Statistical Inference and Methods

David A. Stephens

Department of Mathematics
Imperial College London

d.stephens@imperial.ac.uk
http://stats.ma.ic.ac.uk/~das01/

13th December 2005

Part II

Session 2: Methods of Inference

Session 2: Methods of Inference

- Frequentist considerations
- Likelihood
- Quasi-likelihood
- Estimating Equations
- Generalized Method of Moments
- Bayesian

Session 2: Methods of Inference

Repeated observation of random variables $X_1, X_2, \ldots, X_n$ yields data $x_1, x_2, \ldots, x_n$.

Parametric Probability Model (pdf) $f_{X|\theta}(x;\theta)$.

Objective is inference about parameter θ, a parameter in p dimensions in parameter space $\Theta \subseteq \mathbb{R}^p$.

We seek a procedure for producing *estimators* of θ that have desirable properties.

Estimators: An **estimator**, T_n, derived from a random sample of size n is a statistic, any function of the random variables to be observed $X_1, \ldots, X_n$:

$$T_n = t(X_1, \ldots, X_n)$$

An **estimate**, t_n, is a real value determined as the observed value of an estimator by data $x_1, \ldots, x_n$.

$$t_n = t(x_1, \ldots, x_n)$$

We will assess the worth of an estimator and the procedure used to produce it by inspecting its **frequentist** (empirical) properties assuming, for example, **that the proposed model is correct**.

Desirable Properties of Estimators:
Suppose the true model has $\theta = \theta_0$.

- *Unbiasedness*

$$E_{X|\theta_0}[T_n] = \theta_0$$

Asymptotic Unbiasedness

$$\lim_{n\to\infty} E_{X|\theta_0}[T_n] = \theta_0$$

- *Consistency*: As $n \longrightarrow \infty$,

Weak consistency:

$$T_n \xrightarrow{p} \theta_0$$

Strong consistency:

$$T_n \xrightarrow{a.s.} \theta_0$$

Laws of Large Numbers Under regularity conditions, for function g, as $n \longrightarrow \infty$,

Weak Law:

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{p} E_{X|\theta_0}[g(X)]$$

Strong Law:

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{a.s.}\xrightarrow{p} E_{X|\theta_0}[g(X)]$$

- For unbiased/asymptotically unbiased (but inconsistent) estimators, an estimator $T_n^\star$ is *efficient* if it has smaller **variance** than all other unbiased estimators.

Efficiency

$$Var_{X|\theta_0}[T_n^\star] \leq Var_{X|\theta_0}[T_n]$$

Asymptotic Efficiency

$$\lim_{n\to\infty} Var_{X|\theta_0}[T_n^\star] \leq \lim_{n\to\infty} Var_{X|\theta_0}[T_n]$$

In summary, an estimator should have good frequentist properties

- it should recover the true value of the parameter as the sample size becomes infinite (*consistency*)
- if an estimator is inconsistent, it may be at least asymptotically unbiased, in which case the asymptotic distribution should have low variance (*efficient*)

However, consistency/asymptotic unbiasedness are not in themselves desirable ...

EXAMPLE: Let $X_1, \ldots, X_n \sim N(\theta, 1)$.
Then the two estimators

$$T_{1n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

$$T_{2n} = T_{1n} + \frac{100^{100}}{n}$$

are both consistent and asymptotically unbiased for θ, with the same asymptotic variance.

However, their finite sample behaviours are somewhat different ...

Finite sample behaviour is also crucial. Could consider

- Sampling distribution of T_n for finite n, that is the empirical behaviour T_n over different random samples of size n
- an asymptotic approximation to this distribution suitable for large n

For example, we typically construct an Asymptotic Normal approximation to this distribution

$$T_n \sim AN(\mu_n, \sigma_n^2)$$

for suitable values of μ_n and σ_n.

The **Standard error** of an estimator T_n of parameter θ is

$$s.e.\left(T_n;\theta\right) = \sqrt{Var_{f_{X|\theta}}\left[T_n\right]} = s_e(\theta)$$

for some function s_e.

The **estimated standard error** is

$$e.s.e\left(T_n\right) = s_e\left(\widehat{\theta}_n\right)$$

Likelihood Methods

We seek a general method for producing estimators/estimates from data under a presumed model that utilizes the observed information in the most effective fashion.

The likelihood function:

$$L(\theta; x) = f_{X|\theta}(x_1, \ldots, x_n; \theta)$$

and under independence

$$L(\theta; x) = \prod_{i=1}^{n} f_{X|\theta}(x_i; \theta)$$

The log-likelihood function:

$$l(\theta; x) = \sum_{i=1}^{n} \log f_{X|\theta}(x_i; \theta)$$

Objective: Inference about θ via L or l

Assertion:

The likelihood contains all relevant information about parameter θ represented by the data.

Maximum Likelihood: Estimate θ by $\widehat{\theta}_n = t(x_1, \ldots, x_n)$

$$\widehat{\theta}_n(x_1, \ldots, x_n) = \arg\max_{\theta \in \Theta} l(\theta; x)$$

with corresponding estimator

$$\widehat{\theta}_n(X_1, \ldots, X_n)$$

Maximum likelihood estimate/estimator (mle) $\widehat{\theta}_n$ is often computed as a zero-crossing of the first derivative of $l(\theta, x)$.

Let

$$\dot{l}(\theta; x) = \nabla l(\theta; x) = \left[\frac{\partial l}{\partial \theta_1}, \ldots, \frac{\partial l}{\partial \theta_p}\right]^\mathsf{T}$$

be the vector of first partial derivatives. Then $\widehat{\theta}_n$ solves

$$\dot{l}(\theta; x) = 0.$$

Score function:

$$S_\theta(X) = \dot{l}(\theta; X).$$

Note: in many models

$$E_{X|\theta}[S_\theta(X)] = 0.$$

Hessian Matrix:

$$H(\theta; x) = \left[\ddot{l}(\theta; x)\right]_{ij} = \frac{\partial^2 l}{\partial \theta_i \theta_j}$$

be the $p \times p$ matrix of second partial derivatives.

Define

$$\Psi_\theta^A(X) = -\ddot{l}(\theta; X).$$

Consider also

$$\Psi_\theta^B(X) = S_\theta(X) S_\theta(X)^\mathsf{T}.$$

Then for many models

$$E_{X|\theta}[\Psi_\theta^A(X)] = E_{X|\theta}[\Psi_\theta^B(X)] = n\mathcal{I}(\theta)$$

where $\mathcal{I}(\theta)$ is the unit *Fisher Information* for the model.

$\mathcal{I}(\theta)$ is a positive definite/non-singular and symmetric matrix. Let

$$\mathcal{J}(\theta) = \mathcal{I}(\theta)^{-1}.$$

We can consider sample-based versions of these quantities
Observed Score:

$$S_\theta(x) = \dot{l}(\theta; x).$$

Observed Unit Information:

$$I_n^A(n, \theta) = \frac{1}{n} \sum_{i=1}^{n} \Psi_\theta^A(x_i)$$

or

$$I_n^B(n, \theta) = \frac{1}{n} \sum_{i=1}^{n} S_\theta(x_i) S_\theta(x_i)^\mathsf{T}$$

Note: by the Laws of Large Numbers, as $n \longrightarrow \infty$,

$$I_n^A(n, \theta_0) \xrightarrow{p} \mathcal{I}(\theta_0) \qquad I_n^B(n, \theta_0) \xrightarrow{p} \mathcal{I}(\theta_0)$$

Cramer-Rao Efficiency Bound

An efficiency bound for unbiased estimators: if T_n is unbiased, then under regularity conditions,

$$Var_{X|\theta}[T_n] \geq \left[E_{X|\theta}[\Psi_\theta^A(X)]\right]^{-1} = \left[E_{X|\theta}[\Psi_\theta^B(X)]\right]^{-1}$$

This is the *Cramer-Rao Lower Bound*.

Properties of mles

Under regularity conditions, the mle is

- consistent
- asymptotically unbiased
- asymptotically efficient, with asymptotic variance $\mathcal{J}(\theta_0)$ equal to the Cramer-Rao lower bound.
- invariant: if $\widehat{\theta}_n$ estimates θ, and $\phi = g(\theta)$, then $\widehat{\phi}_n = g(\widehat{\theta}_n)$.

