14.30 Summary Sheet

Sample Space & Events

- Sample Space S is a collection of possible outcomes of a trial
- An event A is a subset of all possible outcomes

* Events are sets

1. $\subset$ = subset
2. $\cap$ = Intersection
3. $\cup$ = Union

Properties:

- $A \subset B$, $A \cup B = B$
- $A \subset B$ & $B \subset A$, $A = B$
- $A \cup A^c = S$
  - ↳ complement set

Probability

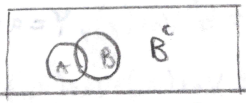
- Every event A assigned Probability $P(A)$

We Require:

1. $P(A) \geq 0$ for all $A \subseteq S$
2. $P(S) = 1$
3. For disjoint sets A_1, A_2, $P(A_1 \cup A_2) = \sum_{i=1}^{2} P(A_i)$

Properties of Probability:

- $P(A^c) = 1 - P(A)$
- $P(\phi) = 0$
- if $A \subset B$, $P(A) \leq P(B)$
- * $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- * $P(A \cap B^c) = P(A) - P(A \cap B)$



Simple Sample Space

- All events Equally Likely

$$P(A) = n(A)/n(S)$$

Combinatorics

1. If experiment happens in multiple parts
   - part 1: N outcomes
   - part 2: M outcomes
   - Total: $M \times N$
2. Permutation → # of Ordered ways to count objects

K objects out of N: $\frac{N!}{(N-K)!}$

3. Combination → # of unordered ways to count object

K objects out of N: $\frac{N!}{(N-K)!K!} = \binom{N}{K}$

Independence

- Events A & B independent if $P(AB) = P(A)P(B)$
  - ↳ Knowing one event occurs doesn't give you intuition about the other

Probability A happens exactly x times in next N Trials: $\left(\frac{N!}{(N-x)!x!}\right) P(A)^x P(A^c)^{N-x}$

Conditional Probability

- One event happening does tell you something about the other

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Versions of Bayes' Theorem

1. $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
2. $P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$

Random Variables

- X a random variable is a real-valued function whose domain is the sample space

Binomial Distribution: $f_x(x) = \binom{n}{x} p^x (1-p)^{n-x}$

Probability Function (Discrete RV)

- Discrete random variable X

Probability Function: $f_x(x) = P(X = x)$

Properties:

1. $0 \leq f_x(x) \leq 1$
2. $\sum_x^x f_x(x) = 1$
3. $P(A) = \sum_{x \in A} f_x(x_i)$

Probability Density Function (PDF)

- Function for Continuous RV that x lies in particular Region

$$P(X \in A) = \int_A f_x(x)\,dx$$

Cumulative Distribution Function (CDF)

- CDF of Random Variable x is

$$F_x(x) = P(X \leq x)$$

Properties:

1. $F_x(x)$ is non-decreasing in x
2. $\lim_{x \to -\infty} F_x(x) = 0$
3. $\lim_{x \to \infty} F_x(x) = 1$

$F_x(x) = \int_{-\infty}^{x} f_x(x)\,dx$

$F'(x) = f_x(x)$

CDF → Discrete RV

- Always Right-continuous

ex. $F_x(x) = \begin{cases} 0, & x < 0 \\ 1-p, & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases}$

example:

$f_x(x) = 1\{a < x < b\} \frac{1}{b-a}$

↳ indicator Function

$F_x(x) = \begin{cases} 0, & x < a \\ (x-a)/(b-a), & a \leq x \leq b \\ 1, & x > b \end{cases}$

Normal Distribution

$f_x(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} * e^{-x^2/2}$ → $N(0,1)$

Joint Distributions

Function: $P((x,y) \in A) = \iint f_{xy}(x,y)$

Joint + Marginal Distributions

- we can Recover Marginal distributions from Joint

Discrete	$f_x(x) = \sum_j f_{xy}(x, y_j)$
Continuous	$f_x(x) = \int_y f_{xy}(x,y)\,dy$

Independence of RV

- If X & Y are independent

$$F_{xy}(x,y) = F_x(x) F_y(y)$$

↳ If continuous:

$$f_{xy}(x,y) = f_x(x) f_y(y)$$

Conditional Distribution

- Conditional PDF of Y given X

$$f_{y|x}(y|x) = f_{xy}(x,y) / f_x(x)$$

(when discrete)

