ESTIMATION — SOME EXAMPLES

1. The simple linear regression model

Recall that in this model, we have some non-random design points $x_1 < \cdots < x_n$ with $n \geq 3$, we observe some random variables $Y_1, ..., Y_n$, and the model says that for some real a and b and some ε_i i.i.d. $N(0, \sigma^2)$ for some unknown $\sigma > 0$, we have

$$Y_j = a + bx_j + \varepsilon_j, \quad j = 1, ..., n. \tag{1}$$

In y-on-x regression, one estimates a and b by minimizing $\sum_{j=1}^n (Y_j - a - bx_j)^2$. Gauss was apparently the first to show (in 1809) that maximum likelihood estimation, using the assumption on the ε_j, gives the same estimates of a and b. It also gives us a way of estimating σ^2.

Theorem 1. *(a) Maximum likelihood estimation of the parameters in the model (1) gives the same estimates of a and b as does y-on-x least squares regression.*
(b) Let S be the minimum with respect to a and b of $\sum_{j=1}^n (Y_j - a - bx_j)^2$. Then the maximum likelihood estimate of σ^2 is S/n.

Proof. The model (1) gives that $\varepsilon_j = Y_j - a - bx_j$ are i.i.d. $N(0, \sigma^2)$, so for given x_j and Y_j the likelihood as a function of a, b, and σ^2 is

$$(\sigma\sqrt{2\pi})^{-n} \prod_{j=1}^n \exp\left(-\frac{(Y_j - a - bx_j)^2}{2\sigma^2}\right) \tag{2}$$

$$= (\sigma\sqrt{2\pi})^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^n (Y_j - a - bx_j)^2\right).$$

For any fixed $\sigma > 0$, this is maximized by minimizing the sum in the exponent, proving part (a). By the assumption on the x_j we know that $s_x^2 > 0$, and that the estimates $\widehat{a}$ of a and $\widehat{b}$ of b satisfy $\overline{Y} = \widehat{a} + \widehat{b}\overline{x}$. The estimated slope $\widehat{b}$ equals

$$\widehat{b} = \mathrm{scov}(x, Y)/s_x^2. \tag{3}$$

The minimum value S is attained for this value of $\widehat{b}$ and $\widehat{a} = \overline{Y} - \widehat{b}\overline{x}$. For $n = 2$ we will definitely have $S = 0$ since for $x_1 < x_2$ there is a

Date: 18.650, Sept. 25, 2015.

line through (x_j, Y_j) for $j = 1$ and 2, but for $n \geq 3$ as we assumed, it will be shown that $S > 0$. In the expression that was minimized with respect to b to evaluate b, if $S = 0$ we would have for $b = \widehat{b}$

$$0 = (n-1)\left[s_Y^2 - \frac{2\,\mathrm{scov}(x,Y)^2}{s_x^2} + \frac{(\mathrm{scov}(x,Y))^2}{s_x^2}\right] = s_Y^2 - \frac{\mathrm{scov}(x,Y)^2}{s_x^2}.$$

It follows that

$$\text{(4)} \qquad \mathrm{scov}(x,Y)^2 = s_x^2 s_Y^2.$$

For each $j = 1, ..., n$ we have $Y_j = a + bx_j + \varepsilon_j$, and so $\overline{Y} = a + b\overline{x} + \overline{\varepsilon}$ and

$$\text{(5)} \qquad Y_j - \overline{Y} = b(x_j - \overline{x}) + \varepsilon_j - \overline{\varepsilon}.$$

For all three of the vectors $\xi = \{\xi_j\}_{j=1}^n$ being considered, $\xi = Y$, x, or ε, we have $\sum_{j=1}^n \xi_j - \overline{\xi} = 0$. $\{x_j - \overline{x}\}_{j=1}^n$ is a fixed vector, but $\{\varepsilon_j - \overline{\varepsilon}\}_{j=1}^n$ has a distribution all over the $(n-1)$-dimensional hyperplane

$$\mathbb{R}_0^n := \{\{\eta_j\}_{j=1}^n : \sum_{j=1}^n \eta_j = 0\},$$

where $n - 1 \geq 2$ since $n \geq 3$.

We can view $\mathrm{scov}(x, Y)$ as the dot product of the fixed vector $\{x_j - \overline{x}\}_{j=1}^n$ and the random vector $\{Y_j - \overline{Y}\}_{j=1}^n$.

