§2.5单方程线性模型的区间估计
Interval Estimation of Multiple Linear Regression Model

一、参数估计量的区间估计

二、预测值的区间估计

一、参数估计量的区间估计

1、问题的提出

- 人们经常说，“通过建立生产函数模型，得到资本的产出弹性是0.5”，“通过建立消费函数模型，得到收入的边际消费倾向是0.6”，等等。其中0.5、0.6是模型具有特定经济含义的参数估计值。

这样的说法正确吗？

应该如何表述才是正确的？

- 线性回归模型的参数估计量是随机变量，利用一次抽样的样本观测值，估计得到的只是参数的一个点估计值。

  如果用参数估计量的一个点估计值近似代表参数值，那么，二者的接近程度如何？以多大的概率达到该接近程度？

- 这就要构造参数的一个区间，以点估计值为中心的一个区间（称为置信区间，**confidence interval**），该区间以一定的概率（称为置信水平，**confidence coefficient** ）包含该参数。

$$P(\hat{\beta}_i - a < \beta_i < \hat{\beta}_i + a) = 1 - \alpha$$

- 参数估计量的区间估计的目的就是求得与 α 相对应的a。

2、参数估计量的区间估计

$$t = \frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}} \sim t(n-k-1)$$

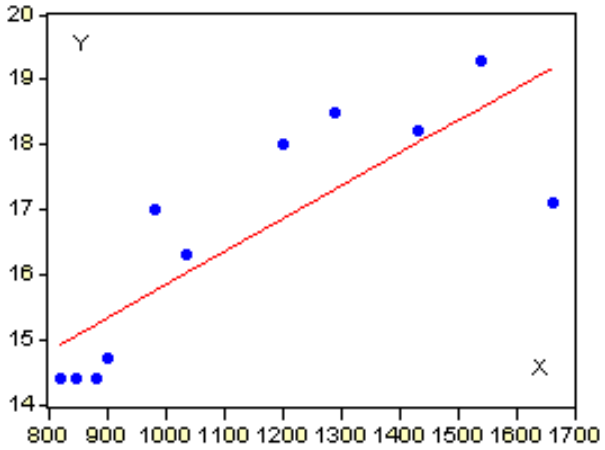
$$P(-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}) = 1-\alpha$$

$$P(-t_{\frac{\alpha}{2}} < \frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}} < t_{\frac{\alpha}{2}}) = 1-\alpha$$

$$P(\hat{\beta}_1 - t_{\frac{\alpha}{2}} \times s_{\hat{\beta}_1} < \beta_1 < \hat{\beta}_1 + t_{\frac{\alpha}{2}} \times s_{\hat{\beta}_1}) = 1-\alpha$$

回归参数的显著性检验与置信区间

例题 人均鲜蛋需求量Y与人均可支配收入X关系



Dependent Variable: Y
Method: Least Squares
Date: 02/12/07 Time: 08:46
Sample: 1988 1998
Included observations: 11

(file: li-2-1)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.76616	1.396736	7.708087	0.0000
X	0.005069	0.001183	4.283328	0.0020

R-squared	0.670895	Mean dependent var	16.57273
Adjusted R-squared	0.634328	S.D. dependent var	1.845042
S.E. of regression	1.115713	Akaike info criterion	3.219829
Sum squared resid	11.20333	Schwarz criterion	3.292174
Log likelihood	-15.70906	F-statistic	18.34690
Durbin-Watson stat	1.320391	Prob(F-statistic)	0.002040

β_1的置信区间：$\hat{\beta}_1 \pm s_{(\hat{\beta}_1)} t_\alpha (T-2) = 0.0051 \pm 2.26 \times 0.0012 = \begin{cases} 0.0024 \\ 0.0078 \end{cases}$

β_0的置信区间：$\hat{\beta}_0 \pm s_{(\hat{\beta}_0)} t_\alpha (T-2) = 10.7662 \pm 2.26 \times 1.3967 = \begin{cases} 7.6097 \\ 13.9227 \end{cases}$

（第2版教材第34页）
（第3版教材第31页）

3、如何缩小置信区间

- 增大样本容量n，因为在同样的样本容量下，n越大，t分布表中的临界值越小，同时，增大样本容量，还可使样本参数估计量的标准差减小；
- 提高模型的拟合优度，因为样本参数估计量的标准差与残差平方和呈正比，模型优度越高，残差平方和应越小。
- 提高样本观测值的分散度。

二、预测值的区间估计

1、问题的提出

计量经济学模型的一个重要应用是经济预测。对于模型

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathrm{B}}$$

如果给定样本以外的解释变量的观测值 $\mathbf{X_0} = (1, X_{10}, X_{20}, \cdots, X_{k0})$，可以得到被解释变量的预测值

$$\hat{Y}_0 = \mathbf{X}_0\hat{\mathrm{B}}$$

但是，严格地说，这只是被解释变量的预测值的估计值，而不是预测值。

为什么？

- 由于随机因素的影响，模型中的参数估计量是不确定的。
- 所以，我们得到的仅能是预测值的一个估计值，预测值仅以某一个置信水平处于以该估计值为中心的一个区间中。
- 于是，又是一个区间估计问题。

- 下面进行置信区间的推导：

2、预测值置信区间的推导

如果已经知道实际的预测值Y_0，那么预测误差为：

$$e_0 = Y_0 - \hat{Y}_0$$

容易证明：e_0服从正态分布，即

$$e_0 \sim N(0, \sigma^2(1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'))$$

取e_0的方差的估计量为

$$\hat{\sigma}_{e_0}^2 = \hat{\sigma}^2(1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'))$$

$$\hat{\sigma}_{e_0} = \hat{\sigma}\sqrt{1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'}$$

构造统计量

$$t = \frac{\hat{Y}_0 - Y_0}{\hat{\sigma}_{e_0}} \sim t(n-k-1)$$

利用该统计量，类似于前面的推导过程，得到在给定（1-α）的置信水平下，预测值$\mathbf{Y}_0$的置信区间为：

$$\hat{y}_0 - t_{\frac{\alpha}{2}} \times \hat{\sigma}_\mu \sqrt{1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'} < y_0$$