Asymptotic Normality

Using the CLT,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathfrak{L}} N(0, \mathcal{J}(\theta_0))$$

or

$$\widehat{\theta}_n \sim AN(\theta_0, n^{-1}\mathcal{J}(\theta_0))$$

i.e. $\widehat{\theta}_n$ converges to θ_0 at rate $\sqrt{n}$.

Hypothesis Testing & Confidence Intervals

We seek a general method for testing a specific *hypothesis* about parameters in probability models.

$$\begin{aligned} H_0 &: \theta = c_0 \\ H_1 &: \theta \neq c_0 \end{aligned}$$

There are five crucial components to a *hypothesis test*, namely

- *Test Statistic*, T_n
- *Null Distribution*, distribution of T_n under H_0.
- *Critical Region*, $\mathcal{R}$, and *Critical Value(s)*(C_{R_1}, C_{R_2})

$$T_n \in \mathcal{R} \quad \Longrightarrow \quad \text{Reject } H_0$$

- *Significance Level*, α,

$$\alpha = P[T_n \in \mathcal{R} | H_0].$$

- *P-Value*, p,

$$p = P[|T_n| \geq |t(x)||H_0].$$

The Likelihood Ratio Test

The **Likelihood Ratio Test** statistic for testing H_0 against H_1 is

$$T_n = \frac{\sup\limits_{\theta \in \Theta_1} f_{X|\theta}(X;\theta)}{\sup\limits_{\theta \in \Theta_0} f_{X|\theta}(X;\theta)}$$

where H_0 is rejected if T_n is too large, that is, if $\mathrm{P}[T_n \geq k | H_0] = \alpha$.

If H_0 imposes q independent constraints on H_1, then, as $n \longrightarrow \infty$

$$2 \log T_n \overset{\mathcal{A}}{\sim} \chi^2_q. \tag{1}$$

The Rao/Score/Lagrange Multiplier Test
The **Rao/Score/Lagrange Multiplier** statistic, R_n, for testing

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}$$

is defined by

$$R_n = Z_n^{\mathsf{T}} \left[\mathcal{I}(\theta_0)\right]^{-1} Z_n \tag{2}$$

where

$$Z_n = \frac{1}{\sqrt{n}} \dot{l}(X; \theta_0).$$

For large n, if H_0 is true,

$$R_n \stackrel{\mathcal{A}}{\sim} \chi_p^2$$

and H_0 is rejected if R_n is too large, that is, if $R_n \geq C$, and where

$$P\left[R_n \geq C | H_0\right] = \alpha$$

for significance level α.

Interpretation and Explanation: The score test uses these results; if H_0 is true,

$$S_{\theta_0}(X) \stackrel{\mathcal{A}}{\sim} N\left(0, n\mathcal{I}(\theta_0)\right)$$

so that the standardized score

$$V_n = L_{\theta_0}^{-1} S_{\theta_0}(X) \;\; \stackrel{\mathcal{A}}{\sim} \;\; N\left(0, \mathbf{I}_p\right)$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix, and where matrix $A(\theta)$ is given by

$$L_{\theta_0} L_{\theta_0}^{\mathsf{T}} = n\mathcal{I}\left(\theta_0\right).$$

Hence, by the usual normal distribution theory

$$R_n = V_n^{\mathsf{T}} V_n = Z_n^{\mathsf{T}} \left\{\mathcal{I}(\theta_0)\right\}^{-1} Z_n \stackrel{\mathcal{A}}{\sim} \chi_p^2$$

so that observed test statistic

$$r_n = z_n^{\mathsf{T}} \left\{\mathcal{I}(\theta_0)\right\}^{-1} z_n \qquad \text{where } z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_{\theta_0}(x_i)$$

should be an observation from a χ_p^2 distribution.

Extension: It is legitimate, if required, to replace $\mathcal{I}(\theta_0)$ by a suitable estimate $\hat{I}_n(\tilde{\theta}_n)$. For example

$$\hat{I}_n(\tilde{\theta}_n) = \begin{cases} \mathcal{I}(\tilde{\theta}_n) \\ \\ I_n^A(n, \tilde{\theta}_n) \\ \\ I_n^B(n, \tilde{\theta}_n) \end{cases} \tag{3}$$

The Wald Test

The **Wald Test** statistic, W_n, for testing H_0 against $H_1 : \theta \neq \theta_0$ is defined by

$$W_n = \sqrt{n}\left(\tilde{\theta}_n - \theta_0\right)^{\mathsf{T}} \left[\hat{I}_n(\tilde{\theta}_n)\right] \sqrt{n}\left(\tilde{\theta}_n - \theta_0\right) \tag{4}$$

Then, for large n, if H_0 is true,

$$W_n \overset{\mathcal{A}}{\sim} \chi_p^2$$

and H_0 is rejected if W_n is too large, that is, if $W_n \geq C$, and where $\mathrm{P}[W_n \geq C | H_0] = \alpha$ for significance level α.

Interpretation and Explanation: the logic of the Wald test depends on the asymptotic Normal distribution of the score equation derived estimates

$$\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \mathcal{I}(\theta_0)^{-1}\right)$$

so that

$$\tilde{\theta}_n \overset{\mathcal{A}}{\sim} N\left(\theta_0, \mathcal{I}(\theta_0)^{-1}\right)$$

Again, estimates of the Fisher Information such as those in (3) can be substituted for $I(\theta_0)$ in (4).

Extension to tests for components of θ.

The theory above concerns tests for the whole parameter vector θ. Often it is of interest to consider components of θ, that is, if $\theta = (\theta_1, \theta_2)$, we might wish to test

$$\begin{aligned} H_0 &: \theta_1 = \theta_{10}, \text{ with } \theta_2 \text{ unspecified} \\ H_1 &: \theta_1 \neq \theta_{10}, \text{ with } \theta_2 \text{ unspecified} \end{aligned}$$

The Rao Score and Wald tests can be developed to allow for testing in this slightly different context.

Suppose that θ_1 has dimension m and θ_2 has dimension $p-m$. Let the Fisher information matrix $\mathcal{I}(\theta)$ and its inverse be partitioned

$$\mathcal{I}(\theta) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$$

$$\mathcal{J}(\theta) = \begin{bmatrix} [I_{11.2}]^{-1} & -[I_{11.2}]^{-1} I_{12} [I_{22}]^{-1} \\ -[I_{22.1}]^{-1} I_{21} [I_{11}]^{-1} & [I_{22.1}]^{-1} \end{bmatrix}$$

be a partition of the information matrix, where

$$I_{11.2} = I_{11} - I_{12} [I_{22}]^{-1} I_{21}$$

$$I_{22.1} = I_{22} - I_{21} [I_{11}]^{-1} I_{12}$$

and all quantities depend on θ.

- The **Rao/score/LM** statistic is given by

$$R_n = Z_{n0}^{\mathsf{T}} \left[\hat{I}_n(\tilde{\theta}_n^{(0)})\right]^{-1} Z_{n0} \overset{\mathcal{A}}{\sim} \chi_m^2$$

where $\tilde{\theta}_n^{(0)}$ is the estimate of θ under H_0 and $\hat{I}_n(\tilde{\theta}_n^{(0)})$ is the estimated Fisher information I_1, evaluated at $\tilde{\theta}_n^{(0)}$, obtained using any of the estimates in (3).

- The **Wald** statistic is given by

$$W_n = \sqrt{n}(\tilde{\theta}_{n1} - \theta_{10})^{\mathsf{T}} \left[\hat{I}_n^{(11.2)}(\tilde{\theta}_n)\right] \sqrt{n}\left(\tilde{\theta}_{n1} - \theta_{10}\right) \overset{\mathcal{A}}{\sim} \chi_m^2$$

where $\tilde{\theta}_{n1}$ is the vector component of $\tilde{\theta}_n$ corresponding to θ_1 under H_1, and $\hat{I}_n^{(11.2)}(\tilde{\theta}_n)$ is the estimated version of $I_{11.2}$ (using the sample data, under H_1) evaluated at $\tilde{\theta}_n$, obtained using any of the estimates in (3).

