计量经济学(Econometrics)

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

2021-09-07

第06章 多元回归I：代数部分

6.1 估计问题

6.2 推断问题

6.3 受约束的最小二乘法

6.4 检验回归模型的结构或稳定性

6.1 多元回归分析：估计问题

三变量模型：符号与假定

三变量的PRM和PRF为：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$
$$E(Y_i | X_{2i}, X_{3i}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$$

多元回归分析是以多个解释变量的固定值为条件的回归分析。

我们所获得的，是各个自变量X值固定时，Y的平均值或Y的平均响应（mean response）。

- β_1：截距项
- u_i：表示所有未包含到模型中来的变量对Y的平均影响。
- β_2, β_3：偏回归系数（partial regression coefficients）
- i：指第i次观测，当数据为时间序列时用t表示；

三变量模型：CLRM假设

CLRM假设1-1：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）

CLRM假设1-2：模型应该是参数线性的。也即模型中参数必须线性，变量可以不是线性

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

三变量模型：CLRM假设

CLRM假设2-1：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。

$$Cov(X_{2i}, u_i) = Cov(X_{3i}, u_i) = 0, \quad i = 1, 2, \ldots, n$$
$$E(X_i, u_i) = 0$$

CLRM假设2-2：X变量间不存在完全共线性

三变量模型：CLRM假设

CLRM假设3-1：假设随机干扰项均值为零。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。

$$E(u|X_{2i}, X_{3i}) = 0$$

CLRM假设3-2：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：

$$\begin{aligned} Var\left(u_i|(X_{2i}, X_{3i})\right) &= E\left[(u_i - E(u_i))^2|(X_{2i}, X_{3i})\right] \\ &= E(u_i^2|X_i) = E(u_i^2) \equiv \sigma^2 \end{aligned}$$

三变量模型：CLRM假设

CLRM假设3-3：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。记为：

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

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

对多元回归方程的解释

三变量的PRM和PRF为:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$
$$E(Y_i | X_{2i}, X_{3i}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$$

多元回归分析是以多个解释变量的固定值为条件的回归分析。

我们所获得的，是各个自变量X值固定时，Y的平均值或Y的平均响应（mean response）。偏回归系数的含义：

- β_2度量着在保持 X_{3i}不变的情况下， X_{2i}每变化1个单位时，Y的均值的变化。换一句话说， 给出 X_{2i}的单位变化对Y均值的“直接”或“净”影响（净在不染有 X_{3i}的影响）。
- β_3则给出了 X_{3i}的单位变化对Y均值 $E(Y)$的“直接”或“净”影响，净在不沾有 X_{2i}的影响。

儿童死亡率案例：数据

研究关注儿童死亡率（CM，千分数）与人均GNP（PGNP，1980年的人均GNP）和妇女识字率（FLR，百分数）的关系，并构建如下PRM:

$$CM = + \beta_1 + \beta_2 PGNP + \beta_3 FLR + u_i$$

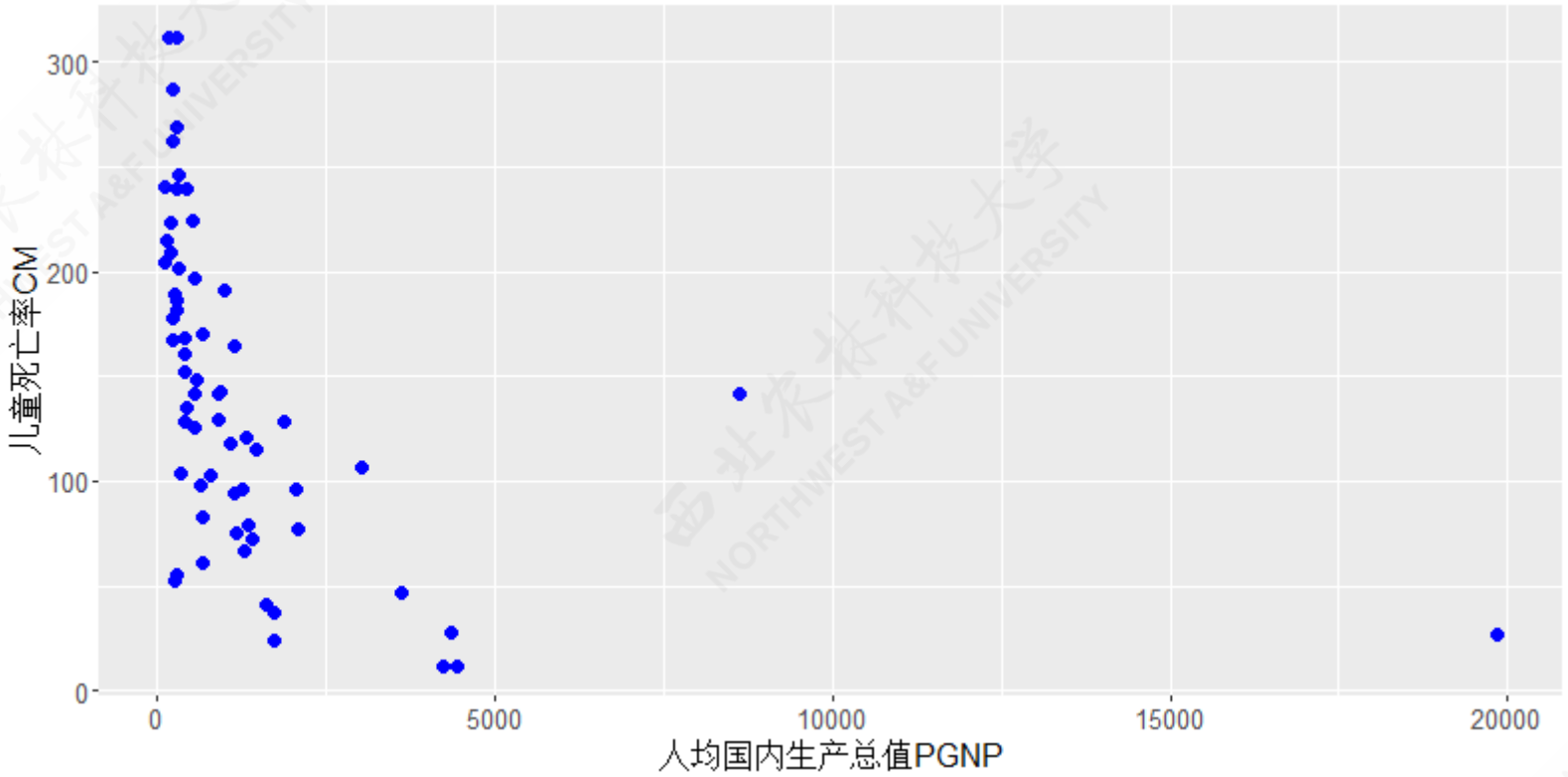
obs	CM	PGNP	FLR
1	128	1870	37
2	204	130	22
3	202	310	16
4	197	570	65
5	96	2050	76
6	209	200	26

