FALL 2021 EC 282 LECTURE NOTES

09/09/2021

LESSON PLAN

- SELF-INTRODUCTION
- COURSE INFO + SYLLABUS + EXPECTATIONS
- COURSE LOGISTICS: R-STUDIO
  &
  STACKOVERFLOW
  GITHUB

EC282: INTRODUCTION to ECONOMETRICS

ECONOMETRICS
↓
ECON THEORY
↳ MEASUREMENT (QUANTIFY)

→ DATA SCIENCE
STATISTICAL LEARNING
↳ REGRESSION ANALYSIS
MACHINE LEARNING
AI

TWO IMPORTANT CHANGES

- BIG DATA REVOLUTION:

WE BECAME CAPABLE OF
↳ COLLECTING MORE DIGITAL INFORMATION.
STORING
MANAGING
↳ SQL, GRAPH DATA BASE

↳ SERVERS, CLOUDS

↳ SMART PHONES

→ COULD BE USED IN ANY MARKET.

- PREDICTION MODELLING CHANGED: MACHINE LEARNING (SUPERVISED OR UNSUPERVISED)

EX: ZILLOW, REDFIN ESTIMATES (NEVER PERFECT)

- CLASSIC ECONOMETRICS ——→ WHY?

$$X \xrightarrow{\text{CORRELATES}} Y$$

CAUSES

↳ WHY? THEORY

COVID-19 VACCINES

CAUSAL RESEARCH DESIGN

FDA APPROVAL AFTER LARGE SCALE RCT.

THIS COURSE: INTRODUCTION TO BOTH:

- LEARN BASIC TOOLS OF REGRESSION:
- DIFFERENTIATE THE CORRELATION & CAUSATION

- SPREAD SHEETS
- ACCESS, EXCEL

} MIGHT NOT BE VERY POPULAR IN THE FUTURE FOR DATA MANAGEMENT & ANALYSIS

- MORE COMPLEX DATA SYSTEMS : PYTHON, R → INDUSTRY STANDARD
  NOT STATA/SPSS

CLASS LOGISTICS

WEB PAGE: PDF/HTML VERSIONS
OFFICE HOURS: ONLINE
TEXT BOOK: GET THE CHEAPEST VERSION

- P & R. STUDIO (I SHOW IF WE HAVE TIME)
- STACKOVERFLOW
  ECONOMETRICS USING R
  POST ON GITHUB

GITHUB ACCOUNT

PROGRAMMING: ~~BEST~~ OF YOUR ~~EFFORT~~ IS FINE
(MORE ON THIS LATER)

GRADING: HIGH STAKES ASSESSMENTS

2 MIDTERMS + FINAL → 70 %

LOW STAKES ASSESSMENTS

7 HOMEWORK ASSIGNMENTS → SUBMITTED THROUGH BLACK BOARD
→ 30% PASS/FAIL

- PASS/FAIL
- GROUP WORK (MAX 2)
- SUBMIT THROUGH BB (EACH STUDENT)
- DIFFERENT DATASETS

NO EXAMS REQUIRE PROGRAMMING SKILLS.

LD RUNNING CLUB

12/09/2021

LESSON PLAN: REVIEW OF PROBABILITY

- RANDOM VARIABLES
- PROBABILITY DISTRIBUTIONS
- $E[X]$. $VAR[X]$. STD. DEV
  μ_x σ_x^2 σ_x
- STANDARDIZED RANDOM VAR $\rightarrow Z$
- 2 RANDOM VARIABLES X, Y
- JOINT & MARGINAL DISTRIBUTIONS
- INDEPENDENCE
- COVARIANCE & CORRELATION

RANDOM VARIABLES: X, Y, Z

NUMERICAL SUMMARY OF A RANDOM OUTCOME

- HEADS/TAILS ——— PURELY RANDOM

. COVID INFECTION → RANDOM COMPONENT + DETERMINISTIC COMPONENT

YES / NO

L_D $\{0,1\}$ → $Y = \{0,1\}$

↓ LOYES

YOUR GRADE FROM ECONOMETRICS: NO (IMPOSSIBLE) → UNLIKELY

$Y = \{0, 0.7, 1, 1.3, 1.7, 2, 2.3, 2.7, 3, 3.3, 3.7, 4\}$

(UNLIKELY)

DISCRETE RANDOM VARIABLES: ↳ ITS Y COUNTABLE NB OF OUTCOME

↳ WE WILL MOSTLY WORK WITH THESE

CONTINUOUS RANDOM VARIABLES: ANY NUMERICAL VALUE IN AN INTERVAL OR COLLECTION OF INTERVALS.

EXPECTED VALUE: LONG-RUN AVERAGE

$E[Y]$, μ_Y

$Y = \{Y_1, Y_2, \ldots, Y_k\}$ ↓ OUTCOMES

↓

$PR(Y=Y_1), PR(Y=Y_2), \ldots, PR(Y=Y_k)$

↳ DISCRETE PROBABILITIES

$$E[Y] = \sum_i Y_i P_i = Y_1 P_1 + Y_2 P_2 + \dots + Y_k P_k$$

↳ WEIGHTED AVERAGE OF OUTCOMES

↳ WEIGHTS → DISCRETE PROBABILITIES

$$\sum_i P_i = 1$$

HANDOUT QUESTION:

$Y = \{0, 1\}$ → ~~BERNOULLI~~ / ~~SUM~~ / ~~BINARY~~

$PR(Y=1)$ → p

$PR(Y=0)$ → $(1-p)$

$$E[Y] = 0 \times (1-p) + 1 \times p$$

$$= p$$

$Y = \{0, 1\}$ → COVID

$PR(Y=1) = 0.01$

$PR(Y=0) = 0.99$

$$E[Y] = 0 \times 0.99 + 1 \times 0.01$$

$$= 0.01 \rightarrow p$$

VARIANCE: $VAR(Y) \rightarrow \sigma_Y^2$

MEASURES THE WEIGHTED SPREAD OF OUTCOMES AROUND THE LONG RUN AVERAGE (μ_Y)

$$\sigma_Y^2 = (Y_1 - \mu_Y)^2 p_1 + (Y_2 - \mu_Y)^2 \cdot p_2 + \dots + (Y_K - \mu_Y)^2 \cdot p_K$$

$$= \sum_{i=1}^{K} (Y_i - \mu_Y)^2 p_i$$

SAME THING!

$\sigma_Y^2 \geq 0 \qquad \sigma_Y \geq 0 \rightarrow \sigma_Y = \sqrt{\sigma_Y^2}$

IF $Y = \{0,1\}$

$Y_1 = 0 \qquad Y_2 = 1 \qquad \mu_Y = p$

$(1-p) \qquad p$

$$\sigma_Y^2 = (0-p)^2 (1-p) + (1-p)^2 p$$

$$= p^2 (1-p) + (1-p)^2 p$$

$$= p(1-p)[p - (1-p)] = p(1-p)$$

σ_Y AS STANDARD DEVIATION

WE LIKE THIS BETTER (UNIT OF MEASUREMENT IS THE SAME AS Y)

HANDOUT QUESTION:

$\sigma_Y^2 = 0.01 \times 0.99$

$\sigma_Y^2 = 0.0099 \qquad \sigma_Y \cong 0.0995$

TWO RANDOM VARIABLES

X & Y : BOTH DISCRETE RANDOM VARIABLES.

UPPER CASE

$PR(X=x)$
$PR(Y=y)$
} PROB OF X & Y EQUAL TO A SPECIFIC VALUE OF x, y

LOWER CASE

MARGINAL PROBABILITY DISTRIBUTIONS

$PR(X=x, Y=y)$ → PROB. THAT $X=x$ AND $Y=y$

$$PR(Y=y) = \sum_i PR(X=x_i, Y=y)$$

HANDOUT

$Y = \$0.1M$

↓

LONG COMMUTE

vs SHORT COMMUTE

$X = \$0.1$

↓

vs NO RAIN

PAIN

$$PR(Y=0) = PR(X=0, Y=0) + PR(X=1, Y=0)$$

$PR(X=0, Y=0) = 0.15$
$PR(X=0, Y=1) = 0.15$
$PR(X=1, Y=0) = 0.07$
$PR(X=1, Y=1) = 0.63$

} MUTUALLY EXCLUSIVE

1

$$\sum_i \sum_j PR(X=x_i, Y=y_j) = 1$$

$$PR(Y=1) = PR(X=0, Y=1) + PR(X=1, Y=1)$$
$$= 0.15 + 0.63$$
$$= 0.78$$

BAYES THEOREM

$$PR(Y=y \mid X=x) = \frac{PR(Y=y, X=x)}{PR(X=x)}$$

THE PROBABILITY THAT Y IS ~~EQUAL~~ TO A VALUE CONDITIONAL ON X IS EQUAL TO x

$$PR(\text{SHORT COMMUTE} \mid \text{RAIN}) \rightarrow \frac{0.15}{0.30} = 0.50$$
$$Y=0 \qquad X=0$$

$\rightarrow PR(Y=0, X=0)$
$\rightarrow PR(X=0)$

09/16/2021

LESSON PLAN:
- BAYES THEOREM + CONDITIONAL
- COND. EXPECTED VAL. PROBABILITY
- LIE: LAW OF ITERATED EXPECTATIONS
- INDEPENDENCE, COVARIANCE, CORRELATION

$$E[Y] = \sum_i y_i \, PR(Y = y_i)$$

$$E[Y \mid X = x] = \sum_i y_i \, PR(Y = y_i \mid X = x)$$

ROLL A DIE → TWO RANDOM VARIABLES

$Y = \{1, 2, 3, 4, 5, 6\}$ $\quad PR(Y = y_i) = 1/6$

$X = \{0, 1\}$ $\quad PR(X = 0) = 1/2$

↓ ODD

EVEN $\quad PR(X = 1) = 1/2$

$$E[Y] = \sum_i y_i \, P(Y = y_i)$$

$$= 1 \times 1/6 + 2 \times 1/6 + \ldots + 6 \, 1/6 = 3.5$$

$$E[Y \mid X = 1] = \sum_i y_i \, PR(Y = y_i \mid X = 1)$$

$$= 1 \times \frac{PR(Y = 1, X = 1)}{PR(X = 1)} + 2 \times \frac{PR(Y = 2, X = 1)}{PR(X = 1)}$$

$$+ \quad - - + \quad 6 \times \frac{PR(Y=6, X=1)}{PR(X=1)}$$

$$= 1 \times 1/2 + 0 + 3 \times 1/3 + 0 + 5 \, 1/2 + 0$$

$$\hookrightarrow E[Y|X=1] = 3$$

LAW OF ITERATED EXPECTATIONS

$$E[Y] = \sum_i E[Y|X=x_i] \, PR(X=x_i)$$

ROLLING A DIE: $Y = \{0,1,2,3,4,5,6\}$

$X = \{0,1\}$

$$E[Y] = E[Y|X=0] PR(X=0) + E[Y|X=1] PR(X=1)$$

$E[Y|X=1] = 3 \quad PR(X=1) = 1/2$

$E[Y|X=0] = 4 \quad PR(X=0) = 1/2$

$\hookrightarrow 1 \times 0 + 2 \times 1/3 + 3 \times 0 + 4 \times 1/3 + 5 \times 0 + 6 \, 1/3$

$= 4$

$$E[Y] = 3 \times 1/2 + 4 \times 1/3 = 3.5$$

SHORT NOTATION FOR LIE:

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

INDEPENDENCE: $PR(Y=y|X=x) = PR(Y=y)$

↳ KNOWING X CARRIES NO INFO ON THE LIKELIHOOD OF $Y=y$

$$PR(Y=y|X=x) = \frac{PR(Y=y, X=x)}{PR(X=x)}$$

↳ IF INDEPENDENT $PR(Y=y)$

$$PR(Y=y, X=x) = PR(Y=y) \times PR(X=x)$$

ONLY HOLDS IF X & Y ARE INDEPENDENT

COVARIANCE:

$$COV(X,Y) = \sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)]$$

$$= \sum_i \sum_j (X_i - \mu_X)(Y_j - \mu_Y) PR(X=x_i, Y=y_j)$$

USUALLY DIRECTION MATTERS NOT THE SCALE $< 0, =0, > 0$

$$COR(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \qquad -1 \leq COR(X,Y) \leq 1$$

-1 → PERFECT NEGATIVE

1 → PERFECT POSITIVE

HANDOUT QUESTION 2b

$A = \{0,1\}$

$M = \{0,1,2,3,4\}$

$E[M \mid A=0]$

$E[M \mid A=1]$

$E[M]$

USE BAYES RULE TO GET

- $PR(M=m \mid A=0)$
- $PR(M=m \mid A=1)$
- $PR(A=0)$
- $PR(A=1)$

} NEED

$PR(A=0) = PR(M=0,A=0) + PR(M=1,A=0) + \ldots + PR(M=4,A=0)$

$PR(A=1) = PR(M=0,A=1) + PR(M=1,A=1) + \ldots + PR(M=4,A=1)$

$PR(A=0) = 0.5 \qquad PR(A=1) = 0.5$

$$PR(M=0 \mid A=0) = PR(M=0, A=0) \,/\, PR(A=0)$$

$$E[M|A=0] = 0 \times PR(M=0|A=0) + 1 \times PR(M=1|A=0) + \dots + 4 \times PR(M=4|A=0)$$

$$= 0.56$$

$$E[M|A=1] = 0.14$$

$$E[M] = E[M|A=0]\,PR(A=0) + E[M|A=1]\,PR(A>1)$$

$$= 0.56 \times 1/2 + 0.14 \times 1/2$$

$$= 0.35$$

09/20/2021

RANDOM SAMPLING & DISTRIBUTION OF THE SAMPLE AVERAGE

POPULATION: IN MOST CASES DATA DO NOT EXIST.

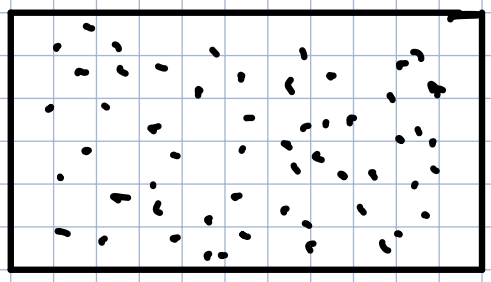
POPULATION PARAMETERS:

$E[Y]$, μ_Y, $VAR(Y)$, σ_Y^2, σ_Y, $COV(X,Y)$

$COR(X,Y)$ ALL UNKNOWN

BUT WE WANT TO LEARN ABOUT THESE PARAMETERS. CONSUMPTION, PRICES, UNEMPLOYMENT, POLITICS, WELLBEING.

RANDOM SAMPLING:



n RANDOM OBS

$\{Y_1, Y_2, \ldots, Y_n\}$

- EACH Y_i IS EQUALLY LIKELY TO BE DRAWN.
- EACH Y_i IS DRAWN FROM THE SAME POP.

iid → IDENTICALLY AND INDEPENDENTLY DISTRIBUTED

② ①

RANDOM SAMPLING INSURES THIS

$$\bar{Y} = \sum_i Y_i / n$$

$\bar{Y}$ → RANDOM VARIABLE

DEF: SIMPLE AVERAGE OF n RANDOMLY DRAWN Y_i

EVERY TIME, A RANDOM DRAWN n WILL GIVE YOU A DIFFERENT $\bar{Y}$

$E[\bar{Y}]$, $\sigma^2_{\bar{Y}}$, PROB. DISTRIBUTION

$$E[\bar{Y}] = E[Y_1 + Y_2 + \dots + Y_n] \,/\, n$$

$$E[\bar{Y}] = 1/n \left(E[Y_1] + E[Y_2] + \dots + E[Y_n] \right)$$

$$= 1/n \left(\mu_Y + \mu_Y + \dots + \mu_Y \right)$$

$$E[\bar{Y}] = \mu_Y$$

$$VAR(\bar{Y}) = VAR\left(1/n \sum_i Y_i \right)$$

$$= VAR\left(1/n \left(Y_1 + Y_2 + \dots + Y_n \right) \right)$$

BECAUSE Y_i, Y_j ARE INDEPENDENT

$$= VAR\left(1/n\, Y_1\right) + VAR\left(1/n\, Y_2\right) + \dots + VAR\left(1/n\, Y_n\right)$$

$$VAR\left(1/n\, Y_i\right) = 1/n^2\, VAR(Y_i)$$

PROOF IS IN THE APPENDIX

$$\sum_i \left(1/n\, Y_i - 1/n\, \mu_Y \right)^2 \rightarrow \text{FOLLOWS THIS LOGIC}$$

$$VAR(\bar{Y}) = 1/n^2\, \sigma_Y^2 \cdot n = \sigma_Y^2 / n = \sigma_{\bar{Y}}^2$$

$$\text{STD.DEV}(\bar{Y}) = \sigma_Y / \sqrt{n} = \sigma_{\bar{Y}}$$

LARGE SAMPLE APPROXIMATIONS:

AS $n \to \infty$

2° $VAR(Y_i) = \sigma_Y^2 < \infty$

1. $Y_1, Y_2, \dots, Y_n$ iid
   ↳ $E[Y_i] = \mu_Y$

LAW OF LARGE NUMBERS: UNDER GENERAL CONDITIONS $\bar{Y}$ IS A GOOD APPROXIMATION FOR μ_Y.

$\bar{Y} \xrightarrow{P} \mu_Y$ AS $n \to \infty$

CENTRAL LIMIT THEOREM: AS $n \to \infty$

$$\bar{Y} \sim N\left[E[\bar{Y}], \sigma_Y^2/n\right]$$

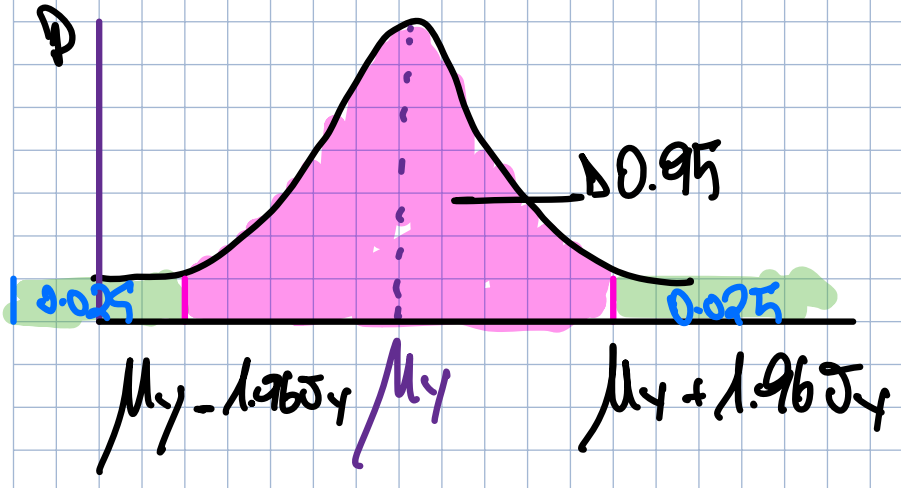
THE INITIAL PROBABILITY DISTRIBUTION OF Y DOES NOT MATTER.

- R SHOWCASE.

