Statistical Models

MIT 18.655

Dr. Kempthorne

Spring 2016

Outline

Statistical Models: Definitions

Def: Statistical Model

- Random experiment with sample space Ω.
- Random vector $X = (X_1, X_2, \ldots, X_n)$ defined on Ω.
  $\omega \in \Omega$: outcome of experiment
  $X(\omega)$: data observations
- Probability distribution of X
  $\mathcal{X}$: Sample Space = {outcomes x}
  $\mathcal{F}_X$: sigma-field of measurable events
  $P(\cdot)$ defined on $(\mathcal{X}, \mathcal{F}_X)$
- Statistical Model
  $\mathcal{P} = \{\text{family of distributions}\}$

Statistical Models: Definitions

Def: Parameters / Parametrization

- Parameter θ identifies/specifies distribution in $\mathcal{P}$.
- $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$
- $\Theta = \{\theta\}$, the Parameter Space

Outline

Statistical Models: Examples

Example 1.1.1 Sampling Inspection

- Shipment of manufactured items inspected for defects
- $N =$ Total number of items
- $N\theta =$ Number of defective items
- Sample $n < N$ items without replacement and inspect for defects
- $X =$ Number of defective items in the sample

Statistical Models: Sampling Inspection Example

Probability Model for X

- $\mathcal{X} = \{x\} = \{0, 1, \ldots, n\}$.
- Parameter θ: proportion of defective items in shipment
$$\Theta = \{\theta\} = \{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N}{N}\}.$$
- Probability distribution of X
$$P(X = k) = \frac{\binom{N\theta}{k}\binom{N - N\theta}{n - k}}{\binom{N}{n}}$$

Statistical Models: Sampling Inspection Example

Probability Model for X (continued)

- Range of X depends on θ, n, and N

$$k \leq n \text{ and } k \leq N\theta$$
$$(n-k) \leq n \text{ and } (n-k) \leq N(1-\theta)$$

$$\Longrightarrow max(0, n - N(1-\theta)) \leq k \leq min(n, N\theta).$$

- $X \sim Hypergeometric(N\theta, N, n)$.

Statistical Models: Examples

Example 1.1.2 One-Sample Model

- $X_1, X_2, \ldots, X_n$ i.i.d. with distribution function $F(\cdot)$.
  E.g., Sample n members of a large population at random and measure attribute X
  E.g., n independent measurements of a physical constant μ in a scientific experiment.
- Probability Model: $\mathcal{P} = \{\text{distribution functions } F(\cdot)\}$
- Measurement Error Model:

  $X_i = \mu + \epsilon_i, \; i = 1, 2, \ldots, n$

  μ is constant parameter (e.g., real-valued, positive)

  $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ i.i.d. with distribution function $G(\cdot)$

  (G does not depend on μ.)

Statistical Models: Examples

Example 1.1.2 One-Sample Model (continued)

- Measurement Error Model:

$$X_i = \mu + \epsilon_i,\ i = 1, 2, \ldots, n$$

μ is constant parameter (e.g., real-valued, positive)

$\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ i.i.d. with distribution function $G(\cdot)$

(G does not depend on μ.)

$\Longrightarrow X_1, \ldots, X_n$ i.i.d. with distribution function

$$F(x) = G(x - \mu).$$

$$\mathcal{P} = \{(\mu, G) : \mu \in R, G \in \mathcal{G}\}$$

where $\mathcal{G}$ is ...

Example: One-Sample Model

Special Cases:

- Parametric Model: Gaussian measurement errors $\{\epsilon_j\}$ are i.i.d. $N(0, \sigma^2)$, with $\sigma^2 > 0$, unknown.
- Semi-Parametric Model: Symmetric measurement-error distributions with mean μ
  $\{\epsilon_j\}$ are i.i.d. with distribution function $G(\cdot)$, where $G \in \mathcal{G}$, the class of symmetric distributions with mean 0.
- Non-Parametric Model: $X_1, \ldots, X_n$ are i.i.d. with distribution function $G(\cdot)$ where
  $G \in \mathcal{G}$, the class of all distributions
  on the sample space $\mathcal{X}$ (with center μ)