↳ $f_{y|x}(y|x) = P(Y=y \mid X=x)$

Example

$$f_{xy}(x,y) = \begin{cases} 2, & 0 \le y \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_{x|y}(x|y) = \frac{1}{1-y} \{y \le x \le 1\}$$

$$f_{y|x}(y|x) = \frac{1}{x} \{0 \le y \le x\}$$

Conditional Distribution / Independence

- If independent

$f_{y|x}(y|x) = f_y(y)$ & $f_{xy} = f_x f_y$

Functions of Random Variables

- X is random variable where $f_x(x)$ known.

we want $Y = h(x)$

$$F_y(y) = \int_{\{x : h(x) \le y\}} f_x(x)\, dx$$

↳ to Find CDF integrate accoreviate region

- If Y continuous $f_y(y) = F_y'(y)$

Example: Function of RV

$$f_x(x) = \begin{cases} 1/2, & \text{for } -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

- $Y = X^2$, what is $f_y(y)$?

Steps: $F_y(y) = P(Y \le y)$

$= P(X^2 \le y)$

$= P(-\sqrt{y} \le X \le \sqrt{y})$

so

$$F_y(y) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2}\, dx = \sqrt{y}$$

so

$$F_y(y) = \begin{cases} 0, & y < 0 \\ \sqrt{y}, & 0 \le y \le 1 \\ 1, & y > 1 \end{cases}$$

$$f_y(y) = \begin{cases} \frac{1}{2\sqrt{y}}, & 0 \le y \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Linear Transformation Example

- X has PDF $f_x(x)$

↳ Let $Y = a + bX$ (How distribute)

case 1: $b > 0$

$F_y(y) = P(Y \le y) = P(a + bX \le y)$

$= P\left(X \le \frac{(y-a)}{b}\right)$

$= \int_{-\infty}^{\frac{y-a}{b}} f_x(x)\, dx$

so

$PDF = f_y(y) = F_y'(y) = \frac{1}{b} f_x\left(\frac{y-a}{b}\right)$

Expectation (or Mean)

continuous: $E(X) = \int x f_x(x)\, dx$

discrete: $E(X) = \sum_{j=1}^{J} x_j f_x(x_j)$

* Properties of Expectation

1. $E(a) = a$, for constant
2. $E(Y) = a + b E(X)$, $Y = a + bX$
3. $E(Y) = E(X_1) + E(X_2) + \dots E(X_n)$
   ↳ X doesnt need independence
4. $E(XY) = E(X) E(Y)$ if Independent

Expectation of RV Function

- we know $f_x(x)$, $Y = g(x)$
- we could figure out how Y is distributed & find $f_y(y)$

so $E(Y) = \int y f_y(y)$

Easier, $E(Y) = E(g(x)) = \int g(x) f_x(x)$

Variance

- $\mathrm{Var}(x) = E\left[(X - \mu)^2\right]$
  ↳ $\mathrm{Var}(x) = \sigma_x^2$

Properties of Variance

1. $\mathrm{Var}(X) \ge 0$
2. $\mathrm{Var}(a) = 0$
3. $\mathrm{Var}(Y) = b^2 \mathrm{Var}(x)$, $Y = a + bX$
4. $\mathrm{Var}(Y) = \mathrm{Var}(X_1) + \mathrm{Var}(X_n)$
   ↳ $X_1, X_2 \dots X_3$ independent
6. $\mathrm{Var}(X) = E[X^2] - (E[X])^2$ **

Variance of a Function

- WANT: $\mathrm{Var}(Y)$, where $Y = r(x)$

$\mathrm{Var}(Y) = E[Y^2] - E[Y]^2$

$= \int r(x)^2 f_x(x)\, dx - \left[\int r(x) f(x)\, dx\right]^2$

Example

- X takes on 0 or 1 (Bernoulli RV)

$P(X=0)=0.22$ | $P(X=1)=.78$

- $\overline{X_n}$ depends on n

if $n=1$: $\overline{X_1}=0,1=0.22$ | $P(\overline{X_1}=1)=0.78$

if $n=2$: $P(\overline{X_2}=0)=0.22^2=.0484$ | $P(\overline{X_2}=0.5)=.3432$ | $P(\overline{X_2}=1)=.6084$

Central Limit Theorem

- Let $X_1,\dots X_n$ be a random sample (i.i.d) of size n w/ finite mean & variance

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

↳ $\Phi(x)$ = Normal CDF

Estimator

what makes "good" Estimator

↳ $\hat{\theta}$ is unbiased if & only if

$$E[\hat{\theta}]=\theta_0$$

↳ $\hat{\theta}$ is consistent if for all $\varepsilon>0$

$$\lim_{n\to\infty} P(|\hat{\theta}-\theta_0|>\varepsilon)=0$$

(Law of large numbers)

* sample mean is consistent estimator of population mean

More Criteria

- If estimator not consistent, it's bad

↳ Let $n/2$ be largest integer less than $n/2$

$\tilde{\theta}=\frac{1}{\lfloor n/2\rfloor}$, this is unbiased & consistent, why prefer $\overline{X}$?