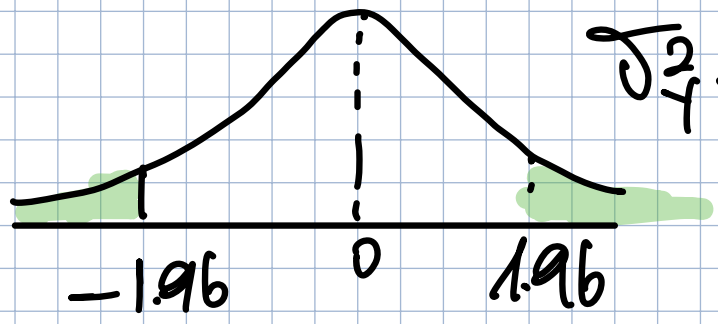
NORMAL DISTRIBUTION: $Y \sim N(\mu_Y, \sigma_Y^2)$

↓

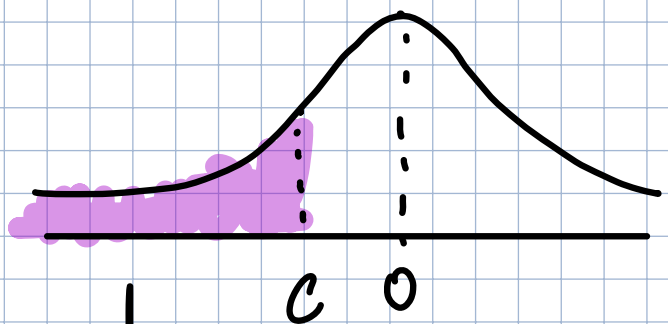
USUALLY BOTH ARE UNKNOWN



$Z \sim N(0,1)$ $\quad E[Z] = 0$



$\sigma_Z^2 = 1 \quad \sigma_Z = 1$



$\Phi(c) = PR(Z \le c)$

If $Y \sim N(\mu_Y, \sigma_Y^2)$

$$\frac{Y - \mu_Y}{\sigma_Y} \sim N(0,1)$$

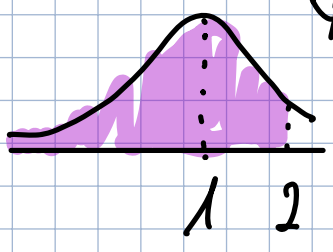
σ_Y^2

So $Y \sim N(1,4)$

μ_Y

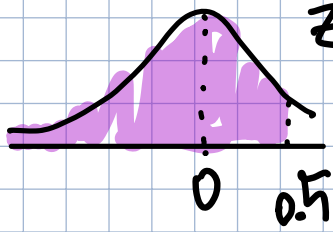
$$\frac{Y-1}{2} \sim N(0,1) \quad \text{OR} \quad Z \sim N(0,1)$$

$$PR(Y \le 2) = PR\left(\frac{Y-1}{2} \le \frac{2-1}{2}\right)$$



$$PR\left(Z \le 1/2\right)$$

$$PR(Z \le 0.5 = 0.691$$

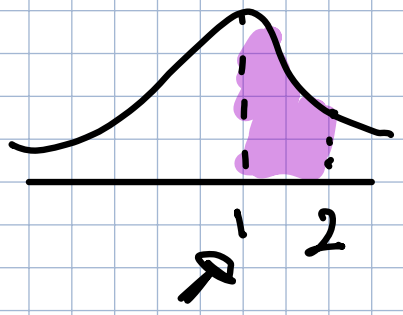


$PR(1 \leq Y \leq 2)$

$PR(Y \leq 2) - PR(Y \leq 1)$

$PR(Z \leq 0.5) - PR(Z \leq 0)$

$0.691 - 0.50 = 0.19$



09/27/2021 REVIEW OF STATISTICS CH 3

ESTIMATOR: FUNCTIONS OF SAMPLE DATA DRAWN FROM AN UNKNOWN POPULATION
- RANDOM VARIABLE (NUMERIC OUTCOME)
- MEAN
- MEDIAN

ESTIMATE: NUMERICAL VALUE OF THE ESTIMATOR

Ex: AVERAGE HOURLY EARNINGS OF COLLEGE GRADS.

↳ Y → RANDOM VARIABLE DEFINED AT THE POPULATION LEVEL

Y — μ_Y → POPULATION AVERAGE IS WHAT WE WOULD LIKE TO KNOW

RANDOM SAMPLE
$\{Y_1, Y_2, \ldots, Y_n\}$

↳ Y_1 TAKE ONLY THE FIRST OBS
$\bar{Y}$ MEAN $\sum_i Y_i / n$ MEDIAN

IN GENERAL AN ESTIMATOR IS DENOTED AS $\hat{\mu}_Y$ ← RANDOM VARIABLE

WHAT IS A GOOD ESTIMATOR?

1. UNBIASED $E[\hat{\mu}_Y] = \mu_Y$

2° CONSISTENCY $\hat{\mu}_Y \xrightarrow{p} \mu_Y$ AS $n \to \infty$ ← CONSISTENCY

3° VARIANCE & EFFICIENCY

$\hat{\mu}_Y$ VS $\tilde{\mu}_Y$ → DIFFERENT ESTIMATORS

$VAR(\hat{\mu}_Y) < VAR(\tilde{\mu}_Y)$

$\hat{\mu}_Y$ IS BETTER

BEST LINEAR UNBIASED ESTIMATOR

$\bar{Y}$ IS BLUE

LINEAR $\Sigma_i Y_i / n$

UNBIASED $E[\bar{Y}] = \mu_Y$

$\bar{Y} \xrightarrow{p} \mu_Y$ AS $n \to \infty$

BEST → SMALLEST VARIANCE

$$VAR(\bar{Y}) = VAR\left(\frac{1}{n}\Sigma_i Y_i\right)$$

$$= \frac{1}{n^2} VAR(\Sigma_i Y_i)$$

$$1/n^2\left[\mathrm{VAR}(Y_1)+\mathrm{VAR}(Y_2)+\ldots+\mathrm{VAR}(Y_n)\right]$$

$$1/n^2 \cdot n\,\mathrm{VAR}(Y_i) = 1/n\,\sigma_Y^2 = \mathrm{VAR}(\bar{Y})$$

$\sum_i (Y_i-m)^2/n$ → VARIANCE TO MINIMIZE BY CHOOSING m

$$\frac{d\sum_i (Y_i-m)^2}{dm}=0 \rightarrow \frac{d\sum_i (Y_i^2-2mY_i+m^2)}{dm}$$

$$\frac{d\sum_i Y_i^2}{dm}-\frac{d\sum_i 2mY_i}{dm}+\frac{dm^2}{dm}=0$$

PROOF:

$$0-2\sum_i Y_i+2m=0$$

$$-2\sum_i (Y_i-m)=0$$

$$\sum_i Y_i = nm \rightarrow m=\sum_i Y_i/n$$

$$m=\bar{Y}$$

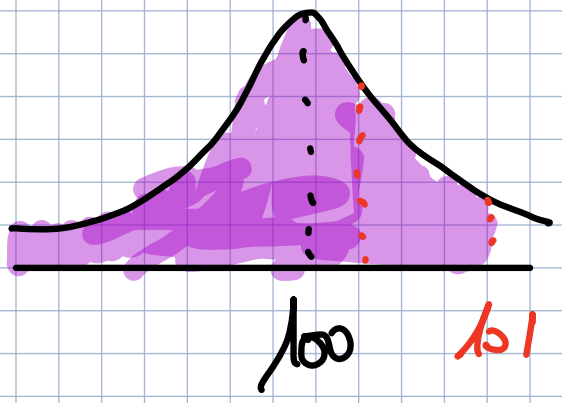
HANDOUT QUESTION 2°

CLT $\bar{Y} \sim N(\mu_Y, \sigma_Y^2/n)$

a. $\mu_Y=100 \quad \sigma_Y^2=43 \quad n=100$

$\bar{Y} \sim N(100, 43/100)$

$\bar{Y} \sim N(100, 0.43)$



$PR(\bar{Y} < 101)$

$\Rightarrow PR\left(\frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}} < \frac{101 - \mu_Y}{\sigma_{\bar{Y}}}\right)$

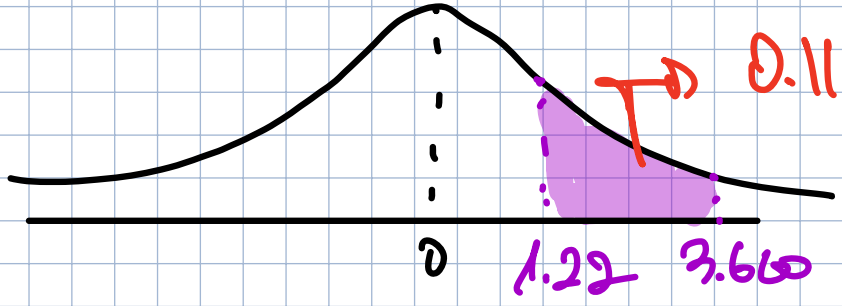
$\Rightarrow PR\left(Z < \frac{101 - 100}{\sqrt{0.43}}\right)$

$\Rightarrow PR(Z < 1.525) = 0.9364$

b. $n = 64 \quad \mu_Y = 100 \quad \sigma^2_Y / n = 0.6719$

$PR(101 < \bar{Y} < 103) = PR\left(\frac{101 - 100}{\sqrt{0.6719}} < Z < \frac{103 - 100}{\sqrt{0.6719}}\right)$

$\Rightarrow PR(1.220 < Z < 3.660)$



c. $n=165$. $\sigma_{\bar{Y}} = \frac{\sigma_Y^2}{n} = 43/165 = 0.266$

$\Pr(\bar{Y} > 98) = 1 - \Pr(\bar{Y} < 98)$

$1 - \Pr\left(Z < \frac{98-100}{\sqrt{0.26}}\right)$

$\sim 100\%$

RANDOM SAMPLING + MEASUREMENT

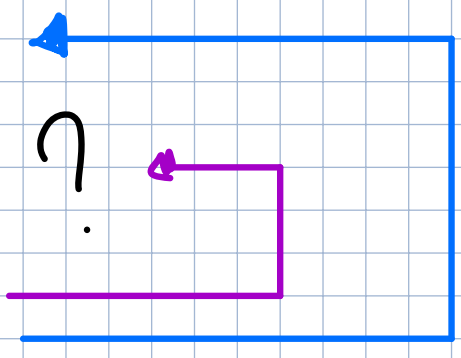
$Y_1, Y_2, \ldots, Y_n \rightarrow$ iid

NON-RANDOM SAMPLING:

UNEMPLOYMENT → SUNDAY ?

CANCER ~~RATE~~ BY AGE

SAMPLE SELECTION BIAS:



HYPOTHESIS TESTING CONCERNING THE POPULATION MEAN

1. SPECIFY THE HYPOTHESIS: WHAT DO YOU WANT TO KNOW ABOUT THE POPULATION?

DO VACCINES WORK?

DO MASKS WORK?

IS THERE A GENDER/RACIAL GLASS CEILING?

1° STATE THE NULL H_0 → TRYING TO REJECT

ALTERNATIVE H_A

$H_0: \mu_Y = \mu_{Y,0}$ → SPECIFIC VALUE

↳ POP. MEAN

$H_A: \mu_Y \neq \mu_{Y,0}$ → TWO-SIDED ALTERNATIVE

$\mu_Y > \mu_{Y,0}$

$\mu_Y < \mu_{Y,0}$

IDEA: TAKE A RANDOM SAMPLE

$\{Y_1, Y_2, \ldots, Y_n\}$

↳ $\sum_i Y_i / n = \bar{Y}$ → SAMPLE MEAN

↳ ACKNOWLEDGE SAMPLING VARIATION

NEVER BE EXACTLY EQUAL

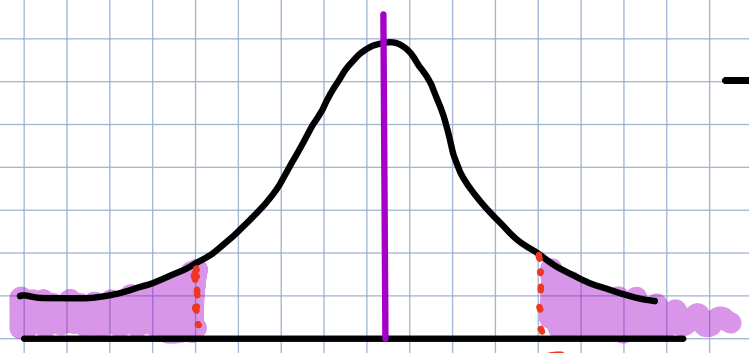
$\bar{Y} = \mu_{Y,0}$

↳ EVERY SAMPLE HAS A DIF. MEAN

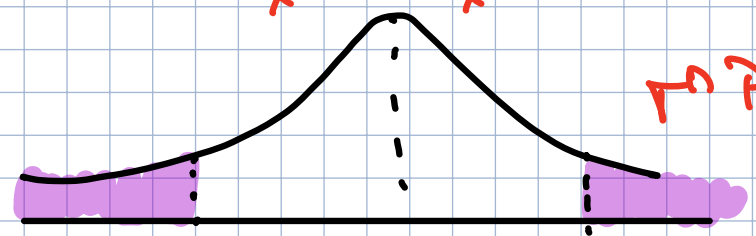
1° ↳ ASSUME H_0 IS TRUE:

$\bar{Y} \sim N(\mu_{Y,0}, \sigma_Y^2/n)$

2° CALC. P-VAL (PROB THAT YOU DRAW A SAMPLE WHOSE MEAN IS AT LEAST AS EXTREME AS OBSERVED $\bar{Y}$



→ IF H_0 IS CORRECT



↳ P-VAL IN 1

↳ TWO-SIDED ALTERNATIVE

$$- \left| \frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}} \right| \quad 0 \quad \left| \frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}} \right|$$

FLIPPING A COIN → μ_Y, σ_Y^2/n

↳ $\sigma_{\bar{Y}}^2$

AVERAGE HOURLY EARNINGS

μ_Y, σ_Y^2 → BOTH UNKNOWN

n → KNOWN

H_0: $\mu_Y = \mu_{Y,0}$

H_A: $\mu_Y \neq \mu_{Y,0}$

} YOU NEED $\bar{Y}$ & σ_Y^2/n

σ_Y^2 UNKNOWN ↩

THE SAMPLE VARIANCE, SAMPLE STANDARD DEV, AND SAMPLE STANDARD ERROR

$$s_Y^2 = \frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2 \quad \xrightarrow{\text{PROXY}} \quad \sigma_Y^2$$

↳ VARIANCE

$$s_Y = \sqrt{s_Y^2}$$ → SAMPLE STD DEV

$$SE[\bar{Y}] = \hat{\sigma}_{\bar{Y}} = s_Y / \sqrt{n} \xrightarrow{\text{PROXY}} \sigma_{\bar{Y}}$$

$$P\text{-VAL} = 2\Phi\left[-\left|\frac{\bar{Y} - \mu_{Y,0}}{SE[\bar{Y}]}\right|\right]$$

$\Updownarrow$

$\sigma_Y / \sqrt{n}$

$$t\text{-stat} = \frac{\bar{Y} - \mu_{Y,0}}{SE[\bar{Y}]} \rightarrow P\text{-VAL} = 2\Phi[-|t\text{-stat}|]$$

IF H_0 IS TRUE $\rightarrow t \overset{a}{\sim} N(0,1)$ WHEN n IS LARGE

1. STATE H_0, H_A
2. CALCULATE TEST STATISTIC → t-stat → p-val
3. COMPARE THE PVAL AGAINST A PRESPECIFIED p THRESHOLD
   ↳ ECON: 1%, 5%, 10%

2 TYPES OF MISTAKES

1. H_0 IS TRUE BUT YOU INCORRECTLY REJECT IT
2. H_0 IS FALSE BUT YOU INCORRECTLY ACCEPT IT OR FAIL TO REJECT IT. → HARD TO JUDGE
   ↳ DEPENDS ON THE DISTRIBUTION UNDER H_A

↳ YOU CAN DETERMINE THE RISK THAT YOU WANT TO TAKE

HANDOUT QUESTION:

a. $Y = \{0,1\}$ $E[Y] = \mu_Y = \sum_i Y_i P_i$

$0 \times 0.5 + 1 \times 0.5 = 0.5 = E[Y]$

b. $VAR(Y) = \sum_i (Y_i - \mu_Y)^2 / 2$

$= \left[(1-0.5)^2 + (0-0.5)^2\right] / 2$

$\sigma_Y^2 = 0.25$ $\quad \sigma_{\bar{Y}}^2 = 0.25 / n$

c. $\sigma_{\bar{Y}}^2 = 0.25 / 100 = 0.0025$

$\sigma_{\bar{Y}} = 0.05$

d. $H_0: E[Y] = 0.5$ → COIN IS FAIR

$H_A: E[Y] \neq 0.5$ → COIN IS BIASED

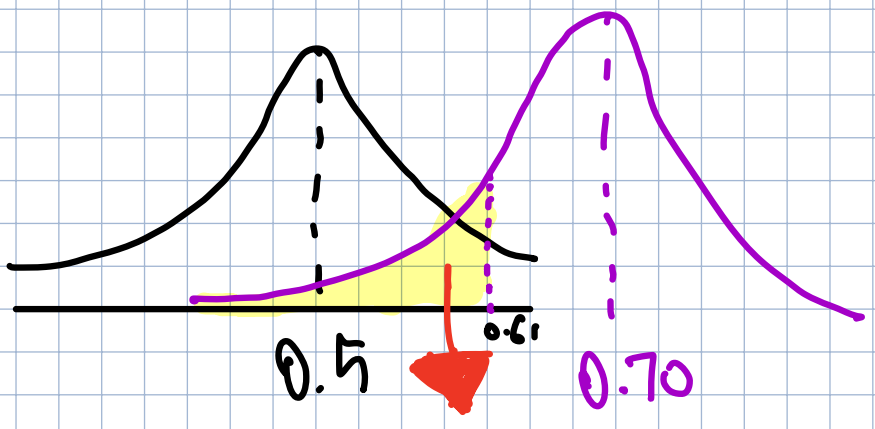
e. $Z = \dfrac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}} = \dfrac{0.61 - 0.5}{0.05} = 2.2$

$$P.\text{JAL} = 2\Phi[-|2\text{-score}|]$$

$$P.\text{JAL} = 0.0278$$

f. 2.78% CHANCE TYPE-I ERROR

g.



↓ TYPE-II ERROR

NULL H_0

ALTERNATIVE

(ONE SIDE ONLY)

TYPE II ERROR

$$2 \times \Phi\left[\left|\frac{0.61 - 70}{\sqrt{0.21/100}}\right|\right] \quad 2 \times \Phi(1.96)$$

$\Rightarrow$ 5%.

