

14.310x Data Analysis for Social Scientists

This is a cheat sheet for data analysis based on the online course given by Prof. Esther Duflo and Prof. Sara Ellison. Compiled by Janus B. Advincula.

Last Updated December 3, 2019

Module 1

Introduction

- Data is plentiful.
- Data is beautiful.
- Data is insightful.
- Data is powerful.
- Data can be deceitful.

Causation vs. Correlation

- Correlation is not causality.
- A causal *story* is not causality either.
- Even more sophisticated data use may still not be causality.

What We Need to Learn

- How do we model the processes that might have generated our data?
 - Probability
- How do we summarize and describe data, and try to uncover what process may have generated it?
 - Statistics
- How do we uncover pattern between variables?
 - Exploratory data analysis
 - Econometrics

Module 2

Fundamentals of Probability

A **sample space** S is a collection of all possible outcomes of an experiment.

An **event** A is any collection of outcomes (including individual outcomes, the entire sample space, the null set).

Useful results:

- If $A \subset B$, then $A \cup B = B$.
- If $A \subset B$ and $B \subset A$, then $A = B$.
- If $A \subset B$, then $A \cap B = AB = A$.
- $A \cup A^c = S$

A and B are **mutually exclusive (disjoint)** if they have no outcomes in common.

A and B are **exhaustive (complementary)** if their union is S .

Probability

We will assign to every event A a number $\mathbb{P}(A)$, which is the probability the event will occur ($\mathbb{P}: S \rightarrow R$).

We require that:

1. $\mathbb{P}(A) \geq 0$ for all $A \subset S$
2. $\mathbb{P}(S) = 1$
3. For any sequence of disjoint sets A_1, A_2, \dots ,

$$\mathbb{P}\left(\bigcap_i A_i\right) = \sum_i \mathbb{P}(A_i)$$

A **probability** on a sample space S is a collection of numbers $\mathbb{P}(A)$ that satisfy axioms 1-3.

Counting

1. If an experiment has two parts, first one having m possibilities and, regardless of the outcome in the first part, the second one having n possibilities, then the experiment has $m \times n$ possible outcomes.
2. Any ordered arrangement of objects is called a **permutation**. The number of different permutations of N objects is $N!$. The number of different permutations of n objects taken from N objects is $\frac{N!}{(N-n)!}$.
3. Any unordered arrangement of objects is called a **combination**. The number of different combinations of n objects taken from N objects is $\frac{N!}{(N-n)!n!}$. We typically denote this $\binom{N}{n}$.

Properties:

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(\emptyset) = 0$
- If $A \subset B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- For all A , $0 \leq \mathbb{P}(A) \leq 1$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$
- $\mathbb{P}(AB^c) = \mathbb{P}(A) - \mathbb{P}(AB)$

Independence Events A and B are *independent* if $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$.

Theorem If A and B are independent, A and B^c are also independent.

Conditional Probability The *probability of A conditional on B* is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}, \quad \mathbb{P}(B) > 0.$$

If A and B are independent and $\mathbb{P}(B) > 0$, then

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A).$$

Bayes' Theorem

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}$$

(A and A^c form a partition of S .)

Random Variables, Distributions, and Joint Distributions

A **random variable** is a real-valued function whose domain is the sample space.

A **discrete** random variable can take on only a finite or countable infinite number of values.

A random variable that can take on any value in some interval, bounded or unbounded, of the real line is called a **continuous** random variable.

Hypergeometric Distribution $X \sim \mathcal{H}(N, K, n)$

$$f_X(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, \quad x = \max(0, n+K-N), \dots, \min(n, K)$$

The hypergeometric distribution describes the number of "successes" in n trials where you're sampling without replacement from a sample of size N whose initial probability of success was K/N .

Binomial Distribution $X \sim \mathcal{B}(n, p)$

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n$$

The binomial distribution describes the number of "successes in n trials where the trials are independent and the probability of success in each is p .

The **probability function** (PF) of X , where X is a discrete random variable, is the function f_X such that for any real number x , $f_X(x) = \mathbb{P}(X = x)$.

Properties:

- $0 \leq f_X(x_i) \leq 1$
- $\sum_i f_X(x_i) = 1$
- $\mathbb{P}(A) = \mathbb{P}(X \subset A) = \sum_A f_X(x_i)$
- $\mathbb{P}(X = x) = 0$ for any x if X is continuous.

The **density** or **probability density function** (PDF) is the continuous analog to the discrete PF in many ways.

A random variable X is **continuous** if there exists a non-negative function f_X such that for any interval $A \subset R$,

$$\mathbb{P}(X \subset A) = \int_A f_X(x) dx.$$

Properties:

- $0 \leq f_X(x)$
- $\int f_X(x) dx = 1$
- $\mathbb{P}(A) = \mathbb{P}(a \leq X \leq b) = \int_A f_X(x) dx$

The **cumulative distribution function** (CDF) F_X of a random variable X is defined for each x as

$$F_X(x) = \mathbb{P}(X \leq x).$$

Properties:

- $0 \leq F_X(x) \leq 1$
- $F_X(x)$ is non-decreasing in x
- $\lim_{x \rightarrow -\infty} F_X(x) = 0$
- $\lim_{x \rightarrow \infty} F_X(x) = 1$
- $F_X(x)$ is right continuous.

A PF/PDF and a CDF for a particular random variable contain exactly the same information about its distribution, just in a different form.

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(x) dx$$

$$F'_X(x) = \frac{dF(x)}{dx} = f_X(x)$$

Joint Distributions

If X and Y are continuous random variables defined on the same sample space S , then the **joint probability density function** of X & Y , $f_{XY}(x, y)$, is the surface such that for any region A of the xy -plane,

$$\mathbb{P}((X, Y) \subset A) = \int \int_A f_{XY}(x, y) dx dy.$$

Gathering and Collecting Data

Where can we find data?

- 1. Existing data libraries
- 2. Collecting your own data
- 3. Extracting data from the internet

What is web scraping?

- Pull data from one page
- Crawl an entire web page
- A set of forms running in the background
- Any of the above in an ongoing fashion

Web scraping in Python

You will work using the request library and the BeautifulSoup library.

Web scraping in R

R has a web scraping package built by Hadley Wickham called rvest.

Human Subject Research

Research A systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge.

Human Subject A living individual about whom an investigator (whether professional or student) conducting research obtains

- 1. data through intervention or interaction with the individual, or
- 2. identifiable private information.

Key Principles of the Belmont Report

1. **Respect for Persons**

- Respect individual autonomy
- Protect individuals with reduced autonomy

2. **Beneficence**

- Maximize benefits and minimize harm

3. **Justice**

- Equitable distribution of research burdens and benefits

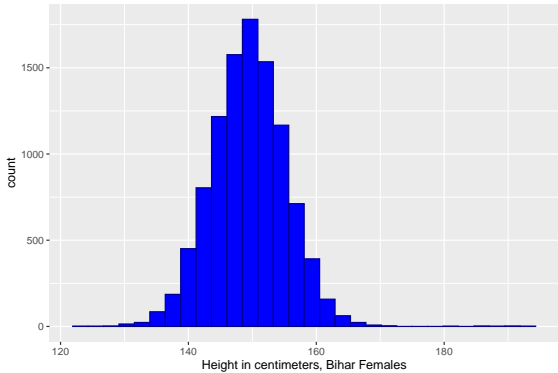
Module 3

Summarizing and Describing Data

Histogram

A histogram is a rough estimate of the probability distribution function of a continuous variable. It is a function that counts the number of observations that fit into each bin.

Example: Women’s height in Bihar



The Kernel Density Estimation

Kernel density estimation is a non-parametric way to estimate the probability density function of a random variable.

