Point Estimates

Empirical and theoretical moments

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \mu_1' = \mathrm{E}(x)$$
$$m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \qquad \mu_2' = \mathrm{E}(x^2)$$
$$\vdots$$
$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \qquad \mu_k' = \mathrm{E}(x^k)$$

Method of moments

$$m_1 = \mu_1'(x;\theta_1,\dots,\theta_k)$$
$$m_2 = \mu_2'(x;\theta_1,\dots,\theta_k)$$
$$\vdots$$
$$m_k = \mu_k'(x;\theta_1,\dots,\theta_k)$$

Solve for $\hat{\theta}$

$$\hat{\theta}_1 = \hat{\theta}_1(m_1,\dots,m_k)$$
$$\hat{\theta}_2 = \hat{\theta}_2(m_1,\dots,m_k)$$
$$\vdots$$
$$\hat{\theta}_k = \hat{\theta}_k(m_1,\dots,m_k)$$

Method of moments

Example:

If $x_1,\dots,x_n$ are iid exponential(λ), then $\mathrm{E}(x_i) = {}^{1}/_{\lambda}$. The method of moments estimator sets

$$\bar{x} = 1/\lambda$$
$$\tilde{\lambda} = 1/\bar{x}$$
$$\overline{x^2} = (1/\lambda)^2 + 1/\lambda^2$$
$$\tilde{\tilde{\lambda}} = \sqrt{2/\overline{x^2}}$$

Method of moments - normal

Suppose $x_1, \dots, x_n$ are iid $N(\mu, \sigma^2)$, then

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

Set

$$\bar{x} = \mu$$
$$\frac{1}{n}\sum x_i^2 = \mu^2 + \sigma^2$$
$$\tilde{\mu} = \bar{x}$$
$$\tilde{\sigma}^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

Method of moments – simple regression

Model:

$$y = \alpha + \beta x + \varepsilon$$
$$E(\varepsilon) = 0$$
$$E\left(\sum x\varepsilon\right) = 0$$

$$\varepsilon = y - \alpha - \beta x$$

Method of moments – simple regression

$$\frac{1}{n}\sum (y - \alpha - \beta x) = 0$$
$$\frac{1}{n}\sum x(y - \alpha - \beta x) = 0$$
$$\tilde{\beta} = \frac{\sum (y - \bar{y})(x - \bar{x})}{\sum (x - \bar{x})^2}$$
$$\tilde{\alpha} = \bar{y} - \tilde{\beta}\bar{x}$$

Method of moments – multiple regression

$$y = X\beta + \varepsilon, X \text{ is } n \times k$$
$$\mathrm{E}(X'\varepsilon) = 0_k$$
$$\varepsilon = y - X\beta$$
$$\frac{1}{n}X'(y - X\beta) = 0_k$$
$$X'y = X'X\tilde{\beta}$$
$$\tilde{\beta} = (X'X)^{-1}X'y$$

More moments than parameters

Suppose we have k parameters but $q \geq k$ moment conditions. Let $\bar{m}(\theta)$ be the $q \times 1$ vector of sample moments, where we want $\bar{m}(\theta) = 0$. Then consider the objective function

$$J = \bar{m}(\theta)' W \bar{m}(\theta)$$

W can be any positive definite function of the data

Generalized method of moments (GMM)

$$J = \bar{m}(\theta)' W \bar{m}(\theta)$$

Pick θ_{GMM} to minimize J.

With some regularity conditions

$$\theta_{GMM} \xrightarrow{p} \theta$$

Convergence in distribution

$$\bar{m}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(y_i, \theta)$$

$$G_{ij} = \mathrm{E}\left[\frac{\partial g(y_i, \theta)}{\partial \theta_j}\right]$$

$$\Omega = \mathrm{E}[g(y_i, \theta) g(y_i, \theta)']$$

$$\sqrt{n}(\theta_{GMM} - \theta) \xrightarrow{d} N[0, (G'WG)^{-1} G'W\Omega W' G (G'WG)^{-1}]$$

Optimal weighting matrix

$$W^* \propto \Omega^{-1}$$

$$\begin{aligned}
&asym.\,\mathrm{var}(\theta_{GMM}) \\
&= (G'WG)^{-1} G'W\Omega W' G (G'WG)^{-1} \\
&= (G'\Omega^{-1}G)^{-1} G'\Omega^{-1}\Omega\Omega^{-1\prime} G (G'\Omega^{-1}G)^{-1} \\
&= (G'\Omega^{-1}G)^{-1} G'\cancel{\Omega^{-1}\Omega}\Omega^{-1\prime} G (G'\Omega^{-1}G)^{-1} \\
&= \cancel{(G'\Omega^{-1}G)^{-1} G'\Omega^{-1\prime} G} (G'\Omega^{-1}G)^{-1} \\
&= (G'\Omega^{-1}G)^{-1}
\end{aligned}$$

