统计学原理(Statistic)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2021-05-16

第五章 相关和回归分析

5.1 变量间关系的度量

5.2 回归分析的基本思想

5.3 OLS方法与参数估计

5.4 假设检验

5.5 拟合优度与残差分析

5.6 回归预测分析

5.7 回归报告解读

5.3 OLS方法与参数估计

普通最小二乘法（OLS）

参数估计

估计精度

区间估计

普通最小二乘法（OLS）：引子

我们如何估计回归函数中的系数？

总体回归：

$$\begin{cases} E(Y|X_i) = \beta_1 + \beta_2 X_i & \text{(PRF)} \\ Y_i = \beta_1 + \beta_2 X_i + u_i & \text{(PRM)} \end{cases}$$

样本回归：

$$\begin{cases} \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i & \text{(SRF)} \\ Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i & \text{(SRM)} \end{cases}$$

首先需要回答的问题是，我们该如何估计得出样本回归函数中的系数？事实上，方法有多种多样：

- 图解法：比较粗糙，但提供了基本的视觉认知
- 最小二乘法(order lease squares, OLS)：最常用的方法
- 最大似然法(maximum likelihood, ML)
- 矩估计方法(Moment method, MM)

普通最小二乘法（OLS）：回顾和比较

总体回归函数PRF：

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

总体回归模型PRM：

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

样本回归函数SRF：

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

样本回归模型SRM：

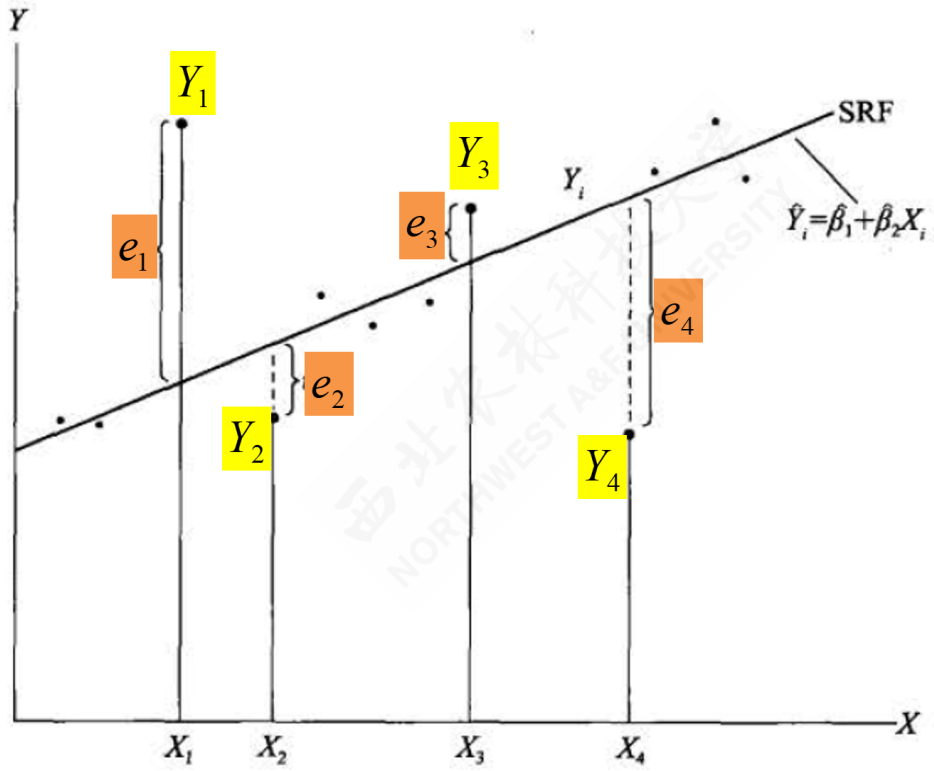
$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

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?