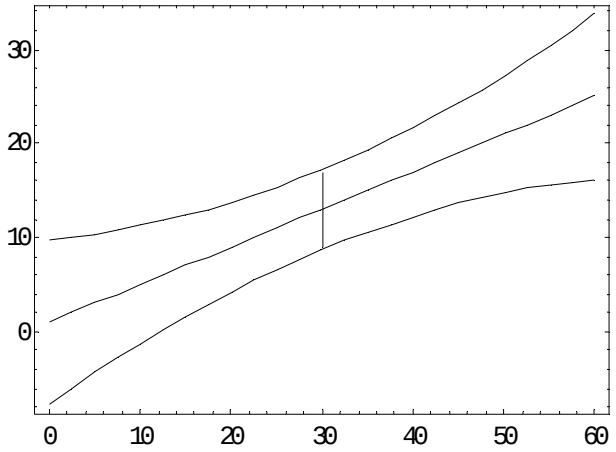
$$< \hat{y}_0 + t_{\frac{\alpha}{2}} \times \hat{\sigma}_\mu \sqrt{1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'}$$

这就是说，当给定解释变量值X_0后，能得到被解释变量Y_0以（1-α）的置信水平处于该区间的结论。

同理，可得均值 $E(\mathbf{Y}\mid\mathbf{X_0})$ 的 $(1-\alpha)$ 预测区间：

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'} < E(\mathbf{Y}\mid\mathbf{X_0}) < \hat{Y}_0 + t_{\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'}$$

y_F 的点预测与区间预测



(1) y_F 的点预测。

根据估计的回归函数，$\hat{y}_F = \hat{\beta}_0 + \hat{\beta}_1 x_F$

(2) 单个 y_F 的区间预测

$\hat{y}_F$ 的分布是

$$\hat{y}_F \sim N(\beta_0 + \beta_1 x_F, \quad \sigma^2 (1 + \frac{1}{T} + \frac{(x_F - \bar{x})^2}{\sum (x_t - \bar{x})^2}))$$

y_F 的区间预测公式是

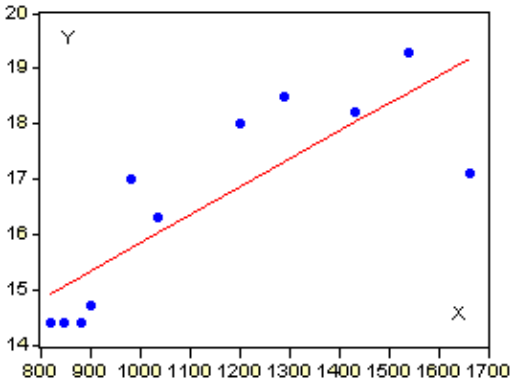
$$\hat{y}_F \pm [t_\alpha(T\text{-}2)\ \hat{\sigma} \sqrt{1 + \frac{1}{T} + \frac{(x_F - \bar{x})^2}{\sum (x_t - \bar{x})^2}}]$$

（第2版教材第38页）
（第3版教材第34页）

y_F 的点预测与区间预测:(演示EViews操作)

例 人均鲜蛋需求量Y与人均可支配收入X关系

(file: li-2-1)



Dependent Variable: Y
Method: Least Squares
Date: 02/12/07 Time: 08:46
Sample: 1988 1998
Included observations: 11

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.76616	1.396736	7.708087	0.0000
X	0.005069	0.001183	4.283328	0.0020

R-squared	0.670895	Mean dependent var	16.57273
Adjusted R-squared	0.634328	S.D. dependent var	1.845042
S.E. of regression	1.115713	Akaike info criterion	3.219829
Sum squared resid	11.20333	Schwarz criterion	3.292174
Log likelihood	-15.70906	F-statistic	18.34690
Durbin-Watson stat	1.320391	Prob(F-statistic)	0.002040

obs	X	Y	YF	YFSE
1998	1663.630	17.10000	19.19836	1.316713
1999	1863.000	NA	20.20887	1.441743
2000	1983.000	NA	20.81710	1.529664

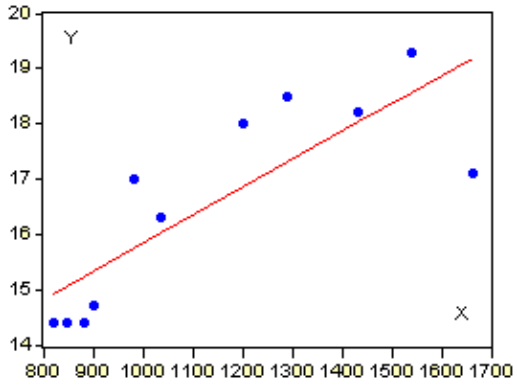
y_F 的点预测

Y_{1999}的点估计值：$Y_{1999} = 10.77 + 0.005069 \times 1863 = 20.21$

Y_{2000}的点估计值：$Y_{2000} = 10.77 + 0.005069 \times 1983 = 20.82$

y_F 的点预测与区间预测

例题2.1 人均鲜蛋需求量Y与人均可支配收入X关系



Dependent Variable: Y
Method: Least Squares
Date: 02/12/07 Time: 08:46
Sample: 1988 1998
Included observations: 11

(file: li-2-1)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.76616	1.396736	7.708087	0.0000
X	0.005069	0.001183	4.283328	0.0020

R-squared	0.670895	Mean dependent var	16.57273
Adjusted R-squared	0.634328	S.D. dependent var	1.845042
S.E. of regression	1.115713	Akaike info criterion	3.219829
Sum squared resid	11.20333	Schwarz criterion	3.292174
Log likelihood	-15.70906	F-statistic	18.34690
Durbin-Watson stat	1.320391	Prob(F-statistic)	0.002040

obs	X	Y	YF	YFSE
1998	1663.630	17.10000	19.19836	1.316713
1999	1863.000	NA	20.20887	1.441743
2000	1983.000	NA	20.81710	1.529664

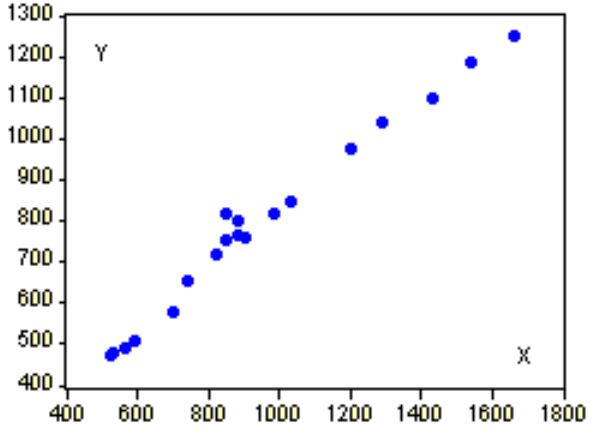
Y_{1999}的点估计值：$Y_{1999} = 10.77 + 0.005069 \times 1863 = 20.21$