Simple linear example

$$y = X\beta + \varepsilon$$

Where β is $k \times 1$ and

$$\mathrm{E}[\varepsilon\varepsilon'] = \sigma^2 I_n$$

And suppose we have the $n \times q$ matrix of instruments Z, $q \geq k$, with moment restrictions

$$Z'\varepsilon = 0$$

$$g(\cdot) = Z_i(y_i - X_i\beta) = Z_i\varepsilon_i$$
$$\bar{m}(\beta) = \frac{1}{n}\sum_{i=1}^{n} Z_i(y_i - X_i\beta)$$
$$G = -ZX$$
$$\Omega = \mathrm{E}[g(y_i, \theta)g(y_i, \theta)']$$
$$= \mathrm{E}[Z'\varepsilon\varepsilon' Z] = Z'\,\mathrm{E}[\varepsilon\varepsilon']Z = \sigma^2 Z'Z$$
$$W^* = (Z'Z)^{-1}$$

$$J = \varepsilon' Z(Z'Z)^{-1}Z'\varepsilon$$
$$J = [y - X\beta]'Z(Z'Z)^{-1}Z'[y - X\beta]$$

Notation:

$$P_Z = Z(Z'Z)^{-1}Z'$$

Note that P_Z is symmetric, idempotent

$$P_Z P_Z = P_Z$$
$$J = [P_Z y - P_Z X\beta]'[P_Z y - P_Z X\beta]$$
$$\beta_{GMM} = (X'P_Z' P_Z X)^{-1}X'P_Z' P_Z y$$
$$= (X'P_Z X)^{-1}X'P_Z y$$

Two-stage least squares

$$\beta_{2SLS} = (X'P_Z X)^{-1}X'P_Z y$$
$$asy\,\mathrm{var}(\beta_{2SLS}) = (X'Z(\sigma^2 Z'Z)^{-1}Z'X)^{-1}$$
$$= \sigma^2(X'P_Z X)^{-1}$$

Likelihood function

If $x_1, \dots x_n$ is a random sample depending on parameters $\theta_1, \dots \theta_k$ then the likelihood function is

$$L(\theta_1, \dots \theta_k | x_1, \dots x_n) = P(x_1, \dots x_n | \theta_1, \dots \theta_k)$$

For small ε

$$\frac{P(x - \varepsilon < X < x + \varepsilon | \theta_A)}{P(x - \varepsilon < X < x + \varepsilon | \theta_B)} \approx \frac{L(\theta_A | x)}{L(\theta_B | x)}$$

Maximum likelihood estimate

Definition 7.2.4 For each sample point x, let $\hat{\theta}(x)$ be a parameter value at which $L(\theta|x)$ attains its maximum as a function of θ with x held fixed. A *maximum likelihood estimator* (*MLE*) of the parameter θ based on a sample X is $\hat{\theta}(X)$.

Invariance property of MLEs

Theorem 7.2.10 *Invariance property of MLEs*. If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Suppose $x_i \sim iid U[0, b]$. Write the joint pdf for x, being careful about edge conditions. Find the value of b that maximizes the joint probability. (Hint: this is a logic question rather than a calculation question.)

If our sample values are iid, then we can write

$$L(\theta|x) = L(\theta_1, \ldots \theta_k | x_1, \ldots x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots \theta_k)$$

If $L(\cdot)$ is differentiable, then maybe $\hat{\theta}$ solves

$$\frac{\partial L(\theta|x)}{\partial \theta} = 0$$

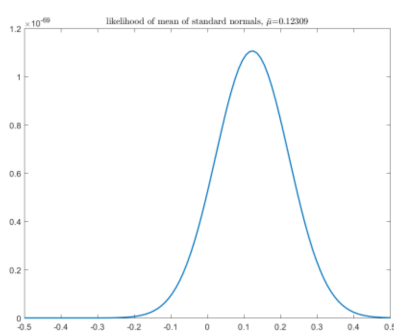
Log-likelihood

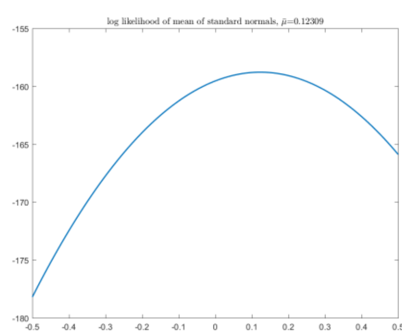
$$\mathcal{L}(\theta|x) = \log L(\theta|x) = \sum_{i=1}^{n} \log f(x_i|\theta_1, \ldots \theta_k)$$

Setting

$$\frac{\partial \mathcal{L}(\theta|x)}{\partial \theta} = 0$$

$$\frac{\partial \mathcal{L}(\theta|x)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log f(x_i|\theta_1, \ldots \theta_k)}{\partial \theta} = 0$$





Normal mle

$$x_i \sim \text{iid } N(\mu, \sigma^2)$$
$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$
$$\log\big(f(x_i)\big) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_i-\mu)^2$$
$$\frac{\partial \log\big(f(x_i)\big)}{\partial\mu} = \frac{1}{\sigma^2}(x_i-\mu)$$
$$\frac{\partial \log\big(f(x_i)\big)}{\partial\sigma^2} = -\frac{1}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}(x_i-\mu)^2$$

F.O.C.

$$\frac{\partial\mathcal{L}}{\partial\mu} = 0 = \frac{1}{\sigma^2}\left(\sum x_i - n\mu\right)$$
$$\frac{\partial\mathcal{L}}{\partial\sigma^2} = 0 = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum(x_i-\mu)^2$$
$$\hat{\mu} = \bar{x}$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum(x_i-\bar{x})^2$$