$\Rightarrow$ TYPE. II ERROR

$$\text{VAR}(\bar{Y}) = \sigma_Y^2/n$$

$$= 0.21/100$$

$$\bar{Y} = 20.1\%$$

$\Rightarrow 0.3$

$$\text{VAR}(0-0.7)^2 \times 0.3 + (1-0.7)^2 \times 0.7 = 0.21$$

04/11/2021

LESSON PLAN: ~~PETIEN~~ HYPOTHESIS TESTING CONCERNING μ_Y

IF POP VARIANCE IS KNOWN σ_Y^2 →

TEST STATISTIC Z-SCORE $= \dfrac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}$

$\sigma_{\bar{Y}} = \sigma_Y / \sqrt{n}$

LARGE CASE

IF POP VARIANCE IS UNKNOWN → T-STAT $= \dfrac{\bar{Y} - \mu_{Y,0}}{S/\sqrt{n}}$

$S^2 = \dfrac{1}{n-1} \sum_i (Y_i - \bar{Y})^2$

LOSAMPLE VALUES

→ CONFIDENCE INTERVALS
→ COMPARING MEANS FROM TWO DIFFERENT POPULATIONS

} NEW MATERIAL

X, Y $\quad H_0: \mu_X = \mu_Y \longrightarrow \mu_X - \mu_Y = d_0$

$H_A: \mu_X \neq \mu_Y \longrightarrow \mu_X - \mu_Y \neq d_0$

→ SCATTER PLOTS, SAMPLE COVARIANCE, SAMPLE CORRELATION

TESTS ON SAMPLE MEAN

1. STATE H_0, H_A $H_0: \mu_Y = 0.5$ $H_A: \mu_Y \neq 0.5$

2. USE CLT TO PREDICT THE DIST. UNDER H_0 IS TRUE

$\bar{Y} \sim N(\mu_Y, \sigma_Y^2/n) \rightarrow \bar{Y} \sim N(0.5, 0.0025)$

3° CALCULATE THE ACTUAL $\bar{Y} = \sum_i Y_i / n = 0.61$

4. IF VARIANCE σ_Y^2/n IS KNOWN:

$$\hookrightarrow Z = \frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}} = \frac{0.61 - 0.5}{\sqrt{0.0025}} = 2.2$$

5. CALCULATE P-VAL

$2 \times \Phi[-|Z\text{-SCORE}|] = 2.78\%$

6. COMPARE P-VALUE TO SIGNIFICANCE LEVEL

1%, 5%, 10% IF P-VAL < SIGNIFICANCE LEVEL

$\hookrightarrow$ REJECT H_0

4° IF σ_Y^2/n IS UNKNOWN? USE S_Y^2 ← SAMPLE VARIANCE $\hookrightarrow \sigma_Y^2$

$S_Y \hookrightarrow \sigma_Y$

$S_Y/\sqrt{n} \hookrightarrow SE[\bar{Y}] \hookrightarrow \sigma_{\bar{Y}}$

$\bar{Y} = 0.61$

$S_Y^2 = \sum_i (Y_i - \bar{Y})^2 / (n-1)$

$\hookrightarrow S_Y = 0.49$ $SE[\bar{Y}] = 0.0049$

5. $t\text{-stat} = \dfrac{\bar{Y} - \mu_{Y,0}}{SE[\bar{Y}]} = \dfrac{0.61 - 0.5}{0.049} \approx 2.244$

↳ $p\text{-value} = 2 \times \Phi\left[-|t\text{-stat}|\right]$

$\sim 2 \times \Phi(-|t\text{-stat}|)$

IF n IS LARGE

$p\text{-val} = 0.0270$

TYPE II ERROR: H_0 IS FALSE

YOU FAIL TO REJECT

YOU NEED AN ALTERNATIVE THAT IS MORE PRECISE

$H_A: \mu_Y = 0.7 \quad Var(Y) = \sigma_Y^2$

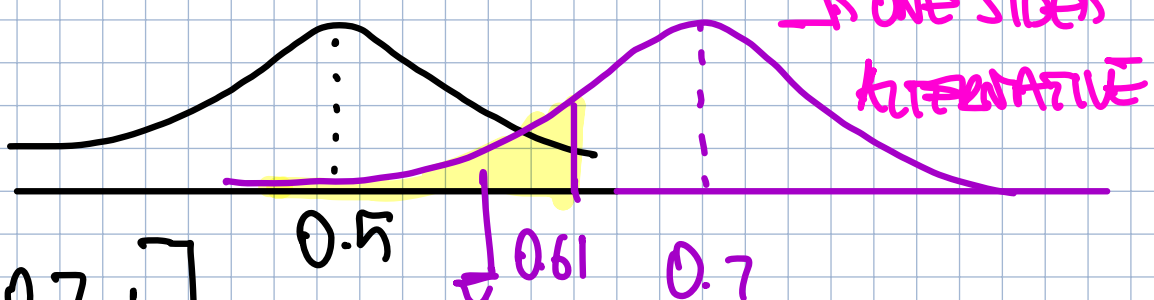
$Var(Y) = \sum_i (Y_i - \mu_Y)^2 \, Pr(Y = y_i)$

$= (0 - 0.7)^2 \times 0.3 + (1 - 0.7)^2 \times 0.70$

$= 0.143 + 0.063 = 0.21$

$\sigma_{\bar{Y}}^2 = \sigma_Y^2 / n = 0.21 / 100 = 0.0021$

TYPE II ERROR:



$\Phi\left[-\left|\dfrac{0.61 - 0.7}{\sqrt{0.0021}}\right|\right]$

$\Phi(-1.96) \sim 2.5\%$

CONFIDENCE INTERVALS:

90%: $\bar{Y} \mp 1.65 \times SE[\bar{Y}]$

95%: $\bar{Y} \mp 1.96 \times SE[\bar{Y}]$

99%: $\bar{Y} \mp 2.576 \times SE[\bar{Y}]$

} LARGER THE INTERVAL GETS MORE CERTAIN THE ESTIMATE IS

$0.6 \mp 1.96 \times 0.049 \rightarrow [0.51\ 0.70]$

INTERPRETATION: 95% CONFIDENT THAT THE TRUE POPULATION MEAN IS WITHIN THE CI

COMPARING MEANS FROM TWO POPULATIONS

μ_W → POPULATION AHE FOR WOMEN

μ_m → POPULATION AHE FOR MEN

$H_0: \mu_m - \mu_W = d_0$ → MOST OFTEN $d_0 = 0$ BUT DOES NOT HAVE TO

$H_A: \mu_m - \mu_W \neq d$

RANDOM SAMPLE OF MEN ⟶ $\bar{Y}_m, n_m$

RANDOM SAMPLE OF WOMEN ⟶ $\bar{Y}_W, n_W$

$\bar{Y}_m - \bar{Y}_W$ → GOOD ESTIMATOR FOR $\mu_m - \mu_W$

$\bar{Y}_M \sim N(\mu_M, \sigma_M^2/n_M)$

$\bar{Y}_W \sim N(\mu_W, \sigma_W^2/n_W)$

↳ $\bar{Y}_M - \bar{Y}_W \sim N\left(\mu_M - \mu_W, \frac{\sigma_M^2}{n_M} + \frac{\sigma_W^2}{n_W}\right)$

σ_M^2, σ_W^2 ↳ TYPICALLY UNKNOWN

YOU ~~NEED TO~~ ESTIMATE s_W^2, s_M^2 ↳ SAMPLE VARIANCE

$$SE[\bar{Y}_M - \bar{Y}_W] = \sqrt{\frac{s_M^2}{n_M} + \frac{s_W^2}{n_W}}$$

IF THE H_0 IS TRUE & n IS LARGE

$$t\text{-stat} = \frac{(\bar{Y}_M - \bar{Y}_W) - d_0}{SE[\bar{Y}_M - \bar{Y}_W]} \sim N(0,1)$$

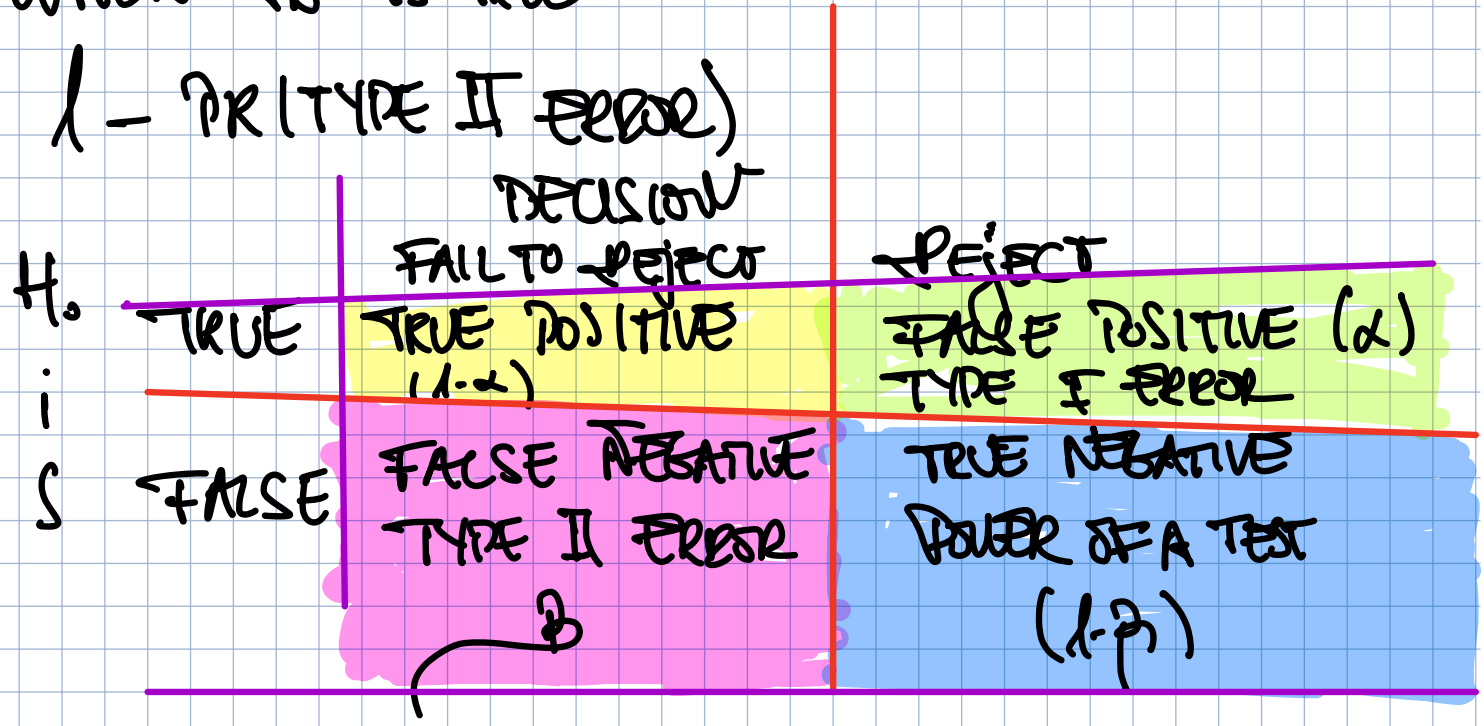
P-VALUE $= 2\Phi(-|t\text{-stat}|)$ / $2t(-|t\text{-stat}|)$

↳ SKIP CASE WHEN $n < 30$

TERMINOLOGY:

- TYPE I ERROR: H_0 IS CORRECT BUT YOU FALSELY REJECT IT
- TYPE II ERROR: H_0 IS INCORRECT BUT YOU FAIL TO REJECT
- SIGNIFICANCE LEVEL: PRESPECIFICIED PROBABILITY OF TYPE-I ERROR
- THE CRITICAL VALUE: VALUE OF THE TEST STATISTIC FOR WHICH THE TEST REJECTS THE NULL
- REJECTION ZONE: AREA UNDER WHICH THE RESEARCHER REJECTS H_0
- SIZE OF THE TEST: PROB. OF INCORRECTLY REJECTING H_0
- POWER OF THE TEST: PROB. OF CORRECTLY REJECTING H_0 WHEN H_0 IS TRUE

$1 - P(\text{TYPE II ERROR})$



10/07/2021

CONFIDENCE INTERVALS: $(\bar{Y}_M - \bar{Y}_W) \pm 1.96 SE[\bar{Y}_M - \bar{Y}_W]$

a. $Y = \{0, 1\}$ ← LOINCUMBENT

$P(Y=0) = P$

$P(Y=1) = 1-P$

EMPLACEMENT

$$\bar{Y} = 215/400 = 0.53375$$

b. $$S_Y^2 = \sum_i (Y_i - \bar{Y})^2 / n-1$$

$$S_Y^2 = \frac{1}{399}\left[(1-0.53375)^2 \times 215 + (0-0.53375)^2 \times 185\right]$$

$$= S_Y^2 \cong 0.249$$

$$S_Y \cong 0.499$$

$$SE[\bar{Y}] = S_Y/\sqrt{n}$$

$$\cong 0.0249$$

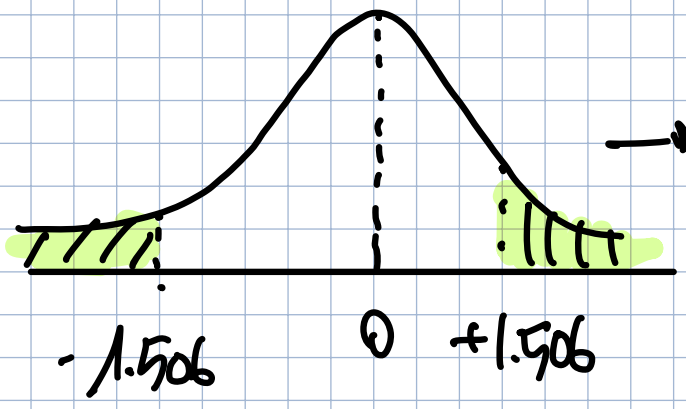
c. $H_0: P = 0.5$

$H_A: P \neq 0.5$

$$-1.564 = \frac{\bar{Y} - \mu_{Y,0}}{SE[\bar{Y}]}$$

$$= \frac{0.53375 - 0.5}{0.0249}$$

$$= 1.506$$

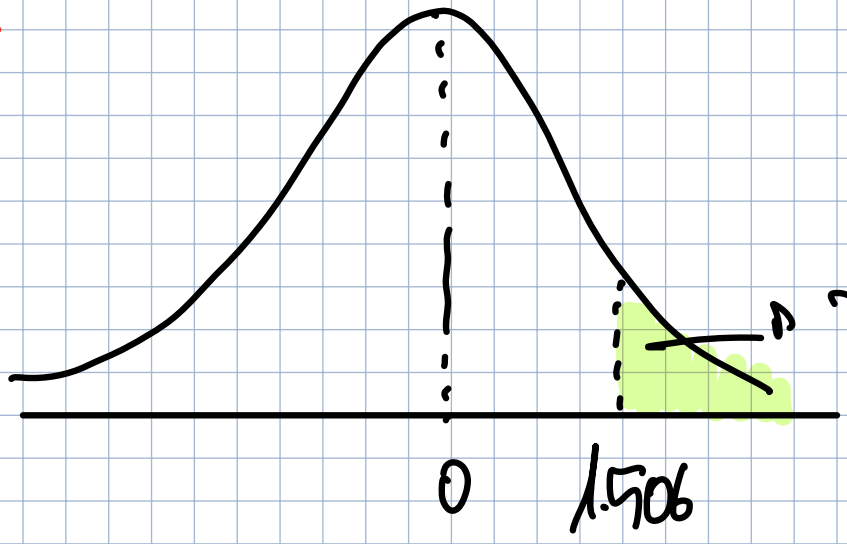


$\rightarrow 2 \times \Phi(-|t\text{-stat}|)$

$2 \times \Phi(-1.506)$

$= 13.2\%$

d.



f. p-val > 0.05 FAIL TO ~~reject~~ THE NULL

SURVEY DOES NOT CONTAIN EVIDENCE

g. 95% CI $\bar{Y} \mp 1.96 \times SE[\bar{Y}]$

$[0.481 \; - \; 0.579]$

95% OF CHANCE THAT TRUE μ_Y IS WITHIN THE 95% CI

↳ MORE CONSERVATIVE

h. $\bar{Y} \mp 2.576 \times SE[\bar{Y}]$ $[0.463 - 0.597]$

Question 2. $n = 420$ $\bar{Y} = 646.2$ $s_Y = 19.5$

a.

$SE[\bar{Y}] = s_Y / \sqrt{n} \cong 0.95$

$\bar{Y} \pm 1.96 \times SE[\bar{Y}]$ $[644.34 - 648.06]$

b. $\bar{Y}_1 - \bar{Y}_2 = 657.4 - 650$

$= 7.4$

$$SE[\bar{Y}_1 - \bar{Y}_2] = \sqrt{\frac{s_{Y_1}^2}{n_1} + \frac{s_Y^2}{n_2}}$$

$$= \sqrt{\frac{19.4}{238} + \frac{17.9}{182}}$$

$= 1.8281$

$[7.4 \mp 1.96 \times 1.828] = [3.82 \quad 10.98]$

H_0 : DIF $= 0$

H_A : DIF > 0

$$t - 1.814 = \frac{7.4 - 0}{1.8281}$$

$= 4.0479$

Reject the null

Districts with smaller classes have better outcomes

P QUESTION : $\bar{Y}_1 - \bar{Y}_2 = -253.2284$

$\bar{Y}_1 = 3178.832$ $\quad S_{Y_1} = 580.0068$

$\bar{Y}_2 = 3432.06$ $\quad S_{Y_2} = 584.622$

$$SE[\bar{Y}_1 - \bar{Y}_2] = \sqrt{\frac{S^2_{Y_1}}{n_1} + \frac{S^2_{Y_2}}{n_2}}$$

$$= 28.82106$$

$$t.stat = -253.2284 \,/\, 28.82106$$

$$= -9.44$$

$\rightarrow$ P-VAL $= 0$

LINEAR REGRESSION

SCATTER PLOT : $X_i : Y_i$ $\cap$ OBS PLOTTING DATA

$$S_{XY} = \frac{1}{n-1}\sum_i (x_i - \bar{x})(Y_i - \bar{Y}) \qquad S_X^2 = \sum_i \frac{(x_i - \bar{x})^2}{(n-1)}$$

$\rightarrow$ SAMPLE COVARIANCE $\quad$ SAMPLE VARIANCE $\leftarrow$

$$\rho_{XY} = \frac{S_{X,Y}}{S_X S_Y}$$ $\rightarrow$ SAMPLE CORRELATION

$-1 \leq \rho_{XY} \leq +1$ $\rightarrow$ TELLS YOU HOW MUCH X,Y ~~ARE~~ ~~RELATED~~

ρ_{XY} → MEASURE OF LINEAR ASSOCIATION
(DO NOT IMPLY CAUSAL INFERENCE)

↳ SHOW STRONG ASSOCIATION BTW X & Y

REGRESSION NOTATION:

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

→ POPULATION REGRESSION FUNCTION

Y → OUTCOME VARIABLE
RESPONSE VARIABLE
LEFT HANDSIDE VARIABLE
DEPENDENT VARIABLE

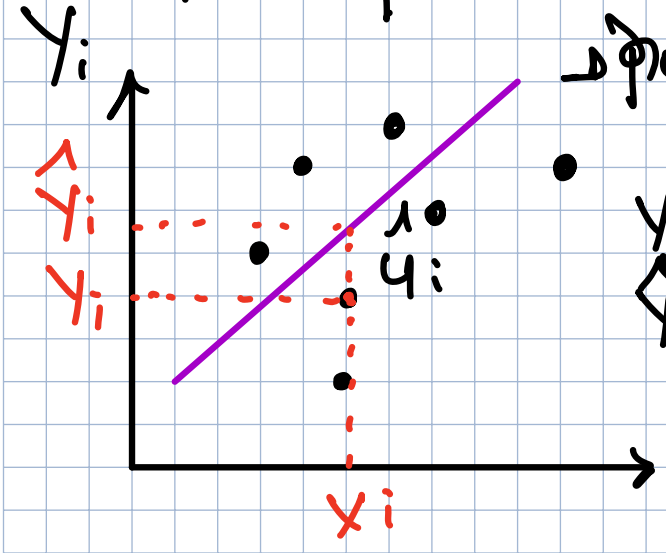
X → INDEPENDENT VAR.
RIGHT HANDSIDE
REGRESSOR

β_0, β_1 → POPULATION COEFFICIENTS / PARAMETERS

$i = \{1, 2, \ldots, n\}$

$\beta_0 + \beta_1 X_i$ → POPULATION REGRESSION LINE
↳ SLOPE
↳ INTERCEPT

u_i → ERROR TERM



$\beta_0 + \beta_1 X_i$

ID	X_i	Y_i
1	5	7
2	6	10
3	9	12
⋮	⋮	⋮

Y_i → OBSERVED
$\hat{Y}_i$ → PREDICTED

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

TWO MAIN PROBLEMS:

1. WE DON'T OBSERVE THE POPULATION
   - BUT CAN USE RANDOM SAMPLES
2. WHICH LINE TO FIT? → FIRST FOCUS ON THIS

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$ → LEFTOVER

- Y_i → EARNINGS
- X_i → EDUCATION, PARENTAL INPUT, SOCIAL NETWORKS

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

ID	X	Y	$\hat{Y}$	u
1	X_1	Y_1	$\hat{Y}_1$	u_1
2	X_2	Y_2	$\hat{Y}_2$	u_2
⋮	⋮	⋮	⋮	⋮
n	X_n	Y_n	$\hat{Y}_n$	u_n

POPULATION

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

SAMPLE

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + u_i$$

- u_i → RESIDUAL
- u_i → ERROR TERM

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$ → PREDICTED VALUE

$$Y_i = \hat{Y}_i + u_i$$ → RESIDUAL

- $\hat{Y}_i$ → PREDICTED VALUE
- Y_i → OBSERVED VALUE

PROBLEM: Choose $\hat{\beta}_0, \hat{\beta}_1 \rightarrow \hat{\beta}_0 + \hat{\beta}_1 x_i \rightarrow \hat{y}_i$

HOW?

