In section so far, we have only dealt with finding MLEs for distributions with one parameter, but some distributions have multiple parameters. In such a case, the parameter is multi-dimensional, leading to a multi-dimensional MLE. Since the MLE is a random variable, this means that we are dealing with a multivariate RV!

So let us first review multivariate RVs, then MLEs for multivariate parameters.

After that, we will review how to compute MLEs for transformations of parameters, which works for univariate and multivariate parameters.

1 Multivariate RVs

For most of the course, we have been working with univariate distributions. For example, if $X \sim N(\mu, \sigma^2)$, X can be viewed as a one-dimensional (or 1×1), vector, since it will only take on a single value like 1.3, -2.5, etc.

But we can also have RVs that are multivariate. For example, if X is a multivariate RV, we would write it as a multi-dimensional (or $p \times 1$) vector:

$$\underset{p\times 1}{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}.$$

or $X = (X_1, X_2, \ldots, X_p)^T$. Its mean is a $p \times 1$ vector, and its variance is a $p \times p$ variance-covariance matrix:

$$\mu = \begin{pmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \\ \vdots \\ \mathbb{E}(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \ldots & \mathrm{Cov}(X_2, X_p) \\ \vdots & & \ldots & \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \ldots & \mathrm{Cov}(X_p, X_p). \end{pmatrix}$$

Recall that the covariance of an RV with itself is just its variance and that the covariance function is symmetric: $\mathrm{Cov}(X_j, X_j) = \mathrm{Var}(X_j) = \sigma_j^2$ and $\mathrm{Cov}(X_j, X_k) = \mathrm{Cov}(X_k, X_j)$.

So we can write Σ with the variances along the diagonal. Also we can see that the matrix is

symmetric, since each entry (j, k) is the same as entry (k, j):

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_1, X_2) & \sigma_2^2 & \dots & \text{Cov}(X_2, X_p) \\ \vdots & & \dots & \\ \text{Cov}(X_1, X_p) & \text{Cov}(X_2, X_p) & \dots & \sigma_p^2. \end{pmatrix}$$

(Exploiting the symmetry of the covariances can be helpful sometimes for computation, since it reduces the number of covariance calculations by half.)

Note that μ and Σ do not necessarily fully characterize a distribution – it's possible that there are higher-order dependencies in the data $X_1, \dots, X_n$, but it can be challenging to analyze those dependencies.

Just like with univariate RVs, the parameters for multivariate distributions are typically unknown in practice. Therefore, we will want to collect a random sample (with the sample size n as large as possible) and try to estimate the parameters using that.

For a random sample of multivariate RVs, we will have two separate indices: $j = 1, 2, \dots, p$ for the dimensions and $i = 1, 2, \dots, n$ for the members of the sample. To keep track of these indices, one can write X_i^j (for double indexing, this is the notation I prefer, with the dimension index j in the superscript). Then we have

$$X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} (\mu, \Sigma).$$

Each X_i is a $p \times 1$ vector, which we may write as

$$X_i = \begin{pmatrix} X_i^1 \\ X_i^2 \\ \vdots \\ X_i^p, \end{pmatrix}$$

and which has mean and covariance

$$\mathbb{E}(X_i) = \begin{pmatrix} \mathbb{E}(X_i^1) \\ \mathbb{E}(X_i^2) \\ \vdots \\ \mathbb{E}(X_i^p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \mu$$

$$\text{and } \text{Cov}(X_i) = \begin{pmatrix} \text{Var}(X_i^1) & \text{Cov}(X_i^1, X_i^2) & \dots & \text{Cov}(X_i^1, X_i^p) \\ \text{Cov}(X_i^2, X_i^1) & \text{Var}(X_i^2) & \dots & \text{Cov}(X_i^2, X_i^p) \\ \vdots & & \dots & \\ \text{Cov}(X_i^p, X_i^1) & \text{Cov}(X_i^p, X_i^2) & \dots & \text{Var}(X_i^p) \end{pmatrix} = \Sigma,$$

since the data are iid.

Why are multivariate distributions helpful? They aim to capture dependencies among RVs by encoding the dependencies within the model. For example, suppose the University of Florida undergraduate admissions office is interested in understanding SAT scores, ACT scores, and GPA among Florida seniors, so that they can figure out how to interpret the applications they receive. Let $X_1, X_2, \ldots, X_n$ be an iid random sample of the academic information from Florida seniors. Let X_i^1 be the SAT score for student i, X_i^2 be their ACT score, and X_i^3 be their GPA. To more easily keep track of things, let's relabel these variables as X_i^S, X_i^A, and X_i^G. Let's compare a univariate and multivariate approach:

