Ratios

one population of size N and two measurements resulting in data

$X_1, X_2, \cdots X_N$ and $Y_1, Y_2, \cdots Y_N$

population parameters:

$$R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\mu_y}{\mu_x}, \quad \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

population covariance

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y).$$

$\mu_x = \frac{1}{N} \sum x_i \quad \mu_y = \frac{1}{N} \sum y_i \quad \sigma_x^2 = \frac{1}{N} \sum (x_i - \mu_x)^2$

two possible goals: ① estimate R via S.R.S

② Often, X values are known, and the ratio estimate can be used to give an estimate of $\mu_y = \frac{\mu_x}{\bar{x}} \bar{y}$

~~$E(R) = r + \frac{1}{n}\left(r \frac{n-1}{N-1}\right)$~~

If $X_1, X_2, \cdots X_n$ and $Y_1, Y_2, \cdots Y_n$ are simple random samples

$$\mathrm{cov}(\bar{x}, \bar{y}) = \frac{\sigma_{xy}}{n}\left(\frac{N-n}{N-1}\right).$$

$$R = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} = \frac{\bar{y}}{\bar{x}} = g(\bar{x}, \bar{y})$$ we need to estimate $\mathrm{Var}(R)$

$\mathrm{Var}(R) = \mathrm{Var}(g(\bar{x}, \bar{y}))$?

In the one variable case:

$$z = g(x) \rightarrow \mathrm{Var}(z)$$

$$z = g(x) = g(\mu_x) + g(x - \mu_x) g'(\mu_x) + \frac{1}{2}(x - \mu_x)^2 g''(\mu_x)$$ — Taylor expansion

$$E(z) = g(\mu_x) + \frac{\sigma_x^2}{2} g''(\mu_x)$$

In the two variables case:

$z = g(x, y)$. we expand $g(x, y)$ around $g(\vec{u})$ where $\vec{u} = (\mu_x, \mu_y)$

$$R = g(\bar{x}, \bar{y}) = g(\vec{u}) + (\bar{x} - \mu_x)\frac{\partial g(\vec{u})}{\partial x} + (\bar{y} - \mu_y)\frac{\partial g(\vec{u})}{\partial y} + \frac{1}{2}(\bar{x} - \mu_x)^2 \frac{\partial^2 g(\vec{u})}{\partial x^2} + \frac{1}{2}(\bar{y} - \mu_y)^2 \frac{\partial^2 g(\vec{u})}{\partial y^2} + (\bar{x} - \mu_x)(\bar{y} - \mu_y)\frac{\partial^2 g(\vec{u})}{\partial x \partial y}$$

$$E[z] = g(\mu_x, \mu_y) + \frac{\sigma_x^2}{2}\frac{\partial^2 g(\vec{u})}{\partial x^2} + \frac{\sigma_y^2}{2}\frac{\partial^2 g(\vec{u})}{\partial y^2} + \sigma_{xy}\frac{\partial^2 g(\vec{u})}{\partial x \partial y}$$

$$E[R] = \frac{\mu_y}{\mu_x} + \frac{\sigma_{\bar{x}}^2}{2}\cdot\frac{2\mu_y}{\mu_x^3} + \frac{\sigma_{\bar{y}}^2}{2}\cdot 0 + \sigma_{\bar{x}\bar{y}}\frac{1}{\mu_x^2}$$

$$= \frac{\mu_y}{\mu_x} + \sigma_{\bar{x}}^2 \cdot \frac{\mu_y}{\mu_x^3} - \sigma_{xy}\frac{1}{\mu_x^2} = \frac{\mu_y}{\mu_x} + \frac{\sigma_{\bar{x}}^2 \mu_y - \sigma_{xy}\mu_x}{\mu_x^3} \quad \cdots$$

(1) $E[R] = r + \frac{1}{n}\left(1-\frac{n-1}{N-1}\right)\cdot\frac{r\sigma_x^2 - \rho\sigma_x\sigma_y}{\mu_x^2}$

(2) $Var[R] = \frac{1}{n}\left(1-\frac{n-1}{N-1}\right)\cdot\frac{r^2\sigma_x^2+\sigma_y^2-2r\rho\sigma_{xy}}{\mu_x^2}$

(1) comes from a 1st order Taylor expansion of $g(X,y)=\frac{y}{X}$ around the point (μ_x,μ_y) and the squaring.

(2) comes from a 2nd order Taylor expansion of $g(X,y)=\frac{y}{X}$ around (μ_x,μ_y).

$g_x(\mu_x,\mu_y) = -\frac{\mu_y}{\mu_x^2}$ $g_y(\mu_x,\mu_y)=\frac{1}{\mu_x}$, $g_{xx}(\mu_x,\mu_y)=\frac{2\mu_y}{\mu_x^3}$ $g_{xy}(\mu_x,\mu_y)=-\frac{1}{\mu_x^2}$

$g_{yy}(\mu_x,\mu_y)=0$

$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n}\left(1-\frac{n-1}{N-1}\right)$ $\Big\}\Rightarrow$ ※ Page. 21. $\Rightarrow$ (1)

$\sigma_{\bar{x}\bar{y}} = \frac{\sigma_{xy}}{n}\left(1-\frac{n-1}{N-1}\right)$

$\Rightarrow$

$Var(R) \doteq (R-r)^2 = \frac{\mu_y^2}{\mu_x^4}\sigma_{\bar{x}}^2 + \frac{1}{\mu_x^2}\sigma_{\bar{y}}^2 - \frac{2\mu_y}{\mu_x^3}\sigma_{\bar{x}\bar{y}} \Rightarrow (2)$

— C.I.

$S_R^2 = \frac{1}{n}\left(1-\frac{n-1}{N-1}\right)\cdot\frac{R^2S_x^2+S_y^2-2RS_{xy}}{\bar{X}^2}$, where $S_x^2 = \frac{1}{n-1}\sum(X_i-\bar{x})^2$ $S_y^2 = \frac{1}{n-1}\sum(Y_i-\bar{y})^2$

$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{x})(Y_i-\bar{y})$

$\frac{R-r}{\sqrt{Var(R)}} \sim N(0,1)$ $R \pm z\left(\frac{\alpha}{2}\right)S_R$ $\cdot 100(1-\alpha)\%$ C.I for r.

$\bar{Y}_R = \mu_x R = \frac{\mu_x}{\bar{x}}\bar{y}$ not an unbiased estimator

$Var(\bar{Y}_R) = \mu_x^2 Var(R) = \frac{1}{n}\left(1-\frac{n-1}{N-1}\right)(r^2\sigma_x^2+\sigma_y^2-2r\sigma_{xy})$ (※)

$= \frac{1}{n}(r^2\sigma_x^2+\sigma_y^2-2r\sigma_{xy})\left(1-\frac{n-1}{N-1}\right)$ ignore finite population correction.

$Var(\bar{y}) = \frac{\sigma_y^2}{n}\left(1-\frac{n-1}{N-1}\right)$

when $r^2\sigma_x^2 < 2r\sigma_{xy}$ $Var(\bar{Y}_R) < Var(\bar{y})$.