Now $\widehat{b} = 0$ is equivalent by (3) and (4) to $s_Y^2 = 0$ and so to equality of all Y_j, but then by (5), $\{\varepsilon_j - \overline{\varepsilon}\}_{j=1}^n$ is a multiple of the fixed vector $\{x_j - \overline{x}\}_{j=1}^n$, which occurs with probability 0 (even though b is a random variable) since $n \geq 3$.

So with probability 1, $\widehat{b} \neq 0$, $\mathrm{scov}(x, Y) \neq 0$ and $s_Y^2 > 0$. But then (4) implies that the two vectors $\{Y_j - \overline{Y}\}_{j=1}^n$ and $\{x_j - \overline{x}\}_{j=1}^n$ are proportional, which by (5) implies that so are $\{\varepsilon_j - \overline{\varepsilon}\}_{j=1}^n$ and $\{x_j - \overline{x}\}_{j=1}^n$, which occurs only with probability 0 since the latter is a fixed vector and the former is distributed over the $n-1$-dimensional subspace $\mathbb{R}_0^n$. It follows that $S > 0$ with probability 1.

Then to maximize with respect to σ, since the (natural) logarithm is a strictly increasing, differentiable function, is equivalent to maximizing

$$\text{(6)} \qquad -(n/2)\log(2\pi) - n\log(\sigma) - \frac{S}{2\sigma^2}.$$

We have $s_x > 0$ and with probability 1, $s_Y > 0$ and $S > 0$. So (6) goes to $-\infty$ as $\sigma \downarrow 0$. It also goes to $-\infty$ as $\sigma \uparrow +\infty$. So to find a maximum in the interior $0 < \sigma < +\infty$ we can differentiate (6) with respect to σ, giving $-n/\sigma + S/\sigma^3$, or $\sigma^2 = S/n$, proving (b). □

2. Estimating parameters of gamma distributions

For $0 < \alpha < \infty$ and $0 < \lambda < \infty$ the $\Gamma(\alpha, \lambda)$ distribution has the density

$$f_{\alpha,\lambda}(x) = \lambda^\alpha x^{\alpha-1} e^{-\lambda x}/\Gamma(\alpha)$$

for $x > 0$ and 0 for $x \leq 0$, where the gamma function is defined by

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx.$$

Suppose we've observed $X_1, ..., X_n$ i.i.d. with a $\Gamma(\alpha, \lambda)$ density and want to estimate the parameters α and λ. The likelihood function is

$$f(X, \alpha, \lambda) = \prod_{j=1}^{n} \lambda^\alpha X_j^{\alpha-1} e^{-\lambda X_j}/\Gamma(\alpha) = \lambda^{n\alpha} T_n^{\alpha-1} \exp(-\lambda S_n)/\Gamma(\alpha)^n$$

where $T_n = \prod_{j=1}^n X_j$ and $S_n = \sum_{j=1}^n X_j$. To maximize this is equivalent to maximizing its log, as $\log(\cdot)$ is an increasing function. The log is $LL(X, \alpha, \lambda)$ defined by

$$n\alpha \log(\lambda) + (\alpha - 1)\log(T_n) - \lambda S_n - n \log \Gamma(\alpha). \tag{7}$$

First, for any given $\alpha > 0$, let's look for a maximum with respect to λ. $LL(X, \alpha, \lambda) \to -\infty$ as $\lambda \downarrow 0$. The gamma distribution implies that all $X_j > 0$ with probability 1, and so $S_n > 0$, which implies that $LL(X, \alpha, \lambda) \to -\infty$ as $\lambda \to +\infty$. So we're looking for an interior maximum with respect to λ given α, for which we set

$$0 = \partial LL(X, \alpha, \lambda)/\partial\lambda = n\alpha/\lambda - S_n,$$

which gives $\alpha/\lambda = S_n/n = \overline{X}$, or $\lambda = \alpha/\overline{X}$. Note that the expectation of X_1 is α/λ, so setting this equal to $\overline{X}$ is as in the method of moments.

