例4　假设模型为 $Y_t = \alpha + \beta X_t + \mu_t$。给定 n 个观察值 (X_1, Y_1)，(X_2, Y_2)，…，(X_n, Y_n)，按如下步骤建立 β 的一个估计量：在散点图上把第 1 个点和第 2 个点连接起来并计算该直线的斜率；同理继续，最终将第 1 个点和最后一个点连接起来并计算该条线的斜率；最后对这些斜率取平均值，称之为 $\hat{\beta}$，即 β 的估计值。

（1）画出散点图，给出 $\hat{\beta}$ 的几何表示并推出代数表达式。

（2）计算 $\hat{\beta}$ 的期望值并对所做假设进行陈述。这个估计值是有偏的还是无偏的？解释理由。

（3）证明为什么该估计值不如我们以前用 OLS 方法所获得的估计值，并做具体解释。

例5　试证：

(1)模型 $y_i=\beta_0+u_i$　$(i=1,2,\cdots,n)$ 中 β_0 的最小二乘估计量为 $\hat{\beta}_0=\bar{y}$；

(2)如(1)中的随机项满足经典回归的基本假定，则有

$$E(\hat{\beta}_0)=\beta_0,V(\hat{\beta}_0)=\frac{1}{n}\sigma_u^2$$

§2.3 一元线性回归模型的统计检验

一、拟合优度检验

二、变量的显著性检验

三、参数的置信区间

- 回归分析是要通过样本所估计的参数来代替总体的真实参数，或者说是用样本回归线代替总体回归线。

- 尽管从统计性质上已知，如果有足够多的重复 抽样，参数的估计值的期望（均值）就等于其总体的参数真值，但在一次抽样中，估计值不一定就等于该真值。
- 那么，在一次抽样中，参数的估计值与真值的差异有多大，是否显著，这就需要进一步进行统计检验。
- 主要包括拟合优度检验、变量的显著性检验及参数的区间估计。