MINIMUM PREDICTION ERROR = MIN $\sum_i \hat{u}_i^2$

$\sum_i \hat{u}_i^2 = \hat{u}_1^2 + \hat{u}_2^2 + \dots + \hat{u}_n^2$

SUM OF THE SQUARED RESIDUALS

$$\text{MIN} \sum_i \hat{u}_i = \text{MIN} \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$

$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

$\hat{u}_i^2 = (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

$\sum_i \hat{u}_i^2 = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}$$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

↳ CAN BE DIRECTLY ESTI ⇒ FROM THE SAMPLE

10/20/2021 LECTURE PLAN

- REVIEW OF OLS → ORDINARY LEAST SQUARES
- MEASURES FOR GOODNESS OF FIT
- ONE BIG EXAMPLE

$Y_i = \beta_0 + \beta_1 X_i + u_i$ → POPULATION REGRESSION FUNCTION ← PREDICTION ERROR

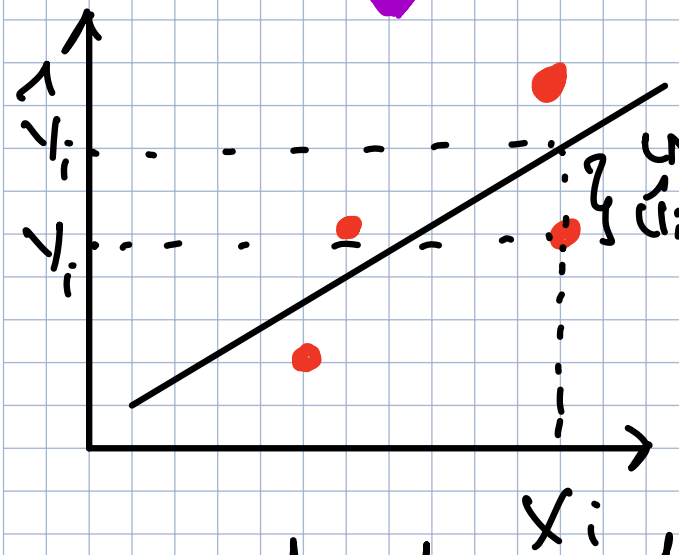
PREDICTS A LINEAR & IMPERFECT RLT B/W X & Y

β_0 → INTERCEPT
β_1 → SLOPE
} POPULATION PARAMETERS

Y IS A LINEAR FUNCTION OF X

} β_0, β_1 ARE UNKNOWN

↓ DRAW A RANDOM SAMPLE:



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

$i = 1, \ldots, n$

↳ RANDOM SAMPLE OF n OBS.

$\hat{Y}_i = Y_i + \hat{u}_i$ → OBSERVED
↳ UNOBSERVED

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
↳ PREDICTED VALUE

$\sum_i \hat{u}_i^2 \rightarrow SSR \rightarrow \min$

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}$$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ → EVERYTHING IS OBSERVED

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

WHAT DO THESE ESTIMATED PARAMETERS $\hat{\beta}_0, \hat{\beta}_1$ TELL US?

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i \quad 1°$$

$$y_i + \Delta y = \hat{\beta}_0 + \hat{\beta}_1 (x_i + \Delta x) + \hat{u}_i \quad 2°$$

$$\Delta y = \hat{\beta}_1 \Delta x \rightarrow \hat{\beta}_1 = \Delta y / \Delta x$$

$\hat{\beta}_1 = \Delta y$ WHEN $\Delta x = 1$

CHANGE IN Y ASSOCIATED WITH ONE UNIT INCREASE IN X

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

$\hat{\beta}_0 = \bar{Y}$ WHEN $\bar{X} = 0$ WHY?

$$\bar{X} = \sum_i X_i / n = 0 \qquad \sum_i X_i = 0$$

$\hat{\beta}_0$ → SOMETIMES IT MAKES SENSE
SOMETIMES IT DOES NOT

HARDER ANSWERS

Y → YOU CAN PREDICT Y FOR ANY VALUE

1° $\hat{\beta}_0$ HAS NO MEANING

2° $\hat{\beta}_1$ → AN INCREASE IN THE SAT BY ONE UNIT IS ASSOCIATED WITH A 2.83 POINTS

IF $X_i = 24$ $\qquad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

$= 625.35 - 2.83.24$

$= 627.38$

IS THE ASSOCIATION LARGE OR SMALL?

SUMMARY THE OUTCOME: $\bar{Y} = 628.8$

$s_Y = 23.8$

$2.83 / 695.3 = 0.4\%$

$2.83 / 23.748 \cong 0.12$ SD

MEASURES OF FIT — R^2 — STD. ERROR OF THE REGRESSION

R^2: $0 \leq R^2 \leq 1$

US SHARE OF VARIATION IN Y EXPLAINED BY THE VARIATION IN X.

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$

$$Y_i = \hat{Y}_i + \hat{u}_i$$

USTSS: TOTAL SUM OF SQUARES

$TSS = \sum_i (Y_i - \bar{Y})^2$ → TOTAL VARIATION IN Y

RSS : RESIDUAL SUM OF SQUARES

$$RSS = \sum_i \hat{u}_i^2$$

ESS : EXPLAINED SUM OF SQUARES

$$ESS = TSS - RSS$$

$$ESS = \sum_i (\hat{y}_i - \bar{y})^2 \rightarrow \text{TOTAL VARIATION IN } \hat{y}$$

$$R^2 = 0 \leq R^2 \leq 1$$

$$R^2 = ESS/TSS = 1 - RSS/TSS$$

FUN FACTS ABOUT OLS

1° SAMPLE AVERAGE OF THE OLS RESIDUALS = 0

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

LD $\quad \hat{u}_i = y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i$

$$\hat{u}_i = (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})$$

$$\sum_i \hat{u}_i = \sum_i (y_i - \bar{y}) - \hat{\beta}_1 \sum_i (x_i - \bar{x})$$

$$= 0 \quad \text{LDO} \qquad \text{LD}$$

$$\sum_i (y_i - \bar{y}) = \sum_i y_i - \sum_i \bar{y}$$

$$= n\bar{y} - n\bar{y} = 0$$

$$\sum_i y_i = y_1 + y_2 + \dots + y_n \quad \times n$$

$$\frac{}{n}$$

$$\sum_i y_i = \bar{y} \cdot n \qquad \text{IF } \sum_i \hat{u}_i = 0$$

$$1/n \sum_i \hat{u}_i = 0$$

2° $1/n \sum_i \hat{y}_i = \bar{y}$

$y_i = \hat{y}_i + \hat{u}_i \rightarrow \sum_i y_i = \sum_i \hat{y}_i + \sum_i \hat{u}_i$

$\sum_i y_i = \sum_i \hat{y}_i + 0$

$N \cdot \bar{y} = \sum_i \hat{y}_i$

$\sum_i \hat{y}_i / N = \bar{y}$ ■

3° → OMITTED

$\sum_i \hat{u}_i x_i = \sum_i \hat{u}_i (x_i - \bar{x}) = 0$

$\sum_i \hat{u}_i x_i = \sum_i \left[(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right] (x_i - \bar{x})$

$\Rightarrow \sum_i (y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}_1 \sum_i (x_i - \bar{x})^2 = 0$

Remember $\hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{(x_i - \bar{x})^2}$

4. $TSS = \sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$

$= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$

$\Rightarrow SSR + ESS + 2\sum_i \hat{u}_i \hat{y}_i$

$\sum_i \hat{u}_i \hat{y}_i = \sum_i \hat{u}_i (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

$= \hat{\beta}_0 \underbrace{\sum_i \hat{u}_i}_{=0} + \hat{\beta}_1 \underbrace{\sum_i \hat{u}_i x_i}_{=0}$

$= 0$

$\hat{\beta}_0, \hat{\beta}_1$ → FIND THESE EQN THAT BEST FIT THE DATA!

↳ WHAT DOES THIS MEAN?

CONTINUUM SUM OF THE SQUARED MISTAKES WHEN YOU USE X TO PREDICT Y.

ANSWER TO THE QUESTION

1. $\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\hat{\beta}_1 = -150.78 / 63.83$

$= -2.83$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\hat{\beta}_0 = 628.8 - (-2.83 \times 23.5)$

$= 628.8 + 66.55$

$= 695.35$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n$

$\hat{u}_i = y_i - \hat{y}_i$

$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ OLS $\sum_i \hat{u}_i^2$

↳ POP. PARAMETERS

$\hat{\beta}_0, \hat{\beta}_1 \rightarrow \beta_0, \beta_1$ $\quad \hat{u}_i \rightarrow u_i$

SAMPLE PARAMETERS RESIDUALS ↳ DEREK TAM

$\widehat{\text{TEST SCORE}} = 695.3 - 2.83 \times \text{STR}$

$\widehat{\text{TEST SCORE}}$ → WHAT MEANS PREDICTED

STR → STUDENT TO TEACHER RATIO.

1. YOU CAN DO PREDICTION BY PLUGGING ANY STR ASSUMING THAT THE RELATIONSHIP IS LINEAR.

2° USE COEFFICIENTS TO INTERPRET THE RELATIONSHIP.

$\hat{\beta}_0$ HAS NO MEANING

$\hat{\beta}_1$ → AN INCREASE IN THE STUDENT TO TEACHER RATIO BY 1 UNIT IS ASSOCIATED WITH A 2.83 POINTS

IF $X_i = 24$ $\quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

$= 695.3 - 2.83 \times 24$

$= 627.38$

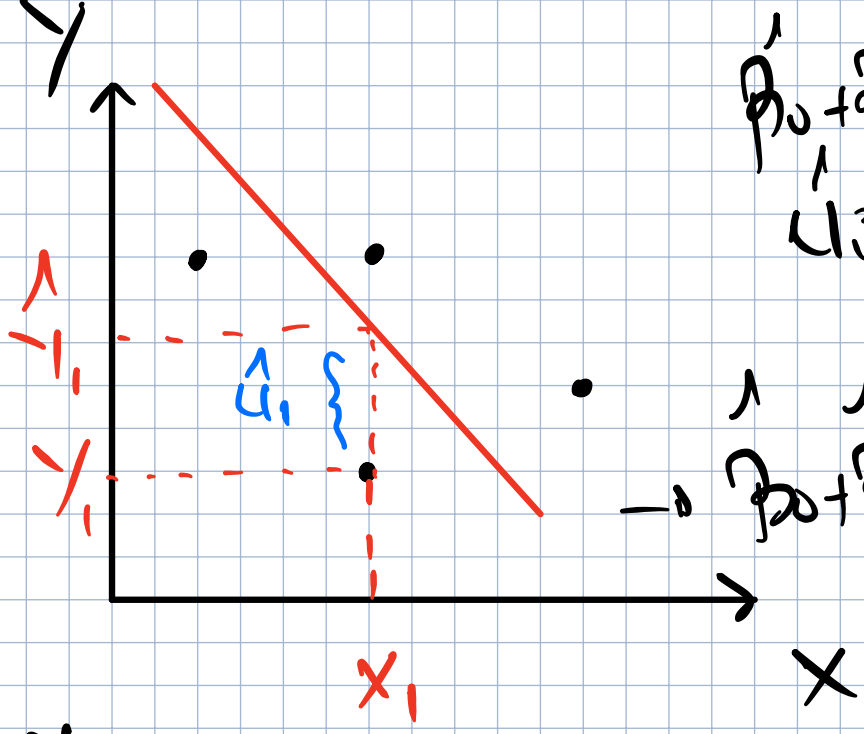
IS THE ESTIMATE SMALL / LARGE?

2 STUDENTS → ALMOST 1 SD A_1

$2 \times 2.83 = 5.66$

A BETTER WAY TO STAND.

$\bar{Y} = 654 \rightarrow S_Y = 19.05$ ← TEST SCORES

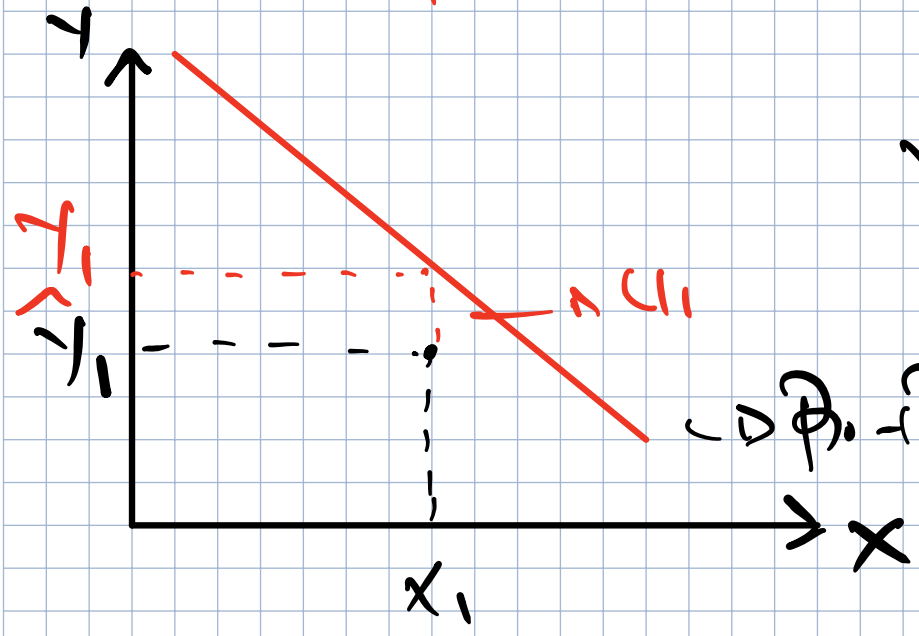


$\hat{\beta}_0 + \hat{\beta}_1 X_i = \hat{Y}_i$

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

$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$

$\rightarrow \hat{\beta}_0 + \hat{\beta}_1 X_i$



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

MEASURES OF FIT : R^2 - STANDARD ERROR OF THE REGRESSION

R^2 : SHARE OF THE VARIATION IN R EXPLAINED BY THE VARIATION IN X.

$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$

$Y_i = \hat{Y}_i + \hat{u}_i$

EXPLAINED SUM OF SQUARES (ESS)

↳ SUM OF THE SQUARED DEVIATIONS OF THE $\hat{Y}_i$ FROM THE AVERAGE

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$ → TOTAL VARIATION IN $\hat{Y}$

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$ → TOTAL VARIATION IN Y

$R^2 = ESS / TSS$ → TOTAL SUM OF SQUARES

SSR → SUM OF SQUARED RESIDUALS

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2 \qquad R^2 = 1 - \frac{SSR}{TSS}$$

X CAN EXPLAIN NO VARIATION IN Y ← $0 \leq R^2 \leq 1$ → VARIATION IN X CAN FULLY EXPLAIN THE VAR IN Y

SOME CUT FACTS ABOUT OLS

1. $1/n \sum_i \hat{u}_i = 0$
2. $1/n \sum_i \hat{Y}_i = \bar{Y}$
3. $\sum_i \hat{u}_i x_i = 0$
4. $\sum_i \hat{u}_i^2$ MINIMUM
5. $TSS = ESS + SSR$

1. SAMPLE AVERAGE OF THE OLS RESIDUALS IS ZERO.

$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

$\Rightarrow \hat{u}_i = (Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X}) \qquad \Rightarrow \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$

$\sum_i \hat{u}_i = \sum_i (Y_i - \bar{Y}) - \hat{\beta}_1 \sum_i (X_i - \bar{X})$

$\sum_i (Y_i - \bar{Y}) = 0 \qquad \hat{\beta}_1 \sum_i (X_i - \bar{X}) = 0$

$\sum_i \hat{u}_i = 0$

$1/N \sum_i \hat{u}_i = 0$

2. $Y_i = \hat{Y}_i + \hat{u}_i \Rightarrow \sum_i Y_i = \sum_i \hat{Y}_i + \sum_i \hat{u}_i$

$\sum_i Y_i = \sum_i \hat{Y}_i + 0$

$N \cdot \bar{Y} = \sum_i \hat{Y}_i$

$\sum_i \hat{Y}_i / N = \bar{Y}$

→ OMITTED

3. $\sum_i \hat{u}_i X_i = \sum_i \hat{u}_i (X_i - \bar{X})$

$\sum_i \hat{u}_i X_i = \sum_i [(Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X})] (X_i - \bar{X})$

$\Rightarrow \sum_i (Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1 \sum_i (X_i - \bar{X})^2 = 0$

Remember $\hat{\beta}_1 = \dfrac{\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{(X_i - \bar{X})^2}$

4° TSS = $\sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$

$= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 + 2\sum_i (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$

$\rightarrow SSR + ESS + 2\sum_i \hat{u}_i \hat{Y}_i$

$\sum_i \hat{u}_i \hat{Y}_i = \sum_i \hat{u}_i (\hat{\beta}_0 + \hat{\beta}_1 X_i)$

$= \hat{\beta}_0 \sum_i \hat{u}_i + \hat{\beta}_1 \sum_i \hat{u}_i X_i$

$\rightarrow 0 \qquad \rightarrow 0$

$= 0$

STANDARD ERROR OF THE REGRESSION

→ MEASURE OF THE SPREAD OF THE OBSERVATIONS AROUND THE REGRESSION LINE.