Confidence Intervals

A $100(1-\gamma)\%$ confidence interval (CI) $\mathcal{C}(X)$ for parameter θ is an interval such that

$$P[\theta \in \mathcal{C}(X)] = 1-\gamma$$

under assumptions made about $X = (X_1, X_2, \ldots, n)$ from model $f_{X|\theta}$. In most cases this corresponds to an interval $\mathcal{C}(X) \equiv (L(X), U(X))$ such that

$$P[L(X) \leq \theta \leq U(X)] = 1-\gamma$$

under $f_{X|\theta}$.

Notice that $\mathcal{C}(X)$ is a *random* interval that can be estimated for real data x by $\mathcal{C}(x)$.

Example: If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, then

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0,1)$$

so that, under this model, if $\gamma = 0.05$,

$$P\left[-1.96 \leq \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \leq 1.96\right] = 1 - \gamma$$

and a $100(1-\gamma)\%$ CI is given by

$$L(X) = \overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \qquad U(X) = \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

Note: *Connection with Testing*

Under $H_0 : \theta = \theta_0$, for test statistic T_n and Critical Region $\mathcal{R}$,

$$\alpha = P[T_n \in \mathcal{R} | H_0].$$

Typically, T_n is a *pivotal quantity* whose form depends on θ, but whose distribution does not.

It can be shown that the $100(1-\gamma)\%$ CI is the range of values of θ_0 that can be hypothesized under H_0 such that the hypothesis **is not rejected** at significance level γ.

The *coverage probability* of any random interval $\mathcal{C}(X)$ is the probability

$$P[\theta \in \mathcal{C}(X)]$$

computed under the true model $f_{X|\theta}$.

Thus the coverage probability for a true $100(1-\gamma)\%$ CI is $(1-\gamma)$.

In complicated estimation problems, confidence intervals and coverage probabilities are typically verified using simulation.

Quasi-Likelihood

Quasi-likelihood (QL) methods were introduced to extend the models that can be fitted to data.

The origin of QL methods lie in the attempts to extend the normal linear model to non-normal data, that is, to extend to

Generalized Linear Models

We again begin with a parametric probability model, $f_{Y|\theta}$.

Exponential Family of Distributions : Suppose

$$f_{Y|\xi}(y;\xi) = \exp\left\{\frac{a_\xi(\xi)b_\xi(y) - c_\xi(\xi)}{\phi} + d(y,\phi)\right\}$$

or equivalently, in *canonical* form, writing $\theta = a_\xi(\xi)$, we have

$$f_{Y|\theta}(y;\theta) = \exp\left\{\frac{\theta b(y) - c(\theta)}{\phi} + d(y,\phi)\right\}$$

Without loss of generality, we assume $b(y) = y$. Then

$$E_{Y|\theta}[Y] = \dot{c}(\theta) = \mu \qquad Var_{Y|\theta}[Y] = \phi\ddot{c}(\theta) = \phi V(\mu),$$

say, that is, expectation and variance are functionally related.

A slight generalization allows different data points to be weighted by potentially different weights, w_i, that is, the likelihood becomes

$$f_{Y|\theta}(y_i;\theta) = \exp\left\{w_i\frac{\theta y_i - c(\theta)}{\phi} + d(y_i,\phi/w_i)\right\}$$

so that w_i is a known constant that changes the scale of datum i. Then

$$E_{Y|\theta}[Y] = \mu \qquad Var_{Y|\theta}[Y] = \phi V(\mu)/w.$$

A Generalized Linear Model is a model such that the expectation is modelled as a function of predictors X, that is

$$\mu = \dot{c}(\theta) = g^{-1}(X\beta)$$

for some *link function*, g, a monotone function onto $\mathbb{R}$. The *canonical link* is the link such that

$$g(\dot{c}(\theta)) = \theta.$$

The term $X\beta$ is the *linear predictor*.

For an exponential family GLM, the log-likelihood in the canonical parameterization is

$$l(\beta;y) \quad = \quad constant + \sum_{i=1}^{n}\left\{w_i\frac{\theta_i y_i - c(\theta_i)}{\phi} + d(y_i,\phi/w_i)\right\}$$

Partial differentiation with respect to β_j yields a *score equation* :

$$\frac{\partial l(\beta;y)}{\partial\beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} w_i\frac{\partial\theta_i}{\partial\beta_j}(y_i - \dot{c}(\theta_i)) = \frac{1}{\phi}\sum_{i=1}^{n} w_i\frac{\partial\theta_i}{\partial\beta_j}(y_i - \mu_i)$$

But, with link function g, we have

$$g(\mu_i) = g(\dot{c}(\theta_i)) = X_i\beta = \eta_i$$

thus

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial g(\dot{c}(\theta_i))}{\partial \beta_j} = \dot{g}(\dot{c}(\theta_i))\ddot{c}(\theta_i)\frac{\partial \theta_i}{\partial \beta_j}$$

and hence, as $\ddot{c}(\theta_i) = V(\mu_i)$,

$$\frac{\partial \eta_i}{\partial \beta_j} = \dot{g}(\mu_i)V(\mu_i)\frac{\partial \theta_i}{\partial \beta_j}.$$

But

$$\frac{\partial \eta_i}{\partial \beta_j} = X_{ij}.$$

Thus we have for the j^{th} $(j = 1, \ldots, p)$ score equation

$$\frac{\partial l(\beta; y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{w_i}{\phi} \frac{(y_i - \mu_i)X_{ij}}{\dot{g}(\mu_i)V(\mu_i)} = 0.$$

where, recall,

$$\mu_i = g^{-1}(X_i\beta).$$

Estimation of $(\beta_1, \ldots, \beta_p)$ can be achieved by solution of this system of equations. Note that ϕ can be omitted from this system as $\phi > 0$ by assumption.

If the canonical link function is used

$$\theta_i = X_i\beta \qquad \Longrightarrow \qquad \frac{\partial \theta_i}{\partial \beta_j} = X_{ij}$$

and the score equations become

$$\frac{\partial l(\beta; y)}{\partial \beta_j} = \sum_{i=1}^{n} w_i(y_i - \mu_i)X_{ij} = 0 \qquad j = 1, \ldots, p.$$

In this model, the assumptions about the specific form of $f_{Y|\theta}$ allowed the construction of the score equations; for different probability models, the different components take different forms:

- $f_{Y|\theta}(y; \theta) \equiv Poisson(\lambda)$
  - canonical parameter $\theta = \log \lambda$,
  - canonical link $g(t) = \log(t)$,
  - $\mu = \lambda = \exp(\theta)$,
  - $V(\mu) = \lambda = \mu$ (so that $V(t) = t$).
  - $w_i = 1$,
  - $\phi = 1$.

- $f_{Y|\theta}(y;\theta) \equiv Binomial(n, \xi)$
  - canonical parameter $\theta = \log(\xi/(1-\xi))$,
  - canonical link $g(t) = \log(t/(1-t))$,
  - $\mu = \xi = \exp(\theta)/(1+\exp(\theta))$,
  - $V(\mu) = \xi(1-\xi) = \mu(1-\mu)$ (so that $V(t) = t(1-t)$).
  - $w_i = n_i$,
  - $\phi = 1$.

Note: y_i presumed to be modelled in proportionate form, that is, if $Z \sim Binomial(n, \xi)$, we model $Y = Z/n$.

- $f_{Y|\theta}(y;\theta) \equiv Normal(\xi, \sigma^2)$
  - canonical parameter $\theta = \xi$,
  - canonical link $g(t) = 1$,
  - $\mu = \xi$,
  - $V(\mu) = 1$ (so that $V(t) = 1$).
  - $w_i = 1$,
  - $\phi = \sigma^2$.

Note: here mean and variance are modelled orthogonally.

Quasi-Likelihood: The score equations are key in the estimation, and are derived directly from the probabilistic assumptions:

$$\sum_{i=1}^{n} w_i \frac{(y_i - \mu_i)X_{ij}}{\dot{g}(\mu_i)V(\mu_i)} = 0 \qquad j = 1, \ldots, p.$$

However, these equations can be used as the basis for estimation **even if they are not motivated by probabilistic modelling**.

We can *propose* forms for $V(\mu_i)$ directly without reference to any specific model. This is the basis of *quasi-likelihood* methods.

