第二章 经典单方程计量经济学模型：一元线性回归模型

The Classical Single Equation Econometric Model：Simple Linear Regression Model

本章内容

- 回归分析概述
- 一元线性回归模型的基本假设**
- 一元线性回归模型的参数估计***
- 一元线性回归模型的检验*
- 一元线性回归模型的预测
- 实例及时间序列问题

§2.1 回归分析概述
（Regression Analysis）

一、变量间的关系及回归分析的基本概念

二、总体回归函数

三*、随机扰动项

四、样本回归函数

一、变量间的关系及回归分析的基本概念

1、变量间的关系

- **确定性关系或函数关系：** 研究的是确定性现象/非随机变量间的关系。

$$\text{圆面积} = f(\pi, \text{半径}) = \pi \cdot \text{半径}^2$$

- **统计依赖或相关关系：** 研究的是非确定性现象/随机变量间的关系。

$$\text{农作物产量} = f(\text{气温}, \text{降雨量}, \text{阳光}, \text{施肥量})$$

- 对变量间统计依赖关系的考察，主要是通过相关分析(correlation analysis)或回归分析(regression analysis)来完成的。

- 相关分析适用于所有统计关系。
  - 相关系数(correlation coefficient)
  - 正相关(positive correlation)
  - 负相关(negative correlation)
  - 不相关(non-correlation)

- 回归分析，仅对存在因果关系而言。

- **注意：**
  - 不存在线性相关，并不意味着不相关。
  - 存在相关关系，并不一定存在因果关系。

  - 相关分析对称地对待任何（两个）变量，两个变量都被看作是随机的。
  - 回归分析对变量的处理方法存在不对称性，即区分应变量（被解释变量y，果）和自变量（解释变量X，因），前者是随机变量，后者不一定是。

2、回归分析的基本概念

- 回归分析(regression analysis)，是研究一个变量关于另一个（些）变量的具体依赖关系的计算方法和理论。
- 其目的在于通过后者的已知或设定值，去估计和（或）预测前者的（总体）均值。
- 两类变量；
  - 被解释变量（Explained Variable）或应变量（Dependent Variable）。
  - 解释变量（Explanatory Variable）或自变量（Independent Variable）。

- 关于变量的术语
  - **Explained Variable ~ Explanatory Variable**
  - **Dependent Variable ~ Independent Variable**
  - **Endogenous Variable ~ Exogenous Variable**
  - **Response Variable ~ Control Variable**
  - **Predicted Variable ~ Predictor Variable**
  - **Regressand (y) ~ Regressor (X)**

- 回归分析构成计量经济学的方法论基础，其主要内容包括：
  - 根据样本观察值对经济计量模型参数进行估计，求得回归方程；
  - 对回归方程、参数估计值进行显著性检验；
  - 利用回归方程进行分析、评价及预测。

二、总体回归函数
Population Regression Function, PRF

1、条件均值（conditional mean）

- **例2.1.1：一个假想的社区有99户家庭组成，欲研究该社区每月家庭消费支出Y与每月家庭可支配收入X的关系。即如果知道了家庭的月收入，能否预测该社区家庭的平均月消费支出水平。**
- **为达到此目的，将该99户家庭划分为组内收入差不多的10组，以分析每一收入组的家庭消费支出。**

