方差分析与回归分析

March 25, 2017

1 方差分析

1.1 单因素方差分析

1.2 多因素方差分析

2 回归分析

确定性关系：变量之间的关系可以用函数关系表达；

非确定性的关系：相关关系；

设随机变量 Y 和普通变量 x 之间存在着相关关系，由于 Y 是随机变量，对于 x 的各个确定值，Y 有它的分布。用 $F(y|x)$ 表示当 x 取确定的 x 值时，所对应的 Y 的分布函数。若知道了 $F(y|x)$ 随着 x 的取值而变化的规律，就完全掌握 Y 与 x 之间的关系了。这样做往往比较复杂。

作为一种近似，考察 Y 的数学期望。若 Y 的数学期望 $E(Y)$ 存在，则其值随 x 的取值而定，它是 x 的函数。将这一函数记为$\mu_{Y|x}$或$\mu(x)$，称为 Y 关于 x 的回归函数。将讨

论 Y 与 x 的相关关系的问题转换为讨论 $E(Y)=\mu(x)$ 与 x 的函数关系。

若 η 是一个随机变量，则 $E[(\eta-c)^2]$ 作为 c 的函数，在 $c=E(\eta)$ 时 $E[(\eta-c)^2]$ 达到最小。这表明在一切 x 的函数中，以回归函数 $\mu(x)$ 作为 Y 的近似，其均方误差 $E[(Y-\mu(x))^2]$ 为最小。

回归分析的任务是根据试验数据去估计回归函数，讨论有关的点估计、区间估计、假设检验等。对随机变量 Y 的观察值作出点预测和区间预测。

对于 x 取定一组不完全相同的值 $x_1,x_2,\cdots,x_n$，设 $Y_1,Y_2,\cdots,Y_n$ 分别是在 $x_1,x_2,\cdots,x_n$ 处对 Y 的独立观察结果，称

$$(x_1,Y_1),(x_2,Y_2),\cdots,(x_n,Y_n) \tag{1}$$

是一个样本，对应的样本值记为

$$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n), \tag{2}$$

利用样本来估计 Y 关于 x 的回归函数 $\mu(x)$。

2.1 一元线性回归

设 Y 关于 x 的回归函数为 $\mu(x)$；利用样本来估计 $\mu(x)$ 的问题称为求 Y 关于 x 的回归问题。设 $\mu(x)$ 为线性函数：

$$\mu(x)=a+bx \tag{3}$$

设对于 x(在某个区间内) 的每一个值有

$$Y\sim N(a+bx,\sigma^2), \tag{4}$$

其中 a,b 及 σ^2 都是不依赖于 x 的未知参数。记 $\varepsilon=Y-(a+bx)$，对 Y 作这样的正态假设，相当于假设

$$Y=a+bx+\varepsilon,\ \ \varepsilon\sim N(0,\sigma^2) \tag{5}$$

其中未知参数 a, b 及 σ^2 都不依赖于 x。(5) 式称为一元线性回归模型，其中 b 称为回归系数。

2.1.1 a, b 的估计

取 x 的 n 个不全相同的值 $x_1, x_2, \cdots, x_n$ 作独立试验，得到样本

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n), \tag{6}$$

由 5 式

$$Y_i = a + bx_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \tag{7}$$

各 ε_i 相互独立。于是

$$Y_i \sim N(a + bx_i, \sigma^2), \quad i = 1, 2, \cdots, n \tag{8}$$

由 $Y_1, Y_2, \cdots, Y_n$ 的独立性，$Y_1, Y_2, \cdots, Y_n$ 的联合概率密度

$$\begin{aligned} L &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_i - a - bx_i)^2\right] \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2\right] \end{aligned} \tag{9}$$

用最大似然估计法来估计参数 a, b。对于任一组观察值 $y_1, y_2, \cdots, y_n$，(9) 式是样本的似然函数。要 L 取最大值，只要 (9) 式右端方括号中的平方和为最小，即

$$Q(a, b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2 \tag{10}$$

取最小值。

(Note: 若 Y 不是正态变量，直接使用 (10) 式估计 a, b，使 Y 的观察值 y_i 与 $a + bx_i$ 偏差的平方和 $Q(a, b)$ 为最小，这种方法称为最小二乘法。若 Y 是正态变量，则最小二乘法与最大似然估计法给出相同的结果)

$$
\begin{aligned}
\frac{\partial Q}{\partial a} &= -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\
\frac{\partial Q}{\partial b} &= -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0
\end{aligned} \tag{11}
$$

得到

$$
\begin{aligned}
na + \left(\sum_{i=1}^{n} x_i\right) b &= \sum_{i=1}^{n} y_i \\
\left(\sum_{i=1}^{n} x_i\right) a + \left(\sum_{i=1}^{n} x_i^2\right) b &= \sum_{i=1}^{n} x_i y_i
\end{aligned} \tag{12}
$$

称为正规方程组。

由于 x_i 不全相同，正规方程组的系数行列式

$$
\begin{vmatrix} n & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^2 \end{vmatrix} = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 = n\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0\ .
$$

