Statistical Learning

Homework #2

Due on March 22, 2022

Shi Yuchen 21210980116

Problem 1

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

Solution

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

Problem 2

试证明: 二元线性回归模型

$$Y_i=\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\mu_i$$

中变量 X_1 与 X_2 的参数的普通最小二乘估计可以写成

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

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。

Solution

$$L=\sum(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}-y_i)^2$$

to minimize the loss,

$$
\begin{aligned}
\sum 2(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}-y_i) &= 0 \\
\sum 2(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}-y_i)x_{i1} &= 0 \\
\sum 2(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}-y_i)x_{i2} &= 0
\end{aligned}
$$

we get $\beta_0=\frac{1}{n}\sum(y_i-\beta_1 x_{i1}-\beta_2 x_{i2})$, plug into above equation:

$$
\begin{aligned}
\beta_1\sum x_{i1}^2+\beta_2\sum x_{i1}x_{i2} &= \sum y_i x_{i1} \\
\beta_1\sum x_{i1}x_{i2}+\beta_2\sum x_{i2}^2 &= \sum y_i x_{i2}
\end{aligned}
$$

and we can easily get

$$\beta_1 = \left((\sum y_i x_{i1})(\sum x_{i2}^2) - (\sum y_i x_{i2})(\sum x_{i1}x_{i2})\right)/c$$
$$\beta_2 = \left((\sum y_i x_{i2})(\sum x_{i1}^2) - (\sum y_i x_{i1})(\sum x_{i1}x_{i2})\right)/c$$

where $c = (\sum x_{i1}^2)(\sum x_{i2}^2) - (\sum x_{i1}x_{i2})^2 = \sum x_{i1}^2 \sum x_{i2}^2(1-r^2)$
When $|r|$ is close to 1, the denominator is close to 0, so the estimator is unstable.

Problem 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 $\mathrm{Var}(\mu_i) = \sigma_i^2 \neq \sigma^2$, 试证明估计的斜率项仍是无偏的, 但方差变为

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

Solution

The estimator is

$$\hat{\beta}_1 = (n\sum x_i y_i - \sum x_i \sum y_i)/(n\sum x_i^2 - (\sum x_i)^2)$$

To show the estimator is unbiased, we can equvilently prove

$$E[(n\sum x_i y_i - \sum x_i \sum y_i) - \beta_1(n\sum x_i^2 - (\sum x_i)^2)] = 0$$

plug $y_i = \beta_0 + \beta_1 x_i + \mu_i$ into the left part:

$$\begin{aligned} E[n\beta_0 \sum x_i + n\beta_1 \sum x_i^2 + n\sum x_i\mu_i - (n\beta_0 \sum x_i + \beta_1(\sum x_i)^2 + \sum x_i \sum \mu_i) \\ -\beta_1(n\sum x_i^2 - (\sum x_i)^2)] \\ = E[n\sum x_i \mu_i - \sum x_i \sum \mu_i] = n\sum E[x_i(\mu_i - \bar{\mu})] = 0 \end{aligned}$$

the last equation use the fact that $E[XY] = E[X]E[Y]$ when X and Y are independent.
Since

$$\begin{aligned} \hat{\beta}_1 &= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \\ &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} \end{aligned}$$

the variance is

$$\begin{aligned} Var(\hat{\beta}_1) &= \frac{\sum(x_i - \bar{x})^2 Var(y_i)}{(\sum(x_i - \bar{x})^2)^2} \\ &= \frac{\sum(x_i - \bar{x})^2 \sigma_i^2}{(\sum(x_i - \bar{x})^2)^2} \end{aligned}$$