3.2 Likelihood equations and errors-in-variables regression; Solari's example..

Here is a case, noted by Solari (1969), where the likelihood equations (3.1.1) have solutions, none of which are maximum likelihood estimates.

Let $(X_i, Y_i),\ i = 1, \ldots, n$, be observed points in the plane. Suppose we want to do a form of "errors in variables" regression, in other words to fit the data by a straight line, assuming normal errors in both variables, so that $X_i = a_i + U_i$ and $Y_i = ba_i + V_i$ where $U_1, \ldots, U_n$ and $V_1, \ldots, V_n$ are all jointly independent, with U_i having distribution $N(0, \sigma^2)$ and V_i distribution $N(0, \tau^2)$ for $i = 1, \ldots, n$. Here the unknown parameters are $a_1, \ldots, a_n,\ b,\ \sigma^2$ and τ^2. Let $c := \sigma^2$ and $h := \tau^2$. Then the joint density is

$$(ch)^{-n/2}(2\pi)^{-n} \exp\left(-\textstyle\sum_{i=1}^n (x_i - a_i)^2/(2c) + (y_i - ba_i)^2/(2h)\right).$$

Let $\sum := \sum_{i=1}^n$. Taking logarithms, the likelihood equations are equivalent to the vanishing of the gradient (with respect to all $n + 3$ parameters) of

$$-(n/2)\log(ch) - \sum[(X_i - a_i)^2/(2c) + (Y_i - ba_i)^2/(2h)].$$

Taking derivatives with respect to c and h gives

$$c \ = \ \sum (X_i - a_i)^2/n, \quad h \ = \ \sum (Y_i - ba_i)^2/n. \tag{3.2.1}$$

For each $i = 1, \ldots, n,\ \partial/\partial a_i$ gives

$$0 \ = \ c^{-1}(X_i - a_i) + h^{-1}b(Y_i - ba_i), \tag{3.2.2}$$

and $\partial/\partial b$ gives

$$\sum a_i(Y_i - ba_i) \ = \ 0. \tag{3.2.3}$$

Next, (3.2.2) implies

$$\sum (X_i - a_i)^2/c^2 \ = \ b^2 \sum (Y_i - ba_i)^2/h^2. \tag{3.2.4}$$

From (3.2.1) it then follows that $1/c \ = \ b^2/h$ and $b \neq 0$. This and (3.2.2) imply $b(X_i - a_i) \ = \ -(Y_i - ba_i)$, so,

$$a_i \ = \ (X_i + b^{-1}Y_i)/2 \ \text{ for } \ i = 1, \ldots, n. \tag{3.2.5}$$

Then from (3.2.1) again,

$$c \ = \ \sum (X_i - b^{-1}Y_i)^2/(4n) \ \text{ and } \ h \ = \ \sum (Y_i - bX_i)^2/(4n). \tag{3.2.6}$$

Using (3.2.5) in (3.2.3) gives

$$\sum (Y_i - bX_i)(X_i + b^{-1}Y_i) \ = \ 0 \ = \ \sum Y_i^2 - b^2 X_i^2.$$

If $\sum Y_i^2 > 0 = \sum X_i^2$ there is no solution for b. Also if $\sum Y_i^2 = 0 < \sum X_i^2$ we would get $b = 0$, a contradiction, so there is no solution in this case. If $\sum X_i^2 = \sum Y_i^2 = 0$ then (3.2.3) gives $a_i = 0$ for all i since $b \neq 0$, but then (3.2.1) gives $c = 0$, a contradiction, so there is no solution. We are left with the general case $\sum Y_i^2 > 0 < \sum X_i^2$. Then $b^2 = \sum Y_i^2 / \sum X_i^2$, so

$$b = \pm \left(\sum Y_i^2 / \sum X_i^2 \right)^{1/2}. \tag{3.2.7}$$

Substituting each of these two possible values of b in (3.2.5) and (3.2.6) then determines values of all the other parameters, giving two points (or one point if all the Y_i are 0) where the likelihood equations hold, in other words *critical points* of the likelihood function, and there are no other critical points.

