ADVANCED ECONOMETRIC METHODS

(R300)

Koen Jochmans

University of Cambridge

Last revised on November 25, 2020

ESTIMATION IN PARAMETRIC PROBLEMS

Reading

Evaluation of estimators:

Casella and Berger, Chapters 7 and 10

Hansen I, Chapter 6

Asymptotics for the sample mean:

Goldberger, Chapter 9

Hansen I, Chapters 7 and 8

Maximum likelihood:

Davidson and MacKinnon, Chapter 8

Hansen I, Chapter 10

Wooldridge, Chapter 13

Linear regression:

Goldberger, Chapters 14–16

Hansen II, Chapters 2–5

Estimation

We are interested in saying something about a population based on a sample

$$x_1, \dots, x_n$$

from it. The x_i may be scalars or vectors.

Want to learn some feature of the population, say θ; the estimand.

For that we use an estimator

$$\theta_n = \theta_n(x_1, \dots, x_n);$$

this is just a function of the sample (it can be any function).

Some questions:

- How do we evaluate estimators, i.e., what is a good estimator?
- How do we construct good estimators?
- Does there exist a best estimator and, if so, is it unique?

This inferential aim is different from a descriptive data analysis that gives means, variances, correlations, regression coefficients, and so on.

The parametric framework

A way to formalize sampling is to see $x_1, \dots, x_n$ as a draw from an (n-variate) probability (mass or density) function, g.

We begin with the random sampling and the parametric framework.

The sample is a random sample if

1. the x_i are independent across i, so that $g(x_1, \dots, x_n) = \prod_{i=1}^n f_i(x_i)$ for probability functions $f_1, \dots, f_n$; and
2. all x_i are identically distributed, so that $f_i = f$ for some f and all i.

We say that the x_i are i.i.d.

The parametric framework says that $f = f_\theta$ is known up to parameter θ which is finite dimensional (and so is a vector, in general).

That is, we know the class

$$\{f_\theta : \theta \in \Theta\},$$

but not the particular θ that generated the data.

We know the whole probability distribution once we know the parameter θ.

We may calculate $P_\theta(x_i \in A) = \int_A f_\theta(x)\,dx$ for any set A. For example,

$$F_\theta(x) = P_\theta(x_i \in (-\infty, x]) = P_\theta(x_i \leq x)$$

for any x (the cumulative distribution function).

We know all raw and centered moments; for example, the mean and variance

$$E_\theta(x_i) = \int x\, f_\theta(x)\,dx, \qquad \mathrm{var}_\theta(x_i) = \int (x - E_\theta(x_i))\,(x - E_\theta(x_i))'\, f_\theta(x)\,dx,$$

and so on.

We know $E_\theta(\varphi(x_i))$ for any chosen function φ and so also parameters ψ defined through

$$E_\theta(\varphi(x_i;\psi)) = 0,$$

(which we call moment conditions). Obvious example is $\psi = E_\theta(x_i)$, which has $\varphi(x_i;\psi) = x_i - \psi$.

For univariate x_i the τth-quantile is $q_\tau = \inf_q\{q : F_\theta(q) \geq \tau\}$, for $\tau \in (0,1)$. It is a solution to the moment condition $E_\theta(\{x_i \leq \psi\} - \tau) = 0$ and so has $\varphi(x_i;\psi) = \{x_i \leq \psi\} - \tau$.

Examples

Let $x_1, \ldots, x_n$ be a sequence of zeros and ones with $P_\theta(x_i = 1) = \theta$. Then x_i is Bernoulli with mass function

$$f_\theta(x) = \theta^x \, (1-\theta)^{1-x}, \qquad \theta \in (0,1),$$

for $x \in \{0,1\}$.

One possible (and sensible) estimator of θ would be the sample frequency of ones, i.e., $\overline{x}_n = n^{-1} \sum_{i=1}^n x_i$

Another simple example has $x_1, \ldots, x_n$ representing the number of arrivals per unit of time. The Poisson distribution has

$$f_\theta(x) = P_\theta(x_i = x) = \frac{\theta^x \, e^{-\theta}}{x!}, \qquad \theta > 0$$

for $x \in \mathbb{N}$.

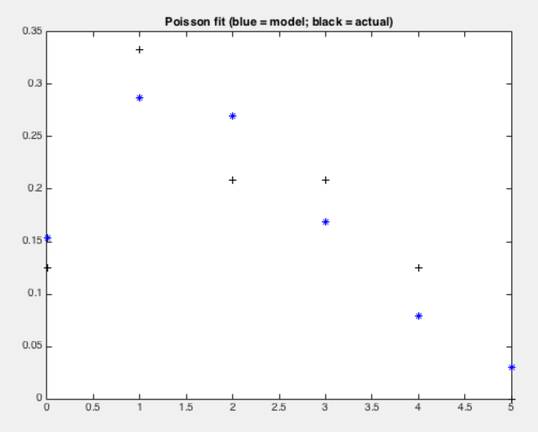
θ is the arrival rate, i.e., the expected number of arrivals per time unit.

A sensible estimator of θ is again the sample mean.

We have data on the number of births per hour over a 24 hour period in Addenbrooke's.

Fitting a Poisson model to such data we estimate the number of births per hour by the sample mean, here 1.875 births/hour.

Given an estimate of θ we can estimate the mass function.



The hospital data also tell us that, of the 44 babies, 18 were boys and 26 where girls.

The maximum likelihood estimator of the probability of giving birth to a boy is $18/44 = .409$.

The estimator is a random variable.

Using arguments to be developed later we can test whether there is a gender bias at Addenbrooke's.

The standard error on our estimate is $\sqrt{(18/44) \times (26/44)/44} = .074$ which gives us the value

$$\frac{.409 - .500}{.074} = -1.23$$

for a test statistic which is (asymptotically) standard normal under the null of no gender bias.

Using a Neyman-Pearson argument (see later) we cannot reject the absence of gender bias (at conventional significance levels).

A continuous example with two parameters is the normal distribution.

The univariate standard-normal density is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2};$$

it has mean zero and variance one. The corresponding distribution function is

$$\Phi(x) = \int_{-\infty}^{x} \phi(u)\, du.$$

The normal distribution is a location/scale family:

If $z_i \sim N(0,1)$, then

$$x_i = \mu + \sigma z_i \sim N(\mu, \sigma^2).$$

Its cumulative distribution function is

$$P(x_i \leq x) = P\left(z_i \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

and its density function is

$$\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}.$$

Obvious estimators for μ, σ^2 would be the sample mean and sample variance.

Less obvious is when

$$x_i^* \sim N(\mu, \sigma^2)$$

but we observe

$$x_i = \begin{cases} x_i^* & \text{if } x_i^* \geq 0 \\ 0 & \text{if } x_i^* < 0 \end{cases} .$$

This is a censored normal variable.

How would we estimate the parameters here?

Some obvious candidates would be

- the sample mean and variance of the x_i;
- the sample mean and variance of the positive x_i.

These turn out not to be very attractive and should not be used.

We will construct a better estimator later on.

As a final example, suppose that $x_i \sim \chi^2_\theta$.

The Chi-squared distribution with (integer) θ degrees of freedom has density

$$f_\theta(x) = \frac{x^{\theta/2-1}\, e^{-x/2}}{2^{\theta/2}\, \Gamma(\theta/2)},$$

where $\Gamma(\theta) = \int_0^\infty x^{\theta-1}\, e^{-x}\, dx$ denotes the Gamma function at θ.

We note without proof that

- $E_\theta(x_i) = \theta$,
- $\text{var}_\theta(x_i) = 2\theta$,
- $E_\theta(x_i^p) = 2^p \Gamma(p + \theta/2)/\Gamma(\theta/2)$.

It follows that the sample mean is an obvious candidate estimator of θ.

But there is also information in higher-order moments so the sample mean may be inefficient (it is here but it need not be, a priori; see the Poisson example).

Change of variable

Let x be a random variable with density f. Let $y = \varphi(x)$ for an invertible function φ.

The density of y is

$$f(\varphi^{-1}(y)) \left|\det\left(\frac{\partial\varphi^{-1}(y)}{\partial y'}\right)\right|.$$

Easiest to see in the univariate case:

If φ is increasing,

$$P(y \leq a) = P(\varphi(x) \leq a) = P(x \leq \varphi^{-1}(a)),$$

and differentiation with respect to a gives

$$f(\varphi^{-1}(a)) \left.\frac{\partial\varphi^{-1}(y)}{\partial y}\right|_{y=a}$$

by the chain rule.

If φ is decreasing, $P(y \leq a) = 1 - P(x \leq \varphi^{-1}(a))$, and differentiation gives $-f(\varphi^{-1}(a))\,(\partial\varphi^{-1}(y)/\partial y\big|_{y=a})$.

Characteristic function

Let x be a continuous univariate random variable with density f. Then

$$\varphi(t) = E(e^{\iota t x}) = \int e^{\iota t x} f(x)\, dx$$

is its characteristic function. (Here, ι is the imaginary unit, i.e., $\iota^2 = -1$)

So, φ is the Fourier transform of f.

Like f, φ completely characterizes the random variable.

f can be recovered from φ through the inverse Fourier transform

$$f(x) = \frac{1}{2\pi} \int e^{-\iota t x} \varphi(t)\, dt.$$

Further, raw moments equal

$$E(x^p) = \iota^{-p} \left. \frac{\partial^p \varphi(t)}{\partial t^p} \right|_{t=0}.$$

For multivariate x, $\varphi(t) = E(e^{\iota t' x})$ for a vector t of conformable dimension.

An example is the standard normal case. Here,

$$f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}, \qquad \varphi(t) = e^{-t^2/2}.$$

We have, using the definition of the cosine function,

$$\begin{aligned}\varphi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{\iota tx}\, e^{-x^2/2}\,dx &= \frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty} \tfrac{1}{2}(e^{\iota tx} + e^{-\iota tx})\, e^{-x^2/2}\,dx \\ &= \frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty} \cos(tx)\, e^{-x^2/2}\,dx.\end{aligned}$$

Next,

$$\begin{aligned}\varphi'(t) &= \frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty} \tfrac{\partial \cos(tx)}{\partial t}\, e^{-x^2/2}\,dx \\ &= -\frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty} \sin(tx)\, x\, e^{-x^2/2}\,dx \\ &= \frac{2}{\sqrt{2\pi}}\left(e^{-x^2/2}\sin(tx)\right)\Big|_{0}^{+\infty} - \frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty} t\,\cos(tx)\, e^{-x^2/2}\,dx \\ &= -t\,\varphi(t).\end{aligned}$$

This implies that $\varphi \propto e^{-t^2/2}$. But because $\int f(x)\,dx = 1$ we must have that $\varphi(0) = 1$ so that, indeed, $\varphi = e^{-t^2/2}$.

To see that

$$\phi(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-\iota tx} e^{-t^2/2}\, dt,$$

we can use the same calculations.

Moreover, note that

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-\iota tx} e^{-t^2/2}\, dt = \tfrac{1}{\pi}\int_{0}^{+\infty} \cos(tx) e^{-t^2/2}\, dt$$

by the same argument as before. We have already computed the last integral.

Moreover,

$$\frac{1}{\pi}\int_{0}^{+\infty} \cos(tx) e^{-t^2/2}\, dt = \tfrac{1}{\pi}\left(\tfrac{\sqrt{2\pi}}{2}\,\varphi(x)\right) = \tfrac{1}{\sqrt{2\pi}} e^{-x^2/2} = \phi(x),$$

as claimed.

Squared standard-normal variable

Let $z \sim N(0,1)$. Then the density of $x = z^2$ at $a > 0$ is

$$\frac{\phi(\sqrt{a}) + \phi(-\sqrt{a})}{2\sqrt{a}} = \frac{1}{2\sqrt{a}} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a} + \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a} \right) = \frac{1}{\sqrt{2\pi}\sqrt{a}} e^{-\frac{1}{2}a}.$$

This is the density of a χ_1^2 random variable. Indeed, use that $\Gamma(1/2) = \sqrt{\pi}$ to rewrite the density as

$$\frac{1}{\sqrt{2\pi}\sqrt{a}} e^{-\frac{1}{2}a} = \frac{a^{1/2-1} e^{-a/2}}{2^{1/2}\sqrt{\pi}} = \frac{a^{1/2-1} e^{-a/2}}{2^{1/2}\Gamma(1/2)},$$

which co-incides with the definition given above.

Sum of squared independent standard-normal variables

The characteristic function of a χ^2_p-variable is

$$\varphi_p(t) = (1 - 2\imath t)^{-p/2}.$$

So, if

$$z_i \sim N(0,1),$$

then $z_i^2 \sim \chi^2_1$ has $\varphi_1(t) = (1 - 2\imath t)^{-1/2}$.

The characteristic function of $\sum_{i=1}^n z_i^2$ is (by independence) equal to

$$\prod_{i=1}^n \varphi_1(t) = \left((1 - 2\imath t)^{-1/2}\right)^n = (1 - 2\imath t)^{-n/2} = \varphi_n(t).$$

Hence,

$$\sum_{i=1}^n z_i^2 \sim \chi^2_n.$$

Sum of independent normal variables

The characteristic function of a $N(\mu, \sigma^2)$ variable is

$$\varphi_{\mu,\sigma^2}(t) = e^{\iota t \mu - \sigma^2 \frac{t^2}{2}}.$$

So, if

$$z_i \sim N(0,1)$$

are independent the characteristic function of $\sum_{i=1}^{n} z_i$ is

$$\prod_{i=1}^{n} \varphi_{0,1}(t) = \prod_{i=1}^{n} (e^{-t^2/2}) = (e^{-t^2/2})^n = e^{-n\,t^2/2} = \varphi_{0,n}(t),$$

i.e., $\sum_{i=1}^{n} z_i \sim N(0,n)$.

By the location/scale properties of the normal we then have that

$$\overline{z}_n \sim N(0, n^{-1})$$

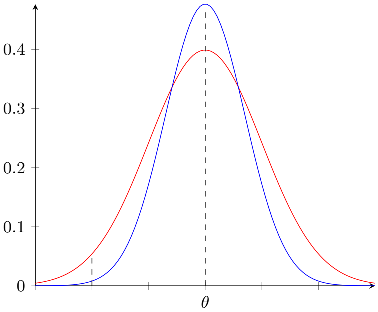
and

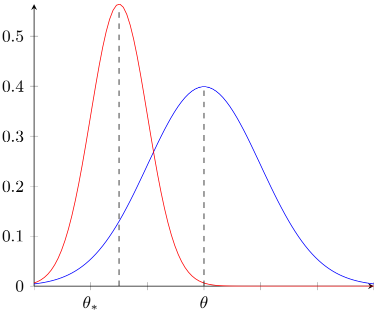
$$\overline{x}_n = \mu + \sigma\,\overline{z}_n \sim N(\mu, \sigma^2/n).$$

Motivating best unbiasedness

Sampling distributions of several estimators.

Blue is better than red.





Best unbiased estimator

An estimator θ_n is best unbiased if $E_\theta(\theta_n) = \theta$ and

$$\mathrm{var}_\theta(\theta_n) \leq \mathrm{var}_\theta(\theta_*)$$

for any other unbiased estimator θ_*.

Here and later, the inequality is to be interpreted in the matrix sense: $A \geq 0$ means that matrix A is positive semi-definite, i.e., $x' A\, x \geq 0$ for any real non-zero vector x.

A lower bound on the variance can be found.

This is called an efficiency bound; here: Cramér-Rao bound.

Very often such an estimator will not exist.

If it exists, it is unique.

Non-existence of bias

The Cauchy distribution with location μ and scale γ has the symmetric density

$$\frac{1}{\pi\gamma\left(1+\left(\frac{x-\mu}{\gamma}\right)^2\right)}.$$

It has no moments.

For example, with $\mu = 0$ and $\gamma = 1$ we have

$$E(|x|) = \lim_{M\to\infty} 2\int_0^M \tfrac{1}{\pi}\tfrac{x}{1+x^2}\,dx = \lim_{M\to\infty} \tfrac{\log(1+M^2)}{\pi} = +\infty.$$

So it is not useful to estimate the location parameter μ via the sample mean. A sensible estimator would be the sample median, which is well defined in spite of the non-existence of moments.

So, if an estimator is Cauchy distributed its bias does not exist.

An example where this happens is with ratios of normal variates, as these are Cauchy.

Ratio of normals

Take independent scalar normal variates $x \sim N(0, \sigma_1^2)$ and $y \sim N(0, \sigma_2^2)$.

Consider the transformation $(x, y) \to (u, v) = (x/y, y)$. The Jacobian of the transformation is v and so the density of u is

$$f_{\sigma_1,\sigma_2}(u) = \int_{-\infty}^{+\infty} \frac{\phi((uv)/\sigma_1)}{\sigma_1} \frac{\phi(v/\sigma_2)}{\sigma_2} \, |v| \, dv.$$

This is (using that $\phi(u) = e^{-u^2/2}/\sqrt{2\pi}$ and that $\phi(u) = \phi(-u)$ for all u)

$$\begin{aligned} f_{\sigma_1,\sigma_2}(u) &= \frac{1}{\pi\sigma_1\sigma_2} \int_0^{+\infty} e^{-\frac{1}{2}v^2((1/\sigma_2)^2+(u/\sigma_1)^2)} \, v \, dv \\ &= \frac{1}{\pi\sigma_1\sigma_2} \int_0^{+\infty} e^{-\frac{1}{2}v^2(1+u^2(\sigma_2/\sigma_1)^2)/\sigma_2^2} \, v \, dv \\ &= \frac{1}{\pi\sigma_1\sigma_2\left((1+u^2(\sigma_2/\sigma_1)^2)/\sigma_2^2\right)} \int_0^{+\infty} \left(-e^{-\frac{1}{2}v^2(1+u^2(\sigma_2/\sigma_1)^2)/\sigma_2^2}\right)' dv \\ &= \frac{1}{\pi\frac{\sigma_1}{\sigma_2}\left(1+\left(\frac{u}{\sigma_1/\sigma_2}\right)^2\right)}, \end{aligned}$$

which is a Cauchy distribution with location zero and scale σ_1/σ_2.

Fisher information

Let

$$\frac{\partial \log f_\theta(x)}{\partial \theta}$$

be the score.

The score has mean zero:

$$E_\theta\left(\frac{\partial \log f_\theta(x_i)}{\partial \theta}\right) = \int \frac{\partial \log f_\theta(x)}{\partial \theta}\, f_\theta(x)\, dx = \int \frac{\partial f_\theta(x)}{\partial \theta}\, dx = 0$$

(where the last step follows from f_θ integrating to one.)

The variance of the score,

$$I_\theta = \operatorname{var}_\theta\left(\frac{\partial \log f_\theta(x_i)}{\partial \theta}\right),$$

is called the Fisher information.

When the score is a vector the information is a (variance-covariance) matrix.

Information inequality

Theorem 1 (Cramér-Rao bound)

Under regularity conditions,

$$\mathrm{var}_\theta(\theta_n) \geq I_\theta^{-1}/n$$

for any unbiased estimator θ_n of θ.

More information reduces the variance bound.

The bound shrinks like n^{-1}.

From the proof (to follow) we have that θ_n attains the bound if and only if

$$n\, I_\theta\, (\theta_n - \theta) = \sum_{i=1}^{n} \left. \frac{\partial \log f_\theta(x_i)}{\partial \theta} \right|_\theta$$

Proof (for the scalar case).

Differentiating the zero-bias condition

$$E_\theta(\theta_n - \theta) = \int \ldots \int (\theta_n(x_1, \ldots, x_n) - \theta) \prod_i f_\theta(x_i)\, dx_1 \ldots dx_n = 0,$$

gives

$$\int \ldots \int \left\{ (\theta_n(x_1, \ldots, x_n) - \theta) \, \frac{\partial \prod_i f_\theta(x_i)}{\partial \theta} - \prod_i f_\theta(x_i) \right\} dx_1 \ldots dx_n = 0.$$

Because densities integrate to one and an identity below we can re-write as

$$\int \cdots \int (\theta_n(x_1, \ldots, x_n) - \theta) \sum_i \frac{\partial \log f_\theta(x_i)}{\partial \theta} \prod_i f_\theta(x_i)\, dx_1 \ldots dx_n = 1.$$

But this is just

$$E_\theta \left((\theta_n - \theta) \sum_i \frac{\partial \log f_\theta(x_i)}{\partial \theta} \right) = 1.$$

Furthermore, this is a covariance (as both terms have zero mean), and so

$$1 = \text{cov}_\theta \left(\theta_n, \sum_i \frac{\partial \log f_\theta(x_i)}{\partial \theta} \right)^2 \leq \text{var}_\theta(\theta_n) \times \text{var}_\theta \left(\sum_i \frac{\partial \log f_\theta(x_i)}{\partial \theta} \right)$$

(by Cauchy-Schwarz). The result then follows from

$$\text{var}_\theta \left(\sum_i \frac{\partial \log f_\theta(x_i)}{\partial \theta} \right) = n\, I_\theta.$$

□

Proof Annex: Identity used in Step 2.

Above we used the following:

$$\begin{aligned}
\sum_{i=1}^{n} \frac{\partial \log f_\theta(x_i)}{\partial \theta} &= \sum_{i=1}^{n} \frac{1}{f_\theta(x_i)} \frac{\partial f_\theta(x_i)}{\partial \theta} \\
&= \sum_{i=1}^{n} \left(\frac{\prod_{j \neq i} f_\theta(x_j)}{\prod_j f_\theta(x_j)} \right) \frac{\partial f_\theta(x_i)}{\partial \theta} \\
&= \frac{\sum_{i=1}^{n} \prod_{j \neq i} f_\theta(x_j) \frac{\partial f_\theta(x_i)}{\partial \theta}}{\prod_j f_\theta(x_j)} = \frac{\frac{\partial \prod_i f_\theta(x_i)}{\partial \theta}}{\prod_j f_\theta(x_j)}
\end{aligned}$$

(using the chain rule on the differentiation of a product), so that we obtain

$$\frac{\partial \prod_{i=1}^{n} f_\theta(x_i)}{\partial \theta} = \left(\sum_{i=1}^{n} \frac{\partial \log f_\theta(x_i)}{\partial \theta} \right) \left(\prod_{j=1}^{n} f_\theta(x_j) \right),$$

the integral of which is an expectation. □

Cauchy-Schwarz inequality

Theorem 2 (Cauchy-Schwarz)

For scalar random variables x_i and y_i

$$E(x_i y_i)^2 \leq E(x_i^2)\ E(y_i^2).$$

Equally, for a sample of size n,

$$\left(n^{-1}\sum_{i=1}^{n} x_i y_i\right)^2 \leq \left(n^{-1}\sum_{i=1}^{n} x_i^2\right)\ \left(n^{-1}\sum_{i=1}^{n} y_i^2\right).$$

Information equality

A useful result is the folowing alternative characterization of the information.

Theorem 3 (Information equality)

$$\operatorname{var}_\theta\left(\frac{\partial \log f_\theta(x_i)}{\partial\theta}\right) = I_\theta = -E_\theta\left(\frac{\partial^2 \log f_\theta(x_i)}{\partial\theta\partial\theta'}\right).$$

We will need this later when establishing optimality of maximum likelihood.

Proof.

Differentiating

$$\int \frac{\partial \log f_\theta(x)}{\partial \theta}\, f_\theta(x)\, dx = 0$$

under the integral sign gives

$$\int \frac{\partial^2 \log f_\theta(x)}{\partial \theta \partial \theta'}\, f_\theta(x)\, dx + \int \frac{\partial \log f_\theta(x)}{\partial \theta}\, \frac{\partial f_\theta(x)}{\partial \theta'}\, dx = 0.$$

Because

$$\frac{\partial \log f_\theta(x)}{\partial \theta} = \frac{1}{f_\theta(x)}\, \frac{\partial f_\theta(x)}{\partial \theta},$$

we have $\frac{\partial f_\theta(x)}{\partial \theta} = \frac{\partial \log f_\theta(x)}{\partial \theta}\, f_\theta(x)$ and so we obtain

$$E_\theta\left(\frac{\partial^2 \log f_\theta(x_i)}{\partial \theta \partial \theta'}\right) + E_\theta\left(\frac{\partial \log f_\theta(x_i)}{\partial \theta}\, \frac{\partial \log f_\theta(x_i)}{\partial \theta'}\right) = 0.$$

Re-arrangement then yields the result. □

If it exists, the best unbiased estimator is unique

Theorem 4 (Uniqueness)

If θ_n^A *and* θ_n^B *are such that*

$$E_\theta(\theta_n^A) = E_\theta(\theta_n^B) = \theta, \qquad \mathrm{var}_\theta(\theta_n^A) = \mathrm{var}_\theta(\theta_n^B) = I_\theta^{-1}/n,$$

then $\theta_n^A = \theta_n^B$.

Proof (for the scalar case).

Define a third estimator θ_n^C through the linear combination

$$\theta_n^c = \lambda\,\theta_n^A + (1-\lambda)\,\theta_n^B, \qquad \lambda \in (0,1).$$

Then $E_\theta(\theta_n^C) = \lambda\,E_\theta(\theta_n^A) + (1-\lambda)\,E_\theta(\theta_n^B) = \theta$, so θ_n^C is also unbiased, and

$$\mathrm{var}_\theta(\theta_n^C) = \lambda^2 \mathrm{var}_\theta(\theta_n^A) + (1-\lambda)^2 \mathrm{var}_\theta(\theta_n^B) + 2\;\lambda(1-\lambda)\,\mathrm{cov}_\theta(\theta_n^A, \theta_n^B).$$

Now, $\mathrm{var}_\theta(\theta_n^A) = \mathrm{var}_\theta(\theta_n^B) = I_\theta^{-1}/n$ by efficiency and

$$|\mathrm{cov}_\theta(\theta_n^A, \theta_n^B)| \le \mathrm{std}_\theta(\theta_n^A)\,\mathrm{std}_\theta(\theta_n^B) = I_\theta^{-1}/n$$

by Cauchy-Schwarz. Thus,

$$\mathrm{var}_\theta(\theta_n^C) \le I_\theta^{-1}/n.$$

The inequality cannot be strict because θ_n^A and θ_n^B are best-unbiased. So we must have that $|\mathrm{corr}_\theta(\theta_n^A, \theta_n^B)| = 1$ which happens iff

$$\theta_n^A = a + b\,\theta_n^B$$

for constants a, b. Now we have that $b = 1$ as $\mathrm{var}_\theta(\theta_n^A) = \mathrm{var}_\theta(\theta_n^B)$ and $a = 0$ as $E_\theta(\theta_n^A) = E_\theta(\theta_n^B)$. □

Bernoulli

With binary data we have

$$\log f_\theta(x) = x\log(\theta) + (1-x)\log(1-\theta),$$

so that

$$\begin{aligned}\frac{\partial \log f_\theta(x)}{\partial\theta} &= \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)},\\ \frac{\partial^2 \log f_\theta(x)}{\partial\theta^2} &= -\frac{\theta(1-\theta)+(x-\theta)(1-2\theta)}{\theta^2(1-\theta)^2} = -\frac{(x-\theta)^2}{\theta^2(1-\theta)^2}.\end{aligned}$$

Clearly,

$$E_\theta\left(\frac{\partial \log f_\theta(x_i)}{\partial\theta}\right) = \frac{E_\theta(x_i-\theta)}{\theta(1-\theta)} = \frac{P_\theta(x_i=1)-\theta}{\theta(1-\theta)} = 0.$$

Further note that, here,

$$\left(\frac{\partial \log f_\theta(x)}{\partial\theta}\right)^2 = -\frac{\partial^2 \log f_\theta(x)}{\partial\theta^2},$$

and so the same holds on taking expectations. This immediately verifies the information equality.

Note that

$$E_\theta(x_i^2) = E_\theta(x_i)$$

when $x_i \in \{0,1\}$.

So,

$$\mathrm{var}_\theta(x_i) = E_\theta((x_i-\theta)^2) = E_\theta(x_i^2 - 2x_i\theta + \theta^2) = \theta(1-\theta)$$

The information thus is

$$I_\theta = \frac{\mathrm{var}_\theta(x_i)}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)}.$$

The efficiency bound for θ is

$$\frac{\theta(1-\theta)}{n}.$$

Note that this is a concave function in θ (and so is maximized at $\theta = 1/2$).

The sample-mean theorem immediately implies that $\overline{x}_n$ is the best unbiased estimator of θ.

Sample-mean theorem

Theorem 5 (Sample-mean theorem)

Let $\overline{x}_n$ be the mean of a random sample $x_1, \ldots, x_n$ from a distribution with finite mean and variance μ, σ^2. Then

$$E(\overline{x}_n) = \mu, \qquad \mathrm{var}(\overline{x}_n) = \sigma^2/n,$$

no matter the distribution of the x_i.

Proof.

By linearity of the expectations operator in the first step and by random sampling in the second step,

$$E(\overline{x}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \mu.$$

Next,

$$\mathrm{var}(\overline{x}_n) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{\mathrm{var}(\sum_{i=1}^{n} x_i)}{n^2} = \frac{\sum_{i=1}^{n} \mathrm{var}(x_i)}{n^2} = \frac{\sigma^2}{n},$$

again by random sampling. □

Poisson

Here we have

$$\log f_\theta(x) = x\,\log(\theta) - \theta + \text{constant}.$$

So,

$$\frac{\partial \log f_\theta(x)}{\partial\theta} = \frac{x}{\theta} - 1, \qquad \frac{\partial^2 \log f_\theta(x)}{\partial\theta^2} = -\frac{x}{\theta^2}.$$

We note the mean/variance equality of a Poisson distribution:

$$E_\theta(x_i) = \textstyle\sum_{x=0}^\infty x\,\frac{e^{-\theta}\theta^x}{x!} = \sum_{x=1}^\infty \frac{e^{-\theta}\theta^x}{(x-1)!} = \sum_{x=0}^\infty \frac{e^{-\theta}\theta^{x+1}}{x!} = \theta\,\sum_{x=0}^\infty \frac{e^{-\theta}\theta^x}{x!} = \theta$$

(because $\sum_{x=0}^\infty \frac{e^{-\theta}\theta^x}{x!} = \sum_{x=0}^\infty f_\theta(x) = 1$), and similarly,

$$E_\theta(x_i^2) = \textstyle\sum_{x=0}^\infty x^2\,\frac{e^{-\theta}\theta^x}{x!} = \theta\,\sum_{x=0}^\infty (x+1)\,\frac{e^{-\theta}\theta^x}{x!} = \theta^2 + \theta,$$

so that $\text{var}_\theta(x_i) = E_\theta(x_i^2) - E_\theta(x_i)^2 = \theta$.

Then $I_\theta = 1/\theta$ and

$$\theta/n$$

is the efficiency bound.

It is again immediate (by the sample-mean theorem) that $\overline{x}_n$ will be best unbiased.

Normal distribution

Here,

$$\log f_\theta(x) = -\frac{1}{2}\left(\log\sigma^2 + \frac{(x-\mu)^2}{\sigma^2}\right) + \text{constant}.$$

So,

$$\begin{aligned}\frac{\partial \log f_\theta(x)}{\partial\mu} &= \frac{(x-\mu)}{\sigma^2},\\ \frac{\partial \log f_\theta(x)}{\partial\sigma^2} &= -\frac{1}{2}\left(\frac{1}{\sigma^2} - \frac{(x-\mu)^2}{\sigma^4}\right),\end{aligned}$$

and

$$\frac{\partial^2 \log f_\theta(x)}{\partial\theta\partial\theta'} = -\begin{pmatrix} \frac{1}{\sigma^2} & \frac{(x-\mu)}{\sigma^4} \\ \frac{(x-\mu)}{\sigma^4} & -\frac{1}{2\sigma^4} + \frac{(x-\mu)^2}{\sigma^6} \end{pmatrix}.$$

The information now is the (diagonal) matrix

$$I_\theta = -E_\theta\left(\frac{\partial^2 \log f_\theta(x_i)}{\partial\theta\partial\theta'}\right) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix},$$

so that the efficiency bounds for μ and σ^2 are σ^2/n and $2\sigma^4/n$, respectively.

The sample mean is again best unbiased for μ.

An unbiased estimator of σ^2 is

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2.$$

However, it does not hit the efficiency bound (see below).

In fact, as

$$\sum_{i=1}^{n}\frac{\partial \log f_\theta(x)}{\partial\sigma^2}=-\frac{1}{2}\left(\frac{n}{\sigma^2}-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^4}\right)=\frac{n}{2\sigma^4}\left(\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n}-\sigma^2\right)$$

depends on μ we cannot have proportionality of the sampling error of any estimator when μ is unknown; the best unbiased estimator is $n^{-1}\sum_i(x_i-\mu)^2$, which is infeasible.

It follows that the efficiency bound is not attainable for σ^2. Moreover, a best unbiased estimator of σ^2 does not exist.

First start with the obvious estimator of σ^2 that is

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2.$$

This estimator is biased:

$$\begin{aligned}
E\left(n^{-1}\textstyle\sum_{i=1}^{n}(x_i - \overline{x}_n)^2\right) &= E((x_i - \overline{x}_n)^2) \\
&= E(((x_i - \mu) - (\overline{x}_n - \mu))^2) \\
&= E((x_i - \mu)^2) - 2E((x_i - \mu)(\overline{x}_n - \mu)) + E((\overline{x}_n - \mu)^2) \\
&= \mathrm{var}_\theta(x_i) - 2\mathrm{cov}_\theta(x_i, \overline{x}_n) + \mathrm{var}_\theta(\overline{x}_n) \\
&= \sigma^2 - 2\sigma^2/n + \sigma^2/n \\
&= \sigma^2 - \sigma^2/n \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}$$

The bias arises from estimating the population mean by the sample mean.

The estimator $\overline{x}_n$ has a variance, σ^2/n, and covaries with each datapoint x_i, with covariance σ^2/n.

An unbiased estimator is therefore

$$\tilde{\sigma}^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2;$$

the change in the numerator is called a degrees of freedom correction.

We can show (see below) that

$$(n-1)\frac{\tilde{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}{\sigma^2} \sim \chi^2_{n-1},$$

and we know the variance of a χ^2_{n-1} is $2(n-1)$.

Hence,

$$\text{var}(\tilde{\sigma}^2) = \left(\frac{\sigma^2}{n-1}\right)^2 \text{var}\left((n-1)\frac{\tilde{\sigma}^2}{\sigma^2}\right) = \frac{2\sigma^4}{n-1}$$

which exceeds the efficiency bound $2\sigma^4/n$.

Sampling distribution of normal variance

First,

$$\begin{aligned}
(n-1)\frac{\tilde{\sigma}^2}{\sigma^2} &= \frac{\sum_{i=1}^n (x_i - \overline{x}_n)^2}{\sigma^2} \\
&= \sum_{i=1}^n \left(\frac{(x_i-\mu)-(\overline{x}_n-\mu)}{\sigma}\right)^2 = \sum_{i=1}^n \left(\frac{x_i-\mu}{\sigma}\right)^2 - \sum_{i=1}^n \left(\frac{\overline{x}_n-\mu}{\sigma}\right)^2 \\
&= \sum_{i=1}^n \left(\frac{x_i-\mu}{\sigma}\right)^2 - n\left(\frac{\overline{x}_n-\mu}{\sigma}\right)^2 = \sum_{i=1}^n \left(\frac{x_i-\mu}{\sigma}\right)^2 - \left(\frac{\overline{x}_n-\mu}{\sigma/\sqrt{n}}\right)^2 .
\end{aligned}$$

The right-hand side terms are χ^2_n and χ^2_1, respectively. The characteristic function of a χ^2_p is $(1-2\iota t)^{-p/2}$.

Second, $\overline{x}_n$ and $\tilde{\sigma}^2$ are independent by Basu's theorem.

Third, the characteristic function of the sum of independent variables is the product of their characteristic functions, so $(n-1)\,\tilde{\sigma}^2/\sigma^2$ has characteristic function

$$(1-2\iota t)^{-n/2}(1-2\iota t)^{1/2} = (1-2\iota t)^{-(n-1)/2},$$

so it is χ^2_{n-1}.

Tobit

$x_i^* \sim N(\mu, \sigma^2)$.

The data are top-coded at c, i.e.,

$$x_i = \begin{cases} x_i^* & \text{if } x_i^* < c \\ c & \text{if } x_i^* \geq c \end{cases} .$$

The density is

$$f_\theta(x) = \left(\frac{1}{\sigma} \phi \left(\frac{x-\mu}{\sigma} \right) \right)^{\{x<c\}} \times \left(1 - \Phi \left(\frac{c-\mu}{\sigma} \right) \right)^{\{x=c\}} .$$

Let us focus on the mean parameter μ here. So we assume that σ is known.

Note that

$$\frac{\partial \log f_\theta(x)}{\partial \mu} = \{x < c\} \frac{(x-\mu)}{\sigma^2} + \{x = c\} \frac{\phi((c-\mu)/\sigma)/\sigma}{1 - \Phi((c-\mu)/\sigma)};$$

both coded and non-coded observations will contribute to the likelihood.