Examples:

- $V(\mu_i) = \mu_i^2$, the **constant coefficient of variation model** where
$$\frac{E[Y_i]}{\sqrt{Var[Y_i]}} = \frac{\mu_i}{\phi^{1/2}\mu_i} = \frac{1}{\phi^{1/2}}$$
- $V(\mu_i) = \phi_i\mu_i(1-\mu_i)$ (an **overdispersed binomial** model)
- $V(\mu_i) = \phi_i\mu_i$ (an **overdispersed Poisson** model)
- $V(\mu_i) = \phi_i\mu_i^2$ (an **overdispersed Exponential** model).

Estimating Equations

The quasi-likelihood approach is a special case of a general approach to estimation based on **estimating equations**.

An *estimating function* is a function

$$\mathbf{G}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{G}(\boldsymbol{\theta}, Y_i) = \sum_{i=1}^{n} \mathbf{G}_i(\boldsymbol{\theta}) \tag{5}$$

of the same dimension as $\boldsymbol{\theta}$ for which

$$E[\mathbf{G}(\boldsymbol{\theta})] = \mathbf{0}. \tag{6}$$

The estimating function $\mathbf{G}(\boldsymbol{\theta})$ is a random variable because it is a function of Y. The corresponding *estimating equation* that defines the estimator $\widehat{\boldsymbol{\theta}}$ has the form

$$\mathbf{G}(\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \mathbf{G}_i(\widehat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{7}$$

For inference, the frequentist properties of the estimating function are derived and are then transferred to the resultant estimator. The estimating function may be constructed to be a simple function of the data, while the estimator of the parameter that solves (7) will often be unavailable in closed form.

The estimating function (5) is a sum of random variables which provides the opportunity to evaluate its asymptotic properties via a central limit theorem. The *art* of constructing estimating functions is to make them dependent on distribution-free quantities, for example, the population moments of the data.

The following theorem that forms the basis for asymptotic inference.

Theorem: Estimator $\widehat{\boldsymbol{\theta}}_n$ which is the solution to

$$\mathbf{G}(\widehat{\boldsymbol{\theta}}_n) = \sum_{i=1}^{n} \mathbf{G}_i(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0},$$

has asymptotic distribution

$$\widehat{\boldsymbol{\theta}}_n \quad \overset{\mathcal{A}}{\sim} \quad N\left(\boldsymbol{\theta}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{\mathsf{T}-1}\right),$$

where

$$\begin{aligned} \mathbf{A} &= \mathbf{A}_n(\boldsymbol{\theta}) = E\left[\frac{\partial \mathbf{G}}{\partial \boldsymbol{\theta}^{\mathsf{T}}}\right] = \sum_{i=1}^{n} E\left[\frac{\partial \mathbf{G}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathsf{T}}}\right] \\ \mathbf{B} &= \mathbf{B}_n(\boldsymbol{\theta}) = Cov(\mathbf{G}) = \sum_{i=1}^{n} Cov\{\mathbf{G}_i(\boldsymbol{\theta})\}. \end{aligned}$$

The form of the covariance of the estimator here, the covariance of the estimating function, flanked by the inverse of the Jacobian of the transformation from the estimating function to the parameter.

In practice, $\mathbf{A} = \mathbf{A}_n(\boldsymbol{\theta})$ and $\mathbf{B} = \mathbf{B}_n(\boldsymbol{\theta})$ are replaced by $\widehat{\mathbf{A}} = \mathbf{A}_n(\widehat{\boldsymbol{\theta}}_n)$ and $\widehat{\mathbf{B}} = \mathbf{B}_n(\widehat{\boldsymbol{\theta}}_n)$, respectively. In this case, we have

$$\widehat{\boldsymbol{\theta}}_n \overset{\mathcal{A}}{\sim} N_p\left(\boldsymbol{\theta}, \widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^{\mathsf{T}-1}\right), \tag{8}$$

since $\widehat{\mathbf{A}} \xrightarrow{p} \mathbf{A}$ and $\widehat{\mathbf{B}} \xrightarrow{p} \mathbf{B}$.

The accuracy of the asymptotic approximation to the sampling distribution of the estimator is dependent on the parameterization adopted. A rule of thumb is to obtain the confidence interval on a reparameterization which takes the parameter onto the real line (for example, the logistic transform for a probability, or the logarithmic transform for a dispersion parameter), and then to transform to the more interpretable scale.

Estimators for functions of interest, $\phi = g(\boldsymbol{\theta})$, may be obtained via $\widehat{\phi} = g(\widehat{\boldsymbol{\theta}})$, and the asymptotic distribution may be found using the delta method.

Sandwich Estimation
A general method of avoiding stringent modelling conditions when the variance of an estimator is calculated is provided by *sandwich estimation*.

The basic idea is to estimate the variance of the data empirically with minimum modelling assumptions, and to incorporate this in the estimation of the variance of an estimator.

We have seen that when the estimating function corresponds to a score equation derived from a probability model, then *under the model*

$$\mathcal{I} = \mathbf{A} = -\mathbf{B}$$

so that

$$Var(\widehat{\boldsymbol{\theta}}) = \mathbf{A}(\boldsymbol{\theta})^{-1}\mathbf{B}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^{\mathsf{T}-1} = \mathcal{I}(\boldsymbol{\theta})^{-1}.$$

If the model is not correct then this equality does not hold, and the variance estimator will be incorrect.

An alternative is to evaluate the variance *empirically* via

$$\widehat{\mathbf{A}} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}(\widehat{\boldsymbol{\theta}}, Y_i),$$

and

$$\widehat{\mathbf{B}} = \sum_{i=1}^{n} \mathbf{G}(\widehat{\boldsymbol{\theta}}, Y_i) \mathbf{G}(\widehat{\boldsymbol{\theta}}, Y_i)^{\mathsf{T}}.$$

This method is general and can be applied to any estimating function, not just those arising from a score equation.

Suppose we assume $E[\mathbf{Y}] = \boldsymbol{\mu}$ and $Var(\mathbf{Y}) = \phi \mathbf{V}$ with $Var(Y_i) = \phi V(\mu_i)$, and $Cov(Y_i, Y_j) = 0$, $i, j = 1, ..., n$, $i \neq j$, as a *working* covariance model.

Under this specification it is natural to take the quasi-likelihood as an estimating function, in which case

$$Cov\{U(\beta)\} = \mathbf{D}^{\mathsf{T}} \mathbf{V}^{-1} Cov(\mathbf{Y}) \mathbf{V}^{-1} \mathbf{D} / \phi^2$$

to give

$$Var_s(\widehat{\beta}) = (\mathbf{D}^{\mathsf{T}} \mathbf{V}^{-1} \mathbf{D})^{-1} \mathbf{D}^{\mathsf{T}} \mathbf{V}^{-1} Cov(\mathbf{Y}) \mathbf{V}^{-1} \mathbf{D} (\mathbf{D}^{\mathsf{T}} \mathbf{V}^{-1} \mathbf{D})^{-1},$$

and so the appropriate variance is obtained by substituting in the correct form for $Cov(\mathbf{Y})$ which is, of course, unknown.

However, a simple "sandwich" estimator of the variance is given by

$$Var_s(\widehat{\boldsymbol{\beta}}) = (\mathbf{D}^{\mathsf{T}} \mathbf{V}^{-1} \mathbf{D})^{-1} \mathbf{D}^{\mathsf{T}} \mathbf{V}^{-1} \mathbf{R}^{\mathsf{T}} \mathbf{R} \mathbf{V}^{-1} \mathbf{D} (\mathbf{D}^{\mathsf{T}} \mathbf{V}^{-1} \mathbf{D})^{-1},$$

where $\mathbf{R} = (R_1, ..., R_n)^{\mathsf{T}}$ is the $n \times 1$ vector with $R_i = Y_i - \mu_i(\widehat{\boldsymbol{\beta}})$.

This estimator is consistent for the variance of $\widehat{\boldsymbol{\beta}}$, under correct specification of the mean, and with uncorrelated data. There is finite sample bias in R_i as an estimate of $Y_i - \mu_i(\boldsymbol{\beta})$ and versions that adjust for the estimation of the parameters $\boldsymbol{\beta}$ are also available

The great advantage of sandwich estimation is that it provides a consistent estimator of the variance in very broad situations. There are two things to bear in mind

- For small sample sizes, the sandwich estimator *may be highly unstable*, and in terms of mean squared error model-based estimators may be preferable for small to medium sized n; *empirical* is a better description of the estimator than *robust*.
- If the model is correct, then the model-based estimators are more *efficient*.

