

第九章 相关与回归分析

- 第一节 相关系数
- 第二节 回归模型与回归方程
- 第三节 估计的回归方程
- 第四节 判定系数
- 第五节 F 检验
- 第六节 回归预测
- 第七节 残差分析
- 第八节 基于SPSS实现回归分析过程



引例

一家超市拟到一个新的地区发展业务，计划部主管需向公司董事会提交有关发展计划的分析报告，其中需要包括该地区居民每月在日用杂货项目上的支出信息，为此他以居民家庭为对象，就每户家庭的“每月日用杂货支出”、“月收入”、“家庭人口数”等变量搜集了一个容量为40的样本数据。并针对“每月日用杂货支出”的数据，估计出“平均每户每月日用杂货支出”为840.63元左右。

在分析报告中已经细化到针对不同收入水平居民的经营策略，而“平均每户每月日用杂货支出”的估计值，还只是一个笼统的估计数字，其中并未包含收入水平差异信息。因此，有必要与“月收入”关联起来，估计出不同收入水平居民的“平均每户每月日用杂货支出”。



40户家庭每月日用杂货支出与月收入

家庭序号	月支出（元）y	月收入（元）x	家庭序号	月支出（元）y	月收入（元）x
1	1148	8882	21	710	6837
2	489	4558	22	937	7263
3	1208	9053	23	1030	8009
4	1065	8094	24	1064	8392
5	1015	8414	25	662	4899
6	1125	8925	26	680	5581
7	1206	9862	27	675	5091
8	613	4856	28	720	5495
9	661	4899	29	913	6688
10	606	5304	30	1140	9329
11	541	5943	31	740	5304
12	1083	7242	32	540	5730
13	839	7540	33	693	5943
14	1090	8989	34	673	6156
15	1217	9138	35	918	6752
16	555	4388	36	1069	8264
17	458	4793	37	1265	9649
18	647	4856	38	1145	9883
19	592	5346	39	549	5133
20	676	6603	40	668	5304

第一节 相关系数

一、散点图与相关关系

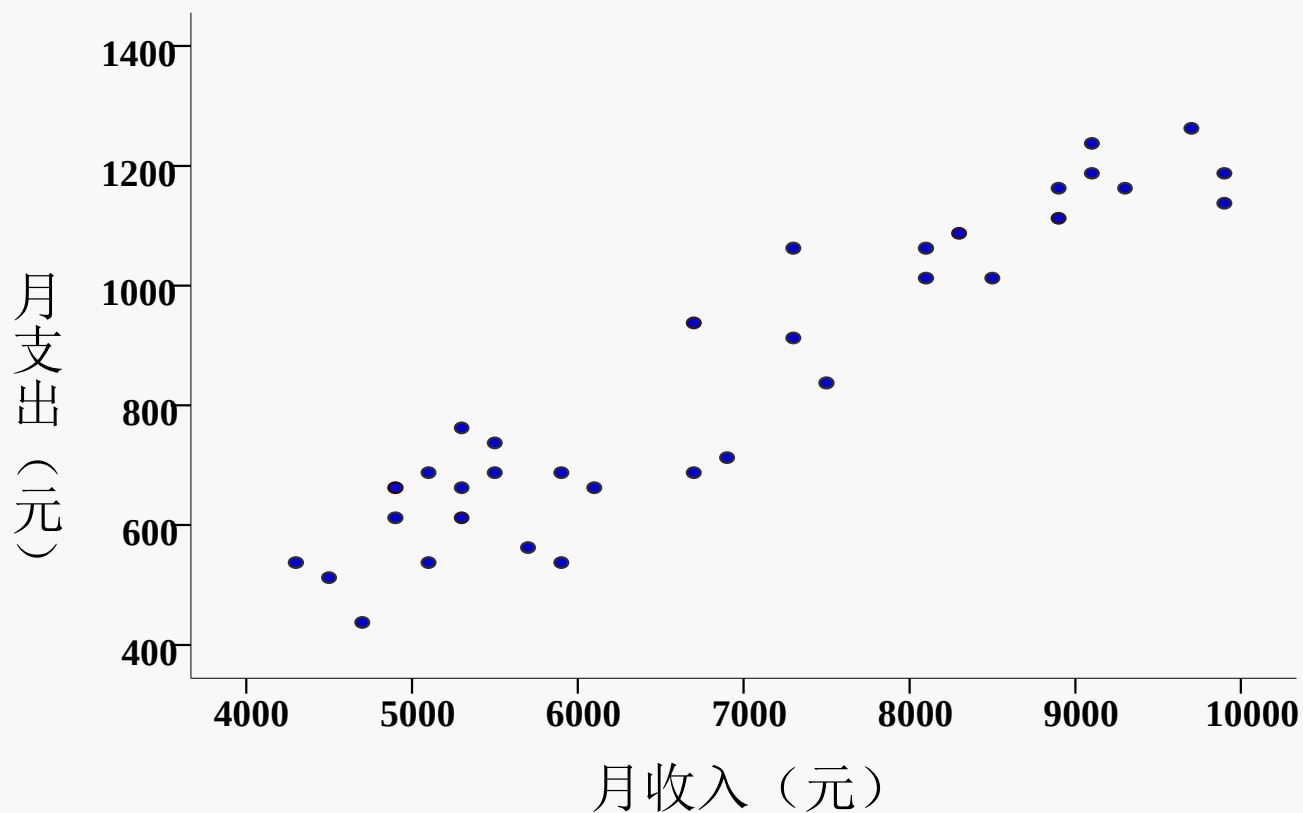
二、相关系数

三、相关系数的显著性检验

四、使用SPSS软件计算相关系数



一、散点图与相关关系



“月支出”与“月收入”样本数据散点图



一、散点图与相关关系

相关关系

指两个变量之间存在的数量上的协变关系，其中一个变量发生变化，另一个变量沿着一定的方向也随之发生变化。

散点图

是从样本数据上直观判断两个数值型变量之间是否具有相关关系的有用工具，图中的各个点，由横轴变量和纵轴变量的每一对观测值所决定。



一、散点图与相关关系

相关关系与函数关系的区别：

- **函数关系**：当一个变量的取值发生变化时，另一个变量有唯一确定的值与之相对应，即**函数关系具有确定性**。
- **相关关系**：当一个变量的取值发生变化时，另一个变量的取值是不确定的，而是遵循某种规律在一定范围内发生取值的变化，即**相关关系具有不确定性**。



一、散点图与相关关系

线性相关关系

一个变量的取值整体上沿着一条直线，随着另一个变量取值的变化而变化，则称两者之间具有线性相关关系。其中，两个变量变动方向相同，为正线性相关；两个变量的变动方向相反，为负线性相关。

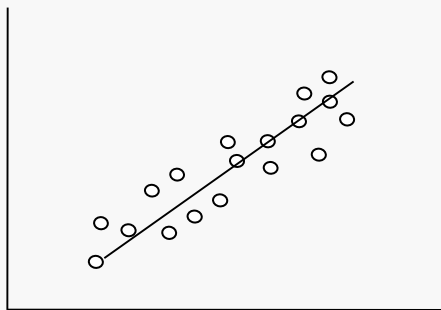
非线性相关关系

如果一个变量的取值整体上是沿着一条曲线变化，则称两者之间是非线性相关关系，或称曲线相关关系。

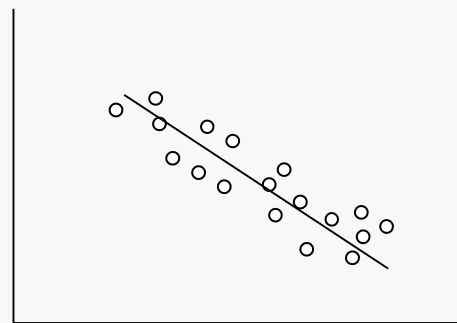


一、散点图与相关关系

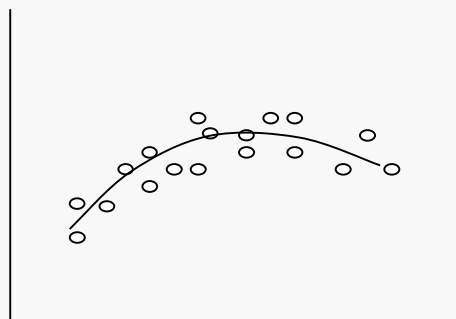
相关关系的类型



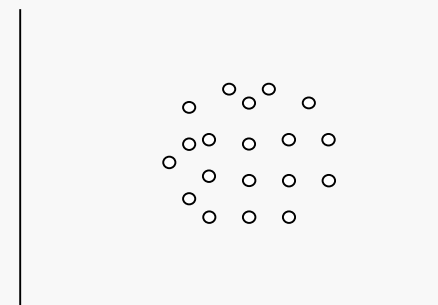
(A) 正线性相关



(B) 负线性相关



(C) 非线性相关



(D) 不相关



二、相关系数

相关系数

度量两个变量之间线性相关关系的类型及其密切程度的统计量。

计算公式:

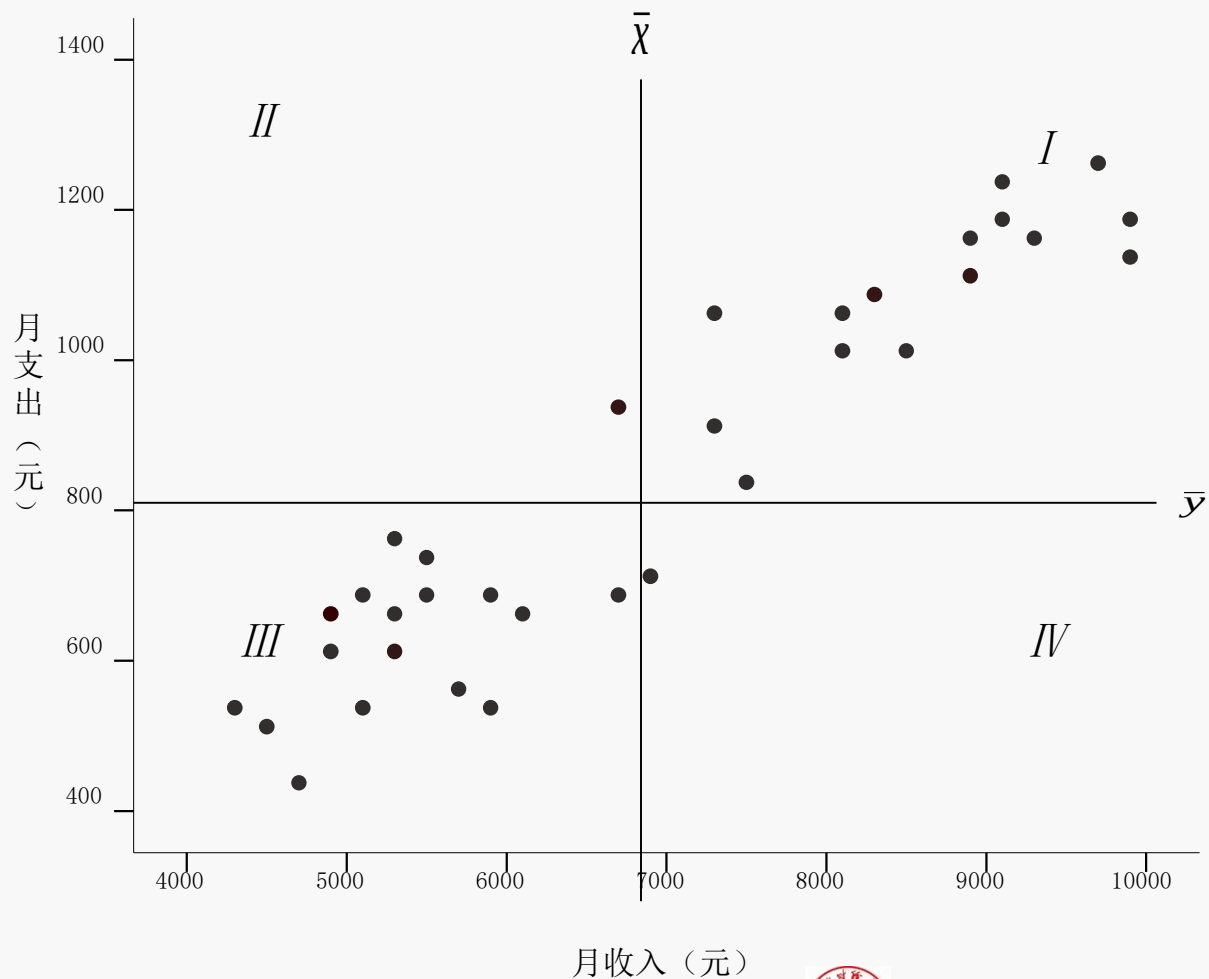
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{\left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right] \left[n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 \right]}}$$



二、相关系数

相关系数的构造原理



二、相关系数

正线性相关呈现的特点是：落在第1象限和第3象限的点，多于落在第2象限和第4象限的点。

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) > 0$$

负线性相关呈现的特点是：落在第2象限和第4象限的点，多于落在第1象限和第3象限的点。

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) < 0$$

把落在第1象限和第3象限的点叫作正相关点，落在第2象限和第4象限的点叫作负相关点，落在均值线 $\bar{y}$ 或 $\bar{x}$ 上的点叫作零相关点。

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 0$$



二、相关系数

积差和 $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ 可作为 x 与 y 之间是否具有线性相关关系以及线性相关关系强弱的一种度量。

由于：

$$0 \leq \left| \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right| \leq \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}$$

所以有：

$$-1 \leq \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \leq 1$$
$$-1 \leq r \leq 1$$



二、相关系数

- 相关系数 r 消除了积差和 $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ 中样本容量和计量单位的影响，以一个系数来衡量两个变量之间线性相关关系的强弱。
- $r=0$ ，表明 x 与 y 之间不存在线性相关关系；
 - $r>0$ ，表明存在正的线性相关关系；
 - $r<0$ ，表明存在负的线性相关关系；
 - r 的绝对值越接近于1，表明线性相关关系越强，当 r 的绝对值等于1时，表明 x 与 y 完全正相关或完全负相关。
 - 一般地，当 $0.5 < |r| < 0.75$ 时，可以认为变量间具有中等强度的线性相关关系；当 $0.75 < |r| < 1$ 时，则认为变量间的线性相关关系很强。



三、相关系数的显著性检验

- ▶ 两变量之间总体相关系数的理论真值是未知的，只能根据随机样本数据计算出样本相关系数，再据此对总体相关系数的理论真值做出具有一定把握程度的推断。
- ▶ 当总体相关系数等于零时，存在一个与样本相关系数有关的统计量，服从自由度为 $n-2$ 的 t 分布：

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

- ▶ 给定显著性水平 α ，在自由度为 $n-2$ 的 t 分布下，若 $|t| \geq t_{\alpha/2}$ 表明相关系数 r 具有统计意义的显著性，即在 $1-\alpha$ 的把握程度下，可以认为总体上两个变量之间是线性相关的。



三、相关系数的显著性检验

【例9.1】根据引例中40户家庭每月日用杂货支出与月收入的样本数据，计算“月支出”与“月收入”的相关系数，并进行显著性检验。

解：计算相关系数：

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = 0.945$$



三、相关系数的显著性检验

提出假设：

$$H_0: \text{总体相关系数} = 0 \quad H_1: \text{总体相关系数} \neq 0$$

构造检验统计量：

$$t = r \sqrt{\frac{n-2}{1-r^2}} = 0.945 \times \sqrt{\frac{40-2}{1-0.945^2}} = 17.8108$$

根据给定的显著性水平 α ，确定临界值：

给定显著性水平为0.05，在自由度为 $n-2=38$ 的 t 分布下，可确定相应的临界值 $t_{\alpha/2} = t_{0.025} = 2.0244$ 。因为 $t = 17.8108 > 2.0244$ ，所以拒绝原假设，即可认为总体中在月支出与月收入之间存在显著的线性相关关系，把握程度为95%。



四、使用SPSS软件计算相关系数

操作步骤:

- 打开数据文件，选择分析方法：

Analyze→Correlate→Bivariate

- 设置变量：

“expenditure” → “variables:” , “income”
→ “variables:” ;

- 点击OK，确认执行程序。



四、使用SPSS软件计算相关系数

输出结果：

Correlations			
		月支出	月收入
月支出	Pearson Correlation	1	.945 ^{***}
	Sig. (2-tailed)		.000
	N	40	40
月收入	Pearson Correlation	.945 ^{***}	1
	Sig. (2-tailed)	.000	
	N	40	40