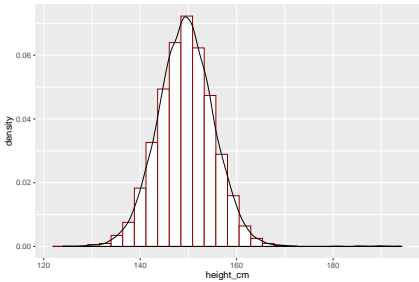
Let (x_1, x_2, \dots, x_n) be an independent and identically distributed sample drawn from some distribution with an unknown PDF f . We are interested in estimating the shape of this function f . Its kernel density estimator is given by

f-hat_h(x) = 1/n * sum_{i=1}^n K_h(x - x_i) = 1/nh * sum_{i=1}^n K((x - x_i)/h)

where $K()$ is the **kernel**, a non-negative function that integrates to 1 and has mean zero, and $h > 0$ is the **bandwidth**.

Things to choose:

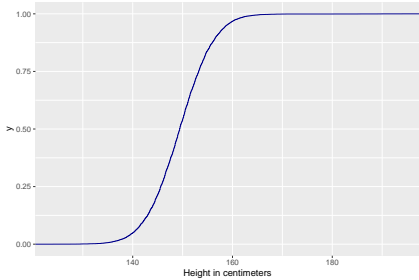
- the Kernel function (Epanechnikov, Normal, etc.)
- the bandwidth (the optimal bandwidth minimizes the Mean Squared Error)



Cumulative Histogram, CDF

Cumulative Histogram the number /frequency of cases that are smaller or equal to the value for a particular bin

You can get a smoothed version of a CDF (using *ecdf* in R)



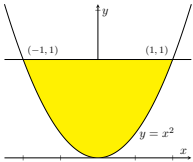
Joint, Marginal and Conditional Distributions

Joint Distribution

Example:

f_{XY}(x, y) = { cx^2y for x^2 <= y <= 1 / 0 otherwise

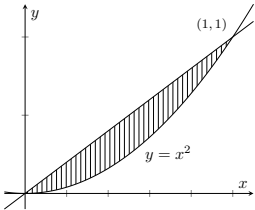
Support:



What is c?

int_{-1}^1 int_{x^2}^1 cx^2y dy dx = 4/21 c = 1 => c = 21/4

What is P(X > Y)?



int_0^1 int_{x^2}^x 21/4 x^2y dy dx = 3/20

Marginal Distribution

For discrete:

f_X(x) = sum_y f_{XY}(x, y)

For continuous:

f_X(x) = int_y f_{XY}(x, y) dy

Independence X & Y are independent if

P(X in A and Y in B) = P(X in A) P(Y in B)

for all regions A and B. Also, X and Y are independent iff

f_{XY}(x, y) = f_X(x) f_Y(y).

Conditional Distribution

The **conditional PDF** of Y given X is

f_{Y|X}(y|x) = f_{XY}(x, y) / f_X(x)

(= P(Y = y | X = x) for X, Y discrete)

Conditional distributions and independence

f_{Y|X}(y|x) = f_Y(y) iff f_{XY}(x, y) = f_X(x) f_Y(y)

iff X & Y independent

Module 4

Functions of Random Variables

X is a random variable with $f_X(x)$ known. We want the distribution of $Y = h(X)$. Then,

F_Y(y) = \int_{\{x:h(x)\leq y\}} f_X(x)dx

If Y is also continuous, then

f_Y(y) = \frac{dF_Y(y)}{dy}.

Example:

f_X(x) = \begin{cases} 1/2 & \text{for } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}

$Y = X^2$. What is $f_Y(y)$? Note that the support of X is $[-1, 1]$ which implies that the induced support of Y is $[0, 1]$.

F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2} dx = \sqrt{y} \text{ for } 0 \leq y \leq 1

f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}} & \text{for } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}

Linear Transformation

Let X have PDF $f_X(x)$. Let $Y = aX + b, a \neq 0$.

f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)

Probability Integral Transformation Let X , continuous, have PDF $f_X(x)$ and CDF $F_X(x)$. Let $Y = F_X(X)$. How is Y distributed?

A continuous random variable transformed by its own CDF will always have a $U[0, 1]$ distribution.

Convolution

A convolution refers to the sum of independent random variables.

Let X be continuous with PDF f_X , Y continuous with PDF f_Y . X and Y are independent. Let Z be their sum. What is the PDF of Z ?

f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy \quad -\infty < z < \infty

Order Statistics

Let X_1, \dots, X_n be continuous, independent, identically distributed, with PDF f_X . Let $Y_n = \max\{X_1, \dots, X_n\}$. This is called the **n^{th} order statistic**.

Distribution:

F_n(y) = F_X(y)^n
f_n(y) = \frac{dF_n(y)}{dy} = n(F_X(y))^{n-1} f_X(y)

Moments of a Distribution

The **mode** is the point where the PDF reaches its highest value.

The **median** is the point above and below which the integral of the PDF is equal to $1/2$.

The **mean**, or **expectation**, or **expected value**, is defined as

E[X] = \int x f_X(x) dx.

Y = g(X)

E[Y] = E[g(X)] = \int g(x) f_X(x) dx

Expectation, Variance and an Introduction to Regression

Properties of Expectation:

- 1. $E[a] = a, a$ constant
- 2. $E[Y] = aE[X] + b, Y = aX + b$
- 3. $E[Y] = E[X_1] + \dots + E[X_n], Y = X_1 + \dots + X_n$
- 4. $E[Y] = a_1E[X_1] + \dots + a_nE[X_n] + b, Y = a_1X_1 + \dots + a_nX_n + b$
- 5. $E[XY] = E[X]E[Y]$ if X, Y independent

Variance

Var(X) = E[(X - \mu)^2]

Properties of Variance:

- 1. $Var(X) \geq 0$
- 2. $Var(a) = 0, a$ constant
- 3. $Var(Y) = a^2Var(X), Y = aX + b$
- 4. $Var(Y) = Var(X_1) + \dots + Var(X_n), Y = X_1 + \dots + X_n, X_1, \dots, X_n$ independent
- 5. $Var(Y) = a_1^2Var(X_1) + \dots + a_n^2Var(X_n), Y = a_1X_1 + \dots + a_nX_n + b, X_1, \dots, X_n$ independent
- 6. $Var(X) = E[X^2] - (E[X])^2$

Standard Deviation

SD(X) = \sigma = \sqrt{Var(X)}

Conditional Expectation

E[Y|X] = \int y f_{Y|X}(y|x) dy

Note that $E[Y|X]$ is a function of X and, therefore, a random variable.

Law of Iterated Expectations

E[E[Y|X]] = E[Y]

Law of Total Variance

Var(E[Y|X]) + E[Var(Y|X)] = Var(Y)

Covariance and Correlation

Covariance

Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]

Correlation

\rho(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{Var(X)}\sqrt{Var(Y)}}

- X & Y are "positively correlated" if $\rho > 0$.
- X & Y are "negatively correlated" if $\rho < 0$.
- X & Y are "uncorrelated" if $\rho = 0$.

Properties of Covariance:

- 1. $Cov(X, X) = Var(X)$
- 2. $Cov(X, Y) = Cov(Y, X)$
- 3. $Cov(X, Y) = E[XY] - E[X]E[Y]$
- 4. X, Y independent $\Rightarrow Cov(X, Y) = 0$
- 5. $Cov(aX + b, cY + d) = ac Cov(X, Y)$
- 6. $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
- 7. $|\rho(X, Y)| \leq 1$
- 8. $|\rho(X, Y)| = 1$ iff $Y = aX + b, a \neq 0$

A Preview of Regression

We have two random variables, X & Y .

E[X] = \mu_X, Var(X) = \sigma_X^2

E[Y] = \mu_Y, Var(Y) = \sigma_Y^2

\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}

If $|\rho_{XY}| < 1$, then we can write $Y = \alpha + \beta X + U$.

Let $\beta = \rho_{XY} \frac{\sigma_Y}{\sigma_X}$.