$SER = S_{\hat{u}} \qquad S_{\hat{u}}^2 = \frac{1}{n-2} \sum_i \hat{u}_i^2 = \frac{SSR}{n-2}$

$SER = \sqrt{\frac{SSR}{n-2}}$ → STANDARD DEVIATION OF THE RESIDUALS

→ $\hat{\beta}_0, \hat{\beta}_1$ SHOULD BE ESTIMATED BEFORE YOU CAN COMPUTE $S_{\hat{u}}$

→ MAGNITUDE OF A TYPICAL REGRESSION ERROR EXPRESSED IN THE UNIT OF THE OUTCOME

THE LEAST SQUARES ASSUMPTIONS

$\hat{\beta}_0, \hat{\beta}_1$ ARE "GOOD" ESTIMATORS FOR β_0, β_1 WHEN

THE CONDITIONAL DISTRIBUTION OF u_i GIVEN X_i HAS A MEAN OF ZERO ← CAUSAL INTERPRETATION OF X

$$E[u_i | X_i] = 0 \iff \text{Corr}(u_i, X_i) = 0$$

↑ 1 STR — ↓ 2.53 POINTS — AUTOMATICALLY SATISFIED WHEN X_i IS RANDOM

→ CONDITIONAL MEAN ASSUMPTION

→ HIGHER STR = 16 } DIFFERENT DISTRICTS
→ LOWER STR = 17 }

→ BUT $\hat{\beta}_0, \hat{\beta}_1$ ARE BIASED

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

+ $\text{Corr}(u_i, X_i) \neq 0$

→ INCLUDES ALL THE OTHER STUFF THAT DETERMINES THE OUTCOME

$\text{Corr}(\text{POVERTY}, \text{STR}) > 0$

MORE POVERTY → LESS

SMALL — LARGE → -1 POINT (IF THEY WERE EQUALLY LIKELY FOR LARGE & SMALL DISTRICTS)

SMALL — LARGE → -2.53 POINTS BECAUSE (DISTRICTS POOR)

↳ CLASSROOM 5

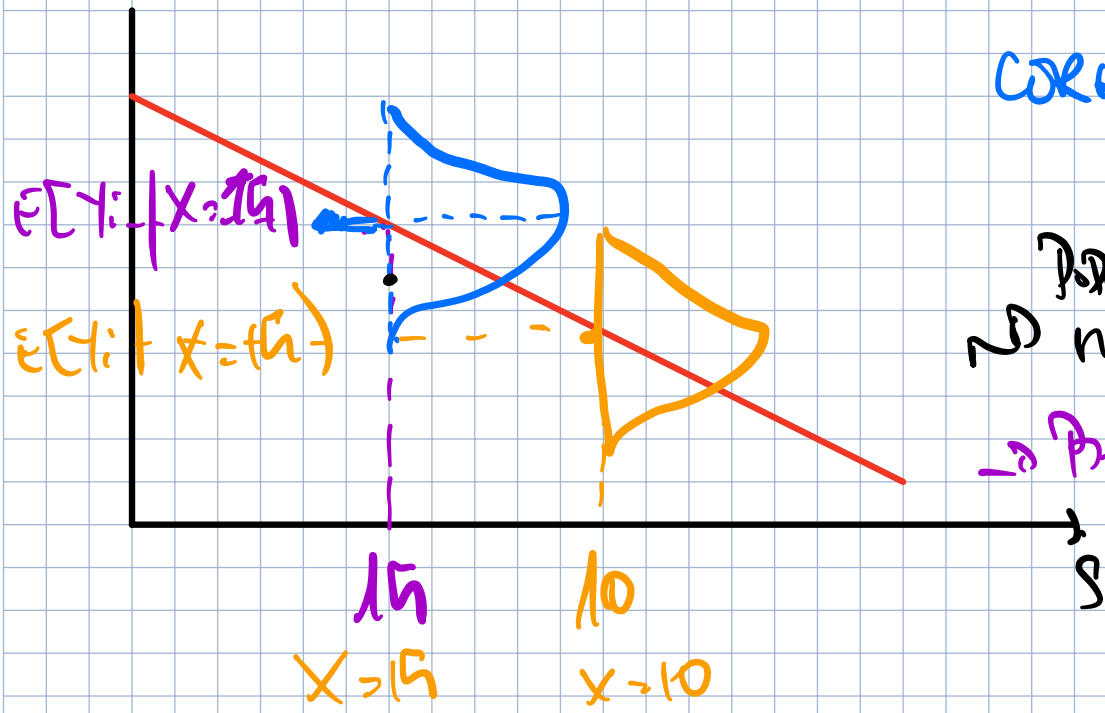
$Y_i = \beta_0 + \beta_1 X + u_i$

$E[Y_i] = \beta_0 + E(X_i) + E[u_i]$

$E[Y_i|X_i] = \beta_0 + \beta_1 X_i + E(u_i|X_i)$

↳ 0

TEST SCORE



$\text{CORR}(u_i, X_i) = 0$

DIRECT RELATIONSHIP
NO MORE OF AVERAGE

$\rightarrow \beta_0 + \beta_1 X$

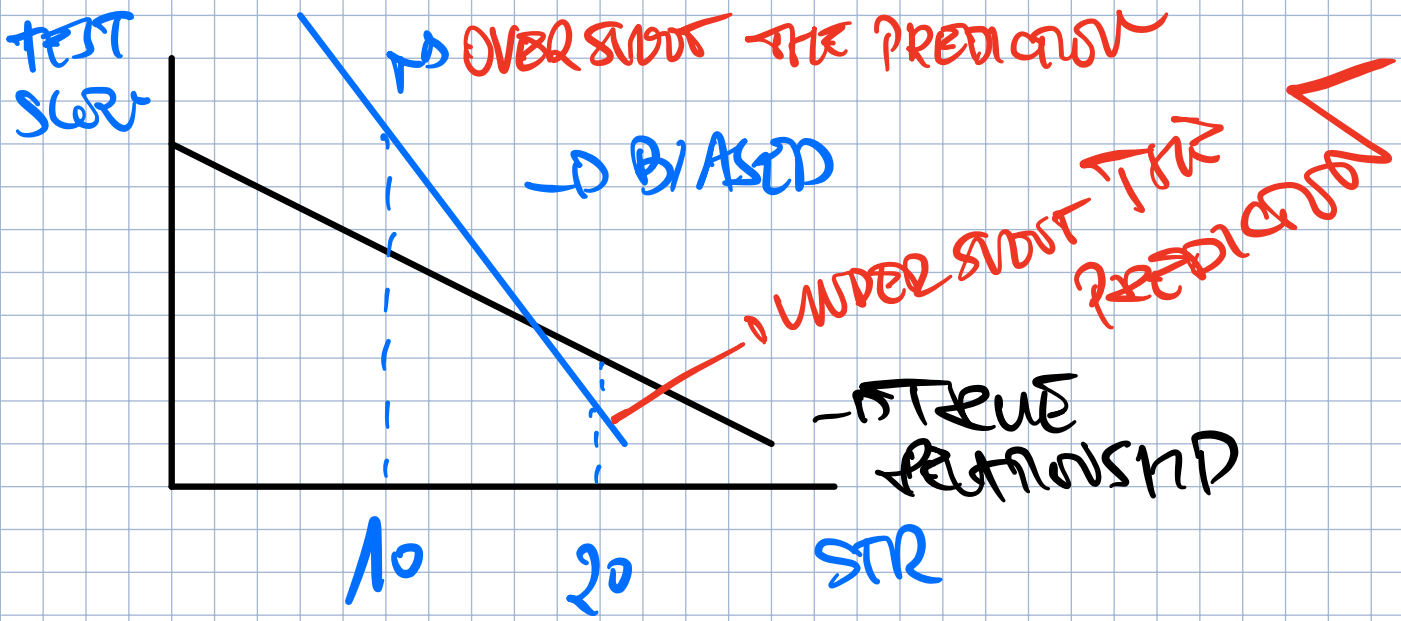
↳ AT A GIVEN VALUE OF X, Y IS DISTRIBUTED AROUND THE REGRESSION LINE, AND THE ERROR $u = Y - (\beta_0 + \beta_1 X)$

↳ HAS A CONDITIONAL MEAN OF ZERO

ERROR: THE DIFFERENCE B/W

$u_i|X_i$ $Y_i|X_i$ & $E(Y_i|X_i]$

ASSUMPTION IS IF $COR(u_i, x_i = 0$ THEN YOU SOMETIMES UNDERPREDICT & SOMETIMES OVER PREDICT BUT ON AVERAGE $E[u_i|X_i]$ MEAN OF THE ERROR IS ZERO.



PROOF -

$$Cov(x,u) = E[(x - E[x])(u - E[u])$$

$$= E[xu - xE[u] - E[x]u + E[x]E[u])$$

$$= E[xu] - E[x]E[u] - E[x]E[u] + E[x]E[u]$$

$$= E[xu] - E[x]E[u]$$

$$= E[xu] - E[x]E[E[u|x]]$$

$$= 0$$

$$= E[xu]$$

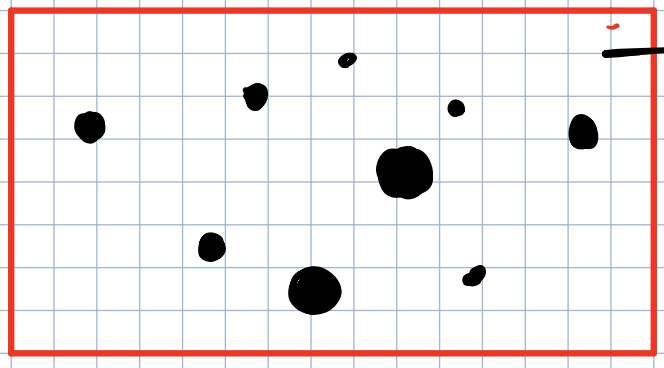
$$= E[E[xu|x]] = E[xE[u|x]] = 0$$

$$\neq 0$$

$$\frac{\text{cov}(x,u)}{\text{Var}(x)} = \text{cor}(x,u) = 0$$

ASSUMPTION 1: (x_i, y_i) $i = 1, \ldots, n$ ARE INDEPENDENTLY & IDENTICALLY DISTRIBUTED. → ASSUMPTION ABOUT HOW SAMPLE IS DRAWN

RANDOM SAMPLING INSURES THIS.



→ IDENTICALLY DISTRIBUTED (COME FROM THE SAME PROBABILITY DISTRIBUTION)

→ INDEPENDENT: DRAWS HAVE NO MEMORY.

ASSUMPTION 3: LARGE OUTLIERS ARE UNLIKELY

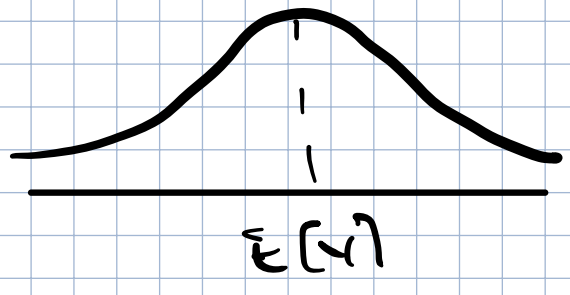
LOLS MIGHT BE MISLEADING IF THERE ARE LARGE OUTLIERS (NEED FOR CONSISTENCY :-)

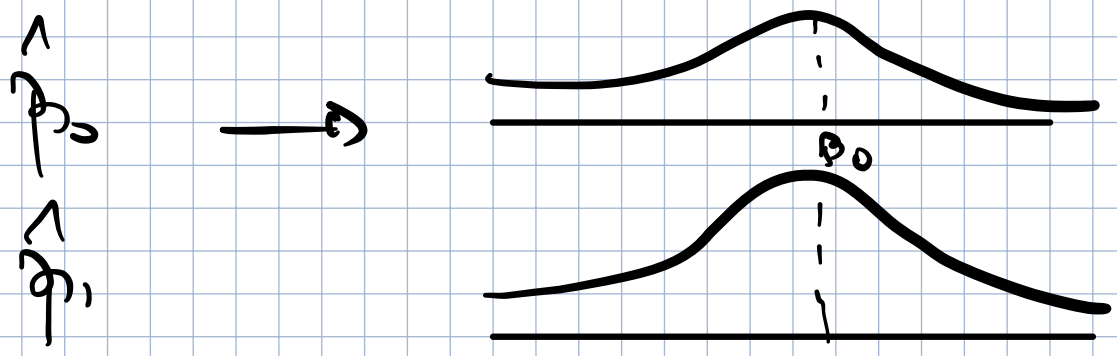
$$s_2^{\,Y} \xrightarrow{p} \sigma_Y^2$$

→ [TECHNICAL ASSUMPTION]

SAMPLING DISTRIBUTION OF THE OLS ESTIMATOR

$\bar{Y}$ →





→ AT LEAST FOR β_1

→ WHEN n IS LARGE & ALL OLS ASSUMPTIONS HOLD!

$E(\hat{\beta}_0) = \beta_0$

$E[\hat{\beta}_1] = \beta_1$

} BY CENTRAL LIMIT THEOREM

→ 3 ASSUMPTIONS HOLD: 1. $E(u_i|X_i) = 0$, $\text{corr}(u_i, x_i) = 0$ 2. (x_i, y_i) iid 3. NO LARGE OUTLIERS

KEY CONCEPT 4.4

Large-Sample Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

If the least squares assumptions in Key Concept 4.3 hold, then in large samples $\hat{\beta}_0$ and $\hat{\beta}_1$ have a jointly normal sampling distribution. The large-sample normal distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma^2_{\hat{\beta}_1})$, where the variance of this distribution, $\sigma^2_{\hat{\beta}_1}$, is

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}. \quad (4.21)$$

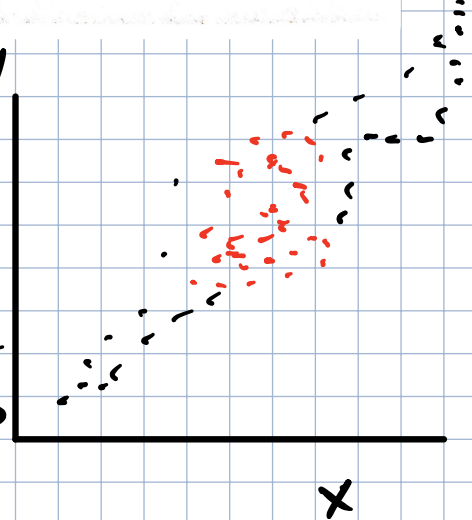
→ UNIMPORTANT

The large-sample normal distribution of $\hat{\beta}_0$ is $N(\beta_0, \sigma^2_{\hat{\beta}_0})$, where

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\text{var}(H_i u_i)}{[E(H_i^2)]^2}, \text{ where } H_i = 1 - \left[\frac{\mu_X}{E(X_i^2)}\right]X_i. \quad (4.22)$$

AS $\text{VAR}(X_i) \uparrow$, $\sigma^2_{\hat{\beta}_1} \downarrow$

→ THIS IS A MEASURE OF PRECISION



Answer to questions

a. $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

$Y_i = \hat{Y}_i + \hat{U}_i$

$\hat{\beta}_0 = 509.384 \qquad X_i = 22$

$\hat{\beta}_1 = -5.614$

$\hat{Y}_i = 509.384 - 5.614 \times 22$

$\hat{Y}_i \cong 385.9$

$\longrightarrow \hat{\beta}_1 = \dfrac{\Delta \hat{Y}_i}{\Delta X_i}$

b.

$\Delta X_i = 23 - 19$

$= 4$

$\Delta \hat{Y}_i = 4 \times -5.64$

$= -22.5$

c. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \bar{Y} = 21.$

$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} \qquad \bar{Y} = ?$

$= 509.384 - 5.614 \times 21 \quad = 389.2$

d. VARIATION IN CS CAN EXPLAIN 54.1. OF THE VARIATION IN TEST 2. IF WE USE CS TO PREDICT TEST SCORES, OUR AVERAGE WE WILL BE OFF BY 105.8 POINTS

e. $SER = \sqrt{\frac{SSR}{n-2}}$

$$105.8 = \sqrt{\frac{SSR}{98}}$$

$$11193.64 = SSR/98$$

$$SSR = 1096977$$

$$R^2 = 1 - \frac{SSR}{TSS} \quad 0.0849 = 1 - \frac{1096977}{TSS}$$

$$S_Y^2 = \frac{TSS}{(n-1)} \quad 0.9151 = \frac{1096977}{TSS}$$

$$TSS = 1198791$$

$$S_Y^2 = 12109 \qquad S_Y = 110$$

CHAPTER 5. HYPOTHESIS TESTS & CONFIDENCE INTERVALS USING A SINGLE REGRESSOR.

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

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$

$$Y_i = \hat{Y}_i + \hat{u}_i$$

— SAMPLING DISTRIBUTION

$$t = \frac{\text{ESTIMATOR} - \text{HYPOTHESIZED VALUE}}{\text{STANDARD ERROR OF THE ESTIMATOR}}$$

IN MOST COMMON CASE: $H_0: \beta_1 = 0$

GENERAL CASE:

$H_0: \beta_1 = \beta_{1,0}$

$H_1: \beta_1 \neq 0$

$H_1: \beta_1 \neq \beta_{1,0}$

— THREE STEPS

1. $SE[\hat{\beta}_1]$ → ESTIMATOR FOR $\sigma_{\hat{\beta}_1}$

$$SE[\hat{\beta}_1] = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$$

IN STANDARD DEVIATION OF THE SAMPLING DISTRIBUTION

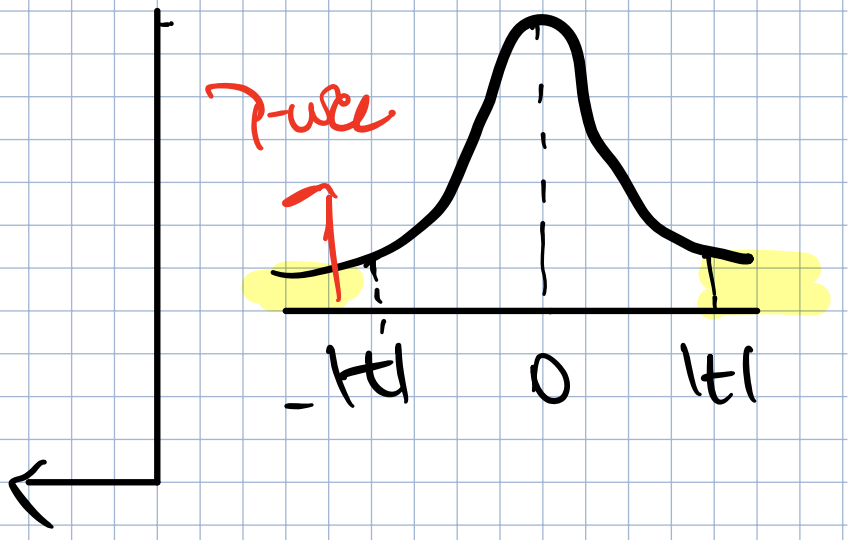
$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}},$$

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}.$$

the variance in Equation (5.4) is discussed
ula for $\hat{\sigma}^2_{\hat{\beta}_1}$ is complicated, in applications the
ession software so that it is easy to use in pra
ep is to compute the ***t*-statistic**,

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}.$$

2. CALCULATE t-STATISTIC



3. COMPUTE p-value: THE PROBABILITY OF OBSERVING A VALUE OF $\hat{\beta}_1$ UNDER THE ASSUMPTION THAT $\beta_1 = \beta_{1,0}$ REJECTED

ANY BETA vs DECIDES ON THIS VALUE

ACTUAL t-STAT

$$\text{p-value} = \Pr_{H_0}\left[|\hat{\beta}_1 - \beta_{1,0}| > |\hat{\beta}_1^{act} - \beta_{1,0}|\right]$$

$$\Rightarrow \Pr_{H_0}\left[\left|\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right| > \left|\frac{\hat{\beta}_1^{act} - \beta_{1,0}}{SE(\hat{\beta}_1)}\right|\right]$$

$$\Rightarrow \Pr_{H_0}(|t| > |t^{act}|)$$

ABSOLUTE => 2/C
TWO t-TEST

IN LARGE SAMPLES ?? $\Pr(|Z| > |t^{act}|)$

$\hookrightarrow 2\Phi(-|t^{act}|)$

t STAT ARE VERY SIMILAR IN LARGE SAMPLES

REJECT THE NULL IF

p-value < pre-specified threshold

→ 1%, 5%, 10%

CRITICAL t-VALUES $2.58^{1\%}$ $1.96^{5\%}$ $1.645^{10\%}$

ANSWER TO QUESTION

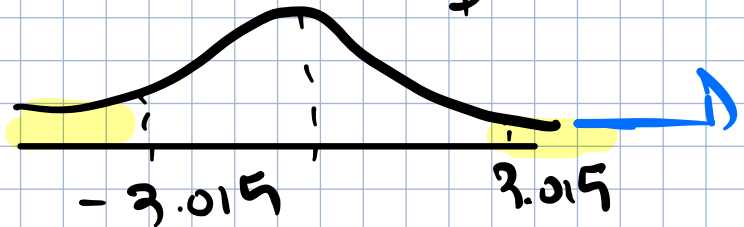
1° $H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

2° $SE[\hat{\beta}_1] = 1.862$

t-stat $= \frac{-5.614 - 0}{1.86} = -3.015$

3° p-value $= 2\Phi(-3.015)$ ⇒



27

p-value < 0.01 → REJECT THE NULL HYPOTHESIS.