Showing 1 to 6 of 64 entries

Previous 1 2 3 4 5 … 11 Next

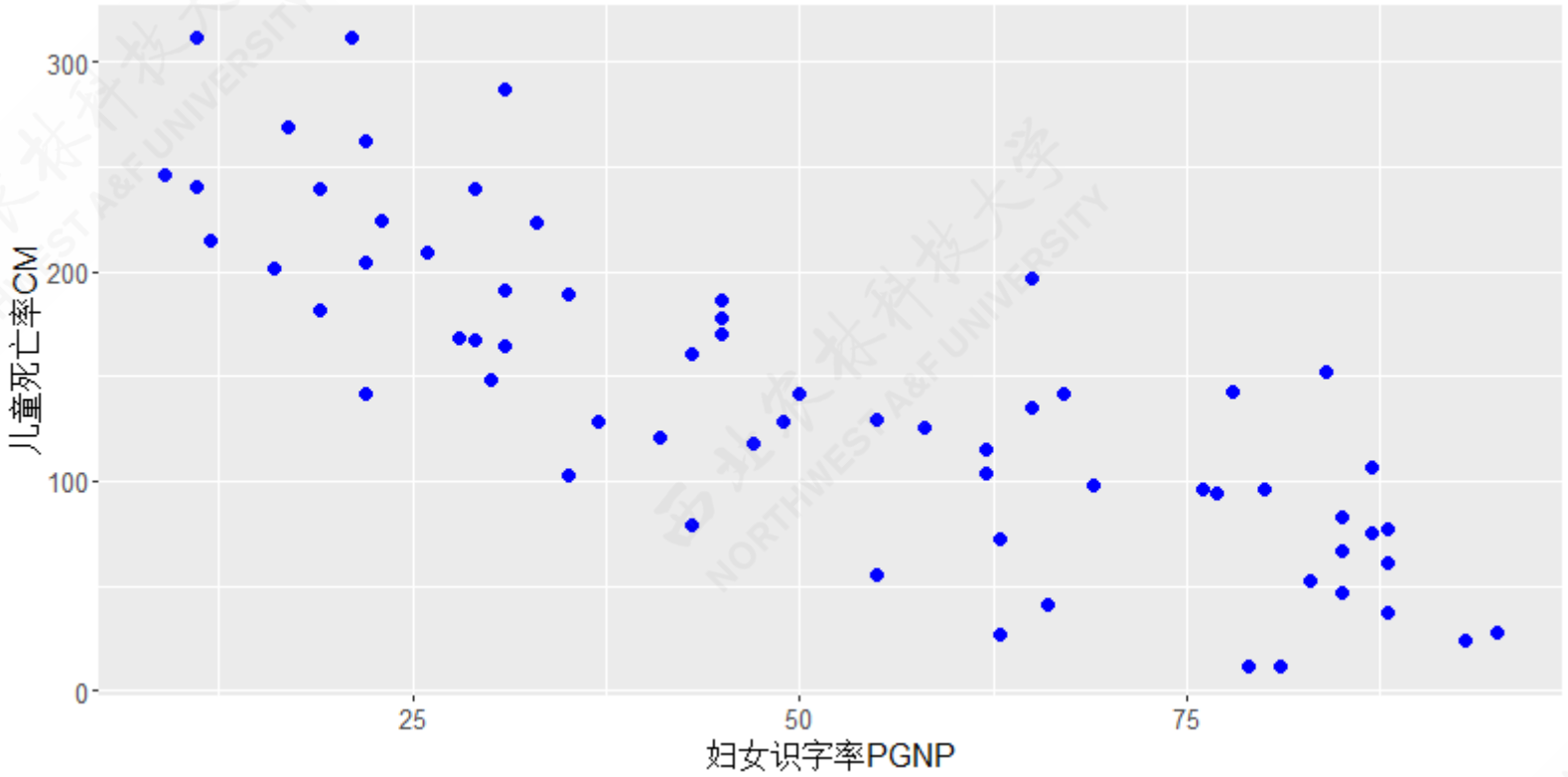
儿童死亡率案例：散点图

绘制散点图1：



儿童死亡率案例：散点图

绘制散点图2：



儿童死亡率案例：二元模型

儿童死亡率的二元回归模型如下：

$$CM = +\beta_1 + \beta_2 PGNP + \beta_3 FLR + u_i$$

以上二元回归模型的OLS估计结果如下：

$$\begin{array}{lllll} \widehat{CM} = & +263.64 & -0.01 PGNP & -2.23 FLR \\ (\text{t}) & (22.7411) & (-2.8187) & (-10.6293) \\ (\text{se}) & (11.5932) & (0.0020) & (0.2099) \\ (\text{fitness}) & R^2 = 0.7077; \bar{R}^2 = 0.6981 & & \\ & F^* = 73.83;\ p = 0.0000 & & \end{array}$$

- 是如何分离出人均国民收入PNGP对CM的“真实”或净影响呢？
- 是如何分离出妇女识字率FLR对CM的“真实”或净影响呢？

儿童死亡率案例：一元回归重现（PGNP纯影响）

步骤1：妇女识字率FLR对儿童死亡率CM的回归**模型1**：

$$CM = + \hat{\beta}_1 + \hat{\beta}_2 FLR + e_i$$

$$\begin{array}{lll} \widehat{CM} = & +263.86 & -2.39FLR \\ (\text{t}) & (21.5840) & (-11.2092) \\ (\text{se}) & (12.2250) & (0.2133) \\ (\text{fitness}) & R^2 = 0.6696; & \bar{R}^2 = 0.6643 \\ & F^* = 125.65; & p = 0.0000 \end{array}$$

步骤2：妇女识字率FLR对人均国民收入PGNP的回归**模型2**：

$$PGNP = + \hat{\beta}_1 + \hat{\beta}_2 FLR + e_i$$

$$\begin{array}{lll} \widehat{PGNP} = & -39.30 & +28.14FLR \\ (\text{t}) & (-0.0535) & (2.1950) \\ (\text{se}) & (734.9526) & (12.8211) \\ (\text{fitness}) & R^2 = 0.0721; & \bar{R}^2 = 0.0571 \\ & F^* = 4.82; & p = 0.0319 \end{array}$$

儿童死亡率案例：一元回归重现（PGNP纯影响）

步骤3：分别得到两次一元线性回归的残差 e_{1i}（`resid1`）和 e_{2i}（`resid2`），然后进行**无截距**回归分析：

obs	CM	FLR	PGNP	TFR	resid1	resid2
1	128	37	1870	6.66	-47.4152	868.0242
2	204	22	130	6.15	-7.2726	-449.8356
3	202	16	310	7	-23.6156	-100.9796
4	197	65	570	6.25	88.5187	-1,219.9707
5	96	76	2050	3.81	13.8142	-49.5402

Showing 1 to 5 of 64 entries

Previous 1 2 3 4 5 … 13 Next

对两个残差序列进一步构造如下的**无截距**回归模型：

$$\widehat{resid1} = + \hat{\beta}_1 resid2$$

儿童死亡率案例：一元回归重现（PGNP纯影响）

残差模型将得到如下回归分析结果：

$$\begin{aligned}
&\widehat{resid1} = -0.01 resid2 \\
&(\text{t}) \qquad (-2.8645) \\
&(\text{se}) \qquad (0.0020) \\
&(\text{fitness}) \ R^2 = 0.1152;\ \bar{R}^2 = 0.1012 \\
&\qquad\qquad F^* = 8.21; \quad p = 0.0057
\end{aligned}$$

对比原来的二元回归模型结果：

$$\begin{aligned}
&\widehat{CM} = \ +263.64 \quad -0.01 PGNP - 2.23 FLR \\
&(\text{t}) \qquad (22.7411) \quad (-2.8187) \quad (-10.6293) \\
&(\text{se}) \qquad (11.5932) \quad (0.0020) \quad (0.2099) \\
&(\text{fitness}) R^2 = 0.7077; \bar{R}^2 = 0.6981 \\
&\qquad\qquad F^* = 73.83;\ p = 0.0000
\end{aligned}$$

儿童死亡率案例：一元回归重现（FLR纯影响）

步骤1：人均国民收入PGNP对儿童死亡率CM的回归**模型3**：

$$CM = +\hat{\beta}_1 + \hat{\beta}_2 PGNP + e_i$$

$$\begin{array}{lll} \widehat{CM} = & +157.42 & -0.01PGNP \\ (\text{t}) & (15.9893) & (-3.5157) \\ (\text{se}) & (9.8456) & (0.0032) \end{array}$$

$$(\text{fitness}) R^2 = 0.1662; \bar{R}^2 = 0.1528$$

$$F^* = 12.36;\ p = 0.0008$$

步骤2：人均国民收入PGNP对妇女识字率FLR的回归**模型4**：

$$FLR = +\hat{\beta}_1 + \hat{\beta}_2 PGNP + e_i$$

$$\begin{array}{lll} \widehat{FLR} = & +47.60 & +0.00PGNP \\ (\text{t}) & (13.3876) & (2.1950) \\ (\text{se}) & (3.5553) & (0.0012) \end{array}$$

$$(\text{fitness}) R^2 = 0.0721; \bar{R}^2 = 0.0571$$

$$F^* = 4.82;\quad p = 0.0319$$

儿童死亡率案例：一元回归重现（GLR纯影响）

步骤3：分别得到两次一元线性回归的残差 e_{3i}（`resid3`）和 e_{4i}（`resid4`），然后进行**无截距**回归分析：

obs	CM	FLR	PGNP	TFR	resid3	resid4
1	128	37	1870	6.66	-8.1729	-15.3886
2	204	22	130	6.15	48.0529	-25.9303
3	202	16	310	7	48.0985	-32.3915
4	197	65	570	6.25	46.0533	15.9424
5	96	76	2050	3.81	-38.1273	23.1502