The probability of not being top-coded and being top coded are

$$\Phi\left(\frac{c-\mu}{\sigma}\right), \qquad 1-\Phi\left(\frac{c-\mu}{\sigma}\right),$$

respectively.

Further,

$$f_\theta(x|\, x<c) = \frac{1}{\sigma}\frac{\phi((x-\mu)/\sigma)}{\Phi((c-\mu)/\sigma)};$$

so that the deviation of the mean (from μ) of this truncated distribution is

$$\frac{\int_{-\infty}^{c} \frac{(x-\mu)}{\sigma}\,\phi\left(\frac{(x-\mu)}{\sigma}\right)\,dx}{\Phi\left(\frac{c-\mu}{\sigma}\right)} = \frac{-\sigma^2\int_{-\infty}^{c} \frac{\partial(\phi((x-\mu)/\sigma)/\sigma)}{\partial x}\,dx}{\Phi\left(\frac{c-\mu}{\sigma}\right)} = -\sigma\frac{\phi(\frac{c-\mu}{\sigma})}{\Phi\left(\frac{c-\mu}{\sigma}\right)}.$$

Using these results it is immediate that

$$E_\theta\left(\frac{\partial \log f_\theta(x_i)}{\partial\mu}\right) = 0.$$

After some more calculus, $\partial^2 \log f_\theta(x)/\partial^2\mu$ is found to be

$$-\{x<c\}\,\frac{1}{\sigma^2}-\{x=c\}\,\frac{1}{\sigma^2}\frac{\phi((c-\mu)/\sigma)}{1-\Phi(c-\mu)/\sigma)}\left(\frac{\phi((c-\mu)/\sigma)}{1-\Phi(c-\mu)/\sigma)}-\frac{c-\mu}{\sigma}\right).$$

The information on μ then becomes

$$\frac{1}{\sigma^2}\Phi\left(\frac{c-\mu}{\sigma}\right)+\frac{1}{\sigma^2}\phi\left(\frac{c-\mu}{\sigma}\right)\left(\frac{\phi((c-\mu)/\sigma)}{1-\Phi(c-\mu)/\sigma)}-\frac{c-\mu}{\sigma}\right).$$

The mean of the underlying random variable is μ.

The mean of the coded data is c.

The mean of the non-coded data is

$$\mu-\sigma\frac{\phi((c-\mu)/\sigma)}{\Phi((c-\mu)/\sigma)}.$$

The marginal mean is

$$\left(\mu-\sigma\frac{\phi((c-\mu)/\sigma)}{\Phi((c-\mu)/\sigma)}\right)\,\Phi\left(\frac{c-\mu}{\sigma}\right)+c\left(1-\Phi\left(\frac{c-\mu}{\sigma}\right)\right).$$

Probit

Again take $x_i^* \sim N(\mu, \sigma^2)$.

Now only observe

$$x_i = \begin{cases} 1 & \text{if } x_i^* \geq 0 \\ 0 & \text{if } x_i^* < 0 \end{cases},$$

which is Bernoulli.

The probability of success and failure are

$$\Phi\left(\mu/\sigma\right), \qquad 1 - \Phi(\mu/\sigma),$$

respectively.

These probabilities depend on μ, σ only through the ratio $\theta = \mu/\sigma$, implying a scale indeterminacy; we can only learn θ.

The mass function becomes

$$f_\theta(x) = \Phi(\theta)^x \times \left(1 - \Phi(\theta)\right)^{1-x}$$

(Could further just focus on success probability $p = \Phi(\theta)$ but this would not extend to the model with covariates.)

Then

$$\frac{\partial \log f_\theta(x)}{\partial \theta} = (x - \Phi(\theta)) \frac{\phi(\theta)}{\Phi(\theta)(1 - \Phi(\theta))},$$

which has mean zero and variance

$$\frac{\phi(\theta)^2}{\Phi(\theta)(1 - \Phi(\theta))},$$

so the efficiency bound for θ becomes

$$\frac{1}{n} \frac{\Phi(\theta)(1 - \Phi(\theta))}{\phi(\theta)^2}.$$

A sensible way to estimate θ would be to first estimate the success probability $p = \Phi(\theta)$ by the sample mean $\overline{x}_n$ and then construct

$$\theta_n = \Phi^{-1}(\overline{x}_n).$$

This estimator is not unbiased (an unbiased estimator of θ does not exist here) but it will hit the efficiency bound in large samples.

Regularity conditions for Cramér-Rao bound

Derivation of best unbiased estimator above required regularity conditions:

- Differentiability of the density/mass function,
- Conditions for interchanging order of differentiation and integration.

An example where this fails is

$$x_i \sim \text{(continuous) uniform}[0, \theta],$$

that is,

$$f_\theta(x) = \frac{\{0 \leq x \leq \theta\}}{\theta}.$$

Nonetheless, a best unbiased estimator exists.

This follows from the Lehmann-Sheffé theorem, which builds on complete sufficient statistics.

▸ Skip the remainder of this section

Sufficiency

A statistic $\gamma_n = \gamma(x_1, \ldots, x_n)$ is sufficient for θ if

$$f_\theta(x_1, \ldots, x_n | \gamma_n) = f(x_1, \ldots, x_n | \gamma_n),$$

i.e., the conditional distribution does not depend on θ.

An obvious example is the Bernoulli distribution, where $P_\theta(x_i = 1) = \theta$.

Here,

$$f_\theta(x_1, \ldots, x_n) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i},$$

so a sufficient statistic will be $\sum_{i=1}^n x_i$, the number of successes in the sample.

Indeed, $\sum_{i=1}^n x_i$ is binomial with

$$f_\theta\left(\sum_{i=1}^n x_i\right) = \binom{n}{\sum_{i=1}^n x_i} \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i},$$

and so

$$f_\theta(x_1, \ldots, x_n | \sum_{i=1}^n x_i) = \binom{n}{\sum_{i=1}^n x_i} = \frac{n!}{(\sum_{i=1}^n x_i)!(n - \sum_{i=1}^n x_i)!}$$

is free of θ.

Sufficiency of the sample mean for a normal population

As another illustration, take $x_i \sim N(\theta, \sigma^2)$ with σ^2 known.

Then the sample mean $\overline{x}_n$ is sufficient for the unknown population mean θ.

We have

$$
\begin{aligned}
f_\theta(x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\sigma} \phi\left(\frac{x_i - \theta}{\sigma}\right) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{\left(-\frac{1}{2} \frac{\sum_{i=1}^{n}(x_i - \theta)^2}{\sigma^2}\right)} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{\left(-\frac{1}{2} \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 + n(\overline{x}_n - \theta)^2}{\sigma^2}\right)} \\
&= \frac{1}{(2\pi\sigma^2)^{(n-1)/2}} e^{\left(-\frac{1}{2} \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}{\sigma^2}\right)} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{\left(-\frac{1}{2} \frac{n(\overline{x}_n - \theta)^2}{\sigma^2}\right)}
\end{aligned}
$$

and

$$
f_\theta(\overline{x}_n) = \frac{1}{\sigma/\sqrt{n}} \phi\left(\frac{\overline{x}_n - \theta}{\sigma/\sqrt{n}}\right) = \frac{n^{1/2}}{(2\pi\sigma^2)^{1/2}} e^{\left(-\frac{1}{2} \frac{n\,(\overline{x}_n - \theta)^2}{\sigma^2}\right)}.
$$

It follows that

$$f_\theta(x_1, \ldots, x_n | \overline{x}_n) = \frac{f_\theta(x_1, \ldots, x_n)}{f_\theta(\overline{x}_n)} = \frac{n^{-1/2}}{(2\pi\sigma^2)^{(n-1)/2}} e^{\left(-\frac{1}{2}\frac{\sum_{i=1}^n (x_i - \overline{x}_n)^2}{\sigma^2}\right)},$$

which does not depend on θ.

When σ^2 is unknown a sufficient statistic for both μ and σ^2 is the pair

$$\overline{x}_n, \qquad \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x}_n)^2,$$

i.e., the sample mean and sample variance.

Improved estimation based on sufficiency

Theorem 6 (Rao-Blackwell theorem)

Let θ_* satisfy $E_\theta(\theta_*) = \theta$ and let γ_n be sufficient for θ. Define the estimator

$$\theta_n = E(\theta_* | \gamma_n)$$

(which is a function of the data through γ_n only). Then θ_n is unbiased and

$$\mathrm{var}_\theta(\theta_n) \leq \mathrm{var}_\theta(\theta_*)$$

holds.

Proof.

Unbiasedness of θ_n follows from iterating expectations on θ_*.

Next, by the law of total variance,

$$\mathrm{var}_\theta(\theta_*) = \mathrm{var}(E(\theta_*|\gamma_n)) + E_\theta(\mathrm{var}(\theta_*|\gamma_n)) = \mathrm{var}_\theta(\theta_n) + \text{non-negative term},$$

and so $\mathrm{var}_\theta(\theta_*) \geq \mathrm{var}_\theta(\theta_n)$.

Finally, θ_n is a statistic (and so computable from data) by sufficiency. □

Rao-Blackwellization for Bernoulli

A simple unbiased estimator of θ is $\theta_* = x_1$; its variance is $\theta(1-\theta)$.

Define

$$\theta_n = E\left(x_1 \left|\sum_{i=1}^n x_i\right.\right) = P_\theta(x_1 = 1 | \sum_{i=1}^n x_i).$$

Note that

$$\begin{aligned}
P_\theta(x_1 = 1 | \sum_{i=1}^n x_i = x) &= \frac{P_\theta(\sum_{i=1}^n x_i = x | x_1 = 1)\, P_\theta(x_1 = 1)}{P_\theta(\sum_{i=1}^n x_i = x)} \\
&= \frac{P_\theta(\sum_{j\neq i}^n x_j = (x-1) | x_1 = 1) P_\theta(x_1 = 1)}{P_\theta(\sum_{i=1}^n x_i = x)} \\
&= \frac{P_\theta(\sum_{j\neq i}^n x_j = (x-1))}{P_\theta(\sum_{i=1}^n x_i = x)}\, P_\theta(x_1 = 1) \\
&= \frac{\binom{n-1}{x-1}\theta^{x-1}(1-\theta)^{n-x}}{\binom{n}{x}\theta^s(1-\theta)^{n-x}}\,\theta = \frac{\frac{(n-1)!}{(n-x)!(x-1)!}}{\frac{n!}{(n-x)!x!}} \\
&= \frac{(n-1)!}{n!}\,\frac{x!}{(x-1)!} = \frac{x}{n}.
\end{aligned}$$

Thus, $\theta_n = n^{-1}\sum_{i=1}^n x_i = \overline{x}_n$, which has variance $\theta(1-\theta)/n$ (and is, in fact, best unbiased).

Completeness

A statistic γ_n is complete (for f_θ) if it holds that

$$\text{if } E_\theta(\varphi(\gamma_n)) = 0 \text{ for all } \theta, \text{ then } P_\theta(\varphi(\gamma_n) = 0) = 1 \text{ for all } \theta,$$

for all φ for which the expectation exists.

To clarify take $x_i \sim N(\theta, \sigma^2)$. Consider the statistic $x_2 - x_1$. We have

$$E_\theta(x_2 - x_1) = \theta - \theta = 0, \qquad \text{for all } \theta.$$

However, $x_2 - x_1 \sim N(0, 2\sigma^2)$, and so

$$P_\theta(x_2 - x_1 = 0) = 0 \qquad \text{for all } \theta.$$

So, this statistic is not complete.

Completeness in the normal problem

A complete statistic here is $\overline{x}_n = n^{-1} \sum_{i=1}^{n} x_i$.

We look for a function φ such that $E_\theta(\varphi(\overline{x}_n)) = 0$ for all θ.

We have

$$\begin{aligned} E_\theta(\varphi(\overline{x}_n)) &= \textstyle\int_{-\infty}^{+\infty} \varphi(x) \, \frac{1}{\sigma/\sqrt{n}} \phi\left(\frac{\overline{x}_n - \theta}{\sigma/\sqrt{n}}\right) \, dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{-\frac{n}{2}(\theta/\sigma)^2} \textstyle\int_{-\infty}^{+\infty} \varphi(x) \, e^{-\frac{n}{2}(x/\sigma)^2} \, e^{n(\theta/\sigma^2)x} \, dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{-\frac{n}{2}(\theta/\sigma)^2} \mathcal{L}\left(\varphi(x) \, e^{-\frac{n}{2}(x/\sigma)^2}\right), \end{aligned}$$

for $\mathcal{L}(g(x))$ the (two-sided) Laplace transform of $g(x)$.

The Laplace transform $\mathcal{L}(g(x))$ cannot be zero unless $g(x)$ is zero (almost everywhere). As the exponential function is non-zero it must be that $\varphi(x) = 0$ (almost everywhere), as claimed.

Completeness in the Bernoulli problem

Remember that, if $P_\theta(x_i = 1) = \theta$ for $\theta \in (0,1)$, then $\gamma_n = \sum_{i=1}^n x_i$ is Binomial with parameters (n, θ).

So, if

$$E_\theta(\varphi(\gamma_n)) = \sum_{\gamma=0}^n \gamma \binom{n}{\gamma} \theta^\gamma (1-\theta)^{n-\gamma} = (1-\theta)^n \sum_{\gamma=0}^n \varphi(\gamma) \binom{n}{\gamma} \left(\frac{\theta}{1-\theta}\right)^\gamma = 0$$

for all $\theta \in (0,1)$ then the following polynomial in $\lambda = \theta/(1-\theta)$

$$\sum_{\gamma=0}^n c_\gamma \lambda^\gamma = 0, \qquad c_\gamma = \varphi(\gamma) \frac{n!}{\gamma(n-\gamma)!},$$

must be zero for all $\theta \in (0,1)$.

But the latter can only hold if $c_\gamma = 0$ for all γ, and so $\varphi(\gamma) = 0$ must hold for all $\gamma \in \{0, 1, \ldots, n\}$.

Hence,

$$P_\theta(\varphi(\gamma_n) = 0) = 1$$

follows.

Best unbiased estimation under sufficiency

Theorem 7 (Lehmann-Scheffé theorem)

Let γ_n *be a complete sufficient statistic for* θ *and consider* $\theta_n = \varphi(\gamma_n)$ *for some function* φ*. If* $E_\theta(\theta_n) = \theta$ *then*

$$\mathrm{var}_\theta(\theta_n) \leq \mathrm{var}(\theta_*)$$

where θ_* *is any unbiased estimator; i.e.,* θ_n *is the best unbiased estimator.*

Proof.

By the Rao-Blackwell result, under sufficiency, any efficient estimator must be a function of γ_n only; so, $\theta_n = \varphi(\gamma_n)$. Then, by assumption, $E_\theta(\theta_n) = \theta$.

It is enough to show that φ is unique. Suppose there exist another ψ such that $E_\theta(\psi(\gamma_n)) = \theta$. Then, by unbiasedness of both estimators,

$$E_\theta(\varphi(\gamma_n) - \psi(\gamma_n)) = 0.$$

But, by completeness, this implies that $P_\theta(\varphi(\gamma_n) = \psi(\gamma_n)) = 1$ (a.e.). □

Bernoulli

We have shown above that

$$\gamma_n = \sum_{i=1}^{n} x_i$$

is both a complete and sufficient statistic for θ.

An unbiased estimator based on it is the sample mean

$$\overline{x}_n = n^{-1} \sum_{i=1}^{n} x_i = \gamma_n / n.$$

This confirms the Cramér-Rao result for Bernoulli that $\overline{x}_n$ is best unbiased.

Estimating the maximum of a uniform distribution

Recall

$$f_\theta(x) = \frac{\{0 \le x \le \theta\}}{\theta}.$$

Easy to see that the maximum-likelihood estimator here is $\max_i(x_i)$.

This estimator is biased.

For all $x \in [0, \theta]$,

$$P_\theta\left(\max_i(x_i) \le x\right) = P_\theta(x_1 \le x, x_2 \le x, \ldots, x_n \le x) = (x/\theta)^n.$$

Further,

$$E_\theta(\max_i(x_i)) = \int_0^\theta 1 - (x/\theta)^n \, dx = \tfrac{n}{n+1}\,\theta.$$

(The first step holds for any non-negative random variable $z \in [0, b]$, say; integrate by parts to see that

$$\int_0^b (1 - F(z)) \, dz = (1 - F(z))\, z|_0^b + \int_0^b z\, f(z)\, dz = E(z),$$

as claimed.)

It follows that

$$\theta_n = \frac{n+1}{n} \max_i(x_i).$$

is unbiased.

Remains to show that $\gamma_n = \max_i(x_i)$ is a complete sufficient statistic for θ.

We already know that $P_\theta(\gamma_n \leq \gamma) = (\gamma/\theta)^n$ and so its density is

$$n \, \frac{\gamma^{n-1}}{\theta^n}$$

for $\gamma \in [0, \theta]$ (and zero elsewhere).

Hence,

$$E_\theta(\varphi(\gamma_n)) = \int_0^\theta \varphi(\gamma) \, n \, \tfrac{\gamma^{n-1}}{\theta^n} \, d\gamma = (n/\theta^n) \, \left(\int_0^\theta \varphi(\gamma) \gamma^{n-1} \, d\gamma \right) =: (n/\theta^n) \, Q(\theta).$$

Note that, by Leibniz' rule,

$$\frac{\partial Q(\theta)}{\partial \theta} = \varphi(\theta) \, \theta^{n-1}.$$

So, if $E_\theta(\varphi(\gamma_n)) = 0$ for all θ then $Q(\theta) = 0$ must hold for all θ, but then its derivative must be zero and so $\varphi(\theta) = 0$ must hold. So, γ_n is indeed complete.

To see sufficiency we look at the ratio of the density of the data,

$$\prod_{i=1}^{n} \frac{\{x_i \leq \theta\}}{\theta} = \frac{\{\gamma_n \leq \theta\}}{\theta^n},$$

and the density of the sample maximum,

$$n\gamma_n^{n-1} \frac{\{\gamma_n \leq \theta\}}{\theta^n}$$

(from above).

As this ratio is

$$\gamma_n^{1-n}/n$$

it is free from θ and so γ_n is indeed sufficient.

Working out the first two moments of γ_n using its density from above gives

$$\frac{1}{n(n+2)}\theta^2$$

as the variance of the unbiased estimator θ_n.

Note that this variance shrinks like n^{-2}, which is faster than the parametric rate of n^{-1}.

Efficiency bound for biased estimators

In many cases an (best) unbiased estimator will not exist. So we need to widen our search to allow for bias.

First generalize Cramér-Rao bound to case where θ_n is biased, i.e.,

$$E_\theta(\theta_n) = \theta + b_n(\theta).$$

Following the same steps as before gives the efficiency bound

$$\mathrm{var}_\theta(\theta_n) \geq I_\theta^{-1}(1 + b_n'(\theta))^2/n.$$

Quite generally, $b_n'(\theta) = O(n^{-1})$, and so

$$\mathrm{var}_\theta(\theta_n) \geq I_\theta^{-1}/n + O(n^{-2});$$

the bias vanishes faster than the standard deviation.

From an asymptotic perspective, this paves the way for best asymptotically unbiased estimators.

Asymptotics (for the univariate sample mean)

Asymptotic analysis is an approximation to the finite-sample behavior of an estimator based on what happens when n becomes large.

While exact small-sample results are few and ad hoc. Large-sample analysis is well established and widely applicable.

The behavior of the sample mean as $n \to \infty$ brings us a long way.

This is so because almost all estimators you will ever look at behave, as $n \to \infty$, like a sample mean.

Such estimators are called asymptotically linear; we can always represent them as

$$(\theta_n - \theta) = n^{-1} \sum_i \varphi_\theta(x_i) + o_p(n^{-1/2})$$

for some function φ_θ for which $E_\theta(\varphi_\theta(x_i)) = 0$ and $\mathrm{var}_\theta(\varphi_\theta(x_i)) < \infty$. We will see many examples.

Slide 25 gives the influence function for the best unbiased estimator (when it exists).

Orders of magnitude (deterministic sequences)

Let h and g be two functions (and $h(x) > 0$ for large x).

We say that $g(x) = O(h(x))$ if and only if there exists a positive number b and a real number $\underline{x}$ such that

$$|g(x)| \leq b\, h(x) \text{ for all } x \geq \underline{x}.$$

That is,

$$\limsup_{x\to\infty} \left| \frac{g(x)}{h(x)} \right| < \infty;$$

$h(x)$ grows at least as fast as $g(x)$.

We say that $g(x) = o(h(x))$ if and only if for every positive number b there exists a real number $\underline{x}$ such that

$$|g(x)| \leq b\, h(x) \text{ for all } x \geq \underline{x}.$$

That is,

$$\lim_{x\to\infty} \left| \frac{g(x)}{h(x)} \right| = 0;$$

$h(x)$ grows faster than $g(x)$.

Orders of magnitude (random sequences)

Let $\{x_n\}$ be a sequence of random variables and let $\{a_n\}$ be a deterministic sequence of numbers.

Consider the limit behavior as $n \to \infty$.

We say that $x_n = O_p(a_n)$ if and only if for every δ there exists a finite number ϵ and an $\underline{n}$ such that

$$P\left(\left|\frac{x_n}{a_n}\right| > \epsilon\right) < \delta \text{ for all } n \geq \underline{n}.$$

That is, $|x_n/a_n|$ is stochastically bounded.

We say that $x_n = o_p(a_n)$ if and only if for every δ and finite number ϵ there exists an $\underline{n}$ such that

$$P\left(\left|\frac{x_n}{a_n}\right| > \epsilon\right) < \delta \text{ for all } n \geq \underline{n}.$$

That is,

$$\lim_{n\to\infty} P\left(\left|\frac{x_n}{a_n}\right| > \epsilon\right) = 0$$

for every $\epsilon > 0$.

Convergence in probability

We say that x_n converges in probability to x if,for every $\epsilon > 0$,

$$\lim_{n \to \infty} P\left(|x_n - x| > \epsilon\right) = 0,$$

i.e., if $x_n - x = o_p(1)$.

We write $x_n \xrightarrow{p} x$ and call x the probability limit of the sequence $\{x_n\}$.

Theorem 8 ((weak) law of large numbers)

Suppose that $\mu = E(x_i)$ *exists. For any* $\epsilon > 0$ *and* $\delta > 0$, *there exists an* $\underline{n}$ *such that*

$$P(|\overline{x}_n - \mu| > \epsilon) < \delta, \qquad \textit{for all} \qquad n > \underline{n}.$$

That is, $\overline{x}_n \xrightarrow{p} \mu$ *as* $n \to \infty$. *Equivalently,* $\overline{x}_n - \mu = o_p(1)$.

Proof.

Suppose that σ^2 exists. Then (by Chebychev's inequality)

$$P(|\overline{x}_n - \mu| > \epsilon) = P((\overline{x}_n - \mu)^2 > \epsilon^2) \leq \frac{E((\overline{x}_n - \mu)^2)}{\epsilon^2} = n^{-1}\frac{\sigma^2}{\epsilon^2}.$$

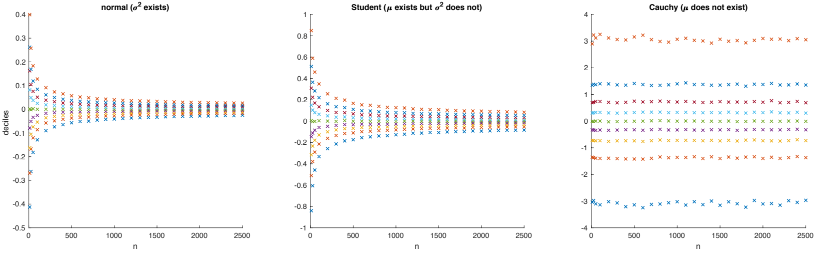
Taking limits gives the result. □

Note that we immediately get the same result for any transformation $\varphi(x_i)$ provided that $E(|\varphi(x_i)|) < \infty$. That is,

$$n^{-1}\textstyle\sum_i \varphi(x_i) \xrightarrow{p} E(\varphi(x_i))$$

as $n \to \infty$.

The below plots give deciles of $\overline{x}_n$ as a function of n.



Consistency

An estimator θ_n is consistent for an estimand θ if $\theta_n \xrightarrow{p} \theta$.

The mean squared error of is

$$E_\theta((\theta_n - \theta)^2) = (E_\theta(\theta_n - \theta))^2 + \mathrm{var}_\theta(\theta_n) = b_n(\theta)^2 + \mathrm{var}_\theta(\theta_n);$$

so a sufficient condition for consistency is that both bias and variance vanish as $n \to \infty$.

Uniform convergence

A more general situation has $\varphi_\theta(x_i)$ indexed by $\theta \in \Theta$ (continuous on Θ compact).

A pointwise convergence result (i.e., for any fixed $\theta \in \Theta$) follows from above:

$$P\left(\left|n^{-1}\textstyle\sum_i \varphi_\theta(x_i) - E(\varphi_\theta(x_i))\right| > \epsilon\right) < \delta, \qquad \text{for all} \qquad n > \underline{n}_\theta,$$

A uniform result is as follows.

Theorem 9 (Uniform law of large numbers)

Suppose that $\varphi_\theta(x) \leq \gamma(x)$ and $E(|\gamma(x_i)|) < \infty$. Then, for all $\theta \in \Theta$,

$$P\left(\left|n^{-1}\sum_i \varphi_\theta(x_i) - E(\varphi_\theta(x_i))\right| > \epsilon\right) < \delta, \qquad \textit{for all} \qquad n > \underline{n},$$

with $\underline{n}$ independent of θ.

We write

$$\sup_{\theta\in\Theta}\left|n^{-1}\sum_i \varphi_\theta(x_i) - E(\varphi_\theta(x_i))\right| \xrightarrow{p} 0$$

as $n \to \infty$.

To appreciate the difference between pointwise and uniform convergence take a simple non-stochastic example:

$$\varphi_\theta(x_i) = n\theta e^{-n\theta}$$

for $\theta \in \Theta = [0, 1]$. This function is continuous in θ.

For any fixed θ,

$$n\theta e^{-n\theta} \to 0$$

as $n \to \infty$. (because the exponential term vanishes more quickly than the linear term grows.)

However, at $\theta = n^{-1}$ the function equals e^{-1} for any n. Hence,

$$\sup_{\theta \in \Theta} |n\theta e^{-n\theta}| \nrightarrow 0$$

as $n \to \infty$.

Note that (in general), uniform convergence implies pointwise convergence.

Continuous-mapping theorem

Theorem 10 (Continuous-mapping theorem)

Suppose that $x_n \xrightarrow{p} x$.

Then

$$\varphi(x_n) \xrightarrow{p} \varphi(x)$$

for (non-stochastic) continuous functions φ.

Convergence in distribution

Let $\{x_n\}$ be a sequence of random variables with distribution $\{F_n\}$ and let $x \sim F$

We say that $x_n \overset{d}{\to} x$ if

$$F_n(a) \to F(a) \text{ as } n \to \infty$$

at all continuity points a of F.

We call F the limit distribution of $\{x_n\}$.

If $x_n \overset{d}{\to} x$ it is stochastically bounded, i.e., $x_n = O_p(1)$.

The central limit theorem

Theorem 11 (Lindeberg-Lévy central limit theorem)

Suppose that $x_i \sim \text{i.i.d.}\,(\mu, \sigma^2)$. *Then*

$$\frac{\overline{x}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

as $n \to \infty$.

This means that the sample distribution of the standardized sample mean approaches the standard-normal distribution.

In practice, this means that

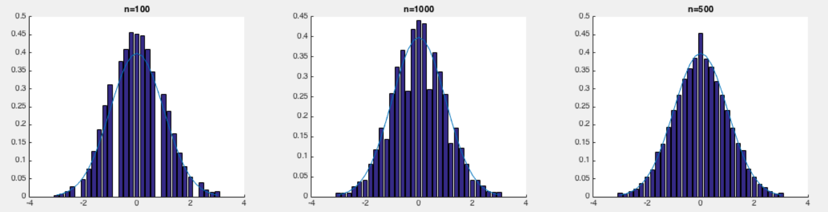
$$\overline{x}_n \overset{a}{\sim} N(\mu, \sigma^2/n),$$

where the a can be interpreted as either ‘asymptotically’ or ‘approximately’.

Observe that this result holds for any distribution, as long as μ, σ^2 exist.

The plots below concern the standardized sample mean of samples of Bernoulli random variables.

Observe how the histogram approaches the standard-normal density as n grows.



Proof.

Let $\varphi_x(t) = E(e^{\iota t x})$ be the characteristic function of x.

Then

$$z = \frac{\overline{x}_n - \mu}{\sigma/\sqrt{n}} = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \frac{x_i - \mu}{\sigma} = \sum_{i=1}^{n} \frac{z_i}{\sqrt{n}} \text{ (say)},$$

has characteristic function

$$\varphi_z(t) = E(e^{\iota t \sum_i z_i/\sqrt{n}}) = \prod_i E(e^{\iota (t/\sqrt{n}) z_i}) = \varphi_{z_i}(t/\sqrt{n})^n,$$

where we used random sampling.

Now, as $\varphi_{z_i}(0) = 1$, $\varphi'_{z_i}(0) = 0$, and $\varphi''_{z_i}(0) = -1$ we have

$$\varphi_{z_i}(t/\sqrt{n}) = \varphi_{z_i}(0) + \varphi'_{z_i}(0)\frac{t}{\sqrt{n}} + \varphi''_{z_i}(0)\frac{t^2}{2n} + o\left(\frac{t^2}{n}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)$$

as $n \to \infty$, and so

$$\lim_{n\to\infty} \varphi_z(t) = \lim_{n\to\infty}\left(1 + \frac{-t^2/2}{n}\right)^n = e^{-t^2/2} (= \varphi \text{ of the standard normal})$$

by definition of the exponential function. □

Slutzky's theorem

Theorem 12 (Slutzky's theorem)

Suppose that $x_n \xrightarrow{p} c$ *(a constant) and* $y_n \xrightarrow{d} y$ *(a random variable). Then*

(i) $x_n + y_n \xrightarrow{d} c + y$*; and*

(ii) $x_n\, y_n \xrightarrow{d} c\, y$.

Take $x_i \sim N(\mu, \sigma^2)$. Best 'estimator' of σ^2 is $n^{-1}\sum_i (x_i - \mu)^2$.

As an example of (i),

$$\hat{\sigma}^2 = n^{-1}\sum_i (x_i - \overline{x}_n)^2 = n^{-1}\sum_i (x_i - \mu)^2 - (\overline{x}_n - \mu)^2.$$

As $(\overline{x}_n - \mu) \xrightarrow{p} 0$ and $(a-\mu)^2$ is continuous in a we have $(\overline{x}_n - \mu)^2 \xrightarrow{p} 0$. Hence,

$$\hat{\sigma}^2 = n^{-1}\sum_i (x_i - \mu)^2 + o_p(1).$$

In fact,

$$(\overline{x}_n - \mu)^2 = (O_p(1/\sqrt{n}))^2 = O_p(1/n) = o_p(1/\sqrt{n}),$$

and so

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) = \frac{1}{\sqrt{n}}\sum_i ((x_i - \mu)^2 - \sigma^2) + o_p(1).$$

Hence, $\hat{\sigma}^2$ and $n^{-1}\sum_i (x_i - \mu)^2$ are asymptotically equivalent; their limit distribution is $\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4)$. This is the same limit distribution as that of (the unbiased) $\tilde{\sigma}^2$.

As an example of (ii),

$$\frac{\overline{x}_n - \mu}{\hat{\sigma}/\sqrt{n}} = \frac{\sigma}{\hat{\sigma}} \frac{\overline{x}_n - \mu}{\sigma/\sqrt{n}} = (1 + o_p(1)) \frac{\overline{x}_n - \mu}{\sigma/\sqrt{n}} = \frac{\overline{x}_n - \mu}{\sigma/\sqrt{n}} + o_p(1) \xrightarrow{d} N(0,1).$$

Now given an asymptotically-linear estimator,

$$(\theta_n - \theta) = n^{-1} \sum_i \varphi_\theta(x_i) + o_p(n^{-1/2})$$

where $E_\theta(\varphi_\theta(x_i)) = 0$ and $\mathrm{var}_\theta(\varphi_\theta(x_i)) < \infty$ our results immediately yield that

(a) $\theta_n - \theta = O_p(n^{-1/2})$; and

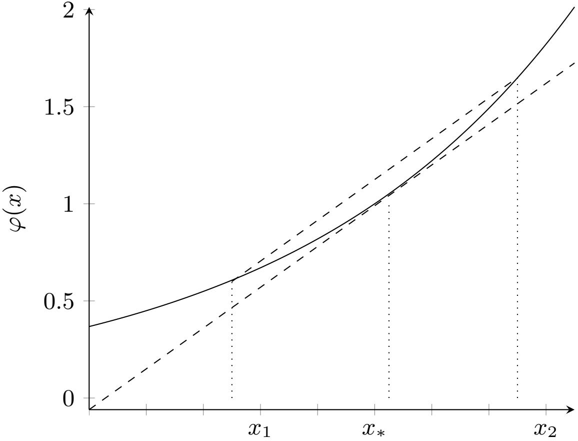
(b) $\sqrt{n}(\theta_n - \theta) \overset{a}{\sim} N(0, \mathrm{var}_\theta(\varphi_\theta(x_i)))$.

We call $\mathrm{var}_\theta(\varphi_\theta(x_i))$ the asymptotic variance.

Mean-value theorem

Let φ be a differentiable function on an interval $[\underline{x}, \overline{x}_n]$. Then, for any $(x_1, x_2) \in [\underline{x}, \overline{x}_n]^2$ there always exists a $x_* \in [\underline{x}, \overline{x}_n]$ (not necessarily unique) so that

$$\varphi(x_2) - \varphi(x_1) = \frac{\partial \varphi(x_*)}{\partial x}(x_2 - x_1).$$



Asymptotics for smooth transformations

Theorem 13 (Delta method)

If $\sqrt{n}(\theta_n - \theta) \xrightarrow{d} N(0, \sigma^2)$, *then*

$$\sqrt{n}(\varphi(\theta_n) - \varphi(\theta)) \xrightarrow{d} N\left(0, \left(\partial\varphi(\theta)/\partial\theta\right)^2 \sigma^2\right)$$

for continuously-differentiable φ.

Proof.

A mean-value expansion gives

$$\varphi(\theta_n) - \varphi(\theta) = \frac{\partial\varphi(\theta_*)}{\partial\theta}\,(\theta_n - \theta).$$

The continuous-mapping theorem yields

$$\frac{\partial\varphi(\theta_*)}{\partial\theta} \xrightarrow{p} \frac{\partial\varphi(\theta)}{\partial\theta}.$$

Slutzky's theorem gives

$$\sqrt{n}(\varphi(\theta_n) - \varphi(\theta)) = \frac{\partial\varphi(\theta)}{\partial\theta}\,\sqrt{n}\,(\theta_n - \theta) + o_p(1) \xrightarrow{d} N\left(0, \left(\partial\varphi(\theta)/\partial\theta\right)^2 \sigma^2\right).$$

The multivariate case

Now suppose that x_i is a vector with mean μ and variance Σ.

The multivariate central limit theorem reads

$$\sqrt{n}\,\Sigma^{-1/2}(\overline{x}_n - \mu) \overset{d}{\to} N(0, I),$$

where I is the identity matrix of conformable dimension. Here, the limit distribution is a multivariate standard normal.

The Delta method extends as follows. Suppose $\sqrt{n}(\theta_n - \theta) \overset{d}{\to} N(0, \Sigma)$. Then

$$\sqrt{n}(\varphi(\theta_n) - \varphi(\theta)) \overset{d}{\to} N(0, \Gamma\Sigma\Gamma')$$

for

$$\Gamma = \frac{\partial \varphi(\theta)}{\partial \theta'}$$

the Jacobian matrix.

A nonsingular matrix A has eigendecomposition

$$A = VDV^{-1}$$

where D is a diagonal matrix of eigenvalues and V is the matrix of associated eigenvectors.

The inverse is

$$A^{-1} = VD^{-1}V^{-1}.$$

A matrix square root is

$$A^{1/2} = VD^{1/2}V^{-1}.$$

Note that

$$A^{-1/2}AA^{-1/2} = (VD^{-1/2}V^{-1})(VDV^{-1})(VD^{-1/2}V^{-1}) = I.$$

So, for example, if $\sqrt{n}(\theta_n - \theta) \xrightarrow{d} N(0, \Sigma)$ for an $m \times m$ nonsingular variance Σ, then

(i) $\sqrt{n}\,\Sigma^{-1/2}(\theta_n - \theta) \xrightarrow{d} N(0, I_m)$; and

(ii) $n(\theta_n - \theta)'\Sigma^{-1}(\theta_n - \theta) \xrightarrow{d} \chi^2_m$.

The multivariate normal distribution

If $x \sim N(\mu, \Sigma)$ its density is

$$\frac{1}{\sqrt{(2\pi)^{\dim x} \det(\Sigma)}} e^{-\frac{(x-\mu)'\Sigma^{-1}(x-\mu)}{2}}.$$

Any subset of x is again normal. All conditional distributions are again normal.