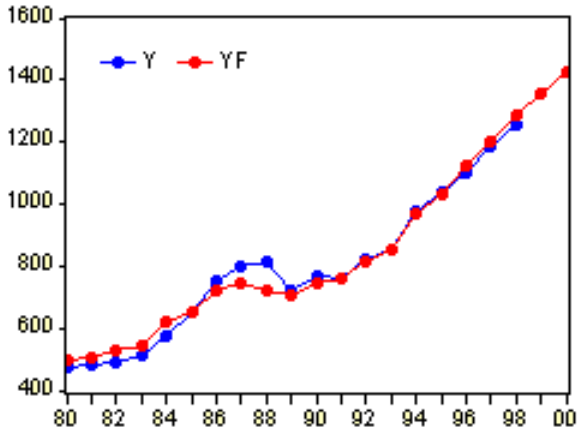
Y_{2000}的点估计值：$Y_{2000} = 10.77 + 0.005069 \times 1983 = 20.82$

Y_{1999}的置信区间：**20.2089±2.26×1.4417 → [16.9507，23.4671]**

Y_{2000}的置信区间：**20.8171±2.26×1.5297 → [17.3600，24.2742]**

案例分析 人均消费性支出与可支配收入关系





(file：li-2-3)

Dependent Variable: Y
Method: Least Squares
Date: 02/12/07 Time: 08:53
Sample: 1980 1998
Included observations: 19

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	135.3063	24.74086	5.468940	0.0000
X	0.691754	0.024671	28.03936	0.0000

R-squared	0.978835	Mean dependent var	790.6132
Adjusted R-squared	0.977590	S.D. dependent var	236.3875
S.E. of regression	35.38729	Akaike info criterion	10.06988
Sum squared resid	21288.43	Schwarz criterion	10.16930
Log likelihood	-93.66389	F-statistic	786.2057
Durbin-Watson stat	0.688666	Prob(F-statistic)	0.000000

整个样本区间预测的EViews操作

补充案例1：用回归模型预测木材剩余物（file:case1）

- 伊春林区位于黑龙江省东北部，有森林面积219万公顷，木材蓄积量为2.3亿m3。森林覆盖率为62.5%，是我国主要的木材工业基地之一。1999年伊春林区木材采伐量为532万m3。按此速度44年之后，1999年的蓄积量将被采伐一空。

- 为缓解森林资源危机，并解决部分职工就业问题，除了做好木材的深加工外，还要充分利用木材剩余物生产林业产品，如纸浆、纸袋、纸板等。因此预测林区的年木材剩余物是安排木材剩余物加工生产的一个关键环节。

表 2.1　年剩余物 y_t 和年木材采伐量 x_t 数据（**file:case1**）

林业局名	年木材剩余物 y_t（万 m^3）	年木材采伐量 x_t（万 m^3）
乌伊岭	**26.13**	**61.4**
东风	**23.49**	**48.3**
新青	**21.97**	**51.8**
红星	**11.53**	**35.9**
五营	**7.18**	**17.8**
上甘岭	**6.80**	**17.0**
友好	**18.43**	**55.0**
翠峦	**11.69**	**32.7**
乌马河	**6.80**	**17.0**
美溪	**9.69**	**27.3**
大丰	**7.99**	**21.5**
南岔	**12.15**	**35.5**
带岭	**6.80**	**17.0**
朗乡	**17.20**	**50.0**
桃山	**9.50**	**30.0**
双丰	**5.52**	**13.8**
合计	**202.87**	**532.00**

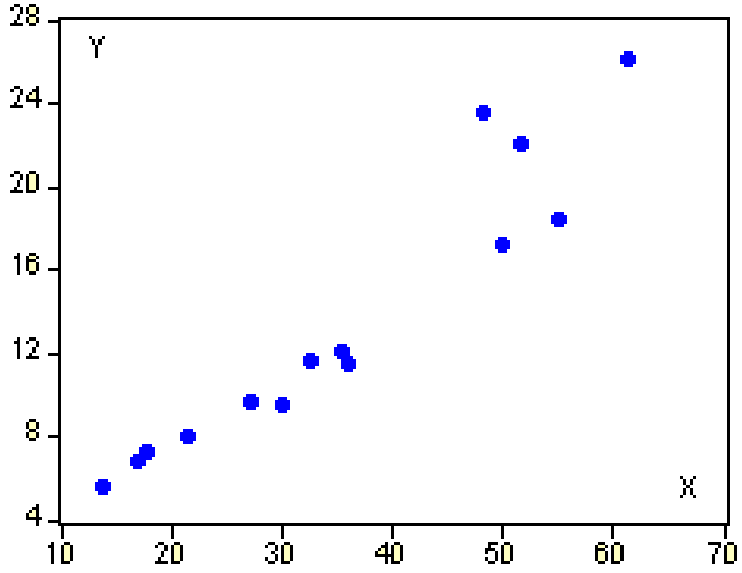
黑龙江省伊春林区

双河大岗
长春
稻田地
嘉荫
(朝阳)
保兴
西克林
乌伊岭区
汤旺河区
红星区
北沟
乌拉嘎
马莲
新青区
五营区
太平沟
上甘岭区
友好区
大黑顶山
1047
先进
兴东
乌马河区
伊春市
延兴
翠峦区
白山狩猎场
美溪区
昆仑气
西林区
金山屯区
鹤北
萝北
(凤翔)
鹤岗市
宝泉岭
带岭森林公园
二股
铁力市
带岭区
南岔区
兴安区
梧桐河
新华
鹤立
(悦来)
桦川
朗乡
神树
桃山
桃山狩猎场
威岭
浩良河
莲江口
汤原
佳木斯市
集贤
平顶山
龙河
龙镇
聚盛
沾北
木沟河
龙山屯
北安市
克东
赵光
南北河
通北
十三井子
海北
双禄
六棵松
四海店
海伦市
东边井
共和
绥棱
海兴
上集
团结
莲花
同乐
双丰
四方台
望奎
庆安
王杨
绥化市
民乐