Showing 1 to 5 of 64 entries

Previous 1 2 3 4 5 … 13 Next

对两个残差序列进一步构造如下的**无截距**回归模型：

$$\widehat{resid3} = + \hat{\beta}_1 resid4$$

儿童死亡率案例：一元回归重现（FLR纯影响）

残差模型将得到如下回归分析结果：

$$\begin{aligned}
&\widehat{resid3} = -2.23 resid4 \\
&(\text{t}) \qquad (-10.8021) \\
&(\text{se}) \qquad (0.2066) \\
&(\text{fitness}) \ R^2 = 0.6494;\ \bar{R}^2 = 0.6438 \\
&\qquad\qquad F^* = 116.69; p = 0.0000
\end{aligned}$$

对比原来的二元回归模型结果：

$$\begin{aligned}
&\widehat{CM} = \ +263.64 \quad -0.01 PGNP - 2.23 FLR \\
&(\text{t}) \qquad (22.7411) \quad (-2.8187) \quad (-10.6293) \\
&(\text{se}) \qquad (11.5932) \quad (0.0020) \quad (0.2099) \\
&(\text{fitness}) R^2 = 0.7077; \bar{R}^2 = 0.6981 \\
&\qquad\qquad F^* = 73.83;\ p = 0.0000
\end{aligned}$$

偏回归系数的OLS估计：模型

样本回归函数SRF、样本回归模型SRM、样本回归函数的离差形式：

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$$
$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + e_i$$
$$\hat{y}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$$

总体回归函数PRF和总体回归模型PRM：

$$\mathrm{E}\left(Y_i | (X_{2i}, X_{3i})\right) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$$
$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

偏回归系数的OLS估计：OLS方法

$$\mathrm{E}\left(Y_i \mid (X_{2i}, X_{3i}\right) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$$
$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_1} = 2 \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}\right)(-1) = 0$$
$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_2} = 2 \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}\right)(-X_{2i}) = 0$$
$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_3} = 2 \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}\right)(-X_{3i}) = 0$$

$$\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}_2 + \hat{\beta}_3 \overline{X}_3$$
$$\sum_i Y_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i}$$
$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2$$

回归系数的OLS估计：回归系数

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X} - \hat{\beta}_3 \overline{X}_3$$

$$\hat{\beta}_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\sum x_{3i}^2\right) - \left(\sum y_i x_{3i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2}$$

$$\hat{\beta}_3 = \frac{\left(\sum y_i x_{3i}\right)\left(\sum x_{2i}^2\right) - \left(\sum y_i x_{2i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2}$$

偏斜率系数 $(\hat{\beta}_2, \hat{\beta}_3)$OLS估计量公式的特点：

- 公式是**对称的**。通过对调 x_{2i}, x_{3i}而得到另一个。
- 两个方程的分母完全相同。
- 三变量情形是双变量情形的自然而然的推广。

回归系数的OLS估计：随机干扰项的方差

对于二元回归模型：

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

在CLRM假设下，$var(u_i) \equiv \sigma^2$，可以证明其OLS估计量（证明略）：

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-3} \qquad (n-3)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-3)$$

$$\begin{aligned}
\sum e_i^2 &= \sum (e_i e_i) = \sum e_i \left(y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}\right) \\
&= \sum e_i y_i \qquad \leftarrow \left[\sum e_i x_{2i} = \sum e_i x_{3i} = 0\right] \\
&= \sum y_i \left(y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}\right) \\
&= \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i}
\end{aligned}$$

回归系数的OLS估计：方差和标准差

$\hat{\beta}_1$的真实方差：

$$\operatorname{var}\left(\hat{\beta}_1\right) \equiv \sigma^2_{\hat{\beta}_1} = \left[\frac{1}{n} + \frac{\overline{X}_2^2 \sum x_{3i}^2 + \overline{X}_3^2 \sum x_{2i}^2 - 2\overline{X}_2\overline{X}_3 \sum x_{2i}x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i}x_{3i}\right)^2}\right] \cdot \sigma^2$$

$\hat{\beta}_1$的样本方差：

$$S^2_{\hat{\beta}_1} = \left[\frac{1}{n} + \frac{\overline{X}_2^2 \sum x_{3i}^2 + \overline{X}_3^2 \sum x_{2i}^2 - 2\overline{X}_2\overline{X}_3 \sum x_{2i}x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i}x_{3i}\right)^2}\right] \cdot \hat{\sigma}^2$$

回归系数的OLS估计：方差和标准差

$\hat{\beta}_2$的真实方差：

$$\begin{aligned}\operatorname{var}\left(\hat{\beta}_2\right) \equiv \sigma^2_{\hat{\beta}_2} &= \frac{\sum x_{3i}^2}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i}x_{3i}\right)^2}\sigma^2 \\ &= \frac{\sigma^2}{\sum x_{2i}^2\left(1 - r_{23}^2\right)} \qquad \leftarrow \left[r_{23}^2 = \frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x_{2i}^2 \sum x_{3i}^2}\right]\end{aligned}$$

$\hat{\beta}_2$的样本方差：

$$\begin{aligned}S^2_{\hat{\beta}_2} &= \frac{\sum x_{3i}^2}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i}x_{3i}\right)^2}\hat{\sigma}^2 \\ &= \frac{\hat{\sigma}^2}{\sum x_{2i}^2\left(1 - r_{23}^2\right)} \qquad \leftarrow \left[r_{23}^2 = \frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x_{2i}^2 \sum x_{3i}^2}\right]\end{aligned}$$

回归系数的OLS估计：方差和标准差

$\hat{\beta}_3$的真实方差：

$$\begin{aligned}\operatorname{var}\left(\hat{\beta}_3\right) \equiv \sigma^2_{\hat{\beta}_3} &= \frac{\sum x_{2i}^2}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i}x_{3i}\right)^2}\sigma^2 \\ &= \frac{\sigma^2}{\sum x_{3i}^2\left(1 - r_{23}^2\right)} \qquad \leftarrow \left[r_{23}^2 = \frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x_{2i}^2 \sum x_{3i}^2}\right]\end{aligned}$$

$\hat{\beta}_3$的样本方差：

$$\begin{aligned}S^2_{\hat{\beta}_3} &= \frac{\sum x_{2i}^2}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i}x_{3i}\right)^2}\hat{\sigma}^2 \\ &= \frac{\hat{\sigma}^2}{\sum x_{3i}^2\left(1 - r_{23}^2\right)} \qquad \leftarrow \left[r_{23}^2 = \frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x_{2i}^2 \sum x_{3i}^2}\right]\end{aligned}$$

回归系数的OLS估计：协方差

随机变量 $\hat{\beta}_2$和 $\hat{\beta}_3$之间的**协方差**为：

$$\operatorname{cov}\left(\hat{\beta}_2, \hat{\beta}_3\right) = \frac{-r_{23}\sigma^2}{\left(1-r_{23}^2\right)\sqrt{\sum x_{2i}^2}\sqrt{\sum x_{3i}^2}} \quad \leftarrow \left[r_{23}^2 = \frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x_{2i}^2 \sum x_{3i}^2}\right]$$