$\mathrm{Var}(\tilde{\theta})=\frac{2\sigma_x^2}{n}$ | $\mathrm{Var}(\overline{X})=\frac{\sigma_x^2}{n}$

↳ we say $\hat{\theta}$ is more efficient than $\tilde{\theta}$

Efficiency of Estimators

- efficiency comparison hard to do for most
- Many Estimators are consistent w/ Normal Distribution in Large samples
- Compare them on variance of Large scale Normal populations

Confidence Intervals

- corresponds to margin of error in polls

say that, $\hat{p}=0.42$ & $n=1500$

Standard Error: $SE(\hat{p})=\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- a 95% confidance interval is

$$CI_{.95}=\left[\hat{p}-1.96\cdot SE(\hat{p}),\ \hat{p}+1.96\cdot SE(\hat{p})\right]$$

↳ $Pr(p_0\in CI_{.95})\to 0.95$ as $n\to\infty$

Properties of Confidance Interval

- It's a random Interval

↳ contains true value of parameter with some probability

- Data changes → But interval stays the same

Z-score

$$Z_n=\frac{\hat{p}-E[\hat{p}]}{SD(\hat{p})}\ \Big|\ \hat{p}=p_0\ \Big|\ \sqrt{\frac{p_0(1-p_0)}{n}}$$

2 Properties:

1. $\hat{p}$ is sample mean so by CLT

↳ $Z_n\sim N(0,1)$

2. $p_0\in CI_{.95}$ if $-1.96\le Z_n\le 1.96$

For General Distributions

$$SD(\hat{x})=\sigma_x/\sqrt{n}$$

$$\frac{\overline{X}-\mu_x}{SD(\overline{x})}=\frac{\sqrt{n}(\overline{x}-\mu_x)}{\sigma_x}\sim N(0,1)$$

$$\hat{\sigma}_x=s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

- If c_{a_1} is the critical value for $N(0,1)$ satisfying

$$Pr(-c_a<N(0,1)<c_a)$$

$$CI_a^{\mu_x}=\left[\overline{X}-c_a\sqrt{\frac{s^2}{n}},\ \overline{X}+c_a\sqrt{\frac{s^2}{n}}\right]$$

Hypothesis Tests

- consider Parameter $\mu=E[Y]$

we test a Null Hypothesis about μ

$H_0:\mu=m$

against alternative Hypothesis $H_1:\mu\neq m$

- we (reject or don't reject) the Null

Discrete Estimators

sample Average: $\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$

standard Error: $\widehat{s.e.}(\overline{Y})=\frac{\hat{\sigma}}{\sqrt{n}}$

↳ $\widehat{Var}=\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2$

t-statistic

$$t=\frac{\overline{Y}-m}{\widehat{s.e}(\overline{Y})}\xrightarrow{d}N(0,1)$$

if $|t|>1.96$, Reject the null

if $|t|<1.96$, Don't reject Null

Equivalently

$p=Pr(|N(0,1)|>|t|)$

$p<0.05$ reject Null

$p\ge 0.05$ Don't reject Null

Terminology

test statistic → stat used for testing null (t)

Null Distribution → sample distribution under H_0

Type I error → rejecting Null even if true

Type II error → Not rejecting Null if false

Critical Region → C_a set of values such that if $t\in C_a$, we reject null $(1.96,\infty)$

Two Sample Test

- Two samples $(X_1,\dots X_m)$ & $(X_{21},\dots X_{2n})$

$H_0:\mu_1-\mu_2=0$

$H_1:\mu_1-\mu_2\neq 0$

(t-test to see if same mean)

Test statistic:

$$t=\frac{\overline{X}_1-\overline{X}_2}{\sqrt{\frac{\hat{\sigma}_1}{n_1}+\frac{\hat{\sigma}_2}{n_2}}}$$