Simple normal regression

$$y = \beta x + \varepsilon, \varepsilon \sim iidN(0, \sigma^2)$$
$$y|x \sim N(\beta x, \sigma^2)$$

$$\log f(y_i) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - \beta x_i)^2$$

Simple normal regression

$$\log f(y_i) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - \beta x_i)^2$$
$$\frac{\partial\mathcal{L}}{\partial\beta} = \frac{1}{\sigma^2}\sum(y_i - \beta x_i)x_i$$
$$\hat{\beta} = \frac{\sum y_i x_i}{\sum x_i^2}$$
$$\hat{\sigma}^2 = \frac{\sum\big(y_i - \hat{\beta}x_i\big)^2}{n}$$

Note: $\sum\big(y_i - \hat{\beta}x_i\big)^2$ is called the sum of squared residuals

Suppose the errors are not iid

$$y = X\beta + \varepsilon, \varepsilon \sim N(0, \Sigma)$$
$$y \text{ is } n \times 1, X \text{ is } n \times k$$
$$f(y)$$
$$= (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(y - X\beta)'\Sigma^{-1}(Y - X\beta)\right]$$
$$\log L \propto (y - X\beta)'\Sigma^{-1}(Y - X\beta)$$
$$\beta_{mle} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$$

IID exponential mle

$$f(y_i|\lambda) = \frac{1}{\lambda}e^{-\frac{y_i}{\lambda}}$$
$$L(\lambda|y) = \prod_{i=1}^{n} \frac{1}{\lambda}e^{-\frac{y_i}{\lambda}}$$
$$\mathcal{L}(\lambda|y) = \log L = -n \log \lambda - \frac{1}{\lambda}\sum_{i=1}^{n} y_i$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{n} y_i$$
$$\lambda_{mle} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Nonlinear regression

$$y_i = f(x_i, \theta) + \varepsilon_i$$
$$\varepsilon \sim iidN(0, \sigma^2)$$

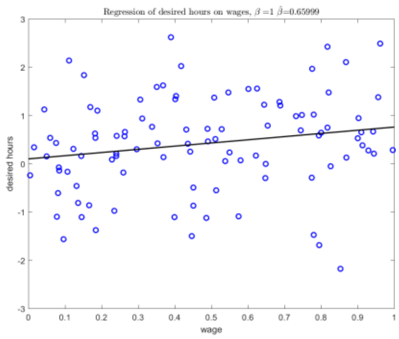
$$L(\theta, x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\big(y_i - f(x_i, \theta)\big)^2\right)$$

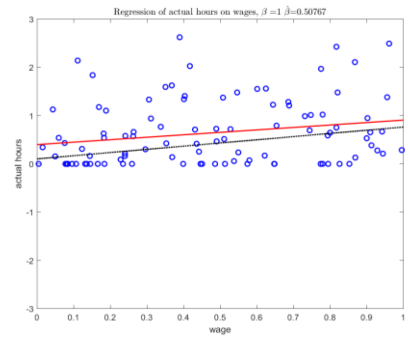
$$\mathcal{L}(\theta, x_i) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\big(y_i - f(x_i, \theta)\big)^2$$

Tobit

We observe wage and desired work hours.

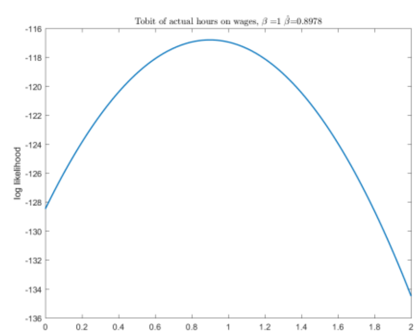
$$H^* = \beta w + \varepsilon$$





Tobit likelihood

$$y|y > 0 \sim N(\beta w, \sigma_\varepsilon^2)$$
$$p(y = 0) = p(\beta w + \varepsilon < 0) = p(\varepsilon < -\beta w)$$
$$= \Phi\left(-\frac{\beta w}{\sigma_\varepsilon}\right)$$
$$L(\beta) = \prod_{y>0} \frac{1}{\sigma}\phi\left(\frac{y - \beta w}{\sigma_\varepsilon}\right) \times \prod_{y=0} \Phi\left(-\frac{\beta w}{\sigma_\varepsilon}\right)$$



EM (Expectations-Maximization) algorithm

E: Make up missing data as expected value of data given current parameter estimates.

M: Get new parameter estimates treating made up data as real.

Then you iterate between the two until convergence.

