Probability & Statistics "Review"

(*) Goal: use data to answer economic questions

↳ almost never can we observe & measure every relevant data point

→ If we're interested in the effect of each year of schooling on income, it's extremely unlikely we'd get to observe every single person's income, much less years of schooling.

↳ And even that wouldn't be the ideal data set.

↳ Statistical Inference — allows us to use a sample of the overall population

→ But in order to do this we need to go down the rabbit hole of theoretical probability & statistics for day.

(*) We often model real world outcomes using random variables.

↳ A variable whose possible values are outcomes of random phenomena.

Examples:

- Your grade in a given class is modeled as a Normal random variable
- The number of baskets you can make is modeled as a Binomial random variable.
- The number of inputs available to a firm is modeled as a Poisson random variable.

(*) The probability of each value the RV can take on is called the probability mass (density) function

→ sometimes you can enumerate each outcome & it's probability.

Example: Dice

X = the result of the roll of a single fair die.

$\sim$ Uniform(1/6)

x	$P(X=x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

→ Sometimes you can't & you have to use a function.

Example: The time until you recieve your next email is modeled as an expon. random var.

$x \in [0, \infty)$

pdf: $f(x) = \lambda e^{-\lambda x} = P(X = x)$

(*) Since it isn't always particularly illuminating to write down a pdf & look at it to understand what's going on with a random variable,

↳ we often talk about:

- the expected value: $E[x]$
  (the value we expect the variable to take on)
- the variance: $V[x]$
  (how spread out the variable's values are)

Discrete Random Variables:

$$E[x] = \sum_{x=x} x\, P[X=x]$$

$$V[x] = \sum_{x=x} (x - E[x])^2 P[X=x]$$

Continuous Random Variables:

$$E[X] = \int_{x \in I} x f(x)\, dx$$

$$V[X] = \int_{x \in I} (x - E[x])^2 f(x)\, dx$$

Examples:

X_1 = Result of a single roll of a fair die.
(varname)

$E[X_1]$? (t)

$$E[X_1] = \sum_{X_1=x} x\, P[X_1 = x]$$ (Always write the formula)

$$= 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6)$$

$$= 1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6$$

$$= 21/6 = \underline{3.5}$$

$V[X_1]$? (i)

$$V[X_1] = \sum_{X_1=x} (x - E[X_1])^2 P[X_1 = x]$$

$$= (1-3.5)^2(1/6) + (2-3.5)^2(1/6) + (3-3.5)^2(1/6)$$
$$+ (4-3.5)^2(1/6) + (5-3.5)^2(1/6) + (6-3.5)^2(1/6)$$

$$= (6.25)(1/6) + (2.25)(1/6) + (0.25)(1/6)$$
$$+ (0.25)(1/6) + (2.25)(1/6) + (6.25)(1/6)$$

$$= 2.91\overline{6}$$

X_2 = the time until your next email

$\sim Exp(\lambda)$

$$E[X_2] = \int_{x \in \Sigma} x f(x)\,dx$$

$$= \int_0^\infty x \lambda e^{-\lambda x}\,dx = \lambda \int_0^\infty x e^{-\lambda x}\,dx$$

$$= \frac{e^{-\lambda x}(\lambda x + 1)}{\lambda}\Bigg|_0^\infty$$

$$= \lim_{a \to \infty} \frac{e^{-\lambda a}(\lambda a + 1)}{\lambda} - \frac{e^{-\lambda(0)}(\lambda(0) + 1)}{\lambda}$$

(1) (0) (+1)

$$= \lambda \left[\lim_{a \to \infty} e^{-\lambda a}(\lambda a + 1)\right] - \frac{(1)(1)}{\lambda}$$

L'Hopital

$$= \frac{1}{\lambda} - 0 = \frac{1}{\lambda}$$

Remember These E.V. Rules:

$E[c] = c$

$E[aX] = aE[X]$

$E[aX + bY] = aE[X] + bE[Y]$

(*) It's one thing to say I think student SAT scores are distributed according to some Distribution with mean μ & variance σ^2

$$Y \sim F(\mu, \sigma^2)$$

→ it's a whole other thing to say what we think μ is.

→ so we need to estimate μ using data

(*) An estimator is a function of observed data used to estimate a statistical parameter

Ex. We observe a bunch of SAT scores

$y_1, y_2, y_3, \cdots, y_n$ } each a realiz. of indep. random vars $Y_i \sim F(\mu)$ "iid"

that's mean estimator

$$\hat{\mu} = f(y_1, y_2, \cdots, y_n)$$

→ There are all sorts of estimators we could. We could use

$$\hat{\mu}' = \min\{y_1, y_2, \cdots, y_n\}$$

↳ It wouldn't be a very good one but it would be a one.