普通最小二乘法（OLS）：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（OLS）：原理

OLS的基本原理：残差平方和最小化。

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

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

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

（示例）普通最小二乘法（OLS）的一个数值试验

假设存在下面所示的4组观测值 (X_i, Y_i)：

X_i	Y_i
(1)	(2)
1	4
4	5
5	7
6	12
Sum：	

数值试验：数据

（示例）普通最小二乘法（OLS）的一个数值试验

假设猜想两个SRF，完成下表计算，并分析哪个SRF给出的 $(\hat{\beta}_1, \hat{\beta}_2)$要更好？

$$SRF1:\ \hat{Y}_{1i} = \hat{\beta}_1 + \hat{\beta}_2 X_i = 1.572 + 1.357 X_i$$

$$SRF2:\ \hat{Y}_{2i} = \hat{\beta}_1 + \hat{\beta}_2 X_i = 3.0 + 1.0 X_i$$

X_i	Y_i	$\hat{Y}_{1i}$	e_{1i}	e_{1i}^2	$\hat{Y}_{2i}$	e_{2i}	e_{2i}^2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	4	2.929	1.071	1.147	4	0	0
4	5	7.000	2.000	4.000	7	-2	4
5	7	8.357	1.357	1.841	8	-1	1
6	12	9.714	2.285	5.226	9	3	9
Sum:			0.000	12.214		0	14

数值试验：计算

参数估计：回归参数的OLS点估计

- 最小化求解：

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

- 方程组变形，得到正规方程组：

$$
\begin{cases}
\sum\left[\hat{\beta}_1 - (Y_i - \hat{\beta}_2 X_i)\right] = 0 \\
\sum\left[X_i^2\hat{\beta}_2 - (Y_i - \hat{\beta}_1)X_i\right] = 0
\end{cases}
$$

$$
\begin{cases}
\sum Y_i - n\hat{\beta}_1 - (\sum X_i)\hat{\beta}_2 = 0 \\
\sum X_i Y_i - (\sum X_i)\hat{\beta}_1 - (\sum X_i^2)\hat{\beta}_2 = 0
\end{cases}
$$

参数估计：回归参数的OLS点估计

进而得到回归系数的计算公式1（Favorite Five，FF）：

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

参数估计：回归参数的OLS点估计

此外我们也可以得到如下的离差公式(favorite five，ff)

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

其中离差计算 $x_i = X_i - \bar{X};\ y_i = Y_i - \bar{Y}$。

以下式子为什么是等价的？你能推导出来么？

$$\begin{cases} \sum x_i y_i = \sum \left[(X_i - \bar{X})(Y_i - \bar{Y}) \right] = \sum X_i Y_i - \frac{1}{n} \sum X_i \sum Y_i \\ \sum x_i^2 = \sum (X_i - \bar{X})^2 \qquad\qquad = \sum X_i^2 - \frac{1}{n} \left(\sum X_i \right)^2 \end{cases}$$

参数估计：随机干扰项参数的OLS点估计

PRM公式变形：

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

残差公式变形：

$$\left.\begin{aligned} e_i &= y_i - \hat{\beta}_2 x_i \\ e_i &= \beta_2 x_i + (u_i - \bar{u}) - \hat{\beta}_2 x_i \end{aligned}\right\} \Rightarrow$$
$$e_i = -(\hat{\beta}_2 - \beta_2) x_i + (u_i - \hat{u})$$

参数估计：随机干扰项参数的OLS点估计

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：

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

参数估计：随机干扰项参数的OLS点估计

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。

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

（附录）A过程证明

$$
\begin{aligned}
A &= \sum x_i^2 E\left[(\hat{\beta}_2 - \beta_2)^2\right] \\
&= \sum \left[x_i^2 \cdot var(\hat{\beta}_2)\right] \\
&= var(\hat{\beta}_2) \cdot \sum x_i^2 \\
&= \frac{\sigma^2}{\sum x_i^2} \cdot \sum x_i^2 \\
&= \sigma^2
\end{aligned}
$$

（附录）B过程证明

$$
\begin{aligned}
B = E\left[\sum (u-\bar{u})^2\right] &= E(\sum u_i^2) - 2E\left[\sum (u_i\bar{u})\right] + nE(\bar{u}^2) \\
&= n \cdot Var(u_i) - 2E\left[\sum (u_i \cdot \frac{\sum u_i}{n})\right] + nE(\frac{\sum u_i}{n})^2 \\
&= n\sigma^2 - 2E\left[\frac{\sum u_i}{n}\sum u_i\right] + E\left[\frac{(\sum u_i)^2}{n}\right] \\
&= n\sigma^2 - E\left[(\sum u_i)^2/n\right] = n\sigma^2 - \frac{E(u_i^2)+E(u_2^2)+\cdots+E(u_n^2))}{n} \\
&= n\sigma^2 - \frac{nVaru_i}{n} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2
\end{aligned}
$$