Generalized Estimating Equations

The models above focus on independent data only. However, the methods can be extended to the dependent data cases.

Recall the Normal Linear (regression) model

$$Y = X\beta + \epsilon$$

where

- Y is $n \times 1$,
- X is $n \times p$,
- β is $p \times 1$,
- ϵ is $n \times 1$,

and $\epsilon \sim N(0, \sigma^2 \mathbf{I}_n)$ for identically distributed errors.

More generally, we can assume $\epsilon \sim N(0, \Sigma)$ and model the observable quantities as *dependent*

- repeated observations on a series of N experimental units that are modelled independently, so that Σ is block diagonal:

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_N \end{bmatrix}$$

- correlated data from one stochastic process

In this model, we have the ML (and GLS) estimates (conditional on Σ) given by

$$\widehat{\beta}_n = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} y$$

and it follows that $\widehat{\beta}_n$ is unbiased, and Normally distributed with variance

$$(X^\top \Sigma^{-1} X)^{-1}$$

Typically Σ is not known, and possibly contains unknown parameters, α. It can be estimated in a number of ways

- ML (distribution-based)
- REML (distribution-based)
- Robust (sandwich) estimation (model-free)

An estimating equation approach can be used to view this form of estimation in a distribution-free context. We consider the *Generalized Estimating Equation* given by

$$\mathbf{G}(\beta) = X^\top W^{-1}(y - X\beta)$$

for symmetric, non-singular matrix W (that is, a matrix version of the independent case given above). Then $E[\mathbf{G}(\beta)] = 0$, and

$$\widehat{\beta}_W = (X^\top W^{-1}X)^{-1}X^\top W^{-1}y$$

is unbiased with

$$Var(\widehat{\beta}_W) = (X^\top W^{-1}X)^{-1}X^\top W^{-1}\Sigma W^{-1}X(X^\top W^{-1}X)^{-1}$$

Could choose

- $W = \Sigma$, so that

$$Var(\widehat{\beta}_W) = (X^\top \Sigma^{-1}X)^{-1}$$

- $W = \mathbf{I}$, so that

$$Var(\widehat{\beta}_W) = (X^\top X)^{-1}X^\top \Sigma X(X^\top X)^{-1}.$$

We still need to estimate $\Sigma = \Sigma(\alpha)$.

We focus on the repeated measures case, where **independent** units $i = 1, \ldots, N$ has $n_1, \ldots, n_N$ observations.

Suppose that $\Sigma_i = W_i$, a known constant "working" covariance matrix. Then we have

$$\widehat{\beta}_W = \left(\sum_{i=1}^{N} X_i^\top W_i^{-1}X_i\right)^{-1}\left(\sum_{i=1}^{N} X_i^\top W_i^{-1}y_i\right)$$

with

$$Var(\widehat{\beta}_W) = (X^\top W^{-1}X)^{-1} = \left(\sum_{i=1}^{N} X_i^\top W_i^{-1}X_i\right)^{-1}$$

If $W = W(\alpha, \beta)$, then the corresponding estimating function is

$$\mathbf{G}(\alpha, \beta) = \sum_{i=1}^{N} X_i^\top W_i^{-1}(\alpha, \beta)(y_i - X_i\beta)$$

If α is consistently estimated by $\widehat{\alpha}_n$, then we can substitute in and leave the estimating function

$$\mathbf{G}(\beta) = \sum_{i=1}^{N} X_i^\top W_i^{-1}(\widehat{\alpha}_n, \beta)(y_i - X_i\beta)$$

and $\widehat{\beta}_W$ can be found using the usual iterative schemes.

In this case, the estimated variance of $\widehat{\beta}_W$ is given by

$$\begin{aligned}\widehat{Var}(\widehat{\beta}_W) &= \left(\sum_{i=1}^{N} X_i^\top W_i^{-1}(\widehat{\alpha}_n, \widehat{\beta}_n) X_i\right)^{-1} \\ &\times \left(\sum_{i=1}^{N} X_i^\top W_i^{-1}(\widehat{\alpha}_n, \widehat{\beta}_n)\widehat{\Sigma}_i W_i^{-1}(\widehat{\alpha}_n, \widehat{\beta}_n) X_i\right) \\ &\times \left(\sum_{i=1}^{N} X_i^\top W_i^{-1}(\widehat{\alpha}_n, \widehat{\beta}_n) X_i\right)^{-1}\end{aligned}$$

where it still remains to estimate Σ by $\widehat{\Sigma}$.

We use the estimate based on the quantities

$$(y_i - X_i\widehat{\beta}_n)(y_i - X_i\widehat{\beta}_n)^\top \qquad i = 1, \ldots, N.$$

For example, for a balanced design (all n_i equal to M), with common covariances, for equally-spaced data, we estimate

$$[\Sigma_i]_{jj} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - X_{ij}\widehat{\beta}_n)^2$$

and

$$[\Sigma_i]_{jk} = \frac{1}{N}\sum_{i=1}^{N}(y_{ij} - X_{ij}\widehat{\beta}_n)(y_{ik} - X_{ik}\widehat{\beta}_n).$$

The model can be extended by the inclusion of a link function h such that

$$\mu_i = h(X_i\beta)$$

in which case the estimating function is

$$\mathbf{G}(\beta) = \sum_{i=1}^{N} D_i^\top W_i^{-1}(\alpha, \beta)(y_i - X_i\beta)$$

where D_i is the $n_i \times p$ matrix of partial derivatives.

$$[D_i]_{jk} = \frac{\partial \mu_{ij}}{\partial \beta_k}$$

for $j = 1, \ldots, n_i$, $k = 1, \ldots, p$.

Summary: GEE given by estimating function

$$\mathbf{G}(\alpha, \beta) = \sum_{i=1}^{N} D_i^\top W_i^{-1}(y_i - \mu_i)$$

where

- $\mu_i = h(X_i\beta)$
- $D_i = \dfrac{\partial \mu_i}{\partial \beta^\top} = X_i^\top$
- W_i is a working covariance model.

- $\widehat{\epsilon} = y_i - \widehat{\mu}_i = y_i - h(X_i\widehat{\beta})$
- $\widehat{D}_i$ is D_i evaluated at $\widehat{\mu}_i$.
- $\widehat{A}$ given by

$$\widehat{A} = \sum_{i=1}^{N} \widehat{D}_i^{\mathsf{T}} W_i^{-1} \widehat{D}_i$$

- $\widehat{B}$ given by

$$\widehat{B} = \left(\sum_{i=1}^{N} \widehat{D}_i^{\mathsf{T}} W_i^{-1} \widehat{\epsilon}_i \widehat{\epsilon}_i^{\mathsf{T}} W_i^{-1} \widehat{D}_i \right)$$

- the variance of $\widehat{\beta}$ is $\widehat{A}^{-1}\widehat{B}\widehat{A}^{-1}$

Generalized Method of Moments

The *Generalized Method of Moments* (GMM) approach to estimation is designed to produce estimates that satisfy *moment conditions* that are appropriate in the context of the modelling situation.

It is an extension to conventional method of moments (MM) in which theoretical and empirical moments are matched.

See Hall (2005), *Generalized Method of Moments*, Oxford Advanced Texts in Econometrics.