Partition $x = (x_1', x_2')'$ and write

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

The marginal distribution of x_1 is normal, $x_1 \sim N(\mu_1, \Sigma_{11})$.

The conditional distribution of x_1 given x_2 is

$$N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

The bivariate normal distribution

The above is particularly tractable in the bivariate case, where x_1 and x_2 are both scalars.

Write

$$\left(\begin{array}{c} x_1 \\ x_2 \end{array} \right) \sim N\left(\left(\begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right) \quad \left(\begin{array}{cc} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{array} \right) \right)$$

for ρ the correlation between x_1 and x_2.

Here,

$$x_1|x_2 \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}\,(x_2 - \mu_2), (1-\rho^2)\,\sigma_1^2\right).$$

Note that

$$E(x_1|x_2) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}\,(x_2 - \mu_2) = \left(\mu_1 - \rho\frac{\sigma_1}{\sigma_2}\mu_2\right) + \rho\frac{\sigma_1}{\sigma_2}\,x_2$$

is linear in x_2.

Also, $\mathrm{var}(x_1|x_2)$ is a constant (i.e., not a function of x_2).

Best asymptotically unbiased estimation

We say that θ_n is best asymptotically unbiased for θ if

$$\sqrt{n}(\theta_n - \theta) \xrightarrow{d} N(0, I_\theta^{-1}),$$

so it achieves the Cramér-Rao bound in large samples.

It exists under weak regularity conditions.

A coherent way of finding it is through the method of maximum likelihood.

The likelihood function

The maximum-likelihood estimator is

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} f_\theta(x_i) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log f_\theta(x_i).$$

The likelihood function, $\prod_{i=1}^{n} f_\theta(x_i)$, represents the density of the sample when sampling from f_θ.

In the discrete case, it is the probability of observing the actual sample, when sampling from f_θ.

Maximize this probability as a function of θ.

Intuitively attractive. Pretty much what anyone without any prior statistical knowledge would do.

Maximization program

Let

$$L_n(\theta) = \sum_i \log f_\theta(x_i)$$

be the log-likelihood function.

The first-order condition is that

$$\frac{\partial L_n(\theta)}{\partial \theta} = \sum_i \frac{\partial \log f_\theta(x_i)}{\partial \theta} = 0;$$

this is the score equation.

The second-order condition for a maximum is that

$$\frac{\partial^2 L_n(\theta)}{\partial \theta \partial \theta'} = \sum_i \frac{\partial^2 \log f_\theta(x_i)}{\partial \theta \partial \theta'} < 0;$$

the Hessian matrix is negative definite.

Note how these derivatives relate to the Cramér-Rao bound.

Numerical maximization: Newton-Raphson

Often we need to tackle the maximization problem numerically.

Newton-Raphson is a popular algorithm for finding the roots of the score equation.

Want to solve $\varphi(x) = 0$. Let x_0 be an initial guess. For a new guess x_1 we have

$$\frac{\varphi(x_1) - \varphi(x_0)}{x_1 - x_0} \approx \left.\frac{\partial \varphi(x)}{\partial x}\right|_{x=x_0} = \varphi'(x_0).$$

So,

$$\varphi(x_0) + (x_1 - x_0)\,\varphi'(x_0) \approx \varphi(x_1).$$

We want that $\varphi(x_1) = 0$. Solving for x_1 yields

$$x_1 = x_0 - \varphi(x_0)/\varphi'(x_0)$$

as our new guess.

In practice, when maximizing a function whose derivative is φ, we start at x_0 and then evaluate φ in x_1. If the function would not improve at x_1 we re-evaluate in $x_1' = x_0 - h(x_0 - x_1)$ for $h \in (0, 1)$ a step size and re-evaluate. We then iterate this procedure untill no further improvement (up to some specified tolerance level) is found.

Bernoulli

When $x_i \in \{0,1\}$ with probability $\theta \in (0,1)$ we have

$$L_n(\theta) = \log\left(\prod_{i=1}^{n} \theta^{x_i}\,(1-\theta)^{1-x_i}\right) = \sum_{i=1}^{n} x_i \log\theta + (1-x_i)\log(1-\theta).$$

So, solving

$$\frac{\partial L_n(\theta)}{\partial\theta} = \sum_{i=1}^{n} \frac{x_i-\theta}{\theta(1-\theta)} = n\frac{\overline{x}_n-\theta}{\theta(1-\theta)} = 0$$

for θ yields $\hat{\theta} = \overline{x}_n$ as the unique solution. This is a global maximum as

$$\frac{\partial^2 L_n(\theta)}{\partial\theta^2} = -\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{\theta^2(1-\theta)^2} < 0$$

for all $\theta \in (0,1)$.

This estimator is best unbiased and so also best asymptotically unbiased.

Invariance

Maximum likelihood is invariant to one-to-one parametrizations, $\beta = \beta(\theta)$.

If $L_n(\theta)$ is the log-likelihood and $L_n^*(\beta)$ is the reparametrized log-likelihood, then

$$\hat{\beta} = \arg\max_{\beta} L_n^*(\beta) = \beta(\arg\max_{\theta} L_n(\theta)) = \beta(\hat{\theta}).$$

This is an interesting property.

A consequence of invariance is that maximum likelihood will not be unbiased, in general.

If $\hat{\theta}$ is unbiased and the transformation $\beta(\theta)$ is nonlinear, then $\hat{\beta} = \beta(\hat{\theta})$ will be biased, in general, by Jensen's inequality.

Jensen's inequality

A univariate function φ is concave if

$$\varphi(\lambda x + (1-\lambda)x') \geq \lambda\,\varphi(x) + (1-\lambda)\varphi(x')$$

and convex if

$$\varphi(\lambda x + (1-\lambda)x') \leq \lambda\,\varphi(x) + (1-\lambda)\varphi(x')$$

for all $\lambda \in [0,1]$.

Theorem 14 (Jensen's inequality)

If φ is concave, $E(\varphi(x_i)) \leq \varphi(E(x_i))$. If φ is convex, $E(\varphi(x_i)) \geq \varphi(E(x_i))$.

Proof.

Take φ concave. Let ψ be the tangent line at $E(x_i)$; i.e., $\psi(x) = a + bx$ for constants a, b such that $\varphi(E(x_i)) = \psi(E(x_i))$.

By concavity $\varphi(x) \leq \psi(x)$ for any x. Hence, using linearity of the tangent,

$$E(\varphi(x_i)) \leq E(\psi(x_i)) = \psi(E(x_i)) = \varphi(E(x_i)).$$

□

Probit

The simplest probit model from above had

$$P_\theta(x_i = 1) = \Phi(\theta).$$

The score equation is

$$\sum_{i=1}^{n} (x_i - \Phi(\theta)) \frac{\phi(\theta)}{\Phi(\theta)(1 - \Phi(\theta))} = 0$$

and the efficiency bound was

$$\frac{1}{n} \frac{\Phi(\theta)(1 - \Phi(\theta))}{\phi(\theta)^2}.$$

Finding $\hat{\theta}$ by solving the score equation requires numerical optimization.

Notice that the success probability $\beta = \Phi(\theta)$ is a one-to-one transformation of θ. The likelihood for β is the ordinary Bernoulli likelihood, with maximizer $\hat{\beta} = \overline{x}_n$.

It follows that $\hat{\theta} = \Phi^{-1}(\overline{x}_n)$.

Further, as

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \beta(1 - \beta))$$

by the central limit theorem,

$$\frac{\partial \Phi^{-1}(\beta)}{\partial \beta} = \frac{1}{\phi(\theta)},$$

and $\beta = \Phi(\theta)$, the Delta method gives

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{\Phi(\theta)(1 - \Phi(\theta))}{\phi(\theta)^2}\right).$$

The asymptotic variance is indeed the Cramér-Rao bound.

The maximum-likelihood estimator may fail to exist in small samples.

In the probit model this happens when complete separation is possible.

In essence, this means we can classify outcomes exactly.

In our model, where

$$P_\theta(x_i = 1) = \beta = \Phi(\theta),$$

a data set consisting of only successes (ones) will have $\hat{\beta} = 1$, and so $\hat{\theta} = +\infty$.

If $\beta \in (0, 1)$ this problem will not occur in large samples.

In an extended model with explanatory variables perfect separation would happen, for example, when all successes can be assigned to one covariate and all failures to another.

This problem may, in principle, persist in large samples.

Why does maximizing the likelihood work: Identification

Note that the expected log-likelihood

$$L(\theta_*) = E_\theta(L_n(\theta_*)) = \sum_{i=1}^{n} E_\theta(\log f_{\theta_*}(x_i)) = n\, E_\theta(\log f_{\theta_*}(x_i))$$

is maximized at θ.

Indeed,

$$E_\theta(L_n(\theta_*) - L_n(\theta)) = n\, E_\theta\left(\log\left(\frac{f_{\theta_*}(x_i)}{f_\theta(x_i)}\right)\right) \leq n\, \log E_\theta\left(\frac{f_{\theta_*}(x_i)}{f_\theta(x_i)}\right) = 0,$$

using Jensen's inequality and the fact that $E_\theta(f_{\theta_*}(x_i)/f_\theta(x_i)) = \int f_{\theta_*}(x)\, dx = 1$.

Crudely put, $L(\theta)$ is the log-likelihood function we would use if we would have an infinitely-large sample.

(Point) identification means that, in that case, we would be able to learn θ; so

$$\theta = \arg\max_{\theta_* \in \Theta} L(\theta_*),$$

and is unique.

Identification may fail (we will give an example below).

Global identification: θ is the unique maximizer of L on Θ.

Local identification: θ is the unique maximizer of L in some neighborhood around θ.

Local identification is

$$\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'} < 0.$$

Note that, as

$$\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'} = n\, E_\theta\left(\frac{\partial^2 \log f_\theta(x_i)}{\partial\theta\partial\theta'}\right) = -nI_\theta$$

this is equivalent to the information matrix being positive definite and, hence, of full rank.

Local identification can be tested.

Why does maximizing the likelihood work: Argmax theorem

By a uniform law of large numbers,

$$\sup_{\theta\in\Theta}|n^{-1}L_n(\theta) - n^{-1}L(\theta)| \xrightarrow{p} 0, \text{ as } n \to \infty,$$

provided f_θ is continuous, $|\log f_\theta(x)| < b(x)$ so that $E(b(x_i)) < \infty$, and Θ is closed and bounded (compact).

Then, if θ is identified as the unique global maximizer of $L(\theta)$, we will have that

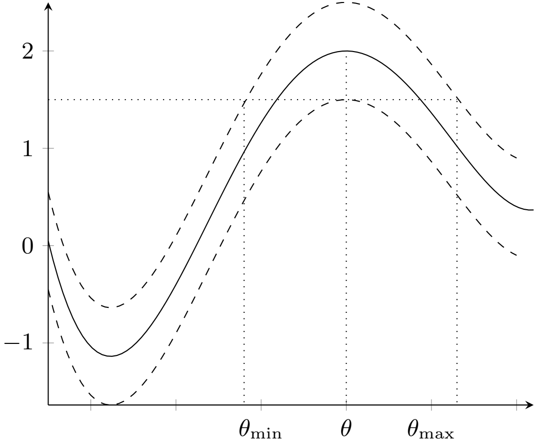
$$\arg\max_{\theta\in\Theta} L_n(\theta) \xrightarrow{p} \arg\max_{\theta\in\Theta} L(\theta),$$

but this is just

$$\hat{\theta} \xrightarrow{p} \theta, \text{ as } n \to \infty,$$

which is consistency.

A uniform ε-band around $L(\theta)$ and the corresponding interval $[\theta_{\min}, \theta_{\max}]$ in which $\hat{\theta}$ must lie.



As $n \to \infty$, the ε-band tightens and so the interval $[\theta_{\min}, \theta_{\max}]$ shrinks to a point. By identification this point must be θ. As $\hat{\theta} \in [\theta_{\min}, \theta_{\max}]$ it must be that $\hat{\theta}$ converges to θ.

Regarding uniform convergence, consider the probit model as an example.

There,

$$\log f_\theta(y|x) = y\,\log\Phi(x'\theta) + (1-y)\log\Phi(-x\theta).$$

We have, by a mean-value expansion, that

$$\log\Phi(x\theta) = \log\Phi(0) + \frac{\phi(x\theta_*)}{\Phi(x\theta_*)}\,x\theta$$

and $0 < \frac{\phi(u)}{\Phi(u)} \leq c\,|1+u|$ for some finite c (visual inspection will help to see this). Consequently,

$$\begin{aligned}|\log\Phi(x\theta)| &\leq |\log\Phi(0)| + \frac{\phi(x\theta_*)}{\Phi(x\theta_*)}\,|x\theta| \leq |\log(2)| + c\,|1+x\theta_*|\,|x||\theta| \\ &\leq |\log(2)| + c\,|x||\theta_*| + c\,|x|^2|\theta|^2\end{aligned}$$

so an integrable upper bound on $\log f_\theta(x_i)$ exists provided $E(x_i^2) < \infty$.

Asymptotic normality

By definition

$$\sum_{i=1}^{n} \left.\frac{\partial \log f_\theta(x_i)}{\partial \theta}\right|_{\hat{\theta}} = 0.$$

A mean-value expansion around the true θ gives

$$\sum_{i=1}^{n} \left.\frac{\partial \log f_\theta(x_i)}{\partial \theta}\right|_{\hat{\theta}} = \sum_{i=1}^{n} \left.\frac{\partial \log f_\theta(x_i)}{\partial \theta}\right|_{\theta} + \sum_{i=1}^{n} \left.\frac{\partial^2 \log f_\theta(x_i)}{\partial \theta \partial \theta'}\right|_{\theta_*} (\hat{\theta} - \theta) = 0,$$

where θ_* is some vector that (elementwise) lies between $\hat{\theta}$ and θ.

Inversion of this equation gives the sampling-error representation

$$(\hat{\theta} - \theta) = - \left(\sum_{i=1}^{n} \left.\frac{\partial^2 \log f_\theta(x_i)}{\partial \theta \partial \theta'}\right|_{\theta_*} \right)^{-1} \sum_{i=1}^{n} \left.\frac{\partial \log f_\theta(x_i)}{\partial \theta}\right|_{\theta}.$$

Now, by invoking a uniform law of large numbers together with consistency,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_\theta(x_i)}{\partial\theta\partial\theta'}\bigg|_{\theta_*} \xrightarrow{p} E_\theta\left(\frac{\partial^2 \log f_\theta(x_i)}{\partial\theta\partial\theta'}\bigg|_{\theta}\right) = -I_\theta.$$

Also, by the central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f_\theta(x_i)}{\partial\theta}\bigg|_{\theta} \xrightarrow{d} N(0, I_\theta).$$

Then, by the continuous-mapping theorem and Slutzky's theorem, we get the influence-function representation

$$\sqrt{n}(\hat{\theta}-\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_\theta^{-1}\,\frac{\partial \log f_\theta(x_i)}{\partial\theta}\bigg|_{\theta} + o_p(1)$$

and we have the following result (note we use the information equality here).

Theorem 15 (Optimality of maximum likelihood)

Under regularity conditions,

$$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0, I_\theta^{-1}).$$

Variance estimation

The information matrix—and, hence, the asymptotic variance of maximum likelihood—can be estimated.

There are two obvious choices.

The first follows from its definition as the variance of the score:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \log f_\theta(x_i)}{\partial\theta}\frac{\partial \log f_\theta(x_i)}{\partial\theta'}\right)\bigg|_{\hat{\theta}}.$$

The second follows from the information equality:

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_\theta(x_i)}{\partial\theta\partial\theta'}\bigg|_{\hat{\theta}}.$$

In both cases, a uniform law of large numbers can be used to show consistency.

The square root of the diagonal entries (of the inverse) give (estimated) standard errors on the maximum-likelihood estimator (after dividing through by $\sqrt{n}$) and so can serve to assess its precision. They will equally serve us in testing later on.

Labor-force participation

Consider the decision of married women to participate to labor market, y_i.

Individuals make decisions based on their own situation/characteristics, x_i. PSID has data on a variety of characteristics (age, education, number of children, and so on).

Standard Bernoulli is too simple to capture this dependence on observable characteristics.

A (possible) specification for a conditional model would be

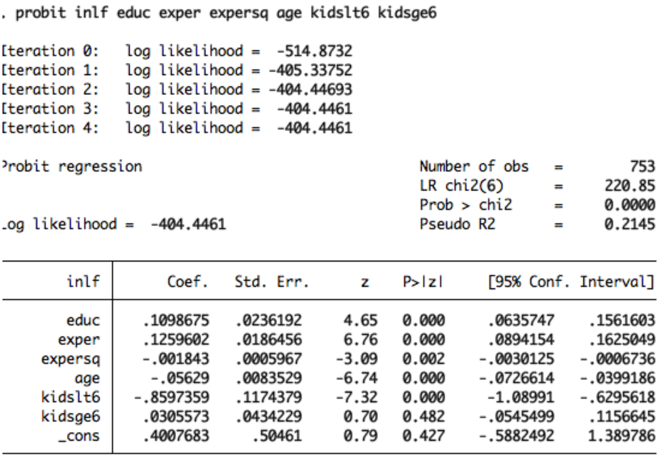
$$p_i = P(y_i = 1|x_i) = \Phi(x_i'\beta).$$

Here, choice probabilities are heterogenous in characteristics.

We can derive an econometric model from a specification of an economic model for the women's decision problem:

$$y_i = 1 \Leftrightarrow u(x_i, \varepsilon_i) \geq 0;$$

$u(x_i, \varepsilon_i)$ is i's utility from working; ε_i is not observed to the econometrician. Our specification has $u(x_i, \varepsilon_i) = x_i'\beta + \varepsilon_i$ for $\varepsilon_i \sim N(0, 1)$ independent of x_i.



```
. probit inlf educ exper expersq age kidslt6 kidsge6

Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -405.33752
Iteration 2:   log likelihood = -404.44693
Iteration 3:   log likelihood =  -404.4461
Iteration 4:   log likelihood =  -404.4461

Probit regression                                 Number of obs   =        753
                                                  LR chi2(6)      =     220.85
                                                  Prob > chi2     =     0.0000
Log likelihood =  -404.4461                       Pseudo R2       =     0.2145

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1098675   .0236192     4.65   0.000     .0635747    .1561603
       exper |   .1259602   .0186456     6.76   0.000     .0894154    .1625049
     expersq |   -.001843   .0005967    -3.09   0.002    -.0030125   -.0006736
         age |    -.05629   .0083529    -6.74   0.000    -.0726614   -.0399186
     kidslt6 |  -.8597359   .1174379    -7.32   0.000     -1.08991   -.6295618
     kidsge6 |   .0305573   .0434229     0.70   0.482    -.0545499    .1156645
       _cons |   .4007683     .50461     0.79   0.427    -.5882492    1.389786
------------------------------------------------------------------------------
```

What are the parameters of interest in the probit model?

Take x_i scalar continuous for a moment.

The average marginal effect is

$$\frac{\partial E(y_i|x_i)}{\partial x_i} = \frac{\partial \Phi(x_i\beta)}{\partial x_i} = \beta\, \phi(x_i\beta).$$

This is nonlinear and heterogenous.

Can look at the distribution of this marginal effect (in x_i), and its functionals.

For example, the mean

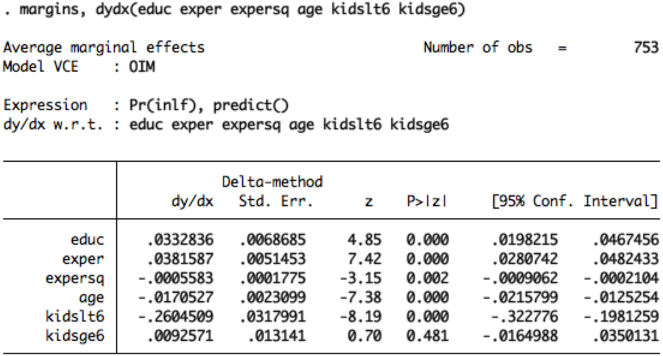
$$\theta = E\left(\frac{\partial E(y_i|x_i)}{\partial x_i}\right) = \beta\, E(\phi(x_i\beta)).$$

The maximum-likelihood estimator is

$$\hat{\theta} = n^{-1} \sum_{i=1}^{n} \hat{\beta}\, \phi(x_i\hat{\beta}).$$

To obtain a standard error, use the Delta method.

We can also look at other functionals of the distribution of the marginal effects.



```
. margins, dydx(educ exper expersq age kidslt6 kidsge6)

Average marginal effects                          Number of obs   =        753
Model VCE    : OIM

Expression   : Pr(inlf), predict()
dy/dx w.r.t. : educ exper expersq age kidslt6 kidsge6

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0332836   .0068685     4.85   0.000     .0198215    .0467456
       exper |   .0381587   .0051453     7.42   0.000     .0280742    .0482433
     expersq |  -.0005583   .0001775    -3.15   0.002    -.0009062   -.0002104
         age |  -.0170527   .0023099    -7.38   0.000    -.0215799   -.0125254
     kidslt6 |  -.2604509   .0317991    -8.19   0.000     -.322776   -.1981259
     kidsge6 |   .0092571    .013141     0.70   0.481    -.0164988    .0350131
------------------------------------------------------------------------------
```

Classical linear regression

The classical linear regression model is an extension of the location/scale model from above in that it adds regressors.

Data on outcome y_i and a (column) vector of regressors (or covariates or explanatory variables) x_i.

The model is

$$y_i|x_i \sim N(x_i'\beta, \sigma^2),$$

and $\theta = (\beta', \sigma^2)'$. Equivalently (and more commonly) we can write the model as

$$y_i = x_i'\beta + \varepsilon_i, \qquad \varepsilon_i|x_i \sim N(0, \sigma^2).$$

Unless explicitely stated otherwise the first covariate is taken to be a constant term.

This is a simple model for analyzing how the distribution of y_i changes with x_i. Here, only impact is through the mean:

$$E(y_i|x_i) = x_i'\beta.$$

Often convenient to look at this model in matrix form.

We have a set of n equations with k regressors, as in

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

which we write as

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}).$$

The log-likelihood (conditional on the regressors) then is

$$-\frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_i \frac{(y_i - x_i'\beta)^2}{\sigma^2} = -\frac{n}{2}\log\sigma^2 - \frac{(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)}{2\sigma^2}.$$

The score equation for β is

$$\frac{\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\beta)}{\sigma^2} = 0.$$

It has the unique solution

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

(independent of σ^2) provided that the inverse of the matrix $\boldsymbol{X}'\boldsymbol{X}$ exists.

This is known as the ordinary least-squares estimator.

So, $\hat{\beta}$ is uniquely defined if $\boldsymbol{X}$ has maximal rank. This is the well-known 'no-multicolinearity condition'. No column of $\boldsymbol{X}$ can be written as a linear combination of the other columns.

A simple counter-example is the dummy-variable trap, where the regressors would be a constant and a collection of dummies for events whose union happens with probability one.

Say, $x_i = (1, d_i, 1 - d_i)'$ where d_i is a binary indicator.

Then $x_{i,1} = x_{i,2} + x_{i,3}$ for all i and the rank condition fails.

The model

$$y_i = \beta_1 + d_i \beta_2 + (1 - d_i)\beta_3 + \varepsilon_i$$

is observationally-equivalent to the three-parameter/two-regressor model

$$y_i = (\beta_1 + \beta_3) + d_i(\beta_2 - \beta_3) + \varepsilon_i = \alpha_1 + d_i \alpha_2 + \varepsilon_i.$$

We can only learn the reduced-form parameters (α_1, α_2). The identified set for β is

$$\{\beta \in \mathbb{R} : \beta_1 + \beta_3 = \alpha_1, \beta_2 - \beta_3 = \alpha_2\}.$$

For example, given β_3, we can back out (β_1, β_2) but, without this knowledge, we can only say things such as $\beta_1 - \beta_2 = \alpha_1 - \alpha_2$.

As another example of identification failure, suppose that we do not observe y_i in the data but, instead, observe variables $\underline{y}_i \leq \overline{y}_i$ for which we know that

$$\underline{y}_i \leq y_i \leq \overline{y}_i$$

(income data in social security records, for example, is often bracketed in this way). Here we cannot even compute the value of the likelihood.

Then the conditional mean is only restricted by

$$E(\underline{y}_i|x_i) \leq x_i'\beta \leq E(\overline{y}_i|x_i)$$

(where we use $E(y_i|x_i) = x_i'\beta$).

An implication is that

$$E(x_i\underline{y}_i) \leq E(x_i x_i')\,\beta \leq E(x_i\overline{y}_i).$$

We can estimate all β compatible with this moment inequality by the set $[\underline{\hat{\beta}}, \overline{\hat{\beta}}]$, with

$$\underline{\hat{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underline{\boldsymbol{y}}, \qquad \overline{\hat{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\overline{\boldsymbol{y}}$$

in obvious notation.

We will write

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}, \qquad \hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta},$$

for fitted values and residuals, respectively.

We have the decomposition

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}},$$

where the fitted values and residuals are uncorrelated, i.e., $\hat{\boldsymbol{y}}'\hat{\boldsymbol{\varepsilon}} = 0$. Indeed, the score equation at $\hat{\beta}$ equals

$$\frac{\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}}}{\sigma^2} = 0,$$

so we can say that $\hat{\beta}$ gives us thát linear combination of the regressors for which residuals and regressors are exactly uncorrelated.

An implication is that

$$\boldsymbol{y}'\boldsymbol{y} = (\hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}})'(\hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}) = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}},$$

and so the uncentered R^2

$$R_u^2 = \frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}}{\boldsymbol{y}'\boldsymbol{y}} = 1 - \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\boldsymbol{y}'\boldsymbol{y}} \in [0,1]$$

gives a relative contribution of the variation in fitted values to the variation in observed outcomes.

More popular to report is the (centered) R^2, which looks at deviations from the mean, as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS},$$

where the total sum of squares decomposes as

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = ESS + SSR,$$

into the explained sum of squares and sum of squared residuals (note that $\overline{\hat{y}} = \overline{y}$ because $\sum_i \hat{\varepsilon}_i = 0$).

The intuition is that we want to compare the improvement in fit of a model with regressors to a model without regressors.

Such a desire for fit comes from the use of the regression model to form linear predictions.

The vector $\boldsymbol{y}$ is a point in $\mathbb{R}^n$. The column space of the $n \times k$ matrix $\boldsymbol{X}$ is the subspace of linear combinations

$$\mathcal{X} = \{\boldsymbol{a} \in \mathbb{R}^n : \boldsymbol{a} = \boldsymbol{X}\boldsymbol{b} \text{ for some vector } \boldsymbol{b}\}.$$

That is, $\mathcal{X}$ is the vector space spanned by the columns of $\boldsymbol{X}$. If $\operatorname{rank} \boldsymbol{X} = k$ the columns of $\boldsymbol{X}$ are basis vectors for $\mathcal{X}$.

The orthogonal projection of $\boldsymbol{y}$ onto $\mathcal{X}$ is the solution to

$$\min_{\boldsymbol{a} \in \mathcal{X}} \|\boldsymbol{y} - \boldsymbol{a}\| = \min_{b \in \mathbb{R}^k} \|\boldsymbol{y} - \boldsymbol{X} b\| = \min_{b \in \mathbb{R}^k} \sqrt{(\boldsymbol{y} - \boldsymbol{X} b)'(\boldsymbol{y} - \boldsymbol{X} b)}$$

and equals

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}.$$

The deviation of $\boldsymbol{y}$ from its projection is

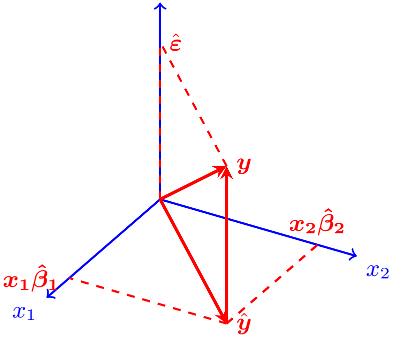
$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y},$$

and lives in the orthogonal complement $\mathcal{X}^{\perp}$, so $\hat{\boldsymbol{y}}'\hat{\boldsymbol{\varepsilon}} = 0$. The projection matrices

$$\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}', \qquad \boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}',$$

will prove convenient. Note that $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}}'$ and $\boldsymbol{P}_{\boldsymbol{X}}^2 = \boldsymbol{P}_{\boldsymbol{X}}$ (and the same for $\boldsymbol{M}_{\boldsymbol{X}}$) and that $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{0}$.

Least squares projection in a three-dimensional space:



Partition $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ so that

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{\varepsilon}$$

and, hence,

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{y} \end{pmatrix}.$$

Some algebra using formulae for partitioned matrix inversion shows that

$$\hat{\beta}_1 = (\boldsymbol{X}_1'\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{X}_1)^{-1}(\boldsymbol{X}_1'\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{y})$$

(and likewise for $\hat{\beta}_2$).

$\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{y}$ is the residual vector of a regression of $\boldsymbol{y}$ on $\boldsymbol{X}_2$.

$\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{X}_1$ is the residual matrix of a regression of the columns of $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$.

These residuals are uncorrelated with $\boldsymbol{X}_2$. Moreover, $\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{y}$ is $\boldsymbol{y}$, and $\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{X}_1$ is $\boldsymbol{X}_1$, respectively, after their linear dependence on $\boldsymbol{X}_2$ has been filtered out. $\hat{\beta}_1$ is the slope coefficient in a regression of these residuals on each other.

This gives (multiple) least squares its partialling-out interpretation. The results is known as the Frisch-Waugh-Lovell theorem.

The estimator $\hat{\beta}$ is (conditionally) unbiased,

$$E(\hat{\beta}|\boldsymbol{X}) = E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{y}|\boldsymbol{X}) = \beta$$

(because $E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\beta$). Its variance is

$$\begin{aligned}
\text{var}(\hat{\beta}|\boldsymbol{X}) &= E((\hat{\beta}-\beta)(\hat{\beta}-\beta)'|\boldsymbol{X}) \\
&= E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned}$$

In fact, its exact (conditional) distribution is normal,

$$\hat{\beta} \sim N(\beta, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}),$$

because, for any conformable non-stochastic matrix $\boldsymbol{A}$, $\boldsymbol{A}\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{A}\boldsymbol{A}')$, and thus also for $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

It is also best unbiased, as $\hat{\beta} - \beta = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$ is proportional to the score equation at the truth ($\boldsymbol{X}'\boldsymbol{\varepsilon}/\sigma^2$), with factor of proportionality equal to $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

The score equation for σ^2 is

$$-\frac{n}{\sigma^2} + \frac{(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)}{\sigma^4} = 0.$$

which, given $\hat{\beta}$, has solution

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}.$$

We already know that this estimator is biased; an unbiased version would be

$$\tilde{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k}.$$

Indeed,

$$E\left(\left.\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}\right| \boldsymbol{X}\right) = \frac{E\left(\boldsymbol{y}'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}|\boldsymbol{X}\right)}{n} = \frac{E\left(\boldsymbol{\varepsilon}'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{\varepsilon}|\boldsymbol{X}\right)}{n} = \sigma^2\frac{\mathrm{tr}(\boldsymbol{M}_{\boldsymbol{X}})}{n} = \sigma^2\frac{n-k}{n}$$

because

$$\mathrm{tr}(\boldsymbol{M}_{\boldsymbol{X}}) = \mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) = \mathrm{tr}(\boldsymbol{I}_n) - \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{X}}) = n - k;$$

using that $\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{X}}) = \mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \mathrm{tr}(\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) = \mathrm{tr}(I_k) = k$.
Finally,

$$(n-k)\,\tilde{\sigma}^2/\sigma^2 \sim \chi^2_{n-k}.$$

Now consider the behavior of the estimators as $n \to \infty$. We let $\Sigma = E(x_i x_i')$.

First,

$$\hat{\beta} - \beta = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon},$$

and

$$\frac{\boldsymbol{X}'\boldsymbol{X}}{n} \xrightarrow{p} \Sigma, \qquad \frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{n} \xrightarrow{p} 0,$$

so that $\hat{\beta} \xrightarrow{p} \beta$ by the continuous-mapping theorem.

Next, we have

$$\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} = \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1} \frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{\sqrt{n}},$$

and

$$\frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{\sqrt{n}} \xrightarrow{d} N(0, \sigma^2\Sigma).$$

So,

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Sigma^{-1} x_i \varepsilon_i + o_p(1) \xrightarrow{d} N(0, \sigma^2\Sigma^{-1}).$$

The influence function here is $\Sigma^{-1} x_i \varepsilon_i$.

For the variance estimator,

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n} \\
&= \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})}{n} \\
&= \frac{(\boldsymbol{X}(\beta - \hat{\beta}) + \boldsymbol{\varepsilon})'(\boldsymbol{X}(\beta - \hat{\beta}) + \boldsymbol{\varepsilon})}{n} \\
&= \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} + \frac{(\hat{\beta} - \beta)'(\boldsymbol{X}'\boldsymbol{X})(\hat{\beta} - \beta)}{n} + 2(\hat{\beta} - \beta)'\frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{n} \\
&= \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} + o_p(n^{-1/2}),
\end{aligned}$$

because $\|\hat{\beta} - \beta\| = O_p(n^{-1/2})$, $\boldsymbol{X}'\boldsymbol{X} = O_p(n)$ and $\boldsymbol{X}'\boldsymbol{\varepsilon} = o_p(n)$.

Therefore,

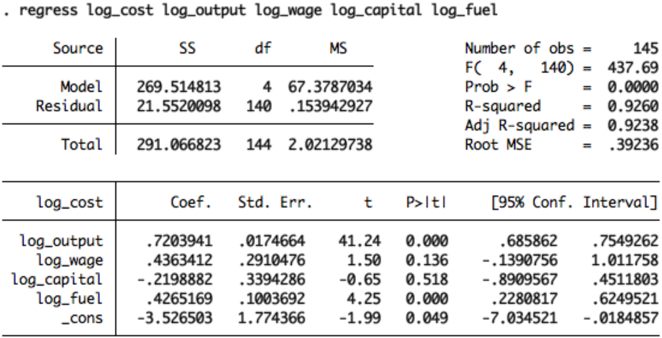
$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) = \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})}{\sqrt{n}} + o_p(1) = \sum_{i=1}^{n} \frac{\varepsilon_i^2 - E(\varepsilon_i^2)}{\sqrt{n}} + o_p(1) \xrightarrow{d} N(0, 2\sigma^4)$$

(recall that, under normality, $E(\varepsilon_i^4) = 3\sigma^4$.)

This estimator is best asymptotically unbiased (and so is $\tilde{\sigma}^2$).

Production and cost function

Suppose a firm creates output according to Cobb-Douglas technology. The associated cost function is linear in logs. The regressors are a constant term, (the log of) total output, and the log of the price of inputs (labor, capital, and so on).



```
. regress log_cost log_output log_wage log_capital log_fuel

      Source |       SS       df       MS              Number of obs =     145
-------------+------------------------------           F(  4,   140) =  437.69
       Model |  269.514813     4  67.3787034           Prob > F      =  0.0000
    Residual |  21.5520098   140  .153942927           R-squared     =  0.9260
-------------+------------------------------           Adj R-squared =  0.9238
       Total |  291.066823   144  2.02129738           Root MSE      =  .39236

------------------------------------------------------------------------------
    log_cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  log_output |   .7203941   .0174664    41.24   0.000      .685862    .7549262
    log_wage |   .4363412   .2910476     1.50   0.136    -.1390756    1.011758
 log_capital |  -.2198882   .3394286    -0.65   0.518    -.8909567    .4511803
    log_fuel |   .4265169   .1003692     4.25   0.000     .2280817    .6249521
       _cons |  -3.526503   1.774366    -1.99   0.049    -7.034521   -.0184857
------------------------------------------------------------------------------
```

The coefficients on the input prices are elasticities.

Poisson regression

The linear regression model will typically be inappropriate when data are not continuous.

An example is $y_i \in \mathbb{N}$, i.e., count data.

Patent-application data fits this framework.

A Poisson regression model has (conditional) mass function

$$\frac{\mu_i^{y_i}\, e^{-\mu_i}}{y_i!}, \qquad \mu_i = e^{x_i'\beta}.$$

Remember that $\mu_i = e^{x_i'\beta}$ is the conditional mean of the outcome variable.

