1 / 7 / 19

# QUANTITATIVE METHODS 1A

* -> EMAIL SCOTT + DESCRIBE MY QUANTITATIVE BACKGROUND
  TA: CAMERON SELLS

WED. LECTURES WHEN MONDAY IS A HOLIDAY

PROJECTS - WEEKS 7-10 (DATASET PRESENTATIONS)

GRADES: HW 50% ; FINAL 50%

RESOURCES: IDRE.UCLA.EDU/STAT
↳ COURSES ON R, STATA, PYTHON, ETC.

---

INDIVIDUALS - UNIT OF ANALYSIS

VARIABLE (FACTOR) - CHARACTERISTICS OF INDIVIDUALS

4 TYPES OF VARIABLES (STEVENS 1946)

- NOMINAL - GENDER, ETHNICITY, POLITICAL PARTY (UNORDERED, CATEGORICAL)
- ORDINAL - DISTANCE BTW. VALUES ISN'T CONSTANT ! BA, MA, PhD ; SPORTS RANKING
- INTERVAL - NO MEANINGFUL ZERO -> TEMPERATURE, TIME, DEMOCRACY SCORE
- RATIO - MEANINGFUL ZERO - AGE

DUMMY VARIABLE - BINARY, CATEGORICAL (0,1)

GRAPHS - PURPOSE: EXAMINE THE CHARACTERISTICS OF THE DISTRIBUTION OF A SINGLE VARIABLE.
↳ SEE TUFTE, THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION.
ILLUMINATE UNDERLYING PATTERNS OF INTEREST, HIGH DATA-TO-INK RATIO, NO CHART JUNK, NO CONFUSING.

"MAXIMUM LIKELIHOOD" ?          GOOGLE NAPOLEON'S ARMY GRAPH "CARTE FIGURATIVE"

GET TUFTE - THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION.

PIE CHARTS - DRAWBACKS: DIFFICULT TO SEE ZERO
HARD TO DISTINGUISH BTW. CATEGORIES

BAR CHARTS - IF X-AXIS IS NOMINAL, THEN GAPS AREN'T MEANINGFUL.

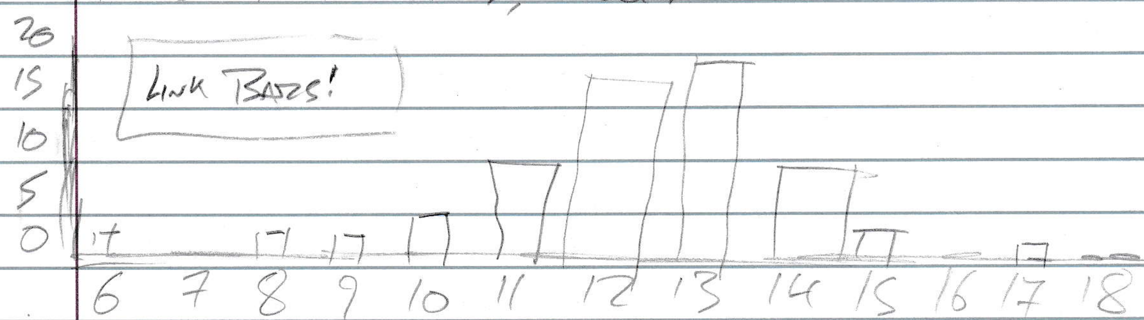
* STEM + LEAF PLOTS - CHOOSE THE STEM (TENS, ONES, ETC)

REVIEW

| | | | |
|---|---|---|---|
| 1 | 3 3 4 | : 13, 13, 14 | |
| 2 | 6 2 | : 26, 22 | PRESERVES RAW DATA |
| 3 | 5 9 | : 35, 39 | |
| 4 | 9 | : 49 | |
| TENS | ONES | | |

DISTINGUISH BETWEEN GRAPHS THAT ARE USEFUL AND THOSE THAT ARE MERELY TECHNICALLY CORRECT.