观测点近似服从线性关系。

建立一元线性回归模型如下：

$$y_t = \beta_0 + \beta_1 x_t + u_t$$



年剩余物y_t和年木材采伐量x_t散点图

Dependent Variable: Y
Method: Least Squares
Date: 09/04/07 Time: 07:41
Sample: 1 16
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.762928	1.220966	-0.624856	0.5421
X	0.404280	0.033377	12.11266	0.0000

R-squared	0.912890	Mean dependent var	12.67938
Adjusted R-squared	0.906668	S.D. dependent var	6.665466
S.E. of regression	2.036319	Akaike info criterion	4.376633
Sum squared resid	58.05231	Schwarz criterion	4.473207
Log likelihood	-33.01306	F-statistic	146.7166
Durbin-Watson stat	1.481946	Prob(F-statistic)	0.000000

$\hat{y}_t$ = **-0.7629 + 0.4043** x_t

(-0.6) (12.1) $R^2 = 0.91, T = 16$

上述模型的经济解释是，对于伊春林区每采伐**1 m³**木材，

将平均产生**0.4 m³**的剩余物。

分析**EViews**输出结果。注意：**S.D.**和**s.e.**的区别。**s.e.**和**SSE**的关系。

点预测：

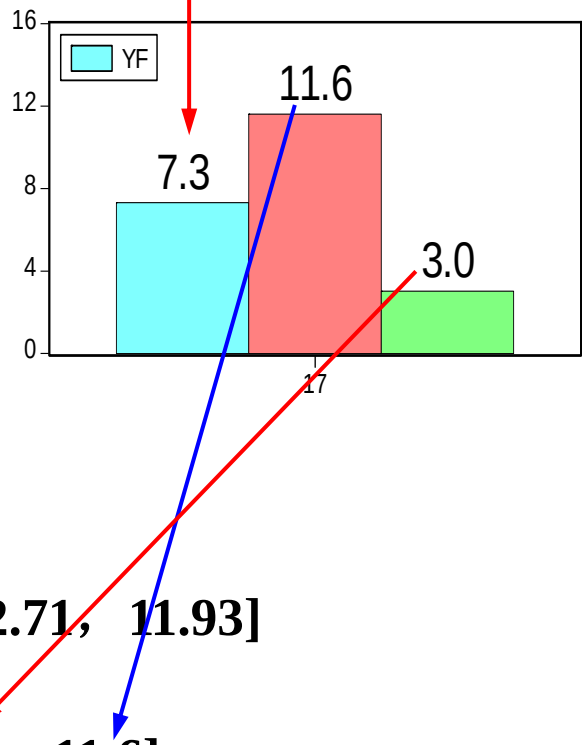
$$\hat{y}_{2000} = -0.7629 + 0.4043\, x_{2000} = -0.7629 + 0.4043 \times 20 = 7.3231 \text{ 万 } \mathrm{m}^3$$

木材剩余物产出量单点 y_t 的置信区间的计算。

$$s^2(\hat{y}_{2000}) = \hat{\sigma}^2 \left(1 + \frac{1}{T} + \frac{(x_F - \bar{x})^2}{\sum (x - \bar{x})^2}\right)$$

$$= 4.1453 \left(1 + \frac{1}{16} + \frac{(20 - 33.25)^2}{3722.2606}\right) = 4.5999$$

$$s(\hat{y}_{2000}) = \sqrt{4.5999} = 2.1447$$



$$\hat{y}_{2000} \pm t_{0.05\,(14)}\, s(\hat{y}_{2000}) = 7.3231 \pm 2.15 \times 2.145 = [2.71,\ 11.93]$$

简单计算（临界值取 2）：$7.3231 \pm 2 \times 2.145 = [3.0,\ 11.6]$

问题 1：估计结果中 $\hat{\beta}_0$ 没有显著性，去掉截距项 β_0 可以吗？

答：依据实际情况处理。一般不做剔除。

本案例剔出截距项后的估计结果是

$$\hat{y}_t = 0.3853\, x_t$$

$$(28.3) \qquad R^2 = 0.91,\ N = 16$$

点预测值是

$$\hat{y}_{2000} = 0.3853\, x_{2000} = 0.3853 \times 20 = 7.7060 \text{ 万 } \mathrm{m}^3$$

问题 2：估计一元线性回归模型，最少需要多少组观测值？

问题3：为什么离群值对回归参数OLS估计量的影响大？

EViews操作

附录1：怎样建立EViews新工作文件。

附录2：怎样用EViews通过键盘输入，复制、粘贴功能输入数据。

注意：

（1）变量命名时，字符不得超过16个。

（2）给变量命名时，避免使用下列名字：ABS，ACOS， AR，ASIN，C，CON，CNORM， COEF，COS，D，DLOG，DNORM，ELSE，ENDIF，EXP，LOG，LOGIT，LPT1， LPT2，MA，NA，NRND，PDL，RESID，RND，SAR， SIN，SMA，SQR，THEN。