EM Tobit

$$\hat{\beta} = \frac{\sum \tilde{h} w}{\sum w^2}$$

$$h > 0, \tilde{h} = h$$
$$h = 0, \tilde{h} = \mathrm{E}\left(TN\left(\hat{\beta} w, \sigma^2\right)\right)$$

Expectation of truncated normal

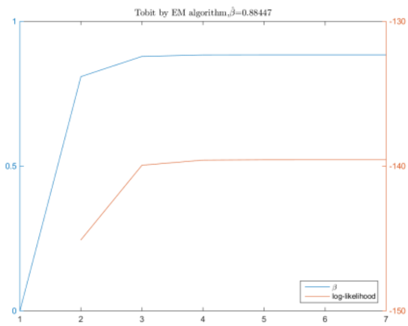
$$x \sim N(\mu, \sigma^2)$$
$$\mathrm{E}(x | l < x < u) = \mu + \frac{\phi\left(\frac{l-\mu}{\sigma}\right) - \phi\left(\frac{u-\mu}{\sigma}\right)}{\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right)} \sigma$$

In our case

$$l = -\infty, u = 0$$
$$\mathrm{E}(x | x < 0) = \mu - \frac{\phi\left(\frac{-\mu}{\sigma}\right)}{\Phi\left(\frac{-\mu}{\sigma}\right)} \sigma$$


```
reset(RandStream.getGlobalStream);
n = 100;
% now let's do tobit
beta = 1;
wage = rand(n,1);
desiredHours = beta*wage + randn(n,1);
actualHours = desiredHours.*(desiredHours>0);
censored = actualHours==0;

maxIt = 100;
betaHat = nan(maxIt,1);
mle = nan(maxIt,1);
betaHat(1) = 0;
mle(1) = -inf;
estimatedHours = actualHours;
for it=2:maxIt
    estimatedHours(censored)...
        = truncatedNormalExpectation(betaHat(it-1)*wage(censored),1);% E step
    betaHat(it) = wage\estimatedHours;                                % M step
    mle(it) = regrLogLike(estimatedHours,wage,betaHat(it));
    if abs(mle(it) - mle(it-1)) <.001
        break; % converged
    end
end
```



Minimum Distance Estimators

Parameters θ, K long
Functions $g(\theta)$, $L \geq K$ long
Sample statistics $\bar{m}_n$, L long defined over n obs.

Assume

$$\operatorname{plim} \bar{m}_n = g(\theta)$$
$$\sqrt{n}\left(\bar{m}_n - g(\theta)\right) \xrightarrow{d} N(0, \Phi)$$

Minimum Distance Estimators

$\hat{\theta}_{MDE}$ solves

$$\min q = \left(\bar{m}_n - g(\theta)\right)' W \left(\bar{m}_n - g(\theta)\right)$$

for positive definite W.

MDE Asymptotics

$$\operatorname{plim}(\hat{\theta}_{MDE}) = \theta$$
$$asym \operatorname{var}\left(\hat{\theta}_{MDE}\right)$$
$$= \frac{1}{n}[\Gamma(\theta)' W \Gamma(\theta)]^{-1}[\Gamma(\theta)' W \Phi W \Gamma(\theta)][\Gamma(\theta)' W \Gamma(\theta)]^{-1}$$
$$\equiv \frac{1}{n} V$$

where

$$\Gamma(\theta) = \operatorname{plim} \frac{\partial g\left(\hat{\theta}_{MDE}\right)}{\partial \hat{\theta}'_{MDE}}$$
$$\hat{\theta}_{MDE} \overset{a}{\sim} N\left(\theta, \frac{1}{n} V\right)$$

Bayesian Estimates

If we have data x and parameters θ,

$$f(\theta|x) = \frac{f(x|\theta) \times f(\theta)}{f(x)}$$

where

$f(\theta|x)$ posterior

$f(x|\theta)$ likelihood function

$f(\theta)$ prior

$f(x) = \int_{-\infty}^{\infty} f(x|\theta) \cdot f(\theta) d\theta$ marginal likelihood

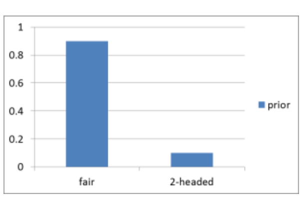
Short-hand

Posterior is proportional to likelihood times prior.

$$f(\theta|x) = \frac{f(x|\theta) \times f(\theta)}{f(x)}$$

$$f(\theta|x) \propto f(x|\theta) \times f(\theta)$$

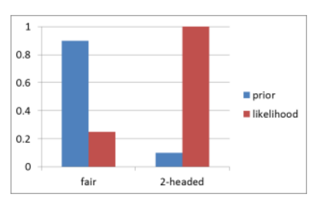
prior



likelihood

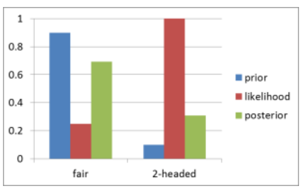
$$f(H = 2|fair) = .5^2 = .25$$

$$f(H = 2|2\ headed) = 1^2 = 1$$

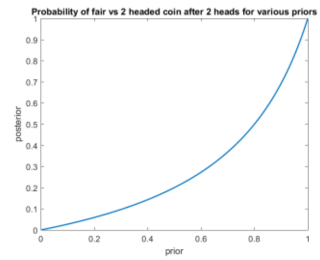


posterior

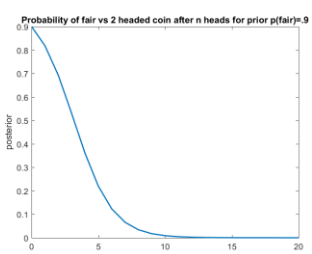
$$f(fair|H=2)$$
$$=\frac{f(H=2|fair)\times f(fair)}{f(H=2|fair)\times f(fair)+f(H=2|2\ headed)\times f(2\ headed)}$$
$$=\frac{.25\times .9}{.25\times .9+1\times .1}=0.69$$



