数据挖掘 Homework 2

蔡育铮 21210980103

问题 1

证明 $E\left(y_0-\hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right)+\left[Bias\left(\hat{f}(x_0)\right)\right]^2+Var(\epsilon)$

证明.

$$
\begin{aligned}
E\left(y_0-\hat{f}(x_0)\right)^2 &= E\left[(y_0-f(x_0))+\left(f(x_0)-\hat{f}(x_0)\right)\right]^2 \\
&= E\left(y_0-f(x_0)\right)^2+2E\left[(y_0-f(x_0))\left(f(x_0)-\hat{f}(x_0)\right)\right]+E\left(f(x_0)-\hat{f}(x_0)\right)^2
\end{aligned}
$$

由于

$$
\begin{aligned}
E\left(y_0-f(x_0)\right)^2 &= Var(\epsilon) \\
E\left[(y_0-f(x_0))\left(f(x_0)-\hat{f}(x_0)\right)\right] &= 0 \\
E\left(f(x_0)-\hat{f}(x_0)\right)^2 &= Var\left(\hat{f}(x_0)\right)+\left[E\left(f(x_0)-\hat{f}(x_0)\right)\right]^2 \\
&= Var\left(\hat{f}(x_0)\right)+\left[Bias\left(\hat{f}(x_0)\right)\right]^2
\end{aligned}
$$

因此

$$
E\left(y_0-\hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right)+\left[Bias\left(\hat{f}(x_0)\right)\right]^2+Var(\epsilon)
$$

□

问题 2

试证明：二元线性回归模型

$$
Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\mu_i
$$

中变量 X_1 与 X_2 的参数的普通最小二乘估计可以写成

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\left(\sum y_ix_{i1}\right)\left(\sum x_{i2}^2\right)-\left(\sum y_ix_{i2}\right)\left(\sum x_{i1}x_{i2}\right)}{\sum x_{i1}^2\sum x_{i2}^2(1-r^2)} \\
\hat{\beta}_2 &= \frac{\left(\sum y_ix_{i2}\right)\left(\sum x_{i1}^2\right)-\left(\sum y_ix_{i1}\right)\left(\sum x_{i1}x_{i2}\right)}{\sum x_{i1}^2\sum x_{i2}^2(1-r^2)}
\end{aligned}
$$

其中, r 为 X_1 与 X_2 的相关系数。讨论 r 等于或接近于 1 时, 该模型的估计问题。

证明. 记：

$$
\begin{aligned}
X_1 &= [x_{11}, x_{21}, \cdots, x_{n1}]^T \in \mathbb{R}^{n\times 1}\\
X_2 &= [x_{12}, x_{22}, \cdots, x_{n2}]^T \in \mathbb{R}^{n\times 1}\\
Y &= [y_1, y_2, \cdots, y_n]^T \in \mathbb{R}^{n\times 1}\\
X &= \left[\mathbf{1}^T, X_1, X_2\right] \in \mathbb{R}^{n\times 3}\\
\beta &= [\beta_0, \beta_1, \beta_2]^T \in \mathbb{R}^{3\times 1}
\end{aligned}
$$

则最小二乘估计为求解如下优化问题：

$$\min_{\beta} (Y - X\beta)^T (Y - X\beta)$$

该问题的解满足：

$$(X^TX)\beta = X^TY$$

等价于如下形式：

$$
\begin{aligned}
\beta_0 &= \bar{Y} - \beta_1\bar{X}_1 - \beta_2\bar{X}_2\\
X_1^TY &= n\bar{X}_1\beta_0 + \beta_1 X_1^TX_1 - \beta_2 X_1^TX_2\\
X_2^TY &= n\bar{X}_2\beta_0 + \beta_1 X_1^TX_2 - \beta_2 X_2^TX_2
\end{aligned}
$$

假设自变量 X_1 与 X_2 均已经过中心化处理，即 $\bar{X}_1 = \bar{X}_2 = 0$（否则不能得到题目中的表达式），则有：

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\left(X_1^TY\right)\left(X_2^TX_2\right) - \left(X_1^TX_2\right)\left(X_2^TY\right)}{\left(X_1^TX_1\right)\left(X_2^TX_2\right) - \left(X_1^TX_2\right)^2}\\
\hat{\beta}_2 &= \frac{\left(X_2^TY\right)\left(X_1^TX_1\right) - \left(X_1^TX_2\right)\left(X_1^TY\right)}{\left(X_1^TX_1\right)\left(X_2^TX_2\right) - \left(X_1^TX_2\right)^2}
\end{aligned}
$$