***. Correlation is significant at the 0.01 level (2-tailed).

可以看出，40户家庭日用杂货支出与月收入的相关系数为0.945，且 p 值即Sig.(2-tailed)为0.00，远小于设定的显著性水平0.05，可以认为家庭每月的日用杂货支出与月收入存在高度正相关关系。



第二节 回归模型与回归方程

一、回归模型

二、回归方程



第二节 回归模型与回归方程

因变量

被估计或者被预测、被解释的变量，一般用 y 表示

自变量

用来估计或者用来预测、用来解释的变量，一般用 x 表示

回归分析

研究因变量对自变量的依赖关系的一种统计分析方法，目的是通过自变量的给定值来估计或预测因变量的均值。



第二节 回归模型与回归方程

回归分析的种类：

按自变量的
个数分为

简单线性回归分析

（模型中只有一个因变量
和一个自变量）

多重线性回归分析

（一个因变量和多个自变量）



一、回归模型

简单线性回归模型

假定自变量 x 与因变量 y 在总体上存在着线性相关关系，构造简单线性回归模型为：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \dots, n$$

式中： β_0 与 β_1 为线性函数的截距和斜率，称作模型参数； ε_i 为误差项，反映了除 x 与 y 之间线性关系之外的随机因素对 y 的影响。



二、回归方程

关于误差项随机变量的假定：

1. ε_i 是期望值为零的随机变量， $E(\varepsilon_i) = 0$ ；
2. ε_i 的方差相等，记为 σ^2 ；
3. ε_i 服从正态分布，即 $\varepsilon_i \sim N(0, \sigma^2)$ ；
4. ε_i 相互独立。

由于 $E(\varepsilon_i) = 0$ 将回归模型两边取数学期望，可得

$$E(y_i) = \beta_0 + \beta_i x_i$$

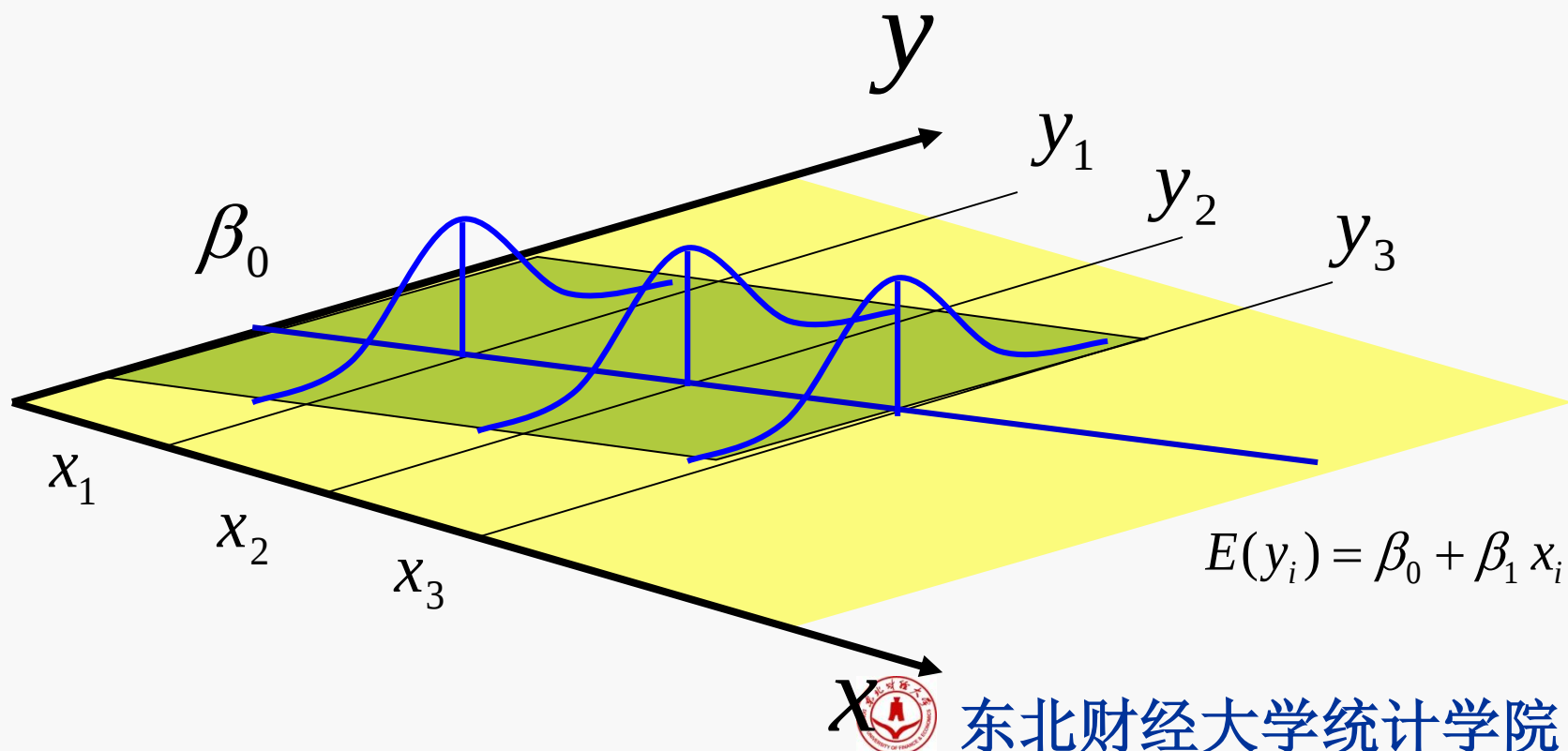
称为简单线性回归方程，表明 y_i 的期望值 $E(y_i)$ 是 x_i 的线性函数。



二、回归方程

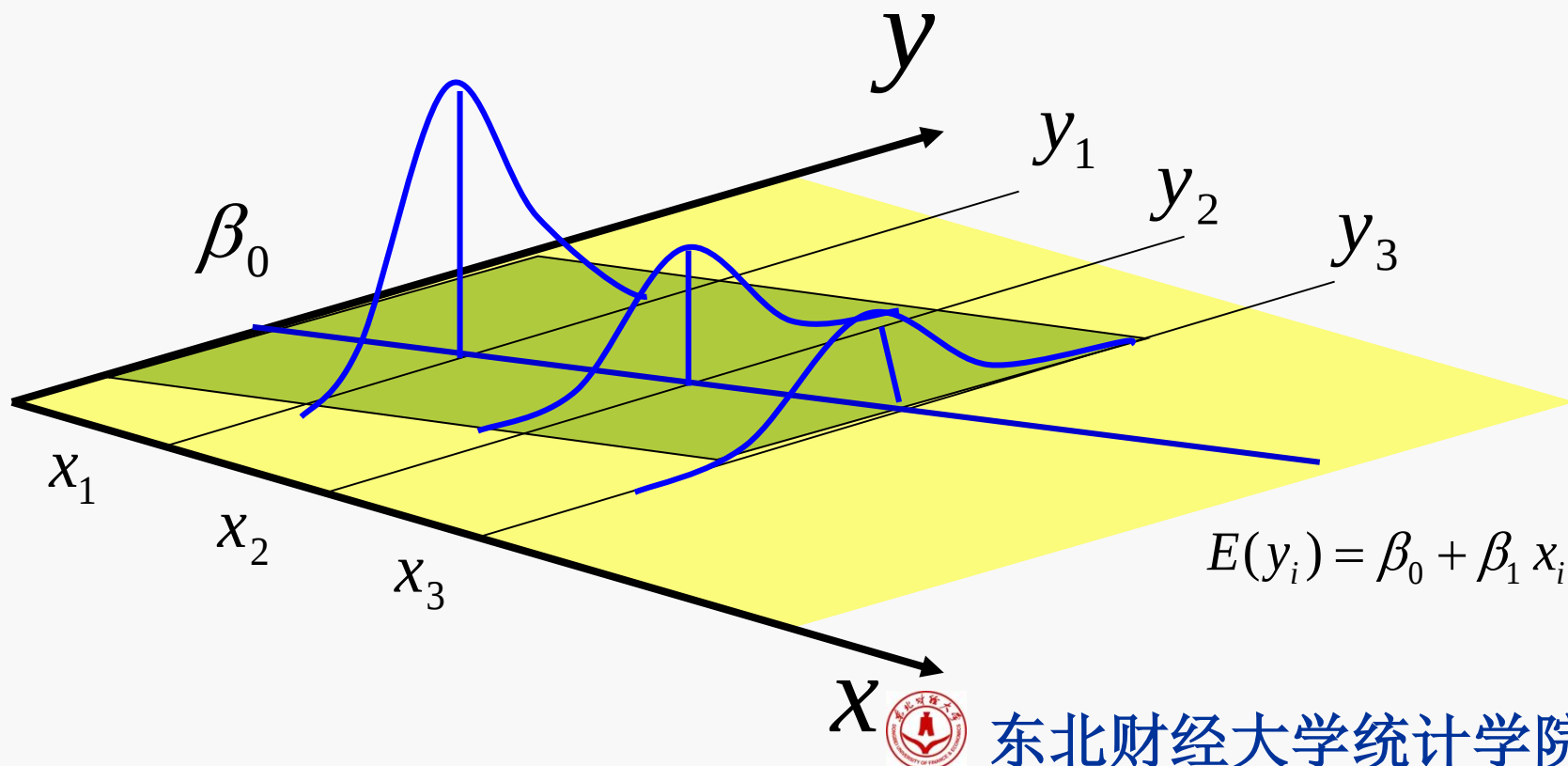
由于 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, 因此 y_i 是随机变量,
且随机性完全由 ε_i 的随机性决定。