$r\sigma_x^2 < 2\rho\sigma_x\sigma_y \Rightarrow \rho > \frac{1}{2}r\frac{\sigma_x}{\sigma_y} = \frac{1}{2}\frac{\frac{\sigma_x}{\mu_x}}{\frac{\sigma_y}{\mu_y}} = \frac{1}{2}\frac{C_x}{C_y} \Rightarrow \bar{Y}_R$ has smaller

$r < 2\rho\frac{\sigma_y}{\sigma_x}$ Standard error than $\bar{y}$

C_x, C_y : Coefficients of variation

Ex. X_i: # beds in hospital i. y_i: # discharges in hospital i in 1968/01

$n = 64$, $N = 393$

$\mu_x = 274.8$ $\quad \mu_y = 814.6$ $\quad C_x = 0.78$

$\sigma_x = 213.2$ $\quad \sigma_y = 589.7$ $\quad C_y = 0.72$

$r = 2.96$ $\quad \rho = 0.91$ $\quad \frac{1}{2}\frac{C_x}{C_y} = 0.54 < \rho$

$$\sigma_{\bar{y}_R} = \frac{\sigma_y}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\mathrm{Var}(\bar{y}_R)} = \text{(*)} \quad \text{page 22 - 30 (book P.227)}$$

$\sigma_{\bar{y}} = 29.9$

$$\sigma_{\bar{y}} = \frac{\sigma_y}{\sqrt{n}}\sqrt{1-\frac{n-1}{N-1}} = 66.3$$

Stratified Sampling

The population is partitioned into L subpopulations $N_1, N_2, \cdots N_L$.

$$N_1 + N_2 + \cdots + N_L = N$$

$W_\ell = \frac{N_\ell}{N}$ weight.

Population measurements and parameters

$X_{11}, X_{21}, \cdots, X_{N_1 1}$ measurement of individuals in subgroup 1.

$$\mu_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} X_{i1} \qquad \sigma_1^2 = \frac{1}{N_1}\sum_{i=1}^{N_1}(X_{i1}-\mu_1)^2$$

$X_{12}, X_{22}, \cdots, X_{N_2 2}$

$$\mu_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} X_{i2} \qquad \sigma_2^2 = \frac{1}{N_2}\sum_{i=1}^{N_2}(X_{i2}-\mu_2)^2$$

$X_{1L}, X_{2L}, \cdots X_{N_L L}$

$$\mu = \frac{1}{N}\sum_{i=1}^{L}\sum_{i=1}^{N_\ell} X_{i\ell} = \frac{1}{N}\sum_{\ell=1}^{L} N_\ell \mu_\ell = \sum_{\ell=1}^{L} \mu_\ell W_\ell$$

— We want to estimate μ and ~~stratify the population~~ give the stratification to select s.r.s in a way that minimizes the standard error of the estimate.

— We select a s.r.s from each strata n_ℓ individuals from N_ℓ.

$X_{1,\ell}, X_{2\ell}, \cdots, X_{n_\ell \ell}$

$\{X_{11}, X_{21}, \cdots X_{n_1 1}\}$
$\{X_{12}, X_{22}, \cdots X_{n_2 2}\}$ } independent across, not independent within strata
$\vdots$
$\{X_{1L}, X_{2L}, \cdots X_{n_L L}\}$.

Our estimate for μ is $\bar{x} = \sum_{\ell=1}^{L} \bar{x}_\ell W_\ell$ $\quad \bar{x}_\ell = \frac{1}{n_\ell}\sum_{i=1}^{n_\ell} X_{i\ell}$ $\quad$ unbiased estimator for $\mu_\ell = E[\bar{x}_\ell]$

$$E[\bar{x}] = \mu.$$

$$VAR(\bar{X}_s) = Var\left(\sum_{l=1}^{L} \bar{X}_l w_l\right) \quad \text{since } \bar{X}_1, \bar{X}_2, \cdots \bar{X}_L \text{ are iid}$$

$$= \sum_{l=1}^{L} Var(w_l \bar{X}_l) = w_l^2 \sum_{l=1}^{L} Var(\bar{X}_l) = \sum w_l^2 \frac{\sigma_l^2}{n_l}\left(1 - \frac{n_l - 1}{N_l - 1}\right)$$

$$Var(\bar{X}_s) \simeq \sigma^2 \sum_{l=1}^{L} W_l^2/n_l \quad \text{if } \sigma^2 \text{ is the same for each strata.}$$

$$Var(\bar{X}_s) = \sum_{l=1}^{L} \sigma_l^2 W_l^2/n_l \quad \text{if } \sigma_1^2, \sigma_2^2, \cdots \sigma_L^2 \text{ are not the same}$$

- choose $n_1, n_2, \cdots n_L$ to minimize $Var(\bar{X}_s)$.

$$\min f(n_1, n_2 \cdots n_L) = \sum_{l=1}^{L} \sigma_l^2 w_l^2/n_l \quad \text{subject to } g(n_1, n_2, \cdots n_L) = n_1 + n_2 + \cdots + n_L = n$$

the method of Lagrange multipliers:

extreme values occur where $\nabla f = \lambda \nabla g$

$$\nabla f = \left(-\frac{\sigma_1^2 w_1^2}{n_1^2}, -\frac{\sigma_2^2 w_2^2}{n_2^2}, \cdots, -\frac{\sigma_L^2 w_L^2}{n_L^2}\right)$$

$$\nabla g = (1, 1, \cdots, 1)$$

$$\frac{\sigma_1^2 w_1^2}{n_1^2} = \cdots \frac{\sigma_l^2 w_l^2}{n_l^2} = \frac{\sigma_L^2 w_L^2}{n_L^2} \quad \Rightarrow \text{if } n_l = C_l n$$

$$\lambda = \frac{n}{\sum_{k=1}^{L} \sigma_k w_k} \quad \Rightarrow n_l = \frac{\sigma_l w_l}{\sum_{k=1}^{L} \sigma_k w_k} n.$$

Ex

Stratum	N_l	w_l	μ_l	σ_l	optimal sampling allocation
A	98	.249	182.9	103.4	$n_1 = 0.106n$
B	98	.249	526.9	204.8	$n_2 = 0.210n$
C	98	.249	958.3	243.5	$n_3 = 0.250n$
D	99	.251	1591.2	419.2	$n_4 = 0.434n$

If you don't know $\sigma_1^2, \cdots \sigma_L^2$ one method is to use proportional sampling

$$n_l = w_l n = \frac{N_l}{N} n.$$

Parameter estimation

It can sometimes be done by ~~incorrect~~ expressing the parameters as functions of moments

① $X \sim P(\lambda)$

$\mu_1 = E(X) = \lambda \quad \lambda = f(\mu_1) = \mu_1$

② $X \sim N(\mu, \sigma^2)$

$\mu_1 = E[X], \quad \mu_2 = E[X^2] \quad \mu = \mu_1, \quad \sigma^2 = \mu_2 - \mu_1^2$

③ $X \sim \Gamma(\alpha, \lambda)$

$\mu_1 = E[X] = \frac{\alpha}{\lambda} \quad \mu_2 = E[X^2] = \frac{(\alpha+1)\alpha}{\lambda^2}$

$\alpha = f_1(\mu_1, \mu_2) = \frac{\mu_1^2}{\mu_2 - \mu_1^2}$

$\lambda = f_2(\mu_1, \mu_2) = \frac{\mu_1}{\mu_2 - \mu_1^2}$

Method of Moments

$\hat{\mu}_1 = \frac{1}{n}\sum_{k=1}^{n} X_k$ estimates μ_1, $\hat{\mu}_2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2$ estimates μ_2 in the sense that

$P(|\hat{\mu}_1 - \mu_1| \geq \varepsilon) \to 0 \quad \text{as } n \to \infty$

$P(|\hat{\mu}_2 - \mu_2| \geq \varepsilon) \to 0$

— Moments can be approximated by sample moments by the law of large numbers

when $P(|\hat{\theta}_j - \theta_j| > \varepsilon) \to 0 \quad \forall \varepsilon > 0$. as $n \to \infty$. $\hat{\theta}_j$ is called a consistent estimator of θ_j.

if f_j is continuous where $\theta_j = f_j(\mu_1, \mu_2, \cdots, \mu_k)$. then $\hat{\theta}_j = f_j(\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_k)$ is a consistent estimator of θ_j.

