Chap 8 线性回归

1 回归问题

- **定义**

$$Y = f(X_1, \cdots, X_n) + \varepsilon.$$

  - Y — 因变量(响应变量).
  - $X_1, \cdots, X_n$ — 自变量(回归变量).
  - ε — 随机误差(无法测量或不重要的因素).

  假定 $E(\varepsilon \mid X_1, \cdots, X_n) = 0$,

$$\Rightarrow E(Y \mid X_1, \cdots, X_n) = f(X_1, \cdots, X_n).$$

  称为 Y 对 $X_1, \cdots, X_n$ 的回归函数. 由样本数据 $X_1, \cdots, X_n, Y$ 获取 f 的过程称为回归(有监督学习).

- **注**

  - $X_1, \cdots, X_n$ 可以是随机的(**e.g.** 随机抽取一人的身高、体重等).
  - $X_1, \cdots, X_n$ 也可以是非随机的控制变量(**e.g.** 施肥量、药品使用剂量).
  - 在应用中, 自变量一律视为非随机的.

- **假设** $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$ (未知).

- **注** 要素是否完全、f 的形式是否准确关乎 σ^2 的大小.

2 简单线性回归

- **定义**

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

  这是**理论模型**, 提供背景作用. 其中回归参数(未知待定):

  - β_0 — 截距.
  - β_1 — 斜率(回归系数).

  对 (X, Y) 进行 n 次独立观测, 得到样本观测值 $(x_1, y_1), \cdots, (x_n, y_n)$. 则

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \, (1 \le i \le n).$$

  其中 ε_i 作为第 i 次观测的随机误差, 无法直接观测得到. 不妨认为

$$\begin{cases} E(\varepsilon_i) = 0, \\ Var(\varepsilon_i) = \sigma^2. \end{cases}$$

这是**简单线性回归模型**. 其中:

- $E(y_i) = \beta_0 + \beta_1 x_i$.
- $Var(y_i) = \sigma^2$.

- **注**
  - **简单**: $n = 1$.
  - **线性**: f 关于参数 β_0, β_1 线性.

3 最小二乘法 (LS) 估计参数

- **定义**

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

最小化 $S(\beta_0, \beta_1)$, 得

- $\hat{\beta}_1 = \dfrac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2} = \dfrac{\sum\limits_{i=1}^{n}(x_i - \overline{x})y_i}{S_{xx}}$ (y_i 的线性组合).
- $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \sum\limits_{i=1}^{n}(\dfrac{1}{n} - \dfrac{(x_i - \overline{x})\overline{x}}{S_{xx}})y_i$ (y_i 的线性组合).
- $y = \hat{\beta}_0 + \hat{\beta}_1 x$ (拟合直线).

- **注**
  - 损失函数: $(y - (\beta_0 + \beta_1 x))^2$.
  - 线性模型是否合理.

- **命题** $\hat{\beta}_0$, $\hat{\beta}_1$ 分别为 β_0, β_1 的无偏估计.

- **证明**

$$\begin{aligned} E(\hat{\beta}_1) &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})E(y_i)}{S_{xx}} \\ &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(\beta_0 + \beta_1 x_i)}{S_{xx}} \\ &= \beta_1 \frac{\sum_{i=1}^{n}(x_i - \overline{x})x_i}{S_{xx}} \\ &= \beta_1 \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{S_{xx}} \\ &= \beta_1. \end{aligned}$$

$$
\begin{aligned}
E(\hat{\beta}_0) &= E(\overline{y} - \hat{\beta}_1 \overline{x}) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(y_i) - E(\hat{\beta}_1)\overline{x} \\
&= \frac{1}{n}\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) - \beta_1 \overline{x} \\
&= \beta_0.
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{\beta}_1) &= Var\Big(\frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{S_{xx}}\Big) \\
&= \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{S_{xx}^2} Var(y_i) \\
&= \frac{\sigma^2}{S_{xx}}.
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{\beta}_0) &= Var\Big(\sum_{i=1}^{n}(\frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{S_{xx}})y_i\Big) \\
&= \sum_{i=1}^{n}(\frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{S_{xx}})^2 Var(y_i) \\
&= (\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}})\sigma^2 \\
&= \frac{\sigma^2}{S_{xx}} \cdot \frac{\sum_{i=1}^{n} x_i^2}{n}.
\end{aligned}
$$

- **注** 中心化处理:

  $$y_i = \beta_0 + \beta_1 \overline{x} + \beta_1(x_i - \overline{x}) + \varepsilon_i.$$

  此时常数项 $\beta_0 + \beta_1\overline{x}$ 的估计 $= \hat{\beta}_0 + \hat{\beta}_1\overline{x} = \overline{y}$.

- **定义(残差)** 当 $X = x_i$ 时, 拟合直线上相应点为 $(x_i, \hat{\beta}_0 + \hat{\beta}_1 x_i)$.