又 $\varepsilon_i \sim N(0, \sigma^2)$, 所以 $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ 。



二、回归方程

一种明显违反模型假定的情况：各个概率密度曲线的形状不尽相同，此即所谓异方差性。



第三节 估计的回归方程



第三节 估计的回归方程

估计的回归方程

指基于样本数据，拟合出的一个确定的代表 x 与 y 之间线性相关关系的直线方程，是对总体回归方程的一个估计。

一般形式：

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

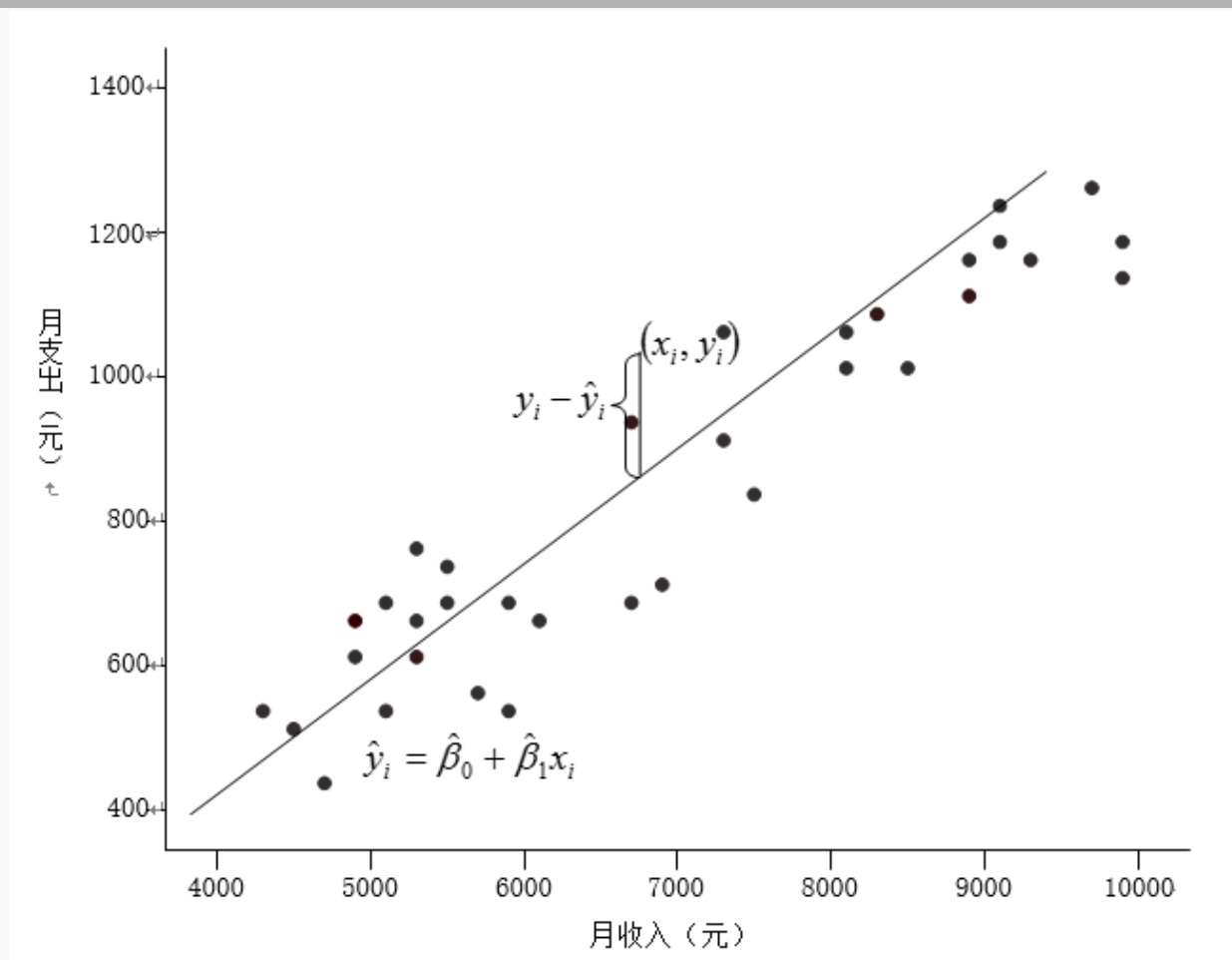
式中： $\hat{\beta}_0$ 是对 β_0 的一个估计， $\hat{\beta}_1$ 是对 β_1 的一个估计，因而 $\hat{y}_i$ 就是对 $E(y_i)$ 的一个估计。

最小平方拟合思想：

$$Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \text{取最小值}$$



第三节 估计的回归方程



最小平方法拟合原理示意图



第三节 估计的回归方程

分别对函数中 $\hat{\beta}_0$ 、 $\hat{\beta}_1$ 求偏导数，并令其为零，有

$$\begin{cases} \frac{\partial Q}{\partial \hat{\beta}_0} = 2 \sum_{i=1}^n [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i](-1) = 0 \\ \frac{\partial Q}{\partial \hat{\beta}_1} = 2 \sum_{i=1}^n [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i](-x_i) = 0 \end{cases}$$

整理得：

$$\begin{cases} \sum_{i=1}^n y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i = \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 \end{cases}$$



第三节 估计的回归方程

进一步求解得到 $\hat{\beta}_0$ 与 $\hat{\beta}_1$:

$$\begin{cases} \hat{\beta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \\ \hat{\beta}_0 = \frac{\sum y_i}{n} - \hat{\beta}_1 \frac{\sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$



第三节 估计的回归方程

【例9.2】以引例中数据为例，求样本数据的线性回归方程。

解：首先计算以下数据： $\sum_{i=1}^n y_i = 33625$ $\sum_{i=1}^n x_i = 273387$

$$\sum_{i=1}^n x_i^2 = 1987875615 \quad \sum_{i=1}^n x_i y_i = 245795835$$

代入公式：

$$\begin{cases} \sum_{i=1}^n y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i = \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 \end{cases}$$



第三节 估计的回归方程

得到：

$$\begin{cases} 33625 = 40\hat{\beta}_0 + 273387\hat{\beta}_1 \\ 245795835 = 33625\hat{\beta}_0 + 1987875615\hat{\beta}_1 \end{cases}$$

解得：

$$\begin{cases} \hat{\beta}_0 = -74.365 \\ \hat{\beta}_1 = 0.134 \end{cases}$$

得到估计的回归方程：

$$\hat{y}_i = -74.365 + 0.134 x_i$$

估计的回归方程的斜率为正（ $\hat{\beta}_1 = 0.134$ ），表明随着月收入的增加，月支出也随之增加。即：月收入每增加1元，月支出将期望增加0.134元。