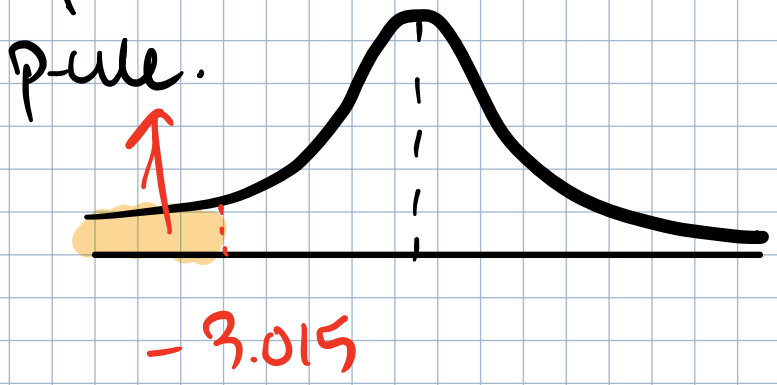
$\hat{\beta}_1$, SE($\hat{\beta}_1$) = 1.162 → t-stat = $\frac{-5.614}{1.862}$

p-value = $\Phi(-3.015)$ = -3.015

p-value = 0.001635

$H_0: \beta_1 = 0$

$H_A: \beta_1 < 0$



t-stat = 3.5 → ? Fail to Reject the Null.

DO NOT OFTEN USED

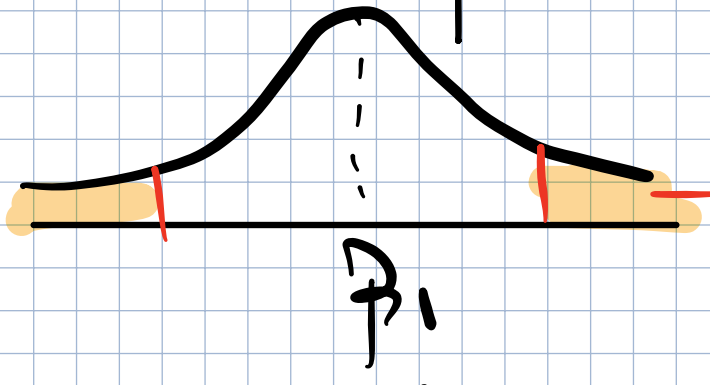
TESTING HYPOTHESES ABOUT THE INTERCEPT β_0 → ALWAYS USE AN INTERCEPT EVEN IF YOU FAIL TO REJECT THE NULL

CONFIDENCE INTERVAL FOR β_1 (95%)

2 DEFINITIONS:

1. SET OF VALUES THAT CAN'T BE REJECTED USING A TWO-SIDED HYP.

2° 95% CONTAINING THE TRUE VALUE OF β_1



→ 5% OF ALL POSSIBLE SAMPLES OF β WILL BE REJECTED

$$\hat{\beta}_1 - 1.96\,SE(\hat{\beta}_1)$$

$$\hat{\beta}_1 + 1.96\,SE(\hat{\beta}_1)$$

99% CI → $\hat{\beta}_1 \mp 2.576 \times SE(\hat{\beta}_1)$

$$-5.614 \mp 2.576 \times 1.162$$

$$[-10.50 \quad -0.7231]$$

CONFIDENCE INTERVALS FOR CHANGES IN X

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + u_i$$

CONSIDER Δx → CHANGE IN X

THE PREDICTED CHANGE IN Y (ΔY)

IN POP. REGRESSION FUNCTION

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

$$\Delta Y_i = \beta_1 \Delta X_i$$

IN 99% CONFIDENCE INTERVAL FOR $\Delta x \cdot \beta_1$

$$[\hat{\beta}_1 \Delta x \pm 1.96 SE[\hat{\beta}_1] \times \Delta x]$$

$$\Delta x \cdot [\hat{\beta}_1 \pm 1.96 SE[\hat{\beta}_1]$$

$$\Delta x [+0.50, +0.7231]$$

$$2 [+0.50, +0.7231]$$

$$\text{CI} [1.01, 1.45] \text{ CLOSE IN TEST SCORE}$$

REGRESSION WHEN X IS A BINARY VARIABLE

LA ONE OF THE MOST COMMON CASES:

- GENDER (MALE FEMALE) $\{0,1\}$
  (1 IF A IS NO-BA)

INDICATOR, DUMMY, BINARY, INDEX VARIABLE

$D_i = \{0,1\}$

$X_i \rightarrow D_i = \begin{cases} \text{SMALL} & \text{STR} < 20 \\ \text{LARGE} & \text{STR} \geq 20 \end{cases}$

$Y_i = \beta_0 + \beta_1 D_i + u_i \quad i = 1, \ldots, n$

$\hat{\beta}_1$ → OLS → HOW TO INTERPRET THIS COEFFICIENT?

EASIEST WAY

LA NOT ASLOPE ANYMORE

$Y_i = \beta_0 + u_i$ (IF $D_i = 0$)

$Y_i = \beta_0 + \beta_1 + u_i$ (IF $D_i = 1$)

$$E[Y_i|D_i=0] = E[\beta_0] + E[u_i|D_i=0]$$

$$E[Y_i|D_i=1] = E[\beta_0] + E[\beta_1] + E[u_i|D_i=1]$$

OLS ASSUMPTION 1.

$$E[u_i|X_i] = 0$$

$$E[Y_i|D_i=0] = \beta_0$$

$$E[Y_i|D_i=1] = \beta_0 + \beta_1$$

POPULATION MEAN OF Y_i WHEN $D_i=0$

$\beta_1 \rightarrow E[Y_i|D_i=1] - E[Y_i|D_i=0]$

$\rightarrow$ POPULATION MEAN OF Y_i WHEN $D_i=1$

$\beta_1 \rightarrow$ DIFFERENCE BTW MEAN TEST-SCORES IN LOW STUDENT-TO-TEACHER RATIO AND HIGH STUDENT-TO-TEACHER RATIO.

$\hat{\beta}_1 \rightarrow$ DIFFERENCE IN SAMPLE AVERAGE OF Y WHEN $D_i=1$ VS $D_i=0$

g. $\bar{Y} | D_i = 0$ → 649.92 → AVERAGE TEST SCORE OF STUDENTS IN LARGE CLASSROOM.

$\bar{Y} | D_i = 1$ → 649.92 + 44.45

→ 414.37 → AVERAGE TEST SCORE FOR STUDENT IN SMALL CLASSROOM.

$\bar{Y} | D_i = 1 - \bar{Y} | D_i = 0$ → 44.45 POINTS OF DIFFERENCE BTW SMALL & LARGE CLASSROOMS:

$44.45 \pm 1.96 \times 22.19$

[0.96 87.94] → 95% CI

CI DOES NOT INCLUDE ~~ZERO~~.

$$t\text{-stat} = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{44.45}{22.191} = 2.003$$

CH 6. LINEAR REGRESSION W/ MULTIPLE REGRESSORS

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i$$

k - independent variables

IF WE ARE INTERESTED IN THE RELATIONSHIP BTW ONE X & Y; WHY DO WE NEED MORE THAN ONE INDEPENDENT VARIABLE?

THE IMPACT OF CLASSROOM SIZE → TEST SCORES (ENGLISH) IN CA

OMITTED VARIABLE BIAS:

CA HAS A LARGE IMMIGRANT POPULATION

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

↓ LOWER RATIOS

TEST SCORE → $\hat{\beta}_1$ MIGHT NOT REFLECT THE

LO PROBABLY WILL NOT REFLECT THE TRUE IMPACT OF THE STR ON TEST SCORE

NEW FACTOR: % OF ENGLISH LEARNERS IN A DISTRICT.

↳ DISTRICTS WITH A HIGHER SHARE OF ENGLISH LEARNERS PERFORM WORSE IN ENGLISH TEST

↳ DISTRICTS WITH LARGE CLASSES MIGHT ALSO HAVE A HIGHER SHARE OF ENGLISH LEARNERS

SO THE RELATIONSHIP THAT IS DESCRIBED BY THE REGRESSION RESULTS MIGHT BE DRIVEN BY THE % OF ENGLISH LEARNERS (IMMIGRANTS) WHILE THE TRUE RELATIONSHIP IS MUCH WEAKER OR EVEN NON-EXISTENT.

OMITTED VARIABLE BIAS: AN IMPORTANT REGRESSOR THAT i. INFLUENCES THE OUTCOME
ii. CORRELATED WITH ONE OR THE INDEPENDENT VARIABLES

OMITTED FROM THE REGRESSION.

TWO CONDITIONS: ↳ OMITTED VARIABLE SHOULD HAVE AN IMPACT ON THE OUTCOME
↳ CORRELATED WITH A REGRESSOR.

$E[u_i|x_i] \neq 0$ $Corr(u_i, x_i) \neq 0$

$Corr(x_i, u_i) = \rho_{xu}$ → REGRESSOR & ERROR TERM

$E[\hat{\beta}_1] = \beta_1 + \rho_{xu} \sigma_u / \sigma_x$ → BIASED DUE TO OMITTED VARIABLE

$E(\hat{\beta}_1) = \beta_1$ → UNBIASED

1. A SAMPLE SIZE WILL NOT HELP.
2. WHETHER THE BIAS IS SMALL OR LARGE DEPENDS ON $Corr(u_i, x_i)$ & $Corr(x_i, y_i)$

LARGER THE COR., LARGER THE BIAS.

ALSO THE DIRECTION DEPENDS ON THE THREE WAY RELATIONSHIP BTW

X_{1i}, X_{2i}, y_i → OUTCOME

↓

OMITTED VAR

CONSIDER ON THIS LATER

ADDRESSING THE OMITTED VARIABLE BIAS:

CD MEASURE THE IMPACT OF STR ON TEST SCORE HOLDING % OF ENGLISH LEARNERS CONSTANT.

IDEA: COMPARE SMALL VS LARGE CLASSIZE AMONG DISTRICTS WITH SIMILAR % OF ENGLISH LEARNERS

THE MULTIPLE REGRESSION MODEL

3/4/2020

IDEA: ESTIMATE THE IMPACT OF X_{1i} (STR) ON TEST SCORE Y_i WHILE HOLDING X_{2i} (% OF ENGLISH LEARNERS) CONSTANT

POPULATION REGRESSION FUNCTION (LINE)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$E[Y_i \mid X_{1i} = x_1, X_{2i} = x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

POPULATION AVERAGE OF TEST SCORES FOR DISTRICTS WITH X_1 STR & X_2 % OF ENG. LEARNERS

β_0 → INTERCEPT

β_1 → SLOPE COEFFICIENT FOR X_{1i}

β_2 → SLOPE COEFFICIENT FOR X_{2i}

X_{1i}, X_{2i} → CONTROL VARIABLES

β_1: CHANGE IN Y INDUCED BY A UNIT CHANGE IN X_{1i}, HOLDING X_{2i} CONSTANT.

AFTER ACCOUNTING FOR X_{2i}
AFTER CONTROLLING FOR X_{2i}

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

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2$$

$$\Delta Y = \beta_1 \Delta X_1 \qquad \beta_1 = \frac{\Delta Y}{\Delta X_1}$$

↳ PARTIAL EFFECT OF X_1 ON Y

$$E[Y \mid X_1 = 0, X_2 = 0] = \beta_0$$

POPULATION AVERAGE WHEN $X_1 = 0, X_2 = 0$ ↩

APPLICATION MULTIPLE REGRESSION

MODEL:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

u_i: Different for each person

u_i: ALL PARTS OF Y_i THAT CAN'T BE EXPLAINED BY $X_{1i} + X_{2i}$

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

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + u_i$$

$X_0 \equiv 1, \ldots, 1$ ($n \times 1$)

CONSTANT REGRESSOR (CONSTANT TERM)

MORE GENERAL

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

Homoskedasticity Assumption:

$$Var(u_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = \sigma_u^2$$

CONSTANT

$E(Y_i | X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{ki} = x_k)$

$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$

$\beta_1 \rightarrow \frac{\Delta Y}{\Delta X_1}$ THE CHANGE IN Y INDUCED BY A UNIT CHANGE IN X HOLDING ALL ELSE CONSTANT.

THE OLS ESTIMATOR

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$

FIND $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k \rightarrow \min \sum_i \hat{u}_i^2$

$k+1$ PARAMETERS TO ESTIMATE

WHERE $Y_i = \hat{Y}_i + \hat{u}_i$

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \beta_k X_{ki}$

$\hat{u}_i = Y_i - \hat{Y}_i \quad \min \sum_i \hat{u}_i^2 \quad \sum_i (Y_i - \hat{Y}_i)^2$

DISTRICTS WITH A HIGH % OF ESL LEARNERS TEND TO HAVE NOT ONLY LOW TE SCORES BUT ALSO A HIGH STUDENT/TEACHER RATIO WHEN WE OMIT ENGLISH LEARNERS (LARGER ESTIMATED COEFFICIENT)

TRUE MODEL:

MODEL 1: $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

↳ WHAT YOU ARE INTERESTED IN → $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

MODEL 2:

ESTIMATED MODEL:

$$y_i = \beta_0 + \beta_1 X_{1i} + u_i \qquad \tilde{\beta}_0, \tilde{\beta}_1$$

X_2 IS OMITTED → NOT AVAILABLE ↳ FORGOT

$\tilde{\beta}_1 = -2.6210$

$\hat{\beta}_1 = -1.28970$

$\hat{\alpha}_1 = 1.8137$

AUXILIARY MODEL:

$X_{2i} = \alpha_0 + \alpha_1 X_{1i} + \epsilon_i$

$\hat{\beta}_2 = -0.73403$ ← ERROR TERM

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \times \hat{\alpha}_1$$

$$-2.6210 = -1.28970 \underbrace{-0.73403 \times 1.8137}$$

$$= -2.6210$$

-1.33

$$\text{BIAS} = \hat{\beta}_2 \times \hat{\alpha}_1 + 0$$

↳ (-)

OLS ESTIMATOR

OBJECTIVE: ESTIMATE

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i$$

FIND $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots \hat{\beta}_k$ THAT MINIMIZE $\sum_i \hat{u}_i^2$

SIMULTANEOUSLY

$(k+1)$ PARAMETERS

k INDEPENDENT VARIABLES

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}$$

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

$$\text{MIN} \sum \hat{u}_i^2 \leftarrow \text{MIN} \sum_i (Y_i - \hat{Y}_i)^2$$

SSR OR RSS

ADJUSTED R^2

MEASURES OF FIT: SER + RMSE + Adj. R^2

SER: STANDARD ERROR OF THE REGRESSION

RMSE: ROOT MEAN SQUARED ERROR

SER & RMSE → VERY SIMILAR

BOTH MEASURE THE SPREAD OF Y_i AROUND Y_i

Y_i → OBSERVED OUTCOME

$\hat{Y}_i$ → PREDICTED OUTCOME

$$SER = \sqrt{\frac{1}{(n-k-1)} \sum_i \hat{u}_i^2}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_i \hat{u}_i^2}$$

k = NO. OF INDEPENDENT VARIABLES

↳ NO. OF ~~DEGREES OF~~ FREEDOM

↳ BOTH ARE MEASURED IN THE UNITS OF OUTPUT.

$$Y_i = \hat{Y}_i + \hat{u}_i$$

$\hat{u}_i$ → RESIDUAL

↳ AVERAGE DISTANCE B/W OBSERVED VALUES VS REGRESSION LINE

AVERAGE PREDICTION ERROR

WHEN n IS LARGE RMSE & SER ARE CLOSE TO EACH OTHER.

Adjusted R^2-squared: R^2 adjusted for the nb of regressors in the equation

$$R^2 = ESS / TSS \qquad ESS = \sum_i (\hat{Y}_i - \bar{Y})^2$$

$$TSS = \sum_i (Y_i - \bar{Y})^2$$

$$SSR = \sum_i \hat{u}_i^2$$

$$TSS = ESS + SSR$$

Problem: everytime you add a new variable R^2 will grow independent of that variable is useful or not.

Solution: Adjusted R^2-squared

$$\bar{R}^2 = 1 - \left[\frac{(n-1)}{(n-k-1)}\right] \frac{SSR}{TSS}$$

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

to penalize for each + regressor

$$\frac{(n-1)}{(n-k-1)} > 1 \longrightarrow \bar{R}^2 \leq R^2$$

IF $n \to \infty$ $\quad \bar{R}^2 \to R^2$

$\bar{R}^2 < 0 \to$ THEORETICALLY

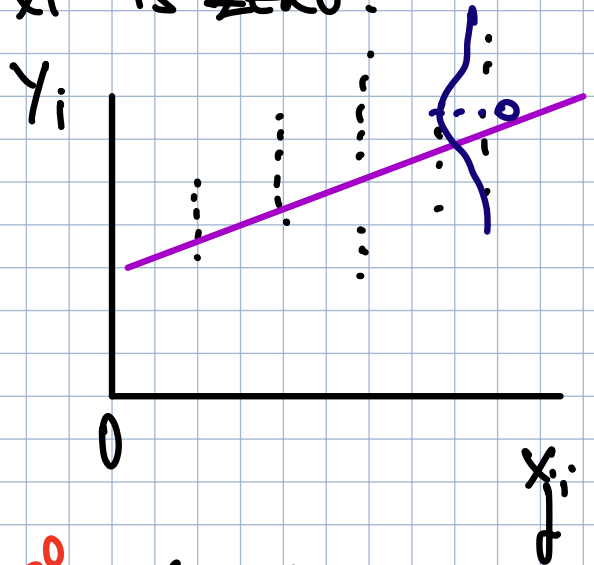
LEAST SQUARES ASSUMPTIONS FOR MULTIVARIATE REGRESSION:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i$$

$i = 1, \dots, n$

1° $E[u_i \mid X_{1i} = x_1, X_{2i} = x_2, \dots, X_{ki} = x_k] = 0$

↳ THE CONDITIONAL DISTRIBUTION OF u_i GIVEN X_i IS ZERO.



NO OMITTED VARIABLE BIAS ← TWO CONDITIONS FOR THE OMV

1° $\text{corr}(u_i, X_{ji}) \neq 0$

2° $u_i \to Y_i$

2° $X_{1i}, X_{2i}, \dots, X_{ki}, Y_i$ i.i.d.

IDENTICALLY & INDEPENDENTLY DISTRIBUTED ← RANDOM SAMPLING INSURES THIS

↳ AUTOMATICALLY SATISFIED WITH A RANDOM

3° LARGE OUTLIERS ARE UNLIKELY.

