4.6 HW 6

作业 6 链接

注 在做回归分析作业，考虑随机性时可以忽略样本自变量 x 带来的随机性，计算期望、方差时均可以看作是固定 x 以后的条件期望、条件方差。这会让许多概念变得更清楚。

注 事实上，在很多时候，x 的随机性也可以考虑到模型中。这时我们通常的做法是先固定 x 求出条件分布、条件期望等等，再考虑 x 的随机性对于结果的影响。例如讲义 *week 06 P20* 的几个结论，在不考虑 x 带来的随机性，即都看作对 x 取条件的版本，那证明都是比较容易的。**其实从这里也可以更进一步推出，当 x 具有随机性时对应的结论。(思考)**

✍ **练习 4.1** 简单随机样本 $y_1, \ldots, y_n$ iid $\sim \left(\mu, \sigma^2\right)$ 是最简单 (不含自变量) 的线性模型, 这是因为该模型可写为:

$$y_i = \mu + \epsilon_i, i = 1, \ldots, n, \quad \epsilon_i, i = 1, \ldots, n \text{ iid } \sim \left(0, \sigma^2\right)$$

其中 μ 是未知参数。使得误差平方和 $\sum_{i=1}^{n}\left(y_i - \mu\right)^2$ 最小的 μ 称为最小二乘估计, 记为 $\hat{\mu}$ 。求 $\hat{\mu}$ 及其方差。y_i 的拟合值 $\hat{y}_i$ 和残差 e_i 应如何定义?

注 建议回顾一下拟合值和残差的定义。拟合值是根据模型和数据可以给出的数值，而扰动项 ε_i 是无法观测的。能利用的数据只有 $y_i, i = 1, \ldots, n$.

解 误差平方和对 μ 求导等于 0, 解得 $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. $var(\bar{y}) = \sigma^2/n$.

拟合值为 $\hat{\mu} = \bar{y}$, 残差为 $y_i - \bar{y}$.

✍ **练习 4.2** 假设随机样本数据 $\left(x_i, y_i\right), i = 1, \ldots, n$ 满足下述过原点的线性回归模型 (无截距项)

$$y_i = bx_i + \epsilon_i, E\left(\epsilon_i\right) = 0, \operatorname{var}\left(\epsilon_i\right) = \sigma^2, \text{且} x_i \text{与} \epsilon_i \text{独立}, i = 1, \ldots, n.$$

通过极小化误差平方和 $\sum_{i=1}^{n}\left(y_i - bx_i\right)^2$, 求未知参数 b 的 LS 估计 $\widehat{b}$, 并求其方差。

解 同上题。误差平方和对 b 等于 0, 解得 $\hat{b} = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2$.$var(\hat{b}) = \sigma^2 / \sum_{i=1}^{n} x_i^2$.

✍ **练习 4.3** 证明给定 $\mathbf{x} = \left(x_1, \ldots, x_n\right)^\top$ 条件下

$$\operatorname{var}(\hat{a}) = \sigma^2/n + \bar{x}^2\sigma^2/s_{xx}, \quad \operatorname{var}(\hat{b}) = \sigma^2/s_{xx}, \quad \operatorname{cov}(\hat{a}, \hat{b}) = -\bar{x}\sigma^2/s_{xx}$$

何时 $\hat{a}, \hat{b}$ 不相关?

证明

$$\begin{aligned}
var(\hat{b}) &= var\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \\
&= var\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 var(y_i)}{s_{xx}^2} \\
&= \sigma^2/s_{xx}, \\
var(\hat{a}) &= var(\bar{y} - \hat{b}\bar{x}) \\
&= var\left(\frac{1}{n}\sum_{i=1}^{n} y_i - \hat{b}\bar{x}\right) \\
&= var\left(\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\bar{x}\right) \\
&= var\left(\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{s_{xx}}\right) y_i\right) \\
&= \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{s_{xx}}\right)^2 var(y_i) \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2\bar{x}^2}{s_{xx}},
\end{aligned}$$

$$
\begin{aligned}
cov(\hat{a},\hat{b}) &= cov(\bar{y}-\hat{b}\bar{x},\hat{b}) \\
&= cov\left(\bar{y},\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)-\bar{x}var(\hat{b}) \\
&= cov\left(\frac{1}{n}\sum_{i=1}^{n}y_i,\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)-\bar{x}var(\hat{b}) \\
&= \sum_{i=1}^{n}\frac{1}{n}\frac{(x_i-\bar{x})}{s_{xx}}var(y_i)-\bar{x}var(\hat{b}) \\
&= 0-\frac{\bar{x}\sigma^2}{s_{xx}}=-\frac{\bar{x}\sigma^2}{s_{xx}}.
\end{aligned}
$$

$\hat{a},\hat{b}$ 不相关当且仅当 $cov(\hat{a},\hat{b})=0$, 即 $\bar{x}=0$ 或 $\sigma=0$.

✍ **练习 4.4** 假设 $x_i=0$ 或 $1, i=1,\ldots,n$, 记 $m=\sum x_i$ 为 1 的个数。当 m 等于或近似等于 $n/2$ 时, 称为是均衡设计 (balanced design), 否则是不均衡的 (unbalanced)。基于 b 的 LS 估计的方差, 讨论均衡设计的优越性。

解 $var(\hat{b})=\sigma^2/s_{xx}$. 均衡设计使 LS 估计参数 $\hat{b}$ 的方差达到最小。

✍ **练习 4.5** 定义最小二乘得到的残差 $e_i=y_i-\hat{y}_i=y_i-\hat{a}-\hat{b}x_i$, 残差平方和定义为 $RSS=\sum_{i=1}^{n}e_i^2$ 。证明 $e_i=(\epsilon_i-\bar{\epsilon})-(x_i-\bar{x})\,s_{x\epsilon}/s_{xx}$, 并利用该表达证明

$$RSS=s_{yy}-s_{xy}^2/s_{xx}=s_{\epsilon\epsilon}-s_{x\epsilon}^2/s_{xx}.$$

证明 证明参考课件 week 06 P16 等.