（附录）C过程证明

$$
\begin{aligned}
C &= -2E\left[(\hat{\beta}_2-\beta_2)\sum x_i(u_i-\bar{u})\right] \\
&= -2E\left[\frac{\sum x_iu_i}{\sum x_i^2}\left(\sum x_iu_i-\bar{u}\sum x_i\right)\right] \\
&= -2E\left[\frac{\left(\sum x_iu_i\right)^2}{\sum x_i^2}\right] \\
&= -2E\left[(\hat{\beta}_2-\beta_2)^2\right]=-2\sigma^2
\end{aligned}
$$

- 其中：

$$
\hat{\beta}_2=\sum k_iY_i=\sum k_i(\beta_1+\beta_2X_i+u_i)=\beta_1\sum k_i+\beta_2\sum k_iX_i+\sum k_iu_i=\beta_2+\sum k_iu_i
$$

$$
\hat{\beta}_2-\beta_2=\sum k_iu_i=\frac{\sum x_iu_i}{\sum x_i^2}
$$

（案例）计算表的和的

obs	`$X_i$`	`$Y_i$`	`$X_iY_i$`	`$X_i^2$`	`$Y_i^2$`	`$x_i$`	`$y_i$`	`$x_iy_i$`	`$x_i^2$`	`$y_i^2$`
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

公式1：（Favorite Five，FF形式）

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

$$= \frac{13 * 1485.04 - 156 * 112.771}{13 * 2054 - 156^2} = 0.7241$$

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

公式2：（离差形式，favorite five，ff形式）

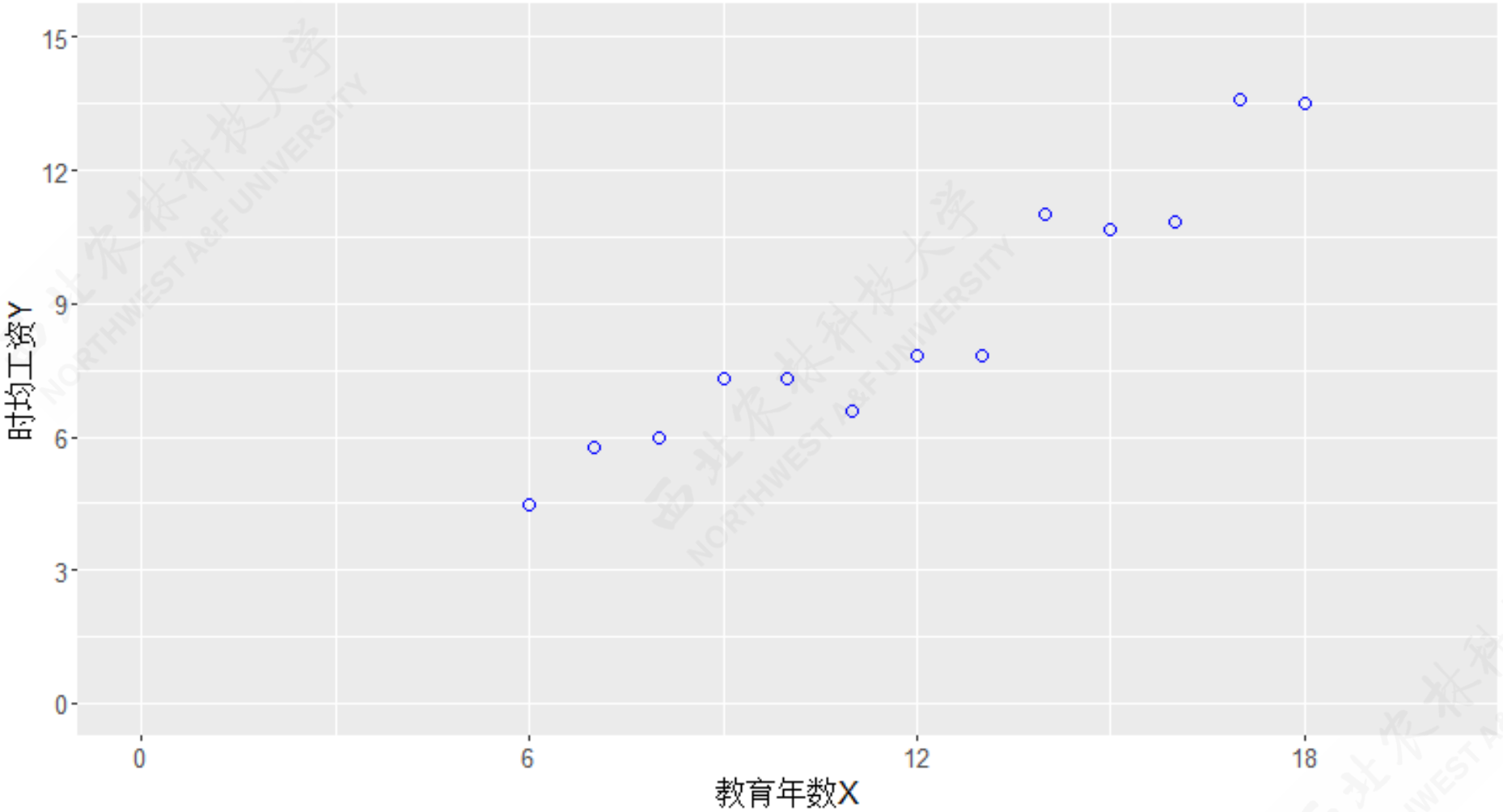
$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{131.786}{182} = 0.7241$$