HISTOGRAM: PICK BINS, ALLOCATE DATA TO BINS




DESCRIBING A GRAPH: CENTER, SHAPE, SPREAD, OUTLIERS

VISIT DOG BEACH IN DEL MAR

1 / 7 / 17

QUANTITATIVE VARIABLES

MEAN : $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$

↳ "SUFFICIENT STATISTIC"
↳ USES ALL DATA POINTS

$X_i$ = INDIVIDUAL OBS.
$n$ = # OF OBS.
$\bar{X}$ = AVERAGE OF $X$

MEDIAN : MIDDLE VALUE (OR AVG. OF 2 MIDDLE VALUES)
↳ LESS VULNERABLE TO OUTLIERS "INSUFFICIENT STATISTIC"

MODE : MOST COMMON ENTRY (REPEATED VALUES)
↳ DEPENDS ON

ALL OF THE ABOVE ARE MEASURES OF CENTRAL TENDENCY

| | | | |
|---|---|---|---|
| | NOMINAL | MODE | NOMINAL |
| MEDIAN | ORDINAL | | ORDINAL |
| MEAN | INTERVAL | | INTERVAL |
| | RATIO | | RATIO |

VARIANCE : $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

STANDARD DEVIATION : $\sigma = \sqrt{\sigma^2}$
("ROUGHLY" THE AVERAGE DISTANCE FROM THE MEAN)

RANGE : $X_M - X_1$

IQR : $X_{75Q} - X_{25Q}$ (INTER-QUARTILE RANGE)

VARIANCE CAN OFTEN BE MORE SIGNIFICANT THAN AVERAGES
(EX. - POLARIZATION)

* THE SAMPLE MEAN IS THE VALUE THAT MINIMIZES THE VARIANCE.

QUANTILES: THE ith QUANTILE OF A VARIABLE IS THE VALUE SUCH THAT i% ARE SMALLER THAN THAT VALUE AND 100 - i% ARE GREATER THAN THAT VALUE

SKEWNESS:

KURTOSIS:

BOXPLOTS:

CROSS TABS: IDENTIFY CAUSE + EFFECT

1 / 14 / 19

# Quantitative Methods Week 2

Probability - the long-run frequency of events
Theme - using simulations to explore probability
↳ Many important results are reached through simulation

Reasonable Doubt: "The probability of seeing a result this far (or further) from the null hypothesis, if in fact the null is true."

Focus: We want to know the probability that our result happened randomly - but we never know for sure.
(But we can replicate.)

Types of Probability: "Classical" ⇒ Equally likely outcomes
Mathematical
Empirical (Frequentist)
Subjective (Bayesian)

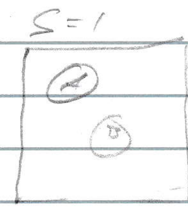
"Count + Divide" → Divide # events by # possible events

Mathematical: 3 Axioms (Kolmogorov)
Given events $A_i$, event space $S$
(See slides)

FREQUENTIST : THE PROBABILITY OF AN EVENT IS ITS LONG-RUN RELATIVE FREQUENCY.

$P(E)$ IS THE PROBABILITY OF EVENT $E$.

SUBJECTIVE : "BELIEFS" OF PROBABILITY (BAYESIAN)




S - SAMPLE SPACE : $S = s_1, \ldots s_n$

TRIAL : RANDOM DRAW

OUTCOME : RESULT

INTERSECTION : $P(A \cap B) = \wedge =$

UNION : $P(A \cup B) = \vee =$

CONDITIONAL : $P(A|B)$

$P(\varnothing) = 0$

$P(A') = P(A^c) = 1 - P(A) \Rightarrow$ COMPLEMENT

$P(A \cap B) = 0$ = MUTUALLY EXCLUSIVE

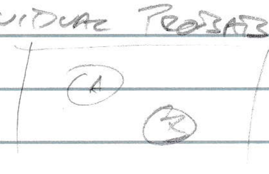
* IF TWO EVENTS ARE MUTUALLY EXCLUSIVE, THE PROB. THAT ONE OR THE OTHER WILL OCCUR IS THE <u>SUM</u> OF THE PROBABILITIES THAT EACH OCCURS.

$P(A \cup B) = P(A) + P(B)$

* IF EVENTS ARE MUTUALLY EXCLUSIVE + COLLECTIVELY EXHAUSTIVE, THE SUM OF THEIR PROBABILITIES MUST ADD TO ONE.

* INDEPENDENCE: IF TWO EVENTS ARE INDEPENDENT, THE PROBABILITY OF BOTH A + B OCCURRING AS THE PRODUCT OF THEIR INDIVIDUAL PROBABILITIES.

