Statistics

Week 3: Data Reduction and Method of Moments

Etienne Wijler

Econometrics and Data Science
Econometrics and Operations Research
Bachelor Program

VRIJE UNIVERSITEIT AMSTERDAM
SCHOOL OF BUSINESS AND ECONOMICS

Course overview: Data Reduction

P4: Estimation

Week 1 Probability Recap

Week 2 Statistical Models

Week 3 Data Reduction and MME

Week 4 MLE and Evaluation

Week 5 Estimator Optimality

Week 6 Consistency

P5: Inference

Week 7 Hypothesis testing

Week 8 Mean and Variance testing

Week 9 Finding test statistics

Week 10 Evaluating tests

Week 11 Interval estimation

Week 12 Asymptotic tests

Dimensionality Reduction

Problem: Datasets that are collected are typically large files containing many measurements that do not seem very informative when observed as a whole.

Idea: Perform data reduction or aggregation to extract meaningful information from the data by calculating statistics.

Definition

A statistic T is any function of the data $\boldsymbol{X}$.

The objective of dimensionality reduction

Goal: When reducing the dimension of the data, we typically aim to:

- discard information that is irrelevant to the parameter of interest,
- retain information that is relevant to this parameter.

Example (Consumer preference)

You are examining consumer preference for different sugar contents in a new soft-drink. You offer participants in your study a high and low sugar version of the drink and record which version they prefer. What would be:

1. research question, random variable, and statistical model,
2. the parameter of interest,
3. a logical statistic.

Sufficient statistics

Sufficiency: A sufficient statistic "contains all relevant information about the parameter of interest".

Definition (6.2.1, sufficient statistics)

Let $\boldsymbol{X}$ be generated by a distribution from the statistical model $\{f(\boldsymbol{x} \mid \theta) \mid \theta \in \Theta\}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ, if the conditional distribution of the sample $\boldsymbol{X}$ given the value of $T(\boldsymbol{X})$ does not depend on θ.

Equivalently: $T(\boldsymbol{X})$ is a sufficient statistic if

$$f_{\boldsymbol{X}}(\boldsymbol{x} \mid T(\boldsymbol{X}), \theta_1) = f_{\boldsymbol{X}}(\boldsymbol{x} \mid T(\boldsymbol{X}), \theta_2), \quad \forall\, \theta_1, \theta_2 \in \Theta.$$

Otherwise: $f_{\boldsymbol{X}}\left(\boldsymbol{x} \mid T(\boldsymbol{X})\right)$ would depend on θ and one could learn additional information about the parameter from observing $\boldsymbol{x}$ instead of $T(\boldsymbol{x})$!

Sufficiency: intuition

Suppose that $f(x|\theta)$ is a pmf and let $T(\boldsymbol{X})$ be a sufficient statistic for θ.

Then, we can rewrite

$$\begin{aligned} f(\boldsymbol{x} \mid \theta) &= \mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x}; T(\boldsymbol{X}) = T(\boldsymbol{x})) \\ &= \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x}))\mathbb{P}_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x})), \end{aligned}$$

Conclusion: all information concerning θ is contained in $\mathbb{P}_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x}))$!

Implication: if $\boldsymbol{x}$ and $\boldsymbol{y}$ are two different samples such that $T(\boldsymbol{x}) = T(\boldsymbol{y})$, then the inference about θ is the same whether $\boldsymbol{X} = \boldsymbol{x}$ or $\boldsymbol{X} = \boldsymbol{y}$ is observed.

Reduction: We can compute $T(\boldsymbol{x})$ and discard $\boldsymbol{x}$ without losing information on θ.

Consumer preference (continued)

Example (Consumer preference continued)

We continue investigating consumer preference for a new soft-drink. Let $X_i \overset{iid}{\sim}$ Bernoulli(p) represent the binary variable indication whether a consumer chooses the low sugar version ($X_i = 1$). The statistic we calculate is $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$. Is this indeed a sufficient statistic for p?

Hint: recall Bayes theorem $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

In practice, we rarely verify sufficiency by direct derivation of $f_{\boldsymbol{X}}(\boldsymbol{x} \mid T(\boldsymbol{x}))$. However, it helps to build intuition!

Factorization Theorem

Problem: The definition of sufficiency is inconvenient, as it requires us to

- use intuition to find the statistic $T(\boldsymbol{X})$, and then
- calculate difficult conditional probabilities to prove sufficiency.

Solution: the Factorization Theorem provides a much easier approach.

Theorem (6.2.6, Factorization Theorem)

A statistic $T(\boldsymbol{X})$ is sufficient for θ if and only if there exist functions $g(x \mid \theta)$ and $h(\boldsymbol{x})$ such that, for all $\boldsymbol{x} \in \mathbb{R}^n$,

$$f(\boldsymbol{x} \mid \theta) = g(T(\boldsymbol{x}) \mid \theta) h(\boldsymbol{x}).$$

Factorization Theorem: proof

Proof of the Factorization Theorem (only if direction).

