Math 40 Probability and Statistical Inference
Winter 2021
Lecture 18 Linear Estimation (6.5)

Estimation of Variance and Correlation Coefficient (6.6)

Yoonsang Lee (yoonsang.lee@dartmouth.edu)

6.5 Linear Estimation

Let $\{Y_i\}$ IID from a distribution with unknown mean α and variance σ^2. That is,

$$Y_i = \alpha + \epsilon_i, \quad i = 1, 2, ..., n.$$

Note that $Var(\epsilon) = \sigma^2$. Using $\{Y_i\}$, can you estimate the unknown mean α?

- Our most intuitive answer would be
$$\hat{\alpha} = \overline{Y} = \frac{1}{n}\sum_i^n Y_i,$$
the sample mean.
- The sample mean is unbiased (section 6.4.1).
- In fact, the sample mean has the smallest MSE (theorem 6.35).

Now we have a data set $\{(x_i, Y_i)\}_i^n$.

- We assume a linear relation between x_i and Y_i

$$Y_i = \beta x_i + \epsilon_i$$

  where ϵ_i has mean zero and variance σ^2.

- Also, they are independent (my assumption is stronger than the one in your textbook).

- We want to estimate β using the data $\{(x_i, Y_i)\}_i^n$.

- Example 6.36 (a) explains how to derive an unbiased estimate of β,

$$\hat{\beta} = \frac{\sum_i^n x_i Y_i}{\sum_i^n x_i^2}$$

  by assuming a linear combination of the data points Y_i,

$$\hat{\beta} = \sum_i^n \lambda_i Y_i.$$

- In the derivation, it uses the idea of the Lagrange multiplier (vector calculus) to minimize MSE.

- Example 6.36 (b) shows that this estimator can be found by minimizing the residual sum of squares

$$RSS = \sum_i^n (Y_i - \beta x_i)^2.$$

- There are other unbiased estimators of β,

$$\hat{\beta}_1 = \frac{1}{n}\sum_i^n \frac{Y_i}{x_i},$$

$$\hat{\beta}_2 = \frac{\sum_i^n Y_i}{\sum_i^n x_i}.$$

- However, they are not optimal (that is, MSEs are larger than the on in Example 6.36). See Example 6.37 for details.

Example 6.38 The instructor gives students a series of n assignments. The maximum number of points in the i-th assignment is x_i. Suppose that the i-th student gains Y_i points in the i-th assignment ($Y_i \leq x_i$). To rank the student in the class, the instructor wants a metric for student' performance by finding the ratio of the number of points received to the maximum number of points. Find an unbiased estimator of the ratio.

Solution The problem asks the coefficient β when

$$Y_i = \beta x_i + \epsilon_i.$$

We have at least three unbiased estimators,

$$\hat{\beta} = \frac{\sum_i^n x_i Y_i}{\sum_i^n x_i^2}, \quad \hat{\beta}_1 = \frac{1}{n}\sum_i^n \frac{Y_i}{x_i}, \quad \hat{\beta}_2 = \frac{\sum_i^n Y_i}{\sum_i^n x_i}.$$

If you are interested in the minimum MSE estimator, choose the first one.

6.6 Estimation of Variance and Correlation Coefficient

6.6.1 Quadratic estimation of the variance

Let $\{Y_i\}_i^n$ be IID from $\mu, \sigma^\in$; μ and σ^2 are unknown.

- From section 6.4, we know that the sample variance

  $$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (Y_i - \overline{Y})^2$$

  is unbiased.

- It is also optimal in the sense that MSE is minimized (theorem 6.41).

- The estimator is also consistent (the variance of the estimator is $\frac{2\sigma^4}{n-1}$, which converges to 0 as $n \to \infty$).

Example 6.43 Let $\{r_i\}_i^n$ be n independent measurements of the radius of a circle, ρ. How do we estimate the area of the circle?

- First estimator, $\hat{A}_1 = \frac{1}{n}\sum_i^n \pi r_i^2$, the mean of the sample area.
- $\hat{A}_1$ is biased as $E(r_i^2) = Var(r_i) + \rho^2$.
- Another estimator, $\hat{A}_2 = \pi \overline{r}^2$ where $\overline{r} = \frac{1}{n}\sum_i^n r_i$.
- This one is also biased,

$$E(\overline{r}^2) = Var(\overline{r}) + E(\overline{r})^2 = \frac{Var(r_i)}{n} + \rho^2$$

- However, its bias converges to 0 as $n \to \infty$. That is, it is asymptotically unbiased.