随机变量 $\hat{\beta}_2$和 $\hat{\beta}_3$之间的**样本协方差**为：

$$S^2_{\hat{\beta}_2\hat{\beta}_3} = \frac{-r_{23}\hat{\sigma}^2}{\left(1-r_{23}^2\right)\sqrt{\sum x_{2i}^2}\sqrt{\sum x_{3i}^2}} \quad \leftarrow \left[r_{23}^2 = \frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x_{2i}^2 \sum x_{3i}^2}\right]$$

OLS估计量的特征

特征1： 三变量回归面通过均值点 $(\bar{Y}, \bar{X}_2, \bar{X}_3)$

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3$$

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3 + \cdots + \hat{\beta}_k \bar{X}_k$$

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($\bar{Y}$)

$$\begin{aligned}
\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3 \\
&= (\bar{Y} - \hat{\beta}_2 \bar{X} - \hat{\beta}_3 \bar{X}_3) + \hat{\beta}_2 X_i + \hat{\beta}_3 X_{3i} \\
&= \bar{Y} - \hat{\beta}_2 (X_{2i} - \bar{X}_2) - \hat{\beta}_3 (X_{3i} - \bar{X}_3)
\end{aligned}$$

$$\Rightarrow 1/n \sum \hat{Y}_i = 1/n \sum \left(\bar{Y} - \hat{\beta}_2 (X_{2i} - \bar{X}_2) - \hat{\beta}_3 (X_{3i} - \bar{X}_3) \right)$$

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

OLS估计量的特征

特征3：残差的均值($\bar{e_i}$)为零：

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_1} = 2\sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}\right)(-1) = 0$$

$$\sum \left[Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i - \hat{\beta}_3 X_{3i}\right] = 0$$

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

$$\sum e_i = 0$$

$$\bar{e_i} = 0$$

OLS估计量的特征

特征4： 残差(e_i)和 Y_i的拟合值($\hat{Y}_i$)不相关

$$
\begin{aligned}
Cov(e_i, \hat{Y}_i) &= E\left[(e_i - E(e_i)) \cdot \left(\hat{Y}_i - E(\hat{Y}_i)\right)\right] \\
&= E(e_i \cdot \hat{y}_i) \\
&= \sum \mathrm{e}_i \left(\hat{y}_i + \bar{Y}\right) \\
&= \sum \mathrm{e}_i \hat{y}_i + \bar{Y} \sum e_i \\
&= 0
\end{aligned}
$$

其中：

$$
\begin{aligned}
\sum \hat{y}_i \mathrm{e}_i &= \hat{\beta}_2 \sum x_{2i} e_i + \hat{\beta}_3 \sum x_{3i} e_i \qquad \leftarrow \left[\hat{y}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}\right] \\
&= \hat{\beta}_2 \sum \left(X_{2i} - \overline{X}_2\right) e_i + \hat{\beta}_3 \sum \left(X_{3i} - \overline{X}_3\right) e_i \\
&= \hat{\beta}_2 \sum X_{2i} e_i - \hat{\beta}_2 \overline{X}_2 \sum e_i + \hat{\beta}_3 \sum X_{3i} e_i - \hat{\beta}_3 \overline{X}_3 \sum e_i \\
&= 0
\end{aligned}
$$

OLS估计量的特征

特征5： 残差(e_i)和自变量(X_{2i}, X_{3i})不相关

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_2} = 2\sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}\right)(-X_{2i}) = 0$$

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_3} = 2\sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}\right)(-X_{3i}) = 0$$

$$\sum e_i X_{2i} = 0$$

$$\sum e_i X_{3i} = 0$$

OLS估计量的特征

特征6： $var(\hat{\beta}_2)$和 $var(\hat{\beta}_2)$的关系。

$$\operatorname{var}\left(\hat{\beta}_2\right)=\sigma_{\hat{\beta}_2}^2=\frac{\sigma^2}{\sum x_{2i}^2\left(1-r_{23}^2\right)}$$

$$\operatorname{var}\left(\hat{\beta}_3\right)=\sigma_{\hat{\beta}_3}^2=\frac{\sigma^2}{\sum x_{3i}^2\left(1-r_{23}^2\right)}$$

$$r_{23}^2=\frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x_{2i}^2\sum x_{3i}^2}$$

$$r_{23}\to 1,\operatorname{var}\left(\hat{\beta}_3\right)\to\infty;\operatorname{var}\left(\hat{\beta}_3\right)\to\infty$$

- 给定 $\sum x_{ki}^2,\sigma^2$：真值 β_i的估计将变得很困难。
- 给定 $\sum x_{ki}^2,r_{23}^2$： $var(\hat{\beta}_i)$与总体方差呈正比。
- 给定 σ^2,r_{23}^2： $var(\hat{\beta}_i)$与 $\sum x_{ki}^2$呈反比。表明 x_{ki}样本值变化越大，真值 β_i的估计精度越高！

多元判定系数

多元判定系数：在三变量（或者更多变量）的模型中，衡量Y的变异由变量(X_{2i}, X_{3i})等联合解释的比重，记作 R^2。

$$
\begin{aligned}
Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + e_i \\
\overline{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \overline{X}_2 + \hat{\beta}_3 \overline{X}_3 \\
y_i &= \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + e_i \\
&= \hat{y}_i + e_i
\end{aligned}
$$

$$
\begin{aligned}
(Y_i - \bar{Y}) &= (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \\
y_i &= \hat{y}_i + e_i \\
\sum y_i^2 &= \sum \hat{y}_i^2 + \sum e_i^2 \\
TSS &= ESS + RSS
\end{aligned}
$$

多元判定系数

$$
\begin{aligned}
RSS &= \sum e_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i} \\
ESS &= \sum \hat{y}_i^2 = \hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i} \\
TSS &= \sum y_i^2 = \sum \hat{y}_i^2 + \left(\sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i} \right)
\end{aligned}
$$

$$
R^2 = \frac{ESS}{TSS} = 1 - \frac{\mathrm{RSS}}{TSS} = 1 - \frac{\sum e_i^2}{\sum y_i^2} = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2}
$$

比较一元回归下的判定系数：

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

多元判定系数

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{\text{RSS}}{TSS} = 1 - \frac{\sum e_i^2}{\sum y_i^2} = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2}$$

- 分母部分与 X_{ki} 的变量数**无关**：

$$\sum y_i^2 = \sum \left(Y_i - \overline{Y}\right)^2$$

- 分子部分与 X_{ki} 的变量数**有关**：

$$\sum e_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i}$$

- 如果 X_{ki} 的变量数**增加**，RSS会**减小**，而TSS总是**不变**，因此判定系数 R^2 自然会**变大**。一般而言，自变量数越多，判定系数约接近于1。
- 启示：模型选择时，较高的 R^2 可能来自解释变量个数的增加，并不能说明模型就一定更好。

调整多元判定系数

调整判定系数（adjusted R square）：利用相应的自由度对平方和进行校正，基于此计算得到的判定系数，记为 $\bar{R}^2$。对于如下的多元回归方程：

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

则**调整判定系数** $\bar{R}^2$ 可以计算为：

$$\overline{R}^2 = 1 - \frac{\sum \mathrm{e}_i^2/(n-k)}{\sum y_i^2/(n-1)}$$

调整多元判定系数

调整判定系数 $\bar{R}^2$ 还可以计算为：

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{S_Y^2} \qquad \leftarrow \left[\hat{\sigma}^2 = \frac{\sum \mathrm{e}_i^2}{n-k}\right]$$

$$\bar{R}^2 = 1 - \left(1 - R^2\right)\frac{n-1}{n-k} \qquad \leftarrow \left[S_Y^2 = \frac{\sum (Y - \bar{Y})^2}{n-1}\right]$$

- 如果 $k = 1$，则 $R^2 = \bar{R}^2$。
- 如果 $k > 1$，则 $R^2 > \bar{R}^2$。表明随着变量数的增多，$\bar{R}^2$ 相对要增大得慢一些。
- $R^2 \geq 0$，但 $\bar{R}^2$ 可以小于0
- 不能单凭最高的 $\bar{R}^2$ 之值来选择模型，可参考的标准还可以有AIC、APC等。