(*) An unbiased estimator is one whose expectation is the true parameter value.

→ those little y_i's are realizations of random variables

↳ functions of random variables are themselves random variables

↳ estimators are random variable & have expected values

(*) An estimator, $\hat{\mu}$, of a true parameter μ, is unbiased if

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

(*) The estimator of μ we usually use is the sample mean, $\bar{x}$.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Is $\bar{x}$ an unbiased estimator of μ.

→ need to show $E[\bar{x}] = \mu$.

$$E[\bar{x}] = E\left[\frac{1}{n}(y_1 + y_2 + \dots + y_n)\right]$$

$$= \frac{1}{n} E[y_1 + y_2 + \dots + y_n]$$

$$= \frac{1}{n}\left(\underbrace{E[y_1]}_{\mu} + \underbrace{E[y_2]}_{\mu} + \cdots + \underbrace{E[y_n]}_{\mu}\right)$$

$$= \frac{1}{n}\left(\underbrace{\mu + \mu + \cdots + \mu}_{n}\right)$$

$$= \frac{n\mu}{n} = \mu \quad \text{Yes!}$$

(*) $\bar{x}$ is itself a random variable & it has a distribution.

Central Limit Theorem:

The sample mean is distributed according to Normal Distribution with mean μ and variance σ^2/n

$$\bar{x} \sim N(\mu, \sigma^2/n)$$

the associated standard deviation, $\sigma/\sqrt{n}$ is called the standard error.

(*) Thing standard error is not.

- sample standard deviation, s.
- the sample standard error, $s/\sqrt{n}$

(*) Reminder of how to calculate s:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$s = \sqrt{s^2}$$

(*) Keeping all our s's straight:

Population Variance: σ^2
↳ Std. Deviation: σ
Sample Variance: s^2
↳ Sample Std. Deviation: s

Population Standard Error (theoretical): $\sigma/\sqrt{n}$
Sample Standard Error (usually what you have): $s/\sqrt{n}$

T-P-S

Book price data
→ differences in prices across gender?
→ genre play a role?

Hypothesis Testing

Suppose we average the SAT (Math) scores of all the students in our class, and get 620. (& std. dev. = 15)

↳ That's higher than the SMC average (610).

↳ That difference could be just due to randomness in sampling. (youre just ~20 realizations of the Saint Mikes SAT random variab

↳ Or it could be that the Economics majors' SAT scores are really higher than average.

(*) Hypothesis Testing is a process that allows us to say whether some characteristic of the data we observe (mean, std. dev, min, max, etc.) are likely due to to randomness or not.

↳ First you specify a null hypothesis (usually the thing you think isn't true ↳ trying to show isn't true)

$\bar{x} = \mu$
$s = \sigma$
$\beta = 0$

$AVG(SAT)_{ECON} = 610$

→ The Hypothesis test tells you whether it's likely (you get to pick your defn of likely) you'd see this data if the null hypothesis is true.

Null Hypothesis: $H_0: \bar{X}^{SAT}_{ECON} = 610$

Alternative Hyp: $H_a: \bar{X}^{SAT}_{ECON} \neq 610$

num obs. = 20 → 16 → t-distribution

Test Statistic: $\dfrac{\bar{X}^{SAT}_{ECON} - 610}{s/\sqrt{n}} \sim t_{15}$

$= \dfrac{620 - 610}{15/\sqrt{20}} = \dfrac{10}{15/4} = \dfrac{10}{15/4}$

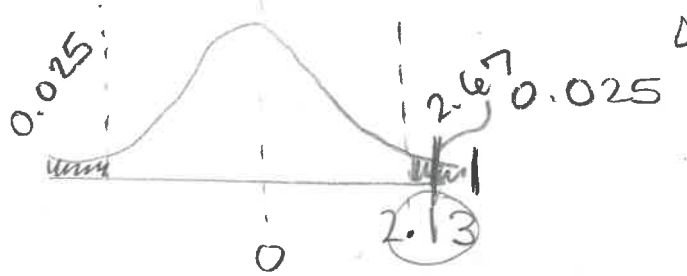
$= \dfrac{620 - 610}{15/\sqrt{16}} = \dfrac{40}{15} = 2.67$

Significance Level (How likely is likely).