(a) Univariate: Three separate analyses would be done. For example, we could write these models:

$$X_1^S, X_2^S, \ldots, X_n^s \overset{\text{iid}}{\sim} (\mathbb{E}(X_i^S), \text{Var}(X_i^S))$$
$$X_1^A, X_1^A, \ldots, X_n^A \overset{\text{iid}}{\sim} (\mathbb{E}(X_i^A), \text{Var}(X_i^A))$$
$$X_1^G, X_1^G, \ldots, X_n^G \overset{\text{iid}}{\sim} (\mathbb{E}(X_i^G), \text{Var}(X_i^G)).$$

(b) Multivariate: One analysis would be done that is more complex. For example, we can write the model

$$X_1, X_2, \ldots, X_p \overset{\text{iid}}{\sim} (\mu, \Sigma),$$

where

$$\underset{3\times 1}{X_i} = \begin{pmatrix} X_i^S \\ X_i^A \\ X_i^G \end{pmatrix}.$$

What's the difference between these two approaches? The multivariate approach takes into account the dependencies among SAT score, ACT score, and GPA through the covariance terms in Σ. Since we would expect there to be some relationship among these variables (e.g., students with high SAT scores would probably be more likely to have high ACT scores), a multivariate approach might make more sense than the univariate here.

Technical Note: If we have a random sample of univariate RVs $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} (\theta, \sigma^2)$, then we could write it with the multivariate vector/matrix notation: $X \sim (\mu, \Sigma)$, where

$$\underset{n\times 1}{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}, \underset{n\times 1}{\mu} = \begin{pmatrix} \theta \\ \theta \\ \vdots \\ \theta, \end{pmatrix} \text{ and } \underset{n\times n}{\Sigma} = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & 0\ldots & 0 \\ \vdots & & \ldots & \\ 0 & \ldots & \ldots & \sigma^2. \end{pmatrix}$$

But it's typically easier to just view/write it as a univariate random sample.

There are many different multivariate distributions. Just like the univariate normal distribution is extremely important in statistics, so is the multivariate normal distribution.

2 Computing MLEs for Multivariate Parameters

Some distributions have multi-dimensional, or $k \times 1$ parameters, e.g., $\theta = (\theta_1, \theta_2, \ldots, \theta_k)'$. To compute the MLE for θ, we have to use multivariate optimization techniques.

One method is multivariate calculus. If this is possible (assuming differentiability, etc.), then the general steps are as follows:

(a) Write out the joint likelihood $L(\theta)$

(b) Get the log-likelihood $l(\theta)$

(c) Compute all the partial derivatives $\frac{\partial l}{\partial \theta_1}, \ldots, \frac{\partial l}{\partial \theta_k}$.

(d) Solve (if possible) the system of equations

$$\begin{cases} \frac{\partial l}{\partial \theta_1} & = 0 \\ \vdots & = 0 \\ \frac{\partial l}{\partial \theta_k} & = 0 \end{cases}$$

for $\theta_1, \ldots, \theta_k$, giving the critical values $\hat{\theta}_1, \ldots, \hat{\theta}_k$.

(e) Conduct a partial derivative test to confirm that the optimizers are maximizers, as opposed to minimizers, saddle points, etc.

For example, suppose we have a (univariate) random sample $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where both μ and σ^2 unknown. Then the multivariate parameter is $\theta = (\mu, \sigma^2)^T$, and we can follow the above steps to find the MLE.

Univariate MLEs had some very desirable properties under some regularity conditions. One of them was the facts about asymptotic normality, which allowed us to calculate approximate confidence intervals for parameters. Multivariate MLEs also have asymptotic normality under some regularity conditions.

Assume the regularity conditions are met. If $\underset{k\times 1}{\hat{\theta}}$ is the MLE for $\underset{k\times 1}{\theta}$, then

$$\hat{\theta} \implies N(\theta, I^{-1}(\theta)),$$

where $\underset{k\times k}{I}^{-1}(\theta)$ is the inverse of the Fisher Information matrix. We can compute the Fisher Information as

$$I(\theta) = -E(H(l(\theta))),$$

where $H(l(\theta))$ is the Hessian of the (full-sample) log-likelihood:

$$H(l(\theta)) = \begin{pmatrix} \frac{\partial}{\partial\theta_1}\left(\frac{\partial l}{\partial\theta_1}\right) & \frac{\partial}{\partial\theta_1}\left(\frac{\partial l}{\partial\theta_2}\right) & \cdots & \frac{\partial}{\partial\theta_1}\left(\frac{\partial l}{\partial\theta_k}\right) \\ \frac{\partial}{\partial\theta_2}\left(\frac{\partial l}{\partial\theta_1}\right) & \frac{\partial}{\partial\theta_2}\left(\frac{\partial l}{\partial\theta_2}\right) & \cdots & \frac{\partial}{\partial\theta_2}\left(\frac{\partial l}{\partial\theta_k}\right) \\ \vdots & & \cdots & \\ \frac{\partial}{\partial\theta_k}\left(\frac{\partial l}{\partial\theta_1}\right) & \frac{\partial}{\partial\theta_k}\left(\frac{\partial l}{\partial\theta_2}\right) & \cdots & \frac{\partial}{\partial\theta_k}\left(\frac{\partial l}{\partial\theta_k}\right) \end{pmatrix}.$$

To get the expectation, simply compute the expectation for each entry of H.