 50



 51



 52

Suppose $\bar{x}|\theta \sim N\left(\theta, \frac{\sigma^2}{n}\right)$—assume σ^2 and n are known constants and $\theta \sim U(-c, c)$. Find an expression for

$$f(\bar{x}) = \int_{-\infty}^{\infty} f(\bar{x}|\theta) f(\theta) d\theta$$

 53

Normal data with normal prior

Let $x \sim N(\theta, \sigma^2)$ and suppose the prior is $N(\mu, \tau^2)$, then $\theta | x \sim N()$ with

$$\mathrm{E}(\theta|x) = \frac{\tau^2}{\tau^2 + \sigma^2} x + \frac{\sigma^2}{\tau^2 + \sigma^2} \mu$$

$$var(\theta|x) = \frac{\sigma^2 \tau^2}{\tau^2 + \sigma^2}$$

Normal prior, likelihood, and posterior

$$x_i \sim iidN(\mu, \sigma^2)$$

Likelihood:

$$f(x|\mu) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(\frac{1}{2\sigma^2} \sum (x_i - \mu)^2\right)$$

Prior

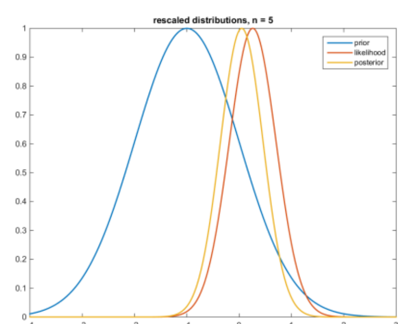
$$\mu \sim N\left(\underline{\mu}, \tau^2\right)$$

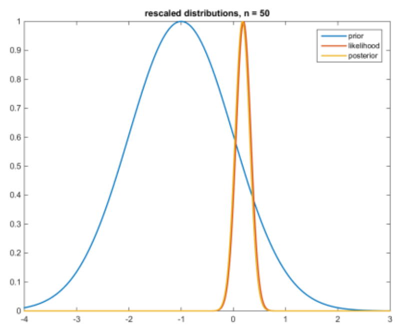
$$f(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(\frac{1}{2\tau^2} \left(\mu - \underline{\mu}\right)^2\right)$$

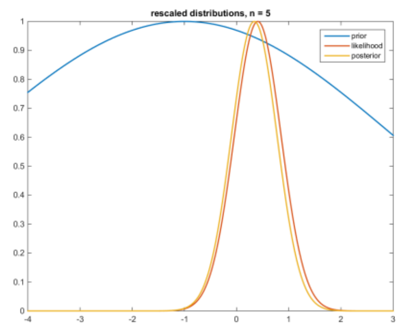
Normal prior, likelihood, and posterior

Posterior:

$$\mu | \bar{x} \sim N\left(\frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}} \bar{x} + \frac{\frac{\sigma^2}{n}}{\tau^2 + \frac{\sigma^2}{n}} \underline{\mu}, \frac{\frac{\sigma^2}{n} \tau^2}{\tau^2 + \frac{\sigma^2}{n}}\right)$$







$n \rightarrow \infty$

$$\mu|\bar{x}\sim N\left(\frac{\tau^2}{\tau^2+\frac{\sigma^2}{n}}\bar{x}+\frac{\frac{\sigma^2}{n}}{\tau^2+\frac{\sigma^2}{n}}\underline{\mu},\frac{\frac{\sigma^2}{n}\tau^2}{\tau^2+\frac{\sigma^2}{n}}\right)$$

As $n \rightarrow \infty$

$$\mu|\bar{x}\sim N\left(\frac{\tau^2}{\tau^2+0}\bar{x}+\frac{0}{\tau^2+0}\underline{\mu},\frac{0\tau^2}{\tau^2+0}\right)$$

$$\mu|x\sim N(\bar{x},0)$$

$n \rightarrow 0$

$$\mu|\bar{x}\sim N\left(\frac{\tau^2}{\tau^2+\frac{\sigma^2}{n}}\bar{x}+\frac{\frac{\sigma^2}{n}}{\tau^2+\frac{\sigma^2}{n}}\underline{\mu},\frac{\frac{\sigma^2}{n}\tau^2}{\tau^2+\frac{\sigma^2}{n}}\right)$$

As $n \rightarrow 0$

$$\mu|\bar{x}\sim N\left(\frac{\tau^2}{\tau^2+\infty}\bar{x}+\frac{\infty}{\tau^2+\infty}\underline{\mu},\frac{\infty\tau^2}{\tau^2+\infty}\right)$$

$$\mu|\bar{x}\sim N\left(\underline{\mu},\tau^2\right)$$

$\tau \to 0$

$$\mu|\bar{x} \sim N\left(\frac{\tau^2}{\tau^2+\frac{\sigma^2}{n}}\bar{x}+\frac{\frac{\sigma^2}{n}}{\tau^2+\frac{\sigma^2}{n}}\underline{\mu}, \frac{\frac{\sigma^2}{n}\tau^2}{\tau^2+\frac{\sigma^2}{n}}\right)$$