Ex. If $X \sim N(\mu, \sigma^2)$. $\mu_1 = E[X]$. $\mu_2 = E[X^2]$

$\mu = \mu_1 = f_1(\mu_1, \mu_2)$

$\sigma^2 = \mu_2 - \mu_1^2 = f_2(\mu_1, \mu_2)$ since f_1, f_2 are continuous,

$\hat{\mu} = \hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$ are consistent estimators.

- the standard error $\sqrt{\text{Var(estimator)}}$ is a measure of accuracy of the estimator.

— In normal dist. the dist'n of $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$ standard error $\frac{\sigma}{\sqrt{n}}$.

Since we don't know σ^2, we resort to $\frac{\hat{\sigma}}{\sqrt{n}}$ as the S.E

The dist'n of $\hat{\sigma}^2$. $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$

$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ where Z_i iid $\sim N(0,1)$

$\text{Var}(X) = E[X^2] - E[X]^2 = 2n$

$E[X^2] = E\left[\left(\sum_{i=1}^{n} Z_i^2\right)^2\right] = E\left[\sum_{i=1}^{n} Z_i^4 + \sum_{i \neq j}^{n} Z_i^2 Z_j^2\right]$

$= n E[Z_1^4] + n(n-1)$

$= 3n + n(n-1) = n^2 + 2n$

$\text{Var}(\hat{\sigma}^2) = \text{Var}\left(\frac{\sigma^2}{n} \cdot \frac{n\hat{\sigma}^2}{\sigma^2}\right) = \frac{\sigma^4}{n^2}\text{Var}\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = \frac{\sigma^4}{n^2} 2(n-1)$

Standard Error $\sqrt{\frac{2(n-1)}{n^2}}\sigma^2 \simeq \sqrt{\frac{2(n-1)}{n^2}}\hat{\sigma}^2$

Maximum Likelihood Estimator (MLE)

$L(\theta_1, \theta_2 \cdots \theta_k) = \log f(x_1, \cdots x_k | \theta_1, \cdots \theta_k)$.

$\text{argmax} L(\theta_1, \theta_2 \cdots \theta_k) = \hat{\theta}_1, \hat{\theta}_2, \cdots \hat{\theta}_k$.

$X_1, X_2 \cdots X_n$ are iid $\sim N(\mu, \sigma^2)$

~~$N(\mu, \sigma^2)$~~.

$\nabla l(\mu, \sigma^2) = \left(\sum_{i=1}^{n} \frac{x_i - \mu}{\sigma}, -\frac{n}{\sigma} + \sigma^{-3}\sum (x_i - \mu)^2\right) \overset{\text{set}}{=} 0$.

$\hat{\mu} = \frac{1}{n}\sum x_i = \bar{X}$ $\quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$

can be used to find C.I for μ and σ^2

$\frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$.

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$

to obtain a $100(1-\alpha)\%$ CI for μ, find $t_{n-1}(\alpha/2)$ and then

$P\left(-t_{n-1}(\frac{\alpha}{2}) \leq \frac{\sqrt{n}(\bar{X} - \mu)}{S} \leq t_{n-1}(\alpha/2)\right) = 1 - \alpha$.

where ~~$t_{n-1}(\alpha)$ is the value for which~~

To obtain a $100(1-\alpha)\%$ CI for σ^2

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$P\left(\chi^2_{n-1}(1-\tfrac{\alpha}{2}) \le \frac{n\hat{\sigma}^2}{\sigma^2} \le \chi^2_{n-1}(\tfrac{\alpha}{2})\right) = 1-\alpha$$

$$\left[\frac{n\hat{\sigma}^2}{\chi^2_{n-1}(1-\frac{\alpha}{2})}, \frac{n\hat{\sigma}^2}{\chi^2_{n-1}(\frac{\alpha}{2})}\right] \text{ is a } 100(1-\alpha)\% \text{ CI for } \sigma^2$$

CI. for MLE's derived from iid samples when sampling dist'n of the MLE is not known.

This method requires large samples since it relies on the CLT.

For an iid sample, $X_1, X_2, \ldots, X_n$ with density $f(x|\theta)$

$$L(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

$\hat{\theta} = \text{argmax } L(\theta)$ · Fish Information

Define $I(\theta) = E\left[\left(\frac{d}{d\theta}\log f(X|\theta)\right)^2\right]$ where X has density $f(x|\theta)$

CLT for MLE (iid case)

If θ_0 is the true value of the parameter, then

$$\sqrt{nI(\theta_0)}(\hat{\theta}-\theta_0) \sim N(0,1) \text{ as } n\to\infty$$

$$P\left(-z(\tfrac{\alpha}{2}) \le \sqrt{nI(\theta_0)}(\hat{\theta}-\theta_0) \le z(\tfrac{\alpha}{2})\right) \approx 1-\alpha$$

$$I(\theta) = -E\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\right] \text{ for calculating } I(\theta) \text{ easier.}$$

Bayesian Approach parameter estimation →

prior distribution θ

view the data to update the opinion of the data.

→ posterior distribution

$f_\theta(\theta) = 1$

$$f_{X|\theta}(x|\theta) = \begin{cases} \theta & x=1 \\ 1-\theta & x=0 \end{cases} \quad \theta^x(1-\theta)^{(1-x)}$$

$$f_{X\theta}(x,\theta) = f_{X|\theta}(x|\theta)\cdot f(\theta) = f_{X|\theta}(x|\theta)$$

Beta distn: $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$.

