

*Lecture Notes*

# Statistics for Data Science

University of Pisa, Italy

Salvatore Ruggieri

April 15, 2022

## Contents

<b>1</b>	<b>On Cramér-Rao's bound and MLE</b>	<b>2</b>
<b>2</b>	<b>Least Square Estimators in Simple Linear Regression</b>	<b>4</b>
2.1	Expectation . . . . .	4
2.2	Variance and Standard Errors of the Coefficients . . . . .	4
2.3	Variance-Covariance Matrix . . . . .	5
2.4	Variance and Standard Errors of Fitted Values . . . . .	5
<b>3</b>	<b>Confidence Intervals for Simple Linear Regression</b>	<b>6</b>
3.1	Confidence Intervals of the Coefficients . . . . .	6
3.2	Confidence and Prediction Intervals of the Fitted Values . . . . .	6
3.3	Hypothesis Testing . . . . .	7
<b>4</b>	<b>Statistical Decision Theory</b>	<b>8</b>

# 1 On Cramér-Rao's bound and MLE

Consider the likelihood and log-likelihood functions:

$$L(\theta) = \prod_{i=1}^n f_{\theta}(X_i) \quad \ell(\theta) = \log L(\theta) = \sum_{i=1}^n \log f_{\theta}(X_i)$$

Since  $X_1, \dots, X_n$  are i.i.d., this is also true for  $Y_1 = \frac{\partial}{\partial \theta} \log f_{\theta}(X_1), \dots, Y_n = \frac{\partial}{\partial \theta} \log f_{\theta}(X_n)$ . The log-likelihood takes its maximum at the zero's of its derivative, which is called the *score function*:

$$S(\theta) = \frac{\partial}{\partial \theta} \ell(\theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_{\theta}(X_i) = \sum_{i=1}^n Y_i$$

The expectation of each  $Y_i$ 's is zero (use Leibniz integral rule):

$$\begin{aligned} \mathbb{E}[Y_i] &= \int \left( \frac{\partial}{\partial \theta} \log f_{\theta}(x) \right) f_{\theta}(x) dx = \int \frac{1}{f_{\theta}(x)} \left( \frac{\partial}{\partial \theta} f_{\theta}(x) \right) f_{\theta}(x) dx \\ &= \int \frac{\partial}{\partial \theta} f_{\theta}(x) dx = \frac{\partial}{\partial \theta} \int f_{\theta}(x) dx = \frac{\partial}{\partial \theta} 1 = 0 \end{aligned}$$

Hence, by linearity of expectation, we have:

$$\mathbb{E}[S(\theta)] = \sum_{i=1}^n \mathbb{E}[Y_i] = 0$$

The variance of  $S(\theta)$  is called the *Fisher information*, and it is the quantity:

$$I(\theta) = \mathbb{E}[S(\theta)^2]$$

It turns out<sup>1,2</sup> that:

$$\begin{aligned} I(\theta) = \mathbb{E}[S(\theta)^2] &= \mathbb{E}\left[\left(\sum_{i=1}^n Y_i\right)\left(\sum_{j=1}^n Y_j\right)\right] \\ &= \mathbb{E}\left[\sum_{i=1}^n Y_i^2 + \sum_{i=1}^n \sum_{j=1, j \neq i}^n Y_i Y_j\right] \\ &= \mathbb{E}\left[\sum_{i=1}^n Y_i^2\right] + \sum_{i=1}^n \sum_{j=1, j \neq i}^n \mathbb{E}[Y_i] \mathbb{E}[Y_j] \end{aligned} \tag{1}$$

$$= \mathbb{E}\left[\sum_{i=1}^n Y_i^2\right] + 0 \tag{2}$$

$$= \mathbb{E}\left[\sum_{i=1}^n \left(\frac{\partial}{\partial \theta} \log f_{\theta}(X_i)\right)^2\right]$$

$$= n \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log f_{\theta}(X)\right)^2\right] \tag{3}$$

where  $X \sim f_{\theta}$ . **Important:** some textbooks define  $I(\theta)$  using a single random variable, i.e., as  $\mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log f_{\theta}(X)\right)^2\right]$ . In such cases, it must be multiplied by  $n$  whenever it is used.

<sup>1</sup>(1) follows since  $\mathbb{E}[Y_i Y_j] = \mathbb{E}[Y_i] \mathbb{E}[Y_j]$  for independent  $Y_i, Y_j$ .