As $\tau \to 0$

$$\mu|\bar{x} \sim N\left(\frac{0}{0+\frac{\sigma^2}{n}}\bar{x}+\frac{\frac{\sigma^2}{n}}{0+\frac{\sigma^2}{n}}\underline{\mu}, \frac{\frac{\sigma^2}{n}0}{0+\frac{\sigma^2}{n}}\right)$$

$$\mu|\bar{x} \sim N\left(\underline{\mu}, 0\right)$$

$\tau \to \infty$

$$\mu|\bar{x} \sim N\left(\frac{\tau^2}{\tau^2+\frac{\sigma^2}{n}}\bar{x}+\frac{\frac{\sigma^2}{n}}{\tau^2+\frac{\sigma^2}{n}}\underline{\mu}, \frac{\frac{\sigma^2}{n}\tau^2}{\tau^2+\frac{\sigma^2}{n}}\right)$$

As $\tau \to \infty$

$$\mu|\bar{x} \sim N\left(\frac{\infty}{\infty+\frac{\sigma^2}{n}}\bar{x}+\frac{\frac{\sigma^2}{n}}{\infty+\frac{\sigma^2}{n}}\underline{\mu}, \frac{\frac{\sigma^2}{n}\infty}{\infty+\frac{\sigma^2}{n}}\right)$$

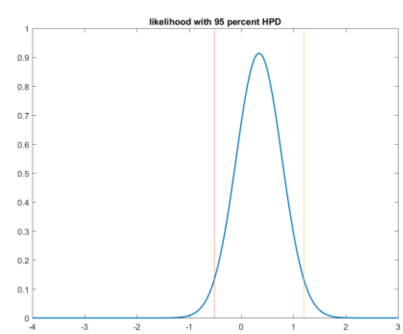
$$\mu|\bar{x} \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right)$$

Highest posterior density (HPD)

Given that the posterior is

$$\mu|x \sim N(\bar{\mu}, \bar{\sigma}^2)$$

we can compute the most compact part of the posterior that contains 95 percent of the probability mass.



Bayesian model comparison

Suppose we have two models, M_0 and M_1.

$$f(\theta|x, M_i) = \frac{f(x|\theta, M_i)f(\theta|M_i)}{f(x|M_i)}$$

where

$$f(x|M_i) = \int f(x|\theta, M_i)f(\theta|M_i)d\theta$$

Posterior odds ratio

$$p(M_i|x) = \frac{f(x|M_i) \times p(M_i)}{f(x)}$$

We can write the *posterior odds ratio* as

$$PO_{01} = \frac{p(M_0|x)}{p(M_1|x)} = \frac{\frac{f(x|M_0) \times p(M_0)}{f(x)}}{\frac{f(x|M_1) \times p(M_1)}{f(x)}}$$

$$PO_{01} = \frac{f(x|M_0)}{f(x|M_1)} \times \frac{p(M_0)}{p(M_1)}$$

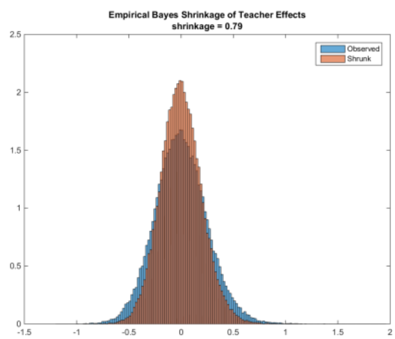
Posterior odds ratio

$$\underbrace{PO_{01}}_{\text{posterior odds}} = \underbrace{\frac{f(x|M_0)}{f(x|M_1)}}_{\text{Bayes factor}} \times \underbrace{\frac{p(M_0)}{p(M_1)}}_{\text{prior odds}}$$

"Empirical Bayes" example

$$x_i = \theta_i + \varepsilon_i$$
$$\theta_i \sim iidN(\bar{\theta}, \sigma^2)$$
$$\varepsilon_i \sim iidN(0, \tau^2)$$

$$\mathrm{var}(x) = \sigma^2 + \tau^2$$
$$\sigma^2 = \mathrm{var}(x) - \tau^2$$
$$\mathrm{E}(\theta_i|x_i) = \frac{\sigma^2}{\sigma^2 + \tau^2}x_i + \frac{\tau^2}{\sigma^2 + \tau^2}\bar{x}$$



Properties of a good estimator

- Unbiased (finite sample)
  - It's nice to get it right on average.
- Consistent (asymptotic)
  - It's nice to know that in a large sample you're going to get it right.

Mean square error

$$mse = \mathrm{E}(\hat{\theta} - \theta)^2$$

$$mse = \mathrm{var}(\hat{\theta}) + \left(\mathrm{E}(\hat{\theta}) - \theta\right)^2$$

Example: s^2 versus $\hat{\sigma}^2_{mle}$

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

$$\sigma^2_{mle} = \frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{n-1}{n}s^2$$

$$\mathrm{E}(s^2) = \sigma^2$$
$$\mathrm{var}(s^2) = \frac{2}{n-1}\sigma^4$$
$$mse(s^2) = \frac{2}{n-1}\sigma^4$$
$$mse(\sigma^2_{mle})$$
$$= \left(\frac{n-1}{n}\right)^2 \left(\frac{2}{n-1}\sigma^4\right) + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2$$
$$= \sigma^4 \left(\frac{2}{n^2}(n-1)\right) + \left(-\frac{1}{n}\right)^2 = \frac{\sigma^4}{n^2}(2n-1)$$
$$mse(s^2) = \frac{2}{n-1}\sigma^4 > \frac{\sigma^4}{n^2}(2n-1) = mse(\sigma^2_{mle})$$

Question

Generate $n = \{10{,}100\}$ $iidU(0, b = 1)$ random variables and find the maximum likelihood estimator of b. Do this many times for each sample size and report on the distribution of b_{mle}. In particular, is it unbiased? Is the distribution approximately normal?