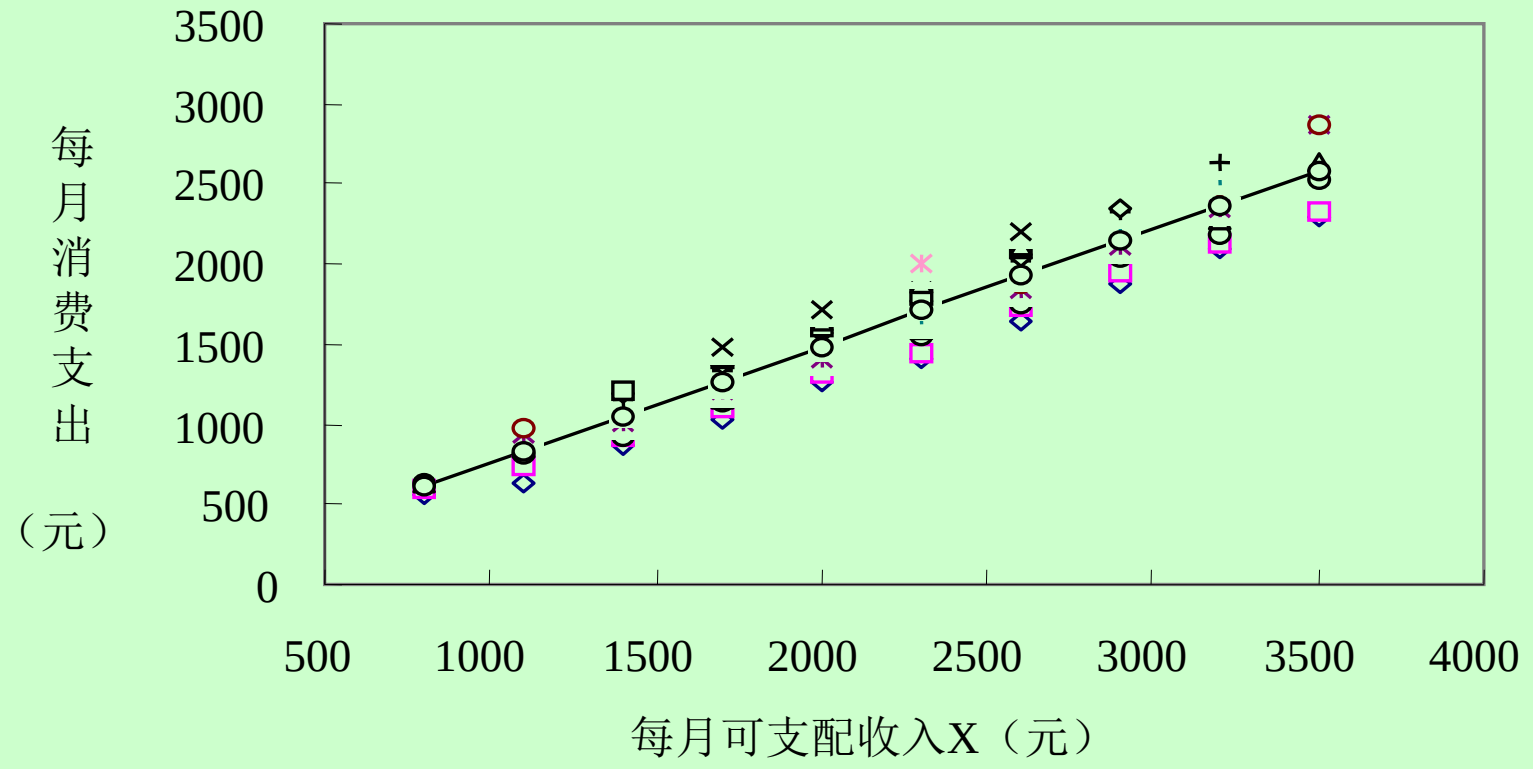
表 2.1.1　某社区家庭每月收入与消费支出统计表

	每月家庭可支配收入X（元）									
	800	1100	1400	1700	2000	2300	2600	2900	3200	3500
每月家庭消费支出Y（元）	561	638	869	1023	1254	1408	1650	1969	2090	2299
	594	748	913	1100	1309	1452	1738	1991	2134	2321
	627	814	924	1144	1364	1551	1749	2046	2178	2530
	638	847	979	1155	1397	1595	1804	2068	2266	2629
		935	1012	1210	1408	1650	1848	2101	2354	2860
		968	1045	1243	1474	1672	1881	2189	2486	2871
			1078	1254	1496	1683	1925	2233	2552	
			1122	1298	1496	1716	1969	2244	2585	
			1155	1331	1562	1749	2013	2299	2640	
			1188	1364	1573	1771	2035	2310		
			1210	1408	1606	1804	2101			
				1430	1650	1870	2112			
				1485	1716	1947	2200			
						2002				
共计	2420	4950	11495	16445	19305	23870	25025	21450	21285	15510

- 由于不确定因素的影响，对同一收入水平X，不同家庭的消费支出不完全相同。

- 但由于调查的完备性，给定收入水平X的消费支出Y的分布是确定的，即以X的给定值为条件的Y的条件分布（Conditional distribution）是已知的，例如：P(Y=561|X=800）=1/4。
- 因此，给定收入X的值X_i，可得消费支出Y的条件均值（conditional mean）或条件期望（conditional expectation）：E(Y|X=Xi)。
- 该例中：E(Y | X=800) = 605

- 描出散点图发现：随着收入的增加，消费“平均地说”也在增加，且Y的条件均值均落在一根正斜率的直线上。



2、总体回归函数

- 在给定解释变量X_i条件下被解释变量Y_i的期望轨迹称为总体回归线（population regression line），或更一般地称为总体回归曲线（population regression curve）。
- 相应的函数称为（双变量）总体回归函数（population regression function, PRF）。

$$E(Y \mid X_i) = f(X_i)$$

- 含义：回归函数（PRF），说明被解释变量Y的平均状态（总体条件期望）、随解释变量X变化的规律。
- 函数形式：可是线性或非线性的。
- 例2.1.1中，将居民消费支出看成是其可支配收入的线性函数时：

$$E(Y \mid X_i) = \beta_0 + \beta_1 X_i$$

为线性函数。

其中，β_0，β_1是未知参数，称为回归系数（regression coefficients）。

三、随机扰动项*
Stochastic Disturbance

- 总体回归函数说明在给定的收入水平X_i下，该社区家庭平均的消费支出水平。
- 但对某一个别的具体家庭，其消费支出可能与该平均水平有偏差。

- 通常称观察值偏离它的期望值的“离差”（**deviation**），是一个不可观测的随机变量，又称为随机干扰项（**stochastic disturbance**）或随机误差项（**stochastic error**）。

$$\mu_i = Y_i - E(Y \mid X_i)$$

- 例2.1.1中，给定收入水平X_i，个别家庭的支出可表示为两部分之和：
  - 该收入水平下所有家庭的平均消费支出$E(Y|X_i)$，称为系统性（systematic）或确定性（deterministic)部分；
  - 其他随机或非确定性（nonsystematic)部分μ_i。

$$Y_i = E(Y \mid X_i) + \mu_i = \beta_0 + \beta_1 X_i + \mu_i$$

• 称为总体回归函数（PRF）的随机设定形式。表明被解释变量除了受解释变量的系统性影响外，还受其他因素的随机性影响。

由于方程中引入了随机项，成为计量经济学模型，因此也称为总体回归模型(PRM)。

- 随机误差项主要包括下列因素：
  - 在解释变量中被忽略因素的影响；
    - ✓影响不太显著的细小因素
    - ✓未知/遗漏的影响因素
    - ✓无法获得数据的因素
  - 变量观测值的观测误差的影响；
  - 模型关系的设定误差的影响；
  - 其它随机因素的影响。

- 关于随机误差项的说明：
  - 将随机项，可区分为“源生的随机扰动”和“衍生的随机误差”。
  - “源生的随机扰动”仅包含“无数对被解释变量影响不显著的微小因素”的影响，服从极限法则（大数定律和中心极限定理），满足基本假设。
  - “衍生的随机误差”包含上述所有内容，并不一定服从极限法则，不一定满足基本假设。