附录3：OLS估计的操作步骤。Quick→Estimate Equation。对话框中输入 y c x 。OK键。

附录4：怎样用EViews预测。

相关系数

相关（**correlation**）：指两个或两个以上变量间相互关系的程度或强度。

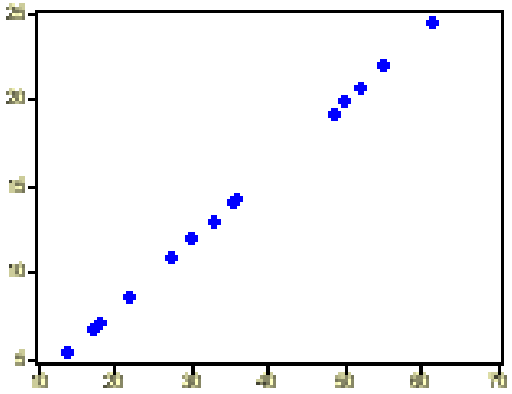
分类：①按强度分

完全相关：变量间存在函数关系。

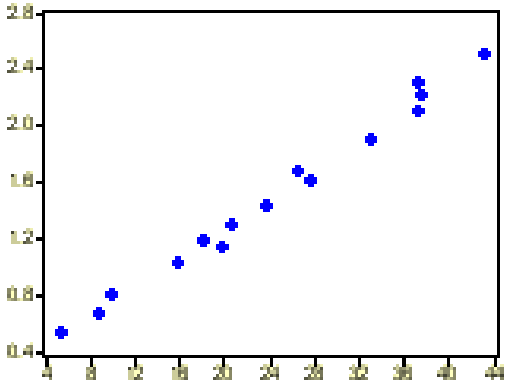
高度相关（强相关）：变量间近似存在函数关系。

弱相关：变量间有关系但不明显。

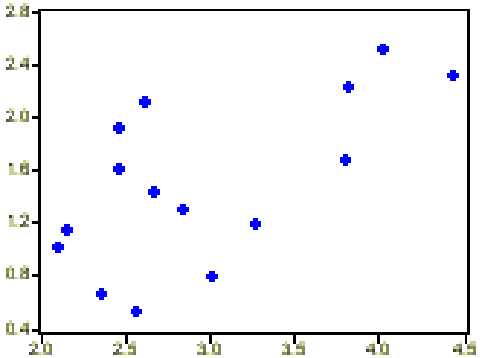
零相关：变量间不存在任何关系。



完全相关



高度相关、线性相关、正相关



弱相关

（第2版教材第28页）

（第3版教材第26页）

相关系数

②按变量个数分

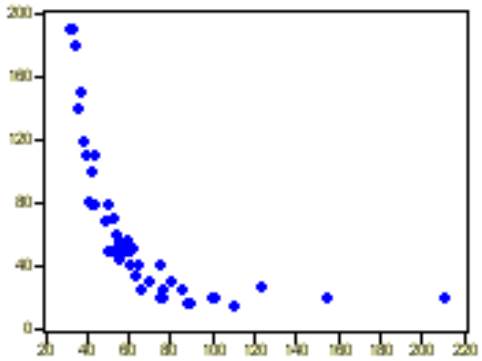
简单相关：指两个变量间相关。

按形式分：线性相关, 非线性相关

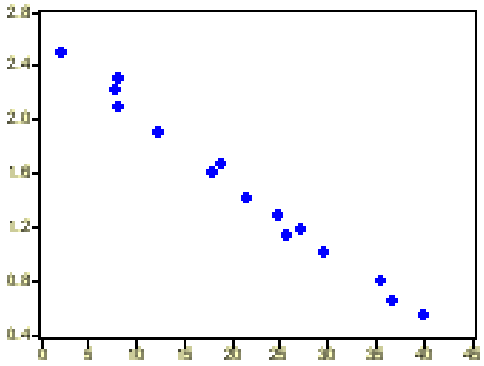
按符号分：正相关, 负相关, 零相关

复相关（多重相关和偏相关）：

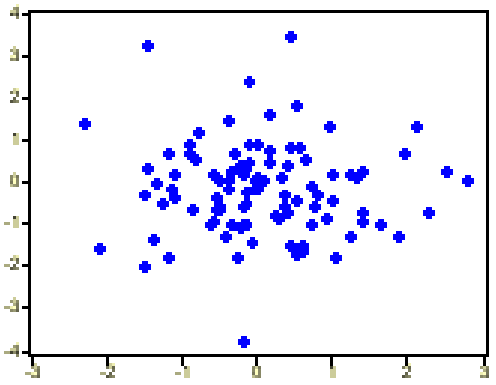
指3个或3个以上变量间的相关。



非线性相关



负相关



零相关

简单线性相关的度量

简单线性相关系数，简称相关系数（**correlation coefficient**）。度量两个变量间的线性相关强度，用 ρ 表示。

ρ 的随机变量表达式是

$$\rho = \frac{Cov(x_t, y_t)}{\sqrt{D(x_t)}\sqrt{D(y_t)}}。$$

ρ 的统计表达式是

$$\rho = \frac{\frac{1}{T}\sum_{t=1}^{T}(x_t - \mu_x)(y_t - \mu_y)}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t - \mu_x)^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \mu_y)^2}}$$

$$= \frac{\sum_{t=1}^{T}(x_t - \mu_x)(y_t - \mu_y)}{\sqrt{\sum_{t=1}^{T}(x_t - \mu_x)^2}\sqrt{\sum_{t=1}^{T}(y_t - \mu_y)^2}}$$

其中 T，总体容量；x_t, y_t，变量的观测值；μ_x, μ_y，变量观测值的均值。

对样本来说，相关系数用 ***r*** 表示， ***r*** 是总体相关系数 ρ 的估计值。

$$r = \hat{\rho} = \frac{\frac{1}{T-1}\sum_{t=1}^{T}(x_t-\bar{x})(y_t-\bar{y})}{\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(x_t-\bar{x})^2}\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(y_t-\bar{y})^2}}$$

$$= \frac{\sum_{t=1}^{T}(x_t-\bar{x})(y_t-\bar{y})}{\sqrt{\sum_{t=1}^{T}(x_t-\bar{x})^2}\sqrt{\sum_{t=1}^{T}(y_t-\bar{y})^2}}$$

其中 T，样本容量；x_t, y_t，变量的观测值；$\bar{x}$，$\bar{y}$，变量观测值的均值。

相关系数的取值范围

(1) 当两个变量严格服从线性关系时，$|\rho| = 1$。

证：设直线斜率为 k， 即 $y = a + kx$。则有

$$\rho = \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum(x_t - \bar{x})^2 \sum(y_t - \bar{y})^2}} = \frac{\sum(x_t - \bar{x})k(x_t - \bar{x})}{\sqrt{\sum(x_t - \bar{x})^2 k^2 \sum(x_t - \bar{x})^2}} = 1$$