6.6.2 Estimation of the covariance and correlation coefficient

Let $\{(X_i, Y_i)\}_i^n$ be IID from a normal distribution with unknown mean $\boldsymbol{\mu}$ and covariance matrix variance $\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$.

- As in the variance estimation, the sample covariance

$$\hat{\sigma}_{xy} = \frac{1}{n-1} \sum_i^n (X_i - \overline{X})(Y_i - \overline{Y}),$$

  is an unbiased estimator of the population covariance σ_{xy} (theorem 6.44).

- R command `cov` calculate the sample covariance.

  Check R code `6_6_Thm6.44.R` on Canvas.

- The correlation coefficient is $\rho = \frac{\sigma_{xy}}{\sigma_x}\sigma_y$. We use sample variances and covariance to estimate the correlation coefficient

$$r = \frac{\sum_i^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{(\sum_i^n (X_i - \overline{X}))(\sum_i^j (Y_j - \overline{Y}))}}$$

- R command `cor` calculate the sample (Pearson) correlation coefficient.

- **Note** The Pearson correlation coefficient is **biased**.

6.7 Least squares for simple linear regression

Let $\{(X_i, Y_i\}_i^n$ be IID from an unknown distribution. We are interested in a linear model between X and Y

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, ..., n,$$

where x_i is a specified value of X_i. $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.

- The ordinary least squares (ODS) estimators of the intercept (α) and the slope (β) in the simple linear regression minimize the residual sum of squares (RSS)

$$RSS(\alpha, \beta) = \sum_i^n (Y_i - \alpha - \beta X_i)^2.$$

- In Chapter 3, we learned that the regression is the conditional expectation, which minimizes $E((Y - r(x))^2)$.
- RSS is related to $E((Y - r(x))^2)$.
- Thus, the ODS estimators give us the conditional expectation in a linear form.
- Using Calculus,

$$\hat{\beta} = \frac{\sum_i^n (x_i - \overline{X})(Y_i - \overline{Y})}{\sum_i^n (x_i - \overline{x})^2}, \quad \hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}$$

- Theorem 6.50 says

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum_i^n (x_i - \overline{x})^2}$$

and

$$Var(\hat{\alpha}) = \frac{\sigma^2 \sum_i^n \frac{x_i^2}{n}}{\sum_i^n (x_i - \overline{x})^2}$$

- In particular, when ϵ_i is normal, we have (theorem 6.53)

  1. $\hat{\alpha}$ and $\hat{\beta}$ has the minimum MSE.
  2. $\hat{\alpha} \sim \mathcal{N}(\alpha, \frac{\sigma^2 \sum_i^n \frac{x_i^2}{n}}{\sum_i^n (x_i - \overline{x})^2})$ and $\hat{\beta} \sim \mathcal{N}(\beta, \frac{\sigma^2}{\sum_i^n (x_i - \overline{x})^2})$.
  3. The estimator $\hat{\sigma}^2 = \frac{1}{n-2} \sum_i^n (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$ is unbiased for σ^2 and the normalized sum of squares has a chi-square distribution

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

  4. $\hat{\beta}$ minus its true value divided by its standard error has a t-distribution

  $$\frac{\hat{\beta} - \beta}{\hat{\sigma}\sqrt{\sum_i^n (x_i - \overline{x})^2}} \sim t(n-2).$$

6.7.3 The `lm` function and prediction by linear regression

We will focus on how to use `lm` in R for linear regression and interpretation of the outputs.

- `lm.model = lm(Y~X)`
- `summary(lm.model)`
- or `names(lm.model)`
- For a confidence interval `confint(lm.model)`
- To make a prediction,

  `predict(lm.model, data.frame(X=c(1,2,3)))`

  Additional option:

  `interval="confidence"` for regression prediction or

  `interval="prediction"` for individual prediction
- `plot(X,Y)` to plot
- `abline(lm.model)` to add the regression line.
- Multiple R-squared

$$R^2 = 1 - \frac{\sum_i^n r_i^2}{\sum_i^n (Y_i - \overline{Y})^2},$$

- Adjusted R-squared

$$R_{adj}^2 = 1 - \frac{(n-1)}{(n-2)} \frac{\sum_i^n r_i^2}{\sum_i^n (Y_i - \overline{Y})^2}.$$