$X \sim P(\lambda)$

$$f_{X|\Lambda}(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}. \qquad f_\Lambda(\lambda) = \frac{v^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-v\lambda}$$

we observe $\vec{X} = (X_1, X_2, \cdots X_n)$.

$$f_{\vec{X}|\Lambda}(x_1, x_2, \cdots x_n|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}e^{-n\lambda}.$$

$$f_{\vec{X},\Lambda}(x_1, x_2, \cdots x_n, \lambda) = f_{\vec{X}|\Lambda}(x_1 x_2 \cdots x_n|\lambda)\cdot f_\Lambda(\lambda)$$

$$f_{\Lambda|\vec{X}}(\lambda|x_1, x_2 \cdots x_n) = \frac{f_{\vec{X}\Lambda}(x_1, x_2 \cdots x_n \lambda)}{f_{\vec{X}}(x_1, x_2, \cdots x_n)} = C\,\lambda^{\sum_{i=1}^n x_i + \alpha - 1}e^{-(n+v)\lambda}$$

$$C = \frac{(n+v)^{\sum x_i + \alpha}}{\Gamma(\sum x_i + \alpha)}$$

EX. case 1. $f_X(x|\theta,\xi) = \sqrt{\frac{\xi}{2\pi}}e^{-\frac{\xi(x-\theta)^2}{2}}$ θ unknow. ξ known. ($\xi = \frac{1}{\sigma^2}$)

give θ a prior distribution of $N(\theta_0, \xi_{prior})$

$$f_\theta(\theta) = \sqrt{\frac{\xi_{prior}}{2\pi}}e^{-\frac{(\theta-\theta_0)^2\cdot\xi_{prior}}{2}}$$

Given $\overset{iid}{}$ observations $X_1, X_2, \cdots X_n$. $f_{X|\theta}(x|\theta) = \sqrt{\frac{\xi}{2\pi}}^n \exp\left\{-\frac{\xi}{2}\sum(X_i-\theta)^2\right\}$

We're interested in $f_{\theta|X}(\theta|\vec{X}) \propto f_{X|\theta}(x|\theta)f(\theta)$

$$= \sqrt{\frac{\xi}{2\pi}}^n e^{-\frac{\xi}{2}\sum(X_i-\theta)^2}\sqrt{\frac{\xi_{prior}}{2\pi}}e^{-\frac{\xi_{prior}}{2}(\theta-\theta_0)^2}$$

$$\propto \exp\left\{-\frac{\xi}{2}\sum(X_i-\theta)^2 - \frac{\xi_{prior}}{2}(\theta-\theta_0)^2\right\}$$

$$\propto \exp\left\{-\frac{n\xi+\xi_{prior}}{2}\left(\theta - \frac{n\xi\bar{X}+\xi_{prior}\theta_0}{n\xi+\xi_{prior}}\right)^2\right\} \sim N\left(\frac{n\xi\bar{X}+\xi_{prior}\theta_0}{n\xi+\xi_{prior}}, \left(n\xi+\xi_{prior}\right)^{-1}\right)$$

Case 2: θ Known, ξ unknown.

We give ξ a prior distribution $f_\xi(\xi)$, we derive its posterior given observed data.

book 292.

$$f(\xi) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \xi^{\alpha-1} e^{-\lambda\xi} \qquad \xi > 0$$

$$f(\xi|\vec{X}) \propto \xi^{\frac{n}{2}} e^{-(\frac{1}{2}\sum_i (X_i-\theta)^2)\xi} \xi^{\alpha-1} e^{-\lambda\xi} = \xi^{\frac{n}{2}+\alpha-1} e^{-(\frac{1}{2}\sum(X_i-\theta)^2+\lambda)\xi}$$

$$= \frac{(\frac{n\hat{\sigma}^2}{2}+\lambda)^{\frac{n}{2}+\alpha}}{\Gamma(\frac{n}{2}+\alpha)} \xi^{\frac{n}{2}+\alpha-1} e^{-(\frac{n\hat{\sigma}^2}{2}+\lambda)\xi}$$

Case 3. both unknown — P.293 book

$-1 < X = \cos\theta < 1 \qquad f(x|\alpha) = \frac{1+\alpha x}{2} \quad -1 \le x \le 1.$

— Method Moment Estimator for α.

$$\mu = E[X] = \int_{-1}^{1} x \cdot \frac{1+\alpha x}{2} dx = \frac{\alpha}{3}$$

$\hat{\alpha} = 3\bar{X}^{-1}$ is the MME for α.

— MLE for α.

$$\ell(\alpha) = \sum_{i=1}^{n} \ln f(X_i|\alpha) = \sum_{i=1}^{n} \ln(1+\alpha X_i) - n\ln 2$$

$$\ell'(\alpha) = \sum_{i=1}^{n} \frac{X_i}{1+\alpha X_i} = 0 \Rightarrow \hat{\alpha}$$

Efficiency

$$\frac{Var(\tilde{\alpha})}{Var(\hat{\alpha})} = \text{efficiency of } \hat{\alpha} \text{ relative to } \tilde{\alpha}$$

$$Var(\tilde{\alpha}) = 9\,Var(\bar{X}) = \frac{9}{n} Var(X_1) = \frac{9}{n}\left[E[X_i^2] - \frac{\alpha^2}{9}\right]$$

$$\sqrt{nI(\alpha)}\,(\hat{\alpha}-\alpha) \sim N(0,1)$$

$$Var(\hat{\alpha}) \approx \frac{1}{nI(\alpha)}$$

$$I(\alpha) = E\left[\left(\frac{\partial}{\partial\alpha} \ln f(x|\alpha)\right)^2\right]$$

$$= \int_{-1}^{1} \left(\frac{x}{1+\alpha x}\right)^2 \frac{1+\alpha x}{2} dx$$

MLE for λ in $P(\lambda)$.

$$f(x_1, x_2, \cdots x_n | \lambda) = \frac{\lambda^{\sum x_i}}{\prod_{i=1} x_i!} e^{-n\lambda}$$

$$\ell(\lambda) = \log f(x_1, x_2, \cdots x_n | \lambda) = n\bar{x} \log \lambda - n\lambda - \log \prod x_i!$$

$$\ell'(\lambda) = \frac{n\bar{x}}{\lambda} - n = 0 \qquad \bar{x} = \hat{\lambda}$$

Cramér-Rao

given iid sample $X_1 X_2 \cdots X_n$, $T(X_1 X_2 \cdots X_n)$.

If $\hat{\theta}$ is an unbiased estimator of θ then

$$\boxed{\mathrm{Var}(T) \geq \frac{1}{nI(\theta)}}$$

provided the density $f(x|\theta)$ of the X_i's is smooth.

$$I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \log f(X|\theta)\right)^2\right] = -E\left[\frac{\partial^2}{\partial \theta^2} \log f(X|\theta)\right]$$

$$\mathrm{Var}(\bar{X}) = \frac{\lambda}{n}.$$

$$I(\lambda) = -E\left[-\frac{X}{\lambda^2}\right] = \frac{1}{\lambda^2} E[X] = \frac{1}{\lambda}$$

Sufficiency A statistics $T(X_1, X_2, \cdots X_n)$ is sufficient for θ if the conditional dist'n of $X_1, X_2, \cdots X_n$ given $T(X_1, X_2 \cdots X_n)$ does not depend on θ.

Ex. $f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ the joint density of an iid sample $X_1, X_2 \cdots X_n$.

$$f(X_1, X_2 \cdots X_n | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n X_i^2 - 2\mu \sum_{i=1}^n X_i + n\mu^2\right)}$$

(*) it suggests $T_1, T_2 = (\sum X_i, \sum X_i^2)$ are sufficient stats for (μ, σ^2).

$$= g(T_1(X_1, X_2, \cdots X_n), T_2(X_1, X_2, \cdots X_n), \mu, \sigma)$$

by Factorization theorem. (*) holds

Factorization Theorem

T be a sufficient stats is that $f(x_1, x_2, \cdots x_n|\theta) = g(T(x_1, x_2 \cdots x_n), \theta)\, h(x_1, x_2 \cdots x_n)$ for some function g and h.

Ex $X_1, X_2, \cdots X_n$ iid $P(\lambda)$. λ unknown.

