Data analysis

Ben Graham

MA930, University of Warwick

October 19, 2015

Intro

- MLE
- Confidence intervals
- Bayesian credible intervals
- p-values
- Hypothesis testing

6.2 Sufficient statistics

- Def 6.2.1 A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.
- Thm 6.2.2 If $f(x \mid \theta)/f(T(x) \mid \theta)$ is constant, then $T(X)$ is sufficient.
- Thm 6.2.6 $T(X)$ is sufficient iff $f(x \mid \theta) = g(T(x) \mid \theta)h(x)$ for some g, h
- Example: Independent $X_i \sim$Bernoulli(θ), $\theta \in (0,1)$.
- Example: Independent $X_1, \ldots, X_N \sim$Uniform$(0,\theta)$, $\theta > 0$.
- Example: Independent $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$, $\theta \in \mathbb{R}$
- Example: Independent $X_1, \ldots, X_n \sim N(\theta_1, \theta_2^2)$, $\theta_1 \in \mathbb{R}, \theta_2 > 0$
- Minimal sufficient statistics

6.3 Likelihood principle

- Random sample $X = (X_1, \ldots, X_n)$
- $X_i \sim f(x_i \mid \theta)$ pmf or pdf
- $X \sim \prod_i f(x_i \mid \theta) = f(x \mid \theta)$
- Likelihood function

$$L(\theta \mid x) = f(x \mid \theta)$$

- Likelihood principle: if

$$L(\theta \mid x)/L(\theta \mid y)$$

  is independent of θ, then the conclusions drawn from x and y should be identical.

Chapter 7 Point estimation

7.2.2 Maximum Likelihood Estimator

- $L(\theta \mid x) = \prod_i f(x_i \mid \theta)$, $\theta \in \mathbb{R}^k$
- MLE: Statistic $\hat{\theta}(x) = \arg\max_\theta L(\theta \mid x)$
- Differentiable? Solve $\frac{\partial}{\partial \theta_i} L(\theta \mid x) = 0$, $i = 1, \ldots, k$
- log-likelihood $\ell(\theta) = \log L(\theta)$
- Ex: $N(\mu, \sigma^2)$
- Ex: Uniform$(0, \theta)$, $\theta > 0$.
- Theorem 7.2.10: Invariance: The MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Lagrange multipliers

- For maximizing/minimizing $f : \mathbb{R}^n \to \mathbb{R}$ subject to $g : \mathbb{R}^n \to \mathbb{R}$, $g(x) = 0$.
- Example: multinomial distribution

Newton Raphson method

- For finding roots of an equation

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

- Ex: $\sqrt{2}$ is a root of the equation $x^2 - 2 = 0$
- $f(x) = x^2 - 2$
- $$x_{n+1} = x_n - \frac{x^2 - 2}{2x}$$

7.2.3 Bayes Estimators

- Parameter θ is random with prior distribution $\pi(\theta)$
- Joint distribution $\pi(\theta)f(x \mid \theta)$
- Posterior distribution $\theta \mid X$
  condition joint distribution on observed data X.
- Example:
  $\theta \sim$Beta(α, β), $\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0,1)$
  $X_1, \ldots, X_n \sim$Bernoulli(p).
- Conjugate family of priors/posteriors.
- Example 2: Normal prior, normal data.

Bayes Risk

- Loss function $L(\theta, \delta)$
- Choose $\delta = \delta(X)$ to minimize the posterior expected loss

$$\mathbb{E}_{\theta|X}\left[L(\theta,\delta)\right] = \int_\theta L(\theta,\delta) f(\theta \mid x) d\theta$$

- This will minimize the Bayes risk

$$\mathbb{E}_{\theta,X}[L(\theta,\delta)].$$

- Quadratic loss $L(\theta, \delta) = (\theta - \delta)^2$ $\rightarrow$Posterior mean
- Absolute value loss $L(\theta, \delta) = |\theta - \delta|$$\rightarrow$Posterior median

7.2.4. EM algorithm

Missing data problem: $x = (x_o, x_m)$.