Note however that the joint density goes to $+\infty$ for $a_i = X_i$, fixed b and h, and $c \downarrow 0$. Thus the above two points cannot give an absolute maximum of the likelihood. On the other hand, as $c \downarrow 0$ for any fixed $a_i \neq X_i$, the likelihood approaches 0. So the likelihood behaves pathologically in the neighborhood of points where $a_i = X_i$ for all i and $c = 0$, its logarithm having what is called an essential singularity. Other such singularities occur where $Y_i - ba_i \to 0$ and $h \downarrow 0$. Here in the parametrization for the natural parameter space, some θ_j have σ^2 in the denominator, so that as $\sigma \downarrow 0$, these natural parameters go to $\pm\infty$, where singular behavior is not so surprising.

Note that the family of densities in the example is exponential, as defined in Sec. 2.5, but that it has no maxima for its likelihoods: this is consistent with Theorem 3.2.2 which only guarantees uniqueness of maximum likelihood estimates when they exist, not existence. Also, the family parametrized by $b, a_1, \ldots, a_n, \sigma^2, \tau^2$ is not the full natural parameter space, but a curved submanifold in it. For example if $n = 2$, θ_i are the coefficients of X_i and θ_{i+2} those of Y_i for $i = 1, 2$, we have $\theta_1 \theta_4 \equiv \theta_2 \theta_3$, so that even if an MLE existed in the family, it would not be guaranteed to be unique.

Having noted that noted that maximum likelihood estimation doesn't work in the formulation given above, let's consider some other formulations.

Let Q_η, $\eta \in Y$ be a family of probability laws on a sample space X where Y is a parameter space. The function $\eta \mapsto Q_\eta$ is called *identifiable* if it's one-to-one, i.e. $Q_\eta \neq Q_\xi$ for $\eta \neq \xi$ in Y. If η is a vector, $\eta = (\eta_1, \ldots, \eta_k)$, a component parameter η_j will be called *identifiable* if laws Q_η with different values of η_j are always distinct. Thus, $\eta \mapsto Q_\eta$ is identifiable if and only if each component $\eta_1, \ldots, \eta_k$ is identifiable. Suppose $\theta \mapsto P_\theta$ for $\theta \in \Theta$ is identifiable and $Q_\eta \equiv P_{\theta(\eta)}$ for some function $\theta(\cdot)$ on Y. Then $\eta \mapsto Q_\eta$ is identifiable if and only if $\theta(\cdot)$ is one-to-one.

Example. Let $dQ_\eta(\psi) = ae^{\cos(\psi - \eta)} d\psi$ for $0 \leq \psi < 2\pi$ where a is the suitable constant, a subfamily of the von Mises-Fisher family. Then $\eta \mapsto Q_\eta$ is not identifiable for $\eta \in Y = \mathbb{R}$, but it is for $Y = [0, 2\pi)$ or $Y = [-\pi, \pi)$.

Now consider another form of errors-in-variables regression, where for $i = 1, \ldots, n$, $X_i = x_i + U_i$, $Y_i = a + bx_i + V_i$, $U_1, \ldots, U_n$ are i.i.d. $N(0, \sigma^2)$ and independent of $V_1, \ldots, V_n$ i.i.d. $N(0, \tau^2)$, all independent of $x_1, \ldots, x_n$ i.i.d. $N(\mu, \zeta^2)$ where $a, b, \mu \in \mathbb{R}$ and $\sigma^2 > 0$, $\tau^2 > 0$ and $\zeta^2 > 0$. This differs from the formulation in the Solari example in that the