$P(A \cap B) = P(A)P(B)$




MUTUALLY EXCLUSIVE

1 / 14 / 19

GET A 6-SIDED DIE W/ 3.5 (EXPECTED VALUE) ON ALL 6 SIDES.

IF EVENTS ARE NOT INDEPENDENT, THEN:

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

CALCULATING SIMPLE PROBABILITIES

FINITE SAMPLE SPACE: THE PROBABILITY OF ANY $s_i$ IS $\frac{1}{n}$. $P(s) = \frac{1}{n}$

GIVEN SOME EVENT A (SUBSET OF S) $s_1 \ldots s_m$ ($m < n$)

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

PROBABILITY DISTRIBUTIONS: DISCRETE + CONTINUOUS (SEE SLIDE)

EXPECTED VALUE: "AVERAGE" AFTER MANY TRIALS

$$E(x) = \sum (x_i P(x_i))$$

VARIANCE:

$$\sum (x_i - E(X))^2 \times P(x_i)$$

BERNOULLI DISTRIBUTION: (COIN TOSS)

$$E(\text{TRIAL}) = 1(P) + 0(1-P) \quad (\text{HEADS} = 1)$$

$$V(\text{TOSS}) = P(1-P)$$

RANDOM VARIABLE K - THE SUM OF INDEPENDENT BERNOULLI TRIALS

$$\boxed{\binom{N}{K} P^K (1-P)^{n-K}}$$

$$\binom{N}{K} = \frac{n!}{K!(n-K)!} = \frac{n(n-1)(n-2)\ldots}{K(K-1)\ldots \times (n-K-1)\ldots}$$

CUMULATIVE DISTRIBUTION FUNCTIONS

INSTEAD OF $P(X = x_i)$ WE EXAMINE $P(X \leq x_i)$

PDF $\Rightarrow$ PROB. OF SEEING EXACTLY X

CDF $\Rightarrow$ PROB. OF SEEING LESS THAN X

DISJOINT EVENTS CAN BE ADDED INTO A CDF

CENTRAL LIMIT THEOREM - AS WE INCREASE N,

NORMAL APPROXIMATION:

$$\left.\begin{aligned} &P(K \geq 40) = 1 - \psi(Z) \\ &Z = \frac{40 - 20}{\sqrt{\frac{.5(1-.5)}{41}}} = 256.12\ldots \end{aligned}\right\} \begin{aligned} &\text{SO:} \\ &P(K \geq 40) = \\ &1 - \psi(256) \end{aligned}$$

GEOMETRIC DISTRIBUTION

$\hookrightarrow$ USEFUL FOR PART FAILURE (TIME TO FAILURE)

HYPERGEOMETRIC DISTRIBUTION

$$P(K = k) = \frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}}$$

K = THE NUMBER OF SUCCESSES IN n DRAWS W/O REPLACEMENT FROM A POPULATION OF SIZE N, CONTAINING R SUCCESSES AND N-R FAILURES

1/14/17

POISSON DISTRIBUTION: NUMBER OF EVENTS IN A DISCRETE TIME PERIOD. (OFTEN INFREQUENT).

$$P(K=x) = \frac{\lambda^{x} e^{-\lambda}}{x!}$$

CONTINUOUS DISTRIBUTIONS:

SINCE THE PROBABILITY OF ANY GIVEN (CONTINUOUS) POINT IS ZERO, WE HAVE TO TALK ABOUT DENSITY FUNCTIONS INSTEAD OF PROBABILITY FUNCTIONS.

UNIFORM DISTRIBUTION

WE INTEGRATE THE PDF TO GET THE CDF

PDF - PROB. DENSITY FUNCTION $= f(x)$

CDF : CUMULATIVE DENSITY FUNCTION $= P(X < x) = F(x)$

$$= \int f(x)\,dx$$

$$E(X) = \int x \cdot (f(x)\,dx)$$ (UNIFORM DISTRIBUTION)

$$V(X) = \int (x - E(x))^2 \cdot (f(x)\,dx)$$

EXPONENTIAL DISTRIBUTION:

$$f(x) = \lambda e^{-\lambda x},\ x \geq 0$$

$$F(x) = 1 - e^{\lambda x},\ x \geq 0$$

$$F(x) = 0 x < 0$$

$$E(X) = \lambda^{-1} \quad V(X) = \lambda^{-1}$$

BETA DISTRIBUTION

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(x^{\alpha - 1} (1-x)^{\beta - 1}\right)$$