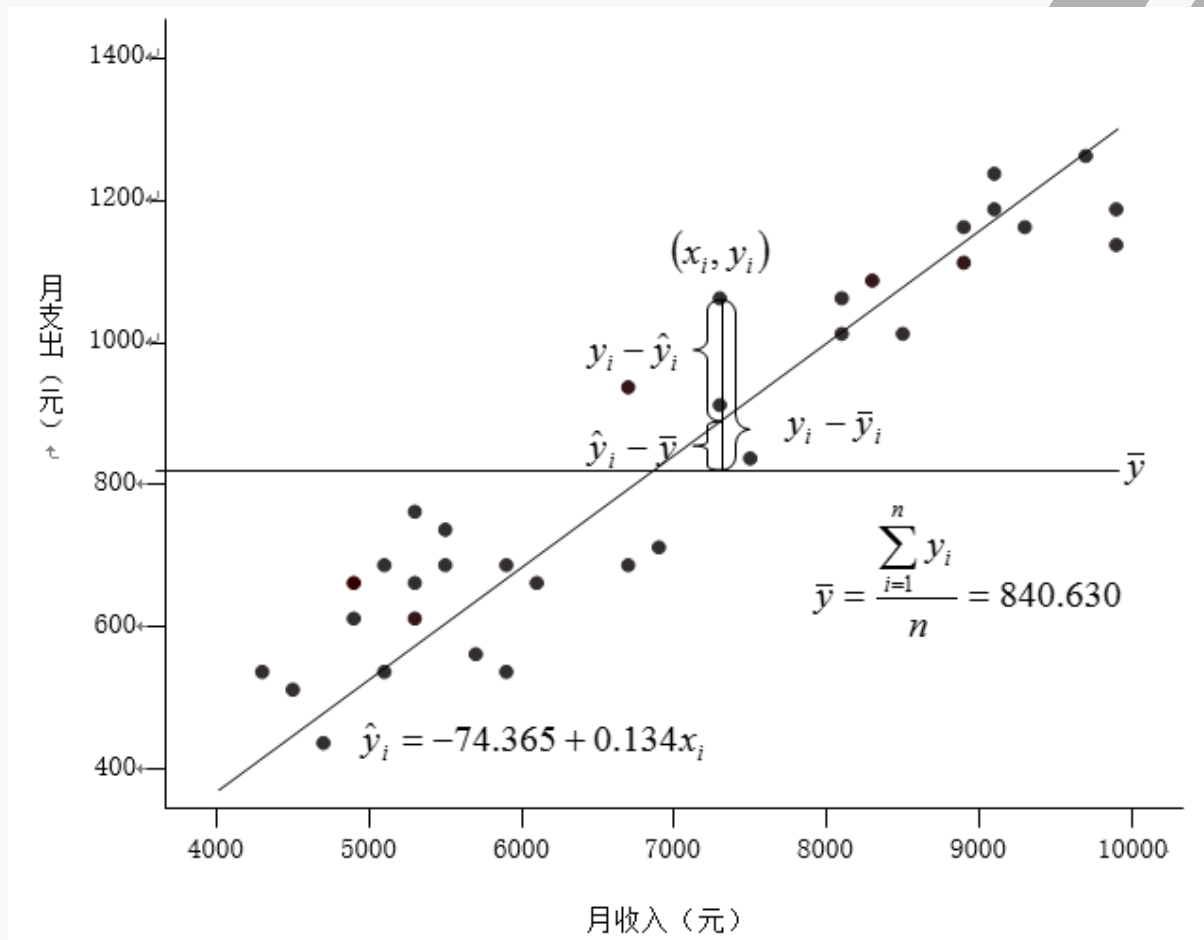
第四节 判定系数



第四节 判定系数

判定系数

用来衡量估计的回归方程对样本数据拟合紧密程度的统计量，记作 r^2 。



r^2 的构造过程：残差、总离差与回归误差



第四节 判定系数

残差平方和（误差平方和）

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

总离差平方和（总平方和）

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

回归平方和

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$



第四节 判定系数

总离差平方和的分解：

$$\begin{aligned} SST &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \end{aligned}$$

由回归分析的独立性假定有： $\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

所以有：

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

即得到： **$SST=SSE+SSR$** ，表明总平方和可以被分解为回归平方和与误差平方和两部分。



第四节 判定系数

判定系数

$$r^2 = \frac{SSR}{SST}$$

判定系数有助于评价估计的回归方程对样本数据拟合效果的好坏： r^2 越接近于1，表明估计的回归方程对样本数据的拟合效果越好； r^2 越接近于0表明拟合效果越差。



第五节 F 检验



第五节 F 检验

均方误差

残差平方和除以其自由度来估计 σ^2 ，记作 MSE 。

自由度 = $n-1-p$ （ n 为样本容量， p 为自变量个数）

均方回归

回归平方和除以其自由度来估计 σ^2 ，记作 MSR 。

自由度等于自变量的个数 p 。

- 假如 $\beta_1 = 0$ ， MSE 和 MSR 都是对 σ^2 的无偏估计，二者的比值接近于 1；
- 假如 $\beta_1 \neq 0$ ， MSE 仍然是 σ^2 的无偏估计， MSR 会出现估计偏高的倾向，二者的比值就会远远大于 1。



第五节 F 检验

提出假设: $H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$

构造检验统计量: $F = MSR/MSE$

拒绝准则:

对于给定的显著性水平 α , 如果 $F > F_\alpha$, 则拒绝 $\beta_1 = 0$ 的原假设, 表明在 $1 - \alpha$ 的把握程度上, 推断总体中 y 与 x 两个变量之间存在线性相关关系。



第五节 F 检验

通常采用方差分析表，来表述 F 检验的具体步骤：

ANOVA表一般形式

方差来源	平方和	自由度	均方	F统计量
回归	SSR	1	MSR	MSR/MSE
误差	SSE	$n-2$	MSE	
总计	SST	$n-1$	—	



第五节 F 检验

【例9.5】以引例为例，计算所拟合回归模型的检验统计量 F 值。

解：原假设为 $\beta_1 = 0$ ，备择假设为 $\beta_1 \neq 0$ ；

$$MSR = SSR/1 = 2139306.14,$$

$$MSE = SSE / (n-2) = 257563.23/38 = 6777.98;$$

$$F = MSR/MSE = \frac{2139306.14/1}{257563.23/38} = 315.626$$

在给定显著性水平0.05下，有 $F_{0.05} = 4.098$ ， $F > F_{0.05}$ ，拒绝原假设。即可认为总体中每月日用杂货支出与月收入之间，存在估计的回归方程 $\hat{y}_i = -74.365 + 0.134 x_i$ 所代表的线性相关关系。



第五节 F 检验

以方差分析表的形式呈现上述计算过程：

方差来源	平方和	自由度	均方	F统计量
回归	2139306.14	1	2139306.14	315.626
误差	257563.23	38	6777.98	
总计	2396869.38	39	—	



第六节 回归预测

一、 $E(y_i)$ 和 y_i 的点估计

二、 $E(y_i)$ 的区间估计

三、 y_i 的区间估计



一、 $E(y_i)$ 和 y_i 的点估计

预测的两种情形：

均值估计 用自变量的值预测因变量的均值，即用 x_i 预测 $E(y_i)$

个别值估计 用自变量的值预测因变量的个别值，即用 x_i 预测 y_i

➤ 两种情形下的点估计值的预测结果相等。



二、 $E(y_i)$ 的区间估计

总体均值 $E(y_i)$ 在显著性水平 α 下的置信区间为：

$$\hat{y}_i \pm t_{\alpha/2} S_{\hat{y}_i}$$

其中： $s_{\hat{y}_i}^2$ 为 $\hat{y}_i$ 关于 $E(y_i)$ 的方差 $\sigma_{\hat{y}_i}^2$ 的无偏估计量，其标准差形式为：

$$s_{\hat{y}_i} = s \sqrt{\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$$

其中 s^2 是 σ^2 的无偏估计 MSE。



二、 $E(y_i)$ 的区间估计

【例9.6】以引例为例，基于前面所建立的回归方程，在0.01显著性水平下，求月收入为8000元的全体家庭平均月日用杂货支出的区间估计。

解：经计算可知 $n=40$ ， $s = \sqrt{s^2} = \sqrt{6777.98} = 82.33$

$$\bar{x} = 6834.68, \quad \sum_{i=1}^n (x_i - \bar{x})^2 = 119364320.78$$

在0.01的显著性水平下，将 $x = 8000$ 代入回归方程中，有：

$$\hat{y}_i = -74.365 + 0.134x_i = -74.365 + 0.134 \times 8000 = 997.635$$



二、 $E(y_i)$ 的区间估计

另外查表可得 $t_{\alpha/2} = t_{0.005} = 2.712$

$$S_{\hat{y}_i} = S \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} = 82.33 \times \sqrt{\frac{1}{40} + \frac{(8000 - 6834.68)^2}{119364320.78}} = 15.70$$