x_i, now random variables, were parameters a_i in the example. In the present model, only the variables (X_i, Y_i) for $i = 1, \ldots, n$ are observed and we want to estimate the parameters. Clearly the (X_i, Y_i) are i.i.d. and have a bivariate normal distribution. The means are $EX_i = \mu$ and $EY_i = \nu := a + b\mu$. The parameters μ and ν are always identifiable. It is easily checked that $C_{11} := \mathrm{var}(X_i) = \zeta^2 + \sigma^2$, $C_{22} := \mathrm{var}(Y_i) = b^2\zeta^2 + \tau^2$, and $C_{12} = C_{21} := \mathrm{cov}(X_i, Y_i) = b\zeta^2$. A bivariate normal distribution is given by 5 real parameters, in this case $C_{11}, C_{12}, C_{22}, \mu, \nu$. A continuous function (with polynomial components in this case) from an open set in $\mathbb{R}^6$ onto an open set in $\mathbb{R}^5$ can't be one-to-one (Appendix B), so the 6 parameters $a, b, \mu, \zeta^2, \sigma^2, \tau^2$ are not all identifiable. Let's examine the situation more closely. The equation $\nu = a + b\mu$ can be solved for a. The equation $C_{11} = \zeta^2 + \sigma^2$ can be solved for $\sigma^2 > 0$ if and only if $\zeta^2 < C_{11}$. The equation $C_{22} = b^2\zeta^2 + \tau^2$ can be solved for $\tau^2 > 0$ if and only if $b^2\zeta^2 < C_{22}$. The equation $C_{12} = b\zeta^2$ can be solved for $\zeta^2 > 0$ if and only if C_{12} and b have the same sign (both are positive, both are negative or both are 0).

If $C_{12} > 0$, all three equations for C_{ij}, $1 \le i \le j \le 2$, can be solved whenever $C_{12}/C_{11} < b < C_{22}/C_{12}$. Specifically, then $bC_{12} > C_{12}^2/C_{11} > 0$, so $b > 0$. Then $\zeta^2 = C_{12}/b$, $b^2\zeta^2 = bC_{12} < C_{22}$, and $\zeta^2 = C_{12}/b < C_{11}$ as desired. Thus b can range over a non-empty open interval if $C_{12}^2 < C_{11}C_{22}$, in other words $\det C > 0$, which is true when the covariance matrix C is positive definite (non-singular). Thus b is not identifiable: we can get the same bivariate normal distribution for different values of b, and some other parameters will also differ.

In the exceptional case $C_{12} = 0$, since $\zeta^2 > 0$ we get $b = 0$ so it is identifiable, as are τ^2 and a, while ζ^2 and σ^2 are not.

If $C_{12} < 0$, we will have $b < 0$. Again we get a non-empty open interval of possible values of b, $C_{12}/C_{11} > b > C_{22}/C_{12}$ for $\det C > 0$, so b is not identifiable.

If the original problem is changed so that $a = 0$ and we are looking for a line through the origin, $b = \nu/\mu$ is identifiable unless $\mu = 0$. If $\mu = 0$ then we have non-identifiability much as in the previous cases.

If, instead, we change the problem so that $\lambda := \tau^2/\sigma^2 > 0$ is assumed known, then all the 5 remaining parameters are identifiable (unless $\lambda C_{11} = C_{22}$), as follows. The equation for C_{22} now becomes $C_{22} = b^2\zeta^2 + \sigma^2\lambda$. After some algebra, we get an equation quadratic in b,

$$b^2C_{12} + (\lambda C_{11} - C_{22})b - \lambda C_{12} = 0. \tag{3.2.8}$$

Since $\lambda > 0$, the equation always has real solutions. If $C_{12} = 0$ the equation becomes linear and either $b = 0$ is the only solution or if $\lambda C_{11} - C_{22} = 0$, b can have any value and in this special case is not identifiable. If $C_{12} \neq 0$ there are two distinct real roots for b, of opposite signs. Since b must be of the same sign as C_{12} to satisfy the original equations, there is a unique solution for b, and one can solve for all the parameters, so they are identifiable in this case.