模型选择的标准：儿童死亡率案例

人均国民收入PGNP和妇女识字率FLR对儿童死亡率CM的二元回归**模型1**：可以在不同回归元之间进行分配吗？

$$
\begin{array}{lllll}
\widehat{CM} = & +263.64 & -0.01PGNP & -2.23FLR \\
(\mathrm{t}) & (22.7411) & (-2.8187) & (-10.6293) \\
(\mathrm{se}) & (11.5932) & (0.0020) & (0.2099) \\
(\mathrm{fitness}) & R^2 = 0.7077; \bar{R}^2 = 0.6981 & & \\
& F^* = 73.83;\ p = 0.0000 & &
\end{array}
$$

妇女识字率FLR对死亡率CM**回归2**：

$$
\begin{array}{lll}
\widehat{CM} = & +263.86 & -2.39FLR \\
(\mathrm{t}) & (21.5840) & (-11.2092) \\
(\mathrm{se}) & (12.2250) & (0.2133) \\
(\mathrm{fitness}) & R^2 = 0.6696; \bar{R}^2 = 0.6643 & \\
& F^* = 125.65; p = 0.0000 &
\end{array}
$$

人均国民收入PGNP对死亡率CM**回归3**：

模型选择的标准：柯布道格拉斯生产曲线案例

以柯布道格拉斯生产模型为例：

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{u_i}$$

$$\begin{aligned} \ln Y_i &= \ln \beta_1 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i \\ &= \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i \qquad \leftarrow [\beta_0 = \ln \beta_1] \end{aligned}$$

模型选择的标准：柯布道格拉斯生产曲线（数据）

state	Y	X2	X3
Alabama	38372840	424471	2689076
Alaska	1805427	19895	57997
Arizona	23736129	206893	2308272
Arkansas	26981983	304055	1376235
California	217546032	1809756	13554116
Colorado	19462751	180366	1790751
Connecticut	28972772	224267	1210229
Delaware	14313157	54455	421064

Showing 1 to 8 of 51 entries

Previous 1 2 3 4 5 6 7 Next

美国51个地区的制造业投入产出数据

模型选择的标准：柯布道格拉斯生产案例（回归）

双对数模型下：

$$log(Y) = +\beta_1 + \beta_2 log(X2) + \beta_3 log(X3) + u_i$$

OLS估计结果为：

$$\begin{array}{lllll}
\widehat{log(Y)} = & +3.89 & +0.47log(X2) & +0.52log(X3) \\
(\text{t}) & (9.8115) & (4.7342) & (5.3803) \\
(\text{se}) & (0.3962) & (0.0989) & (0.0969) \\
(\text{fitness}) & R^2 = 0.9642; \bar{R}^2 = 0.9627 & & \\
& F^* = 645.93; p = 0.0000 & &
\end{array}$$

模型选择的标准：多项式回归

多项式回归模型(polynomial regression models)：

$$
\begin{aligned}
Y &= \beta_1 + \beta_2 X + \beta_3 X^2 \\
Y_i &= \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i \\
Y_i &= \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \cdots + \beta_k X_i^k + u_i \\
\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}
\end{aligned}
$$

- 思考1：上述模型是线性回归模型吗？
- 思考2：X与X的诸多幂函数之间是高度相关的吗？有没有违背自变量无多重共线性的CLRM假设？

模型选择的标准：总生产成本案例（数据）

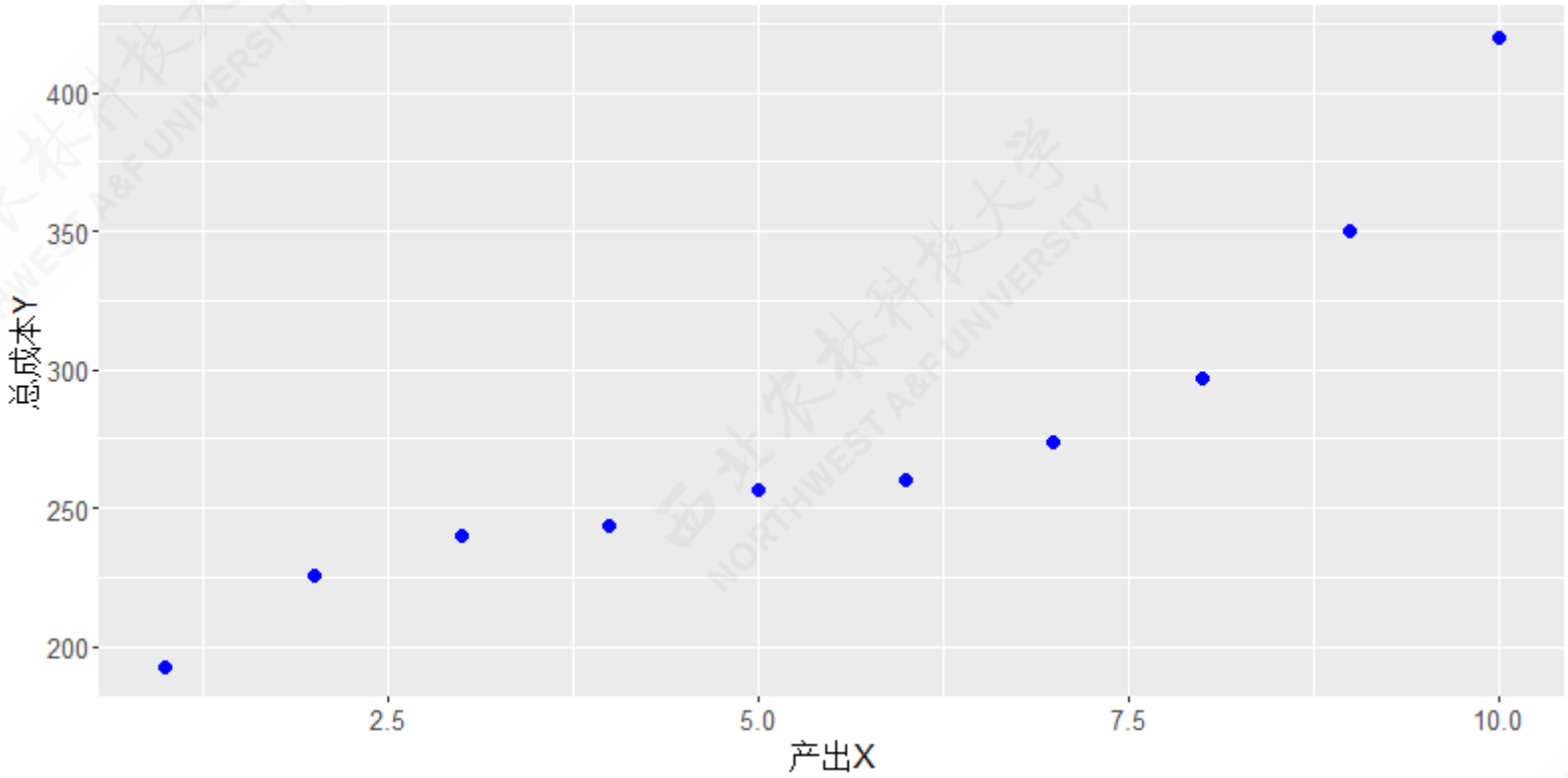
X	Y	XX	XXX
1	193	1	1
2	226	4	8
3	240	9	27
4	244	16	64
5	257	25	125
6	260	36	216
7	274	49	343
8	297	64	512

Showing 1 to 8 of 10 entries

Previous 1 2 Next

模型选择的标准：总生产成本案例（绘图）

产出X与总成本Y的散点图关系如下：



模型选择的标准：总生产成本案例（回归）

多项式模型下：

$$Y = +\beta_1 + \beta_2 X + \beta_3 XX + \beta_4 XXX + u_i$$

OLS估计结果为：

$$\begin{array}{lllll}
\widehat{Y} = & +141.77 & +63.48X & -12.96XX & +0.94XXX \\
(\mathrm{t}) & (22.2368) & (13.2837) & (-13.1501) & (15.8968) \\
(\mathrm{se}) & (6.3753) & (4.7786) & (0.9857) & (0.0591) \\
(\mathrm{fitness}) & R^2 = 0.9983; & \bar{R}^2 = 0.9975 & & \\
 & F^* = 1202.22; & p = 0.0000 & &
\end{array}$$

