Microeconometrics

Based on the textbooks
Verbeek: *A Guide to Modern Econometrics*
and Cameron and Trivedi: *Microeconometrics*

Robert M. Kunst
robert.kunst@univie.ac.at

University of Vienna
and
Institute for Advanced Studies Vienna

December 4, 2013

Outline

The concept of the likelihood

Assume a parametric model as the data-generating law, i.e. a density depending on a parameter vector $\theta \in \mathbb{R}^p$ and the observations $x \in \mathbb{R}^N$:

$$L(\theta; x)$$

is a probability density in the second portion of arguments x for given θ, maybe smooth with a unique mode. In the first portion θ for given x, it is not a density. Thus, it is called the **likelihood**. One may surmise, however, that, for a given x, the value θ that makes the observed x 'most probable' (has it as its mode, maximizes the likelihood) constitutes a good estimate for the generating θ: **maximum likelihood** (ML). This is indeed often the case.

The example of a normal distribution

Assume data are randomly (independently) generated from a normal $\mathcal{N}(\mu, \sigma^2)$ distribution. The full likelihood model reads

$$L(\mu, \sigma^2; x_1, \ldots, x_N) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x_i - \mu)^2}{2\sigma^2}\}.$$

This function is easier to handle in logarithms: the **log likelihood**:

$$\log L(\mu, \sigma^2; x_1, \ldots, x_N) = -\frac{N}{2} \log 2\pi - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2.$$

Taking derivatives and equating them to 0 yields first-order conditions.

Maximum likelihood for the normal distribution

Equating the derivative for μ, $\frac{\partial \log L}{\partial \mu}$, to 0 yields the arithmetic mean

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i = \bar{x}.$$

The first-order condition for σ^2,

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{N} (x_i - \mu)^2 = 0$$

yields, after substituting $\bar{x}$ for μ,

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2.$$

This ML estimator for σ^2 is biased in finite samples.

Maximum likelihood for simple normal regression

The model for simple normal regression has a likelihood very similar to the previous example:

$$\log L(\beta_1, \beta_2, \sigma^2; x_1, \ldots, x_N, y_1, \ldots, y_N) =$$
$$-\frac{N}{2}\log 2\pi - \frac{N}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2.$$

The maximum with respect to β_1, β_2 is taken if the sum in the third term is minimized: ordinary least squares. The maximum with respect to σ^2 appears at

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2 = \frac{1}{N}\sum_{i=1}^{N} e_i^2,$$

the ML estimator for σ^2 with its typical small-sample bias.

The likelihood for multiple normal regression

The likelihood for multiple normal regression extends the simple case slightly:

$$\log L(\beta, \sigma^2; x_1, \ldots, x_N, y_1, \ldots, y_N) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i'\beta)^2$$

The first derivatives of the log-likelihood are called the **scores**:

$$\frac{\partial \log L(\beta, \sigma^2)}{\partial \beta} = \sum_{i=1}^{N} \frac{y_i - x_i'\beta}{\sigma^2} x_i$$

$$\frac{\partial \log L(\beta, \sigma^2)}{\partial \sigma^2} = \sum_{i=1}^{N} -\frac{1}{2\sigma^2} + \frac{(y_i - x_i'\beta)^2}{2\sigma^4}$$

Maximum likelihood for multiple normal regression

Equating the scores to 0 yields the OLS estimators

$$\hat{\beta} = \left(\sum_{i=1}^{N} x_i x_i' \right)^{-1} \sum_{i=1}^{N} x_i y_i = (X'X)^{-1} X'\mathbf{y}$$

and the ML variance estimator

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} e_i^2$$

with a tendency toward a downward bias.

Contributions and information matrix

Note that the log-likelihood as well as the scores are represented as sums of terms indexed i:

$$\log L(\theta|X) = \sum_{i=1}^{N} \log L_i(\theta|x_i), \qquad \left.\frac{\partial \log L_i(\theta|x_i)}{\partial \theta}\right|_{\hat{\theta}} = \sum_{i=1}^{N} s_i(\hat{\theta}) = 0.$$

The terms $\log L_i$ and s_i are called the (log-likelihood and score) **contributions**.

The expectation matrix of the products of score contributions is called the **information in observation** i:

$$I_i(\theta) = \mathrm{E}\left\{\frac{\partial \log L_i(\theta)}{\partial \theta}\frac{\partial \log L_i(\theta)'}{\partial \theta}\right\} = \mathrm{E}\{s_i(\theta)s_i'(\theta)\}.$$

Why maximum likelihood?