Example: $X_1, \ldots, X_n \sim Normal(\mu, \sigma^2)$, independent. We have

$$E_{X|\theta}[X] = \mu \qquad E_{X|\theta}[X^2] = \sigma^2.$$

We equate to the first two empirical moments

$$m_1 = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

yielding equations for estimation

$$m_1 = \mu \qquad m_2 = \mu^2 + \sigma^2,$$

or equivalently

$$\begin{aligned} m_1 - \mu &= 0 \\ m_2 - (\mu^2 + \sigma^2) &= 0 \end{aligned} \qquad (9)$$

Example: $X_1, \ldots, X_n \sim Gamma(\alpha, \beta)$, independent. We have

$$E_{X|\theta}[X] = \frac{\alpha}{\beta} \qquad E_{X|\theta}[X^2] = \frac{\alpha(\alpha+1)}{\beta^2}.$$

We equate to the first two empirical moments

$$m_1 = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

yielding

$$\widehat{\alpha}_n = \frac{m_1^2}{m_2 - m_1^2} \qquad \widehat{\beta}_n = \frac{m_1}{m_2 - m_1^2}$$

A problem with this approach is that typically p is finite (that is, we have a finite number of parameters to estimate), and (often) an infinite number of moments to select from; for example, we could use

$$E_{X|\theta}[X^3] = \frac{\alpha(\alpha+1)(\alpha+2)}{\beta^3} \qquad E_{X|\theta}[X^4] = \frac{\alpha(\alpha+1)(\alpha+2)(\alpha+3)}{\beta^4}$$

as two equations to estimate α and β.

That is, the MM estimator is not uniquely defined.

Econometric Model Suppose, for $t = 1, 2, \ldots,$, we have

$$\begin{aligned} y_t^D &= \alpha_0 x_t + \epsilon_t^D \\ y_t^S &= \beta_{01} n_t + \beta_{02} x_t + \epsilon_t^S \\ y_t^D &= y_t^S (= y_t, \text{ say}) \end{aligned} \tag{10}$$

where, in year t,

- y_t^D is the Demand,
- y_t^S is the Supply,
- x_t is the price,
- n_t is a factor influencing supply.

We wish to estimate α_0, given pairs $(x_t, y_t), t = 1, \ldots, T$.

OLS **does not work effectively** in this model; the estimates are typically biased.

Instead suppose that there is an **observable** variable z_t^D related to x_t, but so that

$$Cov[z_t^D, \epsilon_t^D] = 0$$

e.g. any of the factors that affect supply, n_t.

It is typical to assume that $E[\epsilon_t^D]$, so that

$$E[y_t^D] = \alpha_0 E[x_t].$$

Then taking expectations through equation (10) we have the following relationship

$$E[z_t^D y_t] - \alpha_0 E[z_t^D x_t] = 0 \tag{11}$$

and thus an MM estimate is

$$\widehat{\alpha}_T = \frac{\displaystyle\sum_{t=1}^{T} z_t^D y_t}{\displaystyle\sum_{t=1}^{T} z_t^D x_t}.$$

GMM proceeds as follows: we define a *population moment condition* via vector function g as

$$E_{V|\theta}[g(v_t, \theta)] = 0.$$

For example, in the Normal example above, from equation (9), we have

$$g(V_t, \theta) = \begin{bmatrix} v_t - \mu \\ v_t^2 - \mu - \sigma^2 \end{bmatrix}$$

In the econometric supply/demand model (11) we have

$$g(v_t, \theta) = z_t^D y_t - \alpha_0 z_t^D x_t$$

so that $v_t = (z_t^D, y_t, x_t)^\top$, and $\theta = \alpha_0$.

The **Generalized Method of Moments (GMM) Estimator** of parameter θ based on q *moment conditions*

$$E[g(V_t, \theta)] = 0.$$

is given as the value of θ that minimizes

$$Q_T(\theta) = \overline{g}_T(\theta)^\top W_T \overline{g}_T(\theta)$$

where

$$\overline{g}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} g(v_t, \theta)$$

and W_T is a positive semidefinite matrix such that $W_T \xrightarrow{p} W$, a constant positive definite matrix, as $T \longrightarrow \infty$.

Note: in general $q \geq p$.

- If $q = p$, then the system is *just identified*.
- If $q > p$, then the system is *over-identified*.

Over-identification is what distinguishes GMM from MM.

Some mild regularity conditions are needed to ensure that the estimation procedure works effectively.

Regularity Conditions:

- strict stationarity,
- g is continuous in θ for all v_t, finite expectation that is continuous on Θ,
- q population moment constraints

$$E[g(v_t, \theta_0)] = \mathbf{0} \qquad (q \times 1)$$

- global identification

$$E[g(V_t, \theta^\star)] \neq \mathbf{0} \qquad \theta^\star \neq \theta_0$$

- Conditions on the derivative of g: the $(q \times p)$ matrix of derivatives of g with respect to the elements of θ

$$\frac{\partial g_j}{\partial \theta_k}$$

- θ_0 is an interior point of Θ,
- the expectation matrix

$$E\left[\frac{\partial g}{\partial \theta^\intercal}\right]$$

exists and is finite, and has rank p when evaluated at θ_0.

Recall that the estimator, $\widehat{\theta}_T$ is the value that minimizes

$$Q_T(\theta) = \overline{g}_T(\theta)^\intercal W_T \overline{g}_T(\theta)$$

where

$$\overline{g}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} g(v_t, \theta) \qquad (q \times 1).$$

Under the regularity conditions, we have

$$\overline{D}(v, \widehat{\theta}_T)^\intercal W_T \overline{g}_T(\widehat{\theta}_T) = \mathbf{0} \qquad (p \times 1)$$

where

$$\overline{D}(v, \theta) = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial g(v_t, \theta)}{\partial \theta^\intercal} \qquad (q \times p).$$

Alternative representation: moment condition

$$F(\theta_0)^\intercal W^{1/2} E[g(v_t, \theta_0)] = 0$$

where

$$F(\theta_0) = W^{1/2} E\left[\frac{\partial g(v_t, \theta_0)}{\partial \theta^\intercal}\right]$$

We have $rank\{F(\theta_0)\} = p$.

Identifying Restrictions:

$$F(\theta_0)(F(\theta_0)^\intercal F(\theta_0))^{-1} F(\theta_0)^\intercal W^{1/2} E[g(v_t, \theta_0)] = 0$$

Overidentifying Restrictions:

$$(\mathbf{1}_q - F(\theta_0)(F(\theta_0)^\intercal F(\theta_0))^{-1} F(\theta_0)^\intercal) W^{1/2} E[g(v_t, \theta_0)] = 0$$

Asymptotic Properties: under mild regularity conditions

$$\widehat{\theta}_T \xrightarrow{p} \theta_0$$

(uniformly on Θ), and

$$T^{1/2}(\widehat{\theta}_T - \theta_0) \xrightarrow{\mathfrak{L}} N(0, MSM^{\mathsf{T}})$$

where

$$S = \lim_{T \longrightarrow \infty} Var\left[T^{1/2}\overline{g}_T(\theta_0)\right]$$

and

$$M = (D_0^{\mathsf{T}} W D_0)^{-1} D_0^{\mathsf{T}} W$$

with

$$D_0 = E\left[\frac{\partial g(v_t, \theta_0)}{\partial \theta^{\mathsf{T}}}\right]$$

M can be estimated using

$$\widehat{M}_T = (D_T(\widehat{\theta}_T)^{\mathsf{T}} W_T D_T(\widehat{\theta}_T))^{-1} D_T(\widehat{\theta}_T)^{\mathsf{T}} W_T$$

where

$$D_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial g(v_t, \theta)}{\partial \theta^{\mathsf{T}}}$$

Estimation of S is more complicated; a number of different methods exist to produce an estimate $\widehat{S}$, depending on the context.

Optimal choice of W: It can be shown that the optimal choice of W is S^{-1}, so in the finite sample case we use

$$W_T = \widehat{S}_T^{-1}.$$

In practice, an iterative procedure can be used:

- At step 1, set $W_T = \mathbf{1}_q$. Estimate $\widehat{\theta}_T(1)$, and then $\widehat{S}_T^{-1}(1)$.
- At step $2, 3, \ldots$, set $W_T = \widehat{S}_T^{-1}(i-1)$. Estimate $\widehat{\theta}_T(i)$, and then $\widehat{S}_T^{-1}(i)$.
- Iterate until convergence.

This algorithm typically converges in relatively few steps.

Implementation in the Linear Regression Model.
Consider the model

$$y_t = x_t^{\mathsf{T}}\theta_0 + u_t \qquad t = 1, \ldots, T$$

with *instruments* z_t, where

- x_t is $p \times 1$
- z_t is $q \times 1$.

