第九章、第十章

- 第十九次课

  §9.1 引言

  §9.2 一元线性回归

- 第二十次课

  §9.2 一元线性回归(续)

  §9.3 多元线性回归(简介)

  §10.1 统计决策问题概述

§9.1 引言

- 自变量: x, 因变量: y. 函数关系f (未知): $y=f(x)$.
- 误差: e. 回归关系: $y=f(x)+e$.
- $x=$ 路程(可设定), $y=$ 耗油量.

  $x=$ 父亲身高(不可设定, 只可测量), $y=$ 儿子身高.

  关心f, 不关心自变量如何变化. 将x 视为参数.

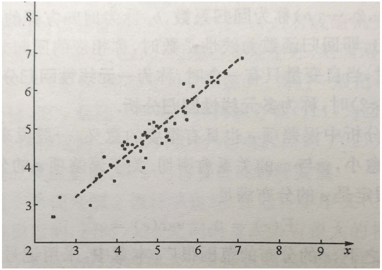
  将y, e 视为随机变量或其取值.
- 正态模型: $y=f(x)+e$, 其中$e\sim N(0,\sigma^2)$, σ^2 未知.
- 数据(x_i, y_i), $i=1,\cdots,n$.

  回归模型: $y_i=f(x_i)+e_i$, $i=1,\cdots,n$.

  x_i 是参数, y_i 是随机变量或其取值(可观测).
- e_i 是随机变量, 但取值未知(不可观测), 因为f 未知.

线性回归: f 为线性函数.

- 一元: $f(x) = a + bx$. 多元: $f(\vec{x}) = a + b_1x_1 + \cdots + b_px_p$. 参数$a, b; b_1, \cdots, b_p$ 未知.
- p 是自变量$\vec{x}$ 的维数, n 是数据量(样本量).
- 例1.1. $y = f(x) + e$, 一元. 数据: (x_i, y_i), $i = 1, \cdots, n = 50$. 观察散点图, 确认f 是否线性.



- 建立回归模型: $y_i = a + bx_i + e_i$, $i = 1, \cdots, n$.

例1.3. $x=$ 某小区人口数, $y=$ 冬季用煤量, $z=$ 室温.

- 预测. 回归关系: $y=a+bx+e$.

  数据(x_i,y_i), $i=1,\cdots,n$.

  小区人口为x, 冬季应储备多少煤?

  自变量→ 因变量.

- 控制. 回归关系: $z=c+dy+\varepsilon$.

  数据(y_i,z_i), $i=1,\cdots,n$.

  为控制室温为18度, 冬季应储备多少煤?

  因变量→ 自变量.

§9.2 一元线性回归

$y=a+bx+e$, $e\sim N(0,\sigma^2)$, σ^2 未知.

数据: (x_i,y_i), $i=1,\cdots,n$.

- 最小二乘拟合系数 指: 使得$Q(a,b)=\sum_{i=1}^n[y_i-(a+bx_i)]^2$ 达到最小的a, b. 记为$\hat{a}$, $\hat{b}$.
- 回归模型: $y_i=a+bx_i+e_i$, $i=1,\cdots,n$.
  $p_{Y_i}(y_i)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y_i-(a+bx_i))^2}$,
  视x_i 为已知参数, a, b 为未知待估参数, σ^2 为讨厌参数.
- 似然函数: $L(a,b,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n[y_i-(a+bx_i)]^2}$.
- 最大似然估计: $L(a,b,\sigma^2)$ 的最大值点.
  a, b 的最大似然估计使得$Q(a,b)$ 达到最小, 即为$\hat{a}$, $\hat{b}$.

定理2.1. $\hat{a}=\bar{y}-\hat{b}\bar{x}$, $\hat{b}=\frac{\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n(x_i-\bar{x})^2}=\frac{\ell_{xy}}{\ell_{xx}}$.