Let $\alpha = \mu_Y - \beta \mu_X$.

Then, $U = Y - \alpha - \beta X$ has the following properties:

- $E[U] = 0$
- $Cov(X, U) = 0$

α & β are the **regression coefficients**.

Inequalities

Markov Inequality X is a random variable that is always non-negative. Then, for any $t > 0$,

P(X \geq t) \leq \frac{E[X]}{t}

Chebyshev Inequality X is a random variable for which $Var(X)$ exists. Then, for any $t > 0$,

P(|X - E[X]| \geq t) \leq \frac{Var(X)}{t^2}

Module 5

Special Distributions

Bernoulli Two possible outcomes: *success* or *failure*. The probability of success is p , failure is q ($= 1 - p$).

Binomial If X_1, \dots, X_n are i.i.d. random variables, all Bernoulli distributed with success probability p , then

$$X = \sum_{k=1}^n X_k \sim \mathcal{B}(n, p) \quad \text{binomial distribution}$$

The binomial distribution is the number of successes in a sequence of n independent (success/failure) trials, each of which yields success with probability p .

Hypergeometric The binomial distribution is used to model the number of successes in a sample of size n *with replacement*. If you sample *without replacement*, you get the hypergeometric distribution.

Negative Binomial Consider a sequence of independent Bernoulli trials, and let X be the number of trials necessary to achieve r successes. X has a negative binomial distribution.

Geometric A negative binomial distribution with $r = 1$ is a geometric distribution. It is the number of failures before the first success.

- The sum of r independent Geometric (p) random variables is a negative binomial (r, p) random variable.
- If X_i are i.i.d. and negative binomial (r_i, p), then $\sum X_i$ is distributed as a negative binomial ($\sum r_i, p$).
- Memorylessness

Poisson The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time if:

1. the events can be counted in whole numbers
2. the occurrences are independent and
3. the average frequency of occurrence for a time period is known.

Relationship between Poisson and Binomial For small values of p , the Poisson distribution can simulate the Binomial distribution.

Exponential

- waiting time between two events in a Poisson process
- memoryless
- It is a special case of a **gamma distribution**, the “waiting time” before a number of occurrences.

Uniform

- quasi-random number generators
- from a uniform distribution, you can use the inverse CDF method to get a sample for many (not all) distributions you are interested in

Normal Distribution

Properties:

- If X_1 is normal, then $X_2 = a + bX_1$ is also normal, with mean $a + b\mathbb{E}[X_1]$ and variance $b^2\text{Var}(X_1)$.
- Normal distributions are symmetric, unimodal, “bell-shaped,” have thin tails, and the support is \mathbb{R} .

Useful R Commands:

	Purpose	Syntax
rnorm	generates random numbers from normal distribution	<i>rnorm</i> (<i>n</i> , <i>mean</i> , <i>sd</i>)
dnorm	probability density function (PDF)	<i>dnorm</i> (<i>x</i> , <i>mean</i> , <i>sd</i>)
pnorm	cumulative distribution function (CDF)	<i>pnorm</i> (<i>q</i> , <i>mean</i> , <i>sd</i>)
qnorm	quantile function – inverse of pnorm	<i>qnorm</i> (<i>p</i> , <i>mean</i> , <i>sd</i>)

The Sample Mean, Central Limit Theorem and Estimation

The **sample mean** is the arithmetic average of the n random variables (or realizations) from a random sample of size n .

$$\begin{aligned}\overline{X}_n &= \frac{1}{n} (X_1 + \dots + X_n) \\ &= \frac{1}{n} \sum_{i=1}^n X_i\end{aligned}$$

Expectation of the sample mean:

$$\mathbb{E} \left[\overline{X}_n \right] = \mu$$

Variance of the sample mean:

$$\text{Var} \left(\overline{X}_n \right) = \frac{\sigma^2}{n}$$

The Central Limit Theorem

Let X_1, \dots, X_n form a random sample of size n from a distribution with finite mean and variance. Then for any fixed number x ,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left[\sqrt{n} \frac{(\overline{X} - \mu)}{\sigma} \leq x \right] = \Phi(x)$$

where $\Phi(x)$ is the CDF of a standard normal random variable.

Statistics

An **estimator** is a function of the random variables in a random sample.

A **parameter** is a constant indexing a family of distributions.

The function of the random sample is the **estimator**. The number, or realization of the function of the random sample, is the **estimate**.

Example: Suppose $X \sim U[0, \theta]$

$$f_X(x) = \begin{cases} \frac{1}{\theta} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

We want to estimate θ .

$$\hat{\theta}_1 = \max\{X_1, \dots, X_n\}$$

$$\hat{\theta}_2 = \frac{2}{n} \sum_{i=1}^n X_i$$

Module 6

Assessing and Deriving Estimators

An estimator is **unbiased** for θ if $\mathbb{E} \left[\hat{\theta} \right] = \theta$ for all θ in Θ .

Example X_i i.i.d. $U [0, \theta]$

$$\hat{\theta} = 2 \frac{1}{n} \sum_{i=1}^n X_i$$

$$\mathbb{E} \left[\hat{\theta} \right] = 2 \frac{1}{n} \sum_{i=1}^n \mathbb{E} [X_i] = 2 \frac{1}{n} \sum_{i=1}^n \frac{\theta}{2} = \theta \quad \text{unbiased}$$

Theorem The sample mean for an i.i.d. sample is unbiased for the population mean.

Theorem The sample variance for an i.i.d. sample is unbiased for the population variance, where the sample variance is

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \overline{X}_n \right)^2$$

Given two unbiased estimators, $\hat{\theta}_1$ & $\hat{\theta}_2$, $\hat{\theta}_1$ is more **efficient** than $\hat{\theta}_2$ if, for a given sample size,

$$\text{Var} \left(\hat{\theta}_1 \right) < \text{Var} \left(\hat{\theta}_2 \right)$$

Mean Squared Error Sometimes we are interested in trading off bias and variance/efficiency.

$$\text{MSE} \left(\hat{\theta} \right) = \mathbb{E} \left[\left(\hat{\theta} - \theta \right)^2 \right] = \text{Var} \left(\hat{\theta} \right) + \left(\mathbb{E} \left[\hat{\theta} \right] - \theta \right)^2$$

$\hat{\theta}$ is a **consistent** estimator for θ if

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\left| \theta - \hat{\theta}_n \right| < \delta \right) = 1.$$

Roughly, an estimator is consistent if its distribution collapses to a single point at the true parameter as $n \rightarrow \infty$.

Method of Moments

Population Moments (about the origin):

$$\mathbb{E} [X], \mathbb{E} \left[X^2 \right], \mathbb{E} \left[X^3 \right], \dots$$

Sample Moments:

$$\frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n X_i^2, \frac{1}{n} \sum_{i=1}^n X_i^3, \dots$$

If you have k parameters to estimate, you will have k **moment conditions**. In other words, you will have k equations in k unknowns to solve.

Maximum Likelihood Estimation

The maximum likelihood estimator of a parameter θ is the value $\hat{\theta}$ which most likely would have generated the observed sample.

Likelihood Function

$$L \left(\theta | x \right) = \prod_{i=1}^n f \left(x_i | \theta \right)$$

We just maximize L over θ in Θ .

Properties:

- If there is an efficient estimator in a class of consistent estimators, MLE will produce it.
- Under certain regularity conditions, MLEs will have asymptotically normal distributions.

Disadvantages:

- They can be biased.
- They might be difficult to compute.
- They can be sensitive to incorrect assumptions about the underlying distribution, more so than other estimators.

Confidence Intervals and Hypothesis Testing

The **standard error** of an estimate is the standard deviation (or estimated standard deviation) of the estimator.

χ² Distribution

The sample variance

S^2 = 1/(n-1) * sum (X_i - X_bar_n)^2

is an unbiased estimator for the variance of a distribution.

