Methods of Estimation

MIT 18.655

Dr. Kempthorne

Spring 2016

Outline

Minimum Contrast Estimates

$X \in \mathcal{X}$, $X \sim P \in \mathcal{P} = \{P_\theta, \theta \in \Theta\}$.
Problem: Finding a function $\hat{\theta}(X)$ which is "close" to θ.
Consider

$$\rho : \mathcal{X} \times \Theta \to R.$$

and define $\mathcal{D}(\theta_0, \theta)$ to measure the *discrepancy* between θ and the true value θ_0.

$$\mathcal{D}(\theta_0, \theta) = E_{\theta_0}\rho(X, \theta).$$

As a discrepancy measure, $\mathcal{D}$ makes sense if the value of θ minimizing the function is $\theta = \theta_0$.
If P_{θ_0} were true, and we knew $D(\theta_0, \theta)$, we could obtain θ_0 as the minimizer.
Instead of observing $D(\theta_0, \theta)$, we observe $\rho(X, \theta)$.

- $\rho(\cdot, \cdot)$ is a contrast function
- $\hat{\theta}(X)$ is a minimum-contrast estimate.

The definition extends to

- Euclidean $\Theta \subset R^d$.
- θ_0 an interior point of Θ.
- Smooth mapping: $\theta \to D(\theta_0, \theta)$.
- $\theta = \theta_0$ solves
  $$\nabla_\theta D(\theta_0, \theta) = 0.$$
  where $\nabla_\theta = (\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_d})^T$
- Substitute $\rho(X, \theta)$ for $D(\theta_0, \theta)$ and solve
  $$\nabla_\theta \rho(X, \theta) = 0 \text{ at } \theta = \hat{\theta}.$$

Estimating Equations:

- $\Psi : \mathcal{X} \times R^d \to R^d$, where $\Psi = (\psi_1, \ldots, \psi_d)^T$.
- For every $\theta_0 \in \Theta$, the expectation of Ψ given P_{θ_0} has a unique solution
$$V(\theta_0, \theta) = E_{\theta_0}[\Psi(X, \theta)] = 0$$
at $\theta = \theta_0$.

Example 2.1.1 Least Squares.

- $\mu(z) = g(\beta, z), \beta \in R^d$.
- $x = \{(z_i, Y_i) : 1 \leq i \leq n\}$, where $Y_1, \ldots, Y_n$ are independent.
- Define $\rho(X, \beta) = |Y - \mu|^2 = \sum_{i=1}^n [Y_i - g(\beta, z_i)]^2$.
- Consider $Y_i = \mu(z_i) + \epsilon_i$, where $\mu(z_i) = g(\beta, z_i)$ and the ϵ_i are iid $N(0, \sigma_0^2)$.

  Then, β parametrizes the model and we can write:

  $$\begin{aligned} D(\beta_0, \beta) &= E_{\beta_0}\rho(X, \beta) \\ &= n\sigma_0^2 + \sum_{i=1}^n [g(\beta_0, z_i) - g(\beta, z_i)]^2. \end{aligned}$$

  This is minimized by $\beta = \beta_0$ and uniquely so iff β identifiable.

- The *least-squares estimate* $\hat{\beta}$ minimizes $\rho(X, \beta)$.
  Conditions to guarantee existence of $\hat{\beta}$:
  - Continuity of $g(\cdot, z_i)$.
  - Minimum of $\rho(X, \cdot)$ existing on compact set $\{\beta\}$
    e.g., $\lim_{|\beta| \to \infty} |g(\beta, z_i)| = \infty$.
  - If $g(\beta, z_i)$ is differentiable in β, then $\hat{\beta}$ satisfies the *Normal Equations* obtained by taking partial derivatives of $\rho(X, \beta) = |Y - \mu|^2 = \sum_{i=1}^{n} [Y_i - g(\beta, z_i)]^2$ and solving:

$$\frac{\partial \rho(X, \beta)}{\partial \beta_j} = 0$$

$$\rho(X, \beta) = |Y - \mu|^2 = \sum_{i=1}^{n} [Y_i - g(\beta, z_i)]^2$$

Solve:

$$\begin{aligned}
\frac{\partial \rho(X, \beta)}{\partial \beta_j} &= 0 \\
\sum_{i=1}^{n} 2[Y_i - g(\beta, z_i)] \frac{\partial g(\beta, z_i)}{\partial \beta_j}(-1) &= 0 \\
\sum_{i=1}^{n} \frac{\partial g(\beta, z_i)}{\partial \beta_j} Y_i - \sum_{i=1}^{n} \frac{\partial g(\beta, z_i)}{\partial \beta_j} g(\beta, z_i) &= 0
\end{aligned}$$