As with the univariate case, if the data are iid, then $I_n(\theta) = nI(\theta)$, so we can use that formula too.

2.1 Matrix Facts

Here are a few simple but helpful facts about matrix inverses. Let A be a square matrix (i.e., A is $p \times p$), and c be a nonzero scalar. Then

(a) $(cA)^{-1} = c^{-1}A^{-1}$

(b) If $p = 2$ and $A = \begin{pmatrix} a & b \\ c & d, \end{pmatrix}$ then the inverse is

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a, \end{pmatrix}$$

where $\det(A) = ad - bc$ is the determinant of A (assuming that $ad - bc \neq 0$, in which case the inverse does not exist). Unfortunately, it is not as easy to invert larger square matrices...

3 MLEs for Transformations of Parameters

Another terrific property of MLEs is known as the invariance or equivalence property.

Theorem 3.1. Invariance Property. Suppose $\hat{\theta}$ is the MLE for θ, and $g(\theta) = \tau$ is a function. The parameters θ and τ could be univariate or multivariate. Then the MLE for τ is $\hat{\tau} = g(\hat{\theta})$.

If g is invertible, then this works because we can simply reparametrize the distribution (substitute $g^{-1}(\tau)$ in for θ in the original distribution). If g is not invertible, then it takes a

bit of care to even define what the likelihood with respect to τ would be. The lecture notes contain the technical details of this.

Under regularity conditions, the MLE of $\hat{\tau}$ is asymptotically normal. This works for both univariate and multivariate parameters, and we use the asymptotic normality to compute approximate confidence intervals for τ.

The univariate case in Theorem 5 of the lecture notes says the following:

Theorem 3.2. Suppose $\tau = g(\theta)$, where $g : \mathbb{R} \to \mathbb{R}$ is differentiable. Assume too that $\frac{dg}{d\theta} = g'(\theta) \neq 0$. Let $\hat{\theta}_n$ be the MLE for θ and $\hat{\tau}_n = g(\hat{\theta})$ the MLE for τ. and assume the regularity conditions hold, so that $\hat{\theta}_n$ is asymptotically normal. Then we have that

$$\hat{\tau}_n \implies N(\tau, \hat{\text{Var}}(\hat{\tau}_n)),$$

where $\hat{\text{Var}}(\hat{\tau}_n) \approx (g'(\hat{\theta}_n))^2 \hat{\text{Var}}(\hat{\theta}_n)$. Equivalently,

$$\frac{\hat{\tau}_n - \tau}{\hat{se}(\hat{\tau}_n)} \implies N(0, 1),$$

where $\hat{se}(\hat{\tau}_n) \approx |g'(\hat{\theta})| \hat{se}(\hat{\theta}_n)$.

Here is an extension of that theorem, which will work for any dimensions of τ and θ (one or both of them could be multivariate):

Theorem 3.3. Suppose $\tau = g(\theta)$, where $g : \mathbb{R}^p \to \mathbb{R}^k$ is differentiable. So

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix} \text{ and}$$

$$\tau = \begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_k \end{pmatrix} = g(\theta) = \begin{pmatrix} g_1(\theta) \\ g_2(\theta) \\ \vdots \\ g_k(\theta) \end{pmatrix}.$$

Assume too that $\Delta(g(\theta)) \neq \mathbf{0}_k$, where $\Delta(g(\theta))$ is the Jacobian and $\mathbf{0}$ is the zero matrix, i.e.,

$$\underset{p \times k}{\Delta(g(\theta))} = \begin{pmatrix} \frac{\partial g_1}{\partial \theta_1} & \frac{\partial g_2}{\partial \theta_1} & \cdots & \frac{\partial g_k}{\partial \theta_1} \\ \frac{\partial g_1}{\partial \theta_2} & \frac{\partial g_2}{\partial \theta_2} & \cdots & \frac{\partial g_k}{\partial \theta_2} \\ \vdots & & \cdots & \\ \frac{\partial g_1}{\partial \theta_p} & \frac{\partial g_2}{\partial \theta_p} & \cdots & \frac{\partial g_k}{\partial \theta_p} \end{pmatrix} \text{ and } \underset{p \times k}{\mathbf{0}} = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & & \ldots & 0 \\ 0 & 0 & \ldots & 0 \end{pmatrix}.$$