注意到

$$
\begin{aligned}
Var(X_1) &= \frac{X_1^TX_1}{n} - \frac{\left(\bar{X}_1\right)^2}{n^2} = \frac{X_1^TX_1}{n}\\
Var(X_2) &= \frac{X_2^TX_2}{n} - \frac{\left(\bar{X}_2\right)^2}{n^2} = \frac{X_2^TX_2}{n}\\
Cov(X_1, X_2) &= \frac{X_1^TX_2}{n} - \frac{\bar{X}_1\bar{X}_2}{n^2} = \frac{X_1^TX_2}{n}\\
r^2 &= \frac{Cov(X_1, X_2)^2}{Var(X_1)Var(X_2)} = \frac{\left(X_1^TX_2\right)^2}{\left(X_1^TX_1\right)\left(X_2^TX_2\right)}
\end{aligned}
$$

因此

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\left(X_1^TY\right)\left(X_2^TX_2\right) - \left(X_1^TX_2\right)\left(X_2^TY\right)}{\left(X_1^TX_1\right)\left(X_2^TX_2\right)(1 - r^2)} = \frac{\left(\sum y_i x_{i1}\right)\left(\sum x_{i2}^2\right) - \left(\sum y_i x_{i2}\right)\left(\sum x_{i1}x_{i2}\right)}{\sum x_{i1}^2 \sum x_{i2}^2 (1 - r^2)}\\
\hat{\beta}_2 &= \frac{\left(X_2^TY\right)\left(X_1^TX_1\right) - \left(X_1^TX_2\right)\left(X_1^TY\right)}{\left(X_1^TX_1\right)\left(X_2^TX_2\right)(1 - r^2)} = \frac{\left(\sum y_i x_{i2}\right)\left(\sum x_{i1}^2\right) - \left(\sum y_i x_{i1}\right)\left(\sum x_{i1}x_{i2}\right)}{\sum x_{i1}^2 \sum x_{i2}^2 (1 - r^2)}
\end{aligned}
$$

□

当 r 等于或接近 1 时，$\hat{\beta}_1$ 和 $\hat{\beta}_2$ 表达式的分母等于或接近于 0，这将导致计算出的参数有可能极大或极小。并且也会导致模型不稳定，因为当训练数据稍微变动时，就可能引起参数的巨大变化。

问题 3

对一元回归模型

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

假如其他基本假设全部满足，但 $Var(\mu_i) = \sigma_i^2 \neq \sigma^2$，试证明估计的斜率项仍是无偏的, 但方差变为

$$Var(\tilde{\beta}_1) = \frac{\sum x_i^2 \sigma_i^2}{\left(\sum x_i^2\right)^2}$$

证明. 记：

$$\begin{aligned}
\mathbf{x} &= [x_1, x_2, \cdots, x_n]^T \in \mathbb{R}^{n\times 1} \\
Y &= [y_1, y_2, \cdots, y_n]^T \in \mathbb{R}^{n\times 1} \\
X &= \left[\mathbf{1}^T, \mathbf{x}\right] \in \mathbb{R}^{n\times 2} \\
\beta &= [\beta_0, \beta_1]^T \in \mathbb{R}^{2\times 1}
\end{aligned}$$

则最小二乘估计为求解如下优化问题：

$$\min_\beta (Y - X\beta)^T (Y - X\beta)$$

该问题的解满足：

$$(X^T X)\beta = X^T Y$$

等价于如下形式：

$$\begin{aligned}
\beta_0 &= \bar{Y} - \beta_1 \bar{\mathbf{x}} \\
\mathbf{x}^T Y &= n\bar{X}_1 \beta_0 + \beta_1 \mathbf{x}^T \mathbf{x}
\end{aligned}$$

假设自变量 X_1 已经过中心化处理，即 $\bar{X}_1 = 0$（否则不可能得到题目中的表达式），则估计的斜率项：

$$\tilde{\beta}_1 = \frac{\mathbf{x}^T Y}{\mathbf{x}^T \mathbf{x}} = \frac{\sum x_i y_i}{\sum x_i^2}$$

其期望满足：

$$E\left(\tilde{\beta}_1\right) = \frac{E(\sum x_i y_i)}{\sum x_i^2} = \frac{\sum x_i E(y_i)}{\sum x_i^2} = \frac{\sum x_i E(x_i \beta_1 + \mu_i)}{\sum x_i^2} = \beta_1$$

因此估计的斜率项是无偏的。其方差为：

$$Var\left(\tilde{\beta}_1\right) = \frac{Var(\sum x_i y_i)}{\sum x_i^2} = \frac{\sum x_i^2 Var(y_i)}{\sum x_i^2} = \frac{\sum x_i \sigma_i^2}{\sum x_i^2}$$

□