<sup>2</sup>(2) follows since  $\mathbb{E}[Y_i] = 0$ .

We can now link Fisher information to the Cramér-Rao inequality from [1, Remark 20.2]:

$$\text{Var}(T) \geq \frac{1}{n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log f_\theta(X)\right)^2\right]} \quad \text{for all } \theta,$$

by observing that, due to (3), the right-hand side is the inverse of  $I(\theta)$ , i.e.:

$$\text{Var}(T) \geq \frac{1}{n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log f_\theta(X)\right)^2\right]} = \frac{1}{I(\theta)} \quad \text{for all } \theta.$$

### Example

The textbook [1, pages 324-325] shows that the unbiased MLE estimator of the mean  $\mu$  of a normal distribution  $N(\mu, \sigma^2)$  is  $\bar{X}_n = (X_1 + \dots + X_n)/n$ . Let  $X \sim \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ . The Fisher information is:

$$\begin{aligned} I(\theta) &= n\mathbb{E}\left[\left(\frac{\partial}{\partial\mu} \log f_\mu(X)\right)^2\right] \\ &= n\mathbb{E}\left[\left(\frac{X-\mu}{\sigma^2}\right)^2\right] \\ &= \frac{n}{\sigma^4}\mathbb{E}\left[(X-\mu)^2\right] \\ &= \frac{n}{\sigma^4}\text{Var}(X) = \frac{n}{\sigma^4}\sigma^2 = \frac{n}{\sigma^2} = \frac{1}{\text{Var}(\bar{X}_n)} \end{aligned}$$

where the last equality follows because for i.i.d. random variables  $\text{Var}(\bar{X}_n) = \sigma^2/n$ . By taking the reciprocals:

$$\text{Var}(\bar{X}_n) = \frac{1}{I(\theta)}$$

we have that the lower bound of the Cramér-Rao inequality is reached, hence  $\bar{X}_n$  is a MVUE (Minimum Variance Unbiased Estimator).

## 2 Least Square Estimators in Simple Linear Regression

Consider the least square estimators:

$$\hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{x}_n \quad \hat{\beta} = \frac{\sum_1^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{SXX} \quad (4)$$

where  $SXX = \sum_1^n (x_i - \bar{x}_n)^2$ . Since  $\sum_1^n (x_i - \bar{x}_n) = 0$ , we can rewrite  $\hat{\beta}$  as:

$$\hat{\beta} = \frac{\sum_1^n (x_i - \bar{x}_n)Y_i - \sum_1^n (x_i - \bar{x}_n)\bar{Y}_n}{SXX} = \frac{\sum_1^n (x_i - \bar{x}_n)Y_i}{SXX} \quad (5)$$

### 2.1 Expectation

$\hat{\beta}$  is an unbiased estimator:

$$\begin{aligned} E[\hat{\beta}] &= \frac{\sum_1^n (x_i - \bar{x}_n)E[Y_i]}{SXX} \\ &= \frac{\sum_1^n (x_i - \bar{x}_n)(\alpha + \beta x_i)}{SXX} \\ &= \frac{\beta \sum_1^n (x_i - \bar{x}_n)x_i}{SXX} = \beta \end{aligned}$$

where the last step follows since  $\sum_1^n (x_i - \bar{x}_n)x_i = \sum_1^n (x_i - \bar{x}_n)x_i - \sum_1^n (x_i - \bar{x}_n)\bar{x} = SXX$ . See the textbook [1, page 331] for a proof that  $\hat{\alpha}$  is also unbiased, and [1, Exercise 22.12] for a different proof for  $\hat{\beta}$ .

### 2.2 Variance and Standard Errors of the Coefficients

We calculate:

$$Var(\hat{\beta}) = \frac{\sum_1^n (x_i - \bar{x}_n)^2 Var(Y_i)}{SXX^2} = \sigma^2 \frac{\sum_1^n (x_i - \bar{x}_n)^2}{SXX^2} = \frac{\sigma^2}{SXX} \quad (6)$$

and:

$$\begin{aligned} Var(\hat{\alpha}) &= Var(\bar{Y}_n - \hat{\beta}\bar{x}_n) \\ &= Var(\bar{Y}_n) + \bar{x}_n^2 Var(\hat{\beta}) - 2\bar{x}_n Cov(\bar{Y}_n, \hat{\beta}) \\ &= \frac{\sigma^2}{n} + \bar{x}_n^2 \frac{\sigma^2}{SXX} - 0 = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}_n^2}{SXX} \right) \end{aligned} \quad (7)$$