Conditional Expectation

- Is the expectation of a conditional distribution of Y given X

$$E(Y|X=x) = \int y f_{y|x}(y|x)\, dy$$

Law of Iterated expectations

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

Conditional Variance

- $Var(Y|X=x)$ is Variance of Distribution of Y given X

$$Var(Y|X=x) = E\left[\left(Y - E[Y|X=x]\right)^2 \middle| X=x\right]$$

Law of Total Variance:

$$Var(Y) = E\left(Var[Y|X]\right) + Var\left(E[Y|X]\right)$$

conditional Variance : Example

- N = # of patents produced in a year
  - $E(N) = 2$; $Var(N) = 2$
  - $p_{success} = 0.2$
- Probality 3 of 5 successful

$$\binom{5}{3}(0.2)^3(1-0.2)^3 = 0.0512$$

- # of expected successes

$$E(S) = E(E(S|N)) = E(Np) = 0.2(E(N)) = 0.4$$

Covariance & Correlation

- Key Moment of Joint Distribution is Covariance

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

$$Corr(X,Y) = \frac{Cov[X,Y]}{\sigma_X\, \sigma_Y}$$

$$\Downarrow$$

$$\rho_{xy}$$

$\rho > 0$	X & Y Positively correlated
$\rho < 0$	X & Y Negatively correlated
$\rho = 0$	Uncorrelated

Properties of Covariance & ρ

1. $Cov(X,X) = Var(X)$
2. $Cov(X,Y) = Cov(Y,X)$
3. $Cov(X,Y) = E(XY) - E(X)E(Y)$
4. $Cov(X,Y) = 0$
   - ↳ if X & Y independent
5. $Cov(aX+b, cY+d) = ac(Cov(X,Y))$
6. $Var(X+Y) = Var(X) + Var(Y) + 2\,Cov(X,Y)$
7. $|\rho(X,Y)| \le 1$
8. $|\rho(X,Y)| = 1$, if $Y = a + bX$

other Moment Metrics

mean = $E(Y)$

Variance = $E\left[(Y-\mu_y)^2\right]$

skewness = $\dfrac{E\left[(Y-\mu_y)^3\right]}{\sigma_y^3}$

Kurtosis = $\dfrac{E\left[(Y-\mu_y)^4\right]}{\sigma_y^4}$

Mode = Point where PDF has highest value

Median = Point where Integral of PDF = 0.5

Median

- Measure of Location where $F_x(m) = \frac{1}{2}$ ↳ halfway through dataset

Quantiles

- Let $a \in (0,1)$, CDF strictly increasing
- a^{th} quantile of X is $q_x(a)$ such that

$$F(q_a(X)) = a$$

- 4 quartiles
- 5 quintiles ~ 0.9^{th} quantile = 90th percentile
- $q_x(0.75) - q_x(0.5)$ = interquantile range
  - ↳ measure of Dispersion

Properties of Median & Quantiles

$$med(bX+c) = b * med(X) + c$$

$$med(X+Y) \neq med(X) + med(Y)$$

In General:

$$q_{bX+c}(a) = b\,q_x(a) + c$$

$$q_{X+Y}(a) \neq q_x(a) + q_y(a)$$

Markov inequality

- X is RV that non-Negative
  - ↳ E[X] exists

For any $t > 0$,

$$P(X \ge t) \le E(X)/t$$

Chebyshev Inequality

- X is a random variable for which Var(X) exists

For any $t > 0$

$$P(|X - E(X)| \ge t) \le Var(X)/t^2$$

Simple Random Sampling

- Choose an individual at Random from population
  1. X_1 & X_2 are independent
  2. X_1 & X_2 are identical

Data set $\{X_1, \ldots, X_n\}$ are i.i.d

Statistics

- Assume we have data set $\{x_1, \ldots, x_n\}$
  - ↳ each X_i has $F_x(x)$
- A Parameter characterises $F_x(x)$
  1. Expectation
  2. variance
- θ_0 = True Parameter
  - ↳ Reflects F_x data drawn from

* goal is to learn θ_0 from data & quantify uncertainty

Estimator

↳ uses Data to construct a "guess" of parameter value

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

↳ function of data

Sample Mean

- Arithmetic average of n random variables

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Linear Algebra Review

- A Matrix is a rectangular array

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & & & \\ a_{n1} & a_{n2} & & a_{nK} \end{pmatrix}$$

- Only Matrices of same Dimensions can be added / subtracted
- A scalar [c] can be Multiplied w/ matrix
- Transpose of a matrix achieved by turning Rows → columns

A = Matrix
A' = Transpose of Matrix

- a Column vector → matrix of single column
- Matrices must be conformable to be multiplied
- Transpose of a Product

$$(AB)' = B'A'$$

Square Matrices

- Square Matrix has same # of rows & columns
- Square matrix is Symmetric if for all elements $a_{ij} = a_{ji}$
- Diagonal Matrix is square matrix w/ only zeroes outside Main Diagonal