一、拟合优度检验

拟合优度检验：对样本回归直线与样本观测值之间拟合程度的检验。

度量拟合优度的指标：判定系数（可决系数）R^2

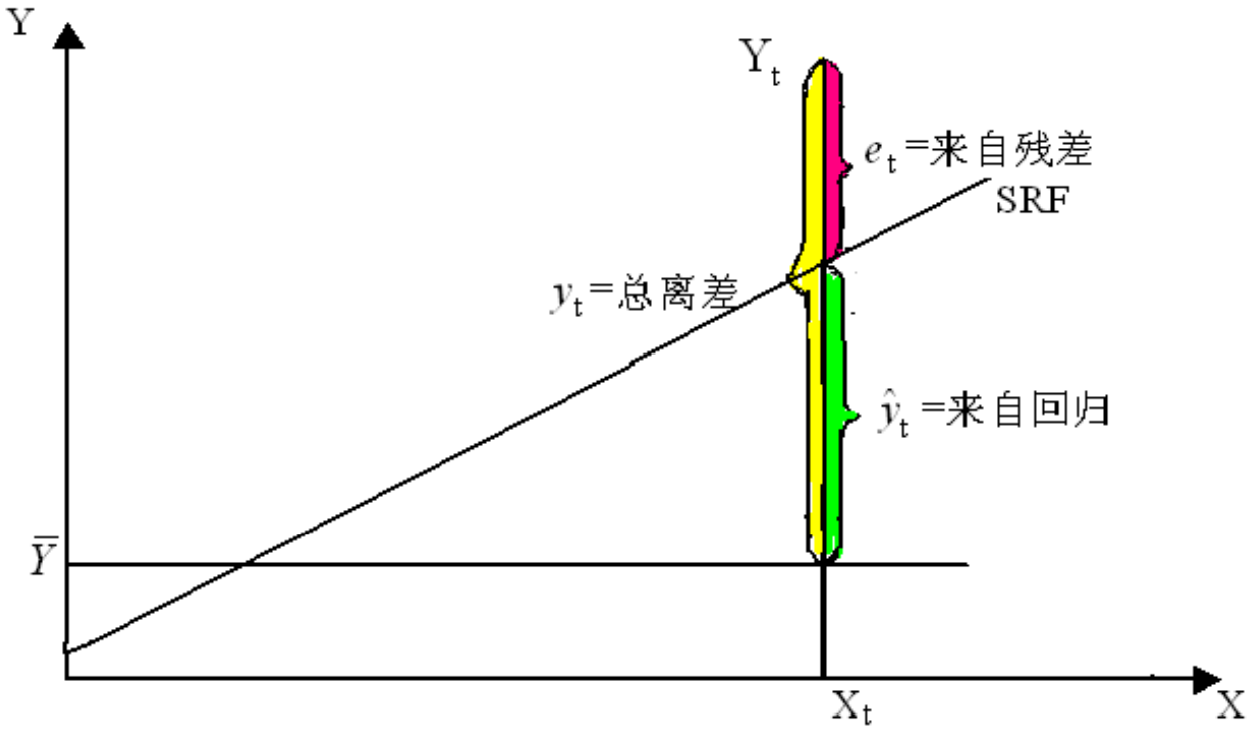
1、总离差平方和的分解

已知由一组样本观测值（X_i, Y_i），i=1,2…,n
得到如下样本回归直线

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

而Y的第*i*个观测值与样本均值的离差 $y_i = (Y_i - \bar{Y})$
可分解为两部分之和

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



$\hat{y}_i = (\hat{Y}_i - \bar{Y})$是样本回归拟合值与观测值的平均值之差，可认为是**由回归直线解释的部分**；

$e_i = (Y_i - \hat{Y}_i)$是实际观测值与回归拟合值之差，是**回归直线不能解释的部分**。

对于所有样本点，则需考虑这些点与样本均值离差的平方和,可以证明：

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

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

记 $TSS = \sum y_i^2 = \sum (Y_i - \bar{Y})^2$ 总体平方和（**Total Sum of Squares**）

$ESS = \sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{Y})^2$ 回归平方和（**Explained Sum of Squares**）

$RSS = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$ 残差平方和（**Residual Sum of Squares**）

TSS=ESS+RSS

Y的观测值围绕其均值的**总离差(total variation)**可分解为两部分：**一部分来自回归线(ESS)，另一部分则来自随机势力(RSS)**。

2、可决系数R^2统计量

记 $$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

称 R^2 为（样本）可决系数/判定系数（coefficient of determination)。

可决系数的取值范围：[0，1]

R^2越接近1，说明实际观测点离样本线越近，拟合优度越高。

二、变量的显著性检验

回归分析是要判断解释变量X是否是被解释变量Y的一个显著性的影响因素。

在一元线性模型中，就是要判断X是否对Y具有显著的线性性影响。这就需要进行变量的显著性检验。

计量经计学中，主要是针对变量的参数真值是否为零来进行假设检验的。

1、假设检验

- 所谓假设检验，就是事先对总体参数或总体分布形式作出一个假设，然后利用样本信息来判断原假设是否合理，即判断样本信息与原假设是否有显著差异，从而决定是否接受或否定原假设。
- 假设检验采用的逻辑推理方法是反证法。

先假定原假设正确，然后根据样本信息，观察由此假设而导致的结果是否合理，从而判断是否接受原假设。

- 判断结果合理与否，是基于“小概率事件不易发生”这一原理的

2、变量的显著性检验

对于一元线性回归方程中的 $\hat{\beta}_1$，已经知道它服从正态分布

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum x_i^2})$$

由于真实的σ^2未知，在用它的无偏估计量 $\hat{\sigma}^2 = \sum e_i^2 / (n-2)$ 替代时，可构造如下统计量

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / \sum x_i^2}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t(n-2)$$

检验步骤：

（1）对总体参数提出假设

H_0：　$\beta_1=0$，　　H_1：$\beta_1\neq0$

（2）以原假设H0构造t统计量，并由样本计算其值

$$t=\frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

（3）给定显著性水平α，查t分布表，得临界值$t_{\alpha/2}(n-2)$

(4) 比较，判断

若　$|t|>t_{\alpha/2}(n-2)$，则拒绝H_0，接受H_1；

若　$|t|\leq t_{\alpha/2}(n-2)$，则拒绝$H_1$，接受$H_0$；

对于一元线性回归方程中的β_0，可构造如下t统计量进行显著性检验：

$$t=\frac{\hat{\beta}_0-\beta_0}{\sqrt{\hat{\sigma}^2\sum X_i^2/n\sum x_i^2}}=\frac{\hat{\beta}_0}{S_{\hat{\beta}_0}}\sim t(n-2)$$