Let $\hat{\theta}_n$ be the MLE for θ and $\hat{\tau} = g(\hat{\theta})$ the MLE for τ. Assume the regularity conditions hold so that $\hat{\theta}_n$ is asymptotically normal. Then we have that

$$\hat{\tau}_n \implies N(\theta, \hat{\text{Cov}}(\hat{\tau}_n)),$$

where $\hat{\text{Cov}}(\hat{\tau}_n) \approx (\Delta g(\hat{\theta}_n))^T \hat{\text{Cov}}(\hat{\theta}_n)(\Delta g(\hat{\theta}_n))$. To build a confidence interval for component j of τ, then just grab entry j, j from the covariance matrix and take the square root to get the standard error.

Problem 1. MLE Invariance for Normal Distribution

Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma)^2$.

(a) Assume σ^2 is known. Can we reparametrize the original density to be with respect to $\tau = \mu^2$? What is the MLE for τ?

(b) Assume μ is known. Can we reparametrize the original density to be in terms of $\nu = \sigma^4$? What is the MLE for ν?

(c) Assume both σ^2 and μ are unknown. Let $\theta = (\mu, \sigma^2)^T$. What is the MLE for $\theta = (\tau, \nu)^T$?

(d) For the case that σ^2 is known, find an $(1-\alpha)100\%$ approximate confidence interval for $\hat{\tau}$.

(e) For the case that μ is known, find an approximate $(1-\alpha)100\%$ confidence interval for $\hat{\sigma}^2$.

(f) For the case that both σ^2 and μ are unknown, compute an approximate $(1-\alpha)100\%$ confidence interval for $(\tau, \nu)^T$.

Solution

(a) Observe that the function $g(\mu) = \mu^2$ is not one-to-one on the domain of $\mu \in (-\infty, \infty)$, since $g(-m) = g(m)$ for any $m \in (-\infty, \infty)$. Thus, g is not invertible, and we cannot reparametrize in that way. But using the definition of induced likelihood and by invariance, we are able to compute the MLE $\hat{\tau}$ as $g(\hat{\mu}) = \hat{\mu}^2$. We have previously calculated the MLE of μ as $\hat{\mu} = \bar{X}$, so $\hat{\tau} = \bar{X}^2$.

(b) We can reparametrize in that way, since the function $\nu = h(\sigma^2) = (\sigma^2)^2$ is invertible on the domain of σ^2, which is $[0, \infty)$. By invariance, the MLE is $\hat{\nu} = h(\hat{\sigma}^4)$. We have previously calculated the MLE for σ as $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}$, so $\hat{\tau} = \hat{\sigma}^4 = \left(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right)^2$.

(c) The method for calculating MLEs for transformations of parameters applies to multivariate parameters. Let $q(\theta) = (\mu^2, \sigma^4)^T = (\tau, \nu)^T$. It was previously calculated in lecture that the MLE for $\theta = (\mu, \sigma)^T$ is $\hat{\theta} = (\bar{X}, \hat{\sigma})^T$, where $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}$. So the MLE for $(\tau, \nu)^T$ is $q(\hat{\theta}) = \left(\bar{X}^2, \left(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right)^2\right)$.

Note: In this case, the multivariate estimator for $(\tau, \nu)^T$ happens to be the same as the two univariate estimators from parts (a) and (b) stacked together. This wouldn't always happen though, depending on what distribution we have, what function we have, etc. So make sure to always follow the correct formulas/processes!

(d) The confidence interval formula is

$$\hat{\tau} \pm z_{1-\alpha/2}\hat{se}(\hat{\tau}).$$

We can use the univariate theorem, Theorem 3.2. The estimated variance for $\hat{\tau}_n$ will be

$$\hat{Var}(\hat{\tau}_n) = g'(\hat{\mu}_n)^2 \hat{\text{Var}}(\hat{\mu}_n).$$

The derivative is $g'(\mu) = 2\mu$, and we also know that the variance of $\bar{X}$ is σ^2/n. Plugging in everything, we have

$$\begin{aligned} \hat{Var}(\bar{X}^2) &= (2\hat{\mu})^2 \frac{\sigma^2}{n} \\ &= 4\bar{X}^2 \frac{\sigma^2}{n}, \end{aligned}$$

so the confidence interval is $\bar{X}^2 \pm z_{1-\alpha/2} \cdot 2\bar{X}\frac{\sigma}{\sqrt{n}}$.