Best unbiased estimator

Definition 7.3.7

An estimator W^* is a *best unbiased estimator* of $\tau(\theta)$ if it satisfies

1. $\mathrm{E}(W^*) = \tau(\theta)$ for all θ
2. For any other estimator, W, with $\mathrm{E}(W) = \tau(\theta)$, we have

$$\mathrm{var}(W^*) \leq \mathrm{var}(W) \text{ for all } \theta.$$

W^* is also called a *uniform minimum variance unbiased estimator* (*UMVUE*) of $\tau(\theta)$.

Incredible Cramér-Rao Lower Bound (CRLB)

Cramér-Rao Inequality

Let $x_1, \ldots, x_n$ be a sample with pdf $f(x|\theta)$, and let $W(x_1, \ldots, x_n)$ be any estimator satisfying

$$\frac{d}{d\theta}\mathrm{E}\big(W(x)\big) = \int_x \frac{\partial}{\partial\theta}[W(x)f(x|\theta)]dx$$

And

$$\mathrm{var}\big(W(x)\big) < \infty$$

Then

$$\mathrm{var}\big(W(x)\big) \geq \underbrace{\frac{\left(\frac{d}{d\theta}\mathrm{E}\big(W(x)\big)\right)^2}{\mathrm{E}\left(\left(\frac{\partial}{d\theta}\log f\,(x|\theta)\right)^2\right)}}_{CRLB}$$

Cramér-Rao Inequality (iid case)

If in addition to the previous regularity conditions $x_1, \ldots, x_n$ is iid with pdf $f(x_i|\theta)$

$$\mathrm{var}\big(W(x)\big) \geq \frac{\left(\frac{d}{d\theta}\mathrm{E}\big(W(x)\big)\right)^2}{n \cdot \mathrm{E}\left(\left(\frac{\partial}{d\theta}\log f\,(x_i|\theta)\right)^2\right)}$$

Sample mean and CRLB

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

$$W = \frac{x_1 + \cdots + x_n}{n}$$

$$\mathrm{E}\big(W(x)\big) = \theta$$

$$\frac{d}{d\theta}\mathrm{E}\big(W(x)\big) = 1$$

$$\mathrm{var}\big(W(x)\big) \geq \frac{\left(\frac{d}{d\theta}\mathrm{E}\big(W(x)\big)\right)^2}{n \cdot \mathrm{E}\left(\left(\frac{\partial}{d\theta}\log f\,(x_i|\theta)\right)^2\right)}$$

$$\mathrm{var}\big(W(x)\big) \geq \frac{1}{n \cdot \mathrm{E}\left(\left(\frac{\partial}{d\theta}\log f\,(x_i|\theta)\right)^2\right)}$$

$$\log f\,(x_i|\theta) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_i-\theta)^2$$
$$\frac{\partial}{d\theta}\log f\,(x_i|\theta) = \frac{1}{\sigma^2}(x_i-\theta)$$

$$\mathrm{E}\left(\left(\frac{1}{\sigma^2}(x_i-\theta)\right)^2\right) = \left(\frac{1}{\sigma^2}\right)^2\sigma^2 = \frac{1}{\sigma^2}$$
$$\operatorname{var}\big(W(x)\big) \geq \frac{1}{n\cdot\left(\frac{1}{\sigma^2}\right)^2\sigma^2} = \frac{\sigma^2}{n}$$

Information matrix

$$I \equiv \mathrm{E}\left(\left(\frac{\partial}{d\theta}\log f\,(x|\theta)\right)^2\right)$$

is called the *information matrix* or the *Fisher information*.

$$\frac{\partial}{d\theta}\log f\,(x|\theta)$$

is called the score. Note that

$$\mathrm{E}\left(\frac{\partial}{d\theta}log\ f\,(x|\theta)\right) = 0$$

$$\mathrm{E}\left(\frac{\partial}{d\theta}log\ f\,(x|\theta)\right)$$
$$= \int\left[\frac{\partial}{d\theta}log\ f\,(x|\theta)\right]f(x|\theta)dx$$
$$= \int\frac{1}{f(x|\theta)}\left[\frac{\partial}{d\theta}f(x|\theta)\right]f(x|\theta)dx$$
$$= \int\left[\frac{\partial}{d\theta}f(x|\theta)\right]dx$$
$$\frac{\partial}{d\theta}\int[f(x|\theta)]dx$$
$$= \frac{\partial}{d\theta}1 = 0$$