接例2.2.1：在上述家庭可支配收入-消费支出例中，进行显著性检验。

表 2.2.1　　参数估计的计算表

	X_i	Y_i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2	X_i^2	Y_i^2
1	800	594	-1350	-973	1314090	1822500	947508	640000	352836
2	1100	638	-1050	-929	975870	1102500	863784	1210000	407044
3	1400	1122	-750	-445	334050	562500	198381	1960000	1258884
4	1700	1155	-450	-412	185580	202500	170074	2890000	1334025
5	2000	1408	-150	-159	23910	22500	25408	4000000	1982464
6	2300	1595	150	28	4140	22500	762	5290000	2544025
7	2600	1969	450	402	180720	202500	161283	6760000	3876961
8	2900	2078	750	511	382950	562500	260712	8410000	4318084
9	3200	2585	1050	1018	1068480	1102500	1035510	10240000	6682225
10	3500	2530	1350	963	1299510	1822500	926599	12250000	6400900
求和	21500	15674			5769300	7425000	4590020	53650000	29157448
平均	2150	1567							

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{5769300}{7425000} = 0.777$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_0 \bar{X} = 1567 - 0.777 \times 2150 = -103.172$$

因此，由该样本估计的回归方程为：

$$\hat{Y}_i = -103.172 + 0.777 X_i$$

下面进行显著性检验，首先计算σ_2的估计值

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum y_i^2 - \hat{\beta}_1^2 \sum x_i^2}{n-2} = \frac{4590020 - 0.777^2 \times 7425000}{10-2} = 13402$$

于是$\hat{\beta}_1$和$\hat{\beta}_0$的标准差的估计值分别是：

$$S_{\hat{\beta}_1} = \sqrt{\hat{\sigma}^2 / \sum x_i^2} = \sqrt{13402 / 7425000} = \sqrt{0.0018} = 0.0425$$

$$S_{\hat{\beta}_0} = \sqrt{\hat{\sigma}^2 \sum X_i^2 / n \sum x_i^2} = \sqrt{13402 \times 53650000 / 10 \times 7425000} = 98.41$$

t统计量的计算结果分别为：

$$t_1 = \hat{\beta}_1 \big/ S_{\hat{\beta}_1} = 0.777/0.0425 = 18.29$$

$$t_0 = \hat{\beta}_0 \big/ S_{\hat{\beta}_0} = -103.17/98.41 = -1.048$$

给定显著性水平α=0.05，查t分布表得临界值

$$t_{0.05/2}(8) = 2.306$$

$|t_1| > 2.306$，说明家庭可支配收入在95%的置信度下显著，即是消费支出的主要解释变量；

$|t_2| < 2.306$, 表明在95%的置信度下，无法拒绝截距项为零的假设。

三、参数的置信区间

回归分析希望通过样本所估计出的参数$\hat{\beta}_1$来代替总体的参数β_1。

要判断样本参数的估计值在多大程度上可以“近似”地替代总体参数的真值，往往需要通过构造一个以样本参数的估计值为中心的“区间”，来考察它以多大的可能性（概率）包含着真实的参数值。这种方法就是参数检验的置信区间估计。

要判断估计的参数值$\hat{\beta}$离真实的参数值β有多“近”，可预先选择一个概率$\alpha(0<\alpha<1)$，并求一个正数δ，使得随机区间$(\hat{\beta}-\delta,\hat{\beta}+\delta)$包含参数的真值的概率为1-$\alpha$。即：

$$P(\hat{\beta}-\delta\le\beta\le\hat{\beta}+\delta)=1-\alpha$$

如果存在这样一个区间，称之为**置信区间**（**confidence interval**）；1-α称为**置信系数**（置信度）（**confidence coefficient**），α称为**显著性水平**（**level of significance**）；置信区间的端点称为**置信限**（**confidence limit**）或**临界值**（**critical values**）。