四、样本回归函数
Sample Regression Function, SRF

1、样本回归函数

- 问题：能否从一次抽样中、获得总体的近似信息？如果可以，如何从抽样中获得总体的近似信息？

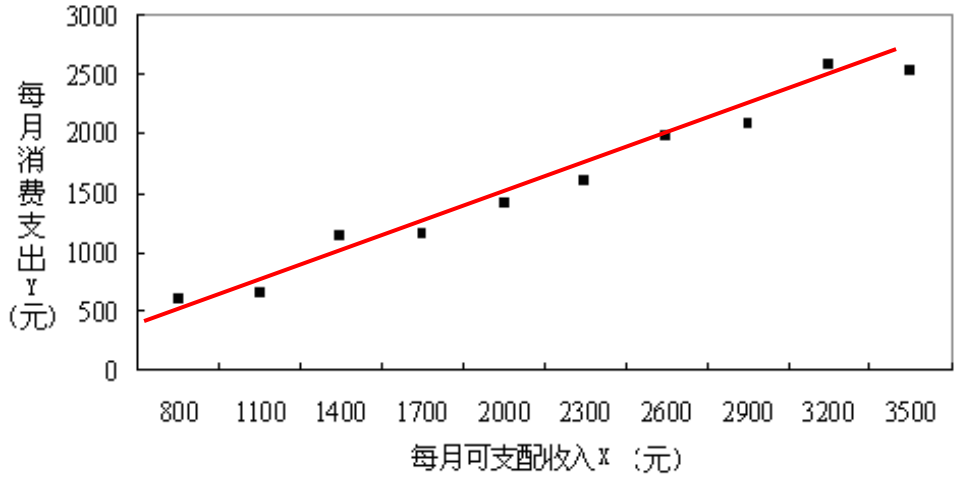
- 在例2.1.1的总体中有如下一个样本，能否从该样本估计总体回归函数？

表 2.1.3　家庭消费支出与可支配收入的一个随机样本

X	800	1100	1400	1700	2000	2300	2600	2900	3200	3500
Y	594	638	1122	1155	1408	1595	1969	2078	2585	2530

回答：能

- 该样本的散点图（scatter diagram)：



- 画一条直线以尽好地拟合该散点图，由于样本取自总体，可以该直线近似地代表总体回归线。

  该直线，称为样本回归线（sample regression lines）。

- 样本回归线的函数形式为：

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

称为样本回归函数（sample regression function，SRF）。

- **注意：**这里将**样本回归线**看成**总体回归线**的近似替代

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

$$\Downarrow$$

$$\begin{aligned} Y_i &= E(Y \mid X_i) + \mu_i \\ &= \beta_0 + \beta_1 X_i + \mu_i \end{aligned}$$

则 $\hat{Y}_i$ 为 $E(Y \mid X_i)$ 的估计量；

$\hat{\beta}_i$ 为 β_i 的估计量，$i = (0,1)$

2、样本回归模型

- **样本回归函数的随机形式：**

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

式中，e_i 称为**（样本）残差**（或**剩余**）**项**（residual），代表了其他影响Y_i 的随机因素的集合，可看成是μ_i 的估计量$\hat{\mu}_i$ 。

•由于方程中引入了随机项，成为计量经济模型，因此也称为样本回归模型（sample regression model）。

- **回归分析的主要目的：**

根据样本回归函数SRF，估计总体回归函数PRF。

$$Y_i = \hat{Y}_i + e_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \rightarrow Y_i = E(Y \mid X_i) + \mu_i = \beta_0 + \beta_1 X_i + \mu_i$$

这就要求：

设计一“方法”构造SRF，以使SRF尽可能“接近”PRF，或者说使$\hat{\beta}_i (i = 0,1)$尽可能接近$\beta_i (i = 0,1)$。

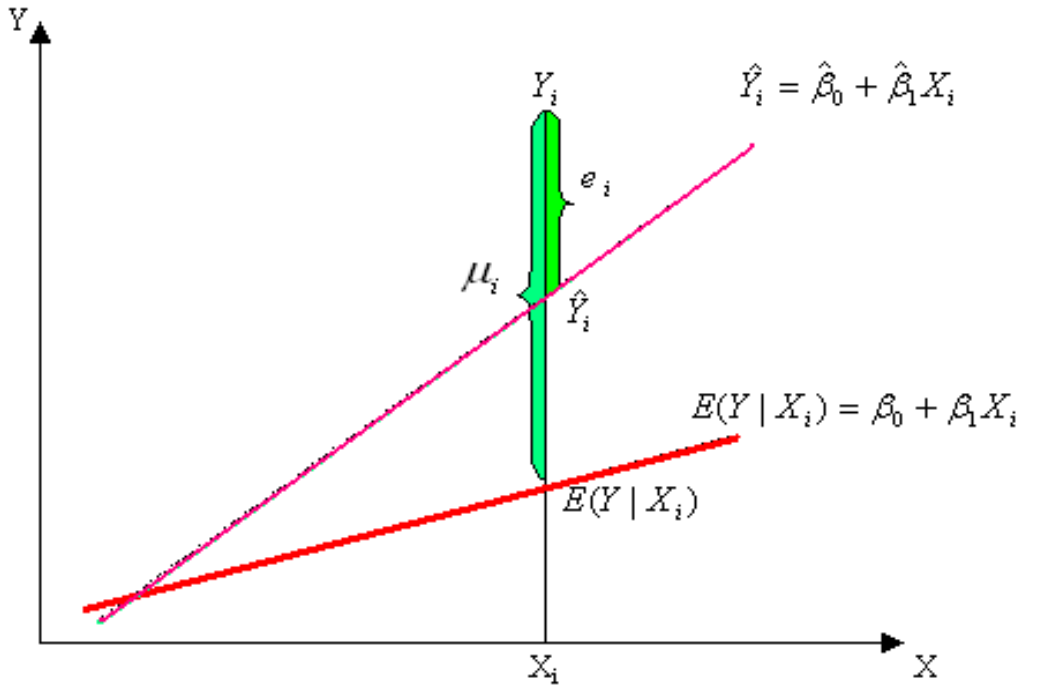


图 2.1.3　总体回归线与样本回归线的基本关系

§2.2** 一元线性回归模型的基本假设 (Assumptions of Simple Linear Regression Model)

一、关于模型设定的假设

二*、关于解释变量X的假设

三**、关于随机项u的假设

说 明

- 为保证参数估计量具有良好的性质，通常对模型提出若干基本假设。
- 实际上这些假设与所采用的估计方法紧密相关。
- 以下假设，主要是针对采用**普通最小二乘法（Ordinary Least Squares, OLS）**估计而提出的；所以，有些教科书称之为**“The Assumption Underlying the Method of Least Squares”**。
- 不同教科书、关于基本假设的陈述略有不同，下面进行了重新归纳。

1、关于模型关系的1个假设

- 模型设定正确假设

  The regression model is correctly specified.