偏相关系数

偏相关系数（partial correlation coefficient）：一个不依赖于X_{2i}的，对X_{3i}和Y_i的影响的一种相关系数。

- 保持X_{3i}不变，Y_i和X_{2i}之间的相关系数：

$$r_{12\cdot 3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{\left(1-r_{13}^2\right)\left(1-r_{23}^2\right)}}$$

- 保持X_{2i}不变，Y_i和X_{3i}之间的相关系数：

$$r_{13.2}=\frac{r_{13}-r_{12}r_{23}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{23}^2\right)}}$$

- 保持Y_i不变，X_{2i}和X_{3i}之间的相关系数：

$$r_{23.1}=\frac{r_{23}-r_{12}r_{13}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{13}^2\right)}}$$

简单相关系数

简单相关系数（simple correlation coefficient）：

Y_i和 X_{2i}之间的相关系数：

$$r_{12} = \frac{\sum y_i x_{2i}}{\sqrt{\sum y_i^2}\sqrt{\sum x_{2i}^2}}$$

Y_i和 X_{3i}之间的相关系数：

$$r_{13} = \frac{\sum y_i x_{3i}}{\sqrt{\sum y_i^2}\sqrt{\sum x_{3i}^2}}$$

X_{2i}和 X_{3i}之间的相关系数：

$$r_{23} = \frac{\sum x_{2i} x_{3i}}{\sqrt{\sum x_{2i}^2}\sqrt{\sum x_{3i}^2}}$$

6.2 多元回归分析：推断问题

N-CLRM假设

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： $Var(u_i)\equiv\sigma^2$
- 无自相关： $E(u_i,u_j)=0$
- 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)。

个别回归系数的显著性检验（t检验理论）

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

二元线性回归模型，在N-CLRM假设容易得到：

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t(n-3)$$

$$T_2 = \frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}} \sim t(n-3)$$

$$T_3 = \frac{\hat{\beta}_3 - \beta_3}{S_{\hat{\beta}_3}} \sim t(n-3)$$

个别回归系数的显著性检验（t检验理论）

其中：

$$S_{\hat{\beta}_1}^2 = \left[\frac{1}{n} + \frac{\overline{X}_2^2 \sum x_{3i}^2 + \overline{X}_3^2 \sum x_{2i}^2 - 2\overline{X}_2\overline{X}_3 \sum x_{2i}x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i}x_{3i}\right)^2}\right] \cdot \hat{\sigma}^2$$

$$S_{\hat{\beta}_2}^2 = \frac{\hat{\sigma}^2}{\sum x_{2i}^2 \left(1 - r_{23}^2\right)} \quad \leftarrow \left[r_{23}^2 = \frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x_{2i}^2 \sum x_{3i}^2}\right]$$

$$S_{\hat{\beta}_3}^2 = \frac{\hat{\sigma}^2}{\sum x_{3i}^2 \left(1 - r_{23}^2\right)} \quad \leftarrow \left[r_{23}^2 = \frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x_{2i}^2 \sum x_{3i}^2}\right]$$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-3} = \frac{1}{n-3}\left(\sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i}\right)$$

个别回归系数的显著性检验（t检验理论）

假设：

$$H_0: \beta_i = 0; \quad H_1: \beta_i \neq 0, \quad i \in (1, 2, 3)$$

基于 H_0 可以得到：

$$t^*_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

$$t^*_{\hat{\beta}_2} = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}}$$

$$t^*_{\hat{\beta}_3} = \frac{\hat{\beta}_3}{S_{\hat{\beta}_3}}$$

个别回归系数的显著性检验（t检验理论）

给定显著性水平 $\alpha = 0.05$下，查出统计量的**理论分布值**。$t_{1-\alpha/2}(n-3)$。

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_i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，认为回归参数 $\beta_i \neq 0$。
- 若 $|t^*_{\hat{\beta}_i}| < t_{1-\alpha/2}(n-2)$，则 β_i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，认为回归参数 $\beta_i = 0$。

个别回归系数的显著性检验（t检验案例）

儿童死亡率的二元回归模型如下：

$$CM = +\beta_1 + \beta_2 PGNP + \beta_3 FLR + u_i$$

以上二元回归模型的OLS估计结果如下：

$$\begin{array}{lllll}
\widehat{CM} = & +263.64 & -0.01PGNP & -2.23FLR & \\
(\text{t}) & (22.7411) & (-2.8187) & (-10.6293) & \\
(\text{se}) & (11.5932) & (0.0020) & (0.2099) & \\
(\text{fitness}) & R^2 = 0.7077; \bar{R}^2 = 0.6981 & & & \\
& F^* = 73.83;\ p = 0.0000 & & &
\end{array}$$

样本回归模型的整体显著性检验（F检验）

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

$$\begin{aligned}
(Y_i - \bar{Y}) &= (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \\
y_i &= \hat{y}_i + e_i \\
\sum y_i^2 &= \sum \hat{y}_i^2 + \sum e_i^2 \\
TSS &= ESS + RSS
\end{aligned}$$

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	$\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}$	2	MSS_{ESS}	$ESS/df_{ESS} = \sum \hat{y}_i^2/2$
残差平方和	RSS	$\sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2$	n-3	MSS_{RSS}	$RSS/df_{RSS} = \frac{\sum e_i^2}{n-3}$
总平方和	TSS	$\sum (Y_i - \bar{Y}_i)^2 = \sum y_i^2$	n-1	MSS_{TSS}	$TSS/df_{TSS} = \frac{\sum y_i^2}{n-1}$

样本回归模型的整体显著性检验（F检验）

在原假设下：

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

有：

$$F^* = \frac{MSS_{ESS}}{MSS_{RSS}} = \frac{ESS/df_{\mathrm{Ess}}}{RSS/df_{\mathrm{RSS}}} = \frac{\left(\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum \mathrm{y}_i x_{3i}\right)/2}{\sum \mathrm{e}_i^2/(n-3)} \sim \mathrm{F}(2, \mathrm{n}-3)$$

k变量回归模型的F检验

k变量回归模型下：

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

给出如下假设：

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

F样本统计量可以表达为：

$$F^* = \frac{ESS/df_{ESS}}{RSS/df_{RSS}} = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{MSS_{ESS}}{MSS_{\text{RSS}}} \sim \text{F}(k-1, n-k)$$

k变量回归模型的F检验

k变量回归模型下，F样本统计量也可以用判定系数 R^2 表达为：

$$
\begin{aligned}
F &= \frac{(n-k)ESS}{(k-1)RSS} \\
&= \frac{n-k}{k-1} \cdot \frac{ESS}{TSS-ESS} \\
&= \frac{n-k}{k-1} \cdot \frac{ESS/TSS}{1-(ESS/TSS)} \\
&= \frac{n-k}{k-1} \cdot \frac{R^2}{1-R^2} \\
&= \frac{R^2/(k-1)}{(1-R^2)/(n-k)}
\end{aligned}
$$

k变量回归模型的F检验

k变量回归模型下，方差分析表（ANOVA）理论上可以写成：

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	$R^2\sum y_i^2$	k-1	MSS_{ESS}	$R^2\sum y_i^2/(k-1)$
残差平方和	RSS	$(1-R^2)\sum y_i^2$	n-3	MSS_{RSS}	$\frac{(1-R^2)\sum y_i^2}{n-3}$
总平方和	TSS	$\sum y_i^2$	n-1	MSS_{TSS}	$TSS/df_{TSS}=\frac{\sum y_i^2}{n-1}$

k变量回归模型的F检验（案例）

研究关注儿童死亡率（CM，千分数）与人均GNP（PGNP，1980年的人均GNP）和妇女识字率（FLR，百分数）的关系，并构建如下PRM:

$$CM = + \beta_1 + \beta_2 PGNP + \beta_3 FLR + u_i$$

可以计算得到如下方差分析表（AVOVA）：

儿童死亡率案例的ANOVA分析表

source	SS	df	MSS
回归平方和ESS	257,362.37	2	128,681.19
残差平方和RSS	106,315.63	61	1,742.88
总平方和TSS	363,678.00	63	5,772.67

k变量回归模型的F检验（案例）

因此可以计算得到样本F统计量值（F^*）为：

$$F^* = \frac{ESS/df_{ESS}}{RSS/df_{RSS}} = \frac{MSS_{ESS}}{MSS_{RSS}} = \frac{128681.1865}{1742.8791} = 73.8325$$