总体均值的区间估计为：

$$\hat{y}_i \pm t_{\alpha/2} S_{\hat{y}_i} = 997.635 \pm 2.712 \times 15.70 = 997.635 \pm 42.58$$

即月收入水平为8000元的全部家庭，其平均月日用杂货支出在955.06元与1040.21元之间，做出这种推断的把握程度为99%。



二、 $E(y_i)$ 的区间估计

注意：当 $x_i = \bar{x}$ 时，有 $x_i - \bar{x} = 0$ ，此时 $\sigma_{\hat{y}_i}$ 有最小的估计量：

$$s_{\hat{y}_i} = s \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = s \sqrt{\frac{1}{n}}$$

当 $x_i = \bar{x}$ 时，能够得到 $E(y_i)$ 的最精确的估计区间。



三、 y_i 的区间估计

总体个别值 y_i 在显著性水平 α 下的置信区间为：

$$\hat{y}_i \pm t_{\alpha/2} S_{ind}$$

其中： s_{ind}^2 为 $\hat{y}_i$ 关于 y_i 的方差 σ_{ind}^2 的无偏估计量，其标准差形式为：

$$s_{ind} = s \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$



三、 y_i 的区间估计

【例9.7】以引例为例，基于前面所建立的回归方程，在0.01显著性水平下，计算某一月收入水平为8000元家庭的月日用杂货支出的区间估计。

解：根据例9.5中的计算结果 $n=40$, $\bar{x} = 6834.68$,

$$s = \sqrt{s^2} = \sqrt{6777.98} = 82.33$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = 119364320.78$$

点估计结果 $\hat{y}_i = 997.635$ 和 $t_{\alpha/2} = t_{0.005} = 2.711$



三、 y_i 的区间估计

则在 $\alpha = 0.01$ 的显著性水平下，估计某一月收入为8000元的家庭的月日用杂货支出区间为：

$$S_{ind} = 82.33 \times \sqrt{1 + \frac{1}{40} + \frac{(8000 - 6834.68)^2}{119364320.78}} = 83.81$$

$$\hat{y}_i \pm t_{\alpha/2} S_{ind} = 997.635 \pm 2.711 \times 83.81 = 997.635 \pm 227.209$$

即该月收入水平为8000元的家庭，其月日用杂货支出在770.43元与1224.84元之间，做出这种推断的把握程度为99%。



第七节 残差分析

一、残差图

二、标准化残差图

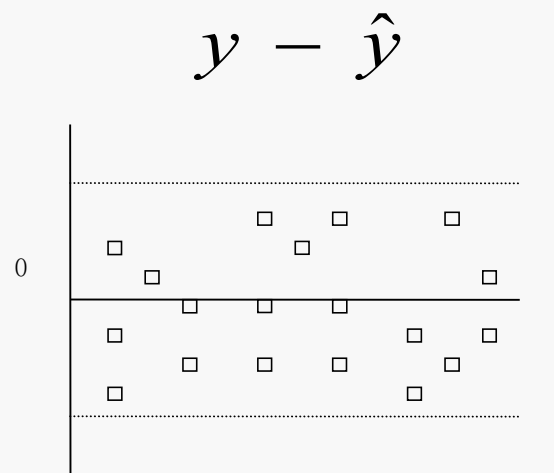
三、正态概率图

四、异常值的检测



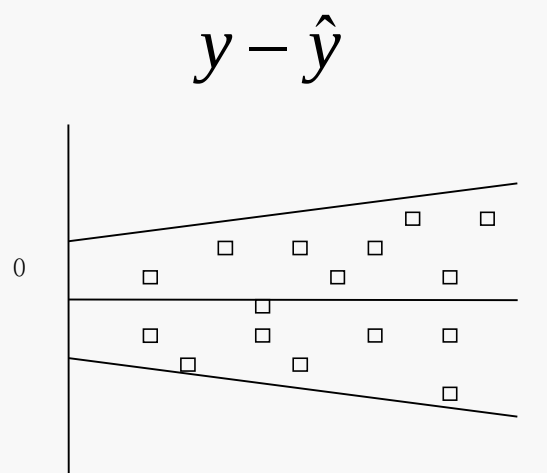
一、残差图

残差图 将样本点 $(y_i, y_i - \hat{y}_i)$ 描绘在以横轴表示因变量 y 、以纵轴表示残差 $y_i - \hat{y}_i$ 的直角坐标系中



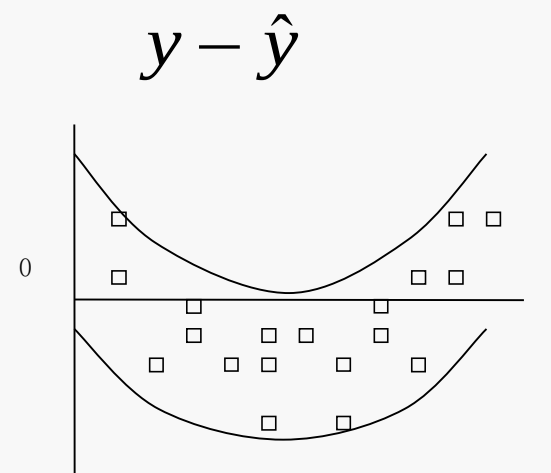
水平的分布带

(模型假定合理)



逐渐加宽的分布带

(模型假定不合理)

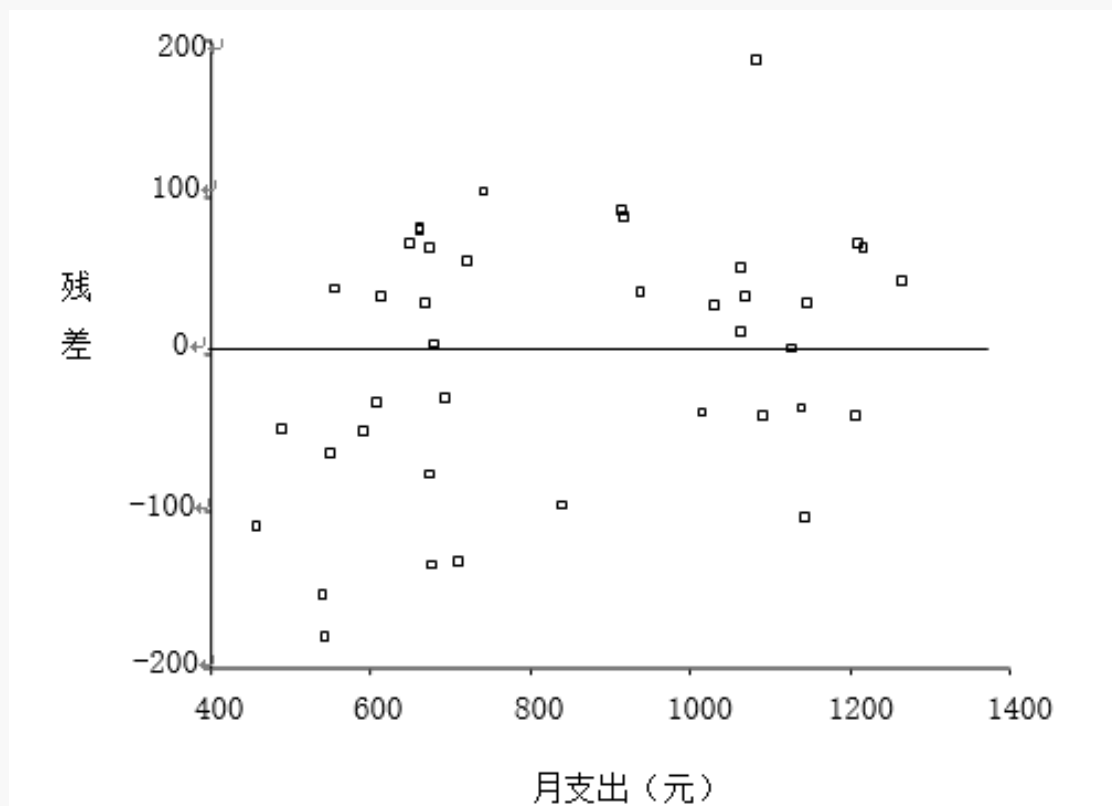


弯曲的分布带

(模型假定不合理)