The log-likelihood (up to a constant) is

$$\sum_{i=1}^{n}(y_i x_i'\beta - e^{x_i'\beta}) + \text{constant}.$$

The score equation is

$$\sum_{i=1}^{n} x_i(y_i - e^{x_i'\beta}) = 0,$$

and the Hessian matrix is

$$-\sum_{i=1}^{n}(x_i x_i')\, e^{x_i'\beta} < 0.$$

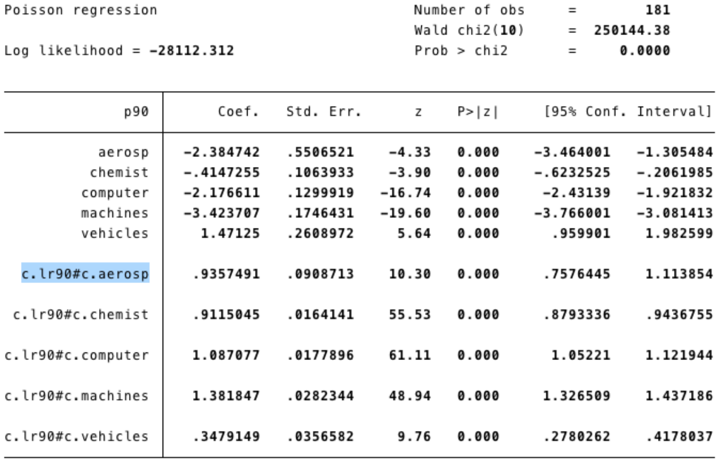
The maximum-likelihood estimator of β is not unbiased.

It is best asymptotically unbiased, however, with limit distribution

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, E(x_i x_i' \mu_i)^{-1}).$$

Patent applications (innovation) and R&D spending

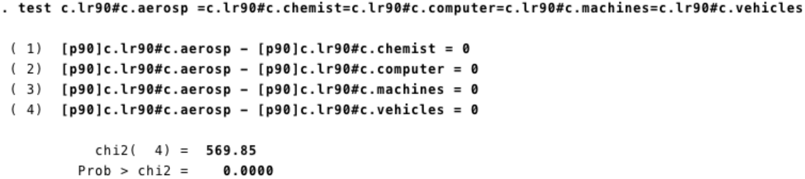
The # of patents applied for on R&D spending, stratified by sector.



```
Poisson regression                              Number of obs     =        181
                                                Wald chi2(10)     =  250144.38
Log likelihood = -28112.312                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------------
                p90 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------+---------------------------------------------------------------
             aerosp |  -2.384742   .5506521    -4.33   0.000    -3.464001   -1.305484
            chemist |  -.4147255   .1063933    -3.90   0.000    -.6232525   -.2061985
           computer |  -2.176611   .1299919   -16.74   0.000     -2.43139   -1.921832
           machines |  -3.423707   .1746431   -19.60   0.000    -3.766001   -3.081413
           vehicles |    1.47125   .2608972     5.64   0.000      .959901    1.982599
                    |
     c.lr90#c.aerosp |   .9357491   .0908713    10.30   0.000     .7576445    1.113854
                    |
    c.lr90#c.chemist |   .9115045   .0164141    55.53   0.000     .8793336    .9436755
                    |
   c.lr90#c.computer |   1.087077   .0177896    61.11   0.000      1.05221    1.121944
                    |
   c.lr90#c.machines |   1.381847   .0282344    48.94   0.000     1.326509    1.437186
                    |
   c.lr90#c.vehicles |   .3479149   .0356582     9.76   0.000     .2780262    .4178037
------------------------------------------------------------------------------------
```

We can test whether the impact of R&D spending on innovation is different across sectors.



```
. test c.lr90#c.aerosp =c.lr90#c.chemist=c.lr90#c.computer=c.lr90#c.machines=c.lr90#c.vehicles

 ( 1)  [p90]c.lr90#c.aerosp - [p90]c.lr90#c.chemist = 0
 ( 2)  [p90]c.lr90#c.aerosp - [p90]c.lr90#c.computer = 0
 ( 3)  [p90]c.lr90#c.aerosp - [p90]c.lr90#c.machines = 0
 ( 4)  [p90]c.lr90#c.aerosp - [p90]c.lr90#c.vehicles = 0

           chi2(  4) =  569.85
         Prob > chi2 =    0.0000
```

Examples of maximum likelihood

Arellano, M. and C. Meghir (1992). Female labour supply and on-the-job search: An empirical model estimated using complementary data sets. *Review of Economic Studies* 59, 537–557.

Blundell, R., P.-A. Chiappori, T. Magnac, and C. Meghir (2007). Collective labour supply: Heterogeneity and non-participation. *Review of Economic Studies* 74, 417–445.

Heckman, J. J. (1974). Shadow prices, market wages, and labor supply. *Econometrica* 42, 679–694.

Magnac, T. (1991). Segmented or competitive labor markets. *Econometrica* 59, 165–187.

Rust, J. (1987). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica* 55, 999–1033.

Tobin, J. (1958). Estimation of relationships from limited dependent variables. *Econometrica* 26, 24–36.

Bayesian estimation

▸ Skip the remainder of this section

Before seeing the data you have beliefs about θ. Suppose we can summarize those beliefs into a distribution function $\pi(\theta)$ on Θ, the prior.

Upon seeing the data we can evaluate the distribution of the sample, $\prod_{i=1}^n f_\theta(x_i)$, at any $\theta \in \Theta$.

When confronted with the data we may alter our beliefs about θ. Bayes rule gives the posterior as

$$\pi(\theta|x_1,\ldots,x_n) = \frac{\prod_{i=1}^n f_\theta(x_i)\,\pi(\theta)}{\int_\Theta \prod_{i=1}^n f_u(x_i)\,\pi(u)\,du}.$$

When prior is a proper distribution, so is the posterior.

The updated beliefs summarized in the posterior distribution can be used to construct a point estimator if desired. An example is

$$\int_\Theta \theta\,\pi(\theta|x_1,\ldots,x_n)\,d\theta,$$

the posterior mean.

Other natural choices would be the posterior median and mode.

Normal example

Take

$$x_i \sim N(\theta, \sigma^2)$$

(with σ^2 known for simplicity), so that

$$f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sigma}\phi\left(\frac{x_i - \theta}{\sigma}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2}\left(\frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}{\sigma^2} + \frac{n(\theta - \overline{x}_n)^2}{\sigma^2}\right)}.$$

Suppose that

$$\pi(\theta) = \frac{1}{\tau}\phi\left(\frac{\theta - \mu}{\tau}\right) = \frac{1}{(2\pi\tau^2)^{1/2}} e^{-\frac{1}{2}\left(\frac{(\theta-\mu)^2}{\tau^2}\right)},$$

i.e., our prior belief is that $\theta \sim N(\mu, \tau^2)$.

A calculation gives the posterior as normal, i.e.,

$$\theta | (x_1, \ldots, x_n) \sim N\left(\frac{\tau^2}{\tau^2 + \sigma^2/n}\overline{x}_n + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\mu, \frac{\tau^2\sigma^2/n}{\tau^2 + \sigma^2/n}\right).$$

The posterior mean is the point estimator

$$\frac{\tau^2}{\tau^2+\sigma^2/n}\overline{x}_n + \frac{\sigma^2/n}{\tau^2+\sigma^2/n}\mu,$$

which is a weighted average of the sample mean $\overline{x}_n$ (the Frequentist estimator) and the prior mean μ.

Note that this estimator is not unbiased.

In fact, Bayesian posterior means are never unbiased.

As $n \to \infty$, the relative contribution of the prior vanishes and

$$\theta|(x_1,\dots,x_n) \overset{a}{\sim} N\left(\overline{x}_n, \sigma^2/n\right),$$

which is the Frequentist asymptotic approximation.

Bernstein-von Mises theorem

The similarity between the Bayesian posterior and the Frequentist asymptotic-distribution approximation in the above example holds much more generally.

This is the Bernstein-von Mises result.

It states that, in an appropriate metric (known as the total-variation norm), the difference between $\pi(\theta|x_1, \ldots, x_n)$ and

$$N(\hat{\theta}, I_\theta^{-1}/n)$$

converges to zero in probability as $n \to \infty$.

One implication is that both procedures are asymptotically equivalent.

James-Stein estimation

Remember the posterior mean in our example was

$$\frac{\tau^2}{\tau^2+\sigma^2/n}\overline{x}_n+\frac{\sigma^2/n}{\tau^2+\sigma^2/n}\mu=\left(1-\frac{(\sigma^2/\tau^2)/n}{1+(\sigma^2/\tau^2)/n}\right)\overline{x}_n+\left(\frac{(\sigma^2/\tau^2)/n}{1+(\sigma^2/\tau^2)/n}\right)\mu.$$

For example, when $\mu=0$ (for notational simplicity) we have

$$\left(1-\frac{(\sigma^2/\tau^2)/n}{1+(\sigma^2/\tau^2)/n}\right)\overline{x}_n=\left(1-\frac{\sigma^2/n}{\tau^2+\sigma^2/n}\right)\overline{x}_n$$

The term in brackets lies in $(0,1)$. So, this estimator is downward biased. The bias is introduced by the shrinkage of $\overline{x}_n$ towards the prior mean of zero.

A multivariate version would have $\overline{\boldsymbol{x}}\sim N(\boldsymbol{\theta},(\sigma^2/n)\,\boldsymbol{I}_m)$ and $\boldsymbol{\theta}\sim N(\boldsymbol{0},\tau^2\boldsymbol{I}_m)$ with shrinkage estimator

$$\left(1-\frac{(\sigma^2/\tau^2)/n}{1+(\sigma^2/\tau^2)/n}\right)\overline{\boldsymbol{x}}.$$

The James-Stein estimator (assuming that σ^2 is known and that $m \geq 2$) is

$$\left(1 - \sigma^2 \frac{m-2}{\|\overline{\boldsymbol{x}}\|^2}\right) \overline{\boldsymbol{x}}.$$

While this estimator is biased, we have

$$E(\|\overline{\boldsymbol{x}} - \mathbf{0}\|^2) > E\left(\left\|\left(1 - \sigma^2 \frac{m-2}{\|\overline{\boldsymbol{x}}\|^2}\right) \overline{\boldsymbol{x}} - \mathbf{0}\right\|^2\right).$$

as soon as $m > 2$.

So, in terms of estimation risk (as measured by expected squared loss), the James-Stein estimator dominates the Frequentist sample mean estimator $\overline{\boldsymbol{x}}$.

The key is that shrinkage reduces variance. Indeed, taking the infeasible estimator for simplicity

$$\begin{aligned} \operatorname{var}\left(\left(1 - \frac{(\sigma^2/\tau^2)/n}{1 + (\sigma^2/\tau^2)/n}\right) \overline{\boldsymbol{x}}\right) &= \left(1 - \frac{(\sigma^2/\tau^2)/n}{1 + (\sigma^2/\tau^2)/n}\right) \tau^2 \, \boldsymbol{I}_m \\ &= \left(\tau^2 - \frac{\sigma^2/\tau^2}{n}\right) \boldsymbol{I}_m + o(n^{-1}). \end{aligned}$$

TESTING IN PARAMETRIC PROBLEMS

Reading

General discussion:

Casella and Berger, Chapter 8

Hansen I, Chapter 13 and 14

Testing in the likelihood framework:

Davidson and MacKinnon, Chapter 13

Hansen II, Chapter 9

Classical linear regression model:

Goldberger, Chapters 19–21

Simple hypothesis and likelihood ratio

Suppose we wish to test the simple null $H_0 : \theta = \theta_0$ against the simple alternative $H_1 : \theta = \theta_1$.

The data distribution is completely specified under both null and alternative.

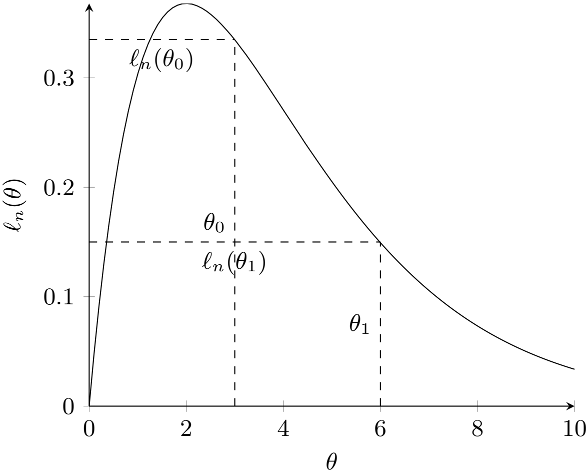
Write

$$\ell_n(\theta) = e^{L_n(\theta)} = \prod_{i=1}^{n} f_\theta(x_i)$$

for the likelihood and define the likelihood ratio as

$$\frac{\ell_n(\theta_0)}{\ell_n(\theta_1)}.$$

If H_0 is false we would expect $\ell_n(\theta_0)/\ell_n(\theta_1)$ to be small.


$\ell_n(\theta_0)$
0.3
0.2
$\ell_n(\theta)$
θ_0
$\ell_n(\theta_1)$
0.1
θ_1
0
0
2
4
6
8
10
θ

A decision rule based on the likelihood ratio is to

Reject the null in favor of the alternative when

$$\frac{\ell_n(\theta_0)}{\ell_n(\theta_1)} < c,$$

Accept the null when

$$\frac{\ell_n(\theta_0)}{\ell_n(\theta_1)} \geq c,$$

for a chosen value c.

We might wrongfully reject the null. This is called a type-I error.

The significance level or size of the test is

$$P_{\theta_0}\left(\frac{\ell_n(\theta_0)}{\ell_n(\theta_1)} < c\right).$$

We might wrongfully accept the null. This is called a type-II error.

The power of the test is

$$P_{\theta_1}\left(\frac{\ell_n(\theta_0)}{\ell_n(\theta_1)} < c\right).$$

Normal

Suppose that

$$x_i \sim N(\theta, \sigma^2)$$

for known σ^2.

(From before; see Slide 49) the density of the data is

$$\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2}\left(\frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}{\sigma^2}\right)} e^{-\frac{1}{2}\left(\frac{n(\theta - \overline{x}_n)^2}{\sigma^2}\right)}$$

The likelihood ratio thus is

$$\frac{e^{-\frac{1}{2}\left(\frac{n(\theta_0 - \overline{x}_n)^2}{\sigma^2}\right)}}{e^{-\frac{1}{2}\left(\frac{n(\theta_1 - \overline{x}_n)^2}{\sigma^2}\right)}} = e^{-\frac{1}{2}\frac{n}{\sigma^2}\left((\overline{x}_n - \theta_0)^2 - (\overline{x}_n - \theta_1)^2\right)} = e^{\frac{\theta_0 - \theta_1}{\sigma/\sqrt{n}}\left(\frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} + \frac{1}{2}\frac{\theta_0 - \theta_1}{\sigma/\sqrt{n}}\right)}.$$

If $\theta_0 < \theta_1$ the likelihood ratio is no greater than c when

$$\frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} \geq c^*,$$

for some c^*.

So a level α test is obtained on choosing c^* so that

$$P_{\theta_0}\left(\frac{\ell_n(\theta_0)}{\ell_n(\theta_1)} < c\right) = P_{\theta_0}\left(\frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} \geq c^*\right) = 1 - \Phi(c^*) = \alpha,$$

which requires that

$$c^* = \Phi^{-1}(1-\alpha) \equiv z_\alpha,$$

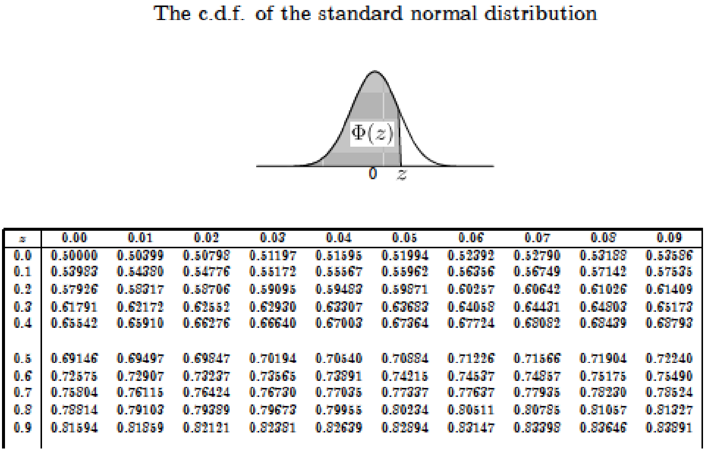
the $(1-\alpha)$th quantile of the standard-normal distribution. These values are tabulated.

Then the decision rule we obtain is that, if,

$$\frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} \geq z_\alpha,$$

we reject the null in favor of the alternative.

The standard-normal distribution



The c.d.f. of the standard normal distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891

The power of the test is

$$P_{\theta_1}\left(\frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} \geq z_\alpha\right) = P_{\theta_1}\left(\frac{\overline{x}_n - \theta_1}{\sigma/\sqrt{n}} + \frac{\theta_1 - \theta_0}{\sigma/\sqrt{n}} \geq z_\alpha\right) = 1-\Phi\left(z_\alpha - \frac{\theta_1 - \theta_0}{\sigma/\sqrt{n}}\right).$$

Note that the power increases when

- the difference $\theta_1 - \theta_0$ (> 0 here) increases; and
- the samples size n increases.

The latter observation implies that

$$P_{\theta_1}\left(\frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} \geq z_\alpha\right) \overset{n\to\infty}{\to} 1,$$

i.e., if the null is false this will be spotted with probability approaching one. This is called consistency of a test.

Note that if, in stead, $\theta_0 > \theta_1$, the decision rule becomes that, if,

$$\frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} \leq -z_\alpha,$$

we reject the null in favor of the alternative.

Now take the reverse situation where

$$x_i \sim N(\mu, \theta)$$

and μ is known.

Wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

The likelihood ratio is

$$(\theta_1/\theta_0)^{n/2} \, e^{-\frac{1}{2} \frac{\theta_1 - \theta_0}{\theta_1} \sum_{i=1}^n \frac{(x_i-\mu)^2}{\theta_0}}.$$

If $\theta_0 < \theta_1$ this is small when

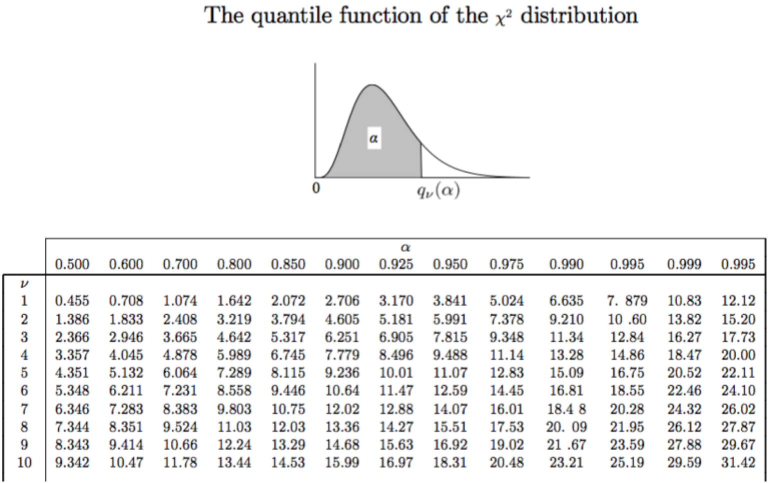
$$\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\theta_0}$$

is large (and vice versa). Now, under the null, this statistic is χ^2_n and so

$$P_{\theta_0}\left(\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\theta_0} \geq \chi^2_{n,\alpha} \right) = \alpha,$$

where $\chi^2_{n,\alpha}$ is the $(1 - \alpha)$th quantile of the χ^2_n distribution. The power is the probability that a χ^2_n is greater than $\chi^2_{n,\alpha}(\theta_0/\theta_1)$.

The χ^2-distribution



The quantile function of the χ^2 distribution

α
0
$q_\nu(\alpha)$

						α						
	0.500	0.600	0.700	0.800	0.850	0.900	0.925	0.950	0.975	0.990	0.995	0.999	0.995
ν													
1	0.455	0.708	1.074	1.642	2.072	2.706	3.170	3.841	5.024	6.635	7. 879	10.83	12.12
2	1.386	1.833	2.408	3.219	3.794	4.605	5.181	5.991	7.378	9.210	10 .60	13.82	15.20
3	2.366	2.946	3.665	4.642	5.317	6.251	6.905	7.815	9.348	11.34	12.84	16.27	17.73
4	3.357	4.045	4.878	5.989	6.745	7.779	8.496	9.488	11.14	13.28	14.86	18.47	20.00
5	4.351	5.132	6.064	7.289	8.115	9.236	10.01	11.07	12.83	15.09	16.75	20.52	22.11
6	5.348	6.211	7.231	8.558	9.446	10.64	11.47	12.59	14.45	16.81	18.55	22.46	24.10
7	6.346	7.283	8.383	9.803	10.75	12.02	12.88	14.07	16.01	18.4 8	20.28	24.32	26.02
8	7.344	8.351	9.524	11.03	12.03	13.36	14.27	15.51	17.53	20. 09	21.95	26.12	27.87
9	8.343	9.414	10.66	12.24	13.29	14.68	15.63	16.92	19.02	21 .67	23.59	27.88	29.67
10	9.342	10.47	11.78	13.44	14.53	15.99	16.97	18.31	20.48	23.21	25.19	29.59	31.42

Exponential

Take x_i to be exponentially distribution, i.e.,

$$f_\theta(x) = \frac{e^{-x/\theta}}{\theta}, \qquad x \geq 0, \quad \theta > 0.$$

Note that the likelihood is

$$\frac{1}{\theta^n} \prod_{i=1}^{n} e^{-x_i/\theta} = \frac{1}{\theta^n} e^{-n\overline{x}_n/\theta},$$

and so the likelihood-ratio statistic for simple null and alternative equals

$$\frac{\frac{1}{\theta_0^n} e^{-n\overline{x}_n/\theta_0}}{\frac{1}{\theta_1^n} e^{-n\overline{x}_n/\theta_1}} = \left(\frac{\theta_1}{\theta_0}\right)^n e^{-n\overline{x}_n \frac{\theta_1 - \theta_0}{\theta_0 \theta_1}}.$$

If $\theta_1 > \theta_0$ this statistic is small when $n\overline{x}_n$ is large. Now,

$$n\overline{x}_n = \sum_{i=1}^{n} x_i \sim \text{Gamma}(n, \theta)$$

(or Erlang$(n, 1/\theta)$ as n is an integer) so size is easily controlled for any n.

Most powerful test

Theorem 16 (Neyman-Pearson lemma)

When both the null and alternative hypothesis are simple, the likelihood ratio test that rejects the null when

$$\frac{\ell_n(\theta_0)}{\ell_n(\theta_1)} < c$$

for a constant c such that

$$P_{\theta_0}\left(\frac{\ell_n(\theta_0)}{\ell_n(\theta_1)} < c\right) = \alpha$$

is the most powerful test among all level-α tests.

Composite alternatives and the power function

Now test the simple null $H_0 : \theta = \theta_0$ against a composite alternative $H_1 : \theta \in \Theta_1$, where $\Theta_1 \subset \Theta$.

We can generalize the likelihood ratio to

$$\frac{\ell_n(\theta_0)}{\sup_{\theta_1 \in \Theta_1} \ell_n(\theta_1)}$$

The data distribution is no longer fully specified under the alternative; there are many possible alternatives.

A test is uniformly most powerful if it is most powerful against all $\theta_1 \in \Theta_1$.

Power is now a function; i.e.,

$$\beta(\theta) = P_\theta\left(\frac{\ell_n(\theta_0)}{\sup_{\theta_1 \in \Theta_1} \ell_n(\theta_1)} < c\right)$$

Unbiased tests

A level-α test is unbiased if

$$\beta(\theta) \geq \alpha$$

for all $\theta \in \Theta_1$.

The null is more likely to be rejected when it is false than when it is true.

Unbiasedness is clearly desirable.

We could consider looking for the uniformly most powerful unbiased test.

Normal (One-sided)

Again take $x_i \sim N(\theta, \sigma^2)$ for known σ^2.

Now test $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

The set of alternatives is thus $\Theta_1 = \{\theta \in \Theta : \theta > \theta_0\}$. This is a one-sided alternative.

Clearly,

$$\hat{\theta}_1 = \arg\max_{\theta_1 \in \Theta_1} \ell_n(\theta_1) = \overline{x}_n \{\overline{x}_n > \theta_0\} + \theta_0 \{\overline{x}_n \leq \theta_0\}.$$

Then,

$$\begin{aligned}
\frac{\ell_n(\theta_0)}{\ell_n(\hat{\theta}_1)} &= \frac{e^{-\frac{1}{2}\frac{n(\overline{x}_n-\theta_0)^2}{\sigma^2}}}{e^{-\frac{1}{2}\frac{n(\overline{x}_n-\hat{\theta}_1)^2}{\sigma^2}}} \\
&= e^{-\frac{n}{2}\frac{(\overline{x}_n-\theta_0)^2-(\overline{x}_n-\theta_0)^2\{\overline{x}_n\leq\theta_0\}-(\overline{x}_n-\overline{x}_n)^2\{\overline{x}_n>\theta_0\}}{\sigma^2}} \\
&= e^{-\frac{1}{2}\frac{(\overline{x}_n-\theta_0)^2}{\sigma^2/n}\left\{\frac{\overline{x}_n-\theta_0}{\sigma/\sqrt{n}}>0\right\}}
\end{aligned}$$

which is no greater than some constant c if

$$\frac{\overline{x}_n-\theta_0}{\sigma/\sqrt{n}} \left\{\frac{\overline{x}_n-\theta_0}{\sigma/\sqrt{n}} > 0\right\} \geq c^*.$$

The random variable

$$\overline{z}\,\{\overline{z} > 0\}, \qquad \overline{z} \sim N(0,1)$$

is truncated standard normal with cumulative distribution function

$$P(\overline{z} < c^* | \overline{z} > 0) = 2\left(\Phi(c^*) - \frac{1}{2}\right).$$

So, noting that only positive values for c^* make sense,

$$\begin{aligned} P(\overline{z}\,\{\overline{z} > 0\} \leq c^*) &= P(\overline{z} \leq c^* | \overline{z} > 0)\, P(\overline{z} > 0) + P(\overline{z} \leq 0) \\ &= \frac{1}{2}\left(2\left(\Phi(c^*) - \frac{1}{2}\right)\right) + \frac{1}{2} \\ &= \Phi(c^*), \end{aligned}$$

and, therefore, the size of our test can be set to $\alpha \in (0,1)$ by setting

$$c^* = \Phi^{-1}(1-\alpha) = z_\alpha.$$

We get the decision rule

Reject $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta > \theta_0$ if

$$\frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} \left\{ \frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} > 0 \right\} \geq z_\alpha;$$

Accept $H_0 : \theta = \theta_0$ if

$$\frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} \left\{ \frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} > 0 \right\} < z_\alpha.$$

With $z_\alpha > 0$ we can just look at

Reject $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta > \theta_0$ if

$$\frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} \geq z_\alpha;$$

Accept $H_0 : \theta = \theta_0$ if

$$\frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} < z_\alpha.$$

This test is uniformly the most powerful.

This conclusion follows from the fact that the decision rule is the same as for the simple alternative $\theta = \theta_1$ from above, and that test was the most powerful for any $\theta_1 > \theta_0$.

We have

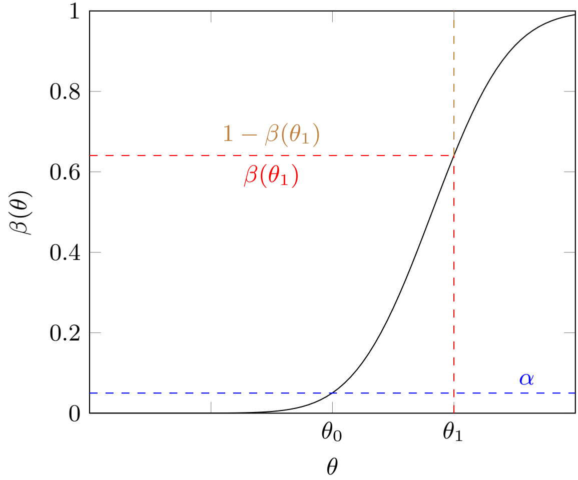
$$P_\theta\left(\frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} \geq z_\alpha\right) = P_\theta\left(\frac{\overline{x}_n - \theta}{\sigma/\sqrt{n}} \geq z_\alpha + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)$$

so the power function is

$$1 - \Phi\left(z_\alpha + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right).$$

This test is consistent.

$\beta(\theta)$ is presented graphically below for a setting where $\theta_0 = 0$ and $\sigma = 1$, with $\alpha = .05$.


1
0.8
0.6
0.4
0.2
0
$\beta(\theta)$
$1-\beta(\theta_1)$
$\beta(\theta_1)$
α
θ_0
θ_1
θ

Normal (Two-sided)

Continue to work with $x_i \sim N(\theta, \sigma^2)$ for known σ^2.

Now test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

The set of alternatives is thus $\Theta_1 = \{\theta \in \Theta : \theta \neq \theta_0\} = \Theta \backslash \{\theta_0\}$. This is a two-sided alternative.

The likelihood-ratio is simply

$$e^{-\frac{1}{2}\left(\frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}}\right)^2},$$

which is no greater than some constant c if

$$\left| \frac{\overline{x}_n - \theta_0}{\sigma/\sqrt{n}} \right| \geq c^*.$$

So,

$$P_{\theta_0}\left(\left|\frac{\overline{x}_n-\theta_0}{\sigma/\sqrt{n}}\right| \geq c^*\right) = 1 - P_{\theta_0}\left(-c^* < \frac{\overline{x}_n-\theta_0}{\sigma/\sqrt{n}} \leq c^*\right)$$

which is simply

$$1-(\Phi(c^*)-\Phi(-c^*)) = 1-(1-\Phi(-c^*)-\Phi(-c^*)) = 2\Phi(-c^*) = 2(1-\Phi(c^*)).$$

Equalizing this probability to α and inverting toward c^* yields

$$c^* = \Phi^{-1}(1-\alpha/2) = z_{\alpha/2},$$

giving the decision rule:

Reject $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$ if

$$\left|\frac{\overline{x}_n-\theta_0}{\sigma/\sqrt{n}}\right| \geq z_{\alpha/2},$$

and accept the null if not.

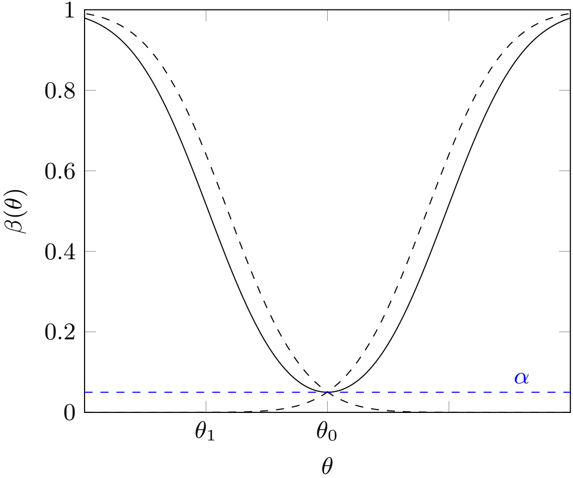
Note that we reject if either

$$\frac{\overline{x}_n-\theta_0}{\sigma/\sqrt{n}} < -z_{\alpha/2} \quad \text{or} \quad \frac{\overline{x}_n-\theta_0}{\sigma/\sqrt{n}} > z_{\alpha/2};$$

each of these events has probability $\alpha/2$ under the null.

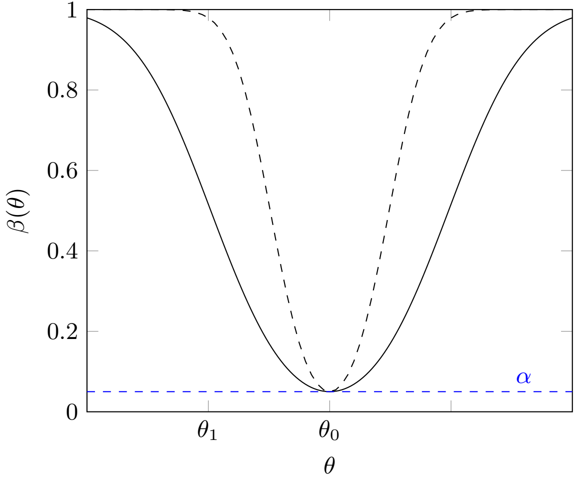
This test is not uniformly most powerful. In fact, for two-sided alternatives, such tests cannot exist.

The one-sided tests with size α are better on their respective sides of the null:



The two-sided test is unbiased and consistent.

Below are the power functions for two sample sizes.



Normal (Two-sided; variance unknown)

Again

$$x_i \sim N(\mu, \sigma^2)$$

but now with both μ, σ^2 unknown.

Consider the hypothesis

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0.$$

The likelihood is

$$\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2}\frac{\sum_{i=1}^n (x_i-\mu)^2}{\sigma^2}}.$$

The unconstrained maximizers are

$$\hat{\mu} = \overline{x}_n, \qquad \hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (x_i - \overline{x}_n)^2,$$

while, when $\mu = \mu_0$, maximizing with respect to σ^2 only yields

$$\check{\sigma}^2 = n^{-1}\sum_{i=1}^n (x_i - \mu_0)^2.$$

The likelihood ratio is simply

$$\left(\frac{\hat{\sigma}^2}{\check{\sigma}^2}\right)^{n/2} = \left(1 + \frac{(\overline{x}_n - \mu_0)^2}{\hat{\sigma}^2}\right)^{-n/2}.$$

This statistic is smaller than some critical value if and only if

$$\left(\frac{\overline{x}_n - \mu_0}{\hat{\sigma}/\sqrt{n}}\right)^2 = \left(\frac{n}{n-1}\right)\left(\frac{\overline{x}_n - \mu_0}{\tilde{\sigma}/\sqrt{n}}\right)^2$$

exceeds some other critical value; where, recall, $\tilde{\sigma}^2 = \hat{\sigma}^2 n/(n-1)$.

But,

$$\frac{\overline{x}_n - \mu_0}{\tilde{\sigma}/\sqrt{n}} = \frac{\sigma}{\tilde{\sigma}}\,\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \Bigg/ \sqrt{\frac{\tilde{\sigma}^2}{\sigma^2}} = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \Bigg/ \frac{\sqrt{(n-1)\,(\tilde{\sigma}^2/\sigma^2)}}{\sqrt{n-1}}.$$

We know that

$$\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1), \qquad (n-1)\frac{\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1};$$

and so the ratio follows a t distribution with $n-1$ degrees of freedom.

The statistic

$$\frac{\overline{x}_n - \mu_0}{\tilde{\sigma}/\sqrt{n}} \sim t_{n-1}.$$

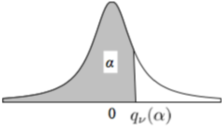
is commonly called the t-statistic.

Exact inference is thus possible on choosing critical values from Student's t distribution with $n-1$ degrees of freedom.

As n grows, t_{n-1} approaches the standard normal. So large-sample theory justifies the use of $z_{\alpha/2}$ as a critical value.

Student's t distribution

The quantile function of the Student's t distribution



							α						
	0.600	0.700	0.750	0.800	0.850	0.900	0.925	0.950	0.975	0.990	0.995	0.999	0.9995
ν													
1	0.325	0.727	1.000	1.376	1.963	3.078	4.165	6.314	12.71	31.82	63.66	318.3	636.6
2	0.289	0.617	0.816	1.061	1.386	1.886	2.282	2.920	4.303	6.965	9.925	22.33	31.60
3	0.277	0.584	0.765	0.978	1.250	1.638	1.924	2.353	3.182	4.541	5.841	10.21	12.92
4	0.271	0.569	0.741	0.941	1.190	1.533	1.778	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.559	0.727	0.920	1.156	1.476	1.699	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.553	0.718	0.906	1.134	1.440	1.650	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.549	0.711	0.896	1.119	1.415	1.617	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.546	0.706	0.889	1.108	1.397	1.592	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.543	0.703	0.883	1.100	1.383	1.574	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.542	0.700	0.879	1.093	1.372	1.559	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.540	0.697	0.876	1.088	1.363	1.548	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.539	0.695	0.873	1.083	1.356	1.538	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.538	0.694	0.870	1.079	1.350	1.530	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.537	0.692	0.868	1.076	1.345	1.523	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.536	0.691	0.866	1.074	1.341	1.517	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.535	0.690	0.865	1.071	1.337	1.512	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.534	0.689	0.863	1.069	1.333	1.508	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.534	0.688	0.862	1.067	1.330	1.504	1.734	2.101	2.552	2.878	3.610	3.922

General composite hypothesis

The more general case has both composite null and alternative, as in

$$H_0 : \theta \in \Theta_0, \qquad H_1 : \theta \in \Theta_1,$$

where Θ_0 and Θ_1 are subsets of the parameter space Θ.

An obvious generalization of the likelihood ratio would be

$$\frac{\sup_{\theta_0 \in \Theta_0} \ell_n(\theta_0)}{\sup_{\theta_1 \in \Theta_1} \ell_n(\theta_1)}.$$

The statistic used above for only the alternative composite is a special case.

Much more common is to work with a likelihood ratio statistic defined as

$$\frac{\sup_{\theta_0 \in \Theta_0} \ell_n(\theta_0)}{\sup_{\theta \in \Theta} \ell_n(\theta)};$$

note that the denominator features the full parameter space. This is often much easier to work with.

Connection to maximum likelihood

By definition

$$\sup_{\theta \in \Theta} \ell_n(\theta) = \ell_n(\hat{\theta}),$$

where $\hat{\theta}$ is the (unconstrained) maximum-likelihood estimator.