IF $\alpha = \beta$, THEN DIST. IS CENTERED ON .5, IF NOT THEN SKEWED

$$E(X) = \frac{\alpha}{\alpha + \beta} \quad V(X) = \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$$

# THE CENTRAL LIMIT THEOREM

THE MEAN OF A SAMPLE OF INDEPENDENT DRAWS FROM THE SAME DISTRIBUTION WITH EXPECTED VALUE $\mu$ AND FINITE VARIANCE $\sigma^2$ WILL BE NORMALLY DISTRIBUTED WITH AN EXPECTED VALUE OF $\mu$ AND A VARIANCE OF $\frac{\sigma^2}{n}$, OR STANDARD DEVIATION OF $\frac{\sigma}{\sqrt{n}}$

THIS MEANS THAT AS LONG AS OUR VARIABLE "LOOKS SOMETHING LIKE A MEAN", WE CAN USE THE NORMAL DIST

NORMAL DISTRIBUTION:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)}$$

(MAXIMIZED WHEN $x = \mu$)

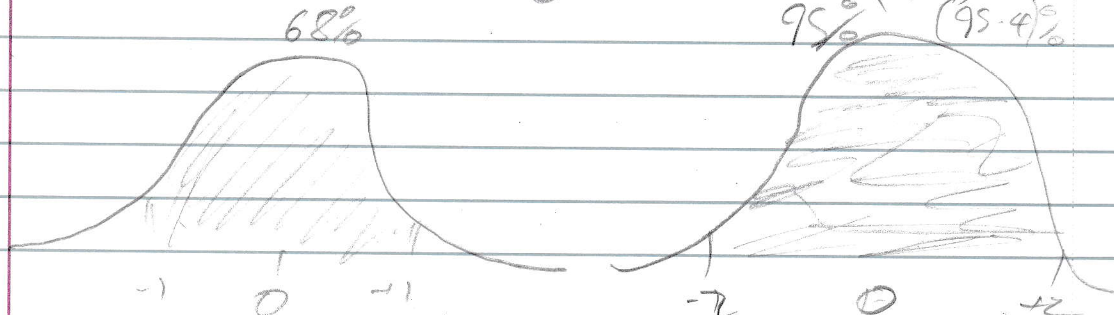
$$E(x) = \mu \qquad V(x) = \sigma^2 \qquad \Phi(Y)$$

$$E(X+Y) = E(X) + E(Y)$$
$$E(cX) = cE(X)$$

IF INDEPENDENT: $V(X+Y) = V(X) + V(Y)$

$$V(cX) = c^2 V(X)$$

$$X \sim N(\mu, \sigma^2) \; ; \; Z = \frac{X-\mu}{\sigma} \; ; \; Z \sim N(0,1)$$



1/16/19

# CHI - SQUARE DISTRIBUTION

SUM OF MANY NORMAL DISTRIBUTIONS (SQUARED)
ALWAYS ≥ 0
APPROACHES NORMAL AS WE ADD MORE NORMAL Ds.

# F DISTRIBUTION

RATIO OF CHI-SQUARE DISTRIBUTIONS

$$K_1 \sim \chi^2_{d_1}$$
$$K_2 \sim \chi^2_{d_2}$$

$$G = \frac{K_1/d_1}{K_2/d_2}$$

$$G \sim F_{d_1, d_2}$$

1/23/17

# QUANTITATIVE METHODS WEEK 3

2 REVOLUTIONS IN SOCIAL SCIENCE
1) EXPERIMENTS
2) CAUSAL EFFECTS
} DRAMATICALLY RAISE BAR

NULL HYPOTHESIS - INNOCENT UNTIL PROVEN GUILTY ($H_0$)
ALTERNATIVE HYPOTHESIS - GUILTY BEYOND A REASONABLE DOUBT ($H_A$)

WHAT IS THE PROBABILITY OF SEEING A SAMPLE LIKE THE ONE WE HAVE, IF THE NULL HYPOTHESIS WERE TRUE?
WE CAN: REJECT NULL
OR FAIL TO REJECT NULL

HYPOTHESIS TESTING:
P-VALUE: THE PROBABILITY OF SEEING DATA AS EXTREME OR MORE EXTREME THAN OUR SAMPLE.
IF THE P-VALUE IS REALLY SMALL, WE PRESUME THAT THE RESULT ISN'T JUST RANDOM ERROR: REASON TO REJECT THE NULL.