Properties of Matrices

- square Matrix is ~~Singular~~ Non-singular if all its Columns are linearly Independent
- If non-singular A^{-1} exists

System of Linear Equations

$$Ax = b \Rightarrow x = A^{-1}b$$

Random Vectors

- Let X be a $(K \times 1)$ Random Vector

$$X=\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{pmatrix} \quad \text{vector of Means} \quad \mu=\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \end{pmatrix}$$

- Variance of X is a $(K \times K)$ matrix

14.30 Summary Sheet #2

$$Var(X)=\begin{pmatrix} Var(X_1) & Cov(X_1,X_2) & \cdots & Cov(X_1,X_K) \\ Cov(X_1,X_2) & Var(X_2) & & \\ \vdots & & & \\ Cov(X_K,X_1) & & & Var(X_K) \end{pmatrix}$$

$$Var(X) = E\left[(X-\mu)(X-\mu)'\right]$$

↳ Variance-Covariance Matrix of X

Properties of Random vectors

- A = (m×K) matrix
- X = (m×1) random vector

$$E[AX] = A\,E[X] \quad \& \quad Var(AX) = A\,var(X)\,A'$$

Example: Functions of Random Vectors

$$X=\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \quad E[X]=\begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad var(X)=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Now Let: $Y_1 = X_1 - X_2$, $Y_2 = X_1 + X_2$

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = AX$$

$$A \to \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

$$E[Y] = E[AX] = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$var[Y] = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

Normal Random vectors

- Let X be a (K×1) random vector

$$X \sim N(\mu, V)$$

↳ X is vector of Jointly

μ = Mean
V = Var-Cov Matrix

The Chi-Square (χ^2_K) Distribution

- Let $X_1, \ldots, X_K$ be independent and $N(0,1)$. Then,

$$Z_K = X_1^2 + \ldots X_K^2 \sim \chi^2_K$$

$$E[Z_K] = K \quad \& \quad Var(Z_K) = 2K$$

Normal Random Vector Properties

1. If (X_1, X_2) are jointly normal & uncorrelated, they are independent
2. If $X \sim N(\mu, V)$ & $Y = AX$

$$Y \sim N(A\mu, AVA')$$

3. If $X \sim N(0, V)$ & $Z = X'V^{-1}X$

Convergence of Random Vectors

- Let $U_1 \ldots U_K$ be (K×1) vectors drawn independently such that $E[U] = 0$ then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i \xrightarrow{d} N(0, V)$$

↳ where $V = E[UU']$

- if $\hat{A}$ is a conformable matrix

$$\hat{A}\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i \xrightarrow{d} N(0, AVA')$$

Population Regression

- Population Regression is the conditional Expectation of Y given X
- when Evaluated at particular x

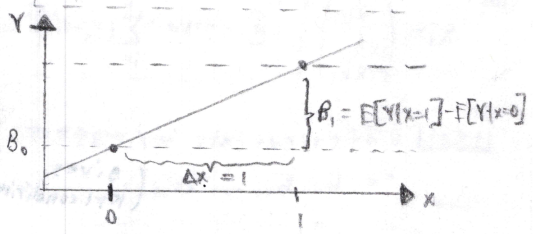
$$E[Y|X=x]$$

Law of Iterated Expectations

$$E(Y) = E\left(E(Y|X)\right)$$

* If X is Binary, population regression is linear

$$E[Y|X] = \beta_0 + \beta_1 X$$



For Any X:

$$\beta_1 = E[Y|X+1] - E[Y|X]$$

- Similarly, For $E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

$$\beta_1 = E\left[Y \mid X_1 = x_1+1, X_2 = x_2\right] - E\left[Y \mid X_1 = x_1, X_2 = x_2\right]$$

Population Regression Function

$X = \begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_K \end{pmatrix}$ & $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$