Likewise, we can think of

$$\check{\theta} = \arg \max_{\theta \in \Theta_0} \ell_n(\theta)$$

as the constrained maximum-likelihood estimator obtained on enforcing the null.

The likelihood ratio is then simply

$$\frac{\ell_n(\check{\theta})}{\ell_n(\hat{\theta})}.$$

Normal (Composite)

$x_i \sim N(\theta, \sigma^2)$ for known σ^2.

Now test $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$.

Here,

$$\arg\max_{\theta_0 \in \Theta_0} \ell_n(\theta) = \overline{x}_n \{\overline{x}_n \leq 0\}, \qquad \arg\max_{\theta_1 \in \Theta_1} \ell_n(\theta) = \overline{x}_n \{\overline{x}_n > 0\},$$

and, also,

$$\arg\max_{\theta \in \Theta} \ell_n(\theta) = \overline{x}_n.$$

So,

$$\frac{\sup_{\theta_0 \in \Theta_0} \ell_n(\theta_0)}{\sup_{\theta_1 \in \Theta_1} \ell_n(\theta_1)} = e^{-\frac{1}{2}\left(\frac{\overline{x}_n}{\sigma/\sqrt{n}}\right)^2 \operatorname{sign}(\overline{x}_n)} = e^{-\frac{1}{2}\left(\frac{\overline{x}_n}{\sigma/\sqrt{n}}\right)\left|\frac{\overline{x}_n}{\sigma/\sqrt{n}}\right|}$$

for $\operatorname{sign}(x) = \{x > 0\} - \{x \leq 0\}$, while

$$\frac{\sup_{\theta_0 \in \Theta_0} \ell_n(\theta_0)}{\sup_{\theta \in \Theta} \ell_n(\theta)} = e^{-\frac{1}{2}\left(\frac{\overline{x}_n}{\sigma/\sqrt{n}}\right)^2 \{\overline{x}_n > 0\}} = e^{-\frac{1}{2}\left(\frac{\overline{x}_n}{\sigma/\sqrt{n}}\right)^2 \left\{\frac{\overline{x}_n}{\sigma/\sqrt{n}} > 0\right\}}.$$

The latter likelihood ratio is smaller then c when

$$\frac{\overline{x}_n}{\sigma/\sqrt{n}}\left\{\frac{\overline{x}_n}{\sigma/\sqrt{n}} > 0\right\} > c^*$$

for some c^*. Note that only positive c^* make sense, otherwise we will never reject.

For any fixed θ, let

$$\overline{z}_\theta = \frac{\overline{x}_n - \theta}{\sigma/\sqrt{n}}.$$

Then

$$P_\theta\left(\tfrac{\overline{x}_n}{\sigma/\sqrt{n}}\left\{\tfrac{\overline{x}_n}{\sigma/\sqrt{n}} > 0\right\} > c^*\right) = P_\theta\left(\overline{z}_\theta > c^* - \tfrac{\theta}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(c^* - \tfrac{\theta}{\sigma/\sqrt{n}}\right)$$

This function is monotone increasing on $\Theta_0 = (-\infty, 0]$. The size of the test is

$$\sup_{\theta\in\Theta_0} P_\theta\left(\tfrac{\overline{x}_n}{\sigma/\sqrt{n}}\left\{\tfrac{\overline{x}_n}{\sigma/\sqrt{n}} > 0\right\} > c^*\right) = 1 - \Phi(c^*) = \alpha$$

so that size control yields the critical value $c^* = \Phi^{-1}(1-\alpha) = z_\alpha$.

The former likelihood ratio is small when either

$$0 < \frac{\overline{x}_n}{\sigma/\sqrt{n}} \text{ and } c^* < \frac{\overline{x}_n}{\sigma/\sqrt{n}}$$

or when

$$\frac{\overline{x}_n}{\sigma/\sqrt{n}} < 0 \text{ and } c^* < \frac{\overline{x}_n}{\sigma/\sqrt{n}}$$

For any $\theta \in \Theta_0$ the probability of this happening is

$$P_\theta\left(\overline{z}_\theta > c^* - \tfrac{\theta}{\sigma/\sqrt{n}}\,,\, \overline{z}_\theta > -\tfrac{\theta}{\sigma/\sqrt{n}}\right) + P_\theta\left(\overline{z}_\theta > c^* - \tfrac{\theta}{\sigma/\sqrt{n}}\,,\, \overline{z}_\theta < -\tfrac{\theta}{\sigma/\sqrt{n}}\right).$$

For $c^* > 0$ this equals

$$P_\theta\left(\overline{z}_\theta > c^* - \tfrac{\theta}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(c^* - \tfrac{\theta}{\sigma/\sqrt{n}}\right)$$

while for $c^* \leq 0$ this equals

$$P_\theta\left(\overline{z}_\theta > -\tfrac{\theta}{\sigma/\sqrt{n}}\right) + P_\theta\left(c^* - \tfrac{\theta}{\sigma/\sqrt{n}} < \overline{z}_\theta \leq -\tfrac{\theta}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(c^* - \tfrac{\theta}{\sigma/\sqrt{n}}\right).$$

In either case the supremum over Θ_0 is achieved at $\theta = 0$ for which we find that $c^* = z_\alpha$ yields size control. Again, for any reasonable size the critical value is positive.

Likelihood-ratio test

Now consider a general setting where θ is a k-dimensional vector and

$$H_0 : r(\theta) = 0, \qquad H_1 : r(\theta) \neq 0,$$

for a continuously-differentiable m-dimensional function r.

We will denote the $m \times k$ Jacobian matrix by $R(\theta)$.

Exact size control is difficult in general.

A general approach that is asymptotically valid is the decision rule

Reject the null if

$$-2 \log \left(\frac{\ell_n(\check{\theta})}{\ell_n(\hat{\theta})} \right) > \chi^2_{m,\alpha};$$

Accept the null if

$$-2 \log \left(\frac{\ell_n(\check{\theta})}{\ell_n(\hat{\theta})} \right) \leq \chi^2_{m,\alpha}.$$

Note that

$$-2 \log \left(\ell_n(\check{\theta}) / \ell_n(\hat{\theta}) \right) = 2(L_n(\hat{\theta}) - L_n(\check{\theta})).$$

Asymptotic distribution under the null

The validity of the test procedure comes from the following theorem.

Theorem 17 (Limit distribution of the Likelihood-ratio statistic)

Under the null,

$$2(L_n(\hat{\theta}) - L_n(\check{\theta})) \xrightarrow{d} \chi^2_m$$

as $n \to \infty$.

Proof.

We work under the null. A Taylor expansion gives

$$L_n(\check{\theta}) - L_n(\hat{\theta}) = -\frac{n}{2}(\check{\theta} - \hat{\theta})' I_\theta (\check{\theta} - \hat{\theta}) + o_p(1).$$

It can be shown that (under the null)

$$\sqrt{n}(\hat{\theta} - \check{\theta}) = I_\theta^{-1} R' (R I_\theta^{-1} R')^{-1} R I_\theta^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left. \frac{\partial \log f_\theta(x_i)}{\partial \theta} \right|_\theta + o_p(1),$$

where $R = R(\theta)$. Plugging this into the expansion gives $2(L_n(\hat{\theta}) - L_n(\check{\theta}))$ as

$$\left(R I_\theta^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left. \frac{\partial \log f_\theta(x_i)}{\partial \theta} \right|_\theta \right)' (R I_\theta^{-1} R')^{-1} \left(R I_\theta^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left. \frac{\partial \log f_\theta(x_i)}{\partial \theta} \right|_\theta \right)$$

(up to $o_p(1)$ terms). But, as

$$R I_\theta^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left. \frac{\partial \log f_\theta}{\partial \theta} \right|_\theta \xrightarrow{d} N(0, R I_\theta^{-1} R'),$$

this quadratric form is asymptotically χ^2_m. □

Analysis of the constrained estimator

Completing the proof requires finding the asymptotic distribution of

$$\check{\theta} = \arg \max_{\theta : r(\theta) = 0} L_n(\theta).$$

This estimator maximizes the Lagrangian problem

$$L_n(\theta) + \lambda' r(\theta).$$

The first-order conditions are

$$\left. \frac{\partial L_n(\theta)}{\partial \theta} \right|_{\check{\theta}} + \check{\lambda}' \, R(\check{\theta}) = 0, \qquad r(\check{\theta}) = 0.$$

We can Taylor expand

$$\begin{aligned} \left. \frac{\partial L_n(\theta)}{\partial \theta} \right|_{\check{\theta}} &= \left. \frac{\partial L_n(\theta)}{\partial \theta} \right|_{\theta} + \left. \frac{\partial^2 L_n(\theta)}{\partial \theta \partial \theta'} \right|_{\theta} (\check{\theta} - \theta) + o_p(1) \\ &= \sum_{i=1}^{n} \left. \frac{\partial \log f_\theta(x_i)}{\partial \theta} \right|_{\theta} - n I_\theta \, (\check{\theta} - \theta) + o_p(1), \end{aligned}$$

and $r(\check{\theta}) = r(\theta) + R\,(\check{\theta} - \theta) + o_p(1) = R\,(\check{\theta} - \theta) + o_p(1)$ (enforcing the null $r(\theta) = 0$).

Plugging the expansions into the first-order conditions and re-arranging yields the system of equations

$$n^{-1/2}\begin{pmatrix} -nI_\theta & R' \\ R & 0 \end{pmatrix}\begin{pmatrix} \check{\theta}-\theta \\ \check{\lambda} \end{pmatrix} = -n^{-1/2}\begin{pmatrix} \sum_{i=1}^n \left.\frac{\partial \log f_\theta(x_i)}{\partial\theta}\right|_\theta \\ 0 \end{pmatrix}$$

(up to $o_p(1)$ terms).

A block-inversion formula shows that

$$\begin{pmatrix} -nI_\theta & R \\ R' & 0 \end{pmatrix}^{-1}$$

equals

$$\begin{pmatrix} -n^{-1}I_\theta^{-1} + n^{-1}I_\theta^{-1}R'(RI_\theta^{-1}R')^{-1}RI_\theta^{-1} & I_\theta^{-1}R'(RI_\theta^{-1}R')^{-1} \\ (RI_\theta^{-1}R')^{-1}RI_\theta^{-1} & n\,(RI_\theta^{-1}R')^{-1} \end{pmatrix}.$$

Then we obtain

$$\sqrt{n}(\check{\theta}-\theta)=(I_\theta^{-1}-I_\theta^{-1}R'(RI_\theta^{-1}R')^{-1}RI_\theta^{-1})\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left.\frac{\partial\log f_\theta(x_i)}{\partial\theta}\right|_\theta+o_p(1),$$

which implies that

$$\sqrt{n}(\hat{\theta}-\check{\theta})=I_\theta^{-1}R'(RI_\theta^{-1}R')^{-1}RI_\theta^{-1})\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left.\frac{\partial\log f_\theta(x_i)}{\partial\theta}\right|_\theta+o_p(1)$$

under the null.

For future reference we also note that

$$\frac{\check{\lambda}}{\sqrt{n}}=-(RI_\theta^{-1}R')^{-1}RI_\theta^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left.\frac{\partial\log f_\theta(x_i)}{\partial\theta}\right|_\theta+o_p(1)\xrightarrow{d}N(0,(RI_\theta^{-1}R')^{-1}).$$

χ^2-statistic

The derivation of the limit distribution of the likelihood-ratio statistic shows that

$$n\,(\check{\theta}-\hat{\theta})' I_\theta (\check{\theta}-\hat{\theta}) \xrightarrow{d} \chi^2_m$$

under the null.

Let $\check{I}_\theta$ be a consistent estimator of the information under the null. Obvious choices are

$$-\frac{1}{n}\; \frac{\partial^2 L_n(\theta)}{\partial\theta\partial\theta'}\bigg|_{\check{\theta}} = -\frac{1}{n}\sum_{i=1}^n \; \frac{\partial^2 \log f_\theta(x_i)}{\partial\theta\partial\theta'}\bigg|_{\check{\theta}}$$

and

$$\frac{1}{n}\sum_{i=1}^n \left(\frac{\partial \log f_\theta(x_i)}{\partial\theta}\bigg|_{\check{\theta}} \frac{\partial \log f_\theta(x_i)}{\partial\theta}\bigg|'_{\check{\theta}} \right) - \left(\frac{1}{n}\sum_{i=1}^n \frac{\partial \log f_\theta(x_i)}{\partial\theta}\bigg|_{\check{\theta}} \right)\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial \log f_\theta(x_i)}{\partial\theta}\bigg|_{\check{\theta}} \right)'$$

note that recentering of the score is needed here as $\check{\theta}$ does not maximize the unconstrained likelihood problem, in general.

Slutzky's theorem gives us the following result.

Theorem 18 (Limit distribution of the χ^2-statistic)

Under the null,

$$n\,(\check{\theta}-\hat{\theta})'\check{I}_\theta(\check{\theta}-\hat{\theta}) \xrightarrow{d} \chi^2_m$$

as $n \to \infty$*.*

This result gives us an alternative, but asymptotically equivalent, testing procedure.

The intuition behind a test based on this result is to look at a distance between the constrained and the unconstrained estimators which, under the null, should be small.

Score statistic

The analysis of the constrained estimator implies the following result.

Theorem 19 (Limit distribution of the Score statistic)

Under the null,

$$\left.\frac{\partial L_n(\theta)}{\partial \theta}\right|_{\check{\theta}}' \left(\frac{\check{I}_\theta^{-1}}{n}\right) \left.\frac{\partial L_n(\theta)}{\partial \theta}\right|_{\check{\theta}} \xrightarrow{d} \chi_m^2,$$

as $n \to \infty$.

This statistic is also known as the Lagrange-multiplier statistic as it can be written as

$$\check{\lambda}' \left(\frac{R(\check{\theta})\check{I}_\theta^{-1}R(\check{\theta})'}{n}\right)\check{\lambda},$$

where $\check{\lambda}$ is the Lagrangian multiplier for the constraint $r(\theta) = 0$.

One interpretation for this is that, if the null is true, the constraint should be ineffective, aside from sampling error, so $\check{\lambda}$ should be small.

Another interpretation is that, under the null, the unconstrained score should be close to zero at $\check{\theta}$.

Wald statistic

Rather than evaluating some distance between $\check{\theta}$ and $\hat{\theta}$ as in the χ^2-statistic

$$n\,(\check{\theta}-\hat{\theta})'\check{I}_\theta(\check{\theta}-\hat{\theta}),$$

we may look at a distance of $r(\hat{\theta})$ from zero (the null). Because we have that

$$r(\hat{\theta}) = r(\hat{\theta}) - r(\check{\theta}) = R\,(\hat{\theta}-\check{\theta}) + o_p(1)$$

under the null, we equally have the following.

Theorem 20 (Limit distribution of the Wald statistic)

Under the null,

$$n\,r(\hat{\theta})'(R(\check{\theta})\check{I}_\theta^{-1}R(\check{\theta})')^{-1}\,r(\hat{\theta}) \overset{d}{\to} \chi^2_m,$$

as $n \to \infty$.

The Wald statistic can equally be derived without reference to a constrained estimation problem.

Because

$$\sqrt{n}(\hat\theta - \theta) \xrightarrow{d} N(0, I_\theta^{-1}) \text{ and } r(\hat\theta) = R\,(\hat\theta - \theta) + o_p(1),$$

under the null, the Delta method gives

$$\sqrt{n}\,r(\hat\theta) \xrightarrow{d} N(0, RI_\theta^{-1}R'),$$

and so also

Theorem 21 (Limit distribution of the Wald statistic (cont'd))

Under the null,

$$n\,r(\hat\theta)'(R(\hat\theta)\hat I_\theta^{-1}R(\hat\theta)')^{-1}\,r(\hat\theta) \xrightarrow{d} \chi^2_m,$$

as $n \to \infty$.

Here it makes sense to use an unconstrained estimator of the information.

Notes

All test statistics can be used in the same way to perform (asymptotically) valid inference.

In small samples they can lead to different test conclusions.

The likelihood-ratio statistic is attractive because

- It does not require an estimator of I_θ;
- It is invariant with respect to one-to one transformations.

The second point is important as it implies that the test conclusion is the same no matter how the null is formulated.

The score statistic is attractive because it requires estimation only under the null, which is often easier.

In the likelihood context there is no strong argument in favor of the Wald statistic. In fact it is not likelihood based. Its power lies in that it can be applied more generally.

Exponential

The exponential distribution is

$$f_\theta(x) = \frac{e^{-x/\theta}}{\theta}.$$

Its mean is θ.

We set up several tests for the null $H_0 : \theta = \theta_0$ against $\theta \neq \theta_0$.

First note that

$$L_n(\theta) = -\sum_{i=1}^{n} (x_i/\theta + \log\theta) = -n\overline{x}_n/\theta - n\log\theta.$$

Hence,

$$\frac{\partial L_n(\theta)}{\partial\theta} = (n/\theta)(\overline{x}_n/\theta - 1), \qquad \frac{\partial^2 L_n(\theta)}{\partial\theta^2} = -(n/\theta^2)(2\overline{x}_n/\theta - 1).$$

Therefore, $\hat{\theta} = \overline{x}_n$ and $I_\theta^{-1}/n = \theta^2/n$.

The likelihood-ratio statistic is

$$-2(L_n(\theta_0) - L_n(\hat{\theta})) = 2n\left(\left(\frac{\overline{x}_n}{\theta_0} - 1\right) - \log\frac{\overline{x}_n}{\theta_0}\right).$$

The score statistic is

$$n(\overline{x}_n/\theta_0 - 1)^2 = \frac{(\overline{x}_n - \theta_0)^2}{\theta_0^2/n}$$

The χ^2-statistic and the Wald statistic are

$$\frac{(\overline{x}_n - \theta_0)^2}{\theta_0^2/n}, \qquad \frac{(\overline{x}_n - \theta_0)^2}{\overline{x}_n^2/n},$$

respectively. The latter is again the usual t-statistic, which should not be surprising here.

Classical linear regression

Recall the setup

$$y_i | x_i \sim N(x_i'\beta, \sigma^2)$$

or, in matrix notation,

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

The log-likelihood (up to a constant) is

$$L_n(\beta, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)}{2\sigma^2}.$$

Let $SSR_\beta = (\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$. Then the profiled log-likelihood for β is

$$L_n(\beta) = -\frac{n}{2} \log \left(SSR_\beta\right) = \log(SSR_\beta^{-n/2})$$

(again up to a constant), and

$$\ell_n(\beta) \propto SSR_\beta^{-n/2}.$$

We consider a set of m linear restrictions on β. We express the null hypothesis as

$$R\beta = r,$$

where R is an $m \times k$ matrix and r and is an m-vector.

The m restrictions are non-redundant, so rank $R = m$.

The unconstrained estimator solves

$$\min_{\beta} SSR_{\beta} = \min_{\beta} (\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$$

and equals

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y},$$

as before.

The constrained estimator solves the Lagrangian problem

$$\min_{\beta} \frac{1}{2} SSR_{\beta} - \lambda'(R\beta - r).$$

The first-order conditions are

$$\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X}\beta)-R'\lambda=0, \qquad R\beta-r=0.$$

Re-arranging the first condition gives

$$(\boldsymbol{X}'\boldsymbol{X})\beta=\boldsymbol{X}'\boldsymbol{y}-R'\lambda$$

and so

$$\check{\beta}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}-(\boldsymbol{X}'\boldsymbol{X})^{-1}R'\lambda=\hat{\beta}-(\boldsymbol{X}'\boldsymbol{X})^{-1}R'\lambda.$$

Further, pre-multiplying by R and enforcing that $R\check{\beta}=r$ gives

$$R\check{\beta}=R\hat{\beta}-R(\boldsymbol{X}'\boldsymbol{X})^{-1}R'\lambda=r,$$

which we solve for λ to obtain

$$\check{\lambda}=(R(\boldsymbol{X}'\boldsymbol{X})^{-1}R')^{-1}(R\hat{\beta}-r).$$

We then find that

$$\check{\beta}=\hat{\beta}-(\boldsymbol{X}'\boldsymbol{X})^{-1}R'(R(\boldsymbol{X}'\boldsymbol{X})^{-1}R')^{-1}(R\hat{\beta}-r).$$

The likelihood ratio statistic is

$$\left(\frac{SSR_{\breve{\beta}}}{SSR_{\hat{\beta}}}\right)^{-n/2},$$

which is small when the ratio in brackets is large. Now,

$$\frac{SSR_{\breve{\beta}}}{SSR_{\hat{\beta}}} - 1 = \frac{SSR_{\breve{\beta}} - SSR_{\hat{\beta}}}{SSR_{\hat{\beta}}}$$

where

$$SSR_{\hat{\beta}} = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{\varepsilon}$$

and, using that $\boldsymbol{y} = \boldsymbol{X}\hat{\beta} + \hat{\boldsymbol{\varepsilon}}$ to simplify $SSR_{\breve{\beta}} = (\boldsymbol{y} - \boldsymbol{X}\breve{\beta})'(\boldsymbol{y} - \boldsymbol{X}\breve{\beta})$ to

$$SSR_{\breve{\beta}} = \boldsymbol{\varepsilon}'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{\varepsilon} + (\hat{\beta} - \breve{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\beta} - \breve{\beta}).$$

Hence,

$$\begin{aligned}\frac{SSR_{\breve{\beta}} - SSR_{\hat{\beta}}}{SSR_{\hat{\beta}}} &= \frac{(\hat{\beta} - \breve{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\beta} - \breve{\beta})}{\boldsymbol{\varepsilon}'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{\varepsilon}} \\ &= \frac{(R\hat{\beta} - r)'(R(\boldsymbol{X}'\boldsymbol{X})^{-1}R')^{-1}(R\hat{\beta} - r)}{\boldsymbol{\varepsilon}'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{\varepsilon}}.\end{aligned}$$

Note that, under the null,

$$R\hat{\beta} - r \sim N(0, \sigma^2 R(\boldsymbol{X}'\boldsymbol{X})^{-1}R')$$

such that

$$\frac{SSR_{\breve{\beta}} - SSR_{\hat{\beta}}}{\sigma^2} \sim \chi^2_m.$$

We also know that

$$\frac{SSR_{\hat{\beta}}}{\sigma^2} = (n-k)\frac{\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k}.$$

Lastly, both terms are independent because they are functions of $\hat{\beta}$ and $\hat{\boldsymbol{\varepsilon}}$, respectively. These variables are jointly normal and independent, as the covariance is

$$E((\hat{\beta} - \beta)\hat{\boldsymbol{\varepsilon}}'|\boldsymbol{X}) = E\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{M}_{\boldsymbol{X}}|\boldsymbol{X}\right) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{X}} = 0$$

(using that $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{X} = 0$)

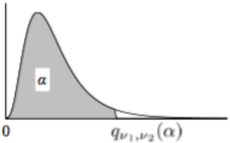
Therefore,

$$\frac{n-k}{m}\,\frac{SSR_{\breve{\beta}} - SSR_{\hat{\beta}}}{SSR_{\hat{\beta}}} \sim F_{m,n-k},$$

where F is Snedecor's F distribution.

Snedecor's F distribution

The quantile function of the F distribution



$\alpha = 0.9$

								ν_1							
	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	∞
ν_2																
1	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.7	61.2	61.7	62.3	62.7	63.3
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9. 41	9.42	9.44	9.46	9.47	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5. 22	5.20	5.18	5.17	5.15	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3. 90	3.87	3.84	3.82	3.80	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3. 27	3.24	3.21	3.17	3.15	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2. 90	2.87	2.84	2.80	2.77	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2. 67	2.63	2.59	2.56	2.52	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2. 50	2.46	2.42	2.38	2.35	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2. 38	2.34	2.30	2.25	2.22	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2 .28	2.24	2.20	2.16	2.12	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2 .21	2.17	2.12	2.08	2.04	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2 .15	2.10	2.06	2.01	1.97	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2 .10	2.05	2.01	1.96	1.92	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2 .05	2.01	1.96	1.91	1.87	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2 .02	1.97	1.92	1.87	1.83	1.76

A particular F test

A popular restriction in a regression model that includes a constant term is that all slopes are zero.

Under the null we only estimate a constant term, i.e., $\check{\beta} = \overline{y}_n$, and so we have

$$SSR_{\check{\beta}} = \sum_{i=1}^{n} (y_i - \overline{y}_n)^2 = TSS.$$

It follows that the F-statistic can be written as

$$\frac{n-k}{m} \frac{SSR_{\check{\beta}} - SSR_{\hat{\beta}}}{SSR_{\hat{\beta}}} = \frac{n-k}{m} \frac{TSS - SSR_{\hat{\beta}}}{SSR_{\hat{\beta}}} = \frac{n-k}{m} \frac{1-R^2}{R^2}$$

with $R^2 = SSR_{\hat{\beta}}/TSS$ the (centered) coefficient of determination of the unrestricted model.

F versus t

When we test only the restriction $\beta_\kappa = \beta_{\kappa_0}$ the F-statistic is

$$\frac{(\hat{\beta}_\kappa - \beta_{\kappa,0})([(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{\kappa,\kappa})^{-1}(\hat{\beta}_\kappa - \beta_{\kappa,0})}{\tilde{\sigma}^2} = \left(\frac{\hat{\beta}_\kappa - \beta_{\kappa,0}}{\sqrt{\tilde{\sigma}^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{\kappa,\kappa}}}\right)^2.$$

This is the square of the usual t-statistic

$$\frac{\hat{\beta}_\kappa - \beta_{\kappa,0}}{\sqrt{\tilde{\sigma}^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{\kappa,\kappa}}}.$$

So the square of a t_{n-k} random variable is $F_{1,n-k}$ distributed.

Joint hypothesis versus multiple single hypotheses

To jointly test the k restrictions $\beta = \beta_0$ we use the F-statistic

$$\frac{1}{k}\,\frac{(\hat{\beta}-\beta_0)'(\boldsymbol{X}'\boldsymbol{X})(\hat{\beta}-\beta_0)}{\tilde{\sigma}^2}.$$

This is not the mean of the t-statistics for the k individual hypotheses that $\beta_\kappa = \beta_{\kappa,0}$. The individual t-statistics are correlated.

Jointly testing hypothesis gives acceptance regions that are ellipsoids. The union of acceptance regions of multiple individual tests is a hypercube.

Multiple testing problems need size corrections which, in turn, lead to low power.

The family-wise error rate is

$$P\left(\cup_\kappa\left\{\left|\frac{\hat{\beta}_\kappa-\beta_{\kappa,0}}{\sqrt{\tilde{\sigma}^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{\kappa,\kappa}}}\right|>t_{n-k,\alpha/2}\right\}\right)\leq\sum_\kappa P\left(\left|\frac{\hat{\beta}_\kappa-\beta_{\kappa,0}}{\sqrt{\tilde{\sigma}^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{\kappa,\kappa}}}\right|>t_{n-k,\alpha/2}\right)=k\times\alpha.$$

To keep the family-wise error rate below α we need to test each of k individual hypothesis at significance level α/k.

p-values

If we follow the Neyman-Pearson decision rule we either accept or reject the null.

We may also look at the p-value of a test statistic.

Consider a test procedure where we reject the null when the statistic ψ_n is large.

If the statistic ψ_n takes on value ψ in the data the p-value is

$$\sup_{\theta \in \Theta_0} P_\theta \left(\psi_n > \psi \right).$$

This is the probability of observing a value of the test statistic greater than ψ if the null holds.

Small p-values suggest the null is likely to be false.

The p-value gives a cut-off of significance levels for which a Neyman-Pearson decision rule would accept/reject...

But the p-value is informative in its own right and need not lead to a decision about the null. This is Fisher's view.

Inverting test statistics

As an alternative to a point estimator, testing procedures can give rise to interval estimators.

Suppose we test $H_0 : \theta = \theta_0$ using a decision rule of the form

$$\text{Accept } H_0 \text{ if } \psi_n(\theta_0) \leq c$$

for some critical value c.

Then the set

$$\hat{\Theta} = \{\theta \in \Theta : \psi_n(\theta) \leq c\}$$

constitutes an interval estimator.

If the original test has size α then

$$P_{\theta_0}(\theta_0 \in \hat{\Theta}) = 1 - \alpha.$$

The interval estimator is also called a $(1 - \alpha)$ confidence set.

Normal

Suppose $x_i \sim N(\theta, \sigma^2)$.

Consider $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$.

The likelihood ratio decision rule goes in favor of the null if

$$\left| \frac{\overline{x}_n - \theta_0}{\tilde{\sigma}/\sqrt{n}} \right| \leq t_{n-1,\alpha/2}.$$

This means that, for any θ in the interval

$$\hat{\Theta} = \left[\overline{x}_n - \frac{\tilde{\sigma}}{\sqrt{n}} t_{n-1,\alpha/2} \, , \, \overline{x}_n + \frac{\tilde{\sigma}}{\sqrt{n}} t_{n-1,\alpha/2} \right]$$

the null would be accepted.

$\hat{\Theta}$ is an interval estimator of θ_0.

We have

$$P_{\theta_0} \left(\theta_0 \in \hat{\Theta} \right) = 1 - \alpha;$$

and so $\hat{\Theta}$ is a $(1 - \alpha)$ confidence interval for θ_0.

Now,

$$x_i \sim N(\mu, \theta)$$

and, say,

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta > \theta_0.$$

Under the null,

$$\frac{\sum_{i=1}^n (x_i - \overline{x}_n)^2}{\theta_0} \sim \chi^2_{n-1},$$

and we would accept the null if the sample variance

$$\tilde{\theta} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2$$

satisfies $\tilde{\theta} \leq (n-1)\theta_0 \chi^2_{n-1,\alpha}$

The corresponding interval estimator thus is

$$\Big[(n-1)\,\tilde{\theta}/\chi^2_{n-1,\alpha}, +\infty\Big)$$

and has coverage probability $1 - \alpha$.

Bayesian credible sets

Given a Bayesian posterior $\pi(\theta|x_1, \ldots, x_n)$ and a region $\mathcal{R}$ of its support, we may calculate

$$P(\theta \in \mathcal{R}|x_1, \ldots, x_n) = \int \{\theta \in \mathcal{R}\}\, \pi(\theta|x_1, \ldots, x_n)\, d\theta.$$

This is a credible probability for the credible set $\mathcal{R}$.

Credible regions can be formed in many ways.

For scalar θ we could, for example, take the interval $[q_{\alpha/2}, q_{1-\alpha/2}]$, where q_τ is the τ quantile of the posterior distribution.

Return to the example where $x_i \sim N(\theta, \sigma^2)$ (with σ^2 known) and we have prior information $\theta \sim N(\mu, \tau^2)$.

Here, the posterior was

$$N\left(m, v^2\right)$$

for mean and variance

$$m = \frac{\tau^2}{\tau^2 + \sigma^2/n}\overline{x}_n + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\mu, \qquad v^2 = \frac{\tau^2\,\sigma^2/n}{\tau^2 + \sigma^2/n}.$$

So,

$$\frac{\theta - m}{v} \sim N(0, 1)$$

and a $1 - \alpha$ credible interval is

$$[m - z_{\alpha/2}v\,;\, m - z_{\alpha/2}v].$$

We can compute the Frequentist coverage probability of this credible set.

The Frequentist framework has $\overline{x}_n \sim N(\theta, \sigma^2/n)$ (here θ is fixed).

The posterior depends on the data only through its mean,

$$m = \frac{1}{1+\delta}\overline{x}_n + \frac{\delta}{1+\delta}\mu, \qquad \delta = \frac{\sigma^2/n}{\tau^2}.$$

A calculation shows that

$$P_\theta\left(m - z_{\alpha/2}v \le \theta \le m + z_{\alpha/2}v\right)$$

equals

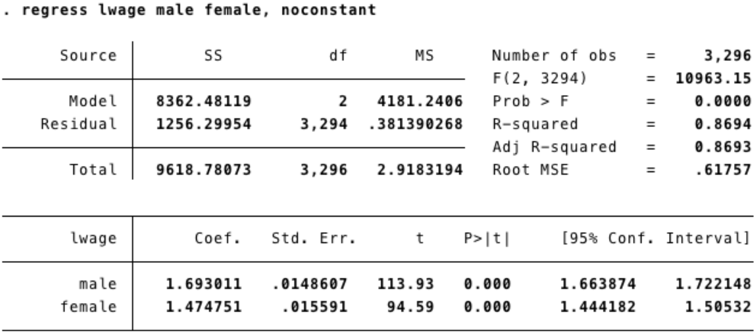
$$\Phi\left(\sqrt{1+\delta}\, z_{\alpha/2} + \delta\frac{\theta-\mu}{\sigma/\sqrt{n}}\right) - \Phi\left(-\sqrt{1+\delta}\, z_{\alpha/2} + \delta\frac{\theta-\mu}{\sigma/\sqrt{n}}\right)$$

which is different from $\Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1 - \alpha$.

Stratifying regressions

log wages are (approximately) normal.

Suppose different means but common variance for males and females.



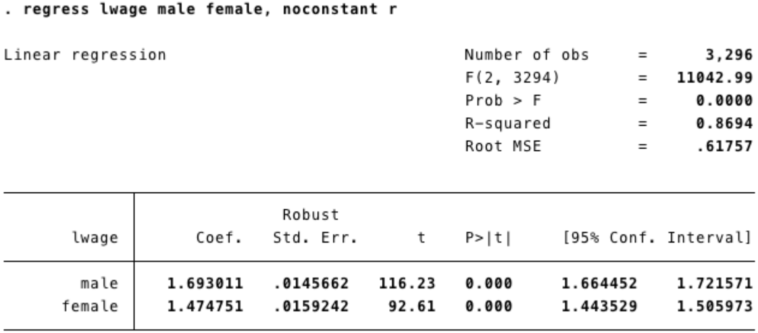
```
. regress lwage male female, noconstant

      Source |       SS           df       MS      Number of obs   =     3,296
-------------+----------------------------------   F(2, 3294)      =  10963.15
       Model |  8362.48119         2  4181.2406    Prob > F        =    0.0000
    Residual |  1256.29954     3,294  .381390268   R-squared       =    0.8694
-------------+----------------------------------   Adj R-squared   =    0.8693
       Total |  9618.78073     3,296  2.9183194    Root MSE        =    .61757

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   1.693011   .0148607   113.93   0.000     1.663874    1.722148
      female |   1.474751    .015591    94.59   0.000     1.444182     1.50532
------------------------------------------------------------------------------
```

Common variance is unrealistic and can be relaxed.

(This will lead us to semiparametric problems; considered below.)

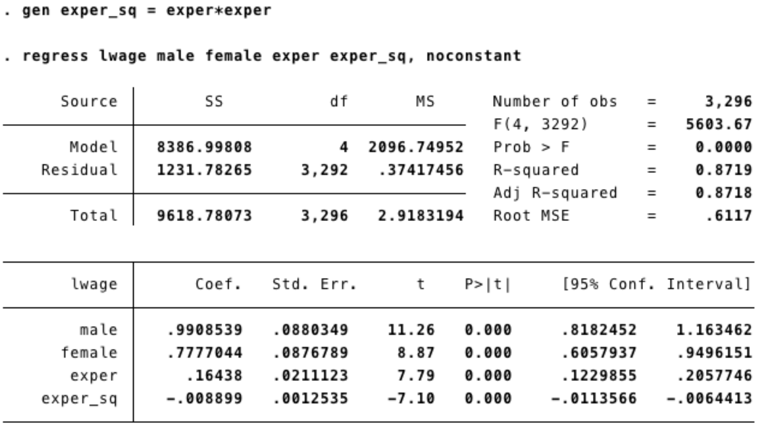


```
. regress lwage male female, noconstant r

Linear regression                               Number of obs     =      3,296
                                                F(2, 3294)        =   11042.99
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8694
                                                Root MSE          =     .61757

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   1.693011   .0145662   116.23   0.000     1.664452    1.721571
      female |   1.474751   .0159242    92.61   0.000     1.443529    1.505973
------------------------------------------------------------------------------
```

Add homogenous impact of experience.