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

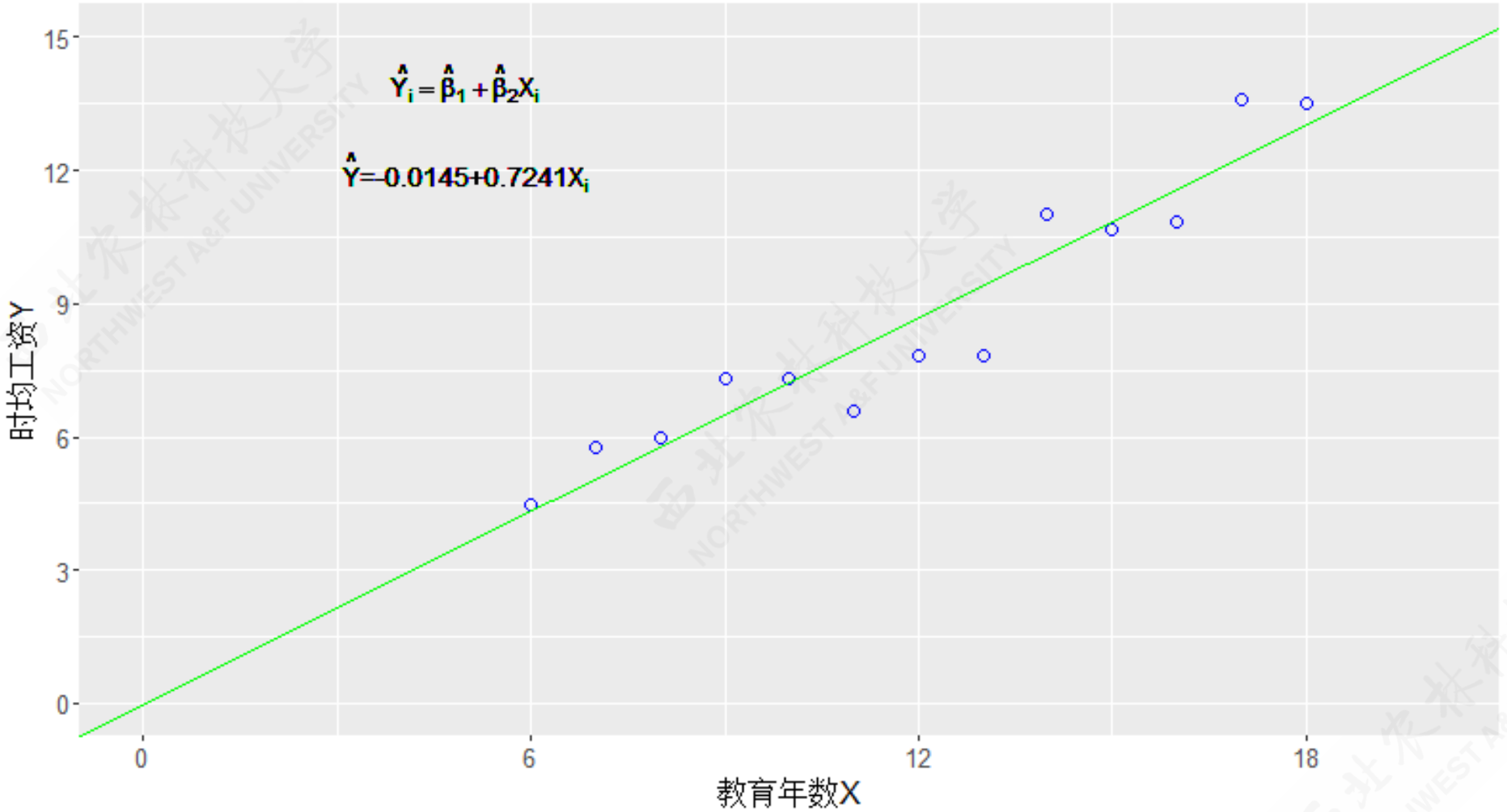
（案例）样本回归方程SRF

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

（案例）样本回归线SRL



（案例）样本回归线SRL


$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
$\hat{Y} = -0.0145 + 0.7241 X_i$
时均工资Y
教育年数X
0
3
6
9
12
15
0
6
12
18

（案例）计算得到拟合值和残差

obs	`$X_i$`	`$Y_i$`	`$\hat{Y}_i$`	`$e_i$`
1	6	4.5	4.3	0.127
2	7	5.8	5.1	0.716
3	8	6.0	5.8	0.200
4	9	7.3	6.5	0.829
5	10	7.3	7.2	0.092
6	11	6.6	8.0	-1.366
7	12	7.8	8.7	-0.857
8	13	7.8	9.4	-1.564
9	14	11.0	10.1	0.899
10	15	10.7	10.8	-0.173
11	16	10.8	11.6	-0.735
12	17	13.6	12.3	1.320
13	18	13.5	13.0	0.512
sum	156	112.8	112.8	0.000

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。

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

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

（案例）计算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 $\hat{\sigma}$：

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

参数估计：“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）：

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

- 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”。

参数估计：SRF和SRM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*。

一旦从样本数据得到OLS估计值，便容易画出样本回归线。

注：我们以后还将考虑区间估计量(interval Estimators)

参数估计：SRF和SRM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（ $\bar{Y}$）