SO

$$E[Y|X] = X'\beta$$

Definition of Error Term

$$u = Y - X'\beta$$

$$Y = X'\beta + u \qquad E[u|X] = 0$$

↳ Y is out come

↳ X's are explanatory variables aka regressors

Bias solution of Minimization Problem

Let,

$m = E[Y]$ & $m_x = E[Y|X]$

Property of Expectations

$$E[(Y-m)^2] \leq E[(Y-c)^2]$$

So if $E[Y|X] = X'\beta$ then

$$\beta = \underset{b \in R^K}{\operatorname{argmin}}\, E[(Y - X'b)^2]$$

- β is argument that minimizes $E[(Y-X'b)^2]$

Estimation: Least Squares

- Suppose there exists Random Sample $(Y_1, X_1), (Y_2, X_2) \ldots (Y_n, X_n)$

$X_i = \begin{pmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{iK} \end{pmatrix}$ $\Big\}$ $\hat{\beta} = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - X_i'b)^2$

steal Differentiate w/ respect to $b_0, b_1, \ldots b_K$ (gives K+1 conditions)

$$\sum_{i=1}^{n} X_i (Y_i - X_i'\hat{\beta}) = 0$$

$$\Downarrow$$

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i Y_i$$

- Co-efficients of $\hat{\beta}$ are called OLS coefficients

Bivariate Regression

$$E[Y|X] = \beta_0 + \beta_1 X$$

OLS coefficients

$$\hat{\beta}_1 = \frac{\widehat{Cov}(Y_i, X_i)}{\widehat{Var}(X_i)} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Residuals: $\hat{u}_i = Y_i - \hat{Y}_i$

Least squares: Simplified Notation

$y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ $X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1K} \\ 1 & X_{12} & \cdots & X_{2K} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1n} & & X_{Kn} \end{pmatrix}$ $u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$

$$y = X\beta + u \qquad E[u|X] = 0$$

$$\hat{\beta} = (X'X)^{-1} X'y$$

Properties of Least Squares

- Regression Residuals $\hat{u}_i$ have sample mean = 0

$$\frac{1}{n} \sum_{i=1}^{n} \hat{u}_i = 0$$

- Regressors (X_i) & regression residuals are perfectly un-coordinated

Multi-Colinearity

- Refers to case where columns of X are linearly dependent

in this case $X'(y - Xb) = 0$ has many solutions

Un biased $\hat{\beta}$ for β

$$\hat{\beta} = (X'X)^{-1} X'y$$

$$= (X'X)^{-1} X' \underbrace{(X\beta + u)}_{y}$$

$$= \beta + (X'X)^{-1} X'u$$

$$\Downarrow$$

$$E[\hat{\beta}] = \beta$$

$\hat{\beta}$ unbiased for β

Example: log & Gini coeff

Let gdpc = log GDP per cap $\Big\}$ gini = log Gini Index

suppose, $E[\text{gini}|\text{gdpc}] = \beta_0 + \beta_1\, \text{gdpc}$

$Y_i = \text{gini}_i$ $X_i = \begin{pmatrix} 1 \\ \text{gdpc}_i \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i Y_i \Rightarrow \begin{matrix} \hat{\beta}_0 = \\ \hat{\beta}_1 = \end{matrix}$$

R^2 Metric

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Large Sample Distribution

Recall

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i Y_i$$

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i'\right)^{-1} * \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i (Y_i - X_i'\beta)$$

(LLN) ↓ CLT ↓

$Q^{-1} = E[X_i X_i']^{-1}$ $N(0, \Sigma)$

so, $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V)$

↳ w/ $V = Q^{-1} \Sigma Q^{-1}$

LSD: Estimator Form

$$\hat{\beta} \sim N(0, V/n)$$

- we estimate V as $\hat{V} = \hat{Q}^{-1} \hat{\Sigma} \hat{Q}^{-1}$

$\hat{Q} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i'$ $\Big|$ $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i (Y_i - X_i'\hat{\beta})^2 X_i'$

- now we can test hypotheses about β

Tests $\left| \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \right| > 1.96$ where $se(\hat{\beta}_1) = \sqrt{\hat{V}_{22}/n}$

homoskedasticity vs. heteroskedasticity

Homo: $var(Y|X) = \sigma^2$

Hetero: $var(Y|X) = \sigma^2(x)$; depends on x

If homoskedasticity

$\Sigma = \sigma^2 Q$ & $V = \sigma^2 Q^{-1}$

Interpretation of Regression coefficients

① $Y = \alpha_0 + \alpha_1 X + v \qquad E[v|X_i] = 0$

② $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \qquad E[u|X_1, X_2] = 0$

α_1: change in Y w/ unit X_1

β_1: change in Y w/ unit X_1, keeping X_2 constant

Omitted Variable Bias

- want to know β_1 but have no data on X_2

↳ can only estimate α_1. what is

$$\alpha_1 - \beta_1$$

suppose: $X_2 = \gamma_0 + \gamma_1 X_1 + w$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

$$= \underbrace{(\beta_0 + \beta_2 \gamma_0)}_{\alpha_0} + \underbrace{(\beta_1 + \beta_2 \gamma_1)}_{\alpha_1} X_1 + \underbrace{(u + \beta_2 w)}_{v}$$