  比如，线性回归假设

  The regression model is linear in the parameters。

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

注意：“**linear in the parameters**”的含义是什么？

2*、关于 解释变量X 的5项假设

- 确定性假设

X values are fixed in repeated sampling. More technically, X is assumed to be non-stochastic.

注意：“**in repeated sampling**”的含义是什么？

- 与随机扰动项不相关假设

The covariances between X_i and μ_i are zero.

$$\text{cov}(X_i, \mu_i) = 0, i = 1, 2, \cdots, n$$
$$E(X_i \mu_i) = 0, i = 1, 2, \cdots, n$$

由确定性假设可以推断。

- 观测值变化假设。**X values in a given sample must not all be the same.**

- 无完全共线性假设。**There is no perfect multicollinearity among the explanatory variables.**

  适用于多元线性回归模型。

时间序列数据作样本时，适用

- 样本方差有界假设。随着样本容量的无限增加，解释变量X的样本方差、趋于一有限常数。

$$\sum (X_i - \overline{X})^2 / n \to Q, \qquad n \to \infty$$

3***、关于 随机项u 的3项假设

- **0均值假设**

The conditional mean value of u_i is zero.

$$E(u_i \mid X_i) = 0, i = 1, 2, \cdots, n$$

由模型设定正确假设推断。

- 同方差假设

The conditional variances of u_i are identical. (Homoscedasticity)

$$Var(u_i \mid X_i) = \sigma^2, i = 1, 2, \cdots, n$$

是否满足需要检验。

- **序列不相关假设**

The correlation between any two u_i and u_j is zero.

$$Cov(u_i, u_j \mid X_i, X_j) = 0, \ \ i, j = 1, 2, \cdots, n, i \neq j$$

是否满足需要检验。

4、随机项的正态性假设

- 采用**OLS**进行参数估计时，不需要正态性假设；

  但在利用参数估计量进行**统计推断**时，需要假设随机项的概率分布。

- 一般假设随机项***u***服从正态分布，可利用中心极限定理（**central limit theorem, CLT**）进行证明。

- 正态性假设

The *u*'s follow the normal distribution.

$$\mu_i \sim N(0,\sigma^2) \rightarrow \mu_i \sim \mathbf{NID}(0,\sigma^2)$$

5、CLRM 和 CNLRM

- 以上假设（正态性假设除外），也称为线性回归模型的“经典假设”或“高斯（Gauss）假设”；

  满足该假设的线性回归模型，也称为经典线性回归模型（**Classical Linear Regression Model, CLRM**）。

- 同时满足正态性假设的线性回归模型，称为经典正态线性回归模型（**Classical Normal Linear Regression Model, CNLRM**）。

§2.3 一元线性回归模型的参数估计*** (Estimation of Simple Linear Regression Model)

一*、 参数的普通最小二乘估计（OLS）

二*、 参数估计的最大或然法(ML)

三**、 最小二乘估计量的性质

四*、 参数估计量的概率分布及随机干扰项方差的估计

一、参数的普通最小二乘估计（OLS）

1、最小二乘原理

- 根据被解释变量的“所有观测值与估计值之差的平方和最小”的原则，求得参数估计量。

$$MinQ = \sum_{1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

- 为什么取平方和？

2、正规方程组

$$\begin{cases} \dfrac{\partial Q}{\partial \hat{\beta}_0} = 0 \\ \dfrac{\partial Q}{\partial \hat{\beta}_1} = 0 \end{cases} \longrightarrow \begin{cases} \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = 0 \end{cases} \quad \text{即：} \quad \begin{cases} \sum e_i = 0 \\ \sum e_i X_i = 0 \end{cases}$$

- 该关于参数估计量的线性方程组，称为正规方程组（**normal equations**）。

3、参数估计量

- 求解正规方程组得到结构参数的普通最小二乘估计量（ordinary least squares estimators）及其离差形式：

$$\begin{cases}\hat{\beta}_0 = \dfrac{\Sigma X_i^2 \Sigma Y_i - \Sigma X_i \Sigma Y_i X_i}{n\Sigma X_i^2 - (\Sigma X_i)^2} \\ \hat{\beta}_1 = \dfrac{n\Sigma Y_i X_i - \Sigma Y_i \Sigma X_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}\end{cases}$$

对任意x_i，均有：

$$\sum x_i = 0$$

$$\begin{cases}\hat{\beta}_1 = \dfrac{\Sigma x_i y_i}{\Sigma x_i^2} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}\end{cases}$$

- 分布参数σ^2的OLS估计量：

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

4、“估计量”（estimator）和“估计值”（estimate）的区别

- 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。

- 如果把上式看成是“参数估计的一个表达式/一个计算规则”，那么它是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”。

二、参数估计的最大似然法(ML)

1、最大似然法

- 最大似然法(Maximum Likelihood, ML)，也称最大或然法，是不同于最小二乘法的另一种参数估计方法，是从最大或然原理出发、发展起来的其它一大类估计方法。

- 基本原理：当从模型总体随机抽取n组样本观测值后，最合理的参数估计量、应使得从模型中“抽取该n组样本观测值的联合概率”最大。

- ML必须已知随机项的分布。

2、估计步骤

$$Y_i \sim N(\hat{\beta}_0 + \hat{\beta}_1 X_i, \sigma^2)$$

Y_i的分布

$$P(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}$$

Y_i的概率函数

$$L(\hat{\beta}_0, \hat{\beta}_1, \sigma^2) = P(Y_1, Y_2, \cdots, Y_n)$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{-\frac{1}{2\sigma^2}\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}$$

Y的“所有样本观测值的联合概率”——**似然函数**

$$L^* = \ln(L)$$

$$= -n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

对数似然函数

$$\begin{cases} \dfrac{\partial}{\partial \hat{\beta}_0}\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0 \\ \dfrac{\partial}{\partial \hat{\beta}_1}\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0 \end{cases}$$