ALPHA ($\alpha$): THE PROBABILITY (P-VALUE) BELOW WHICH WE REJECT THE NULL. TYPICAL: .05
CAN BE MANIPULATED, AVOID VIA PRE-REGISTRATION

Z-SCORE: $Z = \frac{X - \mu}{\sigma}$
STEPS:
1) SET $\alpha$
2) STATE NULL, ALTERNATIVE HYPOTHESES
3) CHECK P-VALUE AGAINST CRITICAL VALUE (OFTEN REQUIRES CALCULATING Z-SCORE)
4) CONCLUDE

IF Z-SCORE ≥ 2, WE'LL REJECT THE NULL AT A .05 CONFIDENCE INTERVAL

ERROR: TYPE I (FALSE POSITIVE)
TYPE II (FALSE NEGATIVE)

NON-PARAMETRIC TESTS - MAKE FEWER ASSUMPTION

POWER: THE POWER OF A TEST IS THE PROBABILITY OF CORRECTLY REJECTING THE NULL WHEN IT IS FALSE.

1/28/17

# 204B - Week 4

Simple Bivariate Regression (OLS)

a) Minimize the sum of vertical squared deviation from the line.

b) Produce unbiased + efficient estimates

$Y = \alpha + \beta x + \epsilon$

↳ Hypotheses involve $\beta$ (Does X affect Y / is Beta 0?)

Regressions give you (approximately) the maximum likelihood + Bayesian + least-squares result.

X: Cause (independent)   Y: Effect (dependent)

Relationship: Type, Direction, Strength (accuracy of prediction)

↳ Positive/negative

Dangerous to project beyond the range of the data

Outliers need not deviate from the trend

Outliers exert a relatively stronger effect than other data points.

Correlation Coefficient: A single number that measures the strength + direction of a straight-line relationship: $-1 \leq r \leq 1$

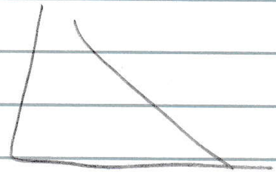
(Only looks at straight lines, vulnerable to outliers)

$r = \frac{\sum Z_{xi} Z_{yi}}{n-1}$ where $Z_{xi} = \frac{(x_i - \bar{x})}{s_x}$

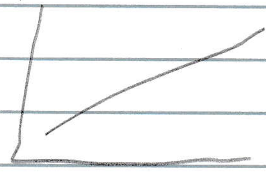
$Z_{yi} = \frac{(y_i - \bar{y})}{s_y}$

OR: $r = \dfrac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{s_x s_y (n-1)}$

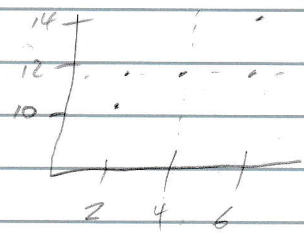
$r = -1$



$r = +1$



START BY RE-SCALING X + Y TO Z-SCORES.
THAT CREATES FOUR QUADRANTS.




$n = 5$

| | X | Y | $(X - \bar{X})^2$ |
|---|---|---|---|
| | 2 | 10 | 4 |
| | 2 | 12 | 4 |
| | 4 | 12 | 0 |
| | 6 | 12 | 4 |
| | 6 | 14 | 4 |
| MEAN: $\bar{X}$ | 4 | 12 | 16 |

$$S_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{16}{4}} = 2$$

| $(Y - \bar{Y})^2$ |
|---|
| 4 |
| 0 |
| 0 |
| 0 |
| 4 |
| 8 |

$$S_y = \sqrt{\frac{\sum (y_i - \bar{y}_i)}{n-1}} = \sqrt{\frac{8}{4}} = \sqrt{2}$$

THEN CALCULATE Z-SCORES

$Z_x$ $\quad Z_Y$ $\quad Z_x * Z_y$

1/28/19

# LINEAR REGRESSION

$Y = \alpha + \beta x + \epsilon \quad = \quad y = \beta_0 + \beta_1 X + \epsilon$

$\alpha$: INTERCEPT $(Y = 0, x = A)$

MINIMIZE THE SUM OF SQUARED VERTICAL DEVIATIONS FROM THE BEST-FIT LINE.