Omitted Variable Bias Formula

$$\alpha_1 - \beta_1 = \beta_2 \gamma_1$$

if $\beta_2 > 0$ & $\gamma_1 > 0$, effect of X_1 will be overstated

Logarithms

- Consider a log-log specification

$$\log(Y) = \beta_0 + \beta_1 \log(X) + u$$

$$E[\Delta \log Y] = \beta_1 \Delta \log X$$

$$E\left[\frac{\Delta Y}{Y}\right] = \beta_1 \frac{\Delta X}{X}$$

β_1 is expected percent change in Y w/ 1% change in X

- consider Log-Linear Specification

$$\log Y = \beta_0 + \beta_1 X_1 + u$$

$$\Downarrow$$

$$E\left[100 * \frac{\Delta Y}{Y}\right] = 100 * \beta_1 \Delta X$$

$100 * \beta_1$ is expected % change in Y w/ 1 unit increase X

Polynomials

Quadratic Specification

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

$$\hookrightarrow \frac{\partial E[Y|X]}{\partial X} = \beta_1 + 2X\beta_2$$

cubic specification

$$E[Y|X] = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

$$\hookrightarrow \frac{\partial E[Y|X]}{\partial X} = \beta_1 + 2X\beta_2 + 3X^2\beta_3$$

Dummy Variables

- we code educational attainment with Dummy Variables (Binary Variables)

hsdrop = { 1 if no high school; 0 otherwise }

hscool = { 1 if HS is highest; 0 otherwise }

college = { 1 if highest is college; 0 otherwise }

* Problem: Note

hsdrop + hschool + college + moretcol = 1

↳ will result in Multicolinearity unless we exclude one dummy variable from Regression

Dummy Variables: Example

Regression:

$$E[\text{Earnings}|\text{schooling}] = \beta_0 + \beta_1 \text{hschool} + \beta_2 \text{college} + \beta_3 \text{moretcol}$$

- "High School Dropout" is omitted category or Base case

* The Intercept = mean of Base Case

$$\beta_0 = E[\text{Earnings}|\text{hsdrop}=1]$$

- coefficients on the dummy variables are difference in mean of that group & base case

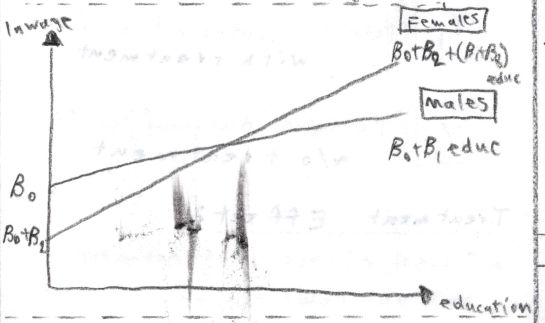
$$\beta_2 = E[\text{Earnings}|\text{college}] - E[\text{Earnings}|\text{hsdrop}]$$

Interactions

- Let Female be a dummy variable for female worker

consider:

$$E[\ln\text{wage}|\text{female}, \text{education}] = \beta_0 + \beta_1(\text{educ}) + \beta_2(\text{female}) + \beta_3(\text{educ} * \text{female})$$



- Here, female * education is the interaction between females & education

Best Linear Predictors

$X'\beta$ is best linear approximation to the regression function $E[Y|X]$

Non-Parametric Regression

Use when:

1. interested in shape
2. we don't want to make functional form assumptions

Kernel Regression

- A Kernel Regression is a weighted average

$$\hat{E}[Y|X=x_0] = \sum_{i=1}^{n} w_i Y_i$$

where,

$$w_i = \frac{K_h(X_i - x_0)}{\sum_{i=1}^{n} K_h(X_i - x_0)}$$

Function K_h is a Kernel

- $K_h(X_i - x_0)$ is Large if $|X_i - x_0|$ small
- $K_h(X_i - x_0)$ is small if $|X_i - x_0|$ large

Kernels

Usual Kernels:

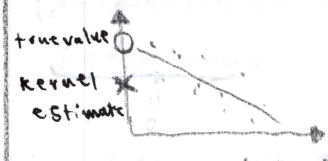
1. Uniform
2. Gaussian
3. Epanechnikov

* h is the bandwidth: Std. Dev of Kernel K_h

- If h is small, only observations close to x_0 get large weights
- If h is large, observations far from x_0 get larger weights