So let's plug $\lambda = \alpha/\overline{X}$ into (7), giving

$$n\alpha \log(\alpha/\overline{X}) + (\alpha - 1)\log(T_n) - (\alpha/\overline{X})S_n - n \log \Gamma(\alpha)$$

$$= n\alpha[\log(\alpha) - \log(\overline{X})] + (\alpha - 1)\log(T_n) - n\alpha - n \log \Gamma(\alpha). \tag{8}$$

This quantity goes to $-\infty$ as $\alpha \to +\infty$ because $\log(\Gamma(\alpha))$ via a Stirling formula for the gamma function is asymptotic to $(\alpha - \frac{1}{2})\log(\alpha)$, so terms $\pm n\alpha \log(\alpha)$ cancel and leave $-n\alpha$ as the dominant term.

As $\alpha \downarrow 0$ we can see how $\Gamma(\alpha)$ behaves as follows. We have the recurrence formula $\Gamma(\alpha + 1) \equiv \alpha\Gamma(\alpha)$ which one gets by integrating by parts in the definition of gamma function. We have $\Gamma(1) = 0! = 1$, and the gamma function is continuous in a neighborhood of 1. Thus as $\alpha \downarrow 0$, $\alpha\Gamma(\alpha) = \Gamma(\alpha + 1)$ converges to 1, and

$$\log(\Gamma(\alpha)) + \log(\alpha) = \log(\Gamma(\alpha) - \log(1/\alpha) \to 0,$$

so $\log(\Gamma(\alpha)) \to +\infty$, and (8) goes to $-\infty$. So to look for an interior maximum, differentiating (8) with respect to α gives

$$n[\log(\alpha) - \log(\overline{X})] + n + \log(T_n) - n - n\Gamma'(\alpha)/\Gamma(\alpha)$$

$$= n[\log(\alpha) - \log(\overline{X})] + \log(T_n) - n\Gamma'(\alpha)/\Gamma(\alpha).$$

Setting this equal to 0 doesn't give a nice closed-form solution. The function $\Gamma'(\alpha)/\Gamma(\alpha)$ is called digamma(α). R has this function, but still, it takes some numerical search work to find the maximum of (8).

So it's much easier to estimate α and λ by the method of moments.

3. A disaster for unbiased estimation

Suppose one can observe a positive integer-valued random variable X which has a Poisson(λ) distribution conditional on $X > 0$. This might be the number of radioactive decay particles of a certain type emitted by a sample of matter. If the number was 0, it could be either that the sample is not radioactive, or that it is, but the number X happened to be 0. So there could be interest in estimating $e^{-\lambda}$, the probability of 0 for a Poisson(λ) distribution.

It will be shown that given $X = k$ for $k \geq 1$, there is a unique unbiased estimator of $e^{-\lambda}$, and it is $(-1)^{k+1}$.

Let $T_k = T(k)$ be the value of an unbiased estimator of $e^{-\lambda}$ when $X = k$ for $k \geq 1$. We have

$$\Pr(X = k | X > 0) = \frac{e^{-\lambda}\lambda^k}{k!(1 - e^{-\lambda})}.$$

Thus unbiasedness says

$$e^{-\lambda} = \sum_{k=1}^{\infty} T_k \frac{e^{-\lambda}\lambda^k}{k!(1 - e^{-\lambda})},$$

or equivalently

$$1 - e^{-\lambda} = \sum_{k=1}^{\infty} T_k \frac{\lambda^k}{k!}.$$

If two power series represent the same function, their coefficients must be equal. So the Taylor series of $e^{-\lambda}$ gives $(-1)^{k+1} = T_k$ for all $k \geq 1$. This is an absurd estimator. A reasonable estimator of $e^{-\lambda}$ should give a number between 0 and 1 which is small when X is large. So unbiasedness may not be a good way to choose an estimator.

4. Notes on history

Stigler (1974) wrote a historical paper on regression (even polynomial regression), and gives a reference to Gauss (1809). Stigler, himself a statistician, has published a number of other works on history of statistics.

REFERENCES

Gauss, Carl Friedrich (1809). *Theorie motus corporum coelestium.* Translated as *Theory of motion of Heavenly Bodies Moving About the Sun in Conic Sections.* Repr. 1963, Dover, New York.

Stigler, Stephen M. (1974). Gergonne's 1815 paper on the design and analysis of polynomial regression experiments. *Historia Mathematica* **1**, 431-447.