↳ CHECK MIN, MAX, MEAN, MEDIAN

4. NO PERFECT MULTICOLLINEARITY

GAP IS LARGER WHEN THERE ARE OUTLIERS

ONE REGRESSOR IS AN EXACT LINEAR FUNCTION OF ANOTHER REGRESSOR

Ex: $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$

Y = WAGE

X_1 = EDUCATIONS IN YEARS

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$X_{2i} = 2 \times X_{1i}$$

PERFECT LINEAR FUNCTION

```
> mo5 <- lm(data=CPS1988, formula=wage~education)
> summary(mo5)

Call:
lm(formula = wage ~ education, data = CPS1988)

Residuals:
    Min      1Q  Median      3Q     Max
 -786.0  -264.9   -54.8   174.5 18035.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.8281    11.8970  -1.078    0.281
education    47.1810     0.8888  53.085   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 432.4 on 28153 degrees of freedom
Multiple R-squared:  0.09099,   Adjusted R-squared:  0.09096
F-statistic:  2818 on 1 and 28153 DF,  p-value: < 2.2e-16

>
> CPS1988$educ2 <- CPS1988$education*2
> mo6 <- lm(data=CPS1988, formula=wage~education + educ2)
> summary(mo6)

Call:
lm(formula = wage ~ education + educ2, data = CPS1988)

Residuals:
    Min      1Q  Median      3Q     Max
 -786.0  -264.9   -54.8   174.5 18035.1

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.8281    11.8970  -1.078    0.281
education    47.1810     0.8888  53.085   <2e-16 ***
educ2            NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 432.4 on 28153 degrees of freedom
Multiple R-squared:  0.09099,   Adjusted R-squared:  0.09096
F-statistic:  2818 on 1 and 28153 DF,  p-value: < 2.2e-16

>
```

DUMMY VARIABLE TRAP: SUPPOSE THAT YOU WANT TO INCLUDE AN INDICATOR VARIABLE D_i IN YOUR REGRESSION:

FEMALE = {0, 1}

↓ 0 FEMALE

MALE

RESPONSE OUTCOME VARIABLE ⟶ WAGE

D_{1i} = {0, 1} (0 MALE, 1 FEMALE)

D_{2i} = {1, 0} (1 MALE, 0 FEMALE)

$Y_i = \alpha_0 + \alpha_1 D_{1i} + u_i$

α_1 = AVERAGE WAGE DIFFERENCE B/W FEMALES AND MALES

$Y_i = \beta_0 + \beta_1 D_{2i} + u_i$

β_1 = AVERAGE WAGE DIFFERENCE B/W MALES AND FEMALES

↳ $\alpha_1 = -\beta_1$

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$$

↳ WHAT IS THE ISSUE?

$D_{1i} + D_{2i} = 1 \qquad D_{1i} = D_{2i} - 1$

PERFECT LINEAR FUNCTION

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 (1 - D_{1i}) + u_i$$

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 - \beta_2 D_{1i} + u_i$$

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 - \beta_2) D_{1i} + u_i$$

1 SINGLE REGRESSOR

TWO PARAMETERS

IMPOSSIBLE ESTIMATE

```
> CPS1988$white <- 0
> CPS1988$white[CPS1988$ethnicity=='cauc'] <- 1
> mo5 <-  lm(data=CPS1988, formula=wage~education+white)
> summary(mo5)

Call:
lm(formula = wage ~ education + white, data = CPS1988)

Residuals:
    Min      1Q  Median      3Q     Max
 -792.0  -256.3   -57.1   173.6 18027.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -123.2581    14.2508  -8.649   <2e-16 ***
education     46.2504     0.8882  52.070   <2e-16 ***
white        133.1458     9.5332  13.967   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 430.9 on 28152 degrees of freedom
Multiple R-squared:  0.09724,   Adjusted R-squared:  0.09718
F-statistic:  1516 on 2 and 28152 DF,  p-value: < 2.2e-16
```

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 D_{1i} + u_i$

$E[Y_i \mid X_{1i}=0, D_{1i}=0] = \beta_0$

β_0

β_1

β_2

```
> CPS1988$non.white <- 0
> CPS1988$non.white[CPS1988$white==0] <- 1
>
> mo6 <-  lm(data=CPS1988, formula=wage~ education + non.white + white)
> summary(mo6)

Call:
lm(formula = wage ~ education + non.white + white, data = CPS1988)

Residuals:
    Min      1Q  Median      3Q     Max
 -792.0  -256.3   -57.1   173.6 18027.3

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.8877    11.9672   0.826    0.409
education     46.2504     0.8882  52.070   <2e-16 ***
non.white   -133.1458     9.5332 -13.967   <2e-16 ***
white             NA         NA      NA       NA

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 430.9 on 28152 degrees of freedom
Multiple R-squared:  0.09724,   Adjusted R-squared:  0.09718
F-statistic:  1516 on 2 and 28152 DF,  p-value: < 2.2e-16
```

(D DROPPED)

IMPERFECT MULTICOLLINEARITY

MAKES A BIG DIFFERENCE.

X_{1i} & X_{2i} ARE HIGHLY CORRELATED

BUT NOT A PERFECT FUNCTION OF EACH OTHER

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

THE OLS WILL WORK BUT POORLY

$\hat{\beta}_1, \hat{\beta}_2$ WILL BE VERY IMPRECISE

(LARGE STANDARD ERRORS)

$\hat{\beta}_1$ → RELATIONSHIP BTW X_{1i} & Y HOLDING X_{2i} CONSTANT.

IF X_{1i} & X_{2i} ARE HIGHLY CORRELATED, THEN IT IS DIFFICULT TO ESTIMATE THE NET EFFECT X_{1i}. → VERY LITTLE VARIATION LEFT TO EXPLOIT

RESULT → $SE[\hat{\beta}_1]$ = LARGE

95% CI $\hat{\beta}_1$ IS LARGE

HOW TO DETECT?

→ PAIRWISE COR

|COR| > 0.80 → NOT A GOOD SIGN

R^2 HIGH BUT INDIVIDUAL COEF. ARE NOT STATISTICALLY SIGNIFICANT DUE TO LARGE STANDARD ERRORS.

HIGH VARIANCE INFLATION FACTOR:

→ DISCARD

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + U_i$$

$$Y_{1i} = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \dots + \alpha_k X_{ki} + \varepsilon_i$$

$$VIF(\beta_1) = 1/(1 - R_1^2)$$

→ CALCULATE R^2

$VIF(\beta_1) > 5$ → INDICATES SEVERE MULTICOLLINEARITY

HOW TO FIX?

- DO NOTHING
- DROP REDUNDANT VARIABLE
- TRANSFORM MULTICOLLINEAR VARIABLES

GDP / PSP CORRELATED → GDP PER CAPITA

LARGER SAMPLE ALSO NEEDS.

```
> mo7 <- lm(data=CASchools, formula=read~str+english+lunch+expenditure)
>
> summary(mo7)

Call:
lm(formula = read ~ str + english + lunch + expenditure, data = CASchools)

Residuals:
    Min      1Q  Median      3Q     Max
-31.181  -5.353  -0.028   5.247  31.735

Coefficients:
             Estimate  Std. Error t value             Pr(>|t|)
(Intercept) 662.0903950   9.0716578  72.984 < 0.0000000000000002 ***
str          -0.2031297   0.2860344  -0.710                0.478
english      -0.2105934   0.0304482  -6.916      0.0000000000176 ***
lunch        -0.5502143   0.0203238 -27.072 < 0.0000000000000002 ***
expenditure   0.0046660   0.0008406   5.551      0.0000000508161 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.544 on 415 degrees of freedom
Multiple R-squared:  0.8212,    Adjusted R-squared:  0.8195
F-statistic: 476.5 on 4 and 415 DF,  p-value: < 0.00000000000000022

> cor(CASchools1)
                read        str     english       lunch expenditure
read         1.0000000 -0.2465930 -0.69028587 -0.87880769  0.21792682
str         -0.2465930  1.0000000  0.18764237  0.13520340 -0.61998216
english     -0.6902859  0.1876424  1.00000000  0.65306072 -0.07139604
lunch       -0.8788077  0.1352034  0.65306072  1.00000000 -0.06103871
expenditure  0.2179268 -0.6199822 -0.07139604 -0.06103871  1.00000000
> vif(mo7)
       str    english      lunch expenditure
  1.680861   1.779490   1.744355   1.630108
>
```

-0.203 ∓ 1.96 × 0.286

[-0.76 95% +0.36]

VARIABLE OF INTEREST: STR

CONTROL VARIABLES: % OF ENGLISH LEARNERS
% OF FREE LUNCH
EXPENDITURE PER STUDENT

CONTROL VARIABLES DO NOT NEED TO HAVE A CAUSAL INTERPRETATION BUT STILL CAN BE USED.

HYPOTHESIS TESTING & CONFIDENCE INTERVALS IN MULTIPLE REGRESSIONS

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i$$

↳ CAN BE ANY NUMBER

1. $H_0: \beta_j = \beta_{j,0}$ $\quad j = 1, \dots, k$
   $H_A: \beta_j \neq \beta_{j,0}$

CLT $\longrightarrow$ $\dfrac{\hat{\beta}_1 - E[\hat{\beta}_1]}{\sqrt{\widehat{VAR}(\hat{\beta}_1)}} \sim N(0,1)$

2° $t\text{-stat} = \dfrac{\hat{\beta}_j - \beta_{j,0}}{SE[\hat{\beta}_j]}$

3° $p\text{-val} = 2 \times \Phi(-|t\text{-stat}|)$

if $p\text{-val} < \text{CRITICAL}$? $\longrightarrow$ Reject

$\hat{\beta}_j \pm 1.96 \times SE(\hat{\beta}_j)$ $\longrightarrow$ 95% CONFIDENCE INTERVAL

TEST OF JOINT HYPOTHESES:

NULL REQUIRES TO TEST THE HYPOTHESES JOINTLY

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

- Y_i: TEST SCORE
- X_{1i}: STR
- X_{2i}: % OF EN
- X_{3i}: % ELIGIBLE FOR LUNCH
- X_{4i}: EXPENDITURE PER PERSON

NULL HYPOTHESIS: SCHOOL RESOURCES DON'T MATTER

ALTERNATIVE: THEY DO

$$H_0: \beta_1 = 0 \text{ y } \beta_4 = 0$$

$$H_A: \text{EITHER } \beta_1 \neq 0 \text{ OR } \beta_4 \neq 0$$

$q = 2$ (NB OF RESTRICTIONS IN A JOINT HYPOTHESIS)

INDIVIDUAL T-TEST WON'T WORK — WHY?

ASSUME $\beta_1 = 0$ → AT 5% SIGNIFICANCE LEVEL CHANCE OF FAILING TO REJECT 95% OF

ASSUME $\beta_4 = 0$ → AT 5% SIGNIFICANCE LEVEL 95% FAIL TO REJECT

PROP. FAIL TO REJECT BOTH 0.95×0.95

$= 0.9025$

PROP. REJECTION $\beta_1 = 0$ & $\beta_4 = 0$

$1 - 0.9025 = 9.75\%$

POTENTIALLY A BIGGER PROBLEM:

$\beta_1 = 0$ & $\beta_4 = 0$

ARE DEPENDENT.

→ REJECT TOO OFTEN IF YOU USE INDIVIDUAL t-tests

$$\text{Sullivan F test} = \frac{1}{2}\left[\frac{t_1^2 + t_4^2 - 2\hat{\rho}_{t_1,t_4}\, t_1 t_4}{1 - \hat{\rho}^2_{t_1,t_2}}\right]$$

ESTIMATED CORRELATION B/W t_1 & t_2 $\hat{\rho}_{t_1,t_4}$ → ACCOUNTS FOR CORRELATION B/W X_1 & X_2

$\frac{\text{cov}(x,y)}{\text{var}(x)}$

IF F IS LARGE, REJECT THE NULL

IF t_1 & t_4 ARE INDEPENDENT

$$F = \frac{1}{2}\left[t_1^2 + t_4^2\right]$$ → ONLY IF t_1 & t_4 ARE INDEPENDENT

+ LARGE SAMPLE → $F \sim \chi^2_q$

If $q = 1$ → 5% CRITICAL-F 3.84

$q = 2$ → 5% CRITICAL-F 3.00

$q = 3$ → 5% CRITICAL-F 2.60

```
>
> linearHypothesis(mo7, c("str=0", "expenditure=0"))
Linear hypothesis test

Hypothesis:
str = 0
expenditure = 0

Model 1: restricted model
Model 2: read ~ str + english + lunch + expenditure

  Res.Df   RSS Df Sum of Sq     F              Pr(>F)
1    417 34575
2    415 30292  2    4283.2 29.34 0.000000000001207 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

SIMPLE FORMULA: HOLDS IN SPECIFIC CONDITIONS (HOMOSCEDASTICITY) BUT USEFUL TO UNDERSTAND WHAT F-TEST IS DOING.

FIT TWO REGRESSIONS:

- RESTRICTED → $\beta_1 = 0, \beta_4 = 0$
- UNRESTRICTED → FULL MODEL

CALCULATE R^2 FROM BOTH

$$F = \frac{(R^2_{UNR} - R^2_{RES})/q}{(1 - R^2_{UNR})/(n-k-1)}$$

IF DIFFERENCES IN R^2 IS BIG, F IS LARGE MORE LIKELY TO REJECT H_0

IF DIF IS NOT BIG, THEN MAY BE THE COEFS JOINTLY DO NOT ADD MUCH PREDICTION POWER TO THE MODEL.

$R^2_{UNR} = 0.8212$ $q = 2$ $(n-k-1) = 420 - 4 - 1 = 415$

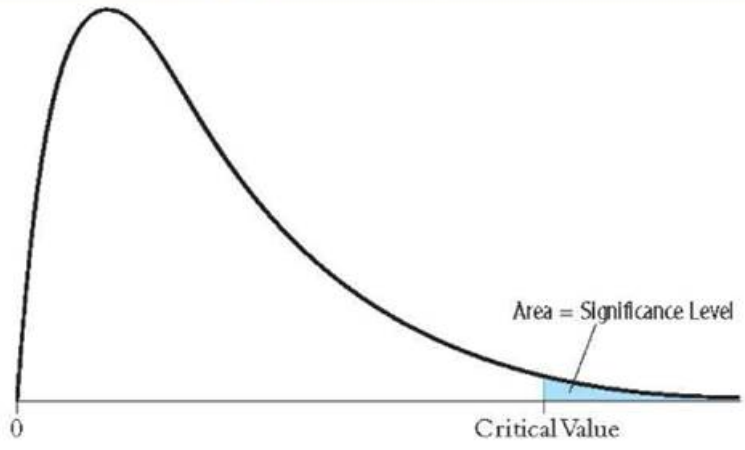
$R^2_{RES} = 0.7959$ $F = \frac{(0.8212 - 0.7959)}{(1 - 0.8212)/415}$

$= 29.36$ REJECT THE NULL > 2.30

→ REJECT THE NULL. CLASSROOM RESOURCES ARE USEFUL

```
> summary(mo8)

Call:
lm(formula = read ~ str + english + lunch + expenditure, data = CASchools)

Residuals:
    Min      1Q  Median      3Q     Max
-31.181  -5.353  -0.028   5.247  31.735

Coefficients:
              Estimate  Std. Error t value             Pr(>|t|)
(Intercept) 662.0903950   9.0716578  72.984 < 0.0000000000000002 ***
str          -0.2031297   0.2860344  -0.710                0.478
english      -0.2105934   0.0304482  -6.916      0.0000000000176 ***
lunch        -0.5502143   0.0203238 -27.072 < 0.0000000000000002 ***
expenditure   0.0046660   0.0008406   5.551      0.0000000508161 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.544 on 415 degrees of freedom
Multiple R-squared:  0.8212,    Adjusted R-squared:  0.8195
F-statistic: 476.5 on 4 and 415 DF,  p-value: < 0.00000000000000022

> summary(mo9)

Call:
lm(formula = read ~ english + lunch, data = CASchools)

Residuals:
    Min      1Q  Median      3Q     Max
-27.130  -5.718   0.185   5.572  33.894

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 683.22276    0.86649 788.497 < 0.0000000000000002 ***
english      -0.22313    0.03212  -6.946      0.0000000000145 ***
lunch        -0.55327    0.02166 -25.547 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.106 on 417 degrees of freedom
Multiple R-squared:  0.7959,    Adjusted R-squared:  0.7949
F-statistic: 813.1 on 2 and 417 DF,  p-value: < 0.00000000000000022