The covariance in the formula is zero because (recall that  $Y_1, \dots, Y_n$  are independent):

$$\begin{aligned} Cov(\bar{Y}_n, \hat{\beta}) &= Cov\left(\frac{1}{n} \sum_1^n Y_i, \frac{\sum_1^n (x_i - \bar{x}_n)Y_i}{SXX}\right) \\ &= \frac{1}{nSXX} Cov\left(\sum_1^n Y_i, \sum_1^n (x_i - \bar{x}_n)Y_i\right) \\ &= \frac{1}{nSXX} \sum_1^n Cov(Y_i, (x_i - \bar{x}_n)Y_i) \\ &= \frac{1}{nSXX} \sum_1^n (x_i - \bar{x}_n) Var(Y_i) = \frac{\sigma^2 \sum_1^n (x_i - \bar{x}_n)}{nSXX} = 0 \end{aligned}$$

The *standard errors* of the coefficient estimators are defined as the estimates of the standard deviations (see (6) and (7)):

$$se(\hat{\alpha}) = \hat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{\bar{x}_n^2}{SXX}\right)} \quad se(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{SXX}} \quad (8)$$

where:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_1^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \quad (9)$$

is the (unbiased) estimate of  $\sigma^2$  (see [1, page 332]).

### 2.3 Variance-Covariance Matrix

The variance-covariance matrix is:

$$\begin{pmatrix} Var(\hat{\alpha}) & Cov(\hat{\alpha}, \hat{\beta}) \\ Cov(\hat{\beta}, \hat{\alpha}) & Var(\hat{\beta}) \end{pmatrix}$$

where the unknown value  $\sigma^2$  is replaced with the estimate  $\hat{\sigma}^2$  from (9). The standard errors can be obtained from the square roots of the diagonal elements<sup>3</sup> The matrix is symmetric, as covariance is symmetric. Moreover, we calculate:

$$\begin{aligned} Cov(\hat{\alpha}, \hat{\beta}) &= Cov(\bar{Y}_n - \hat{\beta}\bar{x}_n, \hat{\beta}) \\ &= Cov(\bar{Y}_n, \hat{\beta}) - \bar{x}_n Cov(\hat{\beta}, \hat{\beta}) \\ &= -\bar{x}_n Var(\hat{\beta}) \end{aligned} \quad (10)$$

### 2.4 Variance and Standard Errors of Fitted Values

For a given value of the explanatory variable, say  $x_0$ , the estimator  $\hat{Y} = \hat{\alpha} + \hat{\beta}x_0$  has expectation  $E[\hat{Y}] = E[\hat{\alpha}] + E[\hat{\beta}]x_0 = \alpha + \beta x_0$ . Hence,  $\hat{Y}$  is unbiased and  $\hat{y} = \hat{\alpha} + \hat{\beta}x_0$  is then the best estimate for the fitted value. We can compute the variance of  $\hat{Y}$  as:

$$\begin{aligned} Var(\hat{Y}) &= Var(\hat{\alpha} + \hat{\beta}x_0) \\ &= Var(\hat{\alpha}) + x_0^2 Var(\hat{\beta}) + 2x_0 Cov(\hat{\alpha}, \hat{\beta}) \\ &= Var(\hat{\alpha}) + (x_0^2 - 2x_0\bar{x}_n) Var(\hat{\beta}) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{SXX}\right) + \frac{(x_0^2 - 2x_0\bar{x}_n)\sigma^2}{SXX} \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{SXX}\right) \end{aligned}$$

where  $Cov(\hat{\alpha}, \hat{\beta})$  has been simplified based on (10). The *standard error* of the fitted value is then the estimate:

$$se(\hat{y}) = \hat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{SXX}\right)} \quad (11)$$

---

<sup>3</sup>In R, with the expression `sqrt(diag(vcov(fit)))` where `fit` is the linear model.

### 3 Confidence Intervals for Simple Linear Regression

In this section, we make the *normality assumption* that  $U_i \sim \mathcal{N}(0, \sigma^2)$  in the simple linear regression model [1, page 257]:

$$Y_i = \alpha + \beta x_i + U_i$$

A fortiori, we have  $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$ .