(e) We can apply the same theorem. The derivative of $h(\sigma)$ is $h'(\sigma) = 4\sigma^3$. The variance of $\hat{\sigma}$ was previously calculated as $\sigma^2/2n$. So we have,

$$\begin{aligned} \hat{Var}(\hat{\nu}_n) &= g'(\hat{\sigma}_n)^2 \hat{\text{Var}}(\hat{\sigma}_n) \\ &= (4\hat{\sigma}^3)^2 \cdot \hat{\sigma}^2/2n \\ &= \frac{8\hat{\sigma}^8}{n}. \end{aligned}$$

So the confidence interval is $\hat{\sigma}^4 \pm z_{1-\alpha/2} \cdot 2\hat{\sigma}^4\sqrt{\frac{2}{n}}$.

(f) Let $\theta = (\mu, \sigma)^T$ and $G(\theta) = (G_1(\theta), G_2(\theta))^T = (\mu^2, \sigma^4)^T$. We need to calculate $\hat{\text{Cov}}(\hat{\theta}) = (\Delta G(\hat{\theta}_n))^T \hat{\text{Cov}}(\hat{\theta}_n)(\Delta G(\hat{\theta}_n))$. First, it was calculated in lecture that $\text{Cov}(\hat{\theta}_n)$ is

$$\text{Cov}(\hat{\theta}_n) = I_n^{-1}(\hat{\theta}_n) = \begin{pmatrix} \hat{\sigma}^2/n & 0 \\ 0 & \hat{\sigma}^2/2n. \end{pmatrix}$$

Now, the (true) Jacobian is calculated as

$$\Delta G(\theta) = \begin{pmatrix} \frac{\partial G_1(\theta)}{\partial \theta_1} & \frac{\partial G_2(\theta)}{\partial \theta_1} \\ \frac{\partial G_1(\theta)}{\partial \theta_2} & \frac{\partial G_2(\theta)}{\partial \theta_2} \end{pmatrix} = \begin{pmatrix} 2\mu & 0 \\ 0 & 4\sigma^3, \end{pmatrix}$$

and the transpose is the same, since it's a diagonal matrix.

So we have

$$\text{Cov}(\hat{\theta}_n) = \begin{pmatrix} 2\bar{X} & 0 \\ 0 & 4\hat{\sigma}^3. \end{pmatrix} \begin{pmatrix} \hat{\sigma}^2/n & 0 \\ 0 & \hat{\sigma}^2/2n. \end{pmatrix} \begin{pmatrix} 2\bar{X} & 0 \\ 0 & 4\hat{\sigma}^3. \end{pmatrix} \tag{1}$$

$$= \begin{pmatrix} 4\bar{X}^2\hat{\sigma}^2 n & 0 \\ 0 & 16\hat{\sigma}^8/2n \end{pmatrix}. \tag{2}$$

To form the confidence intervals, we just need to grab the diagonal entries of the covariance matrix and take the square root. So the approximate confidence interval for τ is

$$\bar{X}^2 \pm z_{1-\alpha/2} \cdot \frac{2\bar{X}\hat{\sigma}}{\sqrt{n}}$$

and for ν is

$$\hat{\sigma}^4 \pm z_{1-\alpha/2} \cdot 2\hat{\sigma}^4 \sqrt{\frac{2}{n}}.$$

Problem 2. Applied Example Using MLE Invariance

The *repose period* of a volcano is the amount of time between two consecutive eruptions. There is a theory that longer repose periods tend to lead to larger eruptions (higher amounts of emission volume).

A geologist Dr. George wants to know the true mean eruption volume of volcanoes in North America. However, he only has a random sample of repose periods $X_1, X_2, \ldots, X_n$, which he assumes are iid Exp(λ). Let $Y_1, Y_2, \ldots, Y_n$ be the corresponding eruption volumes in cubic kilometers (unknown RVs), which are also assumed to be iid. Based on the theory about volcanoes, he proposes the model

$$\mathbb{E}(Y) = \frac{k}{\lambda},$$

where $k > 1$ is a known constant.

(a) Find the MLE for λ.

(b) Find the MLE for $\mathbb{E}(Y)$. Feel free to compute an approximate 90% confidence interval for $\mathbb{E}(Y)$ too.

(c) Suppose the sample average amount of time between eruptions is 200 years and $k = 5$. Compute the value of the MLE for $\mathbb{E}(Y)$.

(d) Optional: Do you see any potential issues with Dr. George's analysis?

Solution

(a) The PDF for an exponential RV is

$$f(x_i; \lambda) = \lambda e^{-\lambda x_i}.$$

The joint likelihood for the sample is

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda X_i} \text{ (since the data are iid)} \\ &= \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}. \end{aligned}$$

The log-likelihood is

$$l(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i.$$

Taking the derivative, we have

$$l'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i.$$

Setting it equal to 0 and solving for λ, we have

$$\begin{aligned} \frac{n}{\lambda} - \sum X_i = 0 &\iff \\ \frac{n - \lambda \sum X_i}{\lambda} = 0 &\iff \\ n - \lambda \sum X_i = 0 &\iff \\ \lambda = \frac{n}{\sum X_i} = \frac{1}{\bar{X}}. & \end{aligned}$$

To verify that this optimum $\hat{\lambda} = \frac{1}{\bar{X}}$ is a maximizer, we can perform the second derivative test:

$$l''(\lambda) = \frac{n}{\lambda^2} < 0.$$

Thus, indeed, the MLE is $\hat{\lambda} = \frac{1}{\bar{X}}$.