It can be shown that, under weak *regularity conditions*, that

1. The ML estimator is **consistent**, i.e. $\operatorname{plim} \hat{\theta} = \theta$;
2. The ML estimator is **asymptotically normally distributed**, i.e.
$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V);$$
3. The ML estimator is **asymptotically efficient**, i.e. it has the smallest variance among all consistent and asymptotically normal estimators. V corresponds to the **Cramèr-Rao lower bound**.

Cramèr-Rao

The celebrated **Cramèr-Rao inequality** says that under some 'regularity conditions', it will hold for any asymptotically normal and consistent estimator $\hat{\theta}$ that

$$\lim_{N\to\infty} N\mathrm{V}(\hat{\theta}) \geq I(\theta)^{-1}.$$

For the ML estimator, $V = I(\theta)^{-1}$, so the lower bound is attained. This (Fisher) information (matrix) $I(\theta)$ is usually defined as the limit for $N \to \infty$ of

$$\bar{I}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\{s_i(\theta)s_i'(\theta)\} = \frac{1}{N}\sum_{i=1}^{N} I_i(\theta),$$

evaluated at the true parameter value θ.

A typical catalog of regularity conditions

The following assumptions suffice for ML asymptotic normality:

- Observations are independent and identically distributed;
- True value is in the interior of the non-empty parameter space;
- True value is identified;
- Log-likelihood is twice continuously differentiable in θ in an open neighborhood of the true value;
- Expectation of the log-likelihood exists at the true value;
- Average log-likelihood converges almost surely to the expectation and uniformly in θ;
- Information matrix exists and is non-singular at the true value.

Some assumptions can be exchanged against others etc.

The information matrix equality

Under some regularity conditions, the information matrix can also be represented by the second derivatives of the log-likelihood:

$$I(\theta) = \lim_{N\to\infty} \mathrm{E}\left\{\frac{1}{N}\frac{\partial \log L(\theta)}{\partial \theta}\frac{\partial \log L(\theta)'}{\partial \theta}\right\} = -\lim_{N\to\infty} \mathrm{E}\left\{\frac{1}{N}\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'}\right\}$$

If the model is incorrectly specified, this equality may not hold. This is why it is often used as a basis for statistical hypothesis tests on correct specification.

Information matrices for OLS estimation

For ML/OLS in multiple regression, the information matrix is easily shown to be block diagonal:

$$\bar{I}_N(\beta, \sigma^2) = \begin{pmatrix} \frac{1}{N\sigma^2} X'X & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

Thus, the asymptotic variance matrix has the form

$$N\mathrm{V}(\hat{\beta}, \hat{\sigma}^2) \to I(\theta)^{-1} = \begin{pmatrix} \sigma^2 \Sigma_{xx}^{-1} & 0 \\ 0 & 2\sigma^4 \end{pmatrix},$$