对数似然函数极大化的一阶条件

$$\begin{cases} \hat{\beta}_0 = \dfrac{\Sigma X_i^2 \Sigma Y_i - \Sigma X_i \Sigma Y_i X_i}{n\Sigma X_i^2 - (\Sigma X_i)^2} \\ \hat{\beta}_1 = \dfrac{n\Sigma Y_i X_i - \Sigma Y_i \Sigma X_i}{n\Sigma X_i^2 - (\Sigma X_i)^2} \end{cases}$$

结构参数的ML估计量

3、讨论

- 在满足一系列基本假设的情况下：模型结构参数的最大似然估计量与普通最小二乘估计量，是相同的。

- 但是，分布参数 σ^2 的估计结果不同。

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

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

三、最小二乘估计量的性质

1、概述

- 当模型参数估计出后，需考虑参数估计值的精度，即能否代表总体参数的真值，或者说需考察参数估计量的统计性质。

- 准则：
  - 线性性(linear)，即它是否是另一随机变量的线性函数；
  - 无偏性(unbiased)，即它的均值或期望值是否等于总体的真实值；
  - 有效性(efficient)，即它是否在所有线性无偏估计量中具有最小方差。

- 这三个准则，也称作估计量的小样本性质。拥有这类性质的估计量称为最佳线性无偏估计量（best liner unbiased estimator, BLUE）。

- 当不满足小样本性质时，需进一步考察估计量的大样本或渐近性质（**asymptotic properties**）：
  - 渐近无偏性，即样本容量趋于无穷大时，是否它的均值序列趋于总体真值；
  - 一致性，即样本容量趋于无穷大时，估计量是否依概率收敛于总体真值；
  - 渐近有效性，即样本容量趋于无穷大时，是否它在所有的一致估计量中具有最小的渐近方差。

2、高斯—马尔可夫定理(Gauss-Markov theorem)

- 在给定经典线性回归的假定下，最小二乘估计量是具有最小方差的线性无偏**BLUE**估计量。

- 下面分别对最小二乘估计量的线性性、无偏性和有效性进行证明，作为不熟悉的同学的自学内容。

1、线性性，即估计量$\hat{\beta}_0$ 、 $\hat{\beta}_1$ 是Y_i 的线性组合。

证：
$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i (Y_i - \bar{Y})}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum x_i^2} + \frac{\bar{Y}\sum x_i}{\sum x_i^2}$$

令 $k_i = \frac{x_i}{\sum x_i^2}$，因 $\sum x_i = \sum (X_i - \bar{X}) = 0$，故有

$$\hat{\beta}_1 = \sum \frac{x_i}{\sum x_i^2} Y_i = \sum k_i Y_i$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{1}{n}\sum Y_i - \sum k_i Y_i \bar{X} = \sum (\frac{1}{n} - \bar{X}k_i) Y_i = \sum w_i Y_i$$

2、无偏性，即估计量$\hat{\beta}_0$ 、$\hat{\beta}_1$ 的均值（期望）等于总体回归参数真值β_0与β_1

证： $\hat{\beta}_1 = \sum k_i Y_i = \sum k_i(\beta_0 + \beta_1 X_i + \mu_i) = \beta_0 \sum k_i + \beta_1 \sum k_i X_i + \sum k_i \mu_i$

易知 $\sum k_i = \dfrac{\sum x_i}{\sum x_i^2} = 0$ $\sum k_i X_i = 1$

故 $\hat{\beta}_1 = \beta_1 + \sum k_i u_i$

$$E(\hat{\beta}_1) = E(\beta_1 + \sum k_i \mu_i) = \beta_1 + \sum k_i E(\mu_i) = \beta_1$$

类似地，容易得出

$$E(\hat{\beta}_0) = E(\beta_0 + \sum w_i \mu_i) = E(\beta_0) + \sum w_i E(\mu_i) = \beta_0$$

3、有效性（最小方差性），即在所有线性无偏估计量中，最小二乘估计量$\hat{\beta}_0$ 、$\hat{\beta}_1$ 具有最小方差。

(1)先求$\hat{\beta}_0$ 与$\hat{\beta}_1$ 的方差

$$\mathrm{var}(\hat{\beta}_1)=\mathrm{var}(\sum k_iY_i)=\sum k_i^2\,\mathrm{var}(\beta_0+\beta_1X_i+\mu_i)=\sum k_i^2\,\mathrm{var}(\mu_i)$$

$$=\sum\left(\frac{x_i}{\sum x_i^2}\right)^2\sigma^2=\frac{\sigma^2}{\sum x_i^2}$$

$$\mathrm{var}(\hat{\beta}_0)=\mathrm{var}(\sum w_iY_i)=\sum w_i^2\,\mathrm{var}(\beta_0+\beta_1X_i+\mu_i)=\sum(1/n-\bar{X}k_i)^2\sigma^2$$

$$=\sum\left[\left(\frac{1}{n}\right)^2-2\frac{1}{n}\bar{X}k_i+\bar{X}^2k_i^2\right]\sigma^2=\left(\frac{1}{n}-\frac{2}{n}\bar{X}\sum k_i+\bar{X}^2\sum\left(\frac{x_i}{\sum x_i^2}\right)^2\right)\sigma^2$$

$$=\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum x_i^2}\right)\sigma^2=\frac{\sum x_i^2+n\bar{X}^2}{n\sum x_i^2}\sigma^2=\frac{\sum X_i^2}{n\sum x_i^2}\sigma^2$$

（2）证明最小方差性

假设$\hat{\beta}_1^*$是其他估计方法得到的关于β_1的线性无偏估计量：

$$\hat{\beta}_1^* = \sum c_i Y_i$$

其中，$c_i=k_i+d_i$，d_i为不全为零的常数

则容易证明 $\mathrm{var}(\hat{\beta}_1^*) \geq \mathrm{var}(\hat{\beta}_1)$

因为由无偏性可得

$$\sum c_i = 0,\ \sum c_i X_i = 1 \Rightarrow \sum k_i d_i = \sum k_i (c_i - k_i) = \sum \frac{x_i c_i}{\sum x_i^2} - \frac{1}{\sum x_i^2} = 0$$

