《线性回归》
—线性回归(7)

杨　　瑛

清华大学　数学科学系
Email: yangying@mail.tsinghua.edu.cn
Tel: 62796887

2019.03.28

主要内容：预测

- 预测
- 习题

预测

- 主要内容见：Draper and Smith (1998). p.80-87

预测

- 回归的主要目的之一是用来预测。假定我们收集到数据$(\mathbf{X}_i, \mathbf{Y}_i), 1 \le i \le n$. 然后建立线性回归模型：

$$\mathbf{Y}_i = \beta_0 + \beta_1 \mathbf{X}_i + \epsilon_i, \epsilon_i \sim iid\ N(0, \sigma^2).$$

  其中β_0，β_1和σ^2是未知的参数。利用LS法，我们可以得到β_0和β_1 的LSE分别为: b_0和b_1.

- 当有新的$x = x_0$时，我们要预测对应随机变量y_0的值。很明显，

$$\widehat{y}_0 = b_0 + b_1 x_0. \tag{1}$$

- 由LSE的性质，$\mathbf{E}[b_0] = \beta_0$, $\mathbf{E}[b_1] = \beta_1$. 因此

$$\mathbf{E}[\widehat{y}_0] = \beta_0 + \beta_1 x_0. \tag{2}$$

$\widehat{y}_0$的方差为：

$$\mathsf{Var}[\widehat{y}_0] = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (\mathbf{X}_i - \overline{X})^2} \right]. \tag{3}$$

这是因为：

$$\begin{aligned}
\mathsf{Var}[\widehat{y}_0] &= \mathbf{E}[\widehat{y}_0 - \mathbf{E}[\widehat{y}_0]]^2] \\
&= \mathbf{E}[(b_0 + b_1 x_0) - (\beta_0 + \beta_1 x_0)]^2 \\
&= \mathbf{E}[(b_0 - \beta_0] + (b_0 - \beta_1) x_0]^2 \\
&= \mathbf{E}\left[(b_1 - \beta_1)(x_0 - \overline{X}) + \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \right]^2 \\
&= \sigma^2 + \sigma^2 (x_0 - \overline{X})^2 / \sum_{i=1}^{n} (\mathbf{X}_i - \overline{X})^2.
\end{aligned}$$

- 可以看出，Var$[\widehat{y}_0]$随着$(x_0-\overline{X})^2$的增加而增加。因而，x_0离$\overline{X}$越远，方差Var$[\widehat{y}_0]$越大。
- 注意y_0是在x_0处未来的观测(不知道具体的值)，我们用$\widehat{y}_0$去预测y_0. 我们现在感兴趣差：

$$y_0-\widehat{y}_0. \tag{4}$$

- 我们现在欲得到(4)中的均值和方差。注意到模型假设，有

$$y_0=\beta_0+\beta_1 x_0+\epsilon_0, \epsilon_0\sim N(0,\sigma^2), \epsilon_0\text{与}\epsilon_i\text{独立}, i=1,\cdots,n.$$

♠

$$\mathbf{E}[y_0] = \beta_0 + \beta_1 x_0.$$

结合(2), 得到：

$$\mathbf{E}[y_0 - \widehat{y}_0] = 0. \tag{5}$$

由此，(4)的方差为：

$$\begin{aligned} \mathrm{Var}[y_0 - \widehat{y}_0] &= \mathrm{Var}[y_0] + \mathrm{Var}[\widehat{y}_0] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{X})^2}{\sum_{i=1}^n (\mathbf{X}_i - \overline{X})^2} \right]. \end{aligned} \tag{6}$$

♠ 因为σ^2是未知的，在实际应用中，要用S^2代替(3)和(6)的σ^2.

♠ 由于未来的观测值y_0未必在真的直线$y = \beta_0 + \beta_1 x$或者拟合直线$y = b_0 + b_1 x$上。由于$y_0 - \beta_0 - \beta_1 x_0$的方差为$\sigma^2$. 因此，$y_0 - \widehat{y}_0$的方差是两部分$\mathrm{Var}[y_0]$和$\mathrm{Var}[\widehat{y}_0]$的组合。

预测区间

- 因为y_0是在x_0处的新的观测值，所以$y_0, \widehat{y}_0$和S^2都是独立的。因此，下面的两个随机变量时独立的：

$$Z = \frac{y - \widehat{y}_0}{\sigma\left[1 + \frac{1}{n} + \frac{(x - \overline{X})^2}{s_x^2}\right]^{1/2}}, W = \frac{(n-2)S^2}{\sigma^2},$$

其中$s_x^2 = \sum_{i=1}^{n}(\mathbf{X}_i - \overline{X})^2$.
因为y_0和$\widehat{y}_0$是独立的正态随机变量，故$Z \sim N(0,1)$.
$W \sim \chi_{n-2}^2$. 于是$Z/(W/(n-2))^{1/2} \sim t_{n-2}$.

$$U_{x_0} = \frac{y_0 - \widehat{y}_0}{S\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{X})^2}{s_x^2}\right]^{1/2}} \sim t_{n-2}. \tag{7}$$

♠ 给定$1-\alpha_0 \in (0,1)$，由7得到

$$P(|U_{x_0}| \le t_{n-2}^{-1}(1-\alpha_0/2)) = 1-\alpha_0.$$

♠ 从而, 随机变量y_0介于两个随机变量

$$\widehat{y}_0 \pm t_{n-2}^{-1}(1-\alpha_0/2) S \left[1+\frac{1}{n}+\frac{(x_0-\overline{X})^2}{s_x^2}\right]^{1/2}. \qquad (8)$$

之间的概率是$1-\alpha_0$.

定义.

*(8)*中给出的随机区间称为y_0的系数为$1-\alpha_0$的**预测区间**。

- 请注意参数的置信区间和随机响应的预测区间的差别。在得到(8)中的所有观测值之后，(8)中的区间与置信区间有相同的解释，但要注意y_0是随机变量！！
- 关于设计点和预测区间进一步内容，请仔细阅读 Draper and Smith (1998) Applied Regression Analysis中p. 79-89的内容，特别是与Figure 3.1, 3.2 3.3有关的内容。这部分内容不在课堂讲解，但要求掌握。

习题

- Draper and Smith (1998) Applied Regression Analysis，p. 96-108中的J, K, T, AA