$$\begin{aligned}\hat{Y}_i&=\hat{\beta}_1+\hat{\beta}_2\bar{X}\\&=(\bar{Y}-\hat{\beta}_2\bar{X})+\hat{\beta}_2X_i\\&=\bar{Y}-\hat{\beta}_2(X_i-\bar{X})\end{aligned}$$

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

$$\Rightarrow \bar{\hat{Y}}_i=\bar{Y}$$

参数估计：SRF和SRM的特征

- 特征3：残差的均值($\bar{e}_i$)为零：

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

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

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

$$\sum e_i = 0$$

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

参数估计：SRF和SRM的特征

- 特征4：SRM和SRF可以写成离差形式：

$$\left.\begin{aligned} Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \\ \bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \end{aligned}\right\} \Rightarrow$$

$$Y_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) + e_i \Rightarrow$$

$$y_i = \hat{\beta}_2 x_i + e_i \qquad \text{(SRM-dev)}$$

$$\left.\begin{aligned} \hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \\ \bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \end{aligned}\right\} \Rightarrow$$

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) \Rightarrow$$

$$\hat{y}_i = \hat{\beta}_2 x_i \qquad \text{(SRF-dev)}$$

参数估计：SRF和SRM的特征

- 特征5：残差（ e_i）和 Y_i的拟合值（ $\hat{Y}_i$）不相关

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

- 特征6：残差（ e_i）和自变量（ X_i）不相关

参数估计：离差公式

- 离差定义与符号：

$$
\begin{aligned}
x_i &= X_i - \bar{X} \\
y_i &= Y_i - \bar{Y} \\
\hat{y}_i &= \hat{Y}_i - \bar{\hat{Y}}_i = \hat{Y}_i - \bar{Y}
\end{aligned}
$$