$T(X_1, X_2, \cdots X_n) = \bar{X}$ is MLE for λ. Is T sufficient? Yes

$$f(x_1, x_2, \cdots x_n|\lambda) = \prod_{i=1}^{n}\left(\frac{\lambda^{x_i}}{x_i!}e^{-\lambda}\right) = \frac{\lambda^{\sum x_i}}{\prod x_i!}e^{-n\lambda}$$

$$= \underbrace{\left(\lambda^{\sum x_i} e^{-n\lambda}\right)}_{g(T,\lambda)\cdot h(T)}\left(\prod_{i=1}^{n} x_i\right)^{-1}$$

— If T is a sufficient stats for θ then so is $K(T)$ for K any function.

Exponential families of densities.

These densities have the form

$$f(\vec{x}|\theta) = \prod_{i=1}^{n} \exp\{c(\theta)K(x_i) + d(\theta) + S(x_i)\}$$

$$= \exp\left\{c(\theta)\sum_{i=1}^{n} k(x_i) + nd(\theta) + \sum_{i=1}^{n} S(x_i)\right\}$$

$$= \exp\{\underbrace{c(\theta)\sum_{i=1}^{n} K(x_i) + nd(\theta)}_{g}\} \cdot \underbrace{\exp\{\sum_{i=1}^{n} S(x_i)\}}_{h}$$

$T(X_1, X_2, \cdots X_n) = \sum_{i=1}^{n} K(x_i)$ is a sufficient statistic.

Ex. $f(x_1, x_2, \cdots x_n|\lambda, \alpha) = \left(\frac{\lambda^\alpha}{\Gamma(\alpha)}\right)^n \cdot e^{-\lambda\sum x_i + (\alpha-1)\sum \log x_i}$

$x_i^{\alpha-1} \cdot e^{-\lambda x_i}$

$= e^{-\lambda x_i + (\alpha-1)\log x_i}$ $= e^{-\lambda\sum x_i + (\alpha-1)\sum \log x_i + n\log\frac{\lambda^\alpha}{\Gamma(\alpha)}}$

$(\sum x_i, \sum \log x_i)$ = sufficient statistic

Sufficient stats and MLEs

$$f(x_1, x_2, \cdots x_n \mid \theta) = g(\theta, T(x_1, x_2, \cdots x_n)) \cdot h(T(x_1, x_2, \cdots x_n))$$

$$\ell'(\theta) = \frac{\frac{d}{d\theta} g(\cdot)}{g(\cdots)} = 0 \qquad \text{we need to solve } \frac{d}{d\theta} g(T(x_1, x_2, \cdots x_n), \theta) = 0$$

$$g_\theta(T(x_1, x_2 \cdots x_n), \hat{\theta}) = 0 \Rightarrow \hat{\theta} = \{(T(x_1, x_2, \cdots x_n))\}$$

Rao-Blackwell Theorem

If $\hat{\theta}$ is an estimator of θ w/ $E[\hat{\theta}^2] < \infty$ for all values of θ and

If T is a sufficient statistic for θ, then.

$$\tilde{\theta} = E[\hat{\theta} \mid T] \text{ satisfies}$$

$$E[(\tilde{\theta} - \theta)^2] \leq E[(\hat{\theta} - \theta)^2] \quad \text{w/ equality holds} \iff \tilde{\theta} = \hat{\theta}$$

Hypothesis testing

H_0 null hypothesis

H_1 alternative hypothesis

Ind. trails are run and based on some test statistic. H_0 is either accepted or rejected

Simple one:
$H_0: \theta = \theta_0$
$H_1: \theta = \theta_1$

Composite one:
$H_0: \theta = \theta_0$
$H_1: \theta \geq \theta_1 > \theta_0$

An iid sample is taken from the unknown dist'n $f(x|\theta)$
$\vec{X} = (X_1, X_2, \cdots X_n)$

In the case of a simple hypothesis, you'd look at

$$\frac{f(\vec{x}|\theta_0)}{f(\vec{x}|\theta_1)}$$

Find some acceptance level X_0 and

Your test says:

if $\frac{f(\vec{x}|\theta_0)}{f(\vec{x}|\theta_1)} \geq X_0$ accept H_0

if $\frac{f(\vec{x}|\theta_0)}{f(\vec{x}|\theta_1)} < X_0$ reject H_0

$\alpha = p(\text{reject } H_0 \mid \text{given } H_0 \text{ is true})$ α = significance level.

Type I error.

$$= p\left(\frac{f(\vec{x}|\theta_0)}{f(\vec{x}|\theta_1)} < X_0 \mid H_0\right)$$

Another error. **Type II error**, is to accept H_0 when H_1 is true.

$\beta = p(\text{accept } H_0 \mid H_1)$

$1-\beta = p(\text{reject } H_0 \mid H_1)$ = power of test

EX. $X_1 X_2 \cdots X_n$ iid $N(\theta, \sigma^2)$. θ unknown. σ^2 known.

$H_0: \theta = \theta_0$
$H_1: \theta = \theta_1$

where $\theta_0 > \theta_1$

$$\frac{f(\vec{x}|\theta_0)}{f(\vec{x}|\theta_1)} = \frac{\frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{1}{2\sigma^2}\sum(x_i-\theta_0)^2}}{\frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{1}{2\sigma^2}\sum(x_i-\theta_1)^2}} = e^{-\frac{1}{2\sigma^2}\left(\sum(x_i-\theta_0)^2+\sum(x_i-\theta_1)^2\right)}$$

$$= \exp\left\{-\frac{n}{2\sigma^2}\left(2\bar{x}(\theta_1-\theta_0)+\theta_0^2-\theta_1^2\right)\right\} \geq \lambda_0$$

$$2\bar{x}(\theta_1-\theta_0)+\theta_0^2-\theta_1^2 \leq y_0 = -\frac{2\sigma^2}{n}\log\lambda_0$$

If $\bar{x} \geq \frac{y_0+\theta_1^2-\theta_0^2}{2(\theta_1-\theta_0)}$, accept H_0.

If $\bar{x} <$ ---- reject H_0.

~~$A = \{(x_1, x_2 \cdots x_n) : \cdots\}$~~ $A = \left(\frac{y_0+\theta_1^2-\theta_0^2}{2(\theta_1-\theta_0)}, \infty\right)$ is called acceptance region.

A^c rejection region.

$\bar{x}$: test statistic.

Given a significance level α, how can we set y_0 so that.

$$P(\text{reject } H_0 | H_0) = \alpha$$

$$P\left(\bar{x} \leq \frac{y_0+\theta_1^2-\theta_0^2}{2(\theta_1-\theta_0)} \middle| H_0\right) = \alpha$$

$\bar{x} \sim N(\theta_0, \sigma^2/n)$

$$P\left(\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}} \leq \underbrace{\left(\frac{y_0+\theta_1^2-\theta_0^2}{2(\theta_1-\theta_0)} - \theta_0\right)/\sigma/\sqrt{n}}_{=-z(\alpha)} \middle| H_0\right) = \alpha.$$

$$\beta = P(\text{accept } H_0 | H_1).$$

$$1-\beta = P(\text{reject } H_0 | H_1) = P\left(\bar{x} \leq \frac{y_0+\theta_1^2-\theta_0^2}{2(\theta_1-\theta_0)} \middle| H_1\right).$$

$\uparrow \sim N(\theta_1, \sigma^2/n)$.