Statistical Models: Examples

Example 1.1.3 Two-Sample Model

- $X_1, X_2, \ldots, X_n$ i.i.d. with distribution function $F(\cdot)$
- $Y_1, Y_2, \ldots, Y_m$ i.i.d. with distribution function $G(\cdot)$
  E.g., Sample n members of population A at random and m members of population B and measure some attribute of population members.
- Probability Model: $\mathcal{P} = \{(F, G),\ F \in \mathcal{F},$ and $G \in \mathcal{G}\}$
  Specific cases relate $\mathcal{F}$ and $\mathcal{G}$
- Shift Model with parameter δ
  - $\{X_i\}$ i.i.d. $X \sim F(\cdot)$, response under Treatment A.
  - $\{Y_j\}$ i.i.d. $Y \sim G(\cdot)$, response under Treatment B.
  - $Y \doteq X + \delta$, i.e., $G(v) = F(v - \delta)$
  - δ is the difference in response with Treatment B instead of Treatment A.

Outline

Statistical Modeling Issues

Issues

- Non-uniqueness of parametrization.
- Varying complexity of equivalent parametrizations
- Possible Non-Identifiability of parameters

  Does $\theta_1 \neq \theta_2$ but $P_{\theta_1} = P_{\theta_2}$?
- Parameters "of interest" vs "Nuisance" parameters
- A vector parametrization that is unidentifiable may have identifiable components.
- Data-based model selection

  How does using the data to select among models affect statistical inference?
- Data-based sampling procedures

  How does the protocol for collecting data observations affect statistical inference?

Regular Models

Notation:

- θ: a parameter specifying a probability distribution P_θ.
- $F(\cdot \mid \theta)$: Distributon function of P_θ
- $E_\theta[\cdot]$: Expectation under the assumption $X \sim P_\theta$. For a measurable function $g(X)$,
  $$E_\theta[g(X)] = \int_{\mathcal{X}} g(x)dF(x \mid \theta).$$
- $p(x \mid \theta) = p(x;\theta)$: density or probability-mass function of X

Assumptions:

- **Either** All of the P_θ are continuous with densities $p(x \mid \theta)$, **Or** All of the P_θ are discrete with pmf's $p(x \mid \theta)$
- The set $\{x : p(x \mid \theta) > 0\}$ is the same for all $\theta \in \Theta$.

Outline

Regression Models

n cases $i = 1, 2, \ldots, n$

- 1 Response (dependent) variable
  $$y_i,\ i = 1, 2, \ldots, n$$
- p Explanatory (independent) variables
  $$\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})^T,\ i = 1, 2, \ldots, n$$

Goal of Regression Analysis:

- Extract/exploit relationship between y_i and $\mathbf{x}_i$.

Examples

- Prediction
- Causal Inference
- Approximation
- Functional Relationships

General Linear Model: For each case i, the conditional distribution $[y_i \mid x_i]$ is given by

$$y_i = \hat{y}_i + \epsilon_i$$

where

- $\hat{y}_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{i,p} x_{i,p}$
- $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ are p regression parameters (constant over all cases)
- ϵ_i Residual (error) variable (varies over all cases)

Extensive breadth of possible models

- Polynomial approximation ($x_{i,j} = (x_i)^j$, explanatory variables are different powers of the same variable $x = x_i$)
- Fourier Series: ($x_{i,j} = sin(jx_i)$ or $cos(jx_i)$, explanatory variables are different sin/cos terms of a Fourier series expansion)
- Time series regressions: time indexed by i, and explanatory variables include lagged response values.

Note: *Linearity* of $\hat{y}_i$ (in regression parameters) maintained with non-linear x.

Steps for Fitting a Model

(1) Propose a model in terms of
- Response variable Y (specify the scale)
- Explanatory variables $X_1, X_2, \ldots X_p$ (include different functions of explanatory variables if appropriate)
- Assumptions about the distribution of ϵ over the cases

(2) Specify/define a criterion for judging different estimators.

(3) Characterize the best estimator and apply it to the given data.

(4) Check the assumptions in (1).

(5) If necessary modify model and/or assumptions and go to (1).