Kernel Regression: Boundary Bias

- Consider x_0 at the boundary of support of x



* all data points far away, so its unreliable

Series Regression

- Fit a Polynomial of order K

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i - b_2 X_i^2 \ldots - b_K X_K)^2$$

Local Linear Regression

- For each $X = x_0$, minimize

$$\sum_{i=1}^{n} K_h(X_i - x_0)(Y_i - b_0 - b_1 X_i)^2$$

Instrumental Variables

- say you collect log of daily prices, log of amount sold

we aim to estimate Demand equation

$$\log q_t = \beta_0 + \beta_1 \log p_t + u_t$$

↳ β_1 = Price elasticity of Demand

Motivation

- Prices & Quantities determined by supply + demand

* Changes in H_t induce change in Price

↳ so $\boxed{cov(\log P_t, H_t) \neq 0}$ { OLS will not work

* we'll consider only supply shocks, to estimate Demand function

Instrumental variable estimation

- we use bad weather as supply shock

$$stormy_t = \begin{cases} 1, \text{if stormy at sea} \\ 0, \text{otherwise} \end{cases}$$

↳ $cov(stormy_t, H_t) = 0$ (1)

↳ $cov(stormy_t, P_t) \neq 0$ (2)

- If (1) & (2) hold, $stormy_t$ is an instrument for $\log P_t$

Instrumental Variable: Procedure

$$cov(stormy_t, \log Q_t) = \beta_1 cov(stormy_t, \log P_t) + \underbrace{cov(stormy_t, H_t)}_{=0}$$

$$\hat{\beta}_1 = \frac{\widehat{cov}(stormy_t, \log Q_t)}{\widehat{cov}(stormy_t, \log P_t)}$$

$$\hat{\beta}_0 = \overline{\log Q_t} - \hat{\beta}_1 \overline{\log P_t}$$

Instrumental variables: General

$$Y_i = \beta_0 + \beta_1 X_i + H_i$$

- X_i is endogenous, when $cov(X_i, H_i) \neq 0$

- If there exists variable $\boxed{Z}$, where

(1) $cov(Z, X) \neq 0$

(2) $cov(Z, H) = 0$

* Z is a valid instrument for endogenous variable X_i

$$\hat{\beta}_1 = \widehat{cov}(Y, Z) / \widehat{cov}(X, Z)$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

2 SLS Interpretation

- Instrumental Variable Estimator has 2-stage least squares interpretation

Purpose, Scope, & Examples

Program Evaluation assesses the causal effects of Public Policy interventions

Effect of:

1. Job training on earnings
2. Class size on test scores
3. Min. wage on employment

Causality w/ Potential outcomes

Treatment

D_i: indicator of treatment intake of unit i

$$D_i = \begin{cases} 1, \text{if unit } i \text{ receives treatment} \\ 0, \text{otherwise} \end{cases}$$

Outcome

Y_i: observed outcome variable of interest for unit i

Potential outcomes

Y_{1i}: Potential outcome for i with treatment

Y_{0i}: Potential outcome for i w/o treatment

Treatment Effect:

- Causal effect of treatment on outcome for unit i

$$Y_{1i} - Y_{0i}$$

Observed Outcomes

- Realized as

$$Y_i = Y_{1i} D_i + Y_{0i}(1 - D_i)$$

Fundamental problem of causal Inference

* cannot observe both potential outcomes

Identification Problem of causal inference

Problem: How to find $Y_{1i} - Y_{0i}$

* homogeneity would solve,

- (Y_{0i}, Y_{1i}) constant over population or time

↳ usually population is of heterogenous nature

Stable Unit Treatment value assumption

Assumptions:

1a. treatment of i doesn't affect j

Quantities of Interest (Estimands)

ATE (average treatment effect)

$$\alpha_{ATE} = E[Y_1 - Y_0]$$

ATET (average treatment effect on treated)

$$\alpha_{ATET} = E[Y_1 - Y_0 | D=1]$$

* Example given on Packet

Selection Bias

$$E[Y|D=1] - E[Y|D=0]$$

$$= \underbrace{E[Y_1 - Y_0 | D=1]}_{ATET} + \underbrace{E[Y_0|D=1] - E[Y_0|D=0]}_{BIAS}$$

Assignment Mechanism

- Procedure that determines which units are selected for treatment intake

1. Random Assignment
2. Selection on Observables
3. Selection on unobservables

* Key ideas *

- causality defined by Potential outcomes

* Observed Association is neither necessary nor sufficient for causation

- Estimation of causal effects starts w/ assignment Mechanism