$e$ = ERROR TERM (DEVIATIONS FROM THE LINE)

$y_i = a + bx_i + e_i$

$\sum_{i=0}^{n} e_i^2$ (MINIMIZE THIS QUANTITY)

$e_i = y_i - a - bx_i$

$\sum_{i=0}^{n} (e_i)^2 = \sum_{i=0}^{n} (y_i - a - bx_i)$

(THIS SAYS THE SUM OF THE ERROR TERMS IS THE SAME AS THE SUM OF THE EQUATION)

$b = r * \frac{s_y}{s_x}$

# LEAST-SQUARES REGRESSION

$\alpha = \beta_0$ = INTERCEPT
$\beta = \beta_1$ = SLOPE

| PARAMETER | ESTIMATES |
|---|---|
| $\beta_0$ | $\hat{\beta}$ or $b_0$ |
| $\beta_1$ | $\beta_1$ or $b_1$ |
| $\sigma$ | |

VERTICAL DISTANCE TO THE REGRESSION LINE:
RESIDUALS - THE PART OF Y WE DIDN'T EXPLAIN

$\epsilon$ = TRUE ERROR (UNKNOWN)
$e$ = ESTIMATE (RESIDUAL)

MODEL: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
ESTIMATE: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$

THE SUM OF SQUARED VERTICAL DEVIATIONS IS <u>ALWAYS</u> TOO SMALL (B/C THAT'S WHY WE CHOSE THE LINE).

$a = \bar{Y} - b\bar{x}$ $\quad$ $\underline{\hat{Y} = a + bx}$ $\quad$ $\underline{Y_i - \hat{Y}_i}$ (ACTUAL MINUS PREDICTED)

$b = r \frac{s_y}{s_x}$

$S_e = \sqrt{\frac{\Sigma e^2}{n-2}}$

$\left( S_e = \sqrt{\frac{\Sigma e^2}{n-2}} \right)$

1 / 28 / 19

MEASURING THE FIT OF THE MODEL:

NULL HYPOTHESIS = X DOESN'T AFFECT Y; $\beta = 0$

TO MEASURE THE FIT, COMPARE THE RESIDUALS FROM BOTH MODELS:

$$\sum e^2_{REGRESS} = \sum (y_i - \hat{y}_i)^2$$
$$\sum e^2_{NULL} = \sum (y_i - \bar{Y})^2$$

OR: PARTITION THE VARIANCE IN Y INTO THE PART WE HAVE EXPLAINED BY X (REG, TSS) WITH THE PART THAT WE HAVEN'T EXPLAINED

$$RSS = \sum (y_i - \hat{Y}_i)^2$$
$$REG\,SS = \sum (\hat{y} - \bar{Y})^2$$
$$TSS = \sum (y_i - \bar{Y})^2$$

$$\frac{\text{EXPLAINED VARIANCE}}{\text{TOTAL VARIANCE}} = \frac{REG\,SS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (y_i - \bar{Y})^2} = R^2$$

$$\frac{REG\,SS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = R^2$$

$R^2$ IS THE PERCENTAGE OF VARIANCE IN Y EXPLAINED BY X.

REGRESSION IS UNBIASED, EFFICIENT, CONSISTENT

1/29/19

# REGRESSION BY HAND

FORMULAE: $S_x = \sqrt{\frac{\Sigma(x-\bar{x})}{n-1}}$ $S_y = \sqrt{\frac{\Sigma(y-\bar{y})}{n-1}}$

$r_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{S_x S_y (n-1)}$ $b = \frac{r_{xy} S_y}{S_x}$ $a = \bar{y} - b\bar{x}$

$a = \bar{y} - b\bar{x}$ $\hat{y} = a + bx$ $\hat{\sigma} = SE = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}}$

STEPS: 1) CALCULATE STANDARD ERRORS

a) FIND $\bar{X}$

b) SUBTRACT $\bar{X}$ FROM X

c) SQUARE RESULTING PRODUCT

d) SUM THE COLUMN OF SQUARES

SAME 4 STEPS

e) FIND $\bar{Y}$

f) SUBTRACT $\bar{Y}$ FROM $Y$

g) SQUARE RESULTING PRODUCT

h) SUM THE COLUMN OF SQUARES

i) MULTIPLY $(x-\bar{x})(y-\bar{y})$

j) CALCULATE STANDARD