# 一元线性回归(一)

## I. 引言

1. 期望与中位数的另一种定义:

$$\min_b E(X-b)^2 \Rightarrow E(X)=b$$

$$\min_b E\,|X-b| \Rightarrow med(X)=b$$

$proof.$

$$E(X-b)^2 = E(X^2-2bX+b^2) = E(X^2)-2bE(X)+b^2 \quad (1)$$

可将 $(1)$ 式看为 $b$ 的一个一元二次函数,现求 $b$ 的最小值。可由公式得: $b=E(X).$

对于中位数，我们只考虑连续的情况，离散时类似可证。令随机变量 $X$ 的概率密度为 $p(x)$, 概率分布为 $F(X)$，则有等式成立:

$$E|X-b| = \int_{-\infty}^{\infty} |x-b|\, p(x)dx = \int_{-\infty}^{b} (b-x)p(x)dx + \int_{b}^{\infty} (x-b)p(x)dx$$

现要求 $b$ 的最小值，两边对 $b$ 求导，并令导数为 $0$ .

$$\int_{-\infty}^{b} p(x)dx - \int_{b}^{\infty} p(x)dx = 0$$

$\Rightarrow F(b)-(1-F(b))=0$ , 即 $F(b)=\dfrac{1}{2}$ , $b=med(X).$

2. Conditional Expectation :
 i.) Law of Total Expectation :

$$E\,[E(X|Y)] = E(X)$$

 ii.) Law of Total Variance

$$Var(X) = Var\,[E(X|Y)] + E\,[Var(X|Y)]$$

$proof.$
设随机变量 $X, Y$的联合密度为$f(x,y)$,边缘密度分别为$f_X(x), f_Y(y)$,则成立等式:
i.) $E(X) = \int x f_X(x)dx = \int\int x f(x,y)dxdy$
$\quad = \int\int x\,[\,f(x|y)\cdot f_Y(y)\,]\,dx\,dy$
$\quad = \int[\int x\cdot f(x|y)\,dx]\,f_Y(y)\,dy$

$$= \int E(X|y) \cdot f_Y(y)\, dy = E[E(X|Y)].$$

ii.) $$\begin{aligned} Var(X) &= E(X-E(X))^2 = E[X - E(X|Y) + E(X|Y) - E(X)]^2 \\ &= E[X-E(X|Y)]^2 + E[E(X|Y)-E(X)]^2 + 2E[(X-E(X|Y))(E(X|Y)-E(X))] \qquad (1) \end{aligned}$$

而 $$\begin{aligned} E[X-E(X|Y)]^2 &= E\left\{ E[(X-E(X|Y))^2|Y] \right\} \\ &= E\left\{ E[(X^2 + (E(X|Y))^2 - 2XE(X|Y))|Y] \right\} \\ &= E\left\{ E(X^2|Y) + (E(X|Y))^2 - 2(E(X|Y))^2 \right\} \\ &= E\left\{ E(X^2|Y) - (E(X|Y))^2 \right\} \\ &= E[Var(X|Y)] \end{aligned}$$

$$\begin{aligned} E[E(X|Y)-E(X)]^2 &= E\left\{ E(X|Y) - E[E(X|Y)] \right\} \\ &= Var[E(X|Y)] \end{aligned}$$

$$\begin{aligned} E[(X-E(X|Y))(E(X|Y)-E(X))] &= E\left\{ [(X-E(X|Y))(E(X|Y)-E(X))]|Y \right\} \\ &= (E(X|Y))^2 - E(X)E(X|Y) - (E(X|Y))^2 + E(X|Y)E(X) = 0 \end{aligned}$$

所以 $(1) = Var(E(X|Y)) + E(Var(X|Y))$.

## II. 一元回归模型

1. **OLSE**下的参数估计

   **OLSE**的主要思想是寻找参数 $\beta_0, \beta_1$ 的估计值使离差平方和达到最小。从而可转化为如下问题：

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

分别对 $\beta_0, \beta_1$ 求偏导，并令偏导为 $0$，结果如下：

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

整理后可得：

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

对 $(1)$ 两边同时除以 $n$，整理后可得：

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

对于 (2) 注意到有如下两个等式：

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

$$= \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

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

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

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

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

由上述公式，(2) 变为：

$$n\bar{x}\hat{\beta}_0 + (\sum_{i=1}^{n}(x_i - \bar{x})^2 + n\bar{x}^2)\hat{\beta}_1 = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + n\bar{x}\bar{y}$$

再由 (1) 整理后可得：

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{cases}$$

2. **MLE**下的参数估计

**MLE**的主要思想是利用总体的概率密度及其样本提供的信息构造似然函数，通过似然函数取值最大得到对应参数的估计。

对于一元线性回归模型参数的MLE，我们假设 $\varepsilon_i \sim N(0, \sigma^2)$ 时，$y_i$ 服从如下正态分布：

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

则 $y_i$的概率密度为

$$f_i(y_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$

构造似然函数:

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} f_i(y_i)$$

$$= \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$

对似然函数取对数:

$$ln(L) = -nln((\sqrt{2\pi}\sigma)) - \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

为了使似然函数 $L$取得最大值,等价的我们只需使 $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ 取得最小值，即