The mean of an exponential distribution is $\frac{1}{\lambda}$, i.e., $= \frac{1}{\mathbb{E}(X_i)}$. So it seems intuitive/reasonable that the MLE for λ is $\frac{1}{\bar{X}}$.

(b) Since the function $g(\lambda) = \frac{k}{\lambda} = \tau = \mathbb{E}(Y)$ is invertible, we can apply the invariance property of the MLE. This means that $\hat{\mathbb{E}(Y)} = \frac{k}{\hat{\lambda}} = \frac{k}{\frac{1}{\bar{X}}} = k\bar{X}$.

(c) The sample average amount of time between eruptions is $\bar{X} = 200$ years. So the observed MLE for λ is $\frac{1}{200}$ eruptions per year.

The observed MLE for $\mathbb{E}(Y)$ is

$$\hat{\mathbb{E}(Y)} = 5(200) = 1000km^3.$$

(d) First of all, he assumed that the repose times, as well as the corresponding volumes, are iid. This is quite unrealistic, considering the geographic/geologic dependence of neighboring volcanoes. For example, geologic activity in a mountain range could trigger multiple volcanoes there to erupt!

Another issue is that there are many outliers in volcano data. Some volcanoes may have extremely large repose periods or even have erupted only once in history! Similarly, some volcanoes could have extraordinarily large eruptions. The mean is known to be sensitive to outliers, so it is not a great measure to use here.

Problem 3. Multivariate MLE for Pareto Distribution

In Section 3, we were only computing MLEs for univariate parameters. The Pareto distribution has two parameters, θ and β, but last time I turned it into a distribution with one parameter by assuming that β was a known constant.

Now let us drop that assumption and find the multivariate MLE.

(a) Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Pareto}(\theta, \beta)$, where θ and β are unknown parameters. Find the MLE for $(\theta, \beta)^T$.

(b) Is it possible to compute an approximate confidence interval for β using a matrix of Fisher Information? Why or why not?

Solution

(a) The density for the Pareto distribution is

$$f(x_i; \theta, \beta) = \theta \beta^\theta x_i^{-\theta-1},$$

where $\theta > 1$, $\beta > 0$, and $x_i \geq \beta$. We will need to take into account the constraint that $x_i \geq \beta$, which I will do using indicator functions similarly to other problems we have seen. The joint likelihood is

$$\begin{aligned} L(\theta, \beta) &= \prod_{i=1}^{n} \theta \beta^\theta X_i^{-\theta-1} 1\{X_i \geq \beta\} (\text{since the data are iid}) \\ &= \theta^n \beta^{n\theta} \{\prod_{i=1}^{n} X_i\}^{-\theta-1} 1\{\min_{i=1,\ldots,n}(X_i) \geq \beta\}. \end{aligned}$$

First, let us find an optimizer for β. Splitting the likelihood into cases, we have

$$L(\theta, \beta) = \begin{cases} \theta^n \beta^{n\theta} \{\prod_{i=1}^{n} X_i\}^{-\theta-1} & \text{if } \beta \leq \min(X_i) \\ 0 \text{ if } \beta > \min(X_i). \end{cases}$$

Within the first case, in the likelihood we have that $\theta > 0$ and $\beta > 0$. And also $\prod_{i=1}^{n} X_i > 0$, since for each i, $0 < \beta \leq X_i$. So $\theta^n \beta^{n\theta} \{\prod_{i=1}^{n} X_i\}^{-\theta-1} > 0$. Also observe that $\beta^{n\theta}$ is a function that is strictly increasing in β. Therefore, the likelihood with respect to β will be maximized when β is as large as possible, subject to the constraint that $\beta \leq \min(X_i)$. Thus, $\hat{\beta} = \min(X_i)$ is the maximizer.