(2) 当两个变量不存在线性关系时，$|\rho| = 0$。

(3)上述是两种极端情形，所以相关系数的取值范围是 **[-1，1]**。

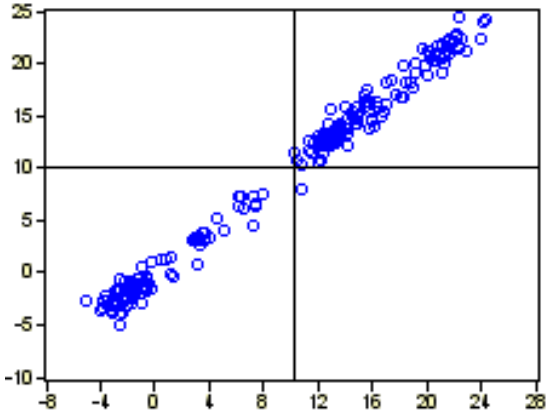


图1 正相关

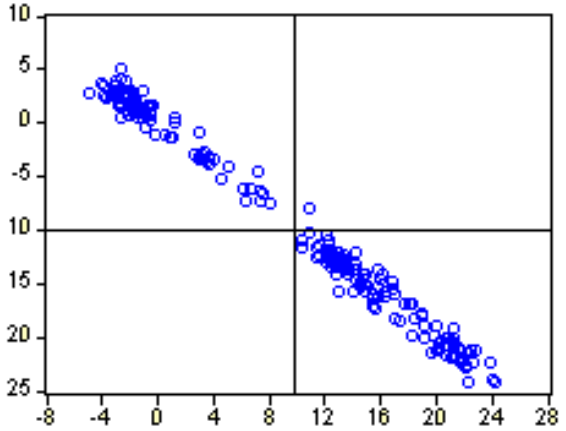


图2 负相关

散点图与相关系数 值的对应关系

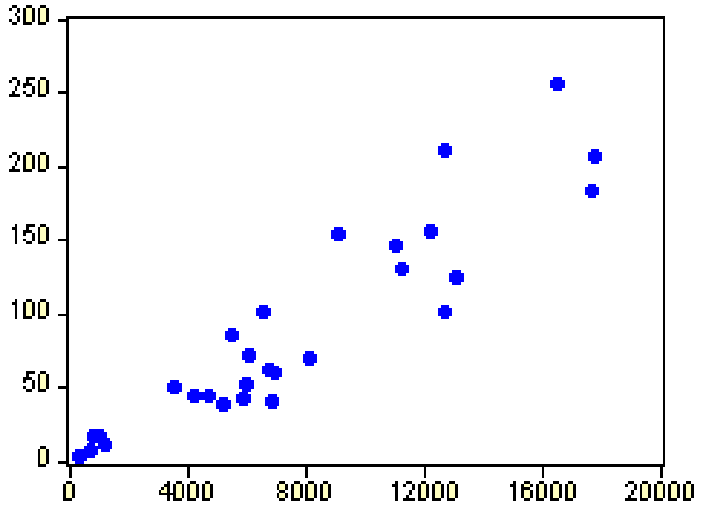


图3 $r = 0.92$

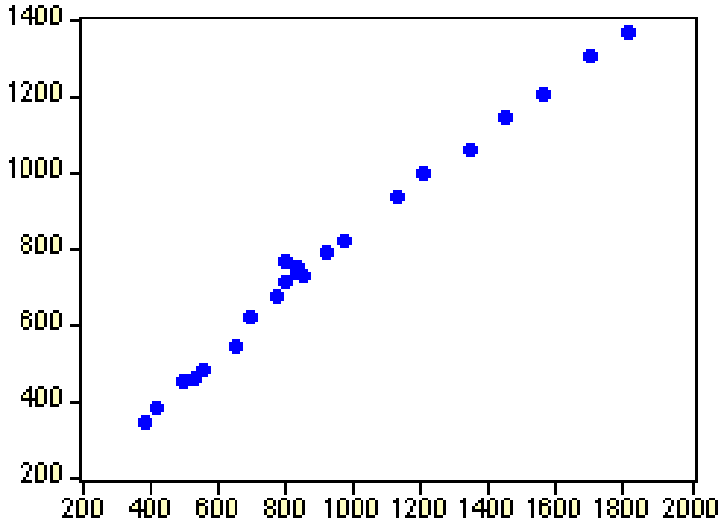


图4 $r = 0.99$

线性相关系数的局限性

(1) 只适用于考察变量间的线性相关关系。变量不相关与变量相互独立在概念上是不同的。

(2) 相关系数的计算是一个数学过程,但不能揭示变量间关系的实质。

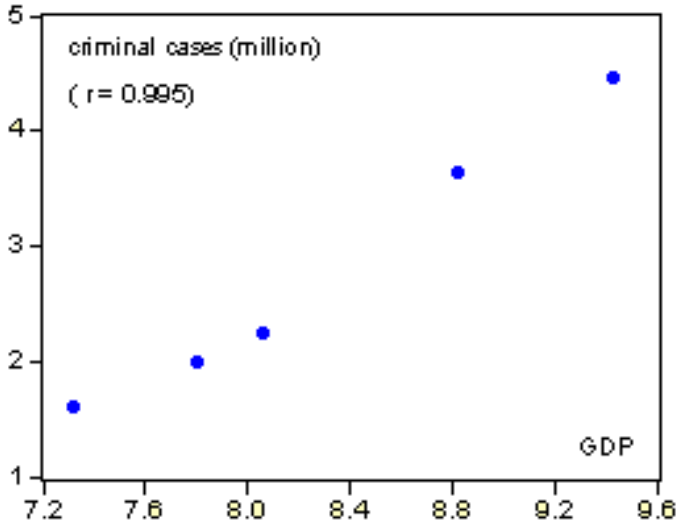
(3) 一般说二变量相关时，可能属于如下一种关系。

单向因果关系。如施肥量与农作物产量；对金属的加热时间与温度值。

双向因果关系。如工业生产与农业生产；商品供给量与商品价格。

另有隐含因素影响二变量变化。

虚假相关。



(1997-2001， file: 5correlation1) 13宗/分

简单相关系数的检验

t 检验

H_0：$\rho = 0$；　H_1：$\rho \neq 0$

$$t = \frac{r-\rho}{s_r} = (r-\rho)\Big/\sqrt{\frac{1-r^2}{T-2}} = r\sqrt{\frac{T-2}{1-r^2}} \sim t(T-2)$$

其中 2 表示涉及两个变量。

若 $|t| > t_\alpha(T\text{-}2)$，则 x_t 和 y_t 相关；

若 $|t| < t_\alpha(T\text{-}2)$，则 x_t 和 y_t 不相关。

相关系数的EViews操作：打开数据窗口。选View/Correlation

补充案例2：刻卜勒（J. Kepler）行星运行第三定律

(file:5kepler3)