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

至此与**OLSE**一致。

3. **OLSE**估计的性质
i. ) 线性

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{j=1}^{n} \frac{(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} y_i - \bar{y} \sum_{j=1}^{n} \frac{(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

注意到:

$$\sum_{j=1}^{n} \frac{(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\bar{x} - n\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0$$

所以

$$\hat{\beta}_1 = \sum_{j=1}^{n} \frac{(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} y_j$$

是关于 $y_j$ 的线性组合。

由于 $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ 且 $\bar{x}$ 是常数，结合上式可得：

$$\beta_0 = \sum_{j=1}^{n}(\frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2})y_j$$

也是关于 $y_j$ 的线性组合。

ii. ) 无偏性

注意到由 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ 可得：

$$E(y_i) = \beta_0 + \beta_1 x_i \ \ (Markov - Gauss\ Condition)$$

所以

$$\begin{aligned}
E(\hat{\beta}_1) &= \sum_{j=1}^{n} \frac{x_j - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} E(y_j) \\
&= \sum_{j=1}^{n} \frac{x_j - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} (\beta_0 + \beta_1 x_j) \\
&= \beta_0 \cdot \sum_{j=1}^{n} \frac{x_j - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \beta_1 \cdot \sum_{j=1}^{n} \frac{x_j^2 - x_j\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (1)
\end{aligned}$$

注意到

$$\sum_{j=1}^{n} \frac{x_j - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\bar{x} - n\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0$$

$$\begin{aligned}
\sum_{j=1}^{n} \frac{x_j^2 - x_j\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} &= \frac{\sum_{j=1}^{n} x_j^2 - n\bar{x}^2}{\sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i} \\
&= \frac{\sum_{j=1}^{n} x_j^2 - n\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = 1
\end{aligned}$$

所以

$$E(\hat{\beta}_1) = (1) = \beta_1.$$

$$
\begin{aligned}
E(\hat{\beta}_0) &= \sum_{j=1}^{n}\left(\frac{1}{n} - \frac{(x_j-\bar{x})\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)E(y_j) \\
&= \sum_{j=1}^{n}\left(\frac{1}{n} - \frac{(x_j-\bar{x})\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)(\beta_0+\beta_1 x_j) \\
&= \beta_0 - \beta_0\bar{x}\cdot\sum_{j=1}^{n}\frac{(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \beta_1\bar{x} - \beta_1\bar{x}\sum_{j=1}^{n}\frac{x_j^2 - x_j\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}
\end{aligned}
$$

由上述两个已证公式,可得：

$$E(\hat{\beta}_0) = \beta_0 - \beta_0\bar{x}\cdot 0 + \beta_1\bar{x} - \beta_1\bar{x}\cdot 1 = \beta_0.$$

iii. ) $\hat{\beta}_0, \hat{\beta}_1$ 的方差：

$$
\begin{aligned}
Var(\hat{\beta}_1) &= Var\left(\sum_{j=1}^{n}\frac{(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}y_j\right) \\
&= \sum_{j=1}^{n}\left(\frac{(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^2 Var(y_j) \qquad (Markov-Gauss\ Condition) \\
&= \frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}\sigma^2 \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{\beta}_0) &= Var\left(\sum_{j=1}^{n}\left(\frac{1}{n} - \frac{(x_j-\bar{x})\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)y_j\right) \\
&= \sum_{j=1}^{n}\left(\frac{1}{n} - \frac{(x_j-\bar{x})\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^2 Var(y_i) \quad (Markov-Gauss\ Condition) \\
&= \sum\left(\frac{1}{n^2} + \left(\frac{(x_j-\bar{x})\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^2 - \frac{2}{n}\frac{(x_j-\bar{x})\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\sigma^2 \\
&= \sigma^2\left(\sum\frac{1}{n^2} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} - \frac{2\bar{x}}{n}\sum\frac{(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right.
\end{aligned}
$$

$$= \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}).$$

协方差:

由于 $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$,故对于统计量 $\bar{y}$ 有:

$$Var(\bar{y}) = Var(\hat{\beta_0} + \hat{\beta_1}\bar{x}) = Var(\hat{\beta_0}) + \bar{x}^2 Var(\hat{\beta_1}) + 2\bar{x}Cov(\hat{\beta_1}, \hat{\beta_0})$$

$$\Rightarrow \quad \frac{\sigma^2}{n} = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}) + \frac{\bar{x}^2\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + 2\bar{x}Cov(\hat{\beta_1}, \hat{\beta_0})$$

$$\Rightarrow \quad Cov(\hat{\beta_1}, \hat{\beta_0}) = -\frac{\bar{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\Rightarrow \quad Cov(\hat{\beta_1}, \hat{\beta_0}) = -\frac{\bar{x}\sigma^2}{L_{xx}}$$

由于 $\hat{\beta_0}$ 和 $\hat{\beta_1}$ 都是 $y_i$ 的线性组合，且 $y_i$ 间独立同分布，故 $\hat{\beta_0}$ 和 $\hat{\beta_1}$ 也服从正态分布，即:

$$\hat{\beta_0} \sim N(\beta_0, (\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}})\sigma^2)$$

$$\hat{\beta_1} \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$$