  记 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, 称为 x_i 处的拟合值. 定义残差 $y_i - \hat{y}_i$. 考虑残差平方和

  $$SSE := \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

- **命题** $\hat{\sigma}^2 := \dfrac{SSE}{n-2}$ 为 σ^2 的无偏估计. 此时
  - $\hat{se}(\hat{\beta}_1) = \dfrac{\hat{\sigma}}{\sqrt{S_{xx}}}$.
  - $\hat{se}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}}}$.

4 回归参数推断

- **追加假设** $\varepsilon_i \sim N(0,\sigma^2)$, $1 \le i \le n$.
- **注**
  - $y_i \sim N(\beta_0+\beta_1 x_i, \sigma^2)$ 独立, $1 \le i \le n$.
  - $\mathbf{MLE}(\beta_0^*, \beta_1^*) = (\hat{\beta}_0, \hat{\beta}_1)$ (**习题课 5**).

    $$L(\beta_0,\beta_1,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-(\beta_0+\beta_1 x_i))^2}{2\sigma^2}}.$$

    $$(\sigma^2)^* = \frac{SSE}{n}.$$
- **定义(假设检验)** $H_0: \beta_1 = 0$ **v.s.** $H_1: \beta_1 \neq 0$. 因为 $\hat{\beta}_1$ 为 y_i 的线性组合. 得到

  $$\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1)$$

  可证明

  $$\frac{SSE}{\sigma^2} = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

  从而

  $$\frac{\hat{\beta}_1 - \beta_1}{\hat{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{\sigma}}{\sqrt{S_{xx}}}} = \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \Big/ \sqrt{\frac{(n-2)\hat{\sigma}^2}{(n-2)\sigma^2}} \sim t(n-2).$$

  检验统计量:

  $$T = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)}.$$

  当 H_0 为真时, $T \sim t(n-2)$. 检验准则为: 当 $|T| \ge t_{\frac{\alpha}{2}}(n-2)$ 时拒绝 H_0.
- **注**
  - 可以对其他的 β_1 可能值进行检验.
  - 可以对 β_1 进行区间估计.
  - 可以对 β_0 进行相应推断, 过程类似.

5 预测

- **例** 当 $X = x_0$ 时, $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$, 其中 $\varepsilon \sim N(0,\sigma^2)$. 令

  $$\mu_0 = E(y_0) = \beta_0 + \beta_1 x_0,$$

  给出对 μ_0 的预测.

- **解答** 用拟合直线上 x_0 处的取值 $\hat{y}_0$ 给出 μ_0 的点估计:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \overline{y} - \hat{\beta}_1(x_0 - \overline{x}) = \sum_{i=1}^{n}(\frac{1}{n} + \frac{(x_i - \overline{x})(x_0 - \overline{x})}{S_{xx}})y_i$$

分别给出

  - $E(\hat{y}_0) = E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0 = \mu_0.$
  - $Var(\hat{y}_0) = \sum_{i=1}^{n}\left(\frac{1}{n} + \frac{(x_i - \overline{x})(x_0 - \overline{x})}{S_{xx}}\right)^2\sigma^2 = \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)\sigma^2.$

从而

$$\frac{\hat{y}_0 - \mu_0}{se(\hat{y}_0)} \sim N(0,1).$$

使用 $\hat{se}(\hat{y}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$ 估计 $se(\hat{y}_0)$, 我们有

$$\frac{\hat{y}_0 - \mu_0}{\hat{se}(\hat{y}_0)} \sim t(n-2).$$

从而 μ_0 的 $(1-\alpha)-$ 置信的双侧区间估计为

$$\left(\hat{y}_0 - t_{\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}, \hat{y}_0 + t_{\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}\right).$$

- **例** 当 $X = x_0$ 时, $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$, 其中 $\varepsilon \sim N(0, \sigma^2)$. 给出对 y_0 的预测.
- **解答** $y_0 \sim N(\mu_0, \sigma^2)$. 若 μ_0 已知, 则 y_0 的(均方意义下最优)估计为 μ_0.

  一般情况下, y_0 的良好点估计为 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

  注意到 y_0 与 $\hat{y}_0$ 相互独立, 从而 $\hat{y}_0 - y_0$ 服从正态分布.

  分别给出

  - $E(\hat{y}_0 - y_0) = E(\hat{y}_0) - E(y_0) = \mu_0 - \mu_0 = 0.$
  - $Var(\hat{y}_0 - y_0) = Var(\hat{y}_0) + Var(y_0) = \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}).$

  从而

$$\frac{\hat{y}_0 - y_0}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}} \sim N(0,1).$$

  进而

$$\frac{\hat{y}_0 - y_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}} \sim t(n-2).$$

从而 y_0 的 $(1-\alpha)-$ 置信的双侧区间估计为

$$\left(\hat{y}_0 - t_{\frac{\alpha}{2}}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}, \hat{y}_0 + t_{\frac{\alpha}{2}}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}\right).$$

- **注** 当 x_0 与 $\overline{x}$ 距离增加时, 估计误差增大.
- **注**
  - 结合实际理解 β.
  - 外推需谨慎.
  - 截距为 0 的回归复杂度 $n-2 \to n-1$.
  - 回归方程不可逆转使用.
  - 常见应用:
    - 描述趋势.
    - 预测均值\取值.
    - 实验控制.