We prove the theorem only for discrete pdfs. We start with the only if direction. By sufficiency we have

$$\begin{aligned} f(\boldsymbol{x} \mid \theta) &= \mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x}, T(\boldsymbol{X}) = T(\boldsymbol{x})) \\ &= \underbrace{\mathbb{P}_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x}))}_{g(T(\boldsymbol{x})|\theta)} \underbrace{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x}))}_{h(\boldsymbol{x})} \\ &= g(T(\boldsymbol{x}) \mid \theta) h(\boldsymbol{x}). \end{aligned}$$

Hence, the factorization always exists if $T(\boldsymbol{X})$ is sufficient.

Proof of the Factorization Theorem (if direction).

To show the if direction, define $A_{T(\boldsymbol{x})} = \{\boldsymbol{y} : T(\boldsymbol{y}) = T(\boldsymbol{x})\}$, such that

$$\mathbb{P}_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x})) = \mathbb{P}_\theta(\boldsymbol{X} \in A_{T(\boldsymbol{x})}) = \sum_{\boldsymbol{y} \in A_{T(\boldsymbol{x})}} \mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{y}).$$

Then, assuming that the factorization exists, we have

$$\begin{aligned}
\mathbb{P}_\theta\left(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x})\right) &= \frac{\mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x}))} = \frac{\mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x})}{\sum_{\boldsymbol{y} \in A_{T(\boldsymbol{x})}} \mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{y})} \\
&= \frac{g(T(\boldsymbol{x}) \mid \theta) h(\boldsymbol{x})}{\sum_{\boldsymbol{y} \in A_{T(\boldsymbol{x})}} g(T(\boldsymbol{y}) \mid \theta) h(\boldsymbol{y})} = \frac{g(T(\boldsymbol{x}) \mid \theta) h(\boldsymbol{x})}{g(T(\boldsymbol{x}) \mid \theta) \sum_{\boldsymbol{y} \in A_{T(\boldsymbol{x})}} h(\boldsymbol{y})} \\
&= \frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{y} \in A_{T(\boldsymbol{x})}} h(\boldsymbol{y})},
\end{aligned}$$

which does not depend on the parameter. $\square$

Factorization Theorem: examples

Example (Uniform(0,θ))

Consider a random sample $X_1, \ldots, X_n$ drawn from a population with pdf

$$g(x \mid \theta) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta.$$

Find the sufficient statistic for θ.

Example (Normal(μ,σ^2))

Consider a random sample $X_1, \ldots, X_n$ drawn from a population with pdf

$$g(x \mid \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Find the sufficient statistic for $\boldsymbol{\theta} = \mu$, i.e. assuming σ^2 is known, and for $\boldsymbol{\theta} = (\mu, \sigma^2)$.

Sufficient statistics and the exponential family

Note: Finding sufficient statistics is further simplified when dealing with members of the exponential family.

Theorem (6.2.10)

Let $X_1, \dots, X_n$ be a random sample from a population belonging to the exponential family

$$g(x \mid \theta) = h(x)c(\theta)e^{\sum_{j=1}^{m} w_j(\theta)t_j(x)}.$$

Then $T(\boldsymbol{X}) = (\sum_{i=1}^{n} t_1(X_i), \dots, \sum_{i=1}^{n} t_m(X_i))$ is a sufficient statistic for θ.

Proof: tutorial exercise!

Non-uniqueness of sufficient statistics

Important: Sufficient statistics are only unique up to an invertible transformation.

Lemma

Let $T(\boldsymbol{X})$ be a sufficient statistic for θ and let ϕ be an invertible function. Then $\phi(T(\boldsymbol{X}))$ is also a sufficient statistic for θ.

Proof.

This follows directly from the factorization theorem since

$$f(\boldsymbol{x} \mid \theta) = g(T(\boldsymbol{x}) \mid \theta)h(\boldsymbol{x}) = g\left(\phi^{-1}\left(\phi\left(T(\boldsymbol{x})\right)\right) \mid \theta\right) h(\boldsymbol{x}) = \tilde{g}\left(\tilde{T}(\boldsymbol{x}) \mid \theta\right) h(\boldsymbol{x}),$$

where $\tilde{g}(x) = g(\phi^{-1}(x))$ and $\tilde{T}(\boldsymbol{x}) = \phi\left(T(\boldsymbol{x})\right)$, the new sufficient statistic. $\square$

Course overview: MME

P4: Estimation

- Week 1 Probability Recap
- Week 2 Statistical Models
- Week 3 Data Reduction and MME
- Week 4 MLE and Evaluation
- Week 5 Estimator Optimality
- Week 6 Consistency

P5: Inference

- Week 7 Hypothesis testing
- Week 8 Mean and Variance testing
- Week 9 Finding test statistics
- Week 10 Evaluating tests
- Week 11 Interval estimation
- Week 12 Asymptotic tests

Parameter estimation

From here on, we are going to focus on parameter estimation.

Note: We will always assume

- We have data $\boldsymbol{x} = (x_1, \ldots, x_n)$ that is a realization from the iid random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ with population $g(x \mid \theta_0)$.
- We are given a statistical model: $\{g(x \mid \theta) \mid \theta \in \Theta\}$.
- The model is correctly specified, i.e. $\theta_0 \in \Theta$.