Specifying Assumptions in (1) for Residual Distribution

- Gauss-Markov: zero mean, constant variance, uncorrelated
- Normal-linear models: ϵ_i are i.i.d. $N(0, \sigma^2)$ r.v.s
- Generalized Gauss-Markov: zero mean, and general covariance matrix (possibly correlated,possibly heteroscedastic)
- Non-normal/non-Gaussian distributions (e.g., Laplace, Pareto, Contaminated normal: some fraction $(1 - \delta)$ of the ϵ_i are i.i.d. $N(0, \sigma^2)$ r.v.s the remaining fraction (δ) follows some contamination distribution).

Normal Linear Regression Model

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

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{p,n} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ and ϵ_j are i.i.d. $N(0, \sigma^2)$
with density $f(\epsilon) = (2\pi\sigma^2)^{-\frac{1}{2}} exp(-\frac{1}{2\sigma^2} \cdot \epsilon^2)$

Multivariate Normal Probability Model

$\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$ where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\sigma^2 > 0$.
$p(Y_1, Y_2, \ldots, Y_n \mid \theta) = \prod_{i=1}^{n} f(Y_j - \mathbf{x}_j^T \beta)$,
with parameter $\theta = (\boldsymbol{\beta}, \sigma^2) \in \Theta = R^p \times R_+$

Outline

1. Statistical Models
   - Definitions
   - Examples
   - Modeling Issues
   - Regression Models
   - Time Series Models

Statistical Models: Dependent Responses

Example 1.1.5 Measurement Model with Autoregressive Errors

- $X_1, X_2, \dots, X_n$ are n successive measurements of a physical constant μ
- $X_i = \mu + e_i,\ i = 1, 2, \dots, n$
- $e_i = \beta e_{i-1} + \epsilon_i,\ i = 2, 3, \dots, n$, and $e_0 = 0$ where ϵ_i are i.i.d. with density $f(\cdot)$.

Note:

- The e_i are not i.i.d. (they are dependent).
- The X_i are dependent

$$X_i = \mu(1-\beta) + \beta X_{i-1} + \epsilon_i,\ i = 2, \dots, n$$
$$X_1 = \mu + \epsilon_1$$

Apply conditional probability theory to compute

$$\begin{aligned} p(e_1,\ldots,e_n) &= p(e_1)p(e_2 \mid e_1)p(e_3 \mid e_1,e_2)\cdots p(e_n \mid e_1,\ldots,e_{n-1}) \\ &= p(e_1)p(e_2 \mid e_1)p(e_3 \mid e_2)\cdots p(e_n \mid e_{n-1}) \\ &= f(e_1)f(e_2-\beta e_1)f(e_3-\beta e_2)\cdots f(e_n-\beta e_{n-1}) \end{aligned}$$

Transform $(e_1,\ldots,e_n)$ to $(X_1,\ldots,X_n)$ where $e_i = X_i - \mu$

$$\begin{aligned} p(x_1,\ldots,x_n) &= f(e_1)f(e_2-\beta e_1)f(e_3-\beta e_2)\cdots f(e_n-\beta e_{n-1}) \\ &= f(x_1-\mu)f(x_2-\mu-\beta(x_1-\mu))\cdots f(x_n-\mu-\beta(x_{n-1}-\mu)) \\ &= f(x_1-\mu)\prod_{j=2}^n f(x_j-\beta x_{j-1}-(1-\beta)\mu) \end{aligned}$$

Gaussian AR(1) Model: f is $N(0,\sigma^2)$ density

$$p(x_1,\ldots,x_n) =$$
$$(2\pi\sigma^2)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}\left[(x_1-\mu)^2+\sum_{j=2}^n(x_j-\beta x_{j-1}-(1-\beta)\mu)^2\right]\right\}$$

Problems

Problem 1.1.3 Identifiable parametrizations.

Problem 1.1.4 Stochastically larger distributions in two-sample Models.

Problem 1.1.7 Symmetric distributions and their properties.

Problem 1.1.9 Collinearity: What conditions on $\mathbf{X}$ are required for the regression parameter β to be identifiable?

Problem 1.1.11 Scale Models and Shift Models.

Problem 1.1.12 Hazard rates and Cox proportional hazard model.

Problem 1.1.14 The Pareto distribution.