由此可得：$$Var(\hat{\beta}_1^*) = \sum (k_i + d_i)^2 \sigma^2 = \sum k_i^2 \sigma^2 + \sum d_i^2 \sigma^2$$

同理，可证明β_0的最小二乘估计量$\hat{\beta}_0$具有最的小方差

由于最小二乘估计量拥有一个“好”的估计量所应具备的小样本特性，它自然也拥有**大样本特性**。

如考察$\hat{\beta}_1$的一致性

$$P\lim(\hat{\beta}_1) = P\lim(\beta_1 + \sum k_i \mu_i) = P\lim(\beta_1) + P\lim(\frac{\sum x_i \mu_i}{\sum x_i^2})$$

$$= \beta_1 + \frac{P\lim(\sum x_i \mu_i / n)}{P\lim(\sum x_i^2 / n)}$$

$$= \beta_1 + \frac{Cov(X, \mu)}{Q} = \beta_1 + \frac{0}{Q} = \beta_1$$

四、参数估计量的概率分布及随机干扰项方差的估计

1、参数估计量的概率分布

普通最小二乘估计量 $\hat{\beta}_0$、$\hat{\beta}_1$ 分别是 Y_i 的线性组合，因此，$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的概率分布取决于 Y 的分布特征

在 μ 是正态分布的假设下，Y 是正态分布，则 $\hat{\beta}_0$、$\hat{\beta}_1$ 也服从正态分布，因此

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

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

2、随机误差项μ的总体方差σ^2的估计

- σ^2又称为总体方差。
- 由于随机项u_i不可观测，只能从“u_i的估计”——“残差e_i”出发，对总体方差进行估计。
- 可以证明，σ^2的最小二乘估计量为：

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

它是关于σ^2的无偏估计量。

Proof of $\hat{\sigma}^2$: $\hat{\beta}_0 + \hat{\beta}_1 \bar{x} + \bar{e} = \bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$

then: $\hat{u}_i = y_i - \hat{y}_i = y_i - \bar{y} - (\hat{y}_i - \bar{y}) = (u_i - \bar{u}) - \left(\hat{\beta}_1 - \beta_1\right)(x_i - \bar{x})$

$$\sum \hat{u}_i^2 = \sum (u_i - \bar{u})^2 - 2\sum (u_i - \bar{u})\left(\hat{\beta}_1 - \beta_1\right)(x_i - \bar{x}) + \sum \left(\hat{\beta}_1 - \beta_1\right)^2 (x_i - \bar{x})^2$$

take expectation: $E(\sum \hat{u}_i^2) = (n-1)\sigma^2 - 2\sigma^2 + \sigma^2 = (n-2)\sigma^2$

Thereinto: 1) $E\sum \left(\hat{\beta}_1 - \beta_1\right)^2 (x_i - \bar{x})^2$

$$= \sum E\left(\hat{\beta}_1 - \beta_1\right)^2 \cdot (x_i - \bar{x})^2$$

$$= Cov(\hat{\beta}_1) \cdot \sum (x_i - \bar{x})^2 = \frac{\sigma^2}{s_x^2} \cdot s_x^2 = \sigma^2$$

2) $E\sum (u_i - \bar{u})^2 = E\sum (u_i^2 - 2u_i\bar{u} + \bar{u}^2)$

a. $E\sum u_i^2 = E(u_1^2 + \cdots + u_n^2) = n\sigma^2$

b. $E\sum u_i\bar{u} = E(u_1\bar{u} + \cdots + u_n\bar{u}) = n \cdot \frac{\sigma^2}{n} = \sigma^2$

c. $E\sum \bar{u}^2 = E[n \cdot (\frac{u_1 + \cdots u_n}{n})^2] = \sigma^2$

3)

$$\mathrm{E}[\sum(u_i-\bar{u})\left(\hat{\beta}_1-\beta_1\right)(x_i-\bar{x})]=\mathrm{E}[\sum(u_i-\bar{u})\left(\hat{\beta}_1-\beta_1\right)d_i]$$

$$=\mathrm{E}[\sum(u_i-\bar{u})\cdot\frac{\sum d_iu_i}{s_x^2}d_i]=\frac{1}{s_x^2}E[\sum u_id_i\cdot\sum d_iu_i-\bar{u}\sum d_iu_i\cdot\sum d_i]$$

$$=\frac{1}{s_x^2}E[\sum u_id_i\cdot\sum d_iu_i]=\frac{1}{s_x^2}E[\sum d_iu_i]^2$$

$$=\frac{1}{s_x^2}E[(d_1u_1)^2+\cdots+(d_nu_n)^2+2\sum_{i\neq j}d_iu_i\cdot d_ju_j]$$

$$=\frac{1}{s_x^2}(d_1^2Eu_1^2+\cdots+d_n^2Eu_n^2)$$

$$=\frac{1}{s_x^2}(d_1^2\cdots+d_n^2)\sigma^2$$

$$=\sigma^2$$

- 最大或然估计法中，将对数似然函数、对 σ^2 求一阶条件：

$$\frac{\partial}{\partial\sigma^2}L^* = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0$$

$$\hat{\sigma}^2 = \frac{1}{n}\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \frac{\sum e_i^2}{n}$$

- σ^2 的最大或然估计量不具无偏性，但却具有一致性。

§2.4 一元线性回归模型的统计检验

Statistical Test of Simple Linear Regression Model

一、拟合优度检验*

二、变量的显著性检验

三、参数的置信区间

说 明

- 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）应该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