((n-1)S^2)/sigma^2 ~ chi^2_{n-1}

t Distribution

If X ~ N(0, 1) and Z ~ chi^2_n and they're independent, then

X/sqrt(Z/n) ~ t_n.

Suppose we are sampling from a N(mu, sigma^2) distribution. Then,

sqrt(n) * (X_bar - mu) / S ~ t_{n-1}

F Distribution

If X ~ chi^2_n and Z ~ chi^2_m and they're independent, then

(X/n) / (Z/m) ~ F_{n,m}.

Confidence Intervals

Case 1 We are sampling from a normal distribution with a known variance and we want a confidence interval for the mean.

P[Phi^-1(alpha/2) < sqrt(n) * (X_bar - mu) / sigma < -Phi^-1(alpha/2)] = 1 - alpha

CI_{1-alpha} = [X_bar + Phi^-1(alpha/2) * sigma/sqrt(n), X_bar - Phi^-1(alpha/2) * sigma/sqrt(n)]

Case 2 We are sampling from a normal distribution with an unknown variance and we want a confidence interval for the mean.

P[t_{n-1}^-1(alpha/2) < sqrt(n) * (X_bar - mu) / sigma < -t_{n-1}^-1(alpha/2)] = 1 - alpha

CI_{1-alpha} = [X_bar + t_{n-1}^-1(alpha/2) * sigma/sqrt(n), X_bar - t_{n-1}^-1(alpha/2) * sigma/sqrt(n)]

Hypothesis Testing

An hypothesis is an assumption about the distribution of a random variable in a population.

A maintained hypothesis is one that cannot or will not be tested.

A testable hypothesis is one that can be tested using evidence from a random sample.

The null hypothesis, H_O, is the one that will be tested.

The alternative hypothesis, H_A, is a possibility (or series of possibilities) other than the null.

We might want to perform a test concerning unknown parameter theta where X_i ~ f(x|theta).

H_O : theta in Theta_O
H_A : theta in Theta_A, where Theta_O and Theta_A disjoint.

A simple hypothesis is one characterized by a single point, i.e., Theta_O = theta_O.
A composite hypothesis is one characterized by multiple points, i.e., Theta_O has multiple values or a range of values.

Example

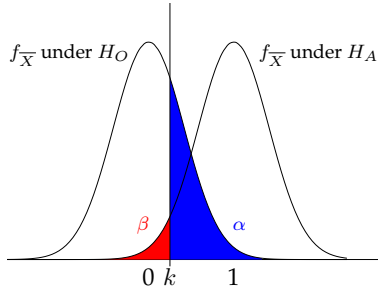
X_i i.i.d. N(mu, sigma^2), where sigma^2 known
Interested in testing whether mu = 0

H_O : mu = 0 null hypothesis, simple
H_A : mu = 1 alternative hypothesis, simple

Table with 2 columns: Action, H_O True, H_O False. Rows: Accept H_O, Reject H_O.

The significance level of the test, alpha, is the probability of type I error.
The operating characteristic of the test, beta, is the probability of type II error.
1 - alpha is the confidence level.
1 - beta is the power.

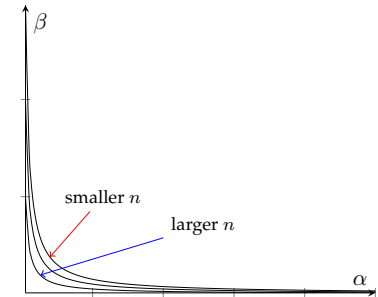
We define the critical region of the test, C or C_X, as the region of the support of the random sample for which we reject the null. The critical region will take the form X_bar > k for some k yet to be determined.



Choice of any one of alpha, beta, or k determines the other two. This involves an explicit trade-off between the probability of type I and type II errors.

- increasing k means alpha down and beta up
- decreasing k means alpha up and beta down

What happens as n increases or decreases?



Power Calculations

We tend to pick alpha low because society does not want to conclude that some treatment work when in fact it really does not.

We want to pick N = N_c + N_t such that, if the average treatment effect is in fact some value tau, the power of the test will be at least 1 - beta for some beta, given that a fraction gamma of the units are assigned to the treatment group.

In addition, we must assume (know) something about the variance of the outcome in each treatment arm: for simplicity, we often assume it is the same, and some parameter sigma^2.

In summary, we know, impose, or assume alpha, beta, tau, sigma, and gamma, and we are looking for N.

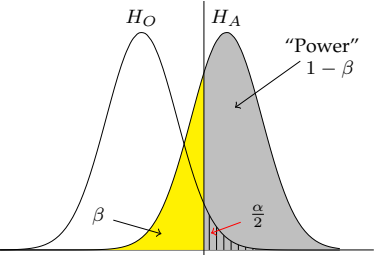
Alternatively, we could be interested in the power for a given sample size: we know alpha, beta, tau, sigma, and N and look for beta.

T = (Y_t^obs - Y_c^obs) / sqrt(V_Neyman) approx (Y_bar_t^obs - Y_bar_c^obs) / sqrt(sigma^2/N_t + sigma^2/N_c)

We reject this hypothesis if |T| > Phi(1 - alpha/2), e.g. if alpha = 0.05, if |T| > 1.96.

(Y_bar_t^obs - Y_bar_c^obs - tau) / sqrt(sigma^2/N_t + sigma^2/N_c) approx N(0, 1)

Statistical Power



P(|T| > Phi(1 - alpha/2)) approx Phi(-Phi^-1(1 - alpha/2) + tau / sqrt(sigma^2/N_t + sigma^2/N_c)) + Phi(-Phi^-1(1 - alpha/2) - tau / sqrt(sigma^2/N_t + sigma^2/N_c))

The second term is very small so we ignore it. We want the first term to be equal to 1 - beta:

Phi^-1(1 - beta) = -Phi^-1(1 - alpha/2) + tau * sqrt(N) * sqrt(gamma(1 - gamma)) / sigma

where gamma = N_t / N.

The required sample size is

N = ((Phi^-1(1 - beta) + Phi^-1(1 - alpha/2))^2) / (gamma^2 * gamma(1 - gamma))

- With stratified design, the variance of the estimated treatment effect is lower.
- With clustered design, the variance of the estimated treatment effect is larger.

Stratified Design

- Take the difference in means within each strata.
- Take a weighted average of the treatment effect with weight the size of the strata:

$$\sum_g \left(\frac{N_g}{N} \right) \hat{\tau}_g$$

- This will be an unbiased estimate of the average treatment effect.
- The variance will be calculated as:

$$\sum_g \left(\frac{N_g}{N} \right)^2 \hat{V}_g$$

- Special case: probability of assignment to control group stays the same in each strata. Then this coefficient is equal to the simple difference between treatment and control, but the variance is always weakly lower.
- Stratification will lower the required sample size for a given power.

Clustered Design

- We need to take into account the fact that the potential outcomes for units within randomization clusters are not independent.
- Conservative way to do this: just average the outcome by unit and treat each one as an observation.
- The number of observations is the number of clusters and you can analyze this data exactly as a completely randomized experiment but with clusters as the unit of analysis.
- A randomization with two clusters is unlikely to go very far!

Module 7

Causality

Definition of Causal Effects For any unit, the causal effect of a treatment is the difference between the potential outcome with and without the treatment.