Lemma 7.3.11 If $f(x|\theta)$ satisfies

$$\frac{d}{d\theta}\mathrm{E}\left(\frac{\partial}{\partial\theta}log(f(x|\theta)\right)$$
$$= \int\frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}log(f(x|\theta)\right)f(x|\theta)\right]dx$$

Then

$$I \equiv \mathrm{E}\left(\left(\frac{\partial}{d\theta}\log f\,(x|\theta)\right)^2\right)$$
$$= -\,\mathrm{E}\left(\frac{\partial^2}{\partial\theta^2}log\ f(x|\theta)\right)$$

$$\frac{\partial}{d\theta}\log f\,(x_i|\theta) = \frac{1}{\sigma^2}(x_i - \theta)$$

$$\mathrm{E}\left(\left(\frac{1}{\sigma^2}(x_i - \theta)\right)^2\right) = \left(\frac{1}{\sigma^2}\right)^2 \sigma^2 = \frac{1}{\sigma^2}$$

$$\frac{\partial^2}{d\theta^2}\log f\,(x_i|\theta) = -\frac{1}{\sigma^2} = -I$$

CRLB for unbiased estimators

If $W(x)$ is an unbiased estimator, then (subject to suitable regularity conditions) the CRLB is

$$\mathrm{var}\big(W(x)\big) \geq I^{-1}(\theta)$$

Proof

$$\mathrm{E}\big(W(x)\big) = \theta \Rightarrow \frac{d}{d\theta}\mathrm{E}\big(W(x)\big) = 1$$

Matrix form

$$I(\theta)_{i,j} = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\log f(x|\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f(x|\theta)\right)\right]$$

Which with appropriate regularity conditions is also

$$I(\theta)_{i,j} = -\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x|\theta)\right]$$

And we can write for any unbiased estimator

$\mathrm{var}\big(W(\theta)\big) - I(\theta)^{-1}$ is p.s.d.

Asymptotic efficiency

A sequence of estimators W_n is *asymptotically efficient* for $\tau(\theta)$ if

$$\sqrt{n}[W_n - \tau(\theta)] \xrightarrow{d} N\big(0, V(\theta)\big)$$

And

$$V(\theta) = \frac{\tau'(\theta)^2}{I(\theta)}$$

That is, if the asymptotic variance attains the CRLB.

MLE achieves the CRLB

Theorem

Under suitable regularity conditions if $x_i \sim iid(\theta, \sigma^2)$, then

$$\hat{\theta}_{mle} \xrightarrow{d} N(\theta, I(\theta)^{-1})$$

Maximum-likelihood is asymptotically efficient.

(Proof: Hansen Appendix B, theorem B.11.2)

$\hat{\theta}_{mle}$ is consistent

Consider

$$\mathrm{E} \log f(x_i|\theta)$$

We just showed this is maximized at θ_0 because the first partial equals zero.

By the law of large numbers, we know that

$$\frac{1}{n}\sum \log f(x_i|\theta) \xrightarrow{p} \mathrm{E} \log f(x_i|\theta)$$

Since θ_{mle} maxes the LHS and θ_0 maxes the RHS, since max is a function, and since plims go through functions, we have

$$\hat{\theta}_{mle} \xrightarrow{p} \theta_0$$

$$l(\theta|x) = \sum \log f(x_i|\theta)$$

Then a first-order Taylor series expansion of the first partial around the true value θ_0 is

$$l'(\theta|x) \approx l'(\theta_0|x) + (\theta - \theta_0) l''(\theta_0|x)$$

Since we have $l'(\hat{\theta}|x) = 0$ we can write

$$0 = l'(\theta_0|x) + (\hat{\theta} - \theta_0) l''(\theta_0|x)$$

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}\frac{-l'(\theta_0|x)}{l''(\theta_0|x)} = \frac{-\frac{1}{\sqrt{n}} l'(\theta_0|x)}{\frac{1}{n} l''(\theta_0|x)}$$

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}\frac{-l'(\theta_0|x)}{l''(\theta_0|x)} = \frac{-\frac{1}{\sqrt{n}} l'(\theta_0|x)}{\frac{1}{n} l''(\theta_0|x)}$$

It can be shown that

$$-\frac{1}{\sqrt{n}} l'(\theta_0|x) \xrightarrow{d} N(0, I(\theta_0))$$

$$\frac{1}{n} l''(\theta_0|x) \xrightarrow{p} I(\theta_0)$$

Using the delta method we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0) I(\theta_0) I^{-1}(\theta_0))$$
$$= N(0, I^{-1}(\theta_0))$$

And that's why people like to use maximum-likelihood. It has a known asymptotic distribution and asymptotically achieves the CRLB.

Loss functions

$$L(\theta, a)$$

And we'd like to minimize the expected loss

$$\min_a \mathrm{E}\big(L(\theta, a)\big)$$

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$$

This leads to using the mean in our standard model.

$$L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$$

Which leads to use of the median.

Bayesian version

In a Bayesian framework, we can set this up using any action as a function of the data, $a(x)$, and take expectations with respect to the Bayesian posterior, $f(\theta|x)$.

$$\min_{a(x)} \int L(a(x), \theta) f(\theta|x) d\theta$$

Investing in a risky asset

W is amount in risky asset

$$\max_W \int U(W|\mu, \sigma^2) f(\mu, \sigma^2|x) d\theta$$