- Linear case:

$$g(\beta, z_i) = \sum_{j=1}^{d} z_{ij}\beta_j = \mathbf{z}_i^T \boldsymbol{\beta}$$

$$\begin{aligned}
\frac{\partial \rho(X,\beta)}{\partial \beta_j} &= 0 \\
\sum_{i=1}^{n} \frac{\partial g(\beta, z_i)}{\partial \beta_j} Y_i - \sum_{i=1}^{n} \frac{\partial g(\beta, z_i)}{\partial \beta_j} g(\beta, z_i) &= 0 \\
\sum_{i=1}^{n} z_{ij} Y_i - \sum_{i=1}^{n} z_{i,j}(\mathbf{z}_i^T \boldsymbol{\beta}) &= 0 \\
\sum_{i=1}^{n} z_{ij} Y_i - \sum_{k=1}^{d}\sum_{i=1}^{n} z_{i,j} z_{i,k} \beta_k &= 0, \quad j = 1, \ldots, d \\
\mathbf{Z}_D^T \mathbf{Y} - \mathbf{Z}_D^T \mathbf{Z}_D \boldsymbol{\beta} &= 0
\end{aligned}$$

where $\mathbf{Z}_D$ is the $(n \times d)$ *design matrix* with (i,j) element $z_{i,j}$

Note:

- Least Squares exemplifies *minimum contrast* and *estimating equation* methodology.
- Distribution assumptions are not necessary to motivate the estimate as a mathematical approximation.

Method of Moments

Method of Moments

- $X_1, \ldots, X_n$ iid $X \sim P_\theta, \theta \in R^d$.
- $\mu_1(\theta), \mu_2(\theta), \ldots, \mu_d(\theta)$:
  $$\mu_j(\theta) = \mu_j = E[X^j \mid \theta] \text{ the } j\text{th moment of } X.$$
- Sample moments:
  $$\hat{\mu}_j = \sum_{i=1}^{n} X_i^j, \; j = 1, \ldots, d.$$
- Method of Moments: Solve for θ in the system of equations
  $$\begin{array}{rcl} \mu_1(\theta) & = & \hat{\mu}_1 \\ \mu_2(\theta) & = & \hat{\mu}_2 \\ \vdots & & \vdots \\ \mu_d(\theta) & = & \hat{\mu}_d \end{array}$$

Note: .

- θ must be identifiable
- Existence of μ_j: $\lim_{n\to\infty} \hat{\mu}_j = \mu_j$ with $|\mu_j| < \infty$.
- If $q(\theta) = h(\mu_1, \ldots, \mu_d)$, then the Method-of-Moments Estimate of $q(\theta)$ is
$$\hat{q}(\theta) = h(\hat{\mu}_1, \ldots, \hat{\mu}_d).$$
- The MOM estimate of θ may not be unique! (See Problem 2.1.11)

Plug-In and Extension Principles

Frequency Plug-In

- Multinomial Sample: $X_1, \ldots, X_n$ with K values $v_1, \ldots, v_K$
  $$P(X_i = v_j) = p_j \; j = 1, \ldots, K$$
- Plug in estimates: $\hat{p}_j = N_j/n$ where $N_j = count(\{i : X_i = v_j\})$
- Apply to any function $q(p_1, \ldots, p_K)$:
  $$\hat{q} = q(\hat{p}_1, \ldots, \hat{p}_K)$$
- Equivalent to substituting the true distribution function
  $$P_\theta(t) = P(X \leq t \mid \theta)$$
  underlying an iid sample with the empirical distribution function:
  $$\hat{P}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{x_i \leq t\}$$
  $\hat{P}$ is an estimate of P, and $\nu(\hat{P})$ is an estimate of $\nu(P)$.