Define

$$u_t(\theta) = y_t - x_t^{\mathsf{T}}\theta_0$$

Assumptions:

- (Strict) Stationarity
- z_T satisfies the *population moment condition* (PMC)

$$E[z_t u_t(\theta_0)] = 0$$

(an *orthogonality condition*).

Fundamental Decomposition:

$$E[z_t u_t(\theta)] = E[z_t u_t(\theta_0)] + E[z_t x_t^{\intercal}](\theta_0 - \theta)$$

so that, via the PMC

$$E[z_t u_t(\theta)] = E[z_t x_t^{\intercal}](\theta_0 - \theta),$$

so θ_0 is *identified* if

$$E[z_t x_t^{\intercal}](\theta_0 - \theta) \neq \mathbf{0}.$$

Note that this is a linear system; $E[z_t x_t^{\intercal}]$ is a $(q \times p)$ matrix - we need

$$rank\{E[z_t x_t^{\intercal}]\} = p.$$

Estimator: We have (in matrix form)

$$Q_T(\theta) = \left\{T^{-1} U(\theta)^{\intercal} Z\right\} W_T \left\{T^{-1} Z^{\intercal} U(\theta)\right\}$$

where now

- y is $(T \times 1)$,
- X is $(T \times p)$,
- Z is $(T \times q)$,
- U is $(T \times 1)$,

$$U(\theta) = y - X\theta$$

Then

$$\widehat{\theta}_T = \left(\left\{T^{-1} X^{\intercal} Z\right\} W_T \left\{T^{-1} Z^{\intercal} X\right\}\right)^{-1} \left\{T^{-1} X^{\intercal} Z\right\} W_T \left\{T^{-1} Z^{\intercal} y\right\}$$

The minimization is equivalent to solving the system

$$\left\{T^{-1} X^{\intercal} Z\right\} W_T T^{-1} Z^{\intercal} U(\widehat{\theta}_T) = 0.$$

Let

$$F^{\intercal} = E[x_t z_t^{\intercal}](W^{1/2})^{\intercal}$$

then GMM estimation is equivalent to solving

$$F(F^{\intercal} F)^{-1} F^{\intercal} W^{1/2} E[z_t u_t(\theta_0)] = 0$$

which are the identifying conditions in this case.

Asymptotic properties: Let

$$M = (F^\top F)^{-1} F^\top W^{1/2}$$

where $F = W^{1/2} E[z_t x_t^\top]$. Note that

$$M = \left(E[x_t z_t^\top] W E[z_t x_t^\top]\right)^{-1} E[x_t z_t^\top] W$$

Then $\widehat{\theta}_T$ is consistent for θ_0, and

$$T^{1/2}(\widehat{\theta}_T - \theta_0) \xrightarrow{\mathfrak{L}} N(0, MSM^\top)$$

$$S = \lim_{T \longrightarrow \infty} Var\left[T^{-1/2} \sum_{t=1}^{T} z_t u_t\right]$$

and where, in the case of independence across time $S = E[u_t^2 z_t z_t^\top]$.

For practical purposes, the expectations are replaced by empirical averages over the T observations, for example, F is replaced by $\widehat{F}_T$, where

$$\widehat{F}_T = W_T^{1/2} \left\{T^{-1} Z^\top X\right\}$$

and, for example,

$$\widehat{S}_T = \frac{1}{T} \sum_{t=1}^{T} \widehat{u}_t^2 z_t z_t^\top$$

where

$$\widehat{u}_t = y_t - x_t^\top \widehat{\theta}_T.$$

Optimal choice of W: As before the optimal choice is

$$W = S^{-1}$$

and so in estimation

$$W_T = \widehat{S}_T^{-1}.$$

An iterative procedure can again be used:

- set $W_T = \mathbf{1}_q$ or $W_T = (T^{-1} Z^\top Z)^{-1}$ and obtain $\widehat{\theta}_T$ and $\widehat{S}_T$
- set $W_T = \widehat{S}_T^{-1}$

and so on.

Test for mis-specification: Using the asymptotic results, it can be shown that

$$J_T = TQ(\widehat{\theta}_T) = T^{-1} U(\widehat{\theta}_T)^\top Z \widehat{S}_T^{-1} Z^\top U(\widehat{\theta}_T) \xrightarrow{\mathfrak{L}} \chi^2_{q-p}$$

under the null hypothesis

$$H_0 : E[z_t u_t(\theta_0)] = 0.$$

This test (*Sargan's Test*) allows assessment of model mis-specification (i.e. assessment of selected instruments).

Asymptotics also yield tests for individual coefficients (Wald-type tests).

Bayesian Methods

The classical view of Statistical Inference Theory contrasts with the alternative **Bayesian** approach.

In Bayesian theory, the likelihood function still plays a central role, but is combined with a **prior** probability distribution to give a **posterior** distribution for the parameters in the model. Inference, estimation, uncertainty reporting and hypothesis testing can be carried out within the Bayesian framework.

Some Reasons To Be Bayesian

- Inference through Probability (coherence, representations of uncertainty for observables)
- Prediction
- Ease of implementation
- Ease of interpretation
- The Logic of Conditional Probability
- Decision Theory (optimal decision making)

Implementation Issues

- Analytic
- Analytic Approximation
- Numerical I : Numerical Integration
- Numerical II: Simulation and Monte Carlo
- Numerical III: Markov chain Monte Carlo

Key Technical Results

- De-Finetti Representation
- Posterior Asymptotic Normality
- Consistency

Different views of Bayesianism

- Subjectivist
- Objectivist
- Regularizers
- Pragmatist
- Opportunist (post-Bayesian)

SOME REASONS NOT TO BE BAYESIAN
(or rather, issues "to be managed" ...)

- Prior specification
- Computation
- Hypothesis Testing
- Model checking
- Model selection

In the Bayesian framework, inference about an unknown parameter θ is carried out via the **posterior probability distribution** that combines prior opinion about the parameter with the information contained in the likelihood $f_{X|\theta}(x;\theta)$ which represents the data contribution. In terms of events, Bayes Theorem says that

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

that is, it relates the two conditional probabilities $P(A|B)$ and $P(B|A)$.

It follows that we can carry out inference via the conditional probability distribution for parameter θ **given** data $X = x$.

Specifically for parameter θ, the **posterior probability distribution** for θ is denoted $p_{\theta|X}(\theta|x)$, and is calculated as

$$p_{\theta|X}(\theta|x) = \frac{f_{X|\theta}(x;\theta)\, p_\theta(\theta)}{\int f_{X|\theta}(x;\theta)\, p_\theta(\theta)\, d\theta} = c(x) f_{X|\theta}(x;\theta)\, p_\theta(\theta) \quad (12)$$

say, where $f_{X|\theta}(x;\theta)$ is the likelihood, and $p_\theta(\theta)$ is the **prior probability distribution** for θ.

The denominator in (12) can be regarded as the **marginal distribution** (or **marginal likelihood**) for data X evaluated at the observed data x

$$f_X(x) = \int f_{X|\theta}(x;\theta)\, p_\theta(\theta)\, d\theta. \tag{13}$$

Inference for the parameter θ via the posterior $p_{\theta|Y}(\theta|y)$ can be carried out once the posterior has been computed. Intuitively appealing methods rely on summaries of this probability distribution, that is, moments or quantiles. For example, one Bayes estimate, $\widehat{\theta}_B$ of θ is the **posterior expectation**

$$\widehat{\theta}_B = E_{p_{\theta|X}}[\theta|X = x] = \int \theta p_{\theta|X}(\theta|x) d\theta$$

whereas another is the **posterior mode** , $\widehat{\theta}_B$, that is, the value of θ at which $p_{\theta|X}(\theta|x)$ is maximized, and finally the **posterior median** that satisfies

$$\int_{-\infty}^{\widehat{\theta}_B} p_{\theta|X}(\theta|x) d\theta = \frac{1}{2}$$

A $\mathbf{100(1-\alpha)}$ **Bayesian Credible Interval** for θ is a subset C of Θ such that

$$\mathrm{P}[\theta \in C] \geq 1-\alpha$$

The $\mathbf{100(1-\alpha)}$**Highest Posterior Density Bayesian Credible Interval** for θ, subject to $\mathrm{P}[\theta \in C] \geq 1-\alpha$ is a subset C of Θ such that $C = \left\{\theta \in \Theta : p_{\theta|X}(\theta|x) \geq k\right\}$ where k is the largest constant such that

$$\mathrm{P}[\theta \in C] \geq 1-\alpha.$$

Bayesian Inference and Decision Making
Suppose that, in an inference setting, a decision is to be made, and the decision is selected from some set $\mathcal{D}$ of alternatives. Regarding the parameter space Θ as a set of potential "states of nature", within which the "true" state θ lies.