方程组有唯一解。b, a 的最大似然估计为

$$
\begin{aligned}
\hat{b} &= \frac{n\sum\limits_{i=1}^{n} x_i y_i - \left(\sum\limits_{i=1}^{n} x_i\right)\left(\sum\limits_{i=1}^{n} y_i\right)}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} \\
\hat{a} &= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{\hat{b}}{n}\sum_{i=1}^{n} x_i = \bar{y} - \hat{b}\bar{x}
\end{aligned} \tag{13}
$$

其中

$$
\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i,
\end{aligned} \tag{14}
$$

在得到 a, b 的估计 $\hat{a}, \hat{b}$ 后，对于给定的 x，取 $\hat{a}+\hat{b}x$ 作为回归函数 $\mu = a + bx$ 的估计，即 $\widehat{\mu(x)} = \hat{a}+\hat{b}x$，称为 Y 关于 x 的经验回归函数。方程

$$\hat{y} = \hat{a}+\hat{b}x \tag{15}$$

称为 Y 关于 x 的经验回归方程，简称回归方程，图形称为回归直线。将 $\hat{a}$ 的表达式代入回归方程，则回归方程可写成

$$\hat{y} = \bar{y} + \hat{b}(x-\bar{x}) \ . \tag{16}$$

即对于样本值 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$，回归直线通过散点图的几何中心 $(\bar{x}, \bar{y})$。

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \\ S_{yy} &= \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \\ S_{xy} &= \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \end{aligned} \tag{17}$$

则

$$\begin{aligned} \hat{b} &= \frac{S_{xy}}{S_{xx}} \ , \\ \hat{a} &= \frac{1}{n}\sum_{i=1}^{n} y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\hat{b} \ . \end{aligned} \tag{18}$$

2.1.2 σ^2 的估计

$$E\{[Y-(a+bx)]^2\} = E(\epsilon^2) = D(\epsilon) + [E(\epsilon)]^2 = \sigma^2$$

利用样本估计 σ^2：

记 $\hat{y}_i = \hat{y}|_{x=x_i} = \hat{a} + \hat{b}x_i$，$y_i - \hat{y}_i$ 记为 x_i 处的残差。

$$Q_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 \tag{19}$$

记为残差平方和。它是经验回归函数在 x_i 处的函数值 $\widehat{\mu(x_i)} = \hat{a} + \hat{b}x_i$ 与 x_i 处的观察值 y_i 的偏差的平方和。

$$\begin{aligned}
Q_e &= \sum_{i=1}^{n}(y - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - \bar{y}_i - \hat{b}(x_i - \bar{x})]^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\hat{b}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \hat{b}^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= S_{yy} - 2\hat{b}S_{xy} + \hat{b}^2 S_{xx} \\
&= S_{yy} - \hat{b}S_{xy}
\end{aligned}$$

b, a 的估计量分别为

$$\begin{aligned}
\hat{b} &= \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} \\
\hat{a} &= \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{\hat{b}}{n}\sum_{i=1}^{n} x_i = \overline{Y} - \hat{b}\bar{x}
\end{aligned} \tag{20}$$

其中 $\overline{Y} = \frac{1}{n}\sum\limits_{i=1}^{n} Y_i$，$\bar{x} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$。将 y_i 改为 $Y_i(i = 1, 2, \cdots, n)$，记为

$$\begin{aligned}
S_{YY} &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \\
S_{xY} &= \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \overline{Y})
\end{aligned}$$

残差平方和 Q_e 的相应的统计量 (仍记为 Q_e) 为

$$Q_e = S_{YY} - \hat{b}S_{xY} \ . \tag{21}$$

残差平方和 Q_e服从分布

$$\frac{Q_e}{\sigma^2} \sim \chi^2(n-2) \ , \tag{22}$$

因此

$$E\left(\frac{Q_e}{\sigma^2}\right) = n-2 \ ,$$

也就是

$$E\left(\frac{Q_e}{n-2}\right) = \sigma^2 \ .$$

于是得到 σ^2 的无偏估计量:

$$\widehat{\sigma^2} = \frac{Q_e}{n-2} = \frac{S_{YY} - \hat{b}S_{xY}}{n-2} \ . \tag{23}$$

2.1.3 线性假设的显著性检验

用 t 检验法检验假设

$$\begin{aligned} &H_0 : b = 0 \ , \\ &H_1 : b \neq 0 \ . \end{aligned} \tag{24}$$

已知

$$\hat{b} \sim N\left(b, \ \frac{\sigma^2}{S_{xx}}\right) \ . \tag{25}$$

又

$$\frac{(n-2)\widehat{\sigma^2}}{\sigma^2} = \frac{Q_e}{\sigma^2} \sim \chi^2(n-2) \ , \tag{26}$$

且 $\hat{b}$ 与 Q_e 独立。故

$$\frac{\hat{b}-b}{\sqrt{\sigma^2/S_{xx}}} \Bigg/ \sqrt{\frac{(n-2)\widehat{\sigma^2}}{\sigma^2} \Bigg/ (n-2)} = \frac{\hat{b}-b}{\hat{\sigma}}\sqrt{S_{xx}} \sim t(n-2) \ .$$