- Example: αth *population quantile*

  $$\nu_\alpha(P) = \frac{1}{2}[F^{-1}(\alpha) + F_U^{-1}(\alpha)], \text{ with } 0 < \alpha < 1:$$

  where

  $$\begin{aligned} F^{-1}(\alpha) &= inf\{x : F(x) \geq \alpha\} \\ F_U^{-1}(\alpha) &= sup\{x : F(x) \leq \alpha\} \end{aligned}$$

  The plug-in estimate is

  $$\hat{\nu}_\alpha(P) = \nu_\alpha(\hat{P}) = \frac{1}{2}[\hat{F}^{-1}(\alpha) + \hat{F}_U^{-1}(\alpha)].$$

- Example: Method of Moments Estimates of jth Moment

  $$\nu(P) = \mu_j = E(X^j)$$
  $$\hat{\nu}(P) = \hat{\mu}_j = \nu(\hat{P}) = \frac{1}{n}\sum_{i=1}^n x_i^j$$

Extension Principle

- Objective: estimate $q(\theta)$, a function of θ.
- Assume $q(\theta) = h(p_1(\theta), \ldots, p_K(\theta))$, where $h()$ is continuous.
- The extension principle estimates $q(\theta)$ with

  $$\hat{q}(\theta) = h(\hat{p}_1, \ldots, \hat{p}_K)$$

- $h()$ may not be unique: what $h()$ is optimal?

Notes on Method-of-Moments/Frequency Plug-In Estimates

- Easy to compute
- Valuable as initial estimates in iterative algorithms.
- Consistent estimates (close to true parameter in large samples).
- Best Frequency Plug-In Estimates are Maximum-Likelihood Estimates.
- In some cases, MOM estimators are foolish (See Example 2.1.7).

Outline

Least Squares

General Model: Only Y Random

- $X = \{(z_i, Y_i) : 1 \leq i \leq n\}$, where
  $Y_1, \ldots, Y_n$ are independent.
  $z_1, \ldots, z_n \in R^d$ are fixed, non-random.
- For cases $i = 1, \ldots, n$
  $Y_i = \mu(z_i) + \epsilon_i$, where
  $\mu(z) = g(\beta, z), \beta \in R^d$.
  ϵ_i are independent with $E[\epsilon_i] = 0$.
- The Least-Squares Contrast function is
  $\rho(X, \beta) = |Y - \mu|^2 = \sum_{i=1}^{n}[Y_i - g(\beta, z_i)]^2$.
- β parametrizes the model and we can write the discrepancy function

$$D(\beta_0, \beta) = E_{\beta_0}\rho(X, \beta)$$

Least Squares: Only Y Random

Contrast Function:

$$\rho(X,\beta) = |Y-\mu|^2 = \sum_{i=1}^n [Y_i - g(\beta, z_i)]^2.$$

Discrepancy Function:

$$\begin{aligned} D(\beta_0,\beta) &= E_{\beta_0}\rho(X,\beta) \\ &= \sum_{i=1}^n Var(\epsilon_i) + \sum_{i=1}^n [g(\beta_0, z_i) - g(\beta, z_i)]^2]. \end{aligned}$$

- The model is semiparametric with unknown parameter β and unknown (joint) distribution P_ϵ of $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$.

Gauss-Markov Assumptions

- Assume that the distribution of ϵ satisfy:

$$\begin{aligned} E(\epsilon_i) &= 0 \\ Var(\epsilon_i) &= \sigma^2 \\ Cov(\epsilon_i, \epsilon_j) &= 0 \quad \text{for } i \neq j \end{aligned}$$