Now, let's consider how the parameters can be estimated, still in case λ is known.

Suppose given a normal distribution $N(\mu, C)$ on $\mathbb{R}^k$ where now $\mu = (\mu_1, \ldots, \mu_k)$, $C := \{C_{ij}\}_{i,j=1}^k$ and we have n i.i.d. observations $X_1, \ldots, X_n \in \mathbb{R}^k$ with $X_r := \{X_{ri}\}_{i=1}^k$.

It's easily seen that the MLE of μ is $\overline{X} := \{\overline{X}_i\}_{i=1}^k$ where $\overline{X}_i := \frac{1}{n}\sum_{r=1}^n X_{ri}$ for $i = 1, \ldots, k$.

3.2.9 Theorem. The MLE of the covariance matrix C of a multivariate normal distribution is the sample covariance matrix

$$\hat{C}_{ij} := \frac{1}{n}\sum_{r=1}^n (X_{ri} - \overline{X}_i)(X_{rj} - \overline{X}_j)$$

for $i, j = 1, \ldots, k$.

Proof. For $k = 1$, this says that the MLE of the variance σ^2 for a normal distribution is $\frac{1}{n}\sum_{r=1}^n (X_r - \overline{X})^2$. (As noted above, this is $(n-1)/n$ times the usual, unbiased estimator of the variance.) This is easily checked, substituting in the MLE $\overline{X}$ of the mean, then finding that the likelihood equation for σ has a unique solution which is easily seen to give a maximum.

Now in k dimensions, consider any linear function f from $\mathbb{R}^k$ into $\mathbb{R}$ such as a coordinate, $f(X_r) = X_{ri}$. Let $\overline{f} := \frac{1}{n}\sum_{r=1}^n f(X_r)$. Then by the one-dimensional facts, $\overline{f}$ is the MLE of the mean of f and the MLE of $\mathrm{var}(f)$ is its sample variance $\frac{1}{n}\sum_{r=1}^n (f(X_r) - \overline{f})^2$.

For any function g on a parameter space, if an MLE $\hat{\theta}$ of the parameter θ exists, then the MLE of $g(\theta)$ is (by definition) $g(\hat{\theta})$. In this sense maximum likelihood estimation is preserved by general functions, unlike unbiased estimation which is not generally preserved except by linear functions. At any rate, we see that the MLE of C_{jj} is $\hat{C}_{jj}$ for $j = 1, \ldots, k$. Moreover, the MLE of $\mathrm{var}(X_{1i} + X_{1j})$ is also its sample variance for any $i, j = 1, \ldots, k$. Subtracting the sample variances of X_{1i} and X_{1j} and dividing by 2, we get the sample covariance of X_{1i} and X_{1j}, namely $\hat{C}_{ij}$. Since the MLE of a sum or difference of functions of the parameters is the sum or difference of the MLEs, we get that the sample covariance $\hat{C}_{ij}$ is indeed the MLE of C_{ij}. □

Returning to the bivariate case and continuing with errors-in-variables regression for a fixed λ, by a change of scale we can assume $\lambda = 1$, in other words $\sigma^2 = \tau^2$. Since $\overline{X}$ and $\overline{Y}$ are the MLEs of the means, we see that $\overline{Y} = \hat{\nu} = a + b\hat{\mu} = a + b\overline{X}$, the MLE regression line will pass through $(\overline{X}, \overline{Y})$, as it also does for the classical regression lines of y on x or x on y.

For a line ℓ and a point $(X, Y) \in \mathbb{R}^2$ the Euclidean distance from (X, Y) to ℓ is

$$d(X, Y), \ell) := \min(|(X, Y) - (\xi, \eta)| : (\xi, \eta) \in \ell).$$

Given $(X_1, Y_1), \ldots, (X_n, Y_n)$, a line $\ell : y = a + bx$ or a vertical line where x is constant will be called an *best-fit-by-squared-distance (bfsd) regression line* if it minimizes $\sum_{j=1}^n d((X_j, Y_j), \ell)^2$.