- 样本: $P(\xi=x_i,\eta=y_i)=\frac{1}{n}$, $i=1,\cdots,n$.
  则$E[\eta-(a+b\xi)]^2=\frac{1}{n}\sum_{i=1}^n[y_i-(a+bx_i)]^2=\frac{1}{n}Q(a,b)$.
- 回顾第八次课, 第三章定理5.3.
  $a=E\eta-bE\xi\Rightarrow$
  $\frac{1}{n}Q(a,b)=E(\tilde{\eta}-b\tilde{\xi})^2$, 其中$\tilde{\eta}=\eta-E\eta$, $\tilde{\xi}=\xi-E\xi$.
- 正交分解: $\tilde{\eta}-b\tilde{\xi}=(\tilde{\eta}-\hat{b}\tilde{\xi})\oplus(\hat{b}-b)\tilde{\xi}$,
  $\hat{b}$ 满足$\oplus$: 即$E(\tilde{\eta}-\hat{b}\tilde{\xi})\tilde{\xi}=0$. 故$\hat{b}=\frac{\mathrm{cov}(\xi,\eta)}{\mathrm{var}(\xi)}$, $\hat{a}=E\eta-\hat{b}E\xi$.
- $E\xi=\bar{x}$, $E\eta=\bar{y}$. $\mathrm{var}(\xi)=\frac{1}{n}\ell_{xx}$, $\mathrm{var}(\eta)=\frac{1}{n}\ell_{yy}$,
  $\mathrm{cov}(\xi,\eta)=\frac{1}{n}\ell_{xy}$, $\rho_{\xi,\eta}=\frac{\ell_{xy}}{\sqrt{\ell_{xx}\ell_{yy}}}=:r$ 样本相关系数(2.15).
- $Q(\hat{a},\hat{b})=nE(\tilde{\eta}-\hat{b}\tilde{\xi})^2=n\mathrm{var}(\eta)(1-\rho_{\xi,\eta}^2)=\ell_{yy}(1-r^2)$.
- $Q(\hat{a},\hat{b})=Q$ 残差平方和, $r^2=1-Q/\ell_{yy}$ (2.20).
  正交分解$\stackrel{b=0}{\longrightarrow}$ $\ell_{yy}=Q+U$ (引理2.1), U: 回归平方和.

正交分解:

$\hat{a} = \bar{y} - \hat{b}\bar{x}$, $\hat{b} = \frac{\ell_{xy}}{\ell_{xx}}$. 回归直线估计$\hat{f}$: $\hat{f}(x) = \hat{a} + \hat{b}x$,

- $\hat{f}$ 过点$(\bar{x}, \bar{y})$, 即$\bar{y} = \hat{f}(\bar{x})$.
- <u>残差</u>平方和: $Q = \sum_{i=1}^{n}(\underline{y_i - \hat{y}_i})^2$, 其中$\hat{y}_i = \hat{f}(x_i) = \hat{a} + \hat{b}x_i$.
  回归平方和: $U = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.
- **引理2.1.** $\ell_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = U + Q$.
  - 正交分解: $\tilde{\eta} = (\tilde{\eta} - \hat{b}\tilde{\xi}) \oplus \hat{b}\tilde{\xi}$, $\eta \to y_i$, $\xi \to x_i$.
    $E\tilde{\eta}^2 = E(\tilde{\eta} - \hat{b}\tilde{\xi})^2 + E(\hat{b}\tilde{\xi})^2$.
  - $\tilde{\eta} = \eta - E\eta \to y_i - \bar{y} \to \ell_{yy}$.
    $\hat{b}\tilde{\xi} \to \hat{b}(x_i - \bar{x}) = \hat{f}(x_i) - \hat{f}(\bar{x}) = \hat{y}_i - \bar{y} \to U$,
    $\tilde{\eta} - \hat{b}\tilde{\xi} \to y_i - \bar{y} - (\hat{y}_i - \bar{y}) = y_i - \hat{y}_i \to Q$.

无偏估计: $E\hat{b} = b,\ E\hat{a} = a$

- $\hat{a} = \bar{y} - \hat{b}\bar{x},\ \hat{b} = \frac{\ell_{xy}}{\ell_{xx}} = \frac{1}{\ell_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$.
- $\hat{a},\ \hat{b}$ 是$(y_1, \cdots, y_n)$ 的线性 函数.
- $y_i = a + bx_i + e_i, \quad \bar{y} = a + b\bar{x} + \bar{e}$,

  $y_i - \bar{y} = b(x_i - \bar{x}) + (e_i - \bar{e})$.
- $\hat{b} = b + \frac{1}{\ell_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) e_i, \quad (\because \sum_{i=1}^{n} (x_i - \bar{x}) \bar{e} = 0)$

  其中$e_1, \cdots, e_n \sim$ i.i.d. $N(0, \sigma^2)$,
- $\hat{a} = \bar{y} - \hat{b}\bar{x} = (a + b\bar{x} + \bar{e}) - \hat{b}\bar{x} = a + (b - \hat{b})\bar{x} + \bar{e}$.
- 无偏: $E\hat{b} = b,\ E\hat{a} = a$.
- **定理2.2.** 假设$x_i,\ i = 1, \cdots, n$ 不全相等.