一元线性模型中，β_i (i=1，2）的置信区间：

在变量的显著性检验中已经知道：

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

意味着，如果给定置信度（1−α），从分布表中查得自由度为(n−2)的临界值，那么t值处在$(-t_{\alpha/2},\ t_{\alpha/2})$的概率是(1−α)。表示为：

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

即

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

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

于是得到:(1−α)的置信度下，β_i的置信区间是

$$(\hat{\beta}_i - t_{\frac{\alpha}{2}} \times s_{\hat{\beta}_i}, \hat{\beta}_i + t_{\frac{\alpha}{2}} \times s_{\hat{\beta}_i})$$

在上述收入−消费支出例中，如果给定α =0.01，查表得:

$$t_{\frac{\alpha}{2}}(n-2) = t_{0.005}(8) = 3.355$$

由于 $S_{\hat{\beta}_1} = 0.042$ $S_{\hat{\beta}_0} = 98.41$

于是，β_1、β_0的置信区间分别为:

(0.6345, 0.9195)

(−433.32, 226.98)

由于置信区间一定程度地给出了样本参数估计值与总体参数真值的“接近”程度，因此置信区间越小越好。

要缩小置信区间，需

（1）增大样本容量n，因为在同样的置信水平下，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；

（2）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，模型拟合优度越高，残差平方和应越小。

§2.4 一元线性回归分析的应用：预测问题

一、$\hat{y}_0$是$\mathbf{Y_0}$的一个无偏估计

二、个值预测值的置信区间

对于一元线性回归模型

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

给定样本以外的解释变量的观测值X_0，可以得到被解释变量的预测值$\hat{y}_0$，可以此作为个别值Y_0的一个近似估计。

一、$\hat{y}_0$是Y_0的一个无偏估计

对总体回归模型$Y=\beta_0+\beta_1X+\mu$，当$X=X_0$时

$$Y_0 = \beta_0 + \beta_1 X_0 + \mu$$

于是

$$E(Y_0) = E(\beta_0 + \beta_1 X_0 + \mu) = \beta_0 + \beta_1 X_0 + E(\mu) = \beta_0 + \beta_1 X_0$$

而 通过样本回归函数$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$，求得拟合值

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

的期望为

$$E(\hat{Y}_0) = E(\hat{\beta}_0 + \hat{\beta}_1 X_0) = E(\hat{\beta}_0) + X_0 E(\hat{\beta}_1) = \beta_0 + \beta_1 X_0$$

$\hat{Y}_0$是个值Y_0的无偏估计。

二、个值预测值的置信区间

由于 $$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum x_i^2}) \qquad \hat{\beta}_0 \sim N(\beta_0, \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2)$$

于是 $$E(\hat{Y}_0) = E(\hat{\beta}_0) + X_0 E(\hat{\beta}_1) = \beta_0 + \beta_1 X_0$$

$$Var(\hat{Y}_0) = Var(\hat{\beta}_0) + 2X_0 Cov(\hat{\beta}_0, \hat{\beta}_1) + X_0^2 Var(\hat{\beta}_1)$$

可以证明 $$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \overline{X} / \sum x_i^2$$

因此

$$Var(\hat{Y}_0) = \frac{\sigma^2 \sum X_i^2}{n\sum x_i^2} - \frac{2X_0\overline{X}\sigma^2}{\sum x_i^2} + \frac{X_0^2\sigma^2}{\sum x_i^2}$$

$$= \frac{\sigma^2}{\sum x_i^2}\left(\frac{\sum X_i^2 - n\overline{X}^2}{n} + \overline{X}^2 - 2X_0\overline{X} + X_0^2\right)$$

$$= \frac{\sigma^2}{\sum x_i^2}\left(\frac{\sum x_i^2}{n} + (X_0 - \overline{X})^2\right) = \sigma^2\left(\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i^2}\right)$$

故

$$\hat{Y}_0 \sim N\left(\beta_0 + \beta_1 X_0, \sigma^2\left(\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i^2}\right)\right)$$