Goodness of Fit, Coefficient of Determination

1、回答一个问题

- **拟合优度检验**：对样本回归线与样本观测值之间拟合程度的检验。

- 问题：

  采用普通最小二乘OLS估计方法，已经保证了模型最好地拟合了样本观测值，为什么还要检验拟合程度？

2、总离差平方和的分解

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

$$\hat{y}_i = (\hat{Y}_i - \overline{Y})$$

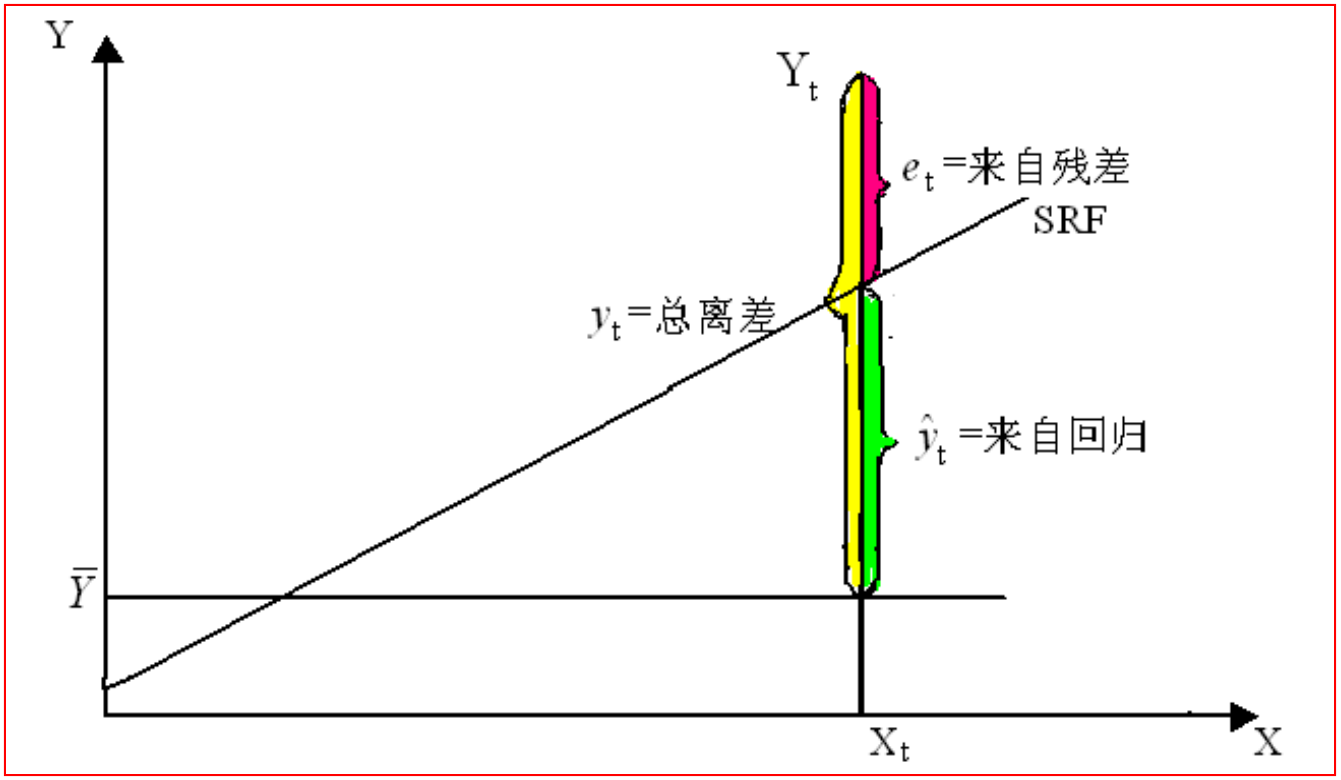
Y的第 *i* 个预测值
与样本均值的离差

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

总离差分解为两部分之和

回归直线不能解释的部分

由回归直线解释的部分



$\hat{y}_i = (\hat{Y}_i - \bar{Y})$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，则需考虑总离差的平方和：

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

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

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)。

在**给定样本**中，**TSS**不变，

如果实际观测点离样本回归线越近，则**ESS**在**TSS**中占的比重越大，因此

拟合优度：回归平方和ESS/Y的总离差TSS

3、可决系数R^2统计量

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

- 是一个非负的统计量。取值范围：[0，1]
- 越接近1，说明实际观测点离回归线越近，拟合优度越高。
- 随着抽样的不同而不同。为此，对可决系数的统计可靠性也应进行检验，这将在第3章中进行。

二、变量的显著性检验

Testing Significance of Variable

说　明

- 在一元线性模型中，变量的显著性检验就是判断X是否对Y具有显著的线性性影响。
- 变量的显著性检验所应用的方法，是数理统计学中的假设检验。
- 通过检验变量的参数真值是否为零来实现显著性检验。

1、假设检验（Hypothesis Testing）

- 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，在一次抽样中是不会发生的”这一原理。

2、变量的显著性检验—t检验

对总体参数提出假设：
H_0：$\beta_1=0$，
H_1：$\beta_1\neq 0$

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

用σ^2的估计量代替，构造t统计量

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

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

1）由样本计算t统计量值；

2）给定显著性水平(level of significance)α，查t分布表得临界值(critical value)$t_{\alpha/2}(n-2)$；

3）拒绝规则：

- 若 $|t| > t_{\alpha/2}(n-2)$，则以（1－α）的置信度（confidence coefficient）拒绝H_0，接受H_1；
- 若 $|t| \le t_{\alpha/2}(n-2)$，则以（1－α）的置信度不拒绝$H_0$。

- 自学教材p48例题，学会检验的全过程。

3、关于常数项的显著性检验

- T检验，同样可以进行。
- 一般不以t检验决定、常数项是否保留在模型中，而是从经济意义方面分析回归线是否应该通过原点。

三、参数的置信区间

Confidence Interval of Parameter

1、概念

- 回归分析希望通过样本得到的参数估计量，能够代替总体参数。
- 假设检验可通过一次抽样的结果检验总体参数可能的假设值的范围（例如是否为零），但它并没有指出在一次抽样中样本参数值到底离总体参数的真值有多“近”。
- 要判断样本参数的估计值在多大程度上“近似”地替代总体参数的真值，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；