Next, let us find an optimizer for θ. We can now drop the indicator from the likelihood function, since optimization of θ is not affected by the indicator, which does not contain θ

in it. We now have

$$L(\theta, \beta) \propto \theta^n \hat{\beta}^{n\theta} \{\prod_{i=1}^{n} X_i\}^{-\theta-1}.$$

Normally, we would just directly compute the log-likelihood $l(\theta, \beta)$, compute the partial derivative $\frac{\partial l}{\partial \theta}$, set it equal to 0, and solve for θ. But in this case we can actually just re-use our work from last time*. The critical value for θ was

$$\theta = \frac{n}{\sum_{i=1}^{n} \log(X_i) - n\log(\beta)}.$$

The multivariate optimizer thus should satisfy the system

$$\begin{cases} \beta &= \min(X_i) \\ \theta &= \frac{n}{\sum_{i=1}^{n} \log(X_i) - n\log(\beta)} \end{cases}.$$

Clearly, this is true iff $\hat{\beta} = \min(X_i)$ and $\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log(X_i) - n\log(\hat{\beta})}$

The last step is to verify that $(\hat{\theta}, \hat{\beta})$ is indeed a maximizer for the likelihood. Ordinarily for multivariate optimization, we would have to perform the partial derivative test. But in this case, we analytically found that the maximizer for β is $\hat{\beta}$ (i.e., we didn't use calculus to get that). So all we have to do is check that $\hat{\theta}$ is a maximizer given $\hat{\beta}$. Using a regular second derivative test (similar to last time), we have

$$\left.\frac{\partial l^2}{\partial \theta^2}\right|_{\beta=\hat{\beta}} = \frac{-n}{\theta^2} < 0.$$

Thus, $(\hat{\theta}, \hat{\beta})^T$ is the MLE for $(\theta, \beta)^T$.

* Technical detail: It's because we had the same likelihood last time, just where we assumed β was a known constant. And by definition of the partial derivative, $\frac{dl(\theta)}{d\theta}$ with β treated as a constant is exactly the same as $\frac{\partial l(\theta,\beta)}{\partial \theta}$.

(b) No, we cannot compute an approximate confidence interval for β using a matrix of Fisher Information. The Fisher Information is not defined for β, since the likelihood was not differentiable due to the presence of the indicator.

Problem 4. MLE for Beta Distribution (Numerical Optimization)

Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Beta}(\alpha, \beta)$.

(a) Write a system of equations that would be used to solve for the MLE of α and β.

(b) Write formulas for approximate 95% confidence intervals for α and β.

(c) Compute the MLE numerically for the simulated dataset (included in Section 4 R code).

(d) Plug in the MLEs from (c) into the Hessian calculated in (b). Then obtain the Hessian from the R code to verify that these Hessians are the same.

Solution

(a) The density is

$$f(x_i; \alpha, \beta) = \frac{x_i^{\alpha-1}(1-x_i)^{\beta-1}}{B(\alpha, \beta)},$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, $\alpha > 0$, $\beta > 0$, and $x_i \in (0, 1)$.

The joint likelihood is

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{n} B(\alpha, \beta)^{-1} X_i^{\alpha-1}(1-X_i)^{\beta-1} \\ &= B(\alpha, \beta)^{-n} \left(X_i\right)^{\alpha-1} \left(\prod (1-X_i)\right)^{\beta-1}. \end{aligned}$$

The log-likelihood is

$$\begin{aligned} l(\alpha, \beta) &= -n \log(B(\alpha, \beta)) + (\alpha-1)\sum \log(X_i) + (\beta-1)\sum \log(1-X_i) \\ &= -n\left(\log(\Gamma(\alpha)) + \log(\Gamma(\beta)) - \log(\Gamma(\alpha+\beta))\right) + (\alpha-1)\sum \log(X_i) + (\beta-1)\sum \log(1-X_i). \end{aligned}$$

Taking the partial derivatives, we have

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= -n\left(\frac{\partial}{\partial \alpha}\log(\Gamma(\alpha)) - \frac{\partial}{\partial \alpha}\log(\Gamma(\alpha+\beta))\right) + \sum \log(X_i) \\ \frac{\partial l}{\partial \beta} &= -n\left(\frac{\partial}{\partial \beta}\log(\Gamma(\beta)) - \frac{\partial}{\partial \beta}\log(\Gamma(\alpha+\beta))\right) + \sum \log(1-X_i). \end{aligned}$$

The digamma function is defined as $\psi(k) = \frac{\partial \log(\Gamma(k))}{k}$, which we can substitute in.

$$\frac{\partial l}{\partial \alpha} = -n\left(\psi(\alpha) - \psi(\alpha+\beta)\right) + \sum \log(X_i)$$
$$\frac{\partial l}{\partial \beta} = -n\left(\psi(\beta) - \psi(\alpha+\beta)\right) + \sum \log(1-X_i).$$

Setting each partial derivative equal to 0, we have the system

$$-n\left(\psi(\alpha) - \psi(\alpha+\beta)\right) + \sum \log(X_i) = 0$$
$$\frac{\partial l}{\partial \beta} = -n\left(\psi(\beta) - \psi(\alpha+\beta)\right) + \sum \log(1-X_i).$$

Rearranging, we have

$$\psi(\alpha) - \psi(\alpha+\beta) = \frac{\sum \log(X_i)}{n}$$
$$\psi(\beta) - \psi(\alpha+\beta) = \frac{\sum \log(1-X_i)}{n}.$$

We cannot solve this in closed form, so we have to use numerical methods.