其中 $\hat{\sigma} = \sqrt{\widehat{\sigma^2}}$。当 H_0 为真时 $b = 0$,

$$t = \frac{\hat{b}}{\hat{\sigma}}\sqrt{S_{xx}} \sim t(n-2) \ , \tag{27}$$

且 $E(\hat{b}) = b = 0$，即得 H_0 的拒绝域

$$|t| = \frac{|\hat{b}|}{\hat{\sigma}}\sqrt{S_{xx}} \geqslant t_{\alpha/2}(n-2)\ , \tag{28}$$

α 为显著性水平。

当假设 $H_0 : b = 0$ 被拒绝时，认为回归效果是显著的，反之，认为回归效果不显著。

2.1.4 系数 b 的置信区间

当回归效果显著时，需要对系数 b 作区间估计。由

$$\frac{\hat{b}-b}{\hat{\sigma}}\sqrt{S_{xx}} \sim t(n-2)$$

得到 b 的置信水平为 $1-\alpha$ 的置信区间为

$$\left(\hat{b} \pm t_{\alpha/2}(n-2) \times \frac{\hat{\sigma}}{S_{xx}}\right) \tag{29}$$

2.1.5 回归函数 $\mu(x) = a + bx$ 函数值的点估计和置信区间

设 x_0 是自变量 x 的某一指定值。用经验回归函数 $\hat{y} = \widehat{\mu(x)} = \hat{a} + \hat{b}x$ 在 x_0 的函数值 $\hat{y}_0 = \widehat{\mu(x_0)} = \hat{a} + \hat{b}x_0$ 作为 $\mu(x_0) = a + bx_0$ 的点估计，即

$$\hat{y}_0 = \widehat{\mu(x_0)} = \hat{a} + \hat{b}x_0\ . \tag{30}$$

考虑相应的估计量

$$\hat{Y}_0 = \hat{a} + \hat{b}x_0\ , \tag{31}$$

由于 $E(\hat{Y}_0) = a + bx_0$，这一估计量是无偏的。求 $\mu(x_0) = a + bx_0$ 的置信区间：

$$\frac{\hat{Y}_0 - (a + bx_0)}{\sigma\sqrt{\dfrac{1}{n} + \dfrac{(x_0-\bar{x})^2}{S_{xx}}}} \sim N(0,1)\ .$$

$$\frac{(n-2)\widehat{\sigma^2}}{\sigma^2}=\frac{Q_e}{\sigma^2}\sim\chi^2(n-2)\ , \tag{32}$$

且 $Q_e, \hat{Y}_0$ 相互独立。

$$\frac{\hat{Y}_0-(a+bx_0)}{\sigma\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}}\Bigg/\sqrt{\frac{(n-2)\widehat{\sigma^2}}{\sigma^2}\Big/(n-2)}=\frac{\hat{Y}_0-(a+bx_0)}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}}\sim t(n-2)\ ,$$

$\mu(x_0)=a+bx_0$ 的置信水平为 $1-\alpha$ 的置信区间为

$$\left(\hat{Y}_0\pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}\right)\ , \tag{33}$$

或

$$\left(\hat{a}+\hat{b}x_0\pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}\right)\ . \tag{34}$$

这一置信区间的长度是 x_0 的函数，它随 $|x_0-\bar{x}|$ 的增加而增加，当 $x_0=\bar{x}$ 时最短。

2.1.6 Y 的观察值的点预测和预测区间

一元回归模型

$$Y=\mu(x;\theta_1,\theta_2,\cdots,\theta_p)+\varepsilon,\quad \varepsilon\sim N(0,\sigma^2), \tag{35}$$

其中 $\theta_1,\theta_2,\cdots,\theta_p,\sigma^2$ 是与 x 无关的未知参数;

线性回归模型:

若回归函数 $\mu(x;\theta_1,\theta_2,\cdots,\theta_p)$ 是参数 $\theta_1,\theta_2,\cdots,\theta_p$ 的线性函数;

非线性回归模型:

若回归函数 $\mu(x;\theta_1,\theta_2,\cdots,\theta_p)$ 是参数 $\theta_1,\theta_2,\cdots,\theta_p$ 的非线性函数;

2.2 多元线性回归

随机变量 Y 与多个变量 $x_1,x_2,\cdots,x_p(p>1)$ 有关。对于自变量 $x_1,x_2,\cdots,x_p$ 的一组确定的值，Y 有它的分布。若 Y 的数学期望存在，则它是 $x_1,x_2,\cdots,x_p$ 的函数，记为

$\mu_{Y|x_1,x_2,\cdots,x_p}$ 或 $\mu(x_1,x_2,\cdots,x_p)$，它就是 Y 关于 x 的回归函数。假设

$$Y = b_0 + b_1 x_1 + \cdots + b_p x_p + \epsilon, \quad \epsilon \sim N(0,\sigma^2)\ , \tag{36}$$

其中 $b_0, b_1, \cdots, b_p, \sigma^2$ 都是与 $x_1, x_2, \cdots, x_p$ 无关的未知参数。