- PRM及其离差形式：

$$
\left.
\begin{aligned}
Y_i &= \beta_1 + \beta_2 X_i + u_i \\
\bar{Y} &= \beta_1 + \beta_2 \bar{X} + \bar{u}
\end{aligned}
\right\} \Rightarrow
$$

$$
Y_i - \bar{Y} = \beta_2 x_i + (u_i - \bar{u}) \Rightarrow
$$

$$
y_i = \hat{\beta}_2 x_i + (u_i - \bar{u}) \qquad \text{(PRM-dev)}
$$

参数估计：离差公式

- SRM及其离差形式：

$$\left.\begin{aligned} Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \\ \bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \end{aligned}\right\} \Rightarrow$$

$$Y_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) + e_i \Rightarrow$$

$$y_i = \hat{\beta}_2 x_i + e_i$$

- SRF及其离差形式：

$$\left.\begin{aligned} \hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \\ \bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \end{aligned}\right\} \Rightarrow$$

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) \Rightarrow$$

$$\hat{y}_i = \hat{\beta}_2 x_i$$

- 残差的离差形式：

$$y_i = \hat{\beta}_2 x_i + e_i \qquad \text{(SRM-dev)} \Rightarrow$$

$$e_i = y_i - \hat{\beta}_2 x_i \qquad \text{(residual-dev)}$$

参数估计：思考与讨论

内容小结：

- 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。
- 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！

思考讨论：

- 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？
- 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？

估计精度：引子

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）的点估计量：

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

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

> 问题是：我们如何知道OLS方法点估计量是否可靠？OLS方法的点估计量是否稳定？OLS方法的点估计量是否可信？

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！

- 点估计量的方差（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！

估计精度：斜率系数的方差和样本方差

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

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

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

- 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（$\hat{\beta}_2$）的样本方差（$S^2_{\hat{\beta}_2}$）和样本标准差（$S_{\hat{\beta}_2}$）：

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

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

- 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。

（附录）证明过程I

步骤1 $\hat{\beta}_2$的变形：

$$
\begin{aligned}
\hat{\beta}_2 &= \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum \left[x_i(Y_i - \bar{Y})\right]}{\sum x_i^2} \\
&= \frac{\sum x_i Y_i - \sum x_i \bar{Y}}{\sum x_i^2} \\
&= \frac{\sum x_i Y_i - \bar{Y} \sum x_i}{\sum x_i^2} && \leftarrow \left[\sum x_i = \sum (X_i - \bar{X}) = 0\right] \\
&= \sum \left(\frac{x_i}{\sum x_i^2} \cdot Y_i \right) && \leftarrow \left[k_i \equiv \frac{x_i}{\sum x_i^2}\right] \\
&= \sum k_i Y_i
\end{aligned}
$$

- 其中，$k_i \equiv \frac{x_i}{\sum x_i^2}$。

（附录）证明过程2

步骤2：计算 $\hat{\beta}_2$的总体方差（ $\sigma^2_{\hat{\beta}_2}$ ）：

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

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

估计精度：截距系数的方差和样本方差

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

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

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

- 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$(S^2_{\hat{\beta}_1})$和样本标准差$(S_{\hat{\beta}_1})$：

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

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

- 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。

步骤1 $\hat{\beta}_1$的变形：

$$
\begin{aligned}
\hat{\beta}_1 &= \bar{Y}_i - \hat{\beta}_2 \bar{X}_i && \leftarrow \left[\hat{\beta}_2 = \sum k_i Y_i\right] \\
&= \frac{1}{n}\sum Y_i - \sum \left(k_i Y_i \cdot \bar{X}\right) \\
&= \sum \left((\frac{1}{n} - k_i \bar{X}) \cdot Y_i \right) && \leftarrow \left[w_i \equiv \frac{1}{n} - k_i \bar{X}\right] \\
&= \sum w_i Y_i
\end{aligned}
$$