✍ **练习 4.6** 定义点 (x_i,y_i) 与 $(\bar{x},\bar{y})$ 决定的直线的斜率为 $k_i=\frac{y_i-\bar{y}}{x_i-\bar{x}}$, 当 $x_i=\bar{x}$ 时定义 $k_i=0, i=1,\ldots,n$ 。则 b 的最小二乘估计 $\hat{b}=s_{xy}/s_{xx}$ 可表示成所有斜率 $k_i, i=1,\ldots,n$ 的加权和

$$\hat{b}=\frac{1}{s_{xx}}\sum(x_i-\bar{x})(y_i-\bar{y})=\frac{1}{s_{xx}}\sum(x_i-\bar{x})^2\left(\frac{y_i-\bar{y}}{x_i-\bar{x}}\right)=\sum w_{0i}k_i$$

其中权重 $w_{0i}=(x_i-\bar{x})^2/s_{xx}, \sum w_{0i}=1$ 。对任何序列 $w_1,\ldots,w_n$ (只依赖于 $x_1,\ldots,x_n$) , $w_i\geq 0, \sum w_i=1$, 定义 b 的一个估计

$$\tilde{b}=\sum w_ik_i=\sum w_i\left(\frac{y_i-\bar{y}}{x_i-\bar{x}}\right).$$

(a) 证明 $E(\tilde{b})=b$.

(b) 求 $\mathrm{var}(\tilde{b}\mid\mathbf{x})$.

(c) 证明 $\mathrm{var}(\tilde{b}\mid\mathbf{x})\geq\sigma^2/s_{xx}=\mathrm{var}(\hat{b}\mid\mathbf{x})$.

证明

(a)

$$
\begin{aligned}
E(\tilde{b}) &= E(E(\tilde{b}\mid x)) \\
&= E\left(\sum_{i=1}^{n}\frac{w_i}{x_i-\bar{x}}E(y_i-\bar{y}\mid x)\right) \\
&= E\left(\sum_{i=1}^{n}\frac{w_i}{x_i-\bar{x}}(b(x_i-\bar{x}))\right) \\
&= bE\left(\sum_{i=1}^{n}w_i\right) \\
&= b
\end{aligned}
$$

(b) 注意到

$$\sum_{i=1}^{n}a_i(b_i-\bar{b})=\sum_{i=1}^{n}(a_i-\bar{a})b_i$$

$$
\begin{aligned}
\operatorname{Var}(\tilde{b} \mid x) &= \operatorname{Var}\left(\sum_{i=1}^{n} \frac{w_i}{x_i - \bar{x}} (y_i - \bar{y}) \mid x\right) \\
&= \operatorname{Var}\left(\sum_{i=1}^{n} c_i (y_i - \bar{y}) \mid x\right) \qquad \text{记} c_i = \frac{w_i}{x_i - \bar{x}} \\
&= \operatorname{Var}\left(\sum_{i=1}^{n} (c_i - \bar{c}) y_i \mid x\right) \\
&= \sum_{i=1}^{n} (c_i - \bar{c})^2 \sigma^2 \\
&= \left(\sum_{i=1}^{n} c_i^2 - n\bar{c}^2\right) \sigma^2 \\
&= \left(\sum_{i=1}^{n} \frac{w_i^2}{(x_i - \bar{x})^2} - \frac{1}{n}\left(\sum_{i=1}^{n} \frac{w_i}{x_i - \bar{x}}\right)^2\right) \sigma^2
\end{aligned}
$$

(c)

$$
\operatorname{Var}(\tilde{b} \mid x) = \sum_{i=1}^{n} (c_i - \bar{c})^2 \sigma^2
$$

由 Cauchy - Schwarz 不等式

$$
\sum_{i=1}^{n} (c_i - \bar{c})^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \geqslant \left(\sum_{i=1}^{n} (c_i - \bar{c})(x_i - \bar{x})\right)^2
$$

$$
\sum_{i=1}^{n} (c_i - \bar{c})^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\operatorname{Var}(\tilde{b} \mid x) s_{xx}}{\sigma^2}
$$

$$
\begin{aligned}
\sum_{i=1}^{n} (c_i - \bar{c})(x_i - \bar{x}) &= \sum_{i=1}^{n} c_i (x_i - \bar{x}) \\
&= \sum_{i=1}^{n} \omega_i \\
&= 1
\end{aligned}
$$

$$
\operatorname{Var}(\tilde{b} \mid x) \geqslant \frac{\sigma^2}{s_{xx}}
$$

注 本题问题多出在方差的计算上，公式

$$
\operatorname{Var}(X_1 + \cdots + X_n) = \operatorname{Var}(X_1) + \cdots + \operatorname{Var}(X_n)
$$

只能用于独立的随机变量。而形如 $y_i - \bar{y}\ (i = 1, \ldots, n)$ 的随机变量列并不是相互独立的，应该拆成独立的部分计算。

注 另外，本题实质上是简单形式的 *Gauss-Markov* 定理。该定理可以表述为：若 $\tilde{b} = \Sigma_{i=1}^{n} w_i y_i$, ($\{w_i\}_{i=1}^{n}$ 只与 x 有关) 是 b 的无偏估计，首先我们知道通过选取适当的 w $\tilde{b}$ 可以取为最小二乘估计 $\hat{b}$，并且最小二乘估计 $\hat{b}$ 是对应不同 w 的无偏估计 $\tilde{b}$ 中，方差最小的一个。