Define the **loss function** for decision d and state θ as the loss (or penalty) incurred when the true state of nature is θ and the decision made is d. Denote this loss as

$$L(d,\theta)$$

With prior $\pi(\theta)$ and no data, the **expected loss** (or the **Bayes loss**) is defined as

$$\mathrm{E}_\theta\left[L(d,\theta)\right] = \int L(d,\theta) p_\theta\left(\theta\right) d\theta$$

The optimal Bayesian decision is

$$d_B = \arg\min_{d \in \mathcal{D}} \mathrm{E}_{p_\theta}\left[L(d,\theta)\right]$$

that is, the decision that minimizes the Bayes loss.

If data are available, the optimal decision will intuitively become a function of the data. Suppose now that the decision in light of the data is denoted $\delta(x)$ (a function from $\mathbb{X}$ to $\mathcal{D}$, and the associated loss is $L(\delta(x),\theta)$)

The **risk** associated with decision $\delta(X)$ is the expected loss associated with $\delta(X)$, with the expectation taken over the distribution of X given θ

$$R_\theta(\delta) = \mathrm{E}_{X|\theta}\left[L(\delta(X),\theta)\right] = \int L(\delta(X),\theta) f_{X|\theta}(x;\theta) dx$$

The **Bayes risk** expected risk $R_\theta(\delta)$ associated with $\delta(X)$, with the expectation taken over the prior distribution of θ

$$\begin{aligned}
R(\delta) &= \mathrm{E}_\theta\left[R_\theta(\delta)\right] = \mathrm{E}_\theta\left[\mathrm{E}_{X|\theta}\left[L(\delta(X),\theta)\right]\right] \\
&= \int \left\{ \int L(\delta(x),\theta) f_{X|\theta}(x;\theta) dx \right\} p_\theta\left(\theta\right) d\theta \\
&= \int \int L(\delta(x),\theta) f_X(x) p_{\theta|X}(\theta|x) dx d\theta \\
&= \int \left\{ \int L(\delta(x),\theta) p_{\theta|X}(\theta|x) d\theta \right\} f_X(x) dx.
\end{aligned}$$

With prior $p_\theta\left(\theta\right)$ and fixed data x, the optimal Bayesian decision, termed the **Bayes rule** is

$$\begin{aligned}
d_B = \arg\min_{\delta \in \mathcal{D}} R(\delta) &= \arg\min_{\delta \in \mathcal{D}} \int \left\{ \int L(\delta(x),\theta) p_{\theta|X}(\theta|x) d\theta \right\} f_X(x) dx \\
&= \arg\min_{\delta \in \mathcal{D}} \int L(\delta(x),\theta) p_{\theta|X}(\theta|x) d\theta
\end{aligned}$$

that is, the decision that minimizes the Bayes risk, or equivalently **(posterior) expected loss** in making decision δ, with expectation taken with respect to the posterior distribution $p_{\theta|X}(\theta|x)$.

Applications of Decision Theory to Estimation
Under squared error loss

$$L(\delta(x),\theta) = (\delta(x)-\theta)^2$$

the Bayes rule for estimating θ is

$$\delta(x) = \widehat{\theta}_B = \mathrm{E}_{p_{\theta|X}}[\theta|x] = \int \theta p_{\theta|X}(\theta|x)d\theta$$

that is, the **posterior expectation**.

Under absolute error loss

$$L(\delta(x),\theta) = |\delta(x)-\theta|$$

the Bayes rule for estimating θ is the solution of

$$\int_{-\infty}^{\delta(x)} p_{\theta|X}(\theta|x)d\theta = \frac{1}{2}$$

that is, the **posterior median**.

Bayesian Hypothesis Testing
To mimic the Likelihood Ratio testing procedure outlined in previous sections. For two hypotheses H_0 and H_1 define

$$\alpha_0 = \mathrm{P}[H_0|X=x] \qquad \alpha_1 = \mathrm{P}[H_1|X=x]$$

For example,

$$\mathrm{P}[H_0|X=x] = \int_R \pi_{\theta|X}(\theta|x)d\theta$$

where R is some region of Θ. Typically, the quantity

$$\frac{\mathrm{P}[H_0|X=x]}{\mathrm{P}[H_1|X=x]}$$

(the **posterior odds** on H_0) is examined.

Example: To test two simple hypothesis

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1 \end{aligned}$$

define the prior probabilities of H_0 and H_1as p_0 and p_1 respectively. Then, by Bayes Theorem

$$\frac{\mathrm{P}[H_1|X=x]}{\mathrm{P}[H_0|X=x]} = \frac{\dfrac{f_{X|\theta}(x;\theta_1)p_1}{f_{X|\theta}(x;\theta_0)p_0 + f_{X|\theta}(x;\theta_1)p_1}}{\dfrac{f_{X|\theta}(x;\theta_0)p_0}{f_{X|\theta}(x;\theta_0)p_0 + f_{X|\theta}(x;\theta_1)p_1}} = \frac{f_{X|\theta}(x;\theta_1)p_1}{f_{X|\theta}(x;\theta_0)p_0}$$

More generally, two hypotheses or models can be compared via the observed marginal likelihood that appears in (13), that is if

$$\frac{f_X(x;\text{Model } 1)}{f_X(x;\text{Model } 0)} = \frac{\int f^{(1)}_{X|\theta}(x;\theta_1)\,p_{\theta_1}(\theta_1)\,d\theta_1}{\int f^{(0)}_{X|\theta}(x;\theta_0)\,p_{\theta_0}(\theta_0)\,d\theta_0}$$

is greater than one we would favour Model 1 (with likelihood $f^{(1)}_{X|\theta}$ and prior p_{θ_1}) over Model 0 (with likelihood $f^{(0)}_{X|\theta}$ and prior p_{θ_0}).

Prediction The Bayesian approach to prediction follows naturally from probability logic. The *posterior predictive distribution* for random variables $X^\star$, given data $X = x$, is computed as

$$f_{X^\star|X}(x^\star|x) = \int f_{X^\star|\theta}(x^\star;\theta)\,p_{\theta|X}(\theta|x)\;d\theta$$

Point predictions and prediction intervals can be computed from this distribution.

The posterior distribution

$$p_{\theta|X}(\theta|x) = \frac{f_{X|\theta}(x;\theta)\,p_\theta(\theta)}{\displaystyle\int f_{X|\theta}(x;\theta)\,p_\theta(\theta)\,d\theta}$$

is a joint probability distribution in $\mathbb{R}^p$. Computation of posterior summaries, estimates etc. typically requires an integral in a high dimension. This can prove problematic if the likelihood prior combination is not analytically tractable.

When $p_{\theta|X}(\theta|x)$ is not a standard multivariate distribution, integrals with respect to $p_{\theta|X}$ can be approximated in a number of ways:

- numerical integration,
- analytic approximation,
- Monte Carlo/Importance sampling.

In high dimensions, such methods can prove inaccurate.

Simulation-based inference: Inferences can be made from a large (independent) sample from via $p_{\theta|X}$, rather than the analytic form itself.

Using ideas from Monte Carlo, if we can obtain a sample of size M from $p_{\theta|X}$, $\theta^{(1)}, \ldots, \theta^{(M)}$, then we may obtain an approximation to $E_{\theta|X}[h(\theta)|x]$ as follows:

$$\widehat{E}_{\theta|X}[h(\theta)|x] = \frac{1}{M} \sum_{m=1}^{M} h(\theta^{(m)})$$

If $p_{\theta|X}$ is non-standard and high-dimensional, producing a large sample from it may also prove problematic.

This problem has been successfully approached using

Markov Chain Monte Carlo

that is, it is possible to construct a *aperiodic* and *irreducible* Markov chain on the parameter space with *stationary distribution* $p_{\theta|X}$.

This method will be studied in detail later.