1−α为置信系数（置信度）（confidence coefficient）；

α称为显著性水平；

置信区间的端点称为置信限（confidence limit）。

2、一元线性模型中β_i的置信区间

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

构造包含待估参数的已知分布统计量

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

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

$(1-\alpha)$的置信度下，β_i的置信区间是

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

• 在上述**收入-消费支出**例题中，如果给定α =0.01，查表得：

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

由于 $S_{\hat{\beta}_1}=0.019$ $S_{\hat{\beta}_0}=44.45$

于是，**β_1、β_0的置信区间分别为：**

（0.6056, 0.7344）

（-6.719, 291.52）

- 显然，在该例题中，我们对结果的正确陈述应该是：边际消费倾向β_1是以99%的置信度处于以0.670为中心的区间（0.6056,0.7344）中。

- 回答：
  - 边际消费倾向等于0.670的置信度是多少？
  - 边际消费倾向以100%的置信度处于什么区间？

- 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，需要
  - 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，还可使参数估计量的样本标准差减小；
  - 提高模型的拟合优度。因为参数估计量的样本标准差与残差平方和呈正比，模型拟合优度越高，残差平方和越小。

§2.5 一元线性回归分析的应用：预测问题

一、预测值条件均值或个值的一个无偏估计

二、总体条件均值与个值预测值的置信区间

说 明

- 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$\hat{Y}_0$，可以此作为其条件均值$E(Y|X=X_0)$或个别值Y_0的一个近似估计。

- 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。原因：
  - 参数估计量不确定；
  - 随机项的影响。

一、预测值是条件均值或个值的一个无偏估计

1、$\hat{Y}_0$是条件均值$E(Y|X=X_0)$的无偏估计

先对**总体回归函数**$E(Y|X=X_0)=\beta_0+\beta_1X$，$X=X_0$时

$$E(Y|X=X_0)=\beta_0+\beta_1X_0$$

通过样本回归函数 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$，求得的拟合值为

$$\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1X_0$$

$$E(\hat{Y}_0)=E(\hat{\beta}_0+\hat{\beta}_1X_0)=E(\hat{\beta}_0)+X_0E(\hat{\beta}_1)=\beta_0+\beta_1X_0$$

可见，$\hat{Y}_0$是条件均值$E(Y|X=X_0)$的无偏估计。

2、$\hat{Y}_0$是个值Y_0的无偏估计

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

$$Y_0=\beta_0+\beta_1X_0+\mu$$

$$E(Y_0)=E(\beta_0+\beta_1X_0+\mu)=\beta_0+\beta_1X_0+E(\mu)=\beta_0+\beta_1X_0$$

通过样本回归函数$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$，求得的拟合值为

$$\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1X_0$$

$$E(\hat{Y}_0)=E(\hat{\beta}_0+\hat{\beta}_1X_0)=E(\hat{\beta}_0)+X_0E(\hat{\beta}_1)=\beta_0+\beta_1X_0$$

可见，**$\hat{Y}_0$是个值Y_0的无偏估计。**

二、总体条件均值与个值预测值的置信区间：分布性质+分布参数

1、总体均值预测值的置信区间

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

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

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

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

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

参数估计量协方差的补充证明：

$$
\begin{aligned}
Cov(\hat{\beta}_0,\hat{\beta}_1) &= E(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)\\
&= E(\bar{Y}-\hat{\beta}_1\bar{X}-\beta_0)(\hat{\beta}_1-\beta_1)\\
&= E(\beta_0+\beta_1\bar{X}+\bar{u}-\hat{\beta}_1\bar{X}-\beta_0)(\hat{\beta}_1-\beta_1)\\
&= E[-\bar{X}(\hat{\beta}_1-\beta_1)+\bar{u}](\hat{\beta}_1-\beta_1)\\
&= -\bar{X}Var(\hat{\beta}_1)+E[\bar{u}\cdot\sum k_i u_i]\\
&= -\bar{X}Var(\hat{\beta}_1)+E[\frac{(u_1+\cdots u_n)}{n}\cdot(k_1u_1+\cdots k_nu_n)]\\
&= -\bar{X}Var(\hat{\beta}_1)+\frac{1}{n}E[(k_1u_1^2+\cdots k_nu_n^2)]\\
&= -\bar{X}Var(\hat{\beta}_1)+\frac{\sigma^2}{n}\sum k_i\\
&= -\bar{X}Var(\hat{\beta}_1)
\end{aligned}
$$

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

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

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

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

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

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

$$t=\frac{\hat{Y}_0-(\beta_0+\beta_1 X_0)}{S_{\hat{Y}_0}}\sim t(n-2)$$

于是，在1-α的置信度下，总体均值$E(Y|X_0)$的置信区间为

$$\hat{Y}_0-t_{\frac{\alpha}{2}}\times S_{\hat{Y}_0}<E(Y\mid X_0)<\hat{Y}_0+t_{\frac{\alpha}{2}}\times S_{\hat{Y}_0}$$

2、总体个值预测值的预测区间

$$Y_0 \sim N(\beta_0 + \beta_1 X_0, \sigma^2)$$

↓

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

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

从而在1-α的置信度下，Y_0的置信区间为

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

3、例题—收入-消费支出

- 样本回归函数为 $\hat{Y}_i = -142.4 + 0.670X_i$

则在 X_0=1000处，$\hat{Y}_0$ = 142.4+0.670×1000=812.4

$$Var(\hat{Y}_0) = 2734\left[\frac{1}{10} + \frac{(1000-2150)^2}{7425000}\right] = 760.4$$

$$S(\hat{Y}_0) = 27.6$$

- 因此，总体均值E(Y|X=1000)的95%的置信区间为：

（812.4－2.306×27.6，812.4+2.306×27.6）

（748.8, 875.9）

- 同样地，对于Y在X=1000的**个体值**，其95%的置信区间为：

（812.4 - 2.306×59.1，812.4 + 2.306×59.1）

(676.1, 948.7)

§2.6 实例及时间序列问题

说　明

- 本节列举了两个一元线性回归模型实例，完成了建立模型、估计参数、统计检验和预测的过程。
- 适合于课堂演示或者由学生在计算机上完成。

- 从理论上讲，经典线性回归模型理论是以“随机抽样的截面数据或者平稳的时间序列数据”为基础的；

  对非平稳时间序列数据，存在理论方法方面的障碍。如何处理？本书第8章将专门讨论。在2—7章中大量采用非平稳时间序列数据作为实例，暂时不考虑理论方法方面的障碍。