General Model: (Y,Z) Both Random

- $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ are i.i.d. as $X = (Y, Z) \sim P$
- Define $\mu(z) = E[Y \mid Z = z] = g(\beta, z)$, where
  $g(\cdot, \cdot)$ is a known function and
  $\beta \in R^d$ is unknown parameter
- Given $Z_i = z_i$, define $\epsilon_i = Y_i - \mu(z_i)$ for $i = 1, \ldots, n$
- Conditioning on the z_i we can write:
  $$Y_i = g(\beta, z_i) + \epsilon_i,\ i = 1, 2, \ldots, n$$
  where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ has (joint) distribution $P_{\boldsymbol{\epsilon}}$
- The Least-Squares Estimate of $\hat{\beta}$ is the plug-in estimate $\beta(\hat{P})$, where $\hat{P}$ is the empirical distribution for the sample $\{(Z_i, Y_i), i = 1, \ldots, n\}$
- The function $g(\beta, z)$ can be linear in β and z or nonlinear.
- Closed-form solutions exist for $\hat{\beta}$ when g is linear in β.

Outline

Gauss-Markov Theorem: Assumptions

Data $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ and $\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{p,n} \end{bmatrix}$

follow a linear model satisfying the **Gauss-Markov Assumptions** if $\mathbf{y}$ is an observation of random vector $\mathbf{Y} = (Y_1, Y_2, \ldots Y_N)^T$ and

- $E(\mathbf{Y} \mid \mathbf{X}, \beta) = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots \beta_p)^T$ is the p-vector of regression parameters.
- $Cov(\mathbf{Y} \mid \mathbf{X}, \beta) = \sigma^2 \mathbf{I}_n$, for some $\sigma^2 > 0$.
  I.e., the random variables generating the observations are uncorrelated and have constant variance σ^2 (conditional on $\mathbf{X}$, and β).

Gauss-Markov Theorem

For known constants $c_1, c_2, \ldots, c_p, c_{p+1}$, consider the problem of estimating

$$\theta = c_1\beta_1 + c_2\beta_2 + \cdots c_p\beta_p + c_{p+1}.$$

Under the Gauss-Markov assumptions, the estimator

$$\hat{\theta} = c_1\hat{\beta}_1 + c_2\hat{\beta}_2 + \cdots c_p\hat{\beta}_p + c_{p+1},$$

where $\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_p$ are the least squares estimates is

1) An **Unbiased Estimator** of θ
2) A **Linear Estimator** of θ, that is

   $\tilde{\theta} = \sum_{i=1}^{n} b_i y_i$, for some known (given $\mathbf{X}$) constants b_i.

Theorem: Under the Gauss-Markov Assumptions, the estimator $\hat{\theta}$ has the smallest (*Best*) variance among all *Linear Unbiased Estimators* of θ, i.e., $\hat{\theta}$ is *BLUE*.

Gauss-Markov Theorem: Proof

Proof: Without loss of generality, assume $c_{p+1} = 0$ and define $\mathbf{c} = (c_1, c_2, \ldots, c_p)^T$.

The Least Squares Estimate of $\theta = \mathbf{c}^T \beta$ is:

$$\hat{\theta} = \mathbf{c}^T \hat{\beta} = \mathbf{c}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \equiv \mathbf{d}^T \mathbf{y}$$

a linear estimate in $\mathbf{y}$ given by coefficients $\mathbf{d} = (d_1, d_2, \ldots, d_n)^T$.

Consider an alternative linear estimate of θ:

$$\tilde{\theta} = \mathbf{b}^T y$$

with fixed coefficients given by $\mathbf{b} = (b_1, \ldots, b_n)^T$.

Define $\mathbf{f} = \mathbf{b} - \mathbf{d}$ and note that

$$\tilde{\theta} = \mathbf{b}^T \mathbf{y} = (\mathbf{d} + \mathbf{f})^T \mathbf{y} = \hat{\theta} + \mathbf{f}^T \mathbf{y}$$

- If $\tilde{\theta}$ is unbiased then because $\hat{\theta}$ is unbiased

$$0 = E(\mathbf{f}^T \mathbf{y}) = \mathbf{f}^T E(\mathbf{y}) = \mathbf{f}^T (\mathbf{X}\beta) \text{ for all } \beta \in R^p$$

$\Longrightarrow$ $\mathbf{f}$ is orthogonal to column space of $\mathbf{X}$

$\Longrightarrow$ $\mathbf{f}$ is orthogonal to $\mathbf{d} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{c}$