给定显著性水平 $\alpha = 0.05$下，查出F分布的**理论值** $F_{1-\alpha}(1, n-2) = F_{0.95}$(2,61)=3.1478

得到显著性检验的判断结论。因为 $F^* = 73.8325$ 大于 $F_{1-\alpha}(1, n-2) = F_{0.95}$(2,61)=3.1478，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 $\beta_2 = \beta_3 \neq 0$。

6.3 受约束的最小二乘法

线性等式约束条件

线性等式约束条件：根据已有的经济理论，某一回归模型中的系数需满足一些线性等式约束条件。

柯布道格拉斯生产函数(the Cobb-Douglas production function):

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{u_i}$$

其中，Y_i表示产出；X_{2i}表示劳动投入；X_{3i}表示资本投入。

可以将指数模型转换成如下线性模型：

$$\begin{aligned} \ln Y_i &= \ln \beta_1 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i \\ &= \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i \qquad \leftarrow [\beta_0 = \ln \beta_1] \end{aligned}$$

假设所描述的生产是规模报酬不变，由经济理论可得如下的**线性等式约束条件**：

$$\beta_2 + \beta_3 = 1$$

线性等式约束的t检验法

步骤1：先做无约束的或**无限制的回归**（unrestricted or unconstrained regression）

$$log(Y) = + \hat{\beta}_1 + \hat{\beta}_2 log(X2) + \hat{\beta}_3 log(X3) + e_i$$

步骤2：构建T统计量：

$$T = \frac{\left(\hat{\beta}_2 + \hat{\beta}_3\right) - (\beta_2 + \beta_3)}{S_{\left(\hat{\beta}_2 + \hat{\beta}_3\right)}} \sim \mathrm{t}(n-3)$$

线性等式约束的t检验法

其中：

$$S^2_{\left(\hat{\beta}_2+\hat{\beta}_3\right)}=S^2_{\hat{\beta}_2}+S^2_{\hat{\beta}_3}+2S^2_{\hat{\beta}_2\hat{\beta}_3}$$

$$S^2_{\hat{\beta}_2}=\frac{\hat{\sigma}^2}{\sum x^2_{2i}\left(1-r^2_{23}\right)}\qquad\leftarrow\left[r^2_{23}=\frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x^2_{2i}\sum x^2_{3i}}\right]$$

$$S^2_{\hat{\beta}_3}=\frac{\hat{\sigma}^2}{\sum x^2_{3i}\left(1-r^2_{23}\right)}\qquad\leftarrow\left[r^2_{23}=\frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x^2_{2i}\sum x^2_{3i}}\right]$$

$$S^2_{\hat{\beta}_2\hat{\beta}_3}=\frac{-r_{23}\hat{\sigma}^2}{\left(1-r^2_{23}\right)\sqrt{\sum x^2_{2i}}\sqrt{\sum x^2_{3i}}}\qquad\leftarrow\left[r^2_{23}=\frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x^2_{2i}\sum x^2_{3i}}\right]$$

线性等式约束的t检验法

提出理论假设：

$$H_0: \beta_2 + \beta_3 = 1; H_1: \beta_2 + \beta_3 \neq 1$$

在原假设 H_0下，可以计算得到如下样本t统计量：

$$t^* = \frac{\left(\hat{\beta}_2 + \hat{\beta}_3\right) - 1}{S_{\left(\hat{\beta}_2 + \hat{\beta}_3\right)}}$$

给定显著性水平 $\alpha = 0.05$下，查出统计量的**理论分布值**。$t_{1-\alpha/2}(n-3)$。

得到显著性检验的判断结论。

- 若 $|t^*| > t_{1-\alpha/2}(n-2)$，则 β_i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 $\beta_2 + \beta_3 \neq 1$，也即**规模报酬可变**。

线性等式约束的F检验

无约束模型（Unrestricted model）：

$$\ln Y_i = \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i \qquad \leftarrow [\beta_0 = \ln \beta_1]$$

在线性等式约束条件下：

$$\beta_2 = 1 - \beta_3$$

可以将原模型变换为如下的**受约束模型**（Restricted model）：

$$\begin{aligned}
\ln Y_i &= \beta_0 + (1 - \beta_3) \ln X_{2i} + \beta_3 \ln X_{3i} + u_i \\
\ln Y_i &= \beta_0 + \ln X_{2i} + \beta_3 (\ln X_{3i} - \ln X_{2i}) + u_i \\
(\ln Y_i - \ln X_{2i}) &= \beta_0 + \beta_3 (\ln X_{3i} - \ln X_{2i}) + u_i \\
\ln(Y_i / X_{2i}) &= \beta_0 + \beta_3 \ln(X_{3i} / X_{2i}) + u_i
\end{aligned}$$

线性等式约束的F检验

给出原假设 $H_0:\beta_2+\beta_3=1$下，可以得到如下样本F统计量：

$$F^*=\frac{\left(RSS_R-RSS_{UR}\right)/m}{RSS_{UR}/(n-k)}=\frac{\left(\sum e_R^2-\sum e_{\mathrm{UR}}^2\right)/m}{\sum e_{\mathrm{UR}}^2/(n-k)}\sim \mathrm{F(m,n-k)}$$

$$F^*=\frac{\left(R_{UR}^2-R_R^2\right)/m}{\left(1-R_{UR}^2\right)/(n-k)}$$

其中：

- $RSS_U R$表示**无约束回归模型**的RSS；
- RSS_R表示**受约束回归模型**的RSS；
- R_{UR}^2表示**无约束回归模型**的判定系数；
- R_R^2表示**受约束回归模型**的判定系数；
- m表示线性约束条件的个数； n表示样本数； k表示无约束回归模型中回归系数个数（包括截距）；

线性等式约束的F检验

给定显著性水平 $\alpha = 0.05$下，查出统计量的**理论分布值**。 $F_{1-\alpha}(m, n-k)$

得到显著性检验的判断结论。

- 若 $F^* > F_{1-\alpha}(m, n-k)$，则**显著**地拒绝原假设 H_0，接受备择假设 H_1，认为**规模报酬可变**。
- 若 $F^* < F_{1-\alpha}(m, n-k)$，则不能**显著**地拒绝原假设 H_0，暂时接受原假设 H_0，认为**规模报酬不变**。

线性等式约束的F检验（案例：数据）

state	Y	X2	X3	Y.X2	X3.X2
Alabama	38372840	424471	2689076	90.4016	6.3351
Alaska	1805427	19895	57997	90.7478	2.9152
Arizona	23736129	206893	2308272	114.7266	11.1568
Arkansas	26981983	304055	1376235	88.7405	4.5263
California	217546032	1809756	13554116	120.2074	7.4895
Colorado	19462751	180366	1790751	107.9070	9.9284
Connecticut	28972772	224267	1210229	129.1887	5.3964

Showing 1 to 7 of 51 entries

Previous 1 2 3 4 5 … 8 Next

- $Y.X_2$表示 Y/X_2； $X3.X_2$表示 X_3/X_2

线性等式约束的F检验（案例：回归I）

无约束模型回归结果如下：

$$log(Y) = + \beta_1 + \beta_2 log(X2) + \beta_3 log(X3) + u_i$$

$$\begin{array}{lllll}
\widehat{log(Y)} = & +3.89 & +0.47 log(X2) & +0.52 log(X3) \\
(\mathrm{t}) & (9.8115) & (4.7342) & (5.3803) \\
(\mathrm{se}) & (0.3962) & (0.0989) & (0.0969) \\
(\mathrm{fitness}) & R^2 = 0.9642; \bar{R}^2 = 0.9627 & & \\
& F^* = 645.93; p = 0.0000 & &
\end{array}$$