Goal: find the correct value $\theta_0 \in \Theta$. Equivalently: estimate the DGP $g(x \mid \theta_0)$.

Estimates and estimators

Definition (7.1.1)

An estimate for θ_0 in a statistical model is any function $W(\boldsymbol{x})$ of the data. The corresponding estimator is the stochastic variable $W(\boldsymbol{X})$ obtained by plugging in the random vector.

Notation: Statisticians often write $\hat{\theta}$ for $W(\boldsymbol{X})$, when it's clear what estimator we are talking about. We will also adopt this convention.

Note: While by definition any function of the data is an estimator, in practise the term is only used when $W(\boldsymbol{x})$ serves to approximate a quantity of interest (e.g. $h(\theta_0)$).

Finding estimators

Intuition: In some instances, we can find natural estimators by intuition:

Example (Coin wager)

Recall the coin wager example. The statistical model for the coin wager is given by $\{\text{Bernoulli}(p) \mid p \in [0,1]\}$. What would be an intuitive estimator of p_0?

However, often times intuition fails us of finding the estimators we need:

Example (Milk sales)

Recall the milk store example, where we had the statistical model $\{\text{Binomial}(k,p) \mid k \in \mathbb{N}, p \in [0,1]\}$.

- Assume k_0 is known: what would be an intuitive estimator of p_0?
- Assume k_0 is unknown: what would be an intuitive estimator of (k_0, p_0)?

Method of Moments

Note: By the LLN, we have the following natural approximations

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}(X_1) \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1} X_i \approx \mathbb{E}(X_1)$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathbb{E}(X_1^2) \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1} X_i^2 \approx \mathbb{E}(X_1^2)$$

$$\vdots$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^n \xrightarrow{p} \mathbb{E}(X_1^n) \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1} X_i^n \approx \mathbb{E}(X_1^n)$$

Idea: Since $E(X_1^k)$ typically depends on the parameters $\boldsymbol{\theta}_0$, this gives a system of equations that we can solve for $\boldsymbol{\theta}_0$!

MM: Solving this system of equation is called the method of moments (MM).

Method of Moments: examples

Example (Basketball skills)

You are evaluating your basketball skills using statistics (yes, we're *that* nerdy). Let X_i, $i = 1, \ldots, n$, denote the number of throws it took you to score the i-th 3-pointer. You assume that $X \sim \text{Geometric}(p_0)$, with pdf $g(x \mid p) = (1-p)^{x-1}p^x$ for $x = 1, 2, \ldots$. Find an estimator $\hat{p}$ for p_0 using the MM.

Example (Vegan diet health effects)

You're interested in the effects of vegan diets on a person's health. In particular, visceral fat is one of the leading causes of health issues such as diabetes and cancer. Let X_i, $i = 1, \ldots, n$, denote the visceral fat content of a randomly selected vegan. Your statistical model is $\{\text{Normal}(\mu, \sigma^2) \mid \mu, \sigma^2 > 0\}$. Find the MME of $\boldsymbol{\theta}_0 = (\mu_0, \sigma_0^2)$.

Milk Sales: the solution

Example (Milk Sales)

Recall the problem of modelling the daily milk sales for inventory management purposes. Our statistical model is $\{\text{Binomial}(k, p) \mid k \in \mathbb{N}, p \in [0, 1]\}$, with k being the total number of potential customers and p the probability of any individual visiting our store. The DGP is given by

$$g(x \mid \boldsymbol{\theta}_0) = \binom{k_0}{x} p_0^x (1 - p_0)^{k_0 - x}, \quad x = 0, 1, \ldots, k_0.$$

Letting X_i, $i = 1, \ldots, n$, denote the (unobserved) number of customers visiting your store on day i, find the MME of $\boldsymbol{\theta}_0$.

Planning wind turbines

Example (Wind turbines)

You are involved in the planning of a new project for wind turbines. As a statistician, you are tasked with modelling the wind speeds in a certain area. Let X_i, $i = 1, \ldots, n$, denote the average wind speed on day i. The statistical model is given by $\{\text{Weibull}(2, \theta) \mid \theta > 0\}$ with pdf

$$g(x \mid \theta) = \frac{2}{\theta} x e^{-x^2/\theta}, \quad 0 \leq x < \infty, \ \theta > 0.$$

Note (C&B) that $\mathbb{E}_\theta(X_1) = \sqrt{\theta}\Gamma(3/2)$, and recall $\Gamma(1 + a) = a\Gamma(a)$ and $\Gamma(1/2) = \sqrt{\pi}$. Find the MME of θ_0.

Note: The CDF of this distribution is $G(x \mid \theta) = 1 - e^{-x^2/\theta}$. Knowing θ would thus easily allow you to calculate wind speed probabilities.

Method of Moments: pros and cons

Pros: The methods of moments estimator (MME) is often

- intuitive and easy to derive,
- widely applicable,
- possible to apply without specifying the distribution.

Cons: The MME can

- be sub-optimal (does not have the lowest variance of all choices),
- provide estimates outside the parameter space ($\hat{\theta} \notin \Theta$),
- provide many solutions that is difficult to choose from,
- not be applied when moments do not exist.