Lecture Notes

Statistics for Data Science

University of Pisa, Italy

Salvatore Ruggieri

April 8, 2023

Contents

1 On Cramér-Rao's bound and MLE

Consider the log-likelihood function:

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

The MLE principle estimates the unknown parameter(s), given the observations, as the θ which maximizes $\ell(\theta)$. The log-likelihood takes its maximum at the zero's of its derivative, which is called the *score function*:

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

Such a function is relevant beyond the maximization problem. It describes how much a change in θ results into a change of the density, or, equivalently, how much much informative[1] are the observations in estimating a parameter θ.

Let us introduce the random variables $Y_i = \frac{\partial}{\partial\theta} \log f_\theta(X_i)$, for $i = 1, \ldots, n$. The score function can be written as $S(\theta) = \sum_{i=1}^{n} Y_i$. Since $X_1, \ldots, X_n$ are i.i.d., by the propagation of independence, this is also true for $Y_1 = \frac{\partial}{\partial\theta} \log f_\theta(X_1)$, $\ldots$, $Y_n = \frac{\partial}{\partial\theta} \log f_\theta(X_n)$.
The expectation of each Y_i's is zero (use Leibniz integral rule):

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

Hence, by linearity of expectation, we have:

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

We resort then to the variance of $S(\theta)$ as a summary of the information provided by the random sample. The variance of $S(\theta)$ is called the *Fisher information*, and it is the quantity:

$$I(\theta) = Var(S(\theta)) = \mathrm{E}\big[S(\theta)^2\big]$$

It turns out[2][3] that:

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

[1] Recall that information is measured as $-\log f_\theta(X)$, i.e., events with small probability bring more information.

[2] (1) follows since $\mathrm{E}\Big[Y_i Y_j\Big] = \mathrm{E}\Big[Y_i\Big]\mathrm{E}\Big[Y_j\Big]$ for independent Y_i, Y_j.

[3] (2) follows since $\mathrm{E}\Big[Y_i\Big] = 0$.

$$= \mathrm{E}\big[\sum_{i=1}^{n} Y_i^2\big] + 0 \tag{2}$$

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

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

where $X \sim f_\theta$. **Important**: some textbooks define $I(\theta)$ using a single random variable, i.e., as $\mathrm{E}\big[(\frac{\partial}{\partial\theta} \log f_\theta(X))^2\big]$. In such cases, it must be multiplied by n whenever it is used. We can now link Fisher information to the Cramér-Rao inequality from [1, Remark 20.2]:

$$\mathrm{Var}(T) \geq \frac{1}{n\mathrm{E}\big[(\frac{\partial}{\partial\theta} \log f_\theta(X))^2\big]} \qquad \text{for all } \theta,$$

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

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

Example

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

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

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

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

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

Exercise

Show the following equivalent formulation:

$$I(\theta) = -n\mathrm{E}\big[\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta} \log f_\theta(X)\big]$$

2 Least Square Estimators in Simple Linear Regression

Consider the least square estimators:

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

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:

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

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:

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

and:

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

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

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

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

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

where:

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

is the (unbiased) estimate of σ^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 σ^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[4] The matrix is symmetric, as covariance is symmetric. Moreover, we calculate:

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

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:

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

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

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

[4] 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 of the Y_i's, hence it has normal distribution as well. By Sections 1.1 and 1.2, it must be that:

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

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

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

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

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

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 α:

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

where $se(\hat{\beta})$ is the standard error of $\hat{\beta}$ 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 $Var(\hat{V}) = \sigma^2(1 + \frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{SXX})$, and then by defining:

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

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:

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

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

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

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.