Example Consider a single unit contemplating whether or not to take an aspirin for headache. The unit-level causal effect involves one of four possibilities:

1. Headache gone only with aspirin:
 $Y(\text{Aspirin}) = \text{No Headache}, Y(\text{No Aspirin}) = \text{Headache}$
2. No effect of aspirin, with a headache in both cases:
 $Y(\text{Aspirin}) = \text{Headache}, Y(\text{No Aspirin}) = \text{Headache}$
3. No effect of aspirin, with the headache gone in both cases:
 $Y(\text{Aspirin}) = \text{No Headache}, Y(\text{No Aspirin}) = \text{No Headache}$
4. Headache gone only without aspirin:
 $Y(\text{Aspirin}) = \text{Headache}, Y(\text{No Aspirin}) = \text{No Headache}$

Unit	Potential Outcomes		Causal Effect
	$Y(\text{Aspirin})$	$Y(\text{No Aspirin})$	
You	No Headache	Headache	Improvement due to Aspirin

The Problem of Causal Inference

- The definition of the causal effect depends on the potential outcomes, but it does *not* depend on which outcome is actually observed.
- The causal effect is the comparison of potential outcomes, for the same unit, at the same moment in time post-treatment. The fundamental problem of causal inference is therefore the problem that at most one of the potential outcomes can be realized and thus observed.
- We must rely on multiple units to make causal inferences.

Stable Unit Treatment Value Assumption (SUTVA) The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

The Assignment Mechanism Let's assume we have a population of size N , indexed by i . Let the treatment indicator W_i take on the values 0 (the control treatment) and 1 (the active treatment). We have one realized (and possibly observed) potential outcome for each unit, denoted by Y_i^{obs} :

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

For each unit we also have one missing potential outcome, Y_i^{mis} ,

$$Y_i^{\text{mis}} = Y_i(1 - W_i) = \begin{cases} Y_i(1) & \text{if } W_i = 0, \\ Y_i(0) & \text{if } W_i = 1. \end{cases}$$

Comparisons of $Y_i(1)$ and $Y_i(0)$ are unit-level causal effects:

$$Y_i(1) - Y_i(0)$$

Missing data problem Given any treatment assigned to an individual unit, the potential outcome associated with any alternate treatment is missing. A key role is therefore played by the missing data mechanism, or the *assignment mechanism*. How is it determined which units get which treatments or, equivalently, which potential outcomes are realized and which are not?

The Selection Problem Imagine we have a larger group of people who took aspirin and a group who did not, and we decide to take the sample mean of headache for people who got or did not get the pill. We know that this is a good estimator for

$$\mathbb{E}[Y_i|W_i = 1] - \mathbb{E}[Y_i|W_i = 0].$$

$$\begin{aligned} \mathbb{E}[Y_i^{\text{obs}}|W_i = 1] - \mathbb{E}[Y_i^{\text{obs}}|W_i = 0] &= \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] \\ &= \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 1] \\ &\quad + \mathbb{E}[Y_i(0)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] \end{aligned}$$

Treatment effect on the treated $\mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 1]$

Selection bias $\mathbb{E}[Y_i(0)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0]$

Randomization solves the selection problem In a completely randomized experiment, N_t units are randomly drawn to be in the treatment group, and N_c units are drawn to be in the control group. Then, the probability of assignment does not depend on potential outcomes:

$$\mathbb{E}[Y_i(0)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] = 0$$

and

$$\begin{aligned} \mathbb{E}[Y_i^{\text{obs}}|W_i = 1] - \mathbb{E}[Y_i^{\text{obs}}|W_i = 0] &= \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 1] \\ &= \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1] \\ &= \mathbb{E}[Y_i(1) - Y_i(0)] \end{aligned}$$

Types of RCT

- Complete randomization
- Stratified randomization
- Pairwise randomization
- Clustered randomization

Analyzing Randomized Experiments

The Average Treatment Effect

$$\text{ATE} = \mathbb{E}[Y_i^{\text{obs}}|W_i = 1] - \mathbb{E}[Y_i^{\text{obs}}|W_i = 0]$$

Suppose we have a completely randomized experiment with N_t treatment units and N_c control units. The difference in sample average

$$\hat{\tau} = \frac{1}{N_t} \sum_{i:W_i=1} Y_i^{\text{obs}} - \frac{1}{N_c} \sum_{i:W_i=0} Y_i^{\text{obs}} = \bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}.$$

The variance of a difference of two statistically independent variables is the sum of their variances. Thus,

$$V(\hat{\tau}) = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t}.$$

To estimate the variance, $\hat{V}(\hat{\tau})$, replace S_c^2 and S_t^2 by their sample counterpart:

$$\begin{aligned} s_c^2 &= \frac{1}{N_c - 1} \sum_{i:W_i=0} \left(Y_i(0) - \bar{Y}_c^{\text{obs}} \right)^2 \\ s_t^2 &= \frac{1}{N_t - 1} \sum_{i:W_i=0} \left(Y_i(0) - \bar{Y}_t^{\text{obs}} \right)^2 \end{aligned}$$

Confidence Intervals We want to find a function of the random samples A and B such that

$$\mathbb{P}(A(X_1, \dots, X_N) < \theta < B(X_1, \dots, X_N)) > 1 - \alpha$$

The ratio of the difference and the estimated standard error will follow a t -distribution, so

$$CI_{1-\alpha}^{\tau} = \left(\hat{\tau} - t_{\text{crit}} \sqrt{\hat{V}}, \hat{\tau} + t_{\text{crit}} \sqrt{\hat{V}} \right).$$

With small samples, take t_{crit} from a table of t -distribution for the relevant α with $N_t + N_c - 1$ degrees of freedom.

Hypothesis Testing

$$\begin{aligned} H_0 : \frac{1}{N} \sum_{i=1}^N (Y_i(1) - Y_i(0)) &= 0 \\ H_1 : \frac{1}{N} \sum_{i=1}^N (Y_i(1) - Y_i(0)) &\neq 0 \end{aligned}$$

Natural Test Statistic

$$t = \frac{\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}}{\sqrt{\hat{V}}}$$

follows a t -distribution with $N - 1$ degrees of freedom or, with N large enough, a normal distribution.

Fisher Exact Test

Another view of uncertainty

(More) Exploratory Data Analysis: Non-Parametric Comparisons and Regressions

Kolmogorov-Smirnov Test

Let X_1, \dots, X_n be a random sample with CDF F and let Y_1, \dots, Y_n be a random sample with CDF G .

We are interested in testing the hypotheses

$$H_o : F = G$$
$$H_a : F \neq G$$

The Statistic

$$D_{nm} = \max_x |F_n(x) - G_m(x)|$$

where F_n and G_m are the empirical CDF of the first and second sample. The empirical CDF just counts the number of sample points below level x :

$$F_n(x) = \mathbb{P}_n(X < x) = \frac{1}{n} \sum_{i=1}^n I(X < x)$$

First Order Stochastic Dominance: One-sided Kolmogorov-Smirnov Test

- We are interested in testing the hypothesis

$$H_o : F = G$$

against

$$H_a : F > G$$

(which would mean that G FSD F).

- The one-sided KS statistics is:

$$D_{nm}^+ = \max_x [F_n(x) - G_m(x)]$$

Asymptotic Distribution of the KS Statistic

- Under H_o , the limit of KS as n and m go to infinity is 0, so we want to compare the KS statistics to 0. We will reject the hypothesis if the statistic is "large" enough.
- Under H_o , the distribution of

$$\left(\frac{nm}{n+m}\right)^{\frac{1}{2}} D_{nm}$$

has a known distribution (KS) with associated critical values.

- Therefore, we reject the null of equality if

$$D_{nm} > c(\alpha) \left(\frac{nm}{n+m}\right)$$

where $c(\alpha)$ are critical values which we find in tables.

Non-Parametric (Bivariate) Regression

You have two random variables, X and Y , and express the conditional expectation of Y given X as $\mathbb{E}[Y|X] = g(X)$. Therefore, for any x and y ,

$$y = g(x) + \epsilon$$

where ϵ is the prediction error. The problem is to estimate $g(x)$ without imposing a functional form.