由 $Y_0=\beta_0+\beta_1X_0+\mu$ 知：

$$Y_0 \sim N(\beta_0+\beta_1X_0,\sigma^2)$$

于是 $$\hat{Y}_0-Y_0 \sim N(0,\sigma^2(1+\frac{1}{n}+\frac{(X_0-\overline{X})^2}{\sum x_i^2}))$$

将未知的σ^2代以它的无偏估计量 $\hat{\sigma}^2$，可构造**t统计量**

$$t=\frac{\hat{Y}_0-Y_0}{S_{\hat{Y}_0-Y_0}} \sim t(n-2)$$

式中：$$S_{\hat{Y}_0-Y_0}=\sqrt{\hat{\sigma}^2(1+\frac{1}{n}+\frac{(X_0-\overline{X})^2}{\sum x_i^2})}$$

从而在1-α的置信度下，**Y_0的置信区间**为

$$\hat{Y}_0-t_{\frac{\alpha}{2}}\times S_{\hat{Y}_0-Y_0}<Y_0<\hat{Y}_0+t_{\frac{\alpha}{2}}\times S_{\hat{Y}_0-Y_0}$$

对于Y的个体值的预测区间（置信区间）：

（1）样本容量n越大，预测精度越高，反之预测精度越低；

（2）样本容量一定时，置信带的宽度当在X均值处最小，其附近进行预测（插值预测）精度越大；X越远离其均值，置信带越宽，预测可信度下降。

例1 已知回归模型 $E=\alpha+\beta N+\mu$，式中 E 为某类公司一名新员工的起始薪金(元)，N 为所受教育水平（年）。随机扰动项 μ 的分布未知，其他所有假设都满足。

（1）从直观及经济角度解释 α 和 β。

（2）OLS 估计量 $\hat{\alpha}$ 和 $\hat{\beta}$ 满足线性性、无偏性及有效性吗？简单陈述理由。

（3）对参数的假设检验还能进行吗？简单陈述理由。

例2 在例1中，如果被解释变量新员工起始薪金的计量单位由元改为100元，估计的截距项与斜率项有无变化？如果解释变量所受教育水平的度量单位由年改为月，估计的截距项与斜率项有无变化？

例3　考虑如下双变量 PRF 表达式：

模型 I：$Y_i = \beta_1 + \beta_2 X_i + u_i$

模型 II：$Y_i = \alpha_1 + \alpha_2 (X_i - \overline{X}) + u_i$

a. 求 β_1 和 α_1 的估计量。它们是否相同？它们的方差是否相同？

b. 求 β_2 和 α_2 的估计量，它们是否相同？它们的方差是否相同？

c. 如果模型 II 比模型 I 好，好在哪里？

例4　令 r_1 为 n 对 (X_i, Y_i) 值的相关系数，而 r_2 为 n 对 (aX_i+b, cY_i+d) 值的相关系数，其中 a，b，c 和 d 为常数。证明 $r_1 = r_2$，从而证实相关系数对度量单位和原点的改变保持不变的性质。

提示：应用方程 (3.5.13) 中所给的 r 定义。

注：运算 aX_i，X_i+b 和 aX_i+b 分别叫做尺度变换、原点变换和尺度与原点同时变换。

例5　假设在回归 $Y_i=\beta_1+\beta_2 X_i+u_i$ 中，我们将每个 X 值都乘以 2，这会不会改变 Y 的残差及拟合值？为什么？如果我们给每个 X 值都加上一个常数 2，又会怎样？

例6　参考方程（3.7.3）中所给出的手机需求回归。

a. 在5%的显著水平上，截距系数估计值显著吗？你进行检验的虚拟假设是什么？

b. 在5%的显著水平上，斜率系数估计值显著吗？

c. 构造真实斜率系数的95%置信区间。

$$\hat{Y}_i = 14.477\ 3 + 0.002\ 2X_i \qquad (3.7.3)$$

$$se(\hat{\beta}_1) = 6.152\ 3 \qquad se(\hat{\beta}_2) = 0.000\ 32$$

$$r^2 = 0.602\ 3$$

样本个数为34