Neyman-Pearson Lemma

Among all tests w/ significance level α, the likelihood ratio test is ~~most~~ most powerful.

what about composite hypothesis

$H_0 : \mu = \mu_0$

$H : \mu > \mu_0$

Consider an iid sample $X_1, X_2, \cdots, X_n$ from a $N(\mu, \sigma^2)$ dist'n.

w/ σ^2 known, μ unknown, and test

$H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$

accept H_0 if $|\bar{X} - \mu_0| \leq X_0$

reject H_0 if $|\bar{X} - \mu_0| > X_0$

$\alpha = p(\text{reject } H_0 \mid H_0) = p(|\bar{X} - \mu_0| \geq X_0 \mid H_0)$ $\quad \bar{X} \sim N(\mu_0, \frac{\sigma^2}{n})$

$= p\left(\frac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}} \geq \frac{X_0}{\sigma/\sqrt{n}} \mid H_0\right)$

$\mu_0 \in \left[\bar{X} - z(\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}},\ \bar{X} + z(\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}\right]$ $\quad = p(|Z| \geq \frac{X_0}{\sigma/\sqrt{n}}) = \alpha.$ $\quad X_0 = z(\frac{\alpha}{2})\sigma/\sqrt{n}$

$(1-\alpha)\%$ C.I for μ_0

$\beta = p(\text{reject } H_0 \mid H_1)$. — cannot be computed. Since we don't know the dist'n of $\bar{X}$ under H_1.

Sometimes, hypotheses have the form $\quad H_0: \theta \in \omega_0$ $\quad H_1: \theta \in \omega_1$. $\quad \omega_0 \cap \omega_1 = \phi$

Generalized Likelihood Ratio Test

Set $\Omega = \omega_0 \cup \omega_1$.

$\ell(\theta) = \log f(x_1, x_2, \cdots x_n \mid \theta)$

define $\lambda = \dfrac{\max_{\theta \in \omega_0} \ell(\theta)}{\max_{\theta \in \omega_1} \ell(\theta)}$

Then a test would take the form. accept H_0 if $\lambda \geq c$.

reject H_0 if $\lambda \leq c$.

Ex. $X_1, X_2, \cdots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$H_0: \mu \in \omega_0 = \{\mu_0\}$ $\quad (\mu = \mu_0)$

$H_1: \mu \in \omega_1 = \mathbb{R} - \{\mu_0\}$ $\quad (\mu \neq \mu_0)$.

$\max_{\theta \in \omega_0} \ell(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{1}{2\sigma^2}\sum (x_i - \mu_0)^2}$

$\max_{\theta \in \omega_1} \ell(\theta) = \ell(\hat{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{1}{2\sigma^2}\sum (x_i - \bar{x})^2}$ $\quad$ MLE $\hat{\theta} = \bar{x}$

$\lambda = \exp\left\{\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (x_i - \bar{x})^2 - \sum_{i=1}^n (x_i - \mu_0)^2\right]\right\}$

$= \exp\left\{\frac{n}{2\sigma^2}(\bar{x} - \mu_0)^2\right\}$.

$$-2\log\lambda = \left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_1$$

$\lambda \geq c$ becomes $-2\log\lambda \leq -2\log c$

The test becomes $\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 \leq -2\log c$

CLT for Generalized Likelihood Ratio Test

If the underlying densities are smooth for all values of the parameter θ and $\lambda = \frac{\max_{\theta \in \omega_0} l(\theta)}{\max_{\theta \in \Omega} l(\theta)}$ and $\dim\Omega - \dim\omega_0 = k$, then

$-2\log\lambda$ has an approximate χ^2_k for large n.

multinomial

$$f(x_1, x_2, \ldots, x_m \mid p_1, \ldots p_m) = \frac{n!}{\prod_{i=1}^{n} x_i!} \cdot \prod_{i=1}^{m} p_i^{x_i}$$

parameters $\vec{p} = (p_1, p_2, \ldots p_m)$

where $p_i > 0$ and $\sum p_i = 1$

p-Value := smallest significance level at which null hypothesis is rejected.

- probability of an outcome for T at least as extreme as the observed one t.

$$p(|T| > t \mid H_0) \text{ — p-value}$$

Poisson Dispersion Test

X_i : # particles in i^{th} cell

Assume dist'n of particles in each cell is Poisson and we want to test

H_0 : λ is the same in all cells

H_1 : λ_i = parameter in i^{th} cell may vary from cell to cell.

use Generalized Likelihood Ratio Test.

$$\lambda = \frac{\max_{\lambda \in \omega_0} l(\lambda)}{\max_{\Omega} l(\lambda_1, \lambda_2 \ldots \lambda_n)}$$

$\omega_0 = [0, \infty)$ $\dim\omega_0 = 1$

$\Omega = (0, \infty)^n$ $\dim\Omega = n$

$$\prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

$$\hat{\lambda} = \bar{x} = \frac{1}{n}\sum x_i$$

$$\Lambda = \frac{\prod_{i=1}^{n} \frac{\hat{\lambda}^{x_i}}{x_i!} e^{-\hat{\lambda}}}{\max \prod_{i=1}^{n} \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i}}$$

$$\frac{d}{d\lambda_i}\left(\log \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i}\right) = \frac{d}{d\lambda_i}(x_i \log \lambda_i - \lambda_i)$$

$$= \frac{x_i}{\lambda_i} - 1 = 0$$

$$x_i = \lambda_i$$

$$\bar{\lambda}_i = x_i$$

$$= \frac{\prod_{i=1}^{n} \frac{\hat{\lambda}^{x_i}}{x_i!} e^{-\hat{\lambda}}}{\prod_{i=1}^{n} \frac{\bar{\lambda}_i^{x_i}}{x_i!} e^{-\bar{\lambda}}} = \prod_{i=1}^{n} \left(\frac{\hat{\lambda}}{\bar{\lambda}_i}\right)^{x_i} e^{\bar{\lambda}_i - \hat{\lambda}} = \prod_{i=1}^{n} \left(\frac{\bar{x}}{x_i}\right)^{x_i} e^{x_i - \bar{x}}$$

when the number of observations is large.

$$-2\log\Lambda \sim \chi^2_{n-1}$$

$$-2\log\Lambda = -2\log\Lambda$$

$$-2\sum_{i=1}^{n} x_i \log\frac{\bar{x}}{x_i} - 2\sum_{i=1}^{n}(x_i - \bar{x})$$

$$= 2\sum_{i=1}^{n} x_i \log\frac{x_i}{\bar{x}}$$

$$\simeq \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\bar{x}} = \boxed{\frac{n\hat{\sigma}^2}{\bar{x}}}$$

is called dispersion

$$f(x) = x\log\frac{x}{\bar{x}} = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + f''(\bar{x})\frac{(x-\bar{x})^2}{2}$$

$$0 \qquad\qquad = 0$$

Σ of $= 0$

Test: If $\frac{n\hat{\sigma}^2}{\bar{x}} > c$ reject

To obtain a significance level of α, we find c s.t. $P_0\left(\frac{n\hat{\sigma}^2}{\bar{x}} > c\right) = \alpha$

Regression

iid random points (X_i, Y_i), $i = 1, 2, \cdots, n$

$y = a + bx$.