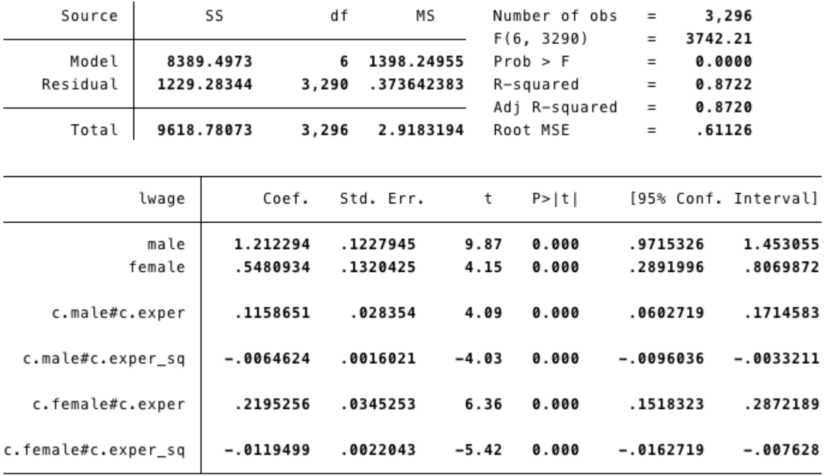
```
. gen exper_sq = exper*exper

. regress lwage male female exper exper_sq, noconstant

      Source |       SS           df       MS      Number of obs   =     3,296
-------------+----------------------------------   F(4, 3292)      =   5603.67
       Model |  8386.99808         4  2096.74952   Prob > F        =    0.0000
    Residual |  1231.78265     3,292   .37417456   R-squared       =    0.8719
-------------+----------------------------------   Adj R-squared   =    0.8718
       Total |  9618.78073     3,296   2.9183194   Root MSE        =     .6117

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .9908539   .0880349    11.26   0.000     .8182452    1.163462
      female |   .7777044   .0876789     8.87   0.000     .6057937    .9496151
       exper |     .16438   .0211123     7.79   0.000     .1229855    .2057746
    exper_sq |   -.008899   .0012535    -7.10   0.000    -.0113566   -.0064413
------------------------------------------------------------------------------
```

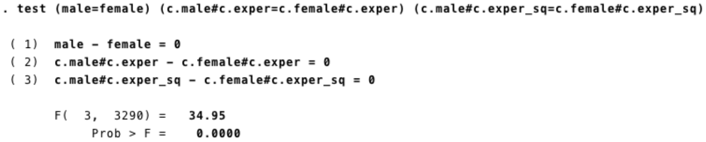
Stratify impact of experience by gender.



```
      Source |       SS           df       MS      Number of obs   =     3,296
-------------+----------------------------------   F(6, 3290)      =   3742.21
       Model |  8389.4973         6  1398.24955   Prob > F        =    0.0000
    Residual | 1229.28344     3,290  .373642383   R-squared       =    0.8722
-------------+----------------------------------   Adj R-squared   =    0.8720
       Total | 9618.78073     3,296   2.9183194   Root MSE        =    .61126

------------------------------------------------------------------------------------
             lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
              male |   1.212294   .1227945     9.87   0.000     .9715326    1.453055
            female |   .5480934   .1320425     4.15   0.000     .2891996    .8069872
                   |
      c.male#c.exper |   .1158651    .028354     4.09   0.000     .0602719    .1714583
                   |
   c.male#c.exper_sq |  -.0064624   .0016021    -4.03   0.000    -.0096036   -.0033211
                   |
    c.female#c.exper |   .2195256   .0345253     6.36   0.000     .1518323    .2872189
                   |
 c.female#c.exper_sq |  -.0119499   .0022043    -5.42   0.000    -.0162719    -.007628
------------------------------------------------------------------------------------
```

Test the equality of the regression lines.

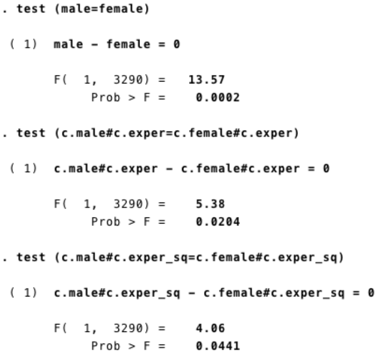


```
. test (male=female) (c.male#c.exper=c.female#c.exper) (c.male#c.exper_sq=c.female#c.exper_sq)

 ( 1)  male - female = 0
 ( 2)  c.male#c.exper - c.female#c.exper = 0
 ( 3)  c.male#c.exper_sq - c.female#c.exper_sq = 0

       F(  3,  3290) =   34.95
            Prob > F =    0.0000
```

Test the equality of the intercept and slopes separately.



```
. test (male=female)

 ( 1)  male - female = 0

       F(  1,  3290) =   13.57
            Prob > F =    0.0002

. test (c.male#c.exper=c.female#c.exper)

 ( 1)  c.male#c.exper - c.female#c.exper = 0

       F(  1,  3290) =    5.38
            Prob > F =    0.0204

. test (c.male#c.exper_sq=c.female#c.exper_sq)

 ( 1)  c.male#c.exper_sq - c.female#c.exper_sq = 0

       F(  1,  3290) =    4.06
            Prob > F =    0.0441
```

SEMIPARAMETRIC PROBLEMS: (GENERALIZED) METHOD OF MOMENTS

Reading

Asymptotic theory:

- Arellano, Appendix A
- Hansen II, Chapter 13
- Hayashi, Chapter 7
- Wooldridge, Chapter 12

Linear instrumental variables:

- Hansen II, Chapter 12
- Hayashi, Chapter 3
- Wooldridge, Chapters 5 and 8

Optimality in conditional moment problems:

- Arellano, Appendix B

Linear model

Recall,

$$y_i = x_i'\theta + \varepsilon_i.$$

Before we had imposed $\varepsilon_i | x_i \sim N(0, \sigma^2)$. but suppose that we only require that

$$E(\varepsilon_i | x_i) = 0.$$

We no longer assume that $y_i | x_i \sim N(x_i'\theta, \sigma^2)$ and so we cannot write down the likelihood.

For example, $\text{var}(\varepsilon_i | x_i)$ is unknown and may depend on x_i.

All the information we have is contained in conditional moment condition

$$E(\varepsilon_i | x_i) = E_\theta(y_i - x_i'\theta | x_i) = 0.$$

This is a semiparametric problem:

The model has a parametric part, the conditional mean, and a nonparametric part, the distribution of $\varepsilon_i | x_i$.

Iterating expectations shows that

$$E_\theta(x_i(y_i - x_i'\theta)) = 0$$

and the analogy principle suggest estimating θ by the solving the empirical moment

$$n^{-1}\sum_{i=1}^{n} x_i\,(y_i - x_i'\theta) = 0.$$

This gives the ordinary least-squares estimator,

$$\left(n^{-1}\sum_{i=1}^{n} x_i\,x_i'\right)^{-1}\left(n^{-1}\sum_{i=1}^{n} x_i\,y_i\right),$$

as unique solution provided $\sum_i x_i x_i'$ has maximal rank.

So, for learning θ here, normality of the errors (and knowledge thereof) is not needed.

The errors can be heteroskedastic, skewed, and so on.

But is ordinary least squares still the best estimator of θ?

Aside from

$$E_\theta(x_i(y_i - x_i'\theta)) = 0$$

we equally have that

$$\begin{aligned} E_\theta((x_i \otimes x_i)(y_i - x_i'\theta)) &= 0, \\ E_\theta((x_i \otimes x_i \otimes x_i)(y_i - x_i'\theta)) &= 0, \\ &\vdots \\ E_\theta((x_i \otimes x_i \otimes \cdots \otimes x_i)(y_i - x_i'\theta)) &= 0, \end{aligned}$$

and, indeed, that

$$E_\theta(\psi(x_i)(y_i - x_i'\theta)) = 0$$

for any vector function ψ.

How do we optimally exploit all this information?

Semiparametric efficiency

In a semiparametric model the distribution of the data is no longer known up to a small number of parameters.

The model has parametric part (θ); and a nonparametric part (say F).

Often (i.e., in these slides), the primary interest lies in the parametric part, θ and all available information on θ is formulated in terms of (conditional) moment conditions.

A general approach to estimation is GMM.

Can be devised to hit the semiparametric efficiency bound.

Intuitively, this bound is

$$\sup_F I_{\theta,F}^{-1};$$

that is, the largest of the Cramér-Rao bounds in the parametric submodels contained in our semiparametric setting.

In the linear regression model from above this would be the Cramér-Rao bound under the least-favorable distribution for $\varepsilon_i|x_i$ that satisfies mean independence.

Method of moments

Suppose all we know is that

$$E_\theta(\varphi(x_i;\theta)) = 0$$

for some known function φ.

A unique solution will generally not exist when $\dim\varphi < \dim\theta$. We say θ is underidentified.

Suppose, for now, that $\dim\varphi = \dim\theta$. We call this the just-identified case.

A method of moment estimator is a solution to

$$n^{-1}\sum_{i=1}^{n}\varphi(x_i;\theta) = 0.$$

The intuition is the analogy principle and similar to the argmax argument.

Identification

The argmax result requires that

$$E_\theta(\varphi(x_i;\theta_*)) \neq 0$$

for any $\theta_* \neq \theta$.

This is global identification.

In contrast, local identification means there is a neighborhood around θ in which it is the unique solution.

A sufficient condition for this is that the Jacobian matrix

$$E_\theta\left(\frac{\partial\varphi(x_i;\theta)}{\partial\theta'}\right)$$

is full rank.

When φ is linear in θ local and global identification are the same.

Limit distribution

Let $\hat{\theta}$ satisfy

$$n^{-1}\sum_{i=1}^{n}\varphi(x_i;\hat{\theta}) = 0.$$

We can use a similar argument as used for maximum likelihood to derive its behavior as $n \to \infty$.

Under smoothness conditions an expansion gives

$$n^{-1}\sum_{i=1}^{n}\varphi(x_i;\hat{\theta}) = n^{-1}\sum_{i=1}^{n}\varphi(x_i;\theta) + n^{-1}\sum_{i=1}^{n}\left.\frac{\partial\varphi(x_i;\theta)}{\partial\theta'}\right|_{\theta_*}(\hat{\theta}-\theta).$$

Re-arrangement gives

$$\sqrt{n}(\hat{\theta}-\theta) = \left(-\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial\varphi(x_i;\theta)}{\partial\theta'}\right|_{\theta_*}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi(x_i;\theta).$$

Under a dominance condition we have

$$-\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial\varphi(x_i;\theta)}{\partial\theta'}\right|_{\theta_*}\xrightarrow{p}-E_\theta\left(\frac{\partial\varphi(x_i;\theta)}{\partial\theta'}\right)=-\Gamma_\theta\ \text{(say)}.$$

Also, $\varphi(x_i;\theta)$ is i.i.d. with zero mean. So we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi(x_i;\theta)\xrightarrow{d}N(0,\Omega_\theta)$$

provided that the asymptotic variance

$$\Omega_\theta=\operatorname{var}_\theta(\varphi(x_i;\theta))=E_\theta(\varphi(x_i;\theta)\varphi(x_i;\theta)')$$

exists.

Combined with Slutzky's theorem we get the following result.

Theorem 22 (Limit distribution of MM estimator)

Under regularity conditions,

$$\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{d}N(0,\Gamma_\theta^{-1}\Omega_\theta\Gamma_\theta^{-\prime}),$$

as $n\to\infty$.

Linear model

Our model is

$$y_i = x_i'\theta + \varepsilon_i, \qquad E_\theta(x_i \varepsilon_i) = 0.$$

Here, $\varphi(x_i;\theta) = x_i(y_i - x_i'\theta)$, which gives the least-squares estimator.

Further,

$$\Omega_\theta = E(\varepsilon_i^2\, x_i x_i'), \qquad \Gamma_\theta = -E(x_i x_i').$$

The asymptotic variance is

$$E(x_i x_i')^{-1} E(\varepsilon_i^2\, x_i x_i')\, E(x_i x_i')^{-1}.$$

The variance would simplify if we additionally have that $\text{var}(\varepsilon_i | x_i) = \sigma^2$.

This is an assumption of homoskedasticity.

Then (by iterating expectations)

$$E(\varepsilon_i^2\, x_i x_i') = \sigma^2\, E(x_i x_i'),$$

so that the asymptotic variance would be

$$\sigma^2 E(x_i x_i')^{-1}.$$

We estimate the asymptotic variance as

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 x_i x_i'\right) \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1},$$

where $\hat{\varepsilon}_i = y_i - x_i'\hat{\theta}$ are the residuals from the least-squares regression.

Under homoskedasticity we can use

$$\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2\right) \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}$$

(could also apply the usual degrees-of-freedom correction to the first term).

Note that least squares is no longer normally distributed for small n because the errors need no longer be normal.

Consequently, the exact distribution of usual t and F statistics is unknown.

Exponential regression

Nonlinear conditional-mean models can be handled in the same way.

For example,

$$E_\theta(y_i|x_i) = e^{x_i'\theta} = \mu_i \text{ (say)}$$

implies the moment condition

$$E_\theta(x_i(y_i - \mu_i)) = E_\theta(x_i\varepsilon_i) = E_\theta(x_i(y_i - e^{x_i'\theta})) = 0$$

(among others) and so the estimator that sets

$$n^{-1}\sum_{i=1}^n x_i(y_i - e^{x_i'\theta}) = 0.$$

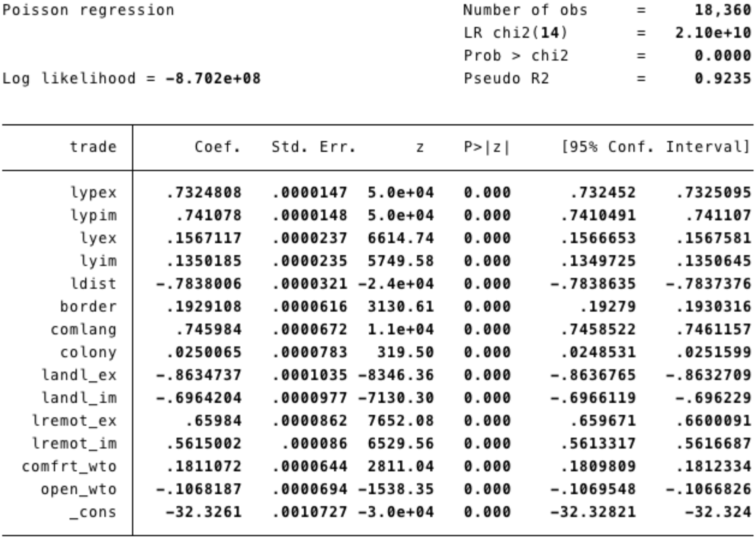
This equals the score equation for Poisson (see Slides 122–123).

Sometimes called the pseudo Poisson estimator.

However, the maximum-likelihood standard errors do not apply because the information equality does not hold here:

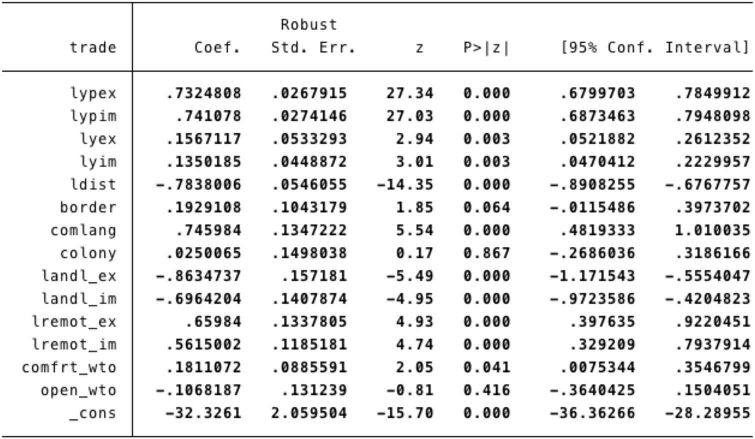
$$\Omega_\theta = E(x_i x_i' \varepsilon_i^2) \neq E(x_i x_i' \mu_i) = -\Gamma_\theta.$$

Pseudo Poisson : gravity equation



```
Poisson regression                              Number of obs     =     18,360
                                                LR chi2(14)       =   2.10e+10
                                                Prob > chi2       =     0.0000
Log likelihood = -8.702e+08                     Pseudo R2         =     0.9235

------------------------------------------------------------------------------
       trade |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lypex |   .7324808   .0000147  5.0e+04   0.000      .732452    .7325095
       lypim |    .741078   .0000148  5.0e+04   0.000     .7410491     .741107
        lyex |   .1567117   .0000237  6614.74   0.000     .1566653    .1567581
        lyim |   .1350185   .0000235  5749.58   0.000     .1349725    .1350645
       ldist |  -.7838006   .0000321 -2.4e+04   0.000    -.7838635   -.7837376
      border |   .1929108   .0000616  3130.61   0.000       .19279    .1930316
     comlang |    .745984   .0000672  1.1e+04   0.000     .7458522    .7461157
      colony |   .0250065   .0000783   319.50   0.000     .0248531    .0251599
    landl_ex |  -.8634737   .0001035 -8346.36   0.000    -.8636765   -.8632709
    landl_im |  -.6964204   .0000977 -7130.30   0.000    -.6966119    -.696229
   lremot_ex |     .65984   .0000862  7652.08   0.000      .659671    .6600091
   lremot_im |   .5615002    .000086  6529.56   0.000     .5613317    .5616687
  comfrt_wto |   .1811072   .0000644  2811.04   0.000     .1809809    .1812334
    open_wto |  -.1068187   .0000694 -1538.35   0.000    -.1069548   -.1066826
       _cons |   -32.3261   .0010727 -3.0e+04   0.000    -32.32821     -32.324
------------------------------------------------------------------------------
```



```
------------------------------------------------------------------------------
             |               Robust
       trade |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lypex |   .7324808   .0267915    27.34   0.000     .6799703    .7849912
       lypim |    .741078   .0274146    27.03   0.000     .6873463    .7948098
        lyex |   .1567117   .0533293     2.94   0.003     .0521882    .2612352
        lyim |   .1350185   .0448872     3.01   0.003     .0470412    .2229957
       ldist |  -.7838006   .0546055   -14.35   0.000    -.8908255   -.6767757
      border |   .1929108   .1043179     1.85   0.064    -.0115486    .3973702
     comlang |    .745984   .1347222     5.54   0.000     .4819333    1.010035
      colony |   .0250065   .1498038     0.17   0.867    -.2686036    .3186166
    landl_ex |  -.8634737    .157181    -5.49   0.000    -1.171543   -.5554047
    landl_im |  -.6964204   .1407874    -4.95   0.000    -.9723586   -.4204823
   lremot_ex |     .65984   .1337805     4.93   0.000      .397635    .9220451
   lremot_im |   .5615002   .1185181     4.74   0.000      .329209    .7937914
  comfrt_wto |   .1811072   .0885591     2.05   0.041     .0075344    .3546799
    open_wto |  -.1068187    .131239    -0.81   0.416    -.3640425    .1504051
       _cons |   -32.3261   2.059504   -15.70   0.000    -36.36266   -28.28955
------------------------------------------------------------------------------
```

Maximum likelihood

The maximum-likelihood estimator is a method-of-moment estimator.

The moment condition is

$$E_\theta \left(\frac{\partial \log f_\theta(x_i)}{\partial \theta} \right) = 0$$

and is always just identified.

Here,

$$\Omega_\theta = -\Gamma_\theta$$

holds by the information equality.

When the distribution of the data is misspecified (so the sample is not drawn from f_θ) the score equation is biased and maximum likelihood inconsistent, in general.

This makes semiparametric alternatives attractive.

Extremum estimators

An Extremum (or M-) estimator is generic terminology for estimators that maximize an objective function, i.e,

$$\arg\max_{\theta} Q_n(\theta),$$

where $Q_n(\theta) = \sum_i q(x_i;\theta)$ need not be a likelihood function.

(Nonlinear) least-squares, for example, has

$$Q_n(\theta) = -\sum_{i=1}^{n} (y_i - \varphi(x_i;\theta))^2,$$

where $E_\theta(y_i|x_i) = \varphi(x_i;\theta)$ (e.g., probit, logit, poisson, etc.).

If Q_n is differentiable, the extremum estimator is a GMM estimator, with moment conditions

$$E_\theta\left(\frac{\partial q(x_i;\theta)}{\partial\theta}\right) = 0.$$

Rank estimator

An example of an M-estimator that is not a GMM estimator is the maximizer of

$$\sum_{i=1}^{n} \sum_{i<j} y_i \left\{x_i'\theta > x_j'\theta\right\} + y_j \left\{x_i'\theta < x_j'\theta\right\}.$$

The objective function is a U-process of order two.

The intuition is that, if

$$E(y_i|x_i) = G(x_i'\theta)$$

is monotonic, then

$$\begin{array}{c} E(y_i|x_i) > E(y_j|x_j) \Rightarrow x_i'\theta > x_j'\theta \\ E(y_i|x_i) < E(y_j|x_j) \Rightarrow x_i'\theta < x_j'\theta \end{array}.$$

However, this objective function is not differentiable in θ.

In fact, the summands in the objective function are not independent. We need a different argument to establish the limit behavior of this estimator.

Quantile regression

Another example of an M-estimator that has a non-smooth objective function is linear quantile regression.

Take an unconditional setting where x_i has continuous (strictly increasing, for simplicity) distribution F. Let

$$\varrho = \operatorname{med}(x_i) = F^{-1}(1/2).$$

We have

$$\varrho = \arg\min_{\rho} E(|x_i - \rho|).$$

Indeed,

$$E(|x_i - \rho|) = \textstyle\int |x - \rho| \, dF(x) = \int_{-\infty}^{\rho} (\rho - x) \, dF(x) + \int_{\rho}^{+\infty} (x - \rho) \, dF(x).$$

Using Leibniz's rule,

$$\frac{\partial E(|x_i - \rho|)}{\partial \rho} = \textstyle\int_{-\infty}^{\rho} dF(x) - \int_{\rho}^{+\infty} dF(x) = F(\rho) - (1 - F(\rho)) = 0$$

has unique solution $\rho = \varrho$.

The sample analog is $n^{-1} \sum_{i=1}^{n} |x_i - \rho|$ and is not differentiable.

An alternative representation of the median follows from

$$F(\varrho) = \frac{1}{2},$$

as

$$E\left(\{x \leq \varrho\} - \frac{1}{2}\right) = 0,$$

which is a moment condition.

This suggest as estimator an (approximate) solution to the empirical moment

$$n^{-1} \sum_{i=1}^{n} \{x_i \leq \rho\} - \frac{1}{2} = 0.$$

The solution, say $\hat{\varrho}$, has 'nice' asymptotic properties,

$$\hat{\varrho} - \varrho \overset{a}{\sim} N\left(0, \frac{1}{n} \frac{1/4}{f(\varrho)^2}\right),$$

but showing this requires different machinery than the one discussed here.

Prediction

You wish to predict y_i based on x_i.

The best predictor depends on how you quantify errors, i.e., the loss function.

If $p(x_i)$ is the predictor,

$$E((y_i - p(x_i))^2)$$

is the expected squared loss.

Under this loss specification the best predictor p minimizes

$$\begin{aligned} E((y_i - p(x_i))^2) &= E(((y_i - E(y_i|x_i)) - (p(x_i) - E(y_i|x_i)))^2) \\ &= E((y_i - E(y_i|x_i))^2) + E((p(x_i) - E(y_i|x_i))^2) \\ &= E(\mathrm{var}(y_i|x_i)) + E((p(x_i) - E(y_i|x_i))^2) \\ &\geq E(\mathrm{var}(y_i|x_i)). \end{aligned}$$

The unique solution is $p(x_i) = E(y_i|x_i)$.

Linear prediction

A linear predictor is a linear function of x_i, i.e., $x_i'\beta$ for any vector β.

The best linear predictor under expected squared loss uses the coefficients

$$\arg\min_{\beta} E((y_i - x_i'\beta)^2).$$

They solve

$$E(x_i\,(y_i - x_i'\beta)) = 0.$$

(uniquely if $E(x_i x_i')$ has full rank) and equal

$$\beta = E(x_i x_i')^{-1} E(x_i y_i).$$

This is the population ordinary least-squares coefficient. By very definition, x_i and $\varepsilon_i = y_i - x_i'\beta$ are uncorrelated.

Consequently, we can always write

$$y_i = x_i'\beta + \varepsilon_i$$

for some vector β such that $E(x_i\varepsilon_i) = 0$.

We call this the linear projection of y_i on x_i and write it as $E^*(y_i|x_i) = x_i'\beta$.

This does not mean that $E(y_i|x_i) = x_i'\beta$.

Endogeneity in a linear model

Again consider

$$y_i = x_i'\theta + \varepsilon_i \text{ but now we allow that } E(x_i\varepsilon) \neq 0.$$

Note that

- $E^*(y_i|x_i) \neq x_i'\theta$,
- so θ is not a regression coefficient;
- $E(y_i|x_i) = x_i'\theta + E(\varepsilon_i|x_i) \neq x_i'\theta$,
- so
$$\frac{\partial E(y_i|x_i)}{\partial x_i} \neq \theta.$$

Omitted variables

Say we have

$$y_i = \alpha_i + x_i'\theta + \eta_i, \qquad E(\eta_i|x_i, \alpha_i).$$

Say an agricultural (log-linearized) Cobb-Douglas production function.

- y_i is output;
- x_i are observable inputs ;
- α_i is soil quality;
- η_i is rainfall.

Farmer observes (α_i, x_i). We only observe x_i. In general, x_i, α_i are not independent.

Estimating

$$y_i = x_i'\theta + (\alpha_i + \eta_i) = x_i'\theta + \varepsilon_i$$

via least-squares suffers from endogeneity bias.

The problem is that α_i is not observed in data. Otherwise, can just include it in x_i.

Measurement error

Suppose that

$$y_i = w_i'\theta + \epsilon_i, \qquad E(\epsilon_i|w_i) = 0$$

but (together with y_i) we only observe a noisy version of w_i, say

$$x_i = w_i + \eta_i,$$

for measurement error η_i.

Then

$$y_i = w_i'\theta + \epsilon_i = (x_i - \eta_i)'\theta + \epsilon_i = x_i'\theta + (\epsilon_i - \eta_i'\theta) = x_i'\theta + \varepsilon_i.$$

Suppose, for simplicity, that $E(\eta_i\epsilon_i) = 0$ and $E(w_i\eta_i) = 0$. Then

$$E(x_i\varepsilon_i) = E(x_i(\epsilon_i - \eta_i'\theta)) = -E(x_i\eta_i')\,\theta = -E(\eta_i\eta_i')\,\theta \neq 0.$$

A least-squares regression would estimate the population quantity

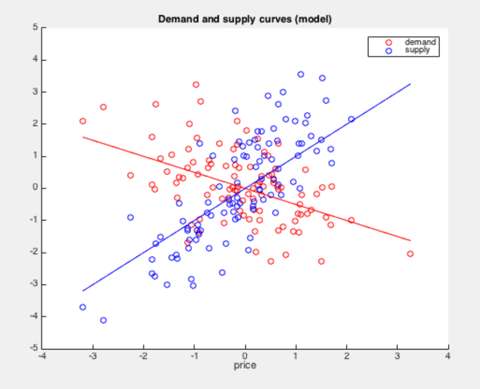
$$E(x_ix_i')^{-1}E(x_iy_i) = \theta + E(x_ix_i')^{-1}E(x_i\varepsilon_i) = \theta - E(x_ix_i')^{-1}E(\eta_i\eta_i')\,\theta.$$

Simultaneity

Temporary deviation from notational conventions to analyze market model

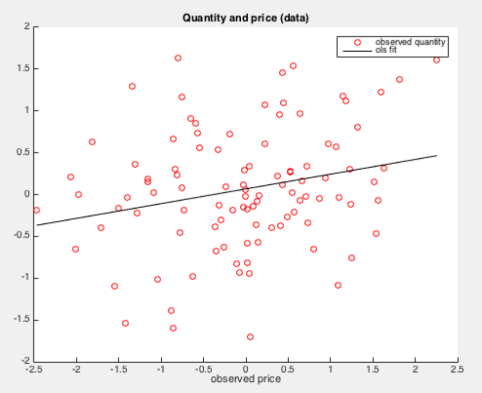
$$d_i = \alpha_d - \theta_d \, p_i + u_i$$
$$s_i = \alpha_s + \theta_s \, p_i + v_i$$

where d_i, s_i, p_i are demand, supply, and price, respectively.



We do not observe supply and demand for any given price.

Collected data is on quantity traded and transaction price, (q_i, p_i).



Data comes from markets in equilibrium.

So, we solve

$$s_i = d_i$$

for the equilibrium price to get

$$p_i = \frac{\alpha_d - \alpha_s}{\theta_d + \theta_s} + \frac{u_i - v_i}{\theta_d + \theta_s}.$$

This gives traded quantity as

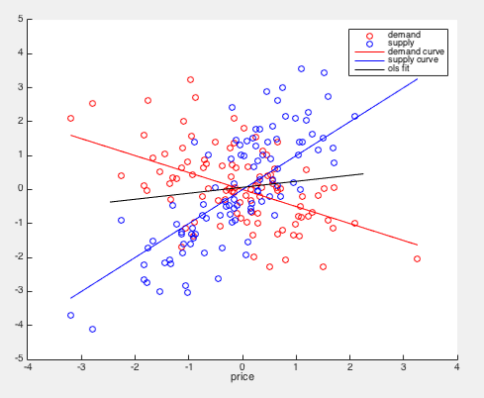
$$q_i = \frac{\alpha_d \theta_s + \alpha_s \theta_d}{\theta_d + \theta_s} + \frac{\theta_s u_i + \theta_d v_i}{\theta_d + \theta_s}.$$

(With $E(u_i v_i) = 0$) the population regression slope of q_i on p_i equals

$$\frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2}\,\theta_s - \frac{\sigma_v^2}{\sigma_u^2 + \sigma_v^2}\,\theta_d,$$

for $\sigma_u^2 = E(u_i^2)$ and $\sigma_v^2 = E(v_i^2)$.

Least-squares estimates a weighted average of supply and demand elasticities.


demand
supply
demand curve
supply curve
ols fit
5
4
3
2
1
0
-1
-2
-3
-4
-5
-4
-3
-2
-1
0
1
2
3
4
price

To see the problem in terms of endogeneity, focus on the estimation of the demand curve.

Then, collecting equations from above,

$$d_i = \alpha_d - \theta_d p_i + u_i, \qquad p_i = \frac{\alpha_d - \alpha_s}{\theta_d + \theta_s} + \frac{u_i - v_i}{\theta_d + \theta_s}.$$

Clearly,

$$E(p_i u_i) = E\left(u_i\left(\frac{u_i - v_i}{\theta_d + \theta_s}\right)\right) = \frac{\sigma_u^2}{\theta_d + \theta_s} \neq 0,$$

as the errors in both equations are correlated.

The same happens for the supply curve, as

$$s_i = \alpha_s + \theta_s p_i + v_i, \qquad p_i = \frac{\alpha_d - \alpha_s}{\theta_d + \theta_s} + \frac{u_i - v_i}{\theta_d + \theta_s}.$$

and

$$E(p_i v_i) = E\left(v_i\left(\frac{u_i - v_i}{\theta_d + \theta_s}\right)\right) = -\frac{\sigma_v^2}{\theta_d + \theta_s} \neq 0.$$

Linear instrumental-variable problem

Now suppose we have

$$y_i = x_i'\theta + \varepsilon_i, \qquad E(z_i\varepsilon_i) = 0$$

for instrumental variables z_i (with $\dim z_i = \dim x_i$).

This gives us the moment conditions

$$E_\theta(z_i\,(y_i - x_i'\theta)) = 0.$$

An instrument is

- valid if $E(z_i\varepsilon_i) = 0$; and
- relevant if $E(z_i x_i')$ is full rank.

We then obtain the instrumental-variable estimator

$$\hat{\theta} = \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i y_i\right).$$

It is useful to proceed in matrix notation:

$$\boldsymbol{y} = \boldsymbol{X}\theta + \boldsymbol{\varepsilon}$$

and we set to zero the sample covariance of the errors and instruments. The solution is

$$\hat{\theta} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{Z}'\boldsymbol{y}).$$

Note that this gives least squares when regressors instrument for themselves.

We need at least as many instruments as we have covariates.

To motivate the sequel suppose $\dim z_i > \dim x_i$. Then the $\dim z_i$ equations

$$\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\theta) = 0$$

involve $\dim x_i < \dim z_i$ unknowns. Then (generically) these equations do not have a solution (for finite n).

The method-of-moment idea fails to provide us with an estimator when we have overidentification.

Resolving simultaneity with instrumental variables

Return to the estimation of a demand curve but now suppose that

$$\begin{aligned} d_i &= \alpha_d - \theta_d\, p_i + u_i \\ s_i &= \alpha_s + \theta_s\, p_i + \pi z_i + v_i \end{aligned}.$$

where $E(z_i u_i) = 0$.

z_i shifts supply (relevance) but not demand (exclusion).

We now have the triangular system of equations

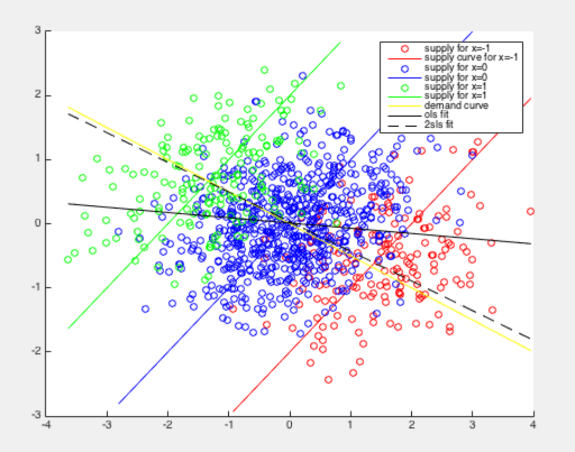
$$\begin{aligned} d_i &= \alpha_d - \theta_d\, p_i + u_i \\ p_i &= \frac{\alpha_d - \alpha_s}{\theta_d + \theta_s} - \frac{\pi}{\theta_d + \theta_s}\, z_i + \frac{u_i - v_i}{\theta_d + \theta_s} \end{aligned}.$$

Further, by relevance and exclusion,

$$\operatorname{cov}(d_i, z_i) = \operatorname{cov}(\alpha_d - \theta_d\, p_i + u_i, z_i) = -\theta_d \operatorname{cov}(p_i, z_i),$$

and so

$$-\theta_d = \frac{\operatorname{cov}(d_i, z_i)}{\operatorname{cov}(p_i, z_i)}.$$


3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4
supply for x=-1
supply curve for x=-1
supply for x=0
supply for x=0
supply for x=1
supply for x=1
demand curve
ols fit
2sls fit

Measurement error

Suppose again that

$$y_i = w_i\theta + \epsilon_i,$$

but that w_i is measured with error, say as $x_i = w_i + \eta_i$. Then

$$y_i = x_i\theta + (\epsilon_i - \eta_i\theta)$$

and a regression of y_i on x_i does not deliver a consistent estimator of θ.

Suppose that we have an additional noisy measurement of w_i,

$$z_i = w_i + \zeta_i.$$

If $E(\zeta_i\eta_i) = 0$ and $E(\zeta_i\epsilon_i) = 0$ we can estimate θ by instrumental variables.

We have

$$E(z_i x_i) = E((w_i + \zeta_i)(w_i + \eta_i)) = \sigma_w^2, \qquad E(z_i(\epsilon_i - \eta_i\theta)) = 0,$$

so z_i is relevant and valid.

Generalized method of moments

In overidentified problems (where we have more equations than unknowns) we cannot satisfy all empirical moment conditions,

$$\hat{g}(\theta) = n^{-1} \sum_{i=1}^{n} \varphi(x_i; \theta) = 0,$$

exactly.

The solution is to minimize the quadratic form

$$\hat{g}(\theta)' A \, \hat{g}(\theta).$$

for some (positive semi-definite) weight matrix A.

This is the generalized method of moments

Intuitively, we minimize the distance $\|\hat{g}(\theta) - 0\|_A$.

Reduction in moments

With

$$\hat{G}(\theta) = \frac{\partial \hat{g}(\theta)}{\partial \theta'} = n^{-1} \sum_{i=1}^{n} \frac{\partial \varphi(x_i; \theta)}{\partial \theta'},$$

the first-order condition to the GMM problem is

$$\hat{G}(\theta)' A \, \hat{g}(\theta) = 0.$$

This is a set of $\dim \theta$ linear combinations of the $\dim \varphi$ original moments.

Linear combination is determined by weight matrix A (which we may choose).

So different A give different estimators.

The optimal weight matrix turns out to be

$$A = \Omega_\theta^{-1}$$

(or a consistent estimator thereof).

Limit distribution

Combine the convergence result $\hat{G}(\theta_*) \xrightarrow{p} \Gamma_\theta$ for any consistent θ_* with the expansion

$$n^{-1} \sum_{i=1}^{n} \varphi(x_i; \hat{\theta}) = n^{-1} \sum_{i=1}^{n} \varphi(x_i; \theta) + n^{-1} \sum_{i=1}^{n} \left. \frac{\partial \varphi(x_i; \theta)}{\partial \theta'} \right|_{\theta_*} (\hat{\theta} - \theta)$$

to see that

$$(\hat{\theta} - \theta) = -(\Gamma'_\theta A \, \Gamma_\theta)^{-1} \, \Gamma'_\theta A \, \frac{1}{n} \sum_i \varphi(x_i; \theta) + o_p(n^{-1/2}).$$

Then, with

$$\frac{1}{\sqrt{n}} \sum_i \varphi(x_i; \theta) \xrightarrow{d} N(0, \Omega_\theta),$$

we get the result.

Theorem 23 (Limit distribution of GMM estimator)

Under regularity conditions,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, (\Gamma'_\theta A \, \Gamma_\theta)^{-1} (\Gamma'_\theta A \Omega_\theta A' \Gamma_\theta)(\Gamma'_\theta A \, \Gamma_\theta)^{-1})$$

as $n \to \infty$.

