannan-Quinn准则。
- Durbin-Watson stat: 回归的德宾沃森统计量d。



回归分析的形式（R软件原始报告）

形式4：原始报告：分析软件如EViews、R、STATA等直接自动生成的多元回归分析报告。R软件原始分析报告形式如下：

```
Call:  
lm(formula = mod_wage, data = data_wage)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-1.5637 -0.7350  0.1266  0.7158  1.3198  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -0.01445    0.87462  -0.017   0.987  
X             0.72410    0.06958   10.406 4.96e-07 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.9387 on 11 degrees of freedom  
Multiple R-squared:  0.9078,    Adjusted R-squared:  0.8994  
F-statistic: 108.3 on 1 and 11 DF,  p-value: 4.958e-07
```

本章结束