$y = a + bx + \varepsilon$ ← everything we don't know

It's not due to randomness.

Goal: Find $\hat{a}, \hat{b}$ that approach a, b.

$X \in \mathbb{R}$ univariate regression

$X \in \mathbb{R}^p$ multivariate regression.

$Y = a + X^T b$.

Why we do this?

ML. relation between X and Y.

Linear function is the simplest.

Ex. $D_i = a + b p_i$

↑ Demand. ↖ price

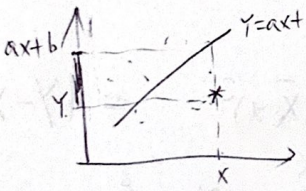
$PV = nRT$ (ideal gas law).

Variable transformation

take { exponential, square, log, et.

Ex. $\log p + \log V = \log nR + \log T$

$\log p = \log nR + \log T - \log V$



$E[Y - (ax+bx)]^2$

function of a and b that measures how close the point (X, Y) is to the line $y = a + bx$ in expectation

assumption: $Var(X) \neq 0$

~~$\tilde{X} = X - E(X)$~~

$\min_{a,b} E[y - (a+bx)]^2$ ~~$= E[y - (a + bE(X) + b\tilde{X})]^2$~~

$\tilde{a} \to a$.

~~$\frac{\partial}{\partial a} = -E[2(y - (\tilde{a} + b\tilde{x}))] = 0$~~

~~$\tilde{a} = E\tilde{E}$~~

~~$\frac{\partial}{\partial b} = -E[2(y - (\tilde{a} + b\tilde{x}))]\tilde{x} = 0$~~

$$\begin{cases} a = E(Y) - bE(X) \\ b = \dfrac{cov(X,Y)}{Var(X)} \end{cases}$$

← replace by $\bar{X}$, S_n

$E(X)$

$Var(X)$

$Y = a + bX + \varepsilon \iff \varepsilon = Y - (a+bX)$. $E(\varepsilon) = E(Y) - [a + bE(X)] = 0$

$cov(X, \varepsilon) = cov(X, Y-(a+bX)) = cov(X,Y) - (a + b\,cov(X,X))$

$= E[X(Y-(a+bX))]$ ~~$= cov(X,Y)$~~

$= 0$

If $X \perp \varepsilon$, then $cov(X, \varepsilon) = 0$

Assumption $\begin{cases} E(\varepsilon) = 0 \\ Cov(X, \varepsilon) = 0 \end{cases}$

We need more when we do more analysis. e.g. t-test we need ε to be Gaussian.

$Y_i = a + bX_i + \varepsilon_i$ $\quad \varepsilon_i$ i.i.d

$E(\varepsilon_i) = 0$

$cov(X_i, \varepsilon_i) = 0$

$\hat{b} = \frac{cov(X,Y)}{Var(X)} = \frac{E(XY) - E(X)E(Y)}{E(X^2) - (E(X))^2} = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2}$

$E[(Y-(a+bX))^2] \rightsquigarrow \frac{1}{n}\sum_{i=1}^{n} (Y-(a+bX_i))^2$

Multivariate case

$Y_i = X_i^T \beta + \varepsilon_i \quad i = 1, 2, \dots n. \quad cov(X_i, \varepsilon_i) = 0$

Y_i: dep. var. / response var. $\quad X_i$: covariates / ind. var.

$\hat{\beta} = \underset{t \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} (Y_i - \bar{X}_i x)^2 = \|Y - X\beta\|_2^2$

Matrix notation

$$\underset{Y}{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}} = \underset{X}{\begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{p-1} \\ & & & & \\ 1 & x_n & x_n^2 & \cdots & x_n^{p-1} \end{pmatrix}} \underset{\beta}{\begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_p \end{pmatrix}} + \underset{\varepsilon}{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}$$

$\min \|Y - X\beta\|^2 = \|Y\|^2 + \|X\beta\|^2 - 2X^T X\beta$

$= \|Y\|^2 + \beta^T X^T X\beta - 2X^T X\beta$

$\frac{\partial}{\partial \beta} = 0 + 2X^T X\beta - 2X^T y \Rightarrow \beta = (X^T X)^{-1} X^T y$

if $p > n$, Least squared doesn't exist.

$\hat{X\beta}$ $\quad X(X^T X)^{-1} X y$ (orthogonal) projection of y onto the subspace spanned by column of X

Statistical Inference

assumptions:

1) X_i deterministic
2) the model is homoscedastic: $\varepsilon_1, \varepsilon_2, \dots \varepsilon_n$ are iid
3) $\varepsilon \sim N_n(0, \sigma^2 I_p)$. Unknown $\sigma^2 > 0$

$$\hat{\beta} = (X^T X)^{-1} X^T y \qquad y = X\beta + \varepsilon$$
$$= (X^T X)^{-1} X^T X \beta + (X^T X)^{-1} X^T \varepsilon$$
$$= \beta + (X^T X)^{-1} X^T \varepsilon$$
$$= \beta + N_p(0, (X^T X)^{-1} X^T X (X^T X)^{-1} \sigma^2)$$
$$= \beta + N_p(0, (X^T X)^{-1} \sigma^2)$$

1) LSE = MLE $\qquad Y \sim N_p(X\beta, \sigma^2 I)$

$$p(y) = \frac{1}{(\sigma^2 2\pi)^{1/2}} \exp\left\{ - \frac{\|y - X\beta\|_2^2}{2\sigma^2} \right\}$$

likelihood

$$\log p(y) = -\frac{n}{2} \log(\sigma^2 2\pi) - \frac{1}{2\sigma^2} \|y - X\beta\|_2^2$$

$$MLE = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - X\beta\|_2^2 = LSE$$

2) Quadratic risk of $\hat{\beta}$:

$$\hat{\beta} - \beta \sim N_p(0, (X^T X)^{-1} \sigma^2)$$

$\varepsilon \sim N(0, \Sigma)$

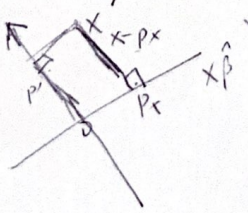
$E\|\varepsilon\|^2 = Tr(\Sigma)$

↖ trace of a square matrix: the sum of elements on the main diagonal

$$E\left[\|\hat{\beta} - \beta\|_2^2\right] = \sum_{i=1}^{n} E(\hat{\beta}_j - \beta_j)^2$$
$$= \sum_{i=1}^{n} \underbrace{Var(\hat{\beta}_j)}_{s_j^2}$$
$$= \sigma^2 Tr(X^T X)^{-1}$$

3) prediction error

$$Y - X\hat{\beta} = \cancel{IY} - X(X^T X)^{-1} X^T y = \underbrace{\left[I - X(X^T X)^{-1} X^T \right]}_{P'} Y$$



P' projection onto orthogonal of column span of X

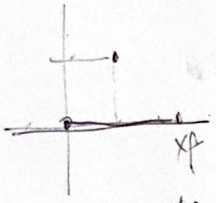
$$E\|y - X\hat{\beta}\|^2 = E\|(I - \underbrace{X(X^TX)^{-1}X^T}_{P'})y\|^2 = E\|N_n(0, \sigma^2 \underbrace{P'(P')^T}_{P'P'=P'})\|$$

$$= \sigma^2 \underbrace{\mathrm{Tr}(P')}_{n-p} \quad \text{rank theorem}$$

$$= \sigma^2(n-p)$$

$y = X\beta + \varepsilon$

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



4). unbiased estimator of σ^2

$$\hat{\sigma}^2 = \frac{1}{n-p}\|y - X\hat{\beta}\|_2^2$$

$\hat{\beta} \perp \hat{\sigma}^2$ ~~two diff~~ projections

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$$

Tests

$H_0: \beta_j = 0$

$H_1: \beta_j \neq 0$

$$\hat{\beta} = \beta + N(0, \sigma^2(X^TX)^{-1})$$

$$\hat{\beta}_j \sim e_j N(0, \sigma^2(X^TX)^{-1}) + \beta_j$$

$$= \beta_j + N(0, \sigma^2 \underbrace{(X^TX)^{-1}_{jj}}_{r_j}).$$

Test: $\left|\frac{\hat{\beta}_j}{\sigma\sqrt{r_j}} \sim N(0,1)\right| > q_{\alpha/2}(N(0,1))$ if σ^2 is known.