3.2.10 Proposition. For errors-in-variables regression with $\sigma^2 = \tau^2 > 0$ ($\lambda = 1$), MLE lines are the same as bfsd regression lines, which always exist. The bfsd line is unique unless $\hat{C}_{11} = \hat{C}_{22} = 0$ or $\hat{C}_{11} = \hat{C}_{22} > 0$ and $\hat{C}_{12} = 0$; in case of non-uniqueness, the bfsd lines are all lines through $(\overline{X}, \overline{Y})$.

Proof. For a more thorough exposition on bfsd lines, see Appendix E, where they are called bfd lines. Here we will concentrate on the cases where uniqueness holds.

The slope of the line joining a point (X, Y) to its nearest point (ξ, η) in a line $\ell : \ y = a + bx$ is $-1/b$, and a little algebra or trigonometry gives

$$d((X,Y),\ell)^2 \ = \ (Y - a - bX)^2/(b^2+1).$$

Thus for bfsd regression we want to find

$$\inf_{a,b} \textstyle\sum_{i=1}^n (a + bX_i - Y_i)^2/(b^2+1).$$

Setting $\partial/\partial a = 0$ gives a unique solution $a = \overline{Y} - b\overline{X}$, which is easily seen to give a minimum with respect to a since the quantity being minimized goes to ∞ as $|a| \to \infty$. So the bfsd regression line, like the others, will pass through $\overline{X}, \overline{Y}$. Plugging in the value for a, we now want to find

$$\inf_b \textstyle\sum_{i=1}^n [b(X_i - \overline{X}) - (Y_i - \overline{Y})]^2/(b^2+1). \tag{3.2.11}$$

Taking d/db gives a quadratic equation in b,

$$0 \ = \ (b^2-1)\hat{C}_{12} + b(\hat{C}_{11} - \hat{C}_{22}). \tag{3.2.12}$$

This equation is the same as (3.2.8) with $\lambda = 1$ and with C_{ij} replaced by their MLEs $\hat{C}_{ij}$. Also, b is a solution of (3.2.12) if and only if $-1/b$ is. If $b\hat{C}_{12} < 0$, then the sum in (3.2.11) becomes smaller if we replace b by $-b$. Thus the infimum can be restricted to values of b having the same sign as $\hat{C}_{12}$. As $b \to \pm\infty$, the quantity being minimized in (3.2.11) approaches $\sum_{i=1}^n (X_i - \overline{X})^2$. If the latter is 0 we do get an infimum, so a vertical line through $(\overline{X}, \overline{Y})$ is a bfsd line. Or, if the Y_i are all equal, $b = 0$ yields a minimum equal to 0 and the horizontal line through $(\overline{X}, \overline{Y})$ is a bfsd line. If all the points (X_i, Y_i) are equal, the bfsd lines are all lines through $(\overline{X}, \overline{Y})$. For further cases of non-uniqueness see Appendix E. In the cases of uniqueness, the coefficients of our quadratic polynomial in b are non-zero. So, bfsd regression does give the same line as maximum likelihood estimation for errors-in-variables regression with $\lambda = 1$. □

PROBLEM

1. In errors-in-variables regression with $\lambda = \tau^2/\sigma^2 = 1$, suppose we observe $\hat{C}_{11} = 1.44$, $\hat{C}_{12} = 0.66$, and $\hat{C}_{22} = 1.21$. Find the MLE of the slope b of the regression line (i.e. the bfsd line).

NOTES

Solari (1969) showed that the critical point at which the likelihood is larger in her example is actually a saddle point, and so not even a local maximum of the likelihood. Some earlier authors had thought it was a maximum.

REFERENCE

Solari, Mary E. (1969). The "maximum likelihood solution" of the problem of estimating a linear functional relationship. *J. Roy. Statist. Soc.* **31**, 372-375.