- 其中：令 $w_i \equiv \frac{1}{n} - k_i \bar{X}$

（附录）证明过程2

步骤2计算 $\hat{\beta}_1$的总体方差（ $\sigma^2_{\hat{\beta}_1}$ ）：

$$
\begin{aligned}
\sigma^2_{\hat{\beta}_1} &\equiv Var(\hat{\beta}_1) = Var(\sum w_i Y_i) \\
&= \sum \left(w_i^2 Var(\beta_1 + \beta_2 X_i + u_i) \right) && \leftarrow \left[w_i \equiv \frac{1}{n} - k_i \bar{X} \right] \\
&= \sum \left(\left(\frac{1}{n} - k_i \bar{X} \right)^2 Var(u_i) \right) \\
&= \sigma^2 \cdot \sum \left(\frac{1}{n^2} - \frac{2\bar{X} k_i}{n} + k_i^2 \bar{X}^2 \right) && \leftarrow \left[\sum k_i = \sum \left(\frac{x_i}{\sum x_i^2} \right) = \frac{\sum x_i}{\sum x_i^2} = 0 \right] \\
&= \sigma^2 \cdot \left(\frac{1}{n} + \bar{X}^2 \sum k_i^2 \right) && \leftarrow \left[k_i \equiv \frac{x_i}{\sum x_i^2} \right] \\
&= \sigma^2 \cdot \left(\frac{1}{n} + \bar{X}^2 \sum \left(\frac{x_i}{\sum x_i^2} \right)^2 \right)
\end{aligned}
$$

（附录）证明过程2（续）

步骤2计算 $\hat{\beta}_1$的总体方差（ $\sigma^2_{\hat{\beta}_1}$ ）（续前）：

$$
\begin{aligned}
&= \sigma^2 \cdot \left(\frac{1}{n} + \bar{X}^2 \frac{\sum x_i^2}{\left(\sum x_i^2\right)^2} \right) \\
&= \sigma^2 \cdot \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2} \right) \\
&= \frac{\sum x_i^2 + n\bar{X}^2}{n \sum x_i^2} \cdot \sigma^2 \qquad \leftarrow \left[\sum x_i^2 + n\bar{X}^2 = \sum (X_i - \bar{X})^2 + n\bar{X}^2 = \sum X_i^2 \right] \\
&= \frac{\sum X_i^2}{n} \cdot \frac{\sigma^2}{\sum x_i^2}
\end{aligned}
$$

估计精度：小结与思考

现在做一个内容小结：

- 为了衡量OLS方法的点估计量是否稳定或是否可信，我们一般采用方差和标准差指标来表达。
- 大家应熟记斜率和截距估计量的总体方差和样本方差最终公式。

请大家思考如下问题：

- 总体方差和样本方差都是确定的数么？
- 二者分别受那些因素的影响？二者又有什么联系？
- 证明过程中，约定的 k_i和 w_i，有什么特征？

$$\begin{cases} \sum k_i = 0 \\ \sum k_i X_i = 1 \end{cases} \qquad \begin{cases} \sum w_i = 1 \\ \sum w_i X_i = 0 \end{cases}$$

（案例）计算回归系数的样本方差

对于“教育程度案例”，利用FF-ff计算表，以及我们已算出的如下计算量：

- 回归误差方差： $\hat{\sigma}^2 = 0.8812$。

则可以进一步计算出，回归系数的样本方差的标准差分别为：

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

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

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

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

区间估计：斜率系数

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

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

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

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

$$\Pr\left[-t_{\alpha/2,(n-2)} \le \mathrm{T} \le t_{\alpha/2,(n-2)}\right] = 1 - \alpha$$

区间估计：斜率系数

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

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

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：

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

区间估计：截距系数

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

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

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

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

$$\Pr\left[-t_{\alpha/2,(n-2)} \le \mathrm{T} \le t_{\alpha/2,(n-2)}\right] = 1 - \alpha$$

区间估计：截距系数

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

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

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：

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

区间估计：随机干扰项的方差

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

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

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

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

因此， σ^2的 $100(1-\alpha)\%$为：

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

我们继续利用样本数据对教育和工资案例进行分析。

> 教育和工资案例的总体回归模型（PRM）如下：

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

> 教育和工资案例的总体回归模型（SRM）如下：

$$\widehat{Wage}_i = \hat{\beta}_1 + \hat{\beta}_2 Edu_i + e_i$$
$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

（案例）相关计算量

我们之前已算出“教育程度案例”中的如下计算量：

- 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：$S^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n} \cdot \frac{\hat{\sigma}^2}{\sum x_i^2} = 0.7650$；$S^2_{\hat{\beta}_2} = \frac{\hat{\sigma}^2}{\sum x_i^2} = 0.0048$；
- 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：

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

（案例）回归系数的区间估计

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：

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

那么，斜率参数 β_2的95%置信区间为：

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

（案例）随机干扰项方差的区间估计

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

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：

$$(n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha}} \leq \sigma^2 \leq (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}$$

$$11 * \frac{0.8812}{21.92} \leq \sigma^2 \leq 11 * \frac{0.8812}{3.8157}$$

$$0.4422 \leq \sigma^2 \leq 2.5403$$

本节结束