一、残差图



40户家庭月日用杂货支出与月收入样本数据残差图

- 各个残差围绕 $y - \bar{y} = 0$ 的水平线上下波动，大体分布在一水平带内，没有证据表明模型假定是不合理的。



二、标准化残差图

利用正态分布的性质，可定义服从标准正态分布的统计量：

$$Z = \frac{y_i - \hat{y}_i}{\sigma}$$

均方误差 MSE 是方差 σ^2 的一个无偏估计，且自由度为 $n-2$ ，以 MSE 估计方差 σ^2 可得到自由度 $n-2$ 的 t 统计量，称作标准化残差，又称为皮尔逊残差。

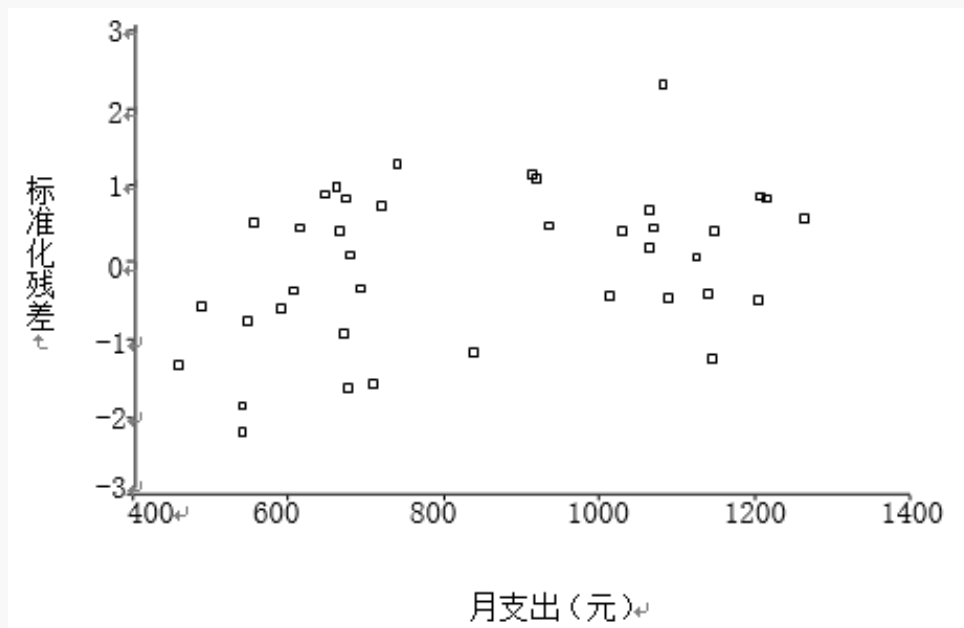
$$zre_i = \frac{y_i - \hat{y}_i}{\sqrt{MSE}}$$



二、标准化残差图

标准化残差图

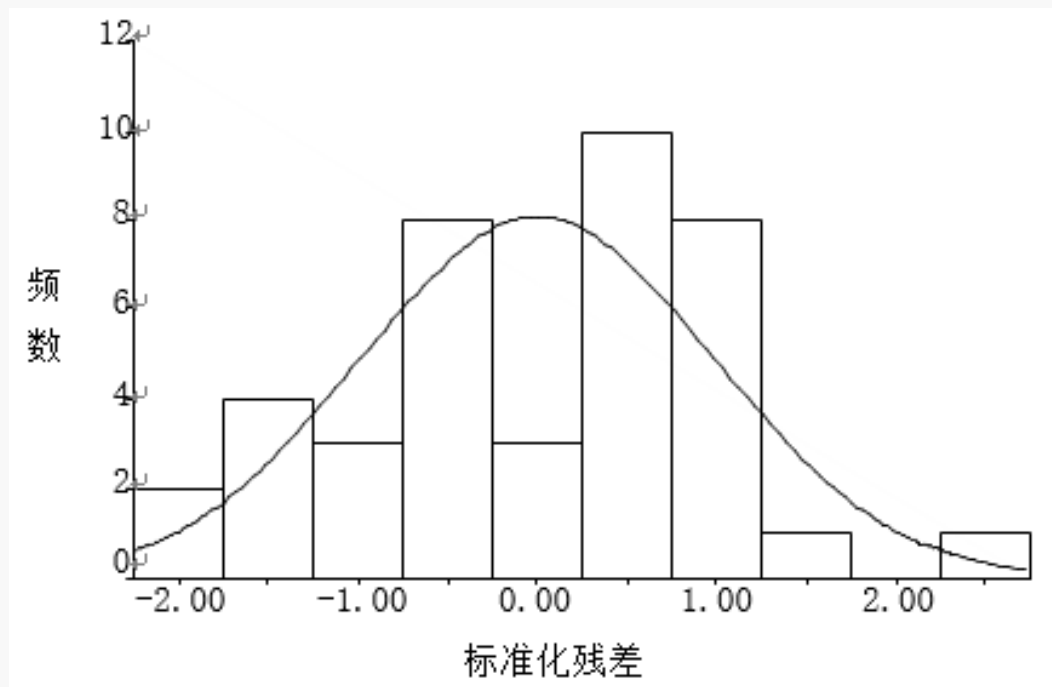
以标准化残差为纵轴、因变量为横轴，描绘出每一个残差所对应的标准化残差



如果模型假定是真实的，在样本容量充分大的前提下，大约会有95%的标准化残差介于-2和+2之间。



二、标准化残差图



如果模型假定是真实的，会看到一个大体以0为中心的对称的钟形分布。该图没有发现明显违反正态性假定的证据。



三、正态概率图

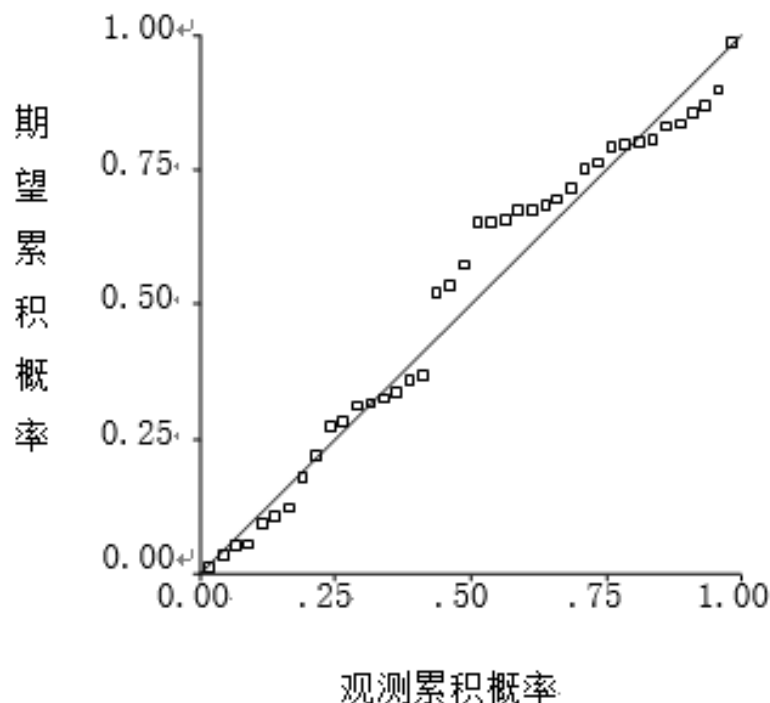
- 从一个标准正态概率分布中随机抽取容量为 n 的样本，每一个可能样本中最小的那个观测是一个随机变量，第二小的那个观测也是一个随机变量，依此类推，第 n 小即最大的那个观测也是一个随机变量，把这种随机变量称为顺序统计量。由于样本容量为 n ，所以有 n 个顺序统计量。
- 各个顺序统计量的数学期望称为正态分数，也共有 n 个正态分数。显然，第 n 小的顺序统计量的正态分数，将大于第 $n-1$ 小的顺序统计量的正态分数，依此类推。而且从负无穷大到第 n 小的正态分数的累积概率，将大于第 $n-1$ 小的正态分数的累积概率，依此类推。