刻卜勒（Johannes Kepler, 1571-1630）

把地球与太阳的距离定为 1 个单位。地球绕太阳公转一周的时间为 1 个单位（年）。那么太阳系 9 个行星绕太阳各公转一周所需时间（*T*）和与太阳的距离（*D*）的数据如下：

obs	水星	金星	地球	火星	木星	土星	天王星	海王星	冥王星
Time	0.24	0.615	1	1.88	11.9	29.5	84	165	248
DISTANCE	0.387	0.723	1	1.52	5.2	9.54	19.2	30.1	39.5

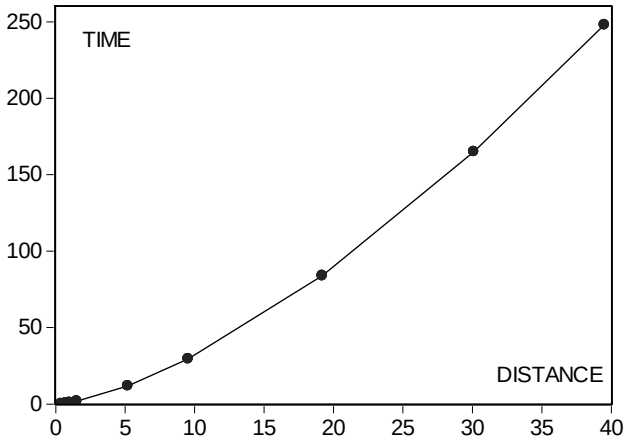
注：第谷（B. Tycho）的观测数据。

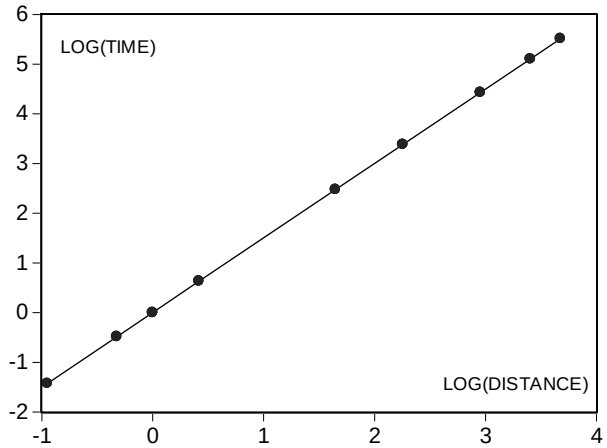
刻卜勒（Johannes Kepler, 1571-1630，德国人）坚信 9 个行星绕太阳运行，一定有规律可循。经过艰苦的努力，他终于 1619 年发表了行星运行三定律。其中行星运行第三定律是：“行星在轨道上运行一周的时间的平方与其至太阳的平均距离的立方成正比例。”即 $T^2 = D^3$。

obs	水星	金星	地球	火星	木星	土星	天王星	海王星	冥王星
Time	0.24	0.615	1	1.88	11.9	29.5	84	165	248
DISTANCE	0.387	0.723	1	1.52	5.2	9.54	19.2	30.1	39.5
T^2	0.057	0.378	1	3.534	141.6	870.2	7056	27225	61504
D^3	0.057	0.377	1	3.512	140.6	868.3	7078	27271	61630

资料来源：《科学发现纵横谈》，王梓坤，上海人民出版社，1978。

用回归分析验证第三定律 （file:6kepler3）





Dependent Variable: LOG(TIME)
Method: Least Squares
Date: 08/25/07 Time: 23:15
Sample: 1 9
Included observations: 9

	Coefficient	Std. Error	t-Statistic	Prob.
LOG(DISTANCE)	1.500033	0.000334	4492.202	0.0000

R-squared	0.999999	Mean dependent var	2.181016
Adjusted R-squared	0.999999	S.D. dependent var	2.587182
S.E. of regression	0.002185	Akaike info criterion	-9.309794
Sum squared resid	3.82E-05	Schwarz criterion	-9.287880
Log likelihood	42.89407	Hannan-Quinn criter.	0.952167

$$\log(T) = 1.5 \log(D) + \hat{u}_t$$

(4492)

$R^2 = 0.999999$, $N = 9$

$\log(T) = (3/2) \log(D)$

$2 \log(T) = 3 \log(D)$

$\log(T^2) = \log(D^3)$

$T^2 = D^3$

3、一点启示

- 计量经济学模型用于预测时，必须严格科学地描述预测结果。
- 如果要求给出一个“准确”的预测值，那么真实值与该预测值相同的概率为0。
- 如果要求以100%的概率给出区间，那么该区间是∞。
- 模型研制者的任务是尽可能地缩小置信区间。

4、如何缩小置信区间

- 增大样本容量n，因为在同样的样本容量下，n越大，t分布表中的临界值越小，同时，增大样本容量，还可使随机误差项的标准差减小；
- 提高模型的拟合优度，模型优度越高，残差平方和应越小。
- 提高样本观测值的分散度。

三、多元线性回归分析计算步骤及主要公式

1、由样本观测值$(Y_i, X_{1i}, X_{2i}, \cdots, X_{ki}), (i = 1,2,\cdots,n)$，写出

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad ; \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}_{n\times(k+1)}$$

2、计算 $\mathbf{X'X}$，$(\mathbf{X'X})^{-1}$，$\mathbf{X'Y}$

3、计算 OLS 估计量

$$\hat{\mathrm{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

4、计算残差及残差平方和

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

$$\mathbf{e}'\mathbf{e} = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}$$

5、计算随机项标准差的估计值

$$\hat{\sigma} = \sqrt{\frac{\mathbf{e}'\mathbf{e}}{n-k-1}}$$

6、作拟合优度检验

$$r^2 = \frac{\hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2}{\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2}$$

$$R^2 = 1 - (1 - R^2)\frac{(n-1)}{(n-k-1)}$$

7、计算样本参数估计值的标准差

$$S(\hat{\beta}_i) = \hat{\sigma}^2 \sqrt{c_{ii}} = \sqrt{\frac{\sum e_i^2}{n-k-1} c_{ii}}$$

其中，$c_{ii} = (\mathbf{X}'\mathbf{X})^{-1}_{ii}$

8、进行 F 检验与 t 检验

$$F = \frac{ESS/k}{RSS/(n-k-1)} = \frac{(\hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2)/k}{(\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{X})/(n-k-1)}$$

$$t = \frac{\hat{\beta}_i}{S(\hat{\beta}_i)}$$

9、在 $X = X_0$ 处进行点预测与区间预测

$$Y_0 = \mathbf{X_0}\hat{\mathbf{B}}$$

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{1 + \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_0} < Y_0 < \hat{Y}_0 + t_{\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{1 + \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_0}$$