(b) The trigamma function is defined as $\psi_1(k) = \frac{\partial}{\partial k}\psi(k)$, which will come in handy.

Now, the second partial derivatives are

$$\frac{\partial}{\partial \alpha}\left(\frac{\partial l}{\partial \alpha}\right) = -n\psi_1(\alpha) + n\psi_1(\alpha+\beta) = H_{11}$$
$$\frac{\partial}{\partial \alpha}\left(\frac{\partial l}{\partial \beta}\right) = n\psi_1(\alpha+\beta) = H_{12}$$
$$\frac{\partial}{\partial \beta}\left(\frac{\partial l}{\partial \alpha}\right) = n\psi_1(\alpha+\beta) = H_{21}$$
$$\frac{\partial}{\partial \beta}\left(\frac{\partial l}{\partial \beta}\right) = -n\psi_1(\beta) + n\psi_1(\alpha+\beta) = H_{22}.$$

Next, we compute the expectations as

$$H_{11} = -n\psi_1(\alpha) + n\psi_1(\alpha+\beta) = H_{11}$$
$$H_{12} = n\psi_1(\alpha+\beta)$$
$$H_{21} = n\psi_1(\alpha+\beta)$$
$$H_{22} = -n\psi_1(\beta) + n\psi_1(\alpha+\beta),$$

which is quite simple in this case since they are all constants.

The Fisher Information can be computed as $I_n(\alpha,\beta) = nI(\alpha,\beta) = -n\mathbb{E}\left(H(l(\alpha,\beta))\right)$, so we have

$$\begin{aligned} I_n(\alpha,\beta) &= -n\begin{pmatrix}\mathbb{E}(H_{11}) & \mathbb{E}(H_{12})\\ \mathbb{E}(H_{21}) & \mathbb{E}(H_{22})\end{pmatrix} \\ &= -n\begin{pmatrix}-n\psi_1(\alpha)+n\psi_1(\alpha+\beta) & n\psi_1(\alpha+\beta)\\ n\psi_1(\alpha+\beta) & -n\psi_1(\beta)+n\psi_1(\alpha+\beta)\end{pmatrix} \\ &= n^2\begin{pmatrix}\psi_1(\alpha)-\psi_1(\alpha+\beta) & -\psi_1(\alpha+\beta)\\ -\psi_1(\alpha+\beta) & \psi_1(\beta)-\psi_1(\alpha+\beta).\end{pmatrix} \end{aligned}$$

To calculate the asymptotic variance, we need to invert $I_n(\alpha,\beta)$. Using the earlier matrix facts in these notes, we have

$$\begin{aligned} I_n^{-1} &= \frac{1}{n^2}\cdot\frac{1}{\left(\psi_1(\alpha)-\psi_1(\alpha+\beta)\right)\left(\psi_1(\beta)-\psi_1(\alpha+\beta)\right)}\begin{pmatrix}\psi_1(\beta)-\psi_1(\alpha+\beta) & \psi_1(\alpha+\beta)\\ \psi_1(\alpha+\beta) & \psi_1(\alpha)-\psi_1(\alpha+\beta)\end{pmatrix} \\ &= \begin{pmatrix}\frac{1}{n^2(\psi_1(\alpha)-\psi_1(\alpha+\beta))} & \frac{\psi_1(\alpha+\beta)}{n^2(\psi_1(\alpha)-\psi_1(\alpha+\beta))(\psi_1(\beta)-\psi_1(\alpha+\beta))}\\ \frac{\psi_1(\alpha+\beta)}{n^2(\psi_1(\alpha)-\psi_1(\alpha+\beta))(\psi_1(\beta)-\psi_1(\alpha+\beta))} & \frac{1}{n^2(\psi_1(\beta)-\psi_1(\alpha+\beta))}\end{pmatrix}. \end{aligned}$$

To get the approximate standard errors, we just need to grab the relevant entries from the covariance matrix and take the square root. The approximate CI for α is

$$\hat{\alpha} \pm z_{.025}\cdot\frac{1}{n\sqrt{\left(\psi_1(\hat{\alpha})-\psi_1(\hat{\alpha}+\hat{\beta})\right)}}$$

and for β is

$$\hat{\beta} \pm z_{.025}\cdot\frac{1}{n\sqrt{\left(\psi_1(\hat{\beta})-\psi_1(\hat{\alpha}+\hat{\beta})\right)}}$$

(c) Solution in the R code.

(d) Solution in the R code.