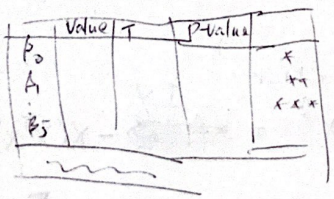
if σ^2 is ~~un~~known, we have $\hat{\sigma}^2$

$$\frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{r_j}} = \frac{\hat{\beta}_j}{\sigma\sqrt{r_j}} \cdot \frac{1}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} \cdot \sqrt{n-p} = \frac{\hat{\beta}_j/\sigma\sqrt{r_j}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}(n-p)}/(n-p)} \sim N(0,1)$$

$$\sim \chi^2_{n-p}$$

$$\sim t_{n-p}$$

$$\left|\frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{r_j}}\right| > q_{\alpha/2}(t_{n-p})$$



$H_0: \beta_j = 0$ for all $j \in S \subset \{1, \dots p\}$

$H_1: \beta_j \neq 0$ for some $j \in S$

P_{H_0} (reject β_1 ∪ reject β_2 ⋯ ∪ reject β_s) $\leq \alpha$

if they disjoint

$\frac{\hat{\beta}}{\hat{\sigma}\sqrt{r_j}} > q_{\alpha/2}(t_{n-p})$.

$\rightarrow q_{\alpha/2p}(t_{n-p}) \rightarrow P\alpha/p = \alpha$.

Bonferroni correction. (multiple testing simultaneous).

$H_0: G\beta = \lambda \quad H_1: G\beta \neq \lambda$.

$G\hat{\beta} = G\beta + \mathcal{N}(0, \sigma^2 (X^T X)^{-1})$.

$$\frac{G\hat{\beta} - \lambda}{\hat{\sigma}} \sim \frac{N(0, G(X^T X)^{-1} G^T)}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}}$$

if $X \sim \mathcal{N}_k(0, \Sigma)$

$X^T \Sigma^{-1} X \sim \chi^2_k$.

$$\hat{\sigma}^{-2} (G\hat{\beta} - \lambda)^T \left(G(X^T X)^{-1} G^T\right)^{-1} (G\hat{\beta} - \lambda) \sim \chi^2_k$$

Let $\frac{\hat{\sigma}^{-2} (G\hat{\beta} - \lambda)^T \left(G(X^T X)^{-1} G^T\right)^{-1} (G\hat{\beta} - \lambda)}{k} \sim \frac{\chi^2_k}{k}$.

Bayesian statistics

Beta distribution $\text{Beta}(a, b)$. $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

$$f(x) = \begin{cases} c x^{a-1} (1-x)^{b-1} & x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

posterior $p(\theta | X)$

Beta is conjugate prior.

$X_1, X_2, \dots X_n \overset{iid}{\sim} \text{Ber}(p)$ $\quad p(\theta | X) \propto p^{a-1}(1-p)^{b-1} \cdot p^{\sum X_i} (1-p)^{n - \sum X_i}$

$= p^{a-1+\sum X_i} \cdot (1-p)^{b-1+n-\sum X_i}$

Non informative priors

- Good candidate. $\pi(\theta) \propto 1$ constant pdf.
- If Θ is bounded, this is uniform prior on Θ
- If Θ is unbounded. this does not define a proper pdf on Θ.

- if $p \sim \mathcal{U}(0,1)$ $p, X_1, X_2 \cdots X_n \overset{iid}{\sim} Ber(p)$

  posterior $B(1+\Sigma X_i, 1+n-\Sigma X_i)$

- if $\pi(\theta) = 1$. given $\theta, X_1, X_2, \cdots X_n \overset{iid}{\sim} N(0,1)$

  posterior $N(\bar{X}_n, \frac{1}{n})$

GLMs

$$Y|X \sim N(\mu(X), \sigma^2 I)$$
$$= N(X^T\beta, \sigma^2 I)$$

$X \longrightarrow X^T\beta$

1. Random component:

$$Y|X \sim N(\mu(X), \sigma^2 I)$$

$\rightarrow Y|X \sim$ Some distribution from the exponential family

2. $g(\mu(X)) = X^T\beta \iff \mu(X) = f(X^T\beta)$.

   g: link function. $f = g^{-1}$.

   log link function

   reciprocal link function

exponential family

$$P_\theta(x) = p(\theta, x) = \exp(\theta \cdot x) \cdot f(x) \cdot g(\theta) = \exp\left[\sum_{i=1}^{K} \eta_i(\theta) T_i(x)\right] c(\theta) h(x)$$

$\eta_1(\theta) \cdot T_1(x)$

$\eta_2(\theta) \cdot T_2(x)$

$\eta_K(\theta) \cdot T_K(x)$

Ex. $X \sim N(\mu, \sigma^2)$ $\theta = (\mu, \sigma^2)$

$$P_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$= \exp\left\{-\frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \log(\sigma\sqrt{2\pi})\right\}$$

$\eta_1 = \frac{\mu}{\sigma^2}$, $\eta_2 = -\frac{1}{2\sigma^2}$ $T_1(x) = x$, $T_2(x) = x^2$.

$B(\theta) = \frac{\mu^2}{2\sigma^2} + \log(\sigma\sqrt{2\pi})$, $h(x) = 1$.

When σ^2 is Known.

$\eta = \frac{\mu}{\sigma^2}$. $T(x) = x$. $B(\theta) = \frac{\mu^2}{2\sigma^2}$. $h(x) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$.

Bernoulli(p). $p^x(1-p)^{1-x}$

Poission $\frac{\lambda^x}{x!}e^{-\lambda}$.

Gamma(a,b) $\frac{1}{\Gamma(a)b^a}x^{a-1}e^{-\frac{x}{b}}$ a: shape parameter, b: scale parameter.

reparametrize: $\mu = ab$

$$\frac{1}{\Gamma(a)}\left(\frac{a}{\mu}\right)^a x^{a-1} e^{-\frac{ax}{\mu}}$$

Inverse Gamma(α, β): $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/b}$.

Inverse Gaussian(μ, σ^2): $\sqrt{\frac{\sigma^2}{2\pi x^3}}\, e^{-\frac{\sigma^2(x-\mu)^2}{2\mu^2 x}}$

chi-square, Beta, Binomial, Negative binomial.