  (关心f, 即, 当x 变化时, y 如何跟着变化).

  那么, $\hat{a},\ \hat{b}$ 是最优线性无偏估计.

统计量计算:

- 最基本统计量: l_{xx}, l_{yy}, l_{xy}
- 其他统计量通过基本统计量计算得到:

$$\hat{b} = l_{xy}/l_{xx},$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x},$$

$$U = \hat{b}l_{xy},$$

$$Q = l_{yy} - U,$$

$$r^2 = U/l_{yy} = 1 - Q/l_{yy}.$$

残差平方和. $Q=\sum_{i=1}^n(y_i-\hat{y}_i)^2$. 则$\frac{1}{\sigma^2}Q\sim\chi^2(n-2)$.

- $\hat{b}=b+\frac{1}{\ell_{xx}}\sum_i\underline{(x_i-\bar{x})(e_i-\bar{e})}=b+\frac{1}{\ell_{xx}}\sum_i(x_i-\bar{x})e_i$,
  $\hat{a}=a+(b-\hat{b})\bar{x}+\bar{e}$.
- $\hat{y}_i-y_i=(\hat{a}+\hat{b}x_i)-(a+bx_i+e_i)=(b-\hat{b})\bar{x}+\bar{e}+(\hat{b}-b)x_i-e_i$
  $\ldots=(\hat{b}-b)(x_i-\bar{x})-(e_i-\bar{e})$.
  $(\ldots)^2=(\hat{b}-b)^2(x_i-\bar{x})^2+(e_i-\bar{e})^2-2(\hat{b}-b)\underline{(x_i-\bar{x})(e_i-\bar{e})}$.
- $Q=(\hat{b}-b)^2\ell_{xx}+\sum_i(e_i-\bar{e})^2-2(\hat{b}-b)\underline{\ell_{xx}(\hat{b}-b)}$,
  $Q=\sum_i(e_i-\bar{e})^2-\ell_{xx}(\hat{b}-b)^2$.
- $\sum_i(e_i-\bar{e})^2=\sum_i e_i^2-[\sum_i\frac{1}{\sqrt{n}}e_i]^2=\underline{\sum_i e_i^2}-[\sum_i a_{1i}e_i]^2$.
- $\ell_{xx}(\hat{b}-b)^2=(\sum_i\frac{x_i-\bar{x}}{\sqrt{\ell_{xx}}}e_i)^2=(\sum_i a_{2i}e_i)^2$.
- $\sum_i a_{1i}^2=\sum_i a_{2i}^2=1$, $\sum_i a_{1i}a_{2i}=0$. 补行得正交矩阵A.
  $(Z_1,\cdots,Z_n)^T=A(e_1,\cdots,e_n)^T\overset{d}{=}(e_1,\cdots,e_n)^T$.
- $Q=\underline{\sum_i Z_i^2}-Z_1^2-Z_2^2=\sum_{i=3}^n Z_i^2$.

回归平方和: $U = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.

- $\hat{y}_i = \hat{f}(x_i)$, $\bar{y} = \hat{f}(\bar{x})$. $\hat{y}_i - \bar{y} = \hat{b}(x_i - \bar{x})$.
  $\hat{b} = b + \frac{1}{\ell_{xx}}\sum_i (x_i - \bar{x})e_i$.
- $U = \hat{b}^2\ell_{xx} = [\sqrt{\ell_{xx}}b + \sum_{i=1}^{n} a_{2i}e_i]^2 = (\sqrt{\ell_{xx}}b + Z_2)^2$.
- 若$b = 0$, 则$\frac{1}{\sigma^2}U = Z^2 \sim \chi^2(1)$, 其中$Z \sim N(0,1)$.
- 若$b \neq 0$, $\frac{1}{\sigma^2}U = (\frac{1}{\sigma}\sqrt{\ell_{xx}}b + Z)^2$.
- 若$Z \sim N(0,1)$, 则$P(|Z + c| \leqslant l) \leqslant P(|Z| \leqslant l)$.
  $U_b := (\sqrt{\ell_{xx}}b + Z_2)^2$, $U_0 := {Z_2}^2$
  $P(U_b \leqslant l^2) \leqslant P(U_0 \leqslant l^2)$ vs $U_b \leqslant l^2 \Rightarrow U_0 \leqslant l^2$, $\forall l$.
  在某种意义下, $U_0 \leqslant U_b$.