The Kernel Regression

$$\mathbb{E}[Y|X = x] = \int y f(y|x) dy$$

By Bayes' rule:

$$\int y f(y|x) dy = \int \frac{y f(x, y)}{f(x)} dy = \frac{\int y f(x, y) dy}{f(x)}$$

Kernel Estimator Replace $f(x, y)$ and $f(x)$ by their empirical estimates:

$$\hat{g}(x) = \frac{\int y \hat{f}(x, y) dy}{\hat{f}(x)}$$

Denominator:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

Numerator:

$$\frac{1}{nh} \sum_{i=1}^n y_i K\left(\frac{x - x_i}{h}\right)$$

where h (the bandwidth) is the kernel estimate of the density of x . $K()$ is a density.

Large sample properties:

- As h goes to zero, the bias goes to zero.
- As nh goes to infinity, variance goes to zero.
- As you increase the number of observation, you *promise* to decrease the bandwidth.

Choices to make:

- Choice of kernel
 1. Histogram: $K(u) = \frac{1}{2}$ if $|u| \leq 1$, $K(u) = 0$ otherwise.
 2. Epanechnikov: $K(u) = \frac{3}{4}(1 - u^2)$ if $|u| \leq 1$, $K(u) = 0$ otherwise.
 3. Quartic: $K(u) = (\frac{3}{4}(1 - u^2))^2$ if $u \leq 1$, $K(u) = 0$ otherwise.
- Choice of bandwidth: trade off bias and variance
 - A large bandwidth will lead to more bias.
 - A small bandwidth will lead to more variance.

Module 8

The Linear Model

Linear Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \dots, n$$

Basic assumptions:

1. X_i, ϵ_i uncorrelated
2. identification

$$\frac{1}{n} \sum_i (X_i - \overline{X})^2 > 0$$

Sample variance is positive.

3. zero mean: $\mathbb{E}[\epsilon_i] = 0$
4. homoskedasticity: $\mathbb{E}[\epsilon_i^2] = \sigma^2$ for all i
5. no serial correlation: $\mathbb{E}[\epsilon_i \epsilon_j] = 0$ if $i \neq j$

Assumptions 3-5 could be subsumed under a stronger assumption: ϵ_i i.i.d. $\mathcal{N}(0, \sigma^2)$.

Properties:

- $\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i$
- $\text{Var}(Y_i) = \mathbb{E}[\epsilon_i^2] = \sigma^2$
- $\text{Cov}(Y_i, Y_j) = 0, \quad i \neq j$

Estimates for β_0 and β_1

- **least squares** (OLS)

$$\min_{\beta} \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2$$

- **least absolute deviations**

$$\min_{\beta} \sum_i |Y_i - \beta_0 - \beta_1 X_i|$$

- **reverse least squares**

$$\min_{\beta} \sum_i \left(X_i - \frac{\beta_0 + Y_i}{\beta_1}\right)^2$$

Under the assumptions of the Classical Linear Regression Model, OLS provides the minimum variance (most efficient) unbiased estimator of β_0 and β_1 . It is the MLE under normality of errors, and the estimates are consistent and asymptotically normal.

Closed-form solutions:

$$\hat{\beta}_1 = \frac{\frac{1}{n} \sum (X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n} \sum (X_i - \overline{X})^2}, \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Fitted Value $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Residual $\hat{\epsilon}_i = Y_i - \hat{Y}_i$

Regression Line or Fitted Line $Y = \hat{\beta}_0 + \hat{\beta}_1 X$

Let $\overline{X} = \frac{1}{n} \sum X_i$ and $\hat{\sigma}_X^2 = \frac{1}{n} \sum (X_i - \overline{X})^2$.

	Mean	Variance	Covariance
$\hat{\beta}_0$	β_0	$\frac{\sigma^2 \overline{X}^2}{n \hat{\sigma}_X^2} + \frac{\sigma^2}{n}$	$-\frac{\sigma^2 \overline{X}}{n \hat{\sigma}_X^2}$
$\hat{\beta}_1$	β_1	$\frac{\sigma^2}{n \hat{\sigma}_X^2}$	

Some comparative statistics:

- A larger σ^2 means larger $\text{Var}(\hat{\beta})$
- A larger $\hat{\sigma}_X^2$ means smaller $\text{Var}(\hat{\beta})$
- A larger n means smaller $\text{Var}(\hat{\beta})$
- If $\overline{X} > 0$, $\text{Cov}(\beta_0, \beta_1) < 0$.

If we use the stronger assumption that the errors are i.i.d. $\mathcal{N}(0, \sigma^2)$, $\hat{\beta}_0$ and $\hat{\beta}_1$ will also have normal distributions.

Analysis of Variance

We want some way to indicate how much of Y 's variation is *explained* by X 's variation. We perform an analysis of variance and that leads us to a measure of goodness-of-fit.

Sum of Squared Residuals (SSR)

$$\text{SSR} = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_i (\hat{\epsilon}_i)^2$$

Total Sum of Squares (SST)

$$\text{SST} = \sum_i (Y_i - \overline{Y})^2$$

Model Sum of Squares (SSM)

$$\text{SSM} = \sum_i (\hat{Y}_i - \overline{Y})^2$$

The fact that the regression line is the *least squares* line ensures that $\text{SSR} \leq \text{SST}$.

$$0 \leq \frac{\text{SSR}}{\text{SST}} \leq 1$$

We want a measure of fit that had larger values when the fit was better so we define

$$R^2 = 1 - \frac{SSR}{SST}.$$

In addition to using R^2 as a basic measure of goodness-of-fit, we can also use it as the basis of a test of the hypothesis that $\beta_1 = 0$. We reject the hypothesis when

$$\frac{(n-1)R^2}{1-R^2},$$

which has an F distribution under the null, is large.

Interpretation $\hat{\beta}_1$ is the estimated effect on Y of a one-unit increase in X .

The Multivariate Linear Model

General Linear Model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i, \quad i = 1, \dots, n$$

Matrix notation:

$$Y = X\beta + \epsilon$$

Assumptions:

- 1. **identification:** $n > k + 1$, X has full column rank $k + 1$ (i.e., regressors are linearly independent; $X^T X$ is invertible)
- 2. **error behavior:** $\mathbb{E}(\epsilon) = 0$, $\mathbb{E}(\epsilon \epsilon^T) = \text{Cov}(\epsilon) = \sigma^2 \mathbb{I}_n$
stronger version $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$

$\hat{\beta}$ is the vector that minimizes the sum of squared errors, i.e.,

$$\hat{\epsilon}^T \hat{\epsilon} = (Y - X\hat{\beta})^T (Y - X\hat{\beta})$$

Solution:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad \text{if } (X^T X) \text{ is invertible}$$

Properties:

- $\mathbb{E}[\hat{\beta}] = \beta$
- $\text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$
- $\hat{\sigma}^2 = \frac{\hat{\epsilon}^T \hat{\epsilon}}{n - k}$

Inference in the Linear Model

Consider the hypotheses

$$H_O : R\beta = c$$
$$H_A : R\beta \neq c.$$

R is a $r \times (k + 1)$ matrix of restrictions. If, for instance, $R = [0 \ 1 \ 0 \ \dots \ 0]$ and $c = [0]$, that corresponds to $H_O : \beta_1 = 0$. If

$$R = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \quad \text{and} \quad c = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

that corresponds to $H_O : \beta_1 = \beta_2 = \dots = \beta_k = 0$.

One thing you cannot do in this framework is test one-sided hypotheses.

Steps:

- 1. We estimate the unrestricted model.

- 2. We impose the restrictions of the null and estimate that model.
- 3. We compare the goodness-of-fit of the models.