- x_o observed
- $x_m$missing
- Joint distribution $f(x_o, x_m \mid \theta)$
- Want $\arg\max_\theta \log L(\theta \mid x_o)$.
- EM algorithm: start at some initial guess $\theta^{(0)}$, then

$$\theta^{(r+1)} \quad = \quad \arg\max_\theta \left[\mathbb{E}_{[x_m|\theta^{(r)},x_o]}\left[\log L(\theta \mid x_o, x_m)\right]\right]$$

$$= \arg\max_\theta \left[\mathbb{E}_{[x_m|\theta^{(r)},x_o]}\left[\log L(\theta \mid x_o)\right] + \underbrace{\mathbb{E}_{[x_m|\theta^{(r)},x_o]}\left[\log f(x_m \mid \theta, x_o)\right]}_{maximised\ \ by\ \ \theta=\theta^{(r)}}\right]$$

The hard EM algorithm

The Hard EM algorithm is as its name suggest, much easier than the general EM-algorithm.

- Split the data $x = (x_o, x_m)$. You have only observed x_o. Start at some $\theta^{(0)}$.
- Iterate $\theta^{(t)} \to \theta^{(t+1)}$ by:
  - sampling $x_m = x_m(t)$ conditional on $\theta^{(t)}$, and then
  - setting $\theta^{(t+1)}$to be the MLE for $x = (x_0, x_m)$.

7.2.17 Multiple Poisson Rates

- Parameters $\beta; \tau_1, \ldots, \tau_n$
- Observe X_i ~Poisson(τ_i) and Y_i ~Poisson($\beta\tau_i$)
- *i.e.* τ_i =population density at place i, β =disease effect size
- Missing data: suppose X_1 is missing.

7.3.1 Mean squared error

- How to measure the quality of estimator W of θ?

$$MSE(W) = \mathbb{E}_\theta[(W-\theta)^2] = Var_\theta[W] + (Bias_\theta W)^2$$

$$Bias_\theta W = \mathbb{E}_\theta W - \theta$$

- An estimator is unbiased if $Bias_\theta W = 0$.
- MLE for σ^2 for $N(\mu, \sigma^2)$ is biased.
- Small MSE more important than unbiasedness
- A sequence of estimator is consistant if $W_n \xrightarrow{P} \mu$ as $n \to \infty$
- $W_n \xrightarrow{P} \mu$ if MSE(W_n) $\to 0$
  - Markov's inequality: $X \geq 0 \to \mathbb{P}(|X| \geq a) \leq E[X]/a$
  - Chebyshev's inequality. r.v. X with mean μ variance σ^2,

$$\mathbb{P}(|X-\mu| \geq k\sigma) \leq k^{-2}$$

Fisher's information

- Sample distribution $f(x \mid \theta)$
- Fisher's information:

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f(X \mid \theta) \right)^2 \right] \overset{\text{regularity}}{=} -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \log f(X \mid \theta) \right]$$

- Theorem 7.3.9 Cramer-Rao Inequality
  Sample $X_1, \ldots, X_n$ with pdf $f(X \mid \theta)$. Estimator $W(X)$ with
  - finite variance
  -
$$\frac{d}{d\theta} \mathbb{E}_\theta W(X) = \int_X \frac{\partial}{\partial \theta} [W(x) f(x \mid \theta)]\, dx$$

  Then

$$Var(W(X)) \geq \frac{\left(\frac{d}{d\theta} \mathbb{E}_\theta W(X) \right)^2}{I(\theta)}$$

- Special case: $X_1, \ldots, X_n$ iidrv

Proof

- Cauchy-Schwarz inequality: Cov$(X, Y)^2 \leq$Var(X)Var(Y).
  - Assume wlog $\mathbb{E}[X] = \mathbb{E}[Y] = 0$
  - For all $t \in \mathbb{R}$,$\mathbb{E}[(tX + Y)^2] = \mathbb{E}[t^2X + 2tXY + Y^2] \geq 0$
- Cramer-Rao

$$Cov\left(W, \frac{\partial}{\partial\theta}f(X \mid \theta)\right)^2 \leq Var(W)Var\left(\frac{\partial}{\partial\theta}f(X \mid \theta)\right)$$

Ch 8 Hypothesis testing

- A hypothesis is a statement about a population parameter
- Null hypothesis H_0
- Alternative hypothesis H_1
- Form a statistical test
- Can you reject H_0?
- Rejecting H_0 does not mean accepting H_1.
- You do not accept H_0.