其中：

- $RSS_{UR} = 3.4155$
- $R^2_{UR} = 0.9642$

线性等式约束的F检验（案例：回归2）

受约束模型回归结果如下：

$$log(Y.X2) = +\beta_1 + \beta_2 log(X3.X2) + u_i$$

$$\begin{array}{lll} \widehat{log(Y.X2)} = +3.76 & & +0.52log(X3.X2) \\ (\text{t}) & (20.2637) & (5.4665) \\ (\text{se}) & (0.1854) & (0.0958) \\ (\text{fitness}) & R^2 = 0.3788; \bar{R}^2 = 0.3661 & \\ & F^* = 29.88;\ p = 0.0000 & \end{array}$$

其中：

- $RSS_R = 3.4256$
- $R_R^2 = 0.3788$

线性等式约束的F检验（案例：F统计量）

利用RSS计算样本F统计量：

$$
\begin{aligned}
F^* &= \frac{\left(RSS_R - RSS_{UR}\right)/m}{RSS_{UR}/(n-k)} \\
&= \frac{\left(3.4256 - 3.4155\right)/1}{3.4155/(51-3)} \\
&= 0.1414
\end{aligned}
$$

利用 R^2计算样本F统计量：

$$
\begin{aligned}
F^* &= \frac{\left(R^2_{UR} - R^2_R\right)/m}{\left(1 - R^2_{UR}\right)/(n-k)} \\
&= \frac{\left(0.9642 - 0.3788\right)/1}{(1 - 0.9642)/(51-3)} \\
&= 784.2919
\end{aligned}
$$

给定显著性水平 $\alpha = 0.05$下，查出F分布的**理论值** $F_{1-\alpha}(m, n-k) = F_{0.95}$ (1,48)=4.0427

因为 $F^* = 0.1414$ 小于 $F_{1-\alpha}(m, n-k) = F_{0.95}$ (1,48)=4.0427，所以，认为 $\beta_2 + \beta_3 = 1$，也即**规模报酬不变**。

6.4 检验回归模型的结构或稳定性

邹至庄检验

Year	Saving	Income	sample
1970	61	727.1	spl1
1971	68.6	790.2	spl1
1972	63.6	855.3	spl1
1973	89.6	965	spl1
1974	97.6	1054.2	spl1
1975	104.4	1159.2	spl1
1976	96.4	1273	spl1
1977	92.5	1401.4	spl1

Showing 1 to 8 of 26 entries

Previous 1 2 3 4 Next

1970-1995年间美国储蓄、可支配收入数据（n=26）

邹至庄检验

按时间分段绘制散点图如下：

其中：spl1表示1970-1981年；spl2表示1982-1995年。

邹至庄检验

根据散点图的情况，我们可以构建如下三个模型：

$$
\begin{aligned}
1970-1981: & Y_t = \lambda_1 + \lambda_2 X_t + u_{1t} & n_1 = 12 \\
1982-1995: & Y_t = \gamma_1 + \gamma_2 X_t + u_{2t} & n_2 = 14 \\
1970-1995: & Y_t = \alpha_1 + \alpha_2 X_t + u_t & n = (n_1 + n_2) = 26
\end{aligned}
$$

邹至庄检验

样本段spl1（1970-1981年）**回归1：**

$$
\begin{array}{lll}
\widehat{Saving} = +1.02 & +0.08Income \\
(\mathrm{t}) \quad (0.0873) & (9.6016) \\
(\mathrm{se}) \quad (11.6377) & (0.0084) \\
(\mathrm{fitness}) \quad R^2 = 0.9021; \bar{R}^2 = 0.8924 \\
F^* = 92.19; \ p = 0.0000
\end{array}
$$

样本段spl2（1982-1995年）**回归2：**

$$
\begin{array}{lll}
\widehat{Saving} = +153.49 & +0.01Income \\
(\mathrm{t}) \quad (4.6923) & (1.7708) \\
(\mathrm{se}) \quad (32.7123) & (0.0084) \\
(\mathrm{fitness}) \quad R^2 = 0.2072; \bar{R}^2 = 0.1411 \\
F^* = 3.14; \quad p = 0.1020
\end{array}
$$

全部样本（1970-1995年）**回归3：**

$$
\begin{array}{lll}
\widehat{Saving} = +62.42 & +0.04Income \\
(\mathrm{t}) \quad (4.8918) & (8.8938) \\
(\mathrm{se}) \quad (12.7607) & (0.0042) \\
(\mathrm{fitness}) \quad R^2 = 0.7672; \bar{R}^2 = 0.7575 \\
F^* = 79.10; \ p = 0.0000
\end{array}
$$

邹至庄检验

邹至庄检验的原理和过程如下：

步骤1：估计在全部样本下的**约束方程**（方程3），得到**约束残差平方和**（记为RSS_R，此处也记为RSS_3）。在全部样本下（1970-1995）的模型，如果参数是稳定的，也即可以认为约束了如下两个条件：

$$\gamma_1 = \lambda_1; \quad \gamma_2 = \lambda_2$$

步骤2：估计分段样本（spl1=1970-1981）下的子方程1，得到其残差平方和RSS_1，其自由度为$df_{RSS_1} = n_1 - k$

步骤3：估计分段样本（spl2=1982-1995）下的子方程2，得到其残差平方和RSS_2，其自由度为$df_{RSS_2} = n_2 - k$

邹至庄检验

步骤4：计算得到**无约束残差平方和**（$RSS_{UR} = RSS_1 + RSS_2$），其自由度为 $df_{RSS_{UR}} = n_1 + n_2 - 2k$

步骤5：如果没有结构性变动（H_0），则构造得到如下样本F统计量：

$$F^* = \frac{\left(\mathrm{RSS_R} - \mathrm{RSS_{UR}}\right)/k}{\left(\mathrm{RSS_{UR}}\right)/\left(n_1 + n_2 - 2k\right)} \sim F_{[k,(n_1+n_2-2k)]}$$

步骤6：得到显著性检验的判断结论。

- 若 $F^* > F_{1-\alpha}(k, n_1 + n_2 - 2k)$，则**显著**地拒绝原假设 H_0，接受备择假设 H_1，认为**存在结构变动**（也即参数不稳定）。
- 若 $F^* < F_{1-\alpha}(k, n_1 + n_2 - 2k)$，则不能**显著**地拒绝原假设 H_0，暂时接受原假设 H_0，认为**不存在结构变动**（也即参数稳定）。

邹至庄检验

在本案例中：

- 约束残差平方和（记为 $RSS_R = 23248.2982$）
- 无约束残差平方和（ $RSS_{UR} = RSS_1 + RSS_2 = 1785.0321+10005=11790.2528$），其自由度为 $df_{RSS_{UR}} = n_1 + n_2 - 2k = 12+14-2*2=22$

$$
\begin{aligned}
F^* &= \frac{\left(RSS_R - RSS_{UR}\right)/k}{RSS_{UR}/(n_1+n_2-2k)} \\
&= \frac{\left(23248.2982 - 11790.2528\right)/2}{11790.2528/(12+14-2*2)} \\
&= 10.6901
\end{aligned}
$$

- 因为 $F^* = 10.6901$ 大于查表值 $F_{1-\alpha}(k, n_1+n_2-2k) = F_{0.95}(2,12+14-2*2)=3.4434$，所以显著拒绝 H_0，接受备择假设 H_1，认为存在结构性变动，也即参数不稳定。

邹至庄检验

牢记关于邹至庄检验的一些警告：

- 必须满足该检验背后的假定。
- 邹至庄检验只是告诉我们子样本模型之间是否有差别，并没有告诉我们差别是来自截距、斜率还是二者都有。
- 邹至庄检验假定我们知道结构转折点。

本章结束