If $\tilde{\theta}$ is unbiased then

- The orthogonality of $\mathbf{f}$ to $\mathbf{d}$ implies

$$\begin{aligned}
Var(\tilde{\theta}) &= Var(\mathbf{b}^T\mathbf{y}) = Var(\mathbf{d}^T\mathbf{y} + \mathbf{f}^T\mathbf{y}) \\
&= Var(\mathbf{d}^T\mathbf{y}) + Var(\mathbf{f}^T\mathbf{y}) + 2Cov(\mathbf{d}^T\mathbf{y}, \mathbf{f}^T\mathbf{y}) \\
&= Var(\hat{\theta}) + Var(\mathbf{f}^T\mathbf{y}) + 2\mathbf{d}^T Cov(\mathbf{y})\mathbf{f} \\
&= Var(\hat{\theta}) + Var(\mathbf{f}^T\mathbf{y}) + 2\mathbf{d}^T(\sigma^2\mathbf{I}_n)\mathbf{f} \\
&= Var(\hat{\theta}) + Var(\mathbf{f}^T\mathbf{y}) + 2\sigma^2\mathbf{d}^T\mathbf{f} \\
&= Var(\hat{\theta}) + Var(\mathbf{f}^T\mathbf{y}) + 2\sigma^2 \times 0 \\
&\geq Var(\hat{\theta})
\end{aligned}$$

Outline

1. Methods of Estimation I
   - Minimum Contrast Estimates
   - Least Squares and Weighted Least Squares
   - Gauss-Markov Theorem
   - Generalized Least Squares (GLS)
   - Maximum Likelihood

Generalized Least Squares (GLS) Estimates

Consider generalizing the Gauss-Markov assumptions for the linear regression model to

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where the random n-vector ϵ: $E[\epsilon] = \mathbf{0}_n$ and $E[\epsilon\epsilon^T] = \sigma^2\Sigma$.

- σ^2 is an unknown scale parameter
- Σ is a known $(n \times n)$ positive definite matrix specifying the relative variances and correlations of the component observations.

Transform the data $(\mathbf{Y}, \mathbf{X})$ to $\mathbf{Y}^* = \Sigma^{-\frac{1}{2}}\mathbf{Y}$ and $\mathbf{X}^* = \Sigma^{-\frac{1}{2}}\mathbf{X}$ and the model becomes

$$\mathbf{Y}^* = \mathbf{X}^*\beta + \epsilon^*, \text{ where } E[\epsilon^*] = \mathbf{0}_n \text{ and } E[\epsilon^*(\epsilon^*)^T] = \sigma^2\mathbf{I}_n$$

By the Gauss-Markov Theorem, the BLUE ('GLS') of β is

$$\hat{\beta} = [(\mathbf{X}^*)^T(\mathbf{X}^*)]^{-1}(\mathbf{X}^*)^T(\mathbf{Y}^*) = [\mathbf{X}^T\Sigma^{-1}\mathbf{X}]^{-1}(\mathbf{X}^T\Sigma^{-1}\mathbf{Y})$$

Outline

Maximum Likelihood Estimation

- $X \sim P_\theta, \theta \in \Theta$ with density or pmf function
  $$p(x \mid \theta).$$
- Given an observation $X = x$, define the likelihood function
  $$L_x(\theta) = p(x \mid \theta) :$$
  a mapping: $\Theta \to R$.
- $\hat{\theta}_{ML} = \hat{\theta}_{ML}(x)$: the Maximum-Likelihood Estimate of θ is the value making $L_x(\cdot)$ a maximum
  $\hat{\theta}$ is the MLE if
  $$L_x(\hat{\theta}) = \max_{\theta \in \Theta} L_x(\theta).$$
- The MLE $\hat{\theta}_{ML}(x)$ identifies the distribution making x "most likely"
- The MLE coincides with the mode of the Posterior Distribution if the Prior Distribution on Θ is uniform:
  $$\pi(\theta \mid x) \propto p(x \mid \theta)\pi(\theta) \propto p(x \mid \theta).$$