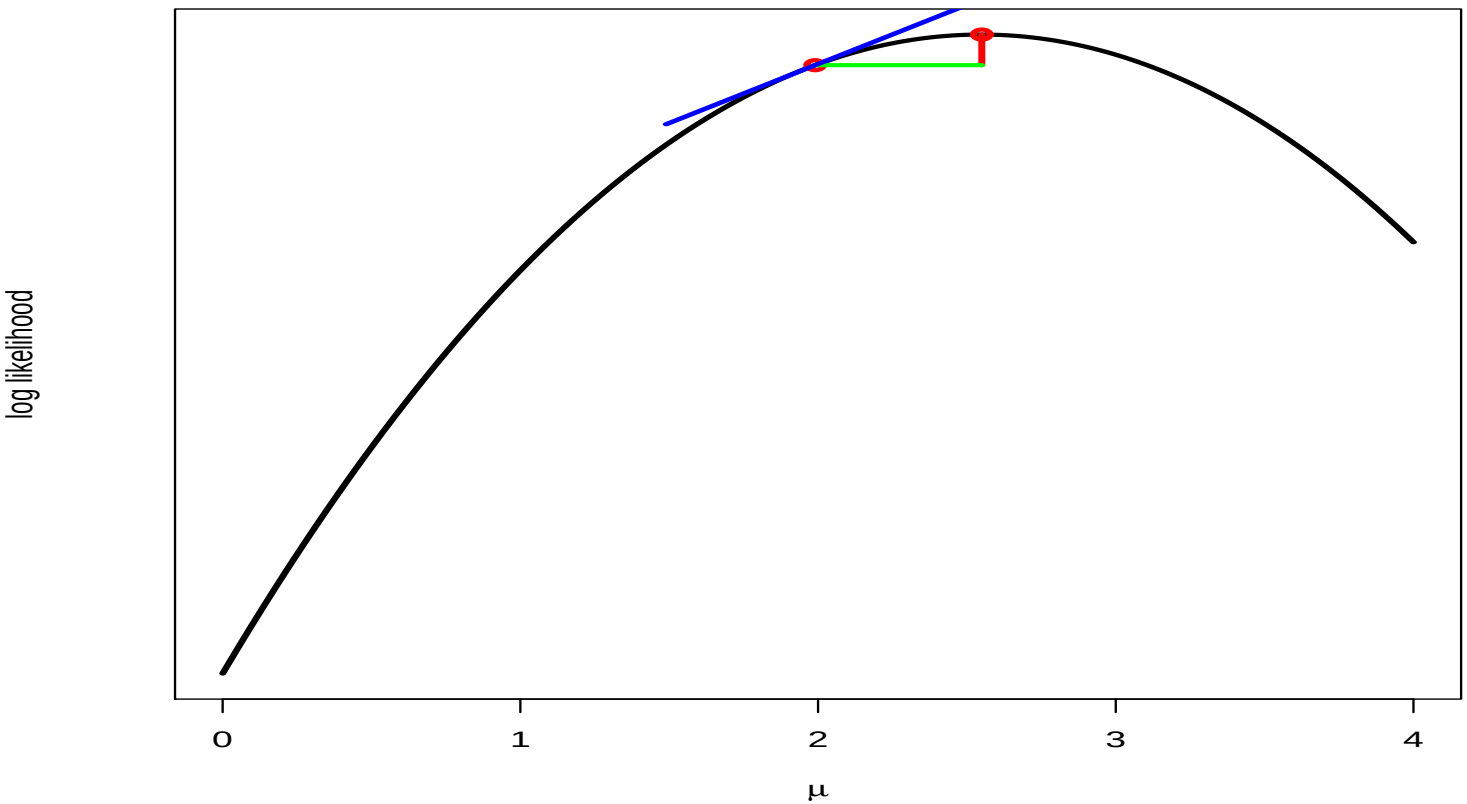
with $\Sigma_{xx} = \lim_{N\to\infty} N^{-1}(X'X)$.

The trinity of test principles

Most parametric hypothesis tests are based on one of three likelihood-based principles:

1. The **likelihood ratio** (LR) principle considers the ratio of likelihoods maximized under the null H_0 and under the general model $H_0 \cup H_A$;
2. The **Wald** principle considers the maximum likelihood under the general model and checks whether this maximum fulfils the restrictions that define H_0. Formally, it does not require maximizing the likelihood under H_0;
3. The **Lagrange multiplier** (LM) or score test considers the maximum likelihood under the null and checks the derivative of the general likelihood (score) around the maximum. Formally, it does not require maximizing the likelihood for the general model.

A visualization of the test trinity



Test $H_0 : \mu = 2$. The LR test statistic (red) is based on the vertical distance, the Wald test statistic (green) on the horizontal distance. The score test statistic (blue) measures the slope at H_0.

The Wald test

Assume the null is defined by a linear restriction $R\theta = q$ within the general space $\Theta \subset \mathbb{R}^K$ for some $J \times K$–matrix R. Then,

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V)$$

implies

$$\sqrt{N}(R\hat{\theta} - R\theta) \xrightarrow{d} \mathcal{N}(0, RVR').$$

Then, the statistic

$$\xi_W = N(R\hat{\theta} - q)'\{R\hat{V}R'\}^{-1}(R\hat{\theta} - q),$$

with $\hat{V}$ consistently estimating V, will be asymptotically distributed as $\chi^2(J)$ under the $H_0 : R\theta = q$.