Optimal weighting

Theorem 24 (Semiparametric efficiency)

The efficiency bound (for a given set of moment conditions) is

$$(\Gamma_\theta' \Omega_\theta^{-1} \Gamma_\theta)^{-1}.$$

We can establish this by showing that the difference

$$(\Gamma_\theta' A\, \Gamma_\theta)^{-1} (\Gamma_\theta' A \Omega_\theta A' \Gamma_\theta)(\Gamma_\theta' A\, \Gamma_\theta)^{-1} - (\Gamma_\theta' \Omega^{-1} \Gamma_\theta)^{-1}$$

is a positive semi-definite matrix.

The bound is achieved if

$$A = \Omega_\theta^{-1}$$

(up to a scale) or if we use a consistent estimator.

Note that in this case we have a generalized information equality.

Proof.

Let

$$C = (\Gamma_\theta' A \Gamma_\theta)^{-1} \Gamma_\theta' A \Omega_\theta^{1/2}, \qquad D = \Omega_\theta^{-1/2} \Gamma_\theta.$$

Then

$$(\Gamma_\theta' A \Gamma_\theta)^{-1} (\Gamma_\theta' A \Omega_\theta A' \Gamma_\theta) (\Gamma_\theta' A \Gamma_\theta)^{-1} - (\Gamma_\theta' \Omega_\theta^{-1} \Gamma_\theta)^{-1}$$

can be written as

$$CC' - CD(D'D)^{-1} D'C'.$$

But this is

$$CM_D C' \geq 0, \qquad M_D = I_m - D(D'D)^{-1} D'.$$

The inequality follows because M_D is an orthogonal projection matrix, and so all its eigenvalues are zero or one. Hence, it is positive semi-definite.

To see that the eigenvalues of an orthogonal projector P are all zero or one, let $\lambda \neq 0$ be an eigenvalue of P. Then $Px = \lambda x$ for some $x \neq 0$. Because P is idempotent we must also have that $P^2 x = Px = \lambda Px = \lambda^2 x$. Therefore it must hold that

$$\lambda x = \lambda^2 x,$$

which can only be true if $\lambda \in \{0, 1\}$. □

χ^2 problem

To illustrate the efficiency gain of combining moments suppose that $x_i \sim \chi^2_\theta$.

We know that

$$E_\theta(x_i - \theta) = 0$$

so we could estimate θ by the sample mean.

But $\mathrm{var}_\theta(x_i) = 2\theta$ so also have the moment condition $E_\theta((x_i - \theta)^2 - 2\theta) = 0$.

Let

$$\varphi(x_i;\theta) = \left(\begin{array}{c} x_i - \theta \\ (x_i - \theta)^2 - 2\theta \end{array} \right).$$

Then

$$\Omega_\theta = E_\theta(\varphi(x_i;\theta)\,\varphi(x_i;\theta)') = 2\theta \left(\begin{array}{cc} 1 & 4 \\ 4 & 6(\theta+4) \end{array} \right).$$

So,

$$\Omega_\theta^{-1} = \frac{1}{\theta(3\theta+4)} \begin{pmatrix} \frac{3(\theta+4)}{2} & -1 \\ -1 & \frac{1}{4} \end{pmatrix}.$$

The Jacobian of the moment conditions is simply $-(1,2)'$ and so we find that the asymptotic variance equals

$$2\theta \, \frac{\theta + 4/3}{\theta + 2}.$$

If we would just use one of the moments the asymptotic variance would be

$$2\theta, \quad \text{and} \quad 3\theta(\theta+4),$$

respectively. Both are larger.

Note that Ω_θ depends on θ. So the optimal GMM estimator will generally be a two-step estimator (see below).

Estimation of the weight matrix introduces additional sampling noise that leads to bias and affects the coverage of confidence intervals/size and power of tests.

Poisson

When x_i is Poisson we similarly have

$$E_\theta(x_i - \theta) = 0, \qquad E_\theta((x_i - \theta)^2 - \theta) = 0$$

by the mean/variance equality.

Here,

$$\Omega_\theta = \theta \begin{pmatrix} 1 & 1 \\ 1 & 1 + 2\theta \end{pmatrix}, \qquad \Omega_\theta^{-1} = \frac{1}{2\theta^2} \begin{pmatrix} 1 + 2\theta & -1 \\ -1 & 1 \end{pmatrix}.$$

But as $\Gamma'_\theta = -(1, 1)$ we get

$$\Gamma'_\theta \Omega_\theta^{-1} \Gamma_\theta = \frac{1}{\theta}.$$

So the asymptotic variance is the same as for the simple estimator $\overline{x}_n$ based on the first moment condition only.

We knew we should have reached this conclusion here because $\overline{x}_n$ is the maximum-likelihood estimator and is best unbiased.

Two-step GMM

We can estimate $\Omega_\theta = E_\theta(\varphi(x_i;\theta)\,\varphi(x_i;\theta)')$ by

$$\hat{\Omega}_{\hat{\theta}} = n^{-1}\sum_i \varphi(x_i;\hat{\theta})\,\varphi(x_i;\hat{\theta})',$$

where we use a first-step GMM estimator $\hat{\theta}$ (constructed using a feasible A).

We then re-estimate θ by

$$\hat{\hat{\theta}} = \arg\min_\theta \hat{g}(\theta)'\hat{\Omega}_{\hat{\theta}}^{-1}\hat{g}(\theta).$$

In principle, this two-step procedure can be iterated.

Could also consider continuously-updated GMM:

$$\arg\min_\theta \hat{g}(\theta)'\hat{\Omega}_\theta^{-1}\hat{g}(\theta), \qquad \hat{\Omega}_\theta = n^{-1}\sum_i \varphi(x_i;\theta)\,\varphi(x_i;\theta)'.$$

This is computationally more challenging; first-order condition features extra terms (use MCMC).

Examples of (nonlinear) method of moments

Avery, R. B., L. P. Hansen, and V. J. Hotz (1983). Multiperiod probit models and orthogonality condition estimation. *International Economic Review* 24, 21–35.

Becker, G. S., M. Grossman, and K. M. Murphy (1994). An empirical analysis of cigarette addiction. *American Economic Review* 84, 396–418.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *RAND Journal of Economics* 25, 242–262.

Goldberg, P. K. and F. Verboven (2001). The evolution of price dispersion in the European car market. *Review of Economic Studies* 68, 811–848.

Hansen, L. P. and K. J. Singleton (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50, 1269–1286.

Pakes, A. (1986). Patents as options: Some estimates of the value of holding European patent stocks. *Econometrica* 54, 755–784.

Two-stage least squares

In the linear instrumental-variable problem (with more instruments than covariates) we minimize

$$(\boldsymbol{y}-\boldsymbol{X}\theta)'\boldsymbol{Z}A\,\boldsymbol{Z}'(\boldsymbol{y}-\boldsymbol{X}\theta).$$

The first-order condition is $(\boldsymbol{X}'\boldsymbol{Z})A\,\boldsymbol{Z}'(\boldsymbol{y}-\boldsymbol{X}\theta)=0$ and the solution is thus

$$\hat{\theta}=(\boldsymbol{X}'\boldsymbol{Z}A\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Z}A\boldsymbol{Z}'\boldsymbol{y}).$$

Under homoskedasticity,

$$\Omega_\theta=E(\varepsilon_i^2 z_i z_i')=\sigma_\varepsilon^2\,E(z_i z_i')$$

so that the optimal weight matrix is simply $\hat{\sigma}_\varepsilon^2\,(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$.

The efficient estimator is a one-step estimator and takes the form

$$\hat{\theta}=(\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y})=(\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{y}).$$

This is the two-stage least squares estimator.

To understand 2SLS recall the model

$$y_i = x_i'\theta + \varepsilon_i, \qquad E(z_i\varepsilon_i) = 0$$

and note that we can always use $E^*(x_i|z_i) = z_i'\pi$ to decompose the covariates as

$$x_i = z_i'\pi + \eta_i = \tilde{x}_i + \eta_i \text{ (say)}.$$

By the validity of z_i as instrument we have $E(z_i\varepsilon_i) = 0$ and so we know that

$$E(x_i\varepsilon_i) = E(\tilde{x}_i\varepsilon_i) + E(\eta_i\varepsilon_i) = E(\eta_i\varepsilon_i),$$

i.e., η_i is the endogenous part of x_i. Also, by virtue of the linear projection, $E(z_i\eta_i) = 0$ and so

$$E(\tilde{x}_i\eta_i) = 0.$$

It follows that, in

$$y_i = x_i'\theta + \varepsilon_i = \tilde{x}_i'\theta + (\varepsilon_i + \eta_i'\theta) = \tilde{x}_i'\theta + \epsilon_i \text{ (say)}.$$

the covariates $\tilde{x}_i$ and error ϵ_i are uncorrelated.

In practice, $\tilde{x}_i$ is unknown. Replacing it with an estimator gives 2SLS:

- Estimate $\tilde{x}_i$ by $\hat{x}_i$, the fitted values from a linear regression of x_i on z_i;
- Estimate θ by regressing y_i on $\hat{x}_i$.

Replacing population projection with sample projection introduces bias.

We have

$$\hat{\pi} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X} = \pi + (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{\eta}$$

and so

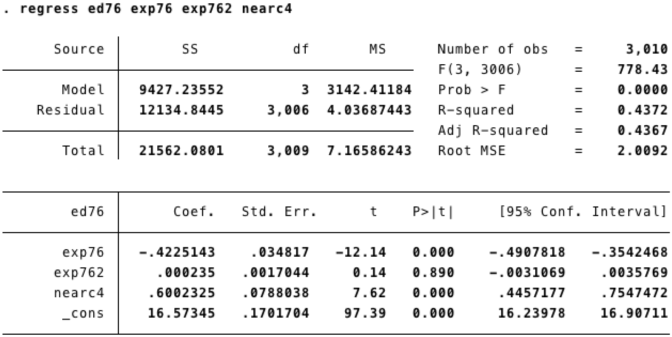
$$\hat{\boldsymbol{X}} = \tilde{\boldsymbol{X}} + \boldsymbol{P_Z}\boldsymbol{\eta}.$$

The second term correlates with $\boldsymbol{\varepsilon}$ and so

$$E(\hat{x}_i\varepsilon_i) \neq 0,$$

which introduces bias.

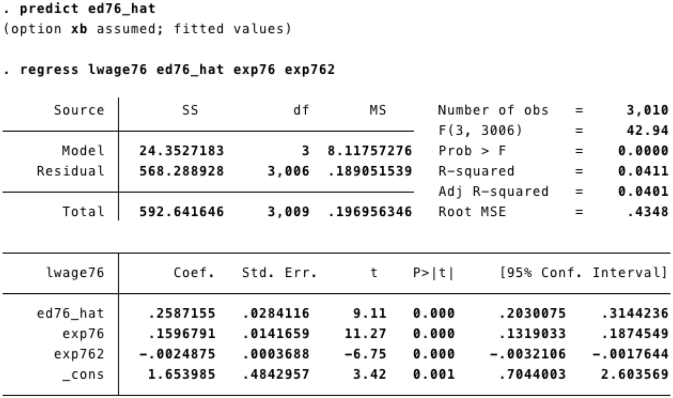
Which vanishes as $n \to \infty$, yielding consistency.



```
. regress ed76 exp76 exp762 nearc4

      Source |       SS           df       MS      Number of obs   =     3,010
-------------+----------------------------------   F(3, 3006)      =    778.43
       Model |  9427.23552         3  3142.41184   Prob > F        =    0.0000
    Residual |  12134.8445     3,006  4.03687443   R-squared       =    0.4372
-------------+----------------------------------   Adj R-squared   =    0.4367
       Total |  21562.0801     3,009  7.16586243   Root MSE        =    2.0092

------------------------------------------------------------------------------
        ed76 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exp76 |  -.4225143    .034817   -12.14   0.000    -.4907818   -.3542468
      exp762 |    .000235   .0017044     0.14   0.890    -.0031069    .0035769
      nearc4 |   .6002325   .0788038     7.62   0.000     .4457177    .7547472
       _cons |   16.57345   .1701704    97.39   0.000     16.23978    16.90711
------------------------------------------------------------------------------
```



```
. predict ed76_hat
(option xb assumed; fitted values)

. regress lwage76 ed76_hat exp76 exp762

      Source |       SS           df       MS      Number of obs   =     3,010
-------------+----------------------------------   F(3, 3006)      =     42.94
       Model |  24.3527183         3  8.11757276   Prob > F        =    0.0000
    Residual |  568.288928     3,006  .189051539   R-squared       =    0.0411
-------------+----------------------------------   Adj R-squared   =    0.0401
       Total |  592.641646     3,009  .196956346   Root MSE        =     .4348

------------------------------------------------------------------------------
     lwage76 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ed76_hat |   .2587155   .0284116     9.11   0.000     .2030075    .3144236
       exp76 |   .1596791   .0141659    11.27   0.000     .1319033    .1874549
      exp762 |  -.0024875   .0003688    -6.75   0.000    -.0032106   -.0017644
       _cons |   1.653985   .4842957     3.42   0.001     .7044003    2.603569
------------------------------------------------------------------------------
```

Precision of instrumental variables

Instrumental-variable estimators are always more variable than least squares.

Suppose we only have one covariate x_i, one instrument z_i, and homoskedastic errors.

The usual first-order approximation to the least-squares estimator is

$$\hat{\theta} - \theta \overset{a}{\sim} N\left(0, n^{-1}\frac{\sigma_\varepsilon^2}{\sigma_x^2}\right)$$

(under exogeneity).

The same first-order approximation to the instrumental-variable estimator is

$$\hat{\theta} - \theta \overset{a}{\sim} N\left(0, \frac{n^{-1}}{\rho_{xz}^2}\frac{\sigma_\varepsilon^2}{\sigma_x^2}\right),$$

where ρ_{xz} is the correlation between x_i and z_i.

The intuition is that x_i is (in terms of relevance/fit) its own best instrument.

The instrument is said to be weak when ρ_{xz} is small.

In this case the first-order approximation becomes poor.

Weak instruments

Take the simple univariate problem, where we only have one covariate x_i and m instruments z_i (treat these as fixed), and suppose we have homoskedastic errors.

We can approximate the mean squared error of 2SLS to second order to get

$$\underset{VARIANCE}{\frac{1}{n}\frac{\sigma_\varepsilon^2/\sigma_\eta^2}{\tau}} + \underset{SQUARED\ BIAS}{\left(\frac{m}{n}\right)^2\left(\frac{\rho\,\sigma_\varepsilon}{\tau}\right)^2} + o(n^{-2}),$$

where

$$n\,\tau = \frac{\pi'\boldsymbol{Z}'\boldsymbol{Z}\pi}{\sigma_\eta^2} = \frac{\overline{R}^2}{1-\overline{R}^2}$$

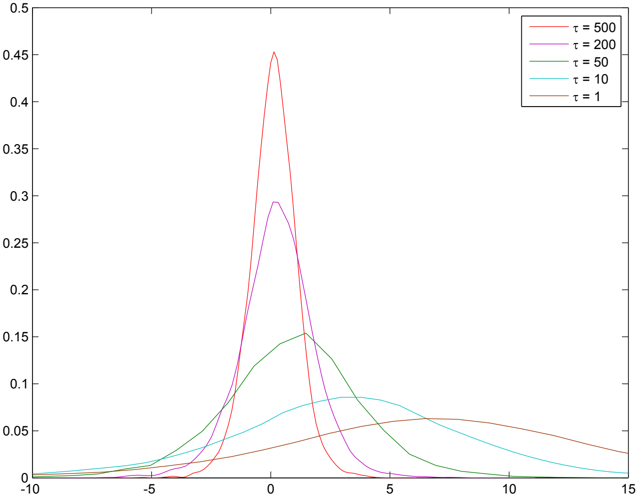
is the concentration parameter.

$\overline{R}^2$ is the (uncentered) population R^2 of the first-stage regression.

This relates directly to the first-stage F-statistic.

When τ is small most of the variation on x_i comes from η_i, and not from z_i.

Sampling distribution of two-stage least squares as a function of the value of the concentration parameter (simulation details omitted).



Many instruments

Note also how

$$\underbrace{\frac{1}{n}\frac{\sigma_\varepsilon^2/\sigma_\eta^2}{\tau}}_{VARIANCE} + \underbrace{\left(\frac{m}{n}\right)^2\left(\frac{\rho\,\sigma_\varepsilon}{\tau}\right)^2}_{SQUARED\ BIAS} + o(n^{-2})$$

depends on the number of instruments (m).

More instruments decrease the relative contribution of η_i to x_i.

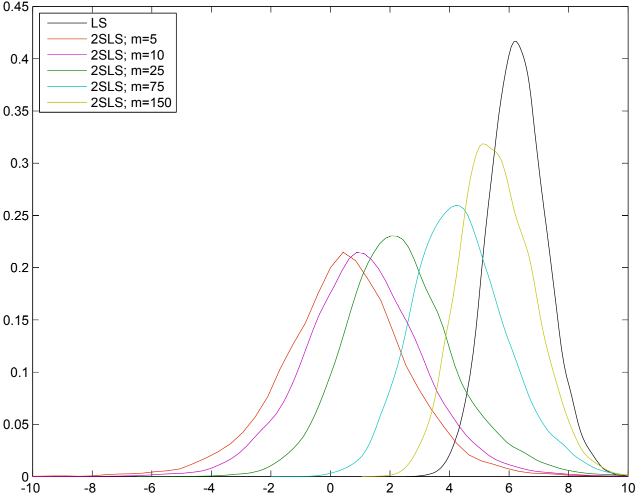
But the fitted values $\tilde{x}_i = z_i'\pi$ have to be estimated.

Under regularity conditions,

$$\hat{x}_i - \tilde{x}_i = O_p(\sqrt{m/n})$$

The noise in the fitted values grows with m.

Sampling distribution of two-stage least squares as a function of the number of instruments (simulation details omitted).



Control-function interpretation

Let

$$\boldsymbol{e} = \boldsymbol{M_Z X}$$

be the residuals from the least-squares regression of $\boldsymbol{X}$ on $\boldsymbol{Z}$ (i.e., from the first stage).

Then 2SLS can be written as

$$\hat{\theta} = (\boldsymbol{X}'\boldsymbol{P_Z X})^{-1}(\boldsymbol{X}'\boldsymbol{P_Z y}) = (\boldsymbol{X}'\boldsymbol{M_e X})^{-1}(\boldsymbol{X}'\boldsymbol{M_e y}).$$

Indeed,

$$\boldsymbol{M_e X} = \boldsymbol{M_e}(\boldsymbol{P_Z X} + \boldsymbol{M_Z X}) = (\boldsymbol{I} - \boldsymbol{P_e})\boldsymbol{P_Z X} + \boldsymbol{M_e e} = \boldsymbol{P_Z X}.$$

So 2SLS can equally be performed in the following two steps:

- Estimate η_i by e_i, the residuals from a linear regression of x_i on z_i;
- Estimate θ by regressing y_i on x_i and e_i.

This view on 2SLS gives us a way to test the null of exogeneity.

Work through the simple model with

$$\begin{aligned} y_i &= x_i\theta + \varepsilon_i \\ x_i &= z_i\pi + \eta_i \end{aligned}$$

where the errors are jointly normal.

Let e_i be the residual from the first stage.

Then 2SLS solves the empirical moments

$$\sum_i \begin{pmatrix} x_i \\ e_i \end{pmatrix} (y_i - x_i\theta - e_i\gamma) = 0.$$

for θ, γ.

As $e_i = x_i - z_i\hat{\pi} = \eta_i - z_i(\hat{\pi} - \pi)$ we can write this as (evaluating at true parameter values)

$$\sum_i \begin{pmatrix} x_i \\ \eta_i \end{pmatrix} u_i + \begin{pmatrix} x_i \\ \eta_i \end{pmatrix} z_i\gamma\,(\hat{\pi} - \pi) + \begin{pmatrix} 0 \\ z_i(\hat{\pi} - \pi) \end{pmatrix} \varepsilon_i + \begin{pmatrix} 0 \\ 1 \end{pmatrix} z_i^2(\hat{\pi} - \pi)^2$$

for $u_i = y_i - x_i\theta - \eta_i\gamma = \varepsilon_i - \eta_i\gamma$ (which does not correlate with x_i or η_i).

Because $E(z_i\eta_i) = 0$ and $E(z_i\varepsilon_i) = 0$, and because $\|\hat{\pi} - \pi\|^2 = O_p(n^{-1})$, this behaves like (as $n \to \infty$ and scaled by n^{-1})

$$n^{-1} \sum_i \begin{pmatrix} x_i \\ \eta_i \end{pmatrix} u_i + n^{-1} \sum_i \begin{pmatrix} \pi\gamma \\ 0 \end{pmatrix} z_i\eta_i,$$

were we have used that $\hat{\pi} - \pi = n^{-1} \sum_{i=1}^{n} z_i\eta_i / E(z_i^2) + o_p(n^{-1/2})$ The first term is standard, it also showed up when η_i was directly observed. The second term is present because we have replaced η_i by an estimator e_i; this introduces additional noise that has to be accounted for.

The variance-covariance matrix of the above random variable is

$$\begin{aligned} \Omega_\theta &= \begin{pmatrix} \sigma_x^2\sigma_u^2 + \gamma^2\pi^2\sigma_z^2\sigma_\eta^2 & \sigma_\eta^2\sigma_u^2 \\ \sigma_\eta^2\sigma_u^2 & \sigma_\eta^2\sigma_u^2 \end{pmatrix} \\ &= \sigma_u^2 \begin{pmatrix} \sigma_x^2 & \sigma_\eta^2 \\ \sigma_\eta^2 & \sigma_\eta^2 \end{pmatrix} + \begin{pmatrix} \gamma^2(\sigma_x^2 - \sigma_\eta^2)\sigma_\eta^2 & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

The limit of the Jacobian of the moment conditions is

$$\Gamma_\theta = \begin{pmatrix} \sigma_x^2 & \sigma_\eta^2 \\ \sigma_\eta^2 & \sigma_\eta^2 \end{pmatrix}, \qquad \Gamma_\theta^{-1} = \frac{1}{\sigma_\eta^2(\sigma_x^2 - \sigma_\eta^2)} \begin{pmatrix} \sigma_\eta^2 & -\sigma_\eta^2 \\ -\sigma_\eta^2 & \sigma_x^2 \end{pmatrix}.$$

The asymptotic variance of the estimator,

$$\Gamma_\theta^{-1}\Omega_\theta\Gamma_\theta^{-1},$$

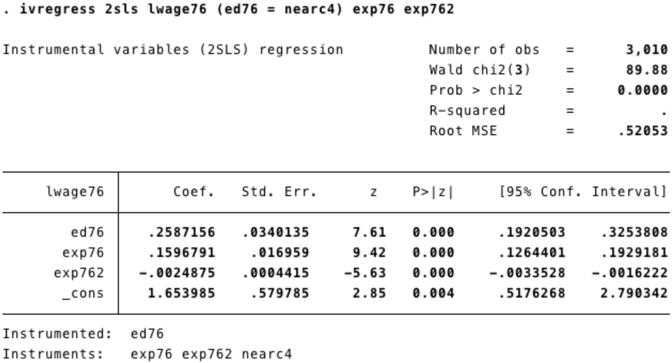
then equals

$$\sigma_u^2\Gamma_\theta^{-1}+\gamma^2\frac{\sigma_\eta^2}{\sigma_x^2-\sigma_\eta^2}\left(\begin{array}{rr}1 & -1\\ -1 & 1\end{array}\right).$$

Under the null of exogeneity ($\gamma = 0$), this is just $\sigma_u^2\Gamma_\theta^{-1}$ and so the usual least-squares standard error will be consistent.

Hence, the reported t-statistic is valid for testing exogeneity.

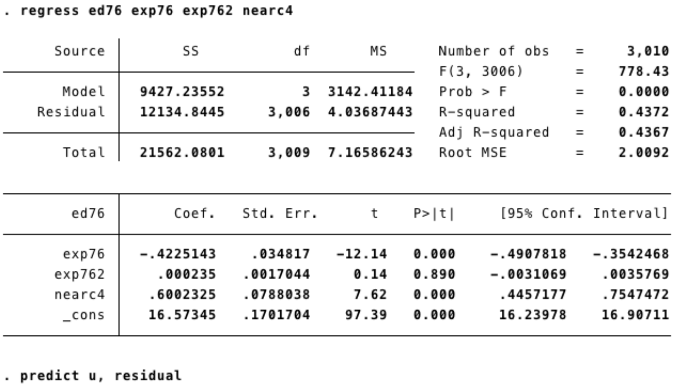
For our estimator of θ we do need a correction to the usual least-squares standard error as we want to allow that $\gamma \neq 0$.



```
. ivregress 2sls lwage76 (ed76 = nearc4) exp76 exp762

Instrumental variables (2SLS) regression          Number of obs   =      3,010
                                                  Wald chi2(3)    =      89.88
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =          .
                                                  Root MSE        =     .52053

------------------------------------------------------------------------------
     lwage76 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ed76 |   .2587156   .0340135     7.61   0.000     .1920503    .3253808
       exp76 |   .1596791    .016959     9.42   0.000     .1264401    .1929181
      exp762 |  -.0024875   .0004415    -5.63   0.000    -.0033528   -.0016222
       _cons |   1.653985    .579785     2.85   0.004     .5176268    2.790342
------------------------------------------------------------------------------
Instrumented:  ed76
Instruments:   exp76 exp762 nearc4
```

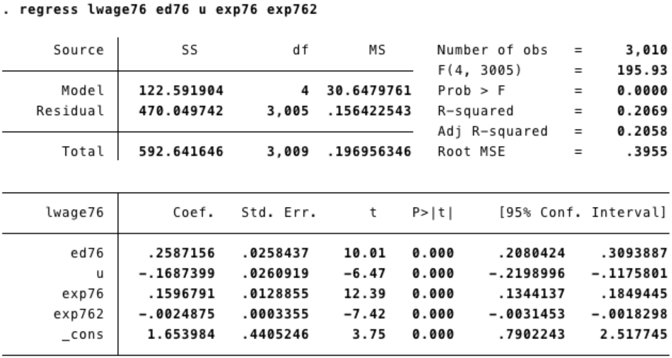


```
. regress ed76 exp76 exp762 nearc4

      Source |       SS           df       MS      Number of obs   =     3,010
-------------+----------------------------------   F(3, 3006)      =    778.43
       Model |  9427.23552         3  3142.41184   Prob > F        =    0.0000
    Residual |  12134.8445     3,006  4.03687443   R-squared       =    0.4372
-------------+----------------------------------   Adj R-squared   =    0.4367
       Total |  21562.0801     3,009  7.16586243   Root MSE        =    2.0092

------------------------------------------------------------------------------
        ed76 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exp76 |  -.4225143    .034817   -12.14   0.000    -.4907818   -.3542468
      exp762 |    .000235   .0017044     0.14   0.890    -.0031069    .0035769
      nearc4 |   .6002325   .0788038     7.62   0.000     .4457177    .7547472
       _cons |   16.57345   .1701704    97.39   0.000     16.23978    16.90711
------------------------------------------------------------------------------

. predict u, residual
```



```
. regress lwage76 ed76 u exp76 exp762

      Source |       SS           df       MS      Number of obs   =     3,010
-------------+----------------------------------   F(4, 3005)      =    195.93
       Model |  122.591904         4  30.6479761   Prob > F        =    0.0000
    Residual |  470.049742     3,005  .156422543   R-squared       =    0.2069
-------------+----------------------------------   Adj R-squared   =    0.2058
       Total |  592.641646     3,009  .196956346   Root MSE        =     .3955

------------------------------------------------------------------------------
     lwage76 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ed76 |   .2587156   .0258437    10.01   0.000     .2080424    .3093887
           u |  -.1687399   .0260919    -6.47   0.000    -.2198996   -.1175801
       exp76 |   .1596791   .0128855    12.39   0.000     .1344137    .1849445
      exp762 |  -.0024875   .0003355    -7.42   0.000    -.0031453   -.0018298
       _cons |   1.653984   .4405246     3.75   0.000     .7902243    2.517745
------------------------------------------------------------------------------
```

Bias correction with many moments

For a fixed weight matrix A, the bias in the GMM objective function is

$$E_\theta\big(\hat{g}(\theta)' A \hat{g}(\theta)\big) = \mathrm{tr}(A\Omega_\theta)/n.$$

The bias shrinks with n but grows (typically linearly) with $\dim \varphi$.

A bias-corrected GMM estimator minimizes

$$\hat{g}(\theta)' A g(\theta) - \mathrm{tr}(A\hat{\Omega}_\theta)/n = \frac{\sum_i \sum_{j\neq i} \varphi(x_i;\theta)' A \varphi(x_j;\theta)}{n^2}.$$

The continuously-updated estimator from above has a similar bias-correction interpretation.

For 2SLS we have $A = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ and so the bias-corrected objective function equals

$$\frac{\sum_i \sum_{j\neq i} (y_i - x_i'\theta)\, p_{ij}\, (y_j - x_j'\theta)}{n^2}$$

for $p_{ij} = (\boldsymbol{P}_{\boldsymbol{Z}})_{ij} = z_i'(\boldsymbol{Z}'\boldsymbol{Z})^{-1} z_j$.

Its minimizer is the jackknife instrumental-variable estimator

$$\left(\sum_i \sum_{j\neq i} x_i p_{ij} x_j'\right)^{-1} \left(\sum_i \sum_{j\neq i} x_i p_{ij} y_j\right) = \left(\sum_i \check{x}_i x_i'\right)^{-1} \left(\sum_i \check{x}_i y_i\right),$$

where

$$\check{x}_i = \sum_{j\neq i} x_j p_{ji} = \sum_{j\neq i} x_j z_j' (\boldsymbol{Z}'\boldsymbol{Z})^{-1} z_i = \hat{\Pi}_{-i} z_i.$$

Recall that the first-stage equation is of the form

$$x_i = \Pi z_i + \eta_i.$$

Here, $\hat{\Pi}_{-i}$ is a leave-one-out estimator of the first-stage coefficient matrix and $\check{x}_i$ is the associated fitted value.

Recall that bias in (feasible) 2SLS arose from the fact that $\hat{\Pi}$ is a function of η_i and η_i correlates with ε_i (See Slide 248). By construction the leave-one-out fitted values do not depend on η_i.

Multiplicative models with endogeneity

As an example of nonlinear instrumental-variable estimation, suppose that

$$y_i = \varphi(x_i;\theta)\,\varepsilon_i, \qquad E(\varepsilon_i|z_i) = 1.$$

We have (conditional) moment condition

$$E_\theta\left(\frac{y_i}{\varphi(x_i;\theta)} - 1 \middle| z_i \right) = 0$$

and so many unconditional moment conditions; for example

$$E_\theta\left(z_i \left(\frac{y_i}{\varphi(x_i;\theta)} - 1\right)\right) = E_\theta\left(\frac{z_i}{\varphi(x_i\theta)}\;(y_i - \varphi(x_i;\theta))\right) = 0.$$

An example is an exponential model.

Additional reading on instrumental variables

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* 62, 657–681.

Bound, J. , D. A. Jaeger, and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443–450.

Staiger, D. and J. H. Stock (1997). Instrumental variabels regression with weak instruments. *Econometrica* 65, 557–586.

Stock, J. H. and M. Yogo (2005). Testing for weak instruments in linear IV regression. In Andrews, D. W. K. and J. H. Stock (Editors), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, Chapter 5, 80—108 (Cambridge UP, Cambridge, UK).

Likelihood-ratio type test statistic

Now consider testing the m-dimensional constraint that $r(\theta) = 0$.

Let

$$\check{\check{\theta}} = \arg \min_{\theta : r(\theta) = 0} \hat{g}(\theta)' \hat{\Omega}_{\hat{\theta}}^{-1} \hat{g}(\theta);$$

the efficient GMM estimator under the constraint.

Theorem 25 (Limit distribution of LR-type statistic)

Under the null,

$$n\, \hat{g}(\check{\check{\theta}})' \hat{\Omega}_{\hat{\theta}}^{-1} \hat{g}(\check{\check{\theta}}) - n\, \hat{g}(\hat{\hat{\theta}})' \hat{\Omega}_{\hat{\theta}}^{-1} \hat{g}(\hat{\hat{\theta}}) \xrightarrow{d} \chi^2_m$$

as $n \to \infty$.

This result requires optimal weighting.

Note that we use the same weight matrix throughout.

Score type test statistic

Similarly, we can look whether the first-order condition of the unconstrained problem,

$$\hat{G}(\theta)'\hat{\Omega}_{\hat{\theta}}^{-1}\hat{g}(\theta) = 0,$$

when evaluated in the constrained estimator $\check{\check{\theta}}$, is far from zero.

Theorem 26 (Limit distribution of LM-type statistic)

Under the null,

$$n\,\hat{g}(\check{\check{\theta}})'\hat{\Omega}_{\hat{\theta}}^{-1}\hat{G}(\check{\check{\theta}})\,(\hat{G}(\check{\check{\theta}})'\hat{\Omega}_{\hat{\theta}}^{-1}\hat{G}(\check{\check{\theta}}))^{-1}\,\hat{G}(\check{\check{\theta}})'\hat{\Omega}_{\hat{\theta}}^{-1}\hat{g}(\check{\check{\theta}}) \xrightarrow{d} \chi^2_m$$

as $n \to \infty$.

This result requires optimal weighting.

Wald test statistic

The Wald statistic works without reference to the constrained problem.

Under optimal weighting we would have the following, where R is again the Jacobian matrix of the constraint vector r.

Theorem 27 (Limit distribution of the Wald statistic)

Under the null,

$$n\, r(\hat{\hat{\theta}})'(R(\hat{G}(\hat{\hat{\theta}})'\hat{\Omega}_{\hat{\theta}}^{-1}\hat{G}(\hat{\hat{\theta}}))^{-1}R')^{-1}r(\hat{\hat{\theta}}) \xrightarrow{d} \chi^2_m,$$

as $n \to \infty$.

More generally, when using estimator $\hat{\theta}$ computed using weight matrix A it equals

$$n\, r(\hat{\theta})'(R((\hat{G}(\hat{\theta})'A\,\hat{G}(\hat{\theta}))^{-1}(\hat{G}(\hat{\theta})'A\hat{\Omega}_{\hat{\theta}}A'\hat{G}(\hat{\theta}))(\hat{G}(\hat{\theta})'A\,\hat{G}(\hat{\theta}))^{-1})^{-1}R')^{-1}r(\hat{\theta}).$$

J-statistic

Note that

$$n\,\hat{g}(\hat{\hat{\theta}})'\hat{\Omega}_{\hat{\theta}}^{-1}\hat{g}(\hat{\hat{\theta}}) \xrightarrow{d} \chi^2_{\dim\varphi - \dim\theta}$$

if all moments hold.

So we can test the specification.

Only possible when we have overidentification, i.e., when

$$\dim\varphi - \dim\theta > 0$$

(Otherwise the test statistic is exactly zero).

If the J-statistic is large relative to the quantiles of the χ^2-distribution at least some of the moment conditions are likely to be invalid.

This does not tell us which moments are troublesome.

We can test subset of the moments as well.

Partition the moments using $\varphi(x;\theta) = (\varphi_1(x;\theta)', \varphi_2(x;\theta))'$.

Also partition

$$\hat{\Omega}_\theta = \begin{pmatrix} (\hat{\Omega}_\theta)_{11} & (\hat{\Omega}_\theta)_{12} \\ (\hat{\Omega}_\theta)_{21} & (\hat{\Omega}_\theta)_{22} \end{pmatrix}.$$

Want to test

$$E_\theta(\varphi_2(x_i;\theta)) = 0$$

assuming that $E_\theta(\varphi_1(x_i;\theta)) = 0$.

If $\dim \varphi_1 \geq \dim \theta$ we can compute

$$\check{\check{\theta}} = \arg\min_\theta \hat{g}_1(\theta)'(\hat{\Omega}_{\hat{\theta}})_{11}^{-1}\hat{g}_1(\theta),$$

where $\hat{g}_1(\theta) = n^{-1}\sum_i \varphi_1(x_i;\theta)$.

We can also compute the estimator using all moment conditions, i.e., the usual

$$\hat{\hat{\theta}} = \arg\min_\theta \hat{g}(\theta)'\hat{\Omega}_{\hat{\theta}}^{-1}\hat{g}(\theta).$$

We then have the following simple result.

Theorem 28 (Testing moment validity)

If all moment conditions hold,

$$n\,\hat{g}(\hat{\hat{\theta}})'\hat{\Omega}_{\hat{\theta}}^{-1}\hat{g}(\hat{\hat{\theta}}) - n\,\hat{g}_1(\check{\check{\theta}})'(\hat{\Omega}_{\hat{\theta}})_{11}^{-1}\hat{g}_1(\check{\check{\theta}}) \xrightarrow{d} \chi^2_{\dim\varphi - \dim\varphi_1}$$

as $n \to \infty$.