Maximum Likelihood

Examples

- Example 2.2.4: Normal Distribution with Known Variance
- Example 2.2.5: Size of a Population
  $X_1, \dots, X_n$ are iid $U\{1, 2, \dots, \theta\}$, with $\theta \in \{1, 2, \dots\}$.
  For $x = (x_1, \dots, x_n)$,

$$
\begin{aligned}
L_x(\theta) &= \textstyle\prod_{i=1}^n \theta^{-1}\mathbf{1}(1 \le x_i \le \theta) \\
&= \theta^{-n} \times \mathbf{1}(max(x_1, \dots, x_n)) \le \theta) \\
&= \begin{cases} 0 & \text{, if } \theta = 0, 1, \dots, max(x_i) - 1 \\ \theta^{-n} & \text{if } \theta \ge max(x_i) \end{cases}
\end{aligned}
$$

Maximum Likelihood As a Minimum Contrast Method

- Define $l_x(\theta) = \log L_x(\theta) = \log p(x \mid \theta)$
- Because $-log(\cdot)$ is monotone decreasing,
  $\hat{\theta}_{ML}(x)$ minimizes $-l_x(\theta)$
- For an iid sample $X = (X_1, \ldots, X_n)$ with densities $p(x_i \mid \theta)$,
  $$\begin{aligned} l_X(\theta) &= \log p(x_1, \ldots, x_n \mid theta) \\ &= \log [\prod_{i=1}^n p(x_i \mid \theta)] \\ &= \sum_{i=1}^n \log p(x_i \mid \theta) \end{aligned}$$
- As a minimum contrast function ,
  $\rho(X, \theta) = -l_X(\theta)$
  yields the MLE $\hat{\theta}_{ML}(x)$
- The discrepancy function corresonding to the contrast function $\rho(X, \theta)$ is
  $D(\theta_0, \theta) = E[\rho(X, \theta) \mid \theta_0] = -E[\log p(x \mid \theta) \mid \theta_0]$

- Suppose that $\theta = \theta_0$ uniquely minimizes $D(\theta_0, \cdot)$. Then

$$
\begin{aligned}
D(\theta_0, \theta) - D(\theta_0, \theta_0) &= -E[\log p(x \mid \theta) \mid \theta_0] - (-E[\log p(x \mid \theta_0) \mid \theta_0]) \\
&= -E[\log \frac{p(x|\theta)}{p(x|\theta_0)} \mid \theta_0] \\
&> 0, \text{ unless } \theta = \theta_0.
\end{aligned}
$$

This difference is the *Kullback-Leibler Information Divergence* between distribution P_{θ_0} and P_θ:

$$K(P_{\theta_0}, P_\theta) = -E[log(\frac{p(x|\theta)}{p(x|\theta_0)}) \mid \theta_0]$$

Lemma 2.2.1 (Shannon, 1948) The mutual entropy $K(P_{\theta_0}, P_\theta)$ is always well defined and

- $K(P_{\theta_0}, P_\theta) \geq 0$
- Equality holds if and only if $\{x : p(x \mid \theta) = p(x \mid \theta_0)\}$ has probability 1 under both P_{θ_0} and P_θ.

Proof Apply Jensen's Inequality (B.9.3)

Likelihood Equations

Suppose:

- $X \sim P_\theta$, with $\theta \in \Theta$, an open parameter space
- the likelihood function $l_X(\theta)$ is differentiable in θ
- $\hat{\theta}_{ML}(x)$ exists

Then: $\hat{\theta}_{ML}(x)$ must satisfy the **Likelihood Equation(s)**

$$\nabla_\theta l_X(\theta) = 0.$$

Important Cases

For independent X_i with densities/pmfs $p_i(x_i \mid \theta)$,

$$\nabla_\theta l_X(\theta) = \sum_{i=1}^n \nabla_\theta \log p_i(x_i \mid \theta) = 0$$

NOTE: $p_i(\cdot \mid \theta)$ may vary with i.

Examples

- Hardy-Weinberg Proportions (Example 2.2.6)
- Queues: Poisson Process Models (Exponential Arrival Times and Poisson Counts) (Example 2.2.7)
- Multinomial Trials (Example 2.2.8)
- Normal Regression Models (Example 2.2.9).

18.655 Mathematical Statistics
Spring 2016