#### 3.1 Confidence Intervals of the Coefficients

By (5), the estimator  $\hat{\beta}$  is a linear combination of the  $Y_i$ 's, hence it has normal distribution as well. By Sections 1.1 and 1.2, it must be that:

$$\hat{\beta} \sim \mathcal{N}(\beta, \text{Var}(\hat{\beta}))$$

where the variance  $\text{Var}(\hat{\beta})$  given in (6) is unknown because  $\sigma^2$  is unknown. The studentized statistics:

$$\frac{\hat{\beta} - \beta}{\sqrt{\text{Var}(\hat{\beta})}} \sim t(n-2) \quad (12)$$

has a t-student distribution with  $n-2$  degrees of freedom ( $n-2$  because 2 parameters are already estimated). The proof of this fact can be found in [2, page 45]. Hence:

$$P\left(-t_{n-2,0.025} \leq \frac{\hat{\beta} - \beta}{\sqrt{\text{Var}(\hat{\beta})}} \leq t_{n-2,0.025}\right) = 0.95$$

where  $t_{n-2,0.025}$  is the critical value of  $t(n-2)$  at 0.025. Hence, a 95% confidence interval is:

$$\hat{\beta} \pm t_{n-2,0.025} se(\hat{\beta})$$

where  $se(\hat{\beta})$  is the standard error of  $\hat{\beta}$  from (8). By following the same reasoning, we obtain the confidence intervals for  $\alpha$ :

$$\hat{\alpha} \pm t_{n-2,0.025} se(\hat{\alpha})$$

where  $se(\hat{\alpha})$  is the standard error of  $\hat{\alpha}$  from (8).

#### 3.2 Confidence and Prediction Intervals of the Fitted Values

Analogously to the previous subsection, for a fitted value  $\hat{y} = \hat{\alpha} + \hat{\beta}x_0$ , a 95% *confidence interval* is:

$$\hat{y} \pm t_{n-2,0.025} se(\hat{Y})$$

where  $se(\hat{y})$  is from (11). In particular, this interval concerns *the expectation of fitted values* at  $x_0$ . For example, we could conclude that the mean of predicted values at  $x_0$  is between  $\hat{y} - t_{n-2,0.025} se(\hat{y})$  and  $\hat{y} + t_{n-2,0.025} se(\hat{y})$ . For a given single prediction, we must also account for the variance of the error term  $U$  in:

$$\hat{V} = \hat{\alpha} + \hat{\beta}x_0 + U$$

Let us assume that  $U \sim \mathcal{N}(0, \sigma^2)$ . By reasoning as in Section 1.3, it can be shown that  $\text{Var}(\hat{V}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{SXX}\right)$ , and then by defining:

$$se(\hat{v}) = \hat{\sigma} \sqrt{\left(1 + \frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{SXX}\right)}$$

we have that the *prediction interval* is:

$$\hat{y} \pm t_{n-2,0.025}se(\hat{v})$$

In this case, we could conclude that the specific predicted value at  $x_0$  is on between  $\hat{y} - t_{n-2,0.025}se(\hat{v})$  and  $\hat{y} + t_{n-2,0.025}se(\hat{v})$ .

### 3.3 Hypothesis Testing

Consider now the two-tailed test of hypothesis:

$$H_0 : \beta = 0 \quad H_1 : \beta \neq 0$$

The p-value of observing  $|\hat{\beta}|$  or a greater value under the null hypothesis, can be calculated from (12) as:

$$p = P(|T| > |t|) = 2 \cdot P\left(T > \left| \frac{\hat{\beta} - 0}{se(\hat{\beta})} \right|\right)$$

for  $T \sim t(n-2)$ . Hence,  $H_0$  can be rejected in favor of  $H_1$  at significance level of 0.05, i.e.  $p < 0.05$ , if  $|t| > t_{n-2,0.025}$ . A similar approach applies to the intercept.

## 4 Statistical Decision Theory

This section will be added later on.

### References

- [1] F.M. Dekking, C. Kraaikamp, H.P. Lopuhaä, and L.E. Meester. *A Modern Introduction to Probability and Statistics*. Springer, 2005.
- [2] M. H. Kutner, C. J. Nachtsheim, J. Neter, and Li W. *Applied Linear Statistical Models*. 5th edition, 2005.