三、正态概率图

正态概率图

纵轴表示正态分数所对应的期望累积概率，横轴表示标准化残差所对应的实际观测的累积概率的散点图



由于散点都分布在45度线附近，没有证据表明误差项服从正态分布的假定是不真实的。



四、异常值的检测

异常值

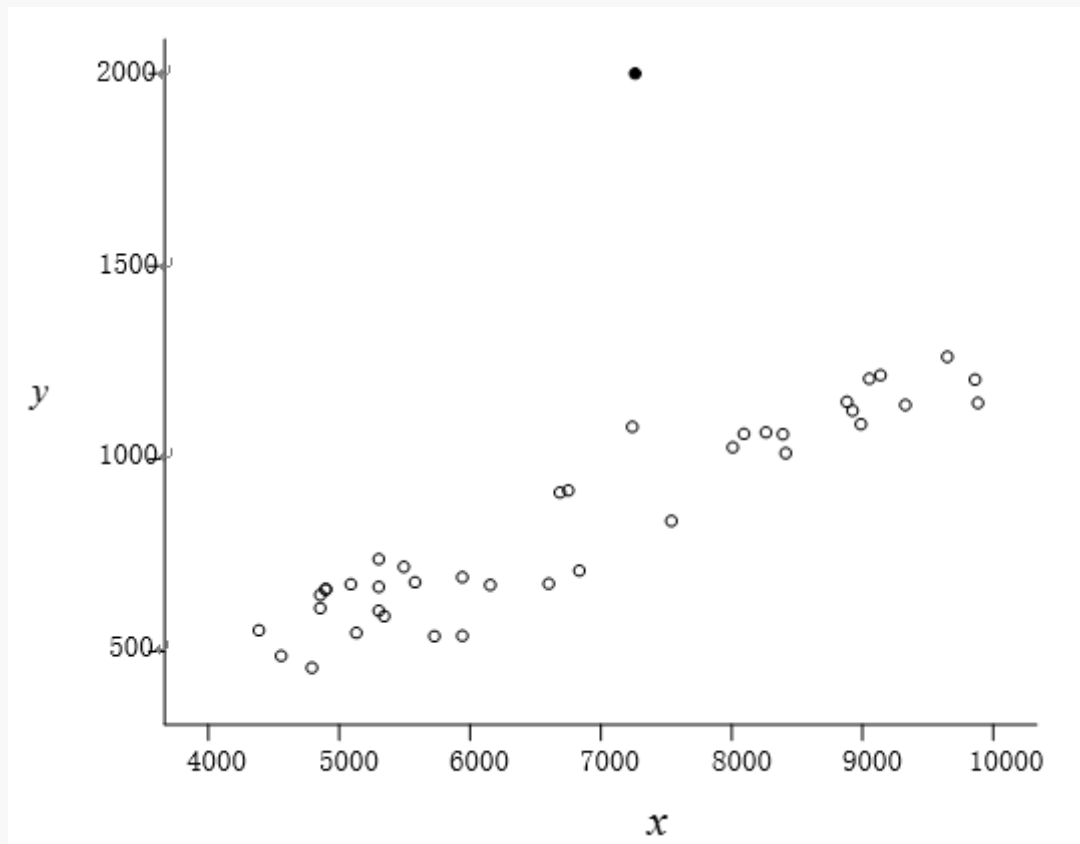
数据集中过大或过小的观测值，其存在对于回归直线方程的拟合、判定系数及显著性检验的结果都将有很大影响。

产生原因及处理办法

- 原始数据的测量或登记错误，应该尽可能地更正；
- 抽样随机性造成，应该保留这些数据，而不能随意将它们剔除；
- 异常值的出现是总体本来数据结构的一种暗示，应该考虑是否增加样本容量，或考虑采用其他形式的模型。



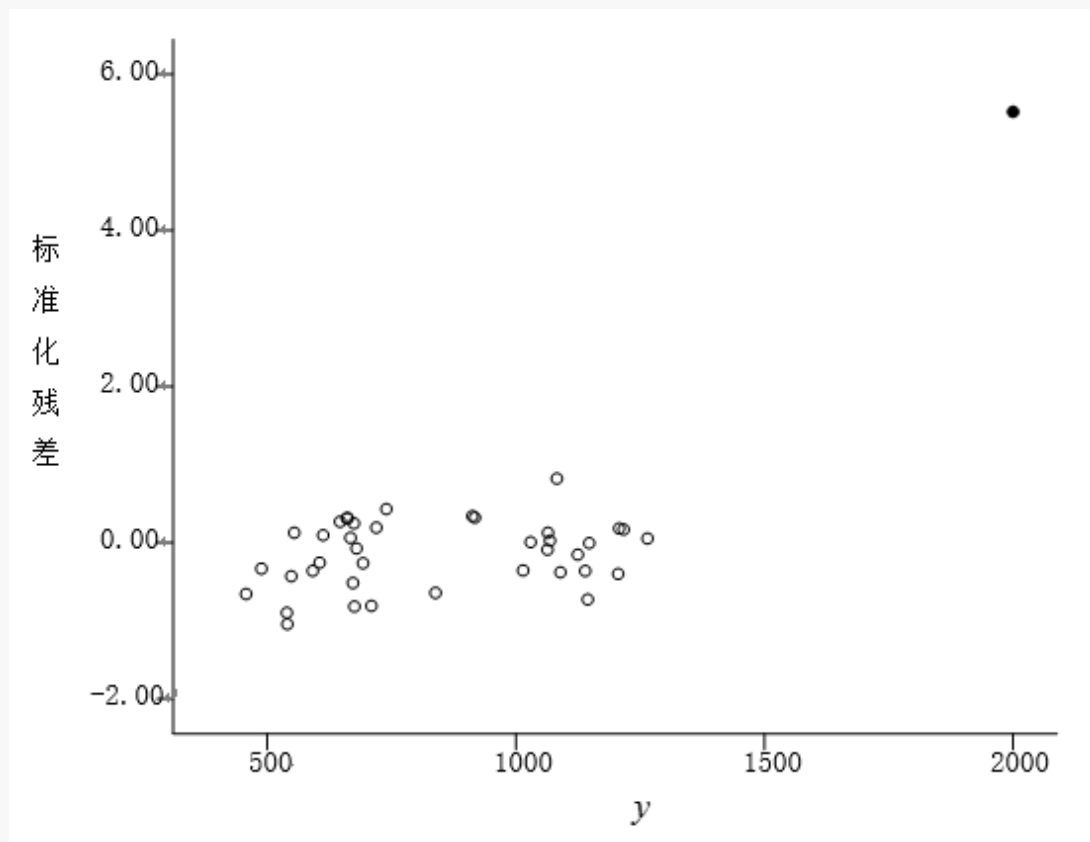
四、异常值的检测



数据集较大时，异常值将很难识别，散点图有助于直观地识别异常值。上图所示数据集中，就存在一个异常值，它表现出与数据整体分布不相吻合的倾向。



四、异常值的检测



如果某一观测值的标准化残差小于-2或大于+2，一般情况下就可以将它识别为异常值。



四、异常值的检测

高杠杆率点 指自变量的观测值是极端值的样本点。高杠杆率点对回归分析的结果影响特别大。

识别高杠杆点的方法——计算杠杆率：

$$h_i = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}$$

若 $h_i > 6/n$ ，可认为该观测值为高杠杆点

- 如果一个高杠杆率点的观测值是有效的，那么可能暗示着总体数据结构的一种不为所知的特征，这时需要进一步扩大样本容量，获取有关变量的一些补充数据，才有可能得到合理的估计模型。



第八节 基于SPSS实现回归分析过程

以引例为例，利用SPSS拟合回归模型并进行显著性检验和识别异常值的操作过程：

- 打开数据文件，选择分析方法：Analyze → Regression → Linear;
- 设置变量：“expenditure” → “Dependent”，“income” → “Independent(s)”；
- 选择回归分析方法：Method → Enter;
- 计算回归模型并进行显著性检验：Statistics → Estimate → Model fit;
- 异常值检验：Casewise diagnostics，并在“Outliers outside”的参数框中键入2;
- Continue → OK。



第八节 基于SPSS实现回归分析过程

- 绘制残差图、直方图和正态概率图：选择plots；
- 设置坐标：“*ZRESID” → x 框内，“DEPENDNT” → y框内；Histogram → Normal probability plot → continue → OK。
- 根据回归方程进行点估计和区间估计的预测：选择Analyze → Regression → Linear → save；
- 输出未标准化的预测值：Predicted Values → Unstandardized, Prediction Intervals → Mean（或Individual）；
- 输出杠杆率：Distances → Leverage values → Continue → OK。