Note that we use the same weight matrix in both terms.

This ensures (in small samples) that the test statistic is non-negative.

Testing instrument validity

In the linear model

$$\boldsymbol{y} = \boldsymbol{X}\theta + \boldsymbol{\varepsilon}$$

with homoskedastic errors, the optimally-weighted GMM estimator is 2SLS and the objective function (scaled up by n and evaluated at its minimizer) equals

$$\frac{\hat{\boldsymbol{\varepsilon}}' \boldsymbol{P}_{\boldsymbol{Z}} \hat{\boldsymbol{\varepsilon}}}{\hat{\sigma}^2},$$

where $\hat{\boldsymbol{\varepsilon}}$ are the 2SLS residuals.

This statistic is known as Sargan's statistic.

Note that $\boldsymbol{P}_{\boldsymbol{Z}} \hat{\boldsymbol{\varepsilon}}$ are the fitted values of a regression of the 2SLS residuals on the instruments

Moreover, as $\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}/n$ we can equivalently write

$$n \, \frac{\hat{\boldsymbol{\varepsilon}}' \boldsymbol{P}_{\boldsymbol{Z}} \hat{\boldsymbol{\varepsilon}}}{\hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}} = n \, \frac{ESS}{TSS} = n \, R^2.$$

Invalid instruments can be detected by looking at correlation between the residuals and the instruments.

Optimal moment conditions in conditional models

Now suppose that we know

$$E_\theta(\varphi(x_i;\theta)|z_i) = 0$$

(a.s.)

This yields an infinite amount of unconditional moments.

We look for the optimal moment conditions, i.e., the function ψ in

$$E_\theta(\psi(z_i)\,\varphi(x_i;\theta)) = 0$$

for which the asymptotic variance of the resulting GMM estimator is minimal.

The optimal instrument turns out to be

$$\psi(z_i) = -E_\theta\left(\left.\frac{\partial\varphi(x_i;\theta)}{\partial\theta'}\right| z_i\right)' E_\theta\left(\left.\varphi(x_i;\theta)\,\varphi(x_i;\theta)'\right| z_i\right)^{-1} = -\Gamma_\theta(z_i)'\,\Omega_\theta(z_i)^{-1}.$$

Note that $\dim\psi = \dim\theta$.

Notice that, now,

$$\Omega_\theta = \mathrm{var}_\theta(\psi(z_i)\,\varphi(x_i;\theta)) = E_\theta\left(\Gamma_\theta(z_i)'\,\Omega_\theta(z_i)^{-1}\Gamma_\theta(z_i)\right),$$

and

$$\Gamma_\theta = E_\theta\left(\psi(z_i)\,\frac{\partial\varphi(x_i;\theta)}{\partial\theta'}\right) = -E_\theta\left(\Gamma_\theta(z_i)'\,\Omega_\theta(z_i)^{-1}\Gamma_\theta(z_i)\right),$$

(use iterated expectations) such that

$$\Omega_\theta = -\Gamma_\theta.$$

Hence, the generic sandwich-form asymptotic variance becomes

$$\mathrm{avar}_\theta(\hat\theta) = (\Gamma_\theta'\Omega_\theta^{-1}\Gamma_\theta)^{-1} = \Omega_\theta^{-1};$$

that is,

$$\sqrt{n}(\hat\theta-\theta) \xrightarrow{d} N(0,\Omega_\theta^{-1}).$$

This is the semiparametric efficiency bound.

Proof.

Let

$$g_i = \psi(z_i)\,\varphi(x_i;\theta), \qquad h_i = \Gamma_\theta'\,A\,\phi(z_i)\,\varphi(x_i;\theta)$$

for arbitrary alternative weight matrix A and instrument vector ϕ.

The asymptotic variances of the associated GMM estimators are

$$E_\theta(g_i g_i')^{-1}, \qquad E_\theta(h_i g_i')^{-1} E_\theta(h_i h_i') E_\theta(g_i h_i')^{-1},$$

respectively.

Rewriting gives

$$E_\theta(h_i g_i')^{-1}\,E_\theta(h_i h_i')\,E_\theta(g_i h_i')^{-1} - E_\theta(g_i g_i')^{-1} = E_\theta(h_i g_i')^{-1} E_\theta(v_i v_i') E_\theta(g_i h_i')^{-1}$$

for

$$v_i = h_i - g_i'\gamma, \qquad \gamma = E_\theta(g_i g_i')^{-1} E_\theta(g_i h_i').$$

This difference is positive semi-definite because $E(v_i v_i') \geq 0$. □

Linear model

With

$$y_i = x_i'\theta + \varepsilon_i, \qquad E_\theta(\varepsilon_i | x_i) = 0$$

we have

$$E_\theta(y_i - x_i'\theta \,|\, x_i) = 0.$$

Here,

$$\Gamma_\theta(x_i) = E_\theta\left(\frac{\partial(y_i - x_i'\theta)}{\partial\theta'} \middle| \, x_i \right) = -x_i', \qquad \Omega_\theta(x_i) = E_\theta(\varepsilon_i^2 \,|\, x_i) = \sigma_i^2 \text{ (say)}.$$

So,

$$\psi(x_i) = -\Gamma_\theta(x_i)'\,\Omega_\theta(x_i)^{-1} = \frac{x_i}{\sigma_i^2},$$

and the optimal estimator solves the empirical moment condition

$$n^{-1} \sum_{i=1}^{n} \frac{x_i(y_i - x_i'\theta)}{\sigma_i^2} = 0.$$

Observation i gets less weight if σ_i^2 is higher. This is weighted least squares.

If we write

$$\boldsymbol{V} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2).$$

Then the optimal estimator is

$$\hat{\theta} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y})$$

Under homoskedasticity, i.e., when $\sigma_i^2 = \sigma^2$ for all i this reduces to the simple

$$\hat{\theta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{y}),$$

which is ordinary least squares.

This is the Gauss-Markov theorem.

Exponential model

We have

$$y_i = e^{x_i'\theta}\,\varepsilon_i, \qquad E_\theta(\varepsilon_i|x_i) = 1,$$

and so

$$E_\theta(y_i - e^{x_i'\theta}|\,x_i) = 0.$$

Here,

$$\Gamma_\theta(x_i) = -e^{x_i'\theta}\,x_i', \qquad \Omega_\theta(x_i) = \sigma_i^2 \text{ (say)}.$$

The optimal empirical moment condition thus is

$$n^{-1}\sum_{i=1}^n \frac{x_i e^{x_i'\theta}(y_i - e^{x_i'\theta})}{\sigma_i^2} = 0$$

With Poisson data, for example, $\sigma_i^2 = e^{x_i'\theta}$ and the estimating equation is

$$n^{-1}\sum_{i=1}^n x_i(y_i - e^{x_i'\theta}) = 0.$$

With homoskedastic errors $\sigma_i^2 = \sigma^2\,(e^{x_i'\theta})^2$ and we solve

$$n^{-1}\sum_{i=1}^n \frac{x_i(y_i - e^{x_i'\theta})}{e^{x_i'\theta}} = 0.$$

Instrumental-variable model

Now if

$$E_\theta(y_i - x_i'\theta|\, z_i) = 0$$

we obtain

$$\Gamma_\theta(z_i)' = -E(x_i|\, z_i), \qquad \Omega_\theta(z_i) = E(\varepsilon_i^2|z_i) = \sigma_i^2 \text{ (say).}$$

So, we solve

$$n^{-1} \sum_{i=1}^{n} \frac{E(x_i|\, z_i)\,(y_i - x_i'\theta)}{\sigma_i^2} = 0.$$

Here, the reduced form is nonlinear, in general.

This is also true under homoskedasticity. So two-stage least squares is not optimal, in general.

Two-stage least squares approximates $E(x_i|z_i)$ by the linear projection $E^*(x_i|z_i)$.

Linear model for panel data

Suppose now that we have repeated measurements, as in

$$y_{it} = x_{it}'\theta + \varepsilon_{it}, \qquad t = 1, \ldots, T.$$

For each i we have a set of T equations.

This fits our framework on stacking observations for each i and writing

$$y_i = x_i'\theta + \varepsilon_i;$$

here, e.g., $y_i = (y_{i1}, \ldots, y_{iT})'$.

Suppose that

$$\varepsilon_{it} = \alpha_i + u_{it}.$$

We may have that α_i and x_i are correlated. Then $E(\varepsilon_i | x_i) \neq 0$.

However, with Δ the first-differencing operator, we have the $T-1$ equations

$$\Delta y_{it} = \Delta x_{it}'\theta + \Delta\varepsilon_{it} = \Delta x_{it}'\theta + \Delta u_{it}$$

that are free of α_i.

A sufficient condition for estimation is $E(u_{it} | x_i, \alpha_i) = 0$.

Define the $(T-1)\times T$ first-differencing matrix D as

$$D = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}.$$

We have the conditional moment conditions

$$E_\theta[D(y_i - x_i'\theta)|x_{i1},\ldots,x_{iT}] = 0.$$

The first-differenced least-squares estimator solves

$$\sum_{i=1}^{n} x_i\, D'\, D(y_i - x_i'\theta) = 0.$$

This is pooled least-squares on first-differenced data. It is inefficient as the Δu_{it} are correlated.

Suppose that $u_i \sim (0, \sigma^2 \, I_T)$. Then $Du_i \sim (0, \sigma^2 \, DD')$.

The optimal unconditional (empirical) moments are

$$\sum_{i=1}^{n} x_i \, D' \, (DD')^{-1} D \, (y_i - x_i'\theta) = 0.$$

This yields a generalized least-squares estimator. It is a pooled least-squares estimator on demeaned data.

A calculation gives

$$M = D' \, (DD')^{-1} D = I_T - \frac{\iota_T \iota_T'}{T},$$

where ι_T is a vector of ones.

The matrix M transforms data into deviations from within-group means. For example, $My_i = y_i - \overline{y}_i$.

Feedback

The above estimator requires that u_{it} is uncorrelated with $x_{i1}, \ldots, x_{iT}$. This rules out dynamics and, more generally, feedback.

A simple model where the problem arises is the (first-order) autoregression

$$y_{it} = y_{it-1}\theta + \alpha_i + u_{it},$$

where the initial value y_{i0} is taken as observed.

First-differencing (and, equivalently, demeaning) sweeps out α_i,

$$\Delta y_{it} = \Delta y_{it-1}\theta + \Delta u_{it},$$

but (taking u_{it} to be homoskedastic and serially uncorrelated for simplicity)

$$E(\Delta y_{it-1}\, \Delta u_{it}) = E(\Delta u_{it-1}\, \Delta u_{it}) = -E(u_{t-1}^2) = -\sigma^2 \neq 0,$$

and so introduces a new endogeneity problem.

An assumption of sequential exogeneity, i.e., $E(u_{it}|y_{i0},\ldots,y_{it-1},\alpha_i)=0$ is enough to obtain a GMM estimator.

It implies (sequential) conditional moments

$$E_\theta(\Delta y_{it}-\Delta y_{it-1}\theta|\, y_{i0},\ldots,y_{it-2})=0.$$

The conventional GMM estimator uses the linear moment conditions

$$E\left(\begin{pmatrix} y_{it-2} \\ y_{it-3} \\ \vdots \\ y_{i0} \end{pmatrix}\bigl(\Delta y_{it}-\Delta y_{it-1}\theta\bigr)\right)=0$$

(for all $t=2,\ldots,T$).

DEALING WITH (WEAK) DEPENDENCE

Reading

Hansen II, Chapters 14 and 15

Hayashi, Chapter 6

Stationary

Random sampling may be too strong a requirement:

- Time series data;
- Interactions and other network data;
- Snowball sampling (and so on).

Consider a scalar sequence $\{x_i\}$.

$\{x_i\}$ is (strictly) stationary if, for any $h \geq 0$, the distribution of $(x_i, \ldots, x_{i+h})$ does not depend on i.

An implication (if the moments exist) is weak stationarity: the mean $E(x_i)$ and covariance $E(x_i x_{i+h}) - E(x_i)E(x_{i+h})$ do not depend on i.

The techniques discussed so far can be adapted to stationary data provided they are weakly dependent.

Dependence and mixing

Weak dependence is a requirement that the overall behavior of $\{x_i\}$ is not driven by the realizations of the initial random variables (or any of the other variables later on).

Blocks of data $(x_i, \ldots, x_{i+j})$ and $(x_{i+j+h}, \ldots, x_{i+j+h+k})$ separated by h units become independent as h grows.

One way to formalize weak dependence is through mixing.

For $h \geq 1$ define the mixing coefficients

$$\alpha_h = \sup_{A \in \mathcal{A}, B \in \mathcal{B}} |P(A \cap B) - P(A)\, P(B)|,$$

where (somewhat crudely stated) the sets $\mathcal{A}$ and $\mathcal{B}$ cover all events involving $x_{-\infty}, \ldots, x_{i-1}, x_i$ and $x_{i+h}, x_{i+h+1}, \ldots, x_{+\infty}$, respectively. Note how these sets depend on h.

The process $\{x_i\}$ is strongly mixing (or alpha mixing) if $\alpha_h \to 0$ as $h \to \infty$. Note that independent data has $\alpha_h = 0$ for any h.

Consistency of the sample mean

Now let

$$\overline{x}_n = n^{-1} \sum_{i=1}^{n} x_i$$

be the sample mean.

Theorem 29 (Law of large numbers)

Suppose that $\{x_i\}$ is stationary and mixing, and that $\mu = E(x_i)$ exists. Then

$$\overline{x}_n \xrightarrow{p} \mu$$

as $n \to \infty$.

This is a substantial generalization of our earlier law of large numbers for random samples.

Note that this also implies that the continuous-mapping theorem generalizes in the same way.

Central limit theorem

Theorem 30 (Central limit theorem)

Suppose that $\{x_i\}$ is stationary and mixing with mixing coefficient satisfying

$$\sum_{h=1}^{\infty} \alpha_h^{\delta/(2+\delta)} < +\infty,$$

that $E(x_i) = \mu$ exists and that $E(\|x_i\|^{2+\delta})$ for some $\delta > 0$. Then

$$\sqrt{n}\Sigma^{-1/2}(\overline{x}_n - \mu) \xrightarrow{d} N(0, I)$$

for

$$\Sigma = \lim_{n\to\infty}\left(\Sigma_0 + \sum_{h=1}^{n-1}\frac{n-h}{n}(\Sigma_h + \Sigma_h')\right) = \Sigma_0 + \sum_{h=1}^{\infty}(\Sigma_h + \Sigma_h') = \sum_{h=-\infty}^{+\infty}\Sigma_h < \infty,$$

where $\Sigma_h = E((x_i - \mu)(x_{i+h} - \mu)')$.

Many special cases of this theorem are available with (complicated) low-level conditions for specific processes.

The summability of the covariances (i.e., the fact that Σ is finite) follows from the restriction on the mixing coefficients. A sufficient condition for summability is that $\Sigma_h \to 0$ faster than $1/h \to 0$.

The variance formula follows from (again for the scalar case)

$$\begin{aligned}
\operatorname{var}\left(\sum_{i=1}^{n} x_i\right) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{cov}(x_i, x_j) \\
&= \sum_{i=1}^{n}\left(\operatorname{cov}(x_i, x_i) + \sum_{h=1}^{i-1} \operatorname{cov}(x_i, x_{i-h}) + \sum_{h=1}^{n-i} \operatorname{cov}(x_i, x_{i+h})\right) \\
&= \sum_{i=1}^{n}\left(\Sigma_0 + \sum_{h=1}^{i-1} \Sigma_{-h} + \sum_{h=1}^{n-i} \Sigma_h\right) \\
&= n\Sigma_0 + (n-1)(\Sigma_1 + \Sigma_{-1}) + \cdots + (\Sigma_{n-1} + \Sigma_{-(n-1)}) \\
&= n\left(\Sigma_0 + \sum_{h=1}^{n-1} \frac{(n-h)}{n}(\Sigma_h + \Sigma_{-h})\right).
\end{aligned}$$

We have $\Sigma_{-h} = \Sigma_h'$.

A (truncated) estimator of the long-run variance Σ can be constructed as

$$\hat{\Sigma} = \hat{\Sigma}_0 + \sum_{h=1}^{\kappa-1} \frac{(\kappa - h)}{\kappa} (\hat{\Sigma}_h + \hat{\Sigma}_h')$$

for chosen integer $\kappa \leq n$ and

$$\hat{\Sigma}_h = \frac{1}{n-h} \sum_{i=1}^{n-h} (x_i - \overline{x}_n)(x_{i+h} - \overline{x}_n)'.$$

The resulting estimator is typically referred to as a HAC estimator.

The truncation at κ lags is needed because $\hat{\Sigma}_h$ becomes increasingly noisy as a function of h (for given n). (Consistency requires that $\kappa \to \infty$ with n, but not too fast.)

$\hat{\Sigma}$ is also called the Newey-West variance estimator. It can be interpreted as a 'kernel' estimator with a triangular kernel. Importantly, an unlike most other such 'kernel' estimators, it is ensured to be positive semi-definite.

Autoregression

Suppose that

$$x_i = \alpha + \rho\, x_{i-1} + \varepsilon_i, \qquad \varepsilon_i \sim \text{i.i.d.}\, N(0, \sigma^2).$$

We impose that $|\rho| < 1$.

We have $E(x_i) = \alpha + \rho\, E(x_{i-1})$ and so

$$\mu = E(x_i) = \frac{\alpha}{1-\rho}.$$

Also, $\operatorname{var}(x_i) = \rho^2 \operatorname{var}(x_{i-1}) + \sigma^2$ and, hence,

$$\Sigma_0 = \frac{\sigma^2}{1-\rho^2}.$$

The univariate stationary distribution, therefore, is $x_i \sim N(\mu, \Sigma_0)$. The covariances are proportional to Σ_0:

$$\Sigma_h = \rho^h\, \Sigma_0;$$

for example,

$$\Sigma_1 = \operatorname{cov}(x_i, x_{i-1}) = \operatorname{cov}(\alpha + \rho\, x_{i-1} + \varepsilon_i, x_{i-1}) = \rho \operatorname{cov}(x_{i-1}, x_{i-1}) = \rho\Sigma_0.$$

$\Sigma_h = \rho^h \, \Sigma_0$ shrinks at a geometric rate. The long-run variance is well-defined and equals

$$\Sigma = \Sigma_0 + 2 \sum_{h=1}^{\infty} \rho^h \Sigma_0 = \frac{1+\rho}{1-\rho} \Sigma_0 .$$

Here, the particularly simple structure of Σ suggests the simple alternative HAC estimator

$$\frac{1+\hat{\rho}}{1-\hat{\rho}} \hat{\Sigma}_0$$

where

$$\hat{\rho} = \frac{\sum_{i=2}^{n} x_{i-1} \, (x_i - \overline{x}_n)}{\sum_{i=2}^{n} (x_i - \overline{x}_n)^2}, \quad \hat{\Sigma}_0 = \frac{\hat{\sigma}^2}{1-\hat{\rho}^2},$$

and

$$\hat{\sigma}^2 = \frac{\sum_{i=2}^{n} ((x_i - \overline{x}_n) - \hat{\rho} \, (x_{i-1} - \overline{x}_n))^2}{n-1}.$$

When $\rho = 0$ $\{x_i\}$ is i.i.d. and $\Sigma = \Sigma_0$ but, for example, when $\rho = .5$ $\{x_i\}$ is dependent and $\Sigma = 3\Sigma_0$.

Ignoring the non-zero covariances can lead to large underestimation of the actual variability of the series.

Moving average

Another example has

$$x_i = \mu + \varepsilon_i + \beta\,\varepsilon_{i-1}, \qquad \varepsilon_i \sim \text{i.i.d. } N(0,\sigma^2).$$

Here, $E(x_i) = \mu$ is immediate. Further,

$$\Sigma_0 = (1+\beta^2)\,\sigma^2 \qquad \text{and} \qquad \Sigma_1 = \beta\,\sigma^2.$$

However,

$$\Sigma_h = 0, \qquad |h| > 1,$$

so the dependence in short-lived and vanishes abruptly beyond the first-order autocovariance.

It follows that

$$\Sigma = \Sigma_0 + \Sigma_{-1} + \Sigma_1 = (1+\beta)^2\sigma^2.$$

A combination of both examples gives

$$x_i = \alpha + \rho\,x_{i-1} + \varepsilon_i + \beta\,\varepsilon_{i-1}, \qquad \varepsilon_i \sim \text{i.i.d. } N(0,\sigma^2).$$

Extensions to higher-order are immediate. This gives a parsimonious way to modelling dependence.

Limit distribution of GMM

In practice the above is important for getting correct standard errors with serially dependent data.

The two-step GMM estimator solves

$$\min_{\theta} \hat{g}(\theta)' \hat{\Omega}_{\hat{\theta}}^{-1} \hat{g}(\theta).$$

where, now, $\hat{\Omega}_{\hat{\theta}}$ is a HAC estimator of the long-run covariance matrix of the moment condition

$$\hat{g}(\theta) = n^{-1} \sum_{i=1}^{n} \varphi(x_i; \theta).$$

The same robust estimator needs to be used when constructing test statistics.

The remainder of the argument for GMM carries over without modification.

Linear model with correlated errors

Consider

$$y_i = x_i'\theta + \varepsilon_i$$

with $E(\varepsilon_i|x_1,\ldots,x_n)=0$. Then, as before,

$$\sqrt{n}(\hat{\theta}-\theta) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i\right).$$

Now,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i \xrightarrow{d} N(0,\Omega), \qquad \Omega = \sum_{h=-\infty}^{+\infty} E((\varepsilon_i\varepsilon_{i+h})(x_i x_{i+h}')).$$

This covariance allows for both heteroskedasticity and autocorrelation in the errors.

Under homoskedasticity, $E(\varepsilon_i\varepsilon_{i+h}|x_1,\ldots,x_n) = E(\varepsilon_i\varepsilon_{i+h})$, and the formula simplifies to

$$\Omega = \sum_{h=-\infty}^{+\infty} E(\varepsilon_i\varepsilon_{i+h})\, E(x_i x_{i+h}').$$

If, in addition, we also have $E(\varepsilon_i\varepsilon_{i+h})=0$ for $h\neq 0$, then $\Omega=\sigma^2 E(x_i x_i')$.

Autoregression

A simple dynamic model is

$$y_i = \rho y_{i-1} + \varepsilon_i, \qquad E(\varepsilon_i | y_1, \ldots, y_{i-1}) = 0.$$

Here the regressor (the lagged outcome) is not strictly exogenous but only weakly exogenous.

Nonetheless,

$$E(y_{i-1}\varepsilon_i) = E_\rho(y_{i-1}(y_i - \rho y_{i-1})) = 0$$

and so least-squares continues to be consistent and asymptotically normal (under the usual regularity conditions).

As

$$y_i = \sum_{h=0}^{\infty} \rho^h \, \varepsilon_{i-h},$$

weak exogeneity implies that the ε_i are uncorrelated.

Autoregression with MA errors

An extension would be

$$y_i = \rho y_{i-1} + \varepsilon_i, \qquad \varepsilon_i = \eta_i + \theta \eta_{i-1},$$

where $\eta_i \sim$ i.i.d. $(0, \sigma^2)$.

Now, least-squares is not consistent. Indeed,

$$E(y_{i-1}\varepsilon_i) = \theta\, \sigma^2,$$

so the usual moment condition is no longer valid.

However, lack of higher-order correlation in the error does imply that

$$E(y_{i-h}\varepsilon_i) = 0, \qquad \text{for all } h \geq 2,$$

opening the way for an instrumental-variable approach.

The extension to higher-order MA processes is immediate.

Note that this approach does not work for autoregressive errors.

Intertemporal CAPM

A consumer chooses consumption stream $\{x_i\}$ to maximize her expected (discounted) utility stream

$$E(\textstyle\sum_{h=0}^{\infty} \alpha^h u(x_{i+h};\beta)|\, z_i),$$

where u is a well-behaved utility function and z_i is the information set at baseline i.

Optimality of the consumption path implies that, for all i, the Euler equation

$$\alpha^i u'(x_i;\beta)\, dx_i = \alpha^{i+1} u'(x_{i+1};\beta)\, r\, dx_i$$

holds. Here, r is the asset return.

Hence,

$$E\,(\,\alpha\, r\, u'(x_{i+1};\beta)/u'(x_i;\beta) - 1|\; z_i) = 0$$

is a valid conditional moment condition for α, β.

NONPARAMETRIC PROBLEMS: CONDITIONAL-MEAN FUNCTIONS

Reading

Hansen II, Chapters 19 and 20

Li and Racine, Chapters 1 and 2

Horowitz, Appendix

Nonparametric specification

Let y_i and x_i be i.i.d. univariate random variables.

Suppose that

$$y_i = m(x_i) + \varepsilon_i, \qquad E(\varepsilon_i | x_i) = 0.$$

We want to estimate m without imposing a functional form.

If x_i takes values $v_1, \ldots, v_k$ for finite $k < n$ this is a parametric problem:

Regress y_i on k dummy variables $d_{i,\kappa} = \{x_i = v_\kappa\}$, i.e.,

$$y_i = \sum_{\kappa=1}^{k} d_{i,\kappa}\, \beta_\kappa + \varepsilon_i;$$

then $\beta_\kappa = m(v_\kappa)$. Equivalently, for a fixed value $x \in \{v_1, \ldots, v_k\}$,

$$\hat{m}(x) = \sum_{i=1}^{n} \omega_i\, y_i, \qquad \omega_i = \frac{\{x_i = x\}}{\sum_{j=1}^{n} \{x_j = x\}};$$

this is the slope of a regression of y_i on $d_{i,x}$, or sample mean in the subsample with $x_i = x$.

Estimation based on binning

When x_i is continuous the probability that it takes on any given value is zero.

We could estimate $m(x)$ by the weighted average

$$\hat{m}_h(x) = \sum_{i=1}^{n} \omega_i y_i, \qquad \omega_i = \frac{\{|x_i - x| \leq h\}}{\sum_j \{|x_j - x| \leq h\}},$$

where h is some chosen positive number, the bandwidth.

This makes sense if we believe m is smooth, so that $m(x)$ does not change too fast when x changes little.

A small choice for the bandwidth h defines a small neighborhood and so decreases bias. But it also increases variance as there will be less observations ‘close’ to x.

Kernel functions

Binning yields a non-smooth estimator of $m(x)$ (as a function of x). Which may not be attractive.

A (second-order) kernel function is any (symmetric) non-negative and bounded function k for which $\int k(u)\,du = 1$, $\int u\,k(u)\,du = 0$, and $\int u^2\,k(u)\,du < \infty$.

Note that any probability density function with finite second moments can be made to satisfy these requirements.

Commonly-used examples are

- Uniform : $\frac{1}{2}\{|u| \leq 1\}$
- Triangular : $(1 - |u|)\,\{|u| \leq 1\}$
- Epanechnikov : $\frac{3}{4}(1 - u^2)\,\{|u| \leq 1\}$
- Gaussian: $\frac{1}{\sqrt{2\pi}} e^{-u^2/2}$

A locally-constant estimator

A kernel estimator of $m(x)$ is

$$\hat{m}_h(x) = \sum_{i=1}^{n} \omega_i y_i, \qquad \omega_i = \frac{k\left(\frac{x_i - x}{h}\right)}{\sum_j k\left(\frac{x_j - x}{h}\right)}.$$

This is the Nadaraya-Watson estimator.

The binning estimator is the special case that uses the uniform kernel. Smooth choices for the kernel k deliver smooth estimators of m.

In practice the choice of h is far more important to the behavior of $\hat{m}_h$ than is the choice of k.

Note that $\hat{m}_h(x)$ solves the weighted least squares problem

$$\min \sum_{i=1}^{n} \omega_i \left(y_i - \alpha\right)^2$$

with respect to α.

The above optimization problem is equivalent to minimizing

$$\frac{1}{n}\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)\,\frac{(y_i-\alpha)^2}{h}.$$

As n grows this sample averages converges to its expectation, which equals

$$E\left(k\left(\frac{x_i - x}{h}\right)\,\frac{(y_i-\alpha)^2}{h}\right) = E\left(k\left(\frac{x_i - x}{h}\right)\,\frac{E((y_i-\alpha)^2|\,x_i)}{h}\right)$$

Let f be the density of x_i and $g(x_i) = E((y_i-\alpha)^2|\,x_i)\,f(x_i)$. A change of variable to $u = (x_i - x)/h$ and a second-order expansion around $u = 0$ show the expectation to equal

$$\begin{aligned}\textstyle\int k\left(\frac{x_i-x}{h}\right)\,\frac{E((y_i-\alpha)^2|\,x_i)\,f(x_i)}{h}\,dx_i &= \textstyle\int k(u)\,g(x+hu)\,du \\ &= \textstyle\int k(u)\left(g(x) + hu\,g'(x) + \frac{1}{2}h^2u^2g''(u^*)\right)du\end{aligned}$$

where u^* lies between u and zero.

Using the properties of the kernel function and assuming that $|g''|_\infty < \infty$,

$$E\left(k\left(\frac{x_i - x}{h}\right)\,\frac{(y_i-\alpha)^2}{h}\right) = E\left(\,(y_i-\alpha)^2\middle|\,x_i = x\right)\,f(x) + O(h^2).$$

Consistency requires that $h \overset{n\uparrow\infty}{\to} 0$.

Can also show that the variance is proportional to $(nh)^{-1}$. So, if f is equally well behaved, and $f(x) > 0$,

$$\sum_{i=1}^{n} \omega_i \left(y_i - \alpha\right)^2 = \frac{\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right) \left(y_i - \alpha\right)^2}{\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)} \xrightarrow{p} E\left(\left(y_i - \alpha\right)^2 \middle| \, x_i = x\right)$$

if $n \to \infty$ provided $h \to 0$ and $nh \to \infty$.

The solution of this limit problem is, of course,

$$\alpha = m(x),$$

justifying the Nadaraya-Watson estimator.

We have implicitely derived the nonparametric kernel density estimator

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)$$

for $f(x)$.

The conditions on h relative to n represent the bias/variance trade-off.

As we need $h \to 0$ the estimator $\hat{m}(x)$ will converge at a slower rate than $n^{-1/2}$ (the parametric rate).

We have

$$\sqrt{nh}\left(\hat{m}_h(x) - m(x) - h^2\, b(x) \int u^2\, k(u)\, du\right) \xrightarrow{d} N\left(0, \frac{\sigma^2(x)}{f(x)} \int k(u)^2\, du\right),$$

where we let

$$b(x) = \frac{m''(x)}{2} + \frac{f'(x)\, m'(x)}{f(x)}, \qquad \sigma^2(x) = \operatorname{var}(y_i|x_i = x) = E(\varepsilon_i^2|x_i = x),$$

be first-order bias and variance, respectively.

The convergence rate can be no faster than $n^{-2/5}$, which happens when bias and standard deviation shrink at the same rate.

With the bias being $O(h^2)$ and the variance $O((nh)^{-1})$ bias vanishes if we choose h such that $nh^5 \to 0$.

This is called undersmoothing; it makes bias small relative to standard error.

Alternatives are bias correction or the use of higher-order kernels. Needed to perform valid inference.

A locally-linear estimator

Rather than a (weighted) regression on a constant alone we may equally fit a linear approximation to m at x.

This amounts to estimating $m(x)$ by the intercept in

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \omega_i \left(y_i - \alpha - (x_i - x)\beta\right)^2.$$

Although it has the same asymptotic behavior, such an estimator tends to perform better than the standard kernel estimator.

In principle, there is no reason to stop at linearity.

Local polynomial regressions, where we add powers of $(x_i - x)$ as regressors, are common practice.

Bandwidth choice

A bandwidth that is 'good' (in an overall sense but not at a point) minimizes

$$\int E((\hat{m}_h(x) - m(x))^2)\, f(x)\, dx,$$

the integrated (with respect to f) mean squared error.

This measure is unknown but can be estimated (up to a constant) by

$$n^{-1} \sum_{i=1}^{n} (y_i - \check{m}_h(x_i))^2,$$

where $\check{m}_h(x_i)$ is the leave-one-out estimator; for Nadaraya-Watson it equals

$$\check{m}_h(x_i) = \frac{\sum_{j \neq i} k\left(\frac{x_j - x_i}{h}\right) y_j}{\sum_{j \neq i} k\left(\frac{x_j - x_i}{h}\right)}$$

for example.

(Least-squares) cross-validation selects h by $\min_h n^{-1} \sum_{i=1}^{n} (y_i - \check{m}_h(x_i))^2$.

This can also be used to select between different estimators (pick the one with lowest IMSE).

Curse of dimensionality

Kernel estimators extend easily to the case where the conditioning variable is the vector $x_i = (x_{i,1}, \dots, x_{i,\kappa})'$.

It suffices to redefine k to be multivariate. A simple choice would be a kernel of the form

$$k\left(\frac{x_{i,1} - x_1}{h_1}\right) \times \cdots \times k\left(\frac{x_{i,\kappa} - x_\kappa}{h_\kappa}\right);$$

a product kernel.

The main problem with multivariate regressors is that the variance of the estimator now becomes inverse-proportional to $n(h_1 \times \cdots \times h_\kappa)$. This implies that the convergence rate decreases with κ. This is known as the curse of dimensionality.

There is a middle-ground between nonparametric and parametric that aims to tackle this issue.

Matching estimators

We wish to infer the average effect of a treatment on an outcome.

If treatment is randomly assigned,

$$\theta = E(y_i|d_i = 1) - E(y_i|d_i = 0)$$

is the average treatment effect.

Estimate the effect by a least-squares regression on a constant and treatment indicator.

Now suppose that treatment is only random conditional on a set of control variables x_i.

Then,

$$\theta = \int \left(E(y_i|d_i = 1, x_i = x) - E(y_i|d_i = 0, x_i = x)\right) f(x)\, dx.$$

Let $m_1(x) = E(y_i|d_i = 1, x_i = x)$ and $m_0(x) = E(y_i|d_i = 0, x_i = x)$. Then

$$n^{-1} \sum_{i=1}^{n} \hat{m}_{1,h}(x_i) - \hat{m}_{0,h}(x_i)$$

is a nonparametric matching estimator of θ.

Matching on the propensity score

Potential problem is limited overlap.

Well-known result says that matching on propensity score,

$$m(x) = P(d_i = 1 | x_i = x),$$

is equivalent.

(Can estimate m nonparametrically and proceed as before but) a convenient alternative follows from observation that

$$\begin{aligned} E(y_i d_i | x_i = x) &= E(y_i | d_i = 1, x_i = x)\, m(x), \\ E(y_i (1-d_i) | x_i = x) &= E(y_i | d_i = 0, x_i = x)\, (1 - m(x)), \end{aligned}$$

so that we can write

$$\theta = E\left(\frac{y_i d_i}{m(x_i)} - \frac{y_i (1-d_i)}{1 - m(x_i)} \right).$$

Still important to have the propensity vary over entire $(0,1)$ for identification.

Regression discontinuity

Now suppose treatment is assigned according to

$$d_i = \begin{cases} 0 & \text{if } x_i < c \\ 1 & \text{if } x_i \geq c \end{cases} ,$$

where the (continuous) running variable x_i cannot be manipulated and c is a known cut-off value.

Identifying assumption is that people around the cut-off are comparable. Can then identify the local treatment effect

$$\theta = \lim_{x \downarrow c} E(y_i | d_i = 1, x_i = x) - \lim_{x \uparrow c} E(y_i | d_i = 0, x_i = x)$$

at the cut-off.

Natural is to fit separate nonparametric regressions to the left and right of cut-off. Using a rectangular kernel, for example, we use observations in the regions $[c - h, c]$ and $[c, c + h]$ only.

Locally-linear estimators are preferable here as they have better properties at the boundary.

Semiparametric binary choice

Consider the binary-choice model

$$y_i = \{x_i'\theta \geq \varepsilon_i\}, \ \varepsilon_i \sim \text{ i.i.d. } F,$$

where, now F, is unknown.

A semiparametric maximum-likelihood estimator is the maximizer of

$$\sum_{i=1}^{n} y_i \log(\check{F}(x_i'\theta)) + (1 - y_i) \log(1 - \check{F}(x_i'\theta))$$

(where we ignore issues of trimming) for

$$\check{F}(x_i'\theta) = \frac{\sum_{j \neq i} k\left(\frac{x_j'\theta - x_i'\theta}{h}\right) y_j}{\sum_{j \neq i} k\left(\frac{x_j'\theta - x_i'\theta}{h}\right)}.$$

Examples of program evaluation

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *American Economic Review* 80, 313–336.

Angrist, J. D. and V. Lavy (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* 114, 533–575.

Card, D. and A. B. Krueger (2000). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* 84, 772–793.

Dehejia, R. H. and S. Wahba (1999). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association* 94, 1053–1062.

Dehejia, R. H. and S. Wahba (1999). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics* 84, 151–161.

Dell, M. (2010). The persistent effects of Peru's mining Mita. *Econometrica* 78, 1863–1903.

Heckman, J. .J., H. Ichimura, and P. E. Todd (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies* 64, 605–654.

Lalonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76, 604–620.