What if the restriction is that some $\beta = c$? This is an F -test.

$$T = \frac{\frac{1}{r} (SSR_R - SSR_U)}{\frac{SSR_U}{n - (k + 1)}}$$

$T \sim F_{r, n - (k + 1)}$ under the null and we reject the null for large values of the test statistic.

$$H_O : \beta_i = c$$

$$T = \frac{\hat{\beta}_i - c}{\text{SE}(\hat{\beta}_i)} \quad \text{where} \quad \text{SE}(\hat{\beta}_i) = \left(\sigma^2 (X^T X)^{-1} \right)_{ii}$$

$$H_O : R\beta = c$$

$$T = \frac{R\hat{\beta} - c}{\text{SE}(R\hat{\beta})} \quad \text{where} \quad \text{SE}(R\hat{\beta}) = \left(\sigma^2 R(X^T X)^{-1} R^T \right)^{1/2}$$

Module 9

Practical Issues in Running Regressions

Dummy Variables

$$Y_i = \alpha + \beta D_i + \epsilon_i$$

D_i is a dummy variable, or an indicator variable, if it takes the value 1 if the observation is in group A and 0 if in group B .

Interpretation Without any control variables, then

$$\hat{\beta} = \bar{Y}_A - \bar{Y}_B.$$

You can always estimate the difference between the treatment and control groups for an RCT using an OLS regression framework.

Categorical Variables If there are more than two groups, you can transform them into dummy variables, one for each group. **Warning:** Omit one category to avoid multi-collinearity.

Interpretation Each coefficient is the difference between the value of that group and the value for the omitted (reference) group.

Other Variables in the Regression

$$Y_i = \alpha + \beta D_i + \gamma X_i + \epsilon_i$$

$\hat{\beta}$ is the difference in intercept between group A and group B . X_i s are “control” variables – things that did not affect the assignment but may have been different at baseline.

Dummy Variables and Interactions Imagine you have two sets of dummy variables, say, Treatment and Control D_i , Male and Female M_i :

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

- $\hat{\alpha}$: an estimate of the mean for women in the control group
- $\hat{\beta}$: an estimate of the difference between treatment and control group means for women (**treatment main effect**)
- $\hat{\gamma}$: an estimate of the difference between males and females (**gender main effect**)
- $\hat{\delta}$: an estimate of the difference between the treatment effect for males and for females (**interaction effect**)

This is the basic **difference-in-differences** model which is used by empirical researchers in a situation where there was a change in the law (or an event) affecting one group but not the other, and you are willing to assume that in the absence of the law, the difference between the two groups would have remained stable over time.

More Generally: Interactions More generally, the coefficient on the interaction between a dummy variable and some variable X tells us the extent to which the dummy variable changes the regression function for that regressor.

$$Y_i = \beta_0 + \beta_0^* D_i + \beta_1 X_{1i} + \beta^* D_i X_{1i} + \cdots + \epsilon_i$$

Transformations of the Dependent Variable

Suppose

$$Y_i = A X_{1i}^{\beta_1} X_{2i}^{\beta_2} e^{\epsilon_i}.$$

Then run the linear regression

$$\log(Y_i) = \beta_0 + \beta_1 \log X_{1i} + \beta_2 \log X_{2i} + \epsilon_i$$

to estimate β_1 and β_2 . Note that β_1 and β_2 are **elasticities**: when X_{1i} changes by 1%, Y_i changes by $\beta_1\%$.

Returns to education formulation

$$\log Y_i = \beta_0 + \beta_1 S_i + \epsilon_i$$

When education increases by 1 year, wages increase by $\beta_1 \times 100\%$.

Box Cox Transformation Suppose

$$Y_i = \frac{1}{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i}.$$

Then run the regression

$$\frac{1}{Y_i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i.$$

Discrete Choice Model Suppose

$$P_i = \frac{e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i}}{1 + e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i}},$$

P_i is the percentage of individuals choosing a particular option (e.g., buying a particular car), then run the regression

$$Y_i = \log \left(\frac{P_i}{1 - P_i} \right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

Polynomial Models

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \cdots + \beta_k X_{1i}^k \epsilon_i$$

- You can choose straight polynomial, series expansion, orthogonal polynomials, etc.
- If you assume that the model is known, this is just standard OLS.
- If you assume that the model is not known, this is a non-parametric method – there is bias (because the shape is never quite perfect) and variance (as you add more X s) so you add more terms as the number of observations increases (**series regression**).

Regression Discontinuity Design

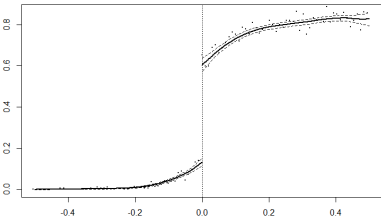
- add polynomials

$$Y_i = \beta_0 + \beta_1 D_{ai} + \beta_2 a_i + \beta_3 a_i^2 + \beta_4 a_i^3 + \epsilon_i$$

- fit a polynomial on each side of the discontinuity:

$$Y_i = \beta_0 + \beta_1 D_{ai} + \beta_2 (a_i - a_0) + \beta_3 (a_i - a_0) * D_{ai} + \cdots + \epsilon_i$$

Centering the variables ensure that β_1 is still the jump at the discontinuity.



Omitted Variable Bias

Suppose that the regression model excludes a key variable (e.g., the data is unavailable).

Example: Consider the model:

log(W_i) = beta_0 + beta_1 E_i + beta_2 X_i + beta_3 A_i + epsilon_i

where W_i is wage, E_i is education, X_i is job experience and A_i is ability. We are interested in measuring the effects of E_i and X_i on W_i with A_i constant. Suppose the A_i is unavailable so we run the regression without it. Next, we, instead, use IQ as proxy to the omitted variable and run the regression. The results are shown below.

log(wage)	Coeff.	SE		Coeff.	SE
Education	0.078	0.007		0.057	0.007
Experience	0.020	0.003		0.020	0.003
IQ	—	—		0.006	0.001
Constant	5.503	0.112		5.198	0.122

The estimated return to education changes from 7.8% to 5.7%.

More Advanced Techniques

Machine Learning: Double Post LASSO

- Suppose that we have lots of variables and we are not sure which ones to include.
- There are machine learning techniques to learn which variables are predictive.
- Three steps:
 - Regress X_1 on all the available variables and see what LASSO picks. Call this X_2.
 - Regress Y on all the available variables and see what LASSO picks. Call this X_3.
 - Run

Y_i = beta_0 + beta_1 X_{1i} + beta_2 X_{2i} + beta_3 X_{3i} + epsilon_i.

Module 10

Machine Learning

Estimation

- Strict assumptions about data generating process
- Back out coefficients
- Low-dimensional

Prediction

- Allow for flexible functional forms
- Get high quality predictions
- Give up on adjudicating between observably similar functions (variables)

Understanding OLS In-sample fit vs. out-of-sample fit

beta_hat^OLS = arg min_beta E_{S_n} (beta; x - y)^2

beta_hat^*_prediction = arg min_beta E_{(y,x)} (beta'x - y)^2

OLS looks good with the sample you have. We overfit in estimation.

Processing of data requires machine learning.

Two kinds of processing:

- Pre-processing
- Processing

Visualizing Data

Two different goals of data visualization

- For yourself: getting a sense of what is in the data – to guide future analysis
- For others: telling a story about the data and your results – to communicate your results

Scientific visualization

What to achieve:

- Show the data.
- Not lie about it.
- Illustrate a story.
- Reduce clutter.
- Visualization must complement the text and have enough information to stand alone.

Tufte’s Principles

- Show the data.
- Maximize data-ink ratio.
- Erase non-data ink (as much as possible).
- Erase redundant data ink.
- Avoid chart junk (moiré, ducks).
- Try to increase the density of data ink.
- Graphics should tend to be horizontal.

Module 11

Endogeneity and Instrumental Variables

Consider a more general model:

{ Y_i = beta_0 + beta_1 X_i + beta_2 T_i + epsilon_i, X_i = alpha_0 + alpha_1 Y_i + alpha_2 Z_i + delta_i }

Endogenous variables (X_i and Y_i) are determined within the system.

We talk about endogeneity when there is mutual relationship, i.e., when reasonable case can be made either way.