The likelihood ratio test

Suppose $\hat{\theta}$ is the ML estimate for θ under the general hypothesis, and $\tilde{\theta}$ is the ML estimate under the null (restricted). Then, under quite general conditions, the statistic

$$\xi_{LR} = 2\{\log L(\hat{\theta}) - \log L(\tilde{\theta})\}$$

will be $\chi^2(J)$ distributed under the null. The **asymptotic equivalence** of the LR, LM, and Wald tests is a general statistical result that will also hold under very general conditions.

The Lagrange multiplier test

Suppose $\theta' = (\theta_1', \theta_2')$ and $H_0 : \theta_2 = 0$. Constrained optimization can be seen as equating the derivatives of the Lagrangian

$$H(\theta, \lambda) = \sum_{i=1}^{N} \log L_i(\theta) - \lambda' \theta_2$$

to 0. Under H_0, the derivatives on θ_2,

$$\lambda = \sum_{i=1}^{N} \frac{\partial \log L_i(\theta)}{\partial \theta_2}\bigg|_{\tilde{\theta}} = \sum_{i=1}^{N} s_{i2}(\tilde{\theta}),$$

will be close to 0. One can show that

$$\xi_{LM} = \sum_{i=1}^{N} s_i(\tilde{\theta})' \left\{ \sum_{i=1}^{N} s_i(\tilde{\theta}) s_i(\tilde{\theta})' \right\}^{-1} \sum_{i=1}^{N} s_i(\tilde{\theta})$$

is, under H_0, asymptotically distributed $\chi^2(J)$.

LM tests in regression form

Explicitly consider LM testing in a regression model. With $H_0 : \theta_2 = 0$, the first part $\sum_{i=1}^{N} s_{i1}(\tilde{\theta}) = 0$, and ξ_{LM} becomes approximately

$$\tilde{e}'X(X'\tilde{e}\tilde{e}'X)^{-1}X'\tilde{e} \approx \sigma^{-2}\tilde{e}'X(X'X)^{-1}X'\tilde{e},$$

as score contributions are $\tilde{e}_i x_i/\sigma^2$ with $\tilde{e}$ denoting residuals from regressing y on the first part X_1. This expression is close to NR^2 from a regression of the $\tilde{e}$ residuals on *all* regressors X_1 and X_2. Only if all additional variables can be assumed to be uncorrelated with the first portion, one might also use

$$\sigma^{-2}\tilde{e}'X_2(X_2'X_2)^{-1}X_2'\tilde{e},$$

a regression of the residuals on just the additional regressors.

Example 1: LM test for omitted variables

This test for the null $H_0 : \theta_2 = 0$, i.e. for

$$H_0 : y = X_1\theta_1 + \varepsilon, \quad H_A : y = X_1\theta_1 + X_2\theta_2 + \varepsilon,$$

is run in two stages:

1. Regress y on X_1; keep residuals $\tilde{e}$;
2. Regress $\tilde{e}$ on X_1 and X_2; keep NR^2 which is under H_0 distributed $\chi^2(J)$ with $J = \dim(X_2)$.

This test can be seen as a *restriction test* (search for simplifications within a correct model) and also as a *specification test* (check correctness of specification, check validity of assumptions).

Example 2: Jarque-Bera test

The test by Jarque and Bera tests for 'H_0: errors normally distributed'. It uses the skewness (third moment) condition

$$\mathrm{E}(\varepsilon^3) = 0$$

and the kurtosis (fourth moment) condition

$$\mathrm{E}(\varepsilon^4) = 3\{\mathrm{E}(\varepsilon^2)\}^2,$$

based on the empirical moments of residuals. It is really an LM–test, its test statistic

$$\xi = N \left\{ \frac{1}{6} \left(\frac{1}{N} \sum_{i=1}^{N} \frac{e_i^3}{\hat{\sigma}^3} \right)^2 + \frac{1}{24} \left(\frac{1}{N} \sum_{i=1}^{N} \frac{e_i^4}{\hat{\sigma}^4} - 3 \right)^2 \right\}$$

is distributed $\chi^2(2)$ under H_0.

Quasi Maximum Likelihood

Estimating parameters by Gaussian ML, although no normal distribution is assumed or even known to be invalid, is called **quasi maximum likelihood** estimation. Under regularity conditions, it can be shown that

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V),$$

with

$$V = I(\theta)^{-1} J(\theta) I(\theta)^{-1},$$

with I an information matrix defined from second derivatives and J defined from products of scores. For correct specification, $I = J$.

White has suggested to test for correct specification by checking the equality of the two definitions of information matrices: **information matrix test**.