>
```

TABLE 4 Critical Values for the $F_{m,\infty}$ Distribution



	Significance Level		
Degrees of Freedom	**10%**	**5%**	**1%**
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78
4	1.94	2.37	3.32
5	1.85	2.21	3.02
6	1.77	2.10	2.80
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.18
13	1.52	1.72	2.13
14	1.50	1.69	2.08
15	1.49	1.67	2.04
16	1.47	1.64	2.00
17	1.46	1.62	1.97
18	1.44	1.60	1.93
19	1.43	1.59	1.90
20	1.42	1.57	1.88
21	1.41	1.56	1.85
22	1.40	1.54	1.83
23	1.39	1.53	1.81
24	1.38	1.52	1.79
25	1.38	1.51	1.77
26	1.37	1.50	1.76
27	1.36	1.49	1.74
28	1.35	1.48	1.72
29	1.35	1.47	1.71
30	1.34	1.46	1.70

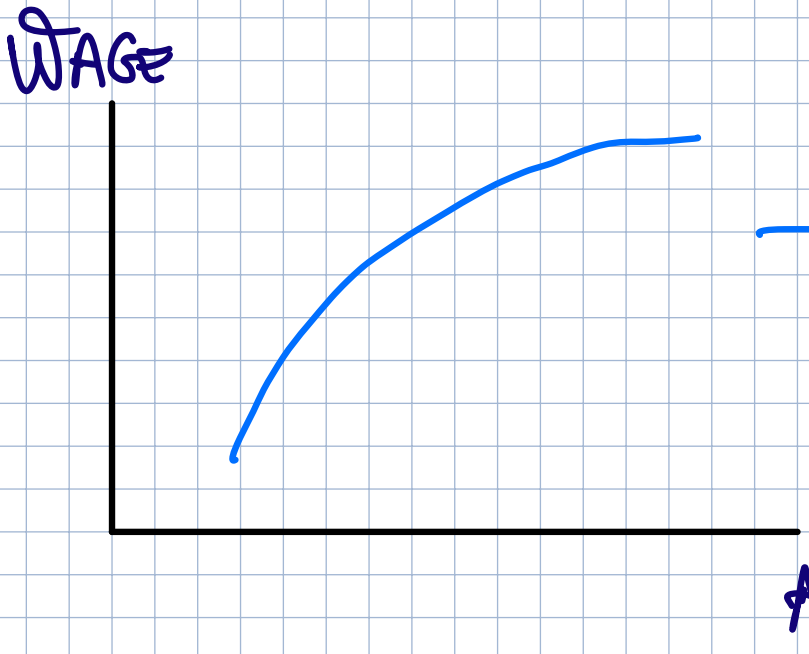
This table contains the 90th, 95th, and 99th percentiles of the $F_{m,\infty}$ distribution. These serve as critical values for tests with significance levels of 10%, 5%, and 1%.

NON-LINEAR MODELS

FORM OF THE RELATIONSHIP B/W Y & X ARE NON-LINEAR

↳ THE IMPACT OF X ON Y DEPENDS ON THE LEVEL OF X → β IS NOT CONSTANT. VALUE

AND IS A FUNCTION OF X. SOMETIMES THEORETICALLY JUSTIFIED.



↳ WAGE ↑ FASTER EARLY IN THE CAREER

GENERAL NON-LINEAR REGRESSION FUNCTION

$$Y_i = f(X_{1i}, X_{2i}, \ldots, X_{ki}) \quad i = 1, 2, \ldots, n$$

↳ NON-LINEAR

SOME CASES: CAN USE OLS AFTER TRANSFORMATION

POLYNOMIALS

US POPULATION REGRESSION FUNCTION CAN BE PROXIED BY A QUADRATIC, CUBIC OR HIGHER ORDER POLYNOMIAL

LOGARITHMIC TRANSFORMATION: Y, X OR BOTH TRANSFORMED TO LOG. WHICH MAKES THEM EASIER TO INTERPRET

POLYNOMIALS:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_k X_i^k + u_i$$

↳ SINGLE X

ALL REGRESSORS ARE POWER OF X

$$Y_i = f(X_i)$$

THE MODEL IS LINEAR IN PARAMETERS SO YOU CAN USE OLS. COEFFICIENTS ARE HARD TO INTERPRET

1. $Y_i = \beta_0 + \beta_1 X_i + u_i$ ⟶ LINEAR
2. $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$ ⟶ QUADRATIC
3. $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$ ⟶ CUBIC

HANDOUT ANSWERS:

1. $\hat{\beta}_1 = 1.815$ p.val < 0.01

$SE[\hat{\beta}_1] = 0.091$

\$1000 ↑ IN INCOME WITH A 1.82 POINTS INCREASE IN TEST SCORES.

2° TEST SCORES ↑ WITH INCOME BUT THE IMPACT OF INCOME IS SMALLER AS THE LEVEL OF INCOME ↑ BUT THE EFFECT OF INCOME IS SMALLER AS THE LEVEL OF INCOME ↑

$\hat{\beta}_1 > 0$ $SE[\hat{\beta}_1] = 0.310$ $\hat{\beta}_1 = 3.073$

$\hat{\beta}_2 < 0$ $SE[\hat{\beta}_2] = 0.006$ $\hat{\beta}_2 = -0.086$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

$$dY_i / dX_i = \beta_1 + 2\beta_2 X_i$$

\+ — LA LEVEL OF X

$\bar{X} \rightarrow dY_i/dX_i = 3.473 + [2 \times (-0.036) \times 15.32]$

$= 3.473 - 1.103$

$= 2.37$

↑ \$1000 IN INCOME WHEN INCOME ≈ \$15317 IS ASSOCIATED WITH A 2.37 POINTS INCREASE IN TEST SCORES.

dY_i/dX_i WHEN $X = \$40{,}000$ → 3.11

$X = \$15{,}000$ → 2.39

$X = \$30{,}000$ → 1.31

CAUTION: NEVER EXTRAPOLATE (MAKE PREDICTIONS/EVALUATE) OUTSIDE THE DATA RANGE OF X.

F-TEST: UNRESTRICTED MODEL:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$$

RESTRICTED

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

H_0: $\beta_2 = 0$ & $\beta_3 = 0$

H_A: EITHER $\beta_2 \neq 0$ OR $\beta_3 \neq 0$

F-stat = 15.702 p-value < 0.01

REJECT H_0 → MODEL SHOULD BE NON-LINEAR.

H_0: $\beta_3 = 0$ $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$

H_A: $\beta_3 \neq 0$ $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$

F-stat = 0.3768 p-value = 0.539 > 0.01

FAIL TO REJECT THE NULL. → OVER-FITTING

X_i^3 SHOULD NOT BE IN THE MODEL.

↳ EQUIVALENT TO AN INDIVIDUAL T-TEST ON β_3

↳ SUMMARY: POLYNOMIAL MODELS CAN BE ESTIMATED USING OLS

↳ INDIVIDUAL COEFFICIENTS ARE HARD TO INTERPRET.

↳ BEST OPTION: TAKE DERIVATIVE + EVALUATE THE IMPACT AT A SPECIFIC X.

↳ DECIDE ON THE APPROPRIATE FORM USING F-TEST & t-TEST.

LOGARITHMIC FUNCTIONS OF Y & X

$\log(x) = \ln(x)$ → THE NATURAL LOG OF x

→ VERY USEFUL FOR MODELING RELATIVE (%) CHANGES.

$$x + \Delta x - x = \Delta x$$

$$\ln(x+\Delta x) - \ln(x) = \ln\left(\frac{x+\Delta x}{x}\right)$$

$$\cong \Delta x / x \rightarrow \text{RELATIVE CHANGE}$$

$\Delta x / x \cdot 100 \rightarrow$ % CHANGE

1° LINEAR-LOG

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$$

2° LOG-LINEAR

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$$

3° LOG-LOG $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$

β_1 IS INTERPRETED VERY DIFFERENTLY IN EACH CASE. USE "BEFORE VS AFTER" RULE

1. LINEAR-LOG

BEFORE: $Y_i = \beta_0 + \beta_1 \cdot \ln(X_i) + u_i$

AFTER: $Y_i + \Delta Y = \beta_0 + \beta_1 \ln(X_i + \Delta X) + u_i$

$$\Delta Y = \beta_1 [\ln(X_i + \Delta X) - \ln(X_i)]$$

$$\Delta Y = \beta_1 \, \Delta X / X_i \qquad \Delta \text{ SMALL}$$

$$\beta_1 = \frac{\Delta Y}{\Delta X / X}$$

$\beta_1 \rightarrow$ A 1% INCREASE IN X IS ASSOCIATED WITH A $\Delta Y / 100$ CHANGE IN Y.

2. LOG-LINEAR FORM

BEFORE : $\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$

AFTER : $\ln(Y+\Delta Y) = \beta_0 + \beta_1 (X_i + \Delta X) + u_i$

$$\ln(Y+\Delta Y) - \ln(Y) = \beta_1 \Delta X$$

$$\Delta Y / Y = \beta_1 \cdot \Delta X$$

$$\beta_1 = \frac{\Delta Y / Y}{\Delta X}$$

β_1 → % CHANGE IN Y ASSOCIATED WITH A 1 UNIT INCREASE IN X

3. LOG-LOG FORM

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$$

$$\ln(Y_i + \Delta Y) = \beta_0 + \beta_1 \ln(X_i + \Delta X) + u_i$$

$$\ln(Y_i + \Delta Y) - \ln(Y_i) = \beta_1 [\ln(X_i + \Delta Y) - \ln(X_i)]$$

$$\beta_1 = \frac{\Delta Y/Y}{\Delta X/X}$$

$\beta_1 \rightarrow$ elasticity

% Δ IN Y ASSOCIATED WITH A 1% Δ IN X

- ESTIMATED USING OLS
- HYPOTHESIS TEST + CONFIDENCE
- INTERVALS INTERPRETED AS USUAL
- WHEN TO USE LOG FORM?

LOG INCOME/WAGES → SKEWED DISTRIBUTION OUTCOMES

- PLOT RELATIONSHIPS
- DIAGNOSTIC TESTS

$$\ln(Y_i) = \beta_0 + \beta_1 D_i + u_i$$

$D_i = 0 \rightarrow \ln(Y_i | D_i = 0) = \beta_0 + u_i$

$D_i = 1 \rightarrow \ln(Y_i | D_i = 1) = \beta_0 + \beta_1 + u_i$

$$\beta_1 = \ln\left(\frac{Y_i \mid D_i = 1}{Y_i \mid D_i = 0}\right)$$

$\therefore \Delta$ IN Y WHEN $D_i = 0 \rightarrow D_i = 1$

HANDOUT ANSWERS:

Question 1.

Interpret the coefficients on the following regression results.

```
> data(CASchools)
>
> CASchools$ln.income<- log(CASchools$income)
> CASchools$ln.math <- log(CASchools$math)
> CASchools$str <- CASchools$students/CASchools$teachers
> CASchools$small <- 0
> CASchools$small[CASchools$str<20] <- 1
>
>
> mo1 <- lm(data=CASchools, formula=math ~ ln.income)
> mo2 <- lm(data=CASchools, formula=ln.math ~ income)
> mo3 <- lm(data=CASchools, formula=ln.math ~ ln.income)
> mo4 <- lm(data=CASchools, formula=math ~ small)
> mo5 <-  lm(data=CASchools, formula=ln.math ~ small)
>
> stargazer::stargazer(mo1,mo2,mo3,mo4,mo5, type='text')

================================================================================
                                        Dependent variable:
                    ----------------------------------------------------------
                          math             ln.math             math     ln.math
                          (1)        (2)        (3)        (4)        (5)
--------------------------------------------------------------------------------
ln.income              34.664***                0.053***
                       (1.610)                  (0.002)

income                            0.003***
                                  (0.0001)

small                                                     5.599***   0.008***
                                                          (1.830)    (0.003)

Constant              561.661***  6.440***   6.342***   650.156***  6.477***
                       (4.304)    (0.002)    (0.007)    (1.380)    (0.002)

--------------------------------------------------------------------------------
Observations              420        420        420        420        420
R2                        0.526      0.481      0.522      0.022      0.022
Adjusted R2               0.525      0.480      0.521      0.020      0.019
Residual Std. Error (df = 418)  12.928     0.021      0.020      18.570     0.028
F Statistic (df = 1; 418)  463.806*** 387.338*** 456.883***  9.364***   9.256***
================================================================================
Note:                                            *p<0.1; **p<0.05; ***p<0.01
>
```

(1): 1 PERCENT ↑ IN INCOME IS ASSOCIATED WITH A 0.347 POINTS INCREASE IN MATH SCORES.

(2): $1000 ↑ IN INCOME IS ASSOCIATED WITH A 0.3 % ↑ IN TEST SCORES

(3): 1 % ↑ IN INCOME IS ASSOCIATED WITH A 0.053 % INCREASE IN TEST SCORES.

(5): ON AVERAGE, THE SMALL CLASSROOMS PERFORM 0.8 % HIGHER.

WARNING!

-APPROXIMATION ONLY WORKS WHEN:

$\Delta Y/Y$ IS SMALL IF YOU WANT THE

EXACT CHANGE:

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

$$Y_i + \Delta Y_i = \beta_0 + \beta_1 (X_i + \Delta X_i) + u_i$$

$$= \beta_1 \Delta X_i$$

$$\ln\left(\frac{\Delta Y_i + Y_i}{Y_i}\right) = \beta_1 \cdot \Delta X_i$$

$$\ln\left(\frac{\Delta Y_i}{Y_i} + 1\right) = \beta_1 \cdot \Delta X_i$$

$\ln(x+1) \approx x$ when x is small

$$\approx \frac{\Delta Y_i}{Y_i} \equiv \%\Delta Y_i$$

ONLY MATTERS WHEN β_1 IS LARGE.

EXACT → $\%\Delta Y_i =$

$$\ln Y_{i0} = \beta_0 + \beta_1 X_{i0} + u_i \qquad X = \exp(\ln(x))$$

$$\ln Y_{i1} = \beta_0 + \beta_1 X_{i1} + u_i$$

$$\frac{\Delta Y_i}{Y_i} = \frac{Y_{i1} - Y_{i0}}{Y_{i0}} = \frac{Y_{i1}}{Y_{i0}} - 1$$

→ $\exp\left(\ln\left(\frac{Y_{i1}}{Y_{i0}}\right)\right) - 1$

→ $\exp(\beta_1 X_{i1} - \beta_1 X_{i0}) - 1$ → $\exp(\Delta X \beta_1) - 1$

IF $\hat{\beta} = 0.3 \rightarrow \% \Delta y = 0.95$

REGRESSION WITH A BINARY DEPENDENT VAR:

$Y = \{0,1\}$ $Y \rightarrow$ CONTINOUS

LATEST SCORES, TRAFFIC ~~FATALITY~~ RATE

GETTING INTO COLLEGE : $\{0,1\}$ SMOKING = $\{0,1\}$

OBESITY = $\{0,1\}$ MORGAGE APLICATION $\{0,1\}$

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

HOW TO INTERPRET $\hat{\beta}_1$ WHEN X_i IS CONTINOUS & Y_i IS BINARY.

LINEAR PROBABILITY MODEL:

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

$E[Y_i | X_i] = 1 \times PR(Y_i = 1 | X_i) +$

$0 \times PR(Y_i = 0 | X_i)$

$E[Y_i|X_i] = PR(Y_i|X_i)$

↳ HANDOUT : PROBABILITY OF DENIAL CONDITIONAL ON PI RATIO

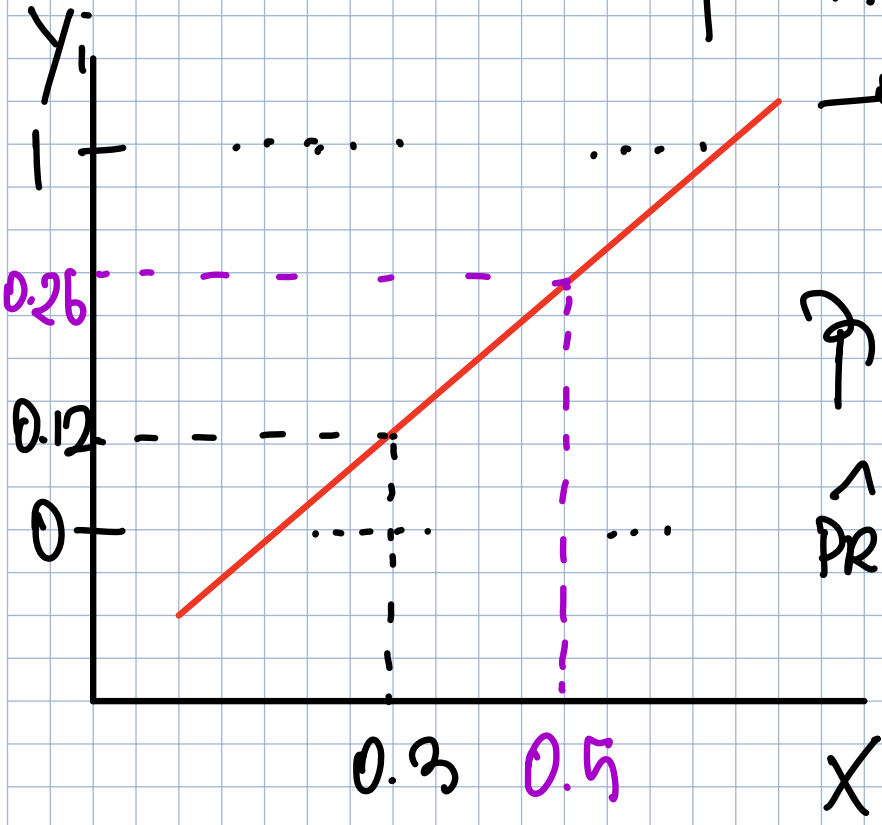
$Y_i = \beta_0 + \beta_1 X_i + u_i$

$E[Y_i|X_i] = PR(Y_i=1|X_i) = E[\beta_0 + \beta_1 X_i + u_i | X_i]$

$E[\beta_0] + E[\beta_1 X_i | X_i] + E[u_i|X_i] = 0$

$PR(Y_i=1|X_i) = \beta_0 + \beta_1 X_i$

$\widehat{PR}(Y_i=1|X_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i$ → SAMPLE



→ $\beta_0 + \beta_1 X_i$

$\beta_1 = \frac{\Delta Y}{\Delta X}$

$\widehat{PR}(Y_i=1|0.3) = 0.12$

$\widehat{PR}(Y_i=1|0.5) = 0.26$

$$\beta_1 = \frac{\widehat{Pr}(Y_i = 1 \mid 0.5) - \widehat{Pr}(Y_i = 1 \mid 0.2)}{0.2}$$

PROBABILITY OF DENIAL GOES UP BY 0.14 OR 14 PERCENTAGE POINTS AS THE PI RATIO INCREASED BY 0.2.

1° $-0.07991 + 0.4 \times 0.60353 = 16.2\%$

$-0.07991 + 0.3 \times 0.60353 = 10.1\%$

0.060353

IF PI ↑ BY 0.1

THE PROB. DENIAL GOES UP BY 6 PERCENTAGE POINTS.

2° $Y_i = \beta_0 + \beta_1 D_i + u_i$

β_0 β_1

$$E[Y_i \mid D_i] = Pr(Y_i = 1 \mid D_i) = \beta_0 + \beta_1 D_i$$

$Pr(Y_i = 1 | D_i = 0) = \beta_0$ → PROBABILITY OF DENIAL OF BEING NON-BLACK

$Pr(Y_i = 1 | D_i = 1) = \beta_0 + \beta_1$ → PROBABILITY OF DENIAL OF BEING BLACK

$\hat{Pr}(Y_i = 1 | D_i = 0) = 9.3\%$

$\hat{Pr}(Y_i = 1 | D_i = 1) = 28.4\%$

$\hat{\beta}_1 = 0.191$

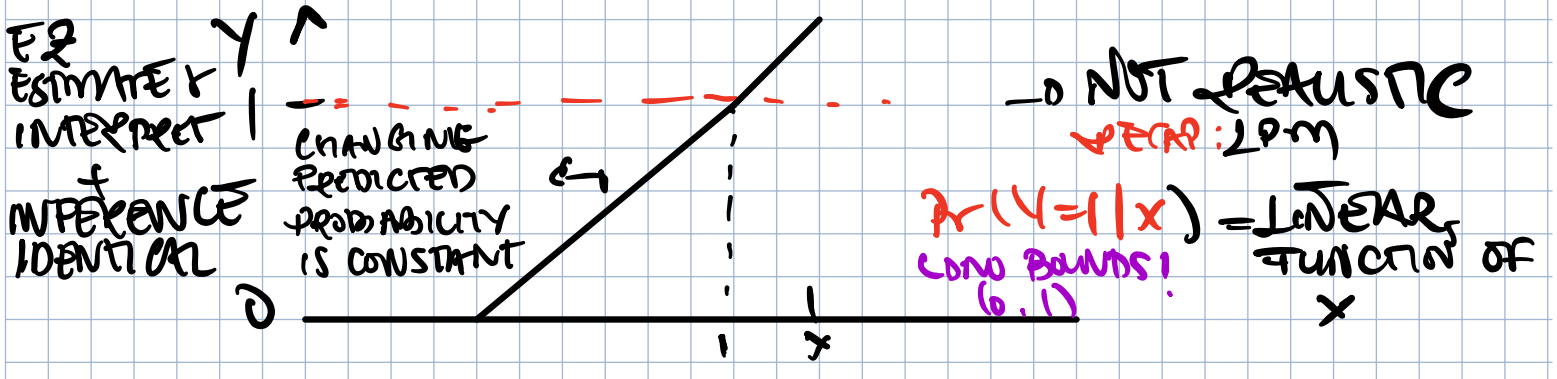
$\hat{\beta}_1 = 0.191$ → BEING BLACK IS ASSOCIATED WITH A 19.1 PP DIFFERENCE DECLINE IN MORTGAGE APPROVAL.

3rd MODEL: BEING BLACK IS ASSOCIATED WITH A 17.7 PP DECLINE IN APPROVAL HOLDING PI RATE CONSTANT.

$Pr(Y=1|x)$ INCREASING WITH x $\beta_1 > 0$

ISSUES:

$Pr(Y=1|x) > 1$ OR $Pr(Y=1|x) < 0$



EG ESTIMATE & INTERPRET INTERCEPT IDENTICAL

→ NOT REALISTIC WEEKLY HRS: 2 P-M

$Pr(Y=1|x)$ = LINEAR FUNCTION OF x

COND BOUNDS: (0, 1)