Instrumental Variables

An instrument for the model

Y_i = beta_0 + beta_1 X_i + epsilon_i

is a variable Z_i such that

Cov(Z, X) != 0 and Cov(Z, epsilon) = 0.

Three conditions:

- It affects X: Cov(Z, X) != 0.
- It is randomly assigned.
- It has no direct effect on Y: Cov(Z, epsilon) = 0 (exclusion restriction)

The IV estimation can be seen as a two-step estimator within a simultaneous equations model.

RCT as IV Let Z_i be a dummy variable equal to 1 if assigned to the treatment group and 0 otherwise. Then,

beta_hat_1 = (E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]) / (E[X_i|Z_i = 1] - E[X_i|Z_i = 0])

- The denominator is the first stage relationship.
- The numerator is the reduced form relationship.

- beta_hat_1 is the Wald estimate.

The interpretation of IV when the treatment effect is not constant Under a fairly mild assumption, the Wald estimate still has a causal interpretation. It captures the effect of the treatment on those who are compelled by the instrument to get treated (Local Average Treatment Effect or LATE).

From the Wald estimate to Two-Stage Least Squares (2SLS)

- We could couch this in a regression framework.

- First stage: pi_hat_1 in

X_i = pi_0 + pi_1 Z_i + delta_i

- Reduced form: gamma_hat_1 in

Y_i = gamma_0 + gamma_1 Z_i + omega_i

- Two-Stage Least Squares: Run the first stage and take the fitted values X_hat_i. In the second stage, run Y_i = beta_0 + beta_1 X_hat_i + epsilon_i.

The 2SLS and the Wald estimates are identical

beta_hat_1 = (Cov(Y_i, X_hat_i) / Var(X_hat_i)) = (Cov(Y_i, pi_0 + pi_1 Z_i) / Var(pi_0 + pi_1 Z_i)) = (pi_1 Cov(Y_i, Z_i) / pi_1^2 Var(Z_i)) = gamma_1 / pi_1

Experimental Design

What is experimental design?

- What is being randomized?
 - the intervention
- Who is being randomized?
 - the level of randomization (schools, individuals, villages, cells)
 - the sample over which you randomize
- How is randomization introduced?
 - method of randomization
 - stratification
- How many units are being randomized?
 - power

Randomization

- Simple randomization: define your sample frame and your unit of randomization, use software to randomly assign one group to treatment, one to control
- Stratification: create groups that are similar ex-ante
- Clustering: randomize at the group level

Introducing Randomization

- Phase-in design
- Randomization “in the bubble”
- Encouragement design

Some R commands

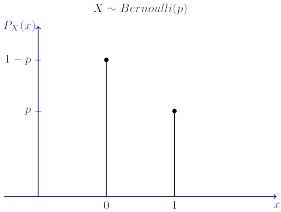
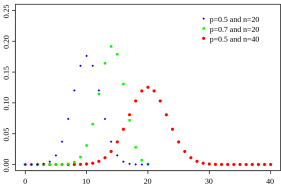
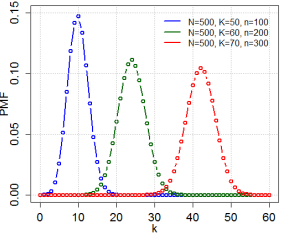
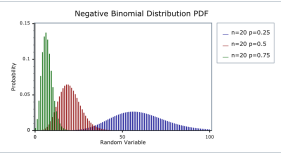
Command	Library	What it does
<code>chooseMatrix(n,m)</code>	<code>perm</code>	Create a matrix of <code>choose(n,m)</code> rows and <code>n</code> columns. The matrix has unique rows with <code>m</code> ones in each row and the rest zeros.
<code>NROW(x)</code> , <code>NCOL(x)</code>		Returns the number of rows or columns in matrix <code>x</code>
<code>var(x)</code>		Computes the variance of <code>x</code> , which is a vector, matrix or dataframe.
<code>covar(x,y)</code>		Computes the covariance of <code>x</code> and <code>y</code> , where both arguments are vectors, matrices or dataframes with comparable dimensions to each other.
<code>apply()</code>		Returns a vector or array or list of values obtained by applying a function to margins of an array or matrix.
<code>lm()</code>		Fits a linear model to the given data.
<code>confint()</code>		Computes confidence intervals for one or more parameters in a fitted model.
<code>felm()</code>	<code>lfe</code>	Fit linear models with multiple group fixed effects, similar to <code>lm</code> .
<code>DCdensity()</code>	<code>rdd</code>	Function to implement the McCrary (2008) sorting test.
<code>RDestimate()</code>	<code>rdd</code>	Function to calculate the Regression Discontinuity estimate.
<code>ivreg()</code>	<code>AER</code>	Fit IV regression by a two-stage least squares method.

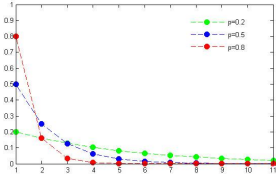
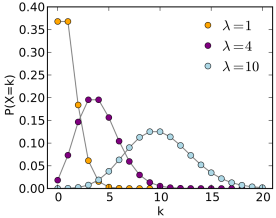
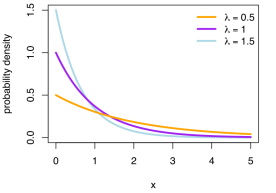
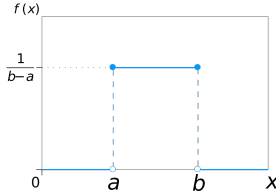
Recommended Resources

- Causal Inference for Statistics, Social, and Biomedical Sciences (Guido W. Imbens and Donald B. Rubin)
- Mastering 'Metrics (Joshua D. Angrist and Jörn-Steffen Pischke)
- Data Analysis for Social Scientists [Lecture Slides] (<http://www.edx.org>)
- R Studio (<https://www.rstudio.com>)

Please share this cheatsheet with friends!

Summary of Special Distributions

Distribution	PDF / PMF	Expectation and Variance	Graph
Bernoulli	$p_X(x) = p^x(1 - p)^{1-x},$ $x \in \{0, 1\}$	$\mathbb{E}[X] = p$ $\text{Var}(X) = p(1 - p)$	
Binomial	$p_X(x) = \binom{n}{x} p^x (1 - p)^{n-x},$ $x = 0, 1, \dots, n$	$\mathbb{E}[X] = np$ $\text{Var}(X) = np(1 - p)$	
Hypergeometric	$p_X(x) = \frac{\binom{A}{x} \binom{B}{n-x}}{\binom{A+B}{n}}$	$\mathbb{E}[X] = \frac{nA}{A+B}$ $\text{Var}(X) = \frac{nAB}{(A+B)^2} \frac{A+B-n}{A+B-1}$	
Negative Binomial	$p_X(k) = \binom{r+k-1}{k} p^k (1-p)^r$	$\mathbb{E}[X] = \frac{k(1-p)}{p}$ $\text{Var}(X) = \frac{r(1-p)}{p^2}$	

Distribution	PDF / PMF	Expectation and Variance	Graph
Geometric	$p_X(k) = (1 - p)^{k-1}p$	$\mathbb{E}[X] = \frac{1}{p}$ $\text{Var}(X) = \frac{1 - p}{p^2}$	
Poisson	$p_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}$	$\mathbb{E}[X] = \lambda$ $\text{Var}(X) = \lambda$	
Exponential	$f_X(x) = \lambda e^{-\lambda x}$	$\mathbb{E}[X] = \frac{1}{\lambda}$ $\text{Var}(X) = \frac{1}{\lambda^2}$	
Uniform	$f_X(x) = \frac{1}{b - a}$	$\mathbb{E}[X] = \frac{a + b}{2}$ $\text{Var}(X) = \frac{(b - a)^2}{12}$	
Normal	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\mathbb{E}[X] = \mu$ $\text{Var}(X) = \sigma^2$	