例：设某中心城市对各地区商品流出量Y取决于各地区的社会商品购买力X_1以及各地区对该市的商品流入X_2，即可能有如下总体回归方程：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2$$

在下列样本下进行回归分析：

地区	该市对各地区销售额 Y（万元）	各地区社会购买力 X 1（亿元）	各地区商品流入该市量 X 2（万元）
1	6800	1300	400
2	1900	350	1200
3	2800	180	700
4	1000	340	400
5	700	70	1600
6	500	200	1200
7	60	30	240
8	50	20	400

（1）估计总体回归模型 $\mathbf{Y}=\mathbf{X}\hat{\mathbf{B}}+\mathbf{N}$

$$\mathbf{X}'\mathbf{X}=\begin{pmatrix}8 & 2490 & 6140\\ 2490 & 2006700 & 1569200\\ 6140 & 1569200 & 6467600\end{pmatrix},\quad \mathbf{X}'\mathbf{Y}=\begin{pmatrix}13810\\ 10500800\\ 9114400\end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1}=\begin{pmatrix}0.64282 & -0.00040 & -0.0005\\ -0.0004 & 8.58299E-07 & 1.6718E-07\\ -0.00051 & 1.67183E-07 & 6.0231E-07\end{pmatrix}$$

参数的最小二乘估计

$$\hat{B}=\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}=\begin{pmatrix}37.085\\ 5.075\\ 0.143\end{pmatrix}$$

（2）统计检验

方差分析计算表

	Y	Yˆ	y	y^2	e=Y-Yˆ	e^2	yˆ	$Yˆ^2$
	6800	6692	5074	25742939	108	11654	4966	24659152
	1900	1985	174	30189	-85	7159	258	66749
	2800	1050	1074	1152939	1750	3060796	-676	456656
	1000	1820	-726	527439	-820	671984	93	8742
	700	621	-1026	1053189	79	6308	-1106	1222517
	500	1223	-1226	1503689	-723	523179	-503	252947
	60	224	-1666	2776389	-164	26757	-1503	2258026
	50	196	-1676	2809814	-146	21213	-1531	2342749
和	13810			35596588		4329050		31267537
均值	1726			TSS		RSS		ESS
			自由度	8-1=7		8-3=5		3-1=2
			均方差			865810		15633769

拟合优度检验：

$$r^2 = \frac{ESS}{TSS} = \frac{31267537}{35596588} = 0.8784 \quad ,$$

$$R^2 = 1 - (1 - 0.8784) \times \frac{8-1}{8-3} = 0.8297$$

总体显著性检验（F检验）：

$$F = \frac{ESS/(3-1)}{RSS(8-3)} = 18.06$$

查表，在 5%的显著性水平下，临界值 F_α (3-1,8-3)=5.79

因为通过样本计算的 F 值大于临界值 F_α，因此模型总体上是显著的。

参数显著性检验（t检验）

参数估计的方差-协方差矩阵

$$\text{var}-\text{cov}(\hat{B}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$=865810\begin{pmatrix} 0.64282 & -0.00040 & -0.0005 \\ -0.0004 & 8.58299E-07 & 1.6718E-07 \\ -0.00051 & 1.67183E-07 & 6.0231E-07 \end{pmatrix}$$

对参数 $\hat{\beta}_0$、$\hat{\beta}_1$、$\hat{\beta}_2$ 分别作 t 检验：

$\hat{\beta}_0$：
$$t = \frac{\hat{\beta}_0}{S(\hat{\beta}_0)} = \frac{\hat{\beta}_0}{\sqrt{\sigma^2 c_{00}}} = \frac{37.09}{\sqrt{865810 \times 0.6428}} = 0.0497$$

$\hat{\beta}_1$：
$$t = \frac{\hat{\beta}_1}{S(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{\sigma^2 c_{11}}} = \frac{5.075}{\sqrt{865810 \times 8.583E-07}} = 5.888$$

$\hat{\beta}_2$：
$$t = \frac{\hat{\beta}_2}{S(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{\sqrt{\sigma^2 c_{22}}} = \frac{0.143}{\sqrt{865810 \times 6.023E-07}} = 0.198$$

查表：在 5%的显著性水平下，$t_{0.05/2}$=2.571，

因此，$\hat{\beta}_1$显著不为 0，而$\hat{\beta}_2$显著为 0，

说明各地区商品流入量X_2不是一个重要的影响因素，而各地区社会购买力X_1是重要的因素。

在模型中册去X_2，重新建立模型

$$Y = \beta_0 + \beta_1 X_1 + \mu$$

利用表中资料通过 OLS 法得到的回归结果如下：

$$\hat{Y} = 158.88 + 5.0357 X_1$$

t:（0.4128）(6.540)

r^2=0.8774　　F=42.9544

Eviews 软件输出

（1）模型中包括X1与X2：

Dependent Variable: Y
Method: Least Squares
Date: 02/07/03 Time: 09:39
Sample: 1 8
Included observations: 8

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	37.08504	746.0306	0.049710	0.9623
X1	5.075314	0.862046	5.887521	0.0020
X2	0.142636	0.722144	0.197518	0.8512

R-squared	0.878386	Mean dependent var	1726.250
Adjusted R-squared	0.829740	S.D. dependent var	2255.045
S.E. of regression	930.4891	Akaike info criterion	16.78929
Sum squared resid	4329050.	Schwarz criterion	16.81908
Log likelihood	-64.15718	F-statistic	18.05681
Durbin-Watson stat	2.718540	Prob(F-statistic)	0.005158

（2）模型中仅包括X1

Dependent Variable: Y
Method: Least Squares
Date: 02/07/03 Time: 09:42
Sample: 1 8
Included observations: 8

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	158.8811	384.8171	0.412874	0.6940
X1	5.035723	0.768348	6.553959	0.0006

R-squared	0.877437	Mean dependent var	1726.250
Adjusted R-squared	0.857010	S.D. dependent var	2255.045
S.E. of regression	852.7239	Akaike info criterion	16.54707
Sum squared resid	4362828.	Schwarz criterion	16.56693
Log likelihood	-64.18827	F-statistic	42.95437
Durbin-Watson stat	2.731033	Prob(F-statistic)	0.000604