$\alpha = 0.05$

$df = 15 = n - 1$

$2.67 > 2.13$



Three Ways
1. Critical Value
2. P-Values
3. Confidence Intervals

Reject the Null

↳ There is a less than 5% chance that we would see this data AND the null be true

(P) μ = the true population mean SAT Math score for SMC students
= 610

s = sample std. dev = 15

$\bar{x}$ = sample mean SAT math score (sampled from some Econ majors)
= 620

α = significance level
= 0.05

~~n = 20~~ n = 16

(H) $H_0: \bar{x} = \mu$ $H_a: \bar{x} \neq \mu$

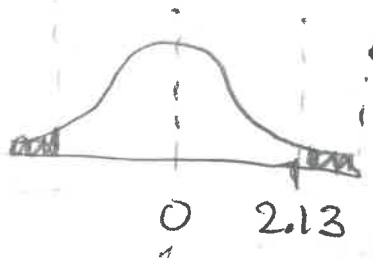
(A) $\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{15}$

(N) Two-sided t-test at a 5% sig. level

(T) test statistic $= \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{620 - 610}{15/\sqrt{16}}$

$= 2.67$

(O)



(M) Reject the null hyp that the sample mean SAT score equals the popul. mean.

(*) Now let's use our data on book prices to do a difference of means test.

85
171 - 86 = 86

→ Question: Is there a statistically significant difference between the price of books by female authors and the price of books by male authors?

→ First: use our own indiv[illegible]s

→ Then: use our combi[illegible]

Open Spreadsheet in STATA

(P) μ_1 = true mean pric[illegible]
female authors

μ_2 = true mean pri[illegible]
male authors

$\bar{x}_1$ = sample mean [illegible]
by female au[illegible]

$\bar{x}_2$ = sample mean price of books by male authors = 15.64

$df = (n_1 - 1) + (n_2 - 2) = 168$ $\quad \alpha = 0.05$ $\quad s_1 = 5.20$

$n_1 = 84$ $\quad n_2 = 86$ $\quad s_2 = 6.48$

(H) $H_0: \mu_1 = \mu_2$ OR $\mu_1 - \mu_2 = 0$ $\quad H_a: \mu_1 - \mu_2 \neq 0$

(A) $$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{168} \downarrow_Z$$

$$df = \overset{84-1}{(n_1 - 1)} + \overset{86-1}{(n_2 - 1)}$$
$$83 + 85$$
$$= 168$$

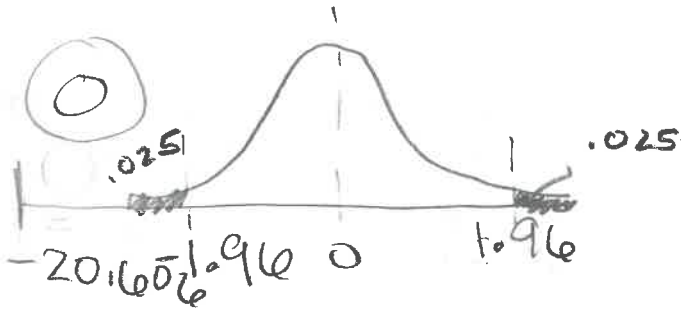
(N) Two-Sided Difference of means (indep) t test

(T) test statistic: $\dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

$$= \frac{\bar{x}_1-\bar{x}_2}{s_p\sqrt{\frac{1}{84}+\frac{1}{86}}} = \frac{14.81-15.64}{s_p\sqrt{0.012+0.012}} = \frac{-0.83}{0.26\sqrt{0.024}} = \frac{-0.83}{0.04} = -20.604$$

$$s_p = \sqrt{\frac{s_1+s_2}{(n_1-1)+(n_2-1)}} = \sqrt{\frac{5.20+6.48}{(83+85)}} = \sqrt{\frac{11.68}{168}} = \sqrt{0.0} = 0.2$$

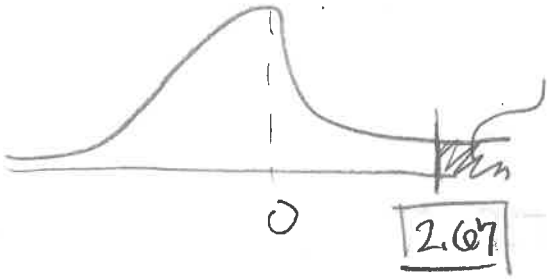
(O)



(M) Reject the null that $\mu_1-\mu_2=0$

(*) Now do it again with the ~~big data set~~ your data set

(*) The p-value tells you exactly how likely it is.



? = .01

→ There is a 1% chance that we would see this data if the null were true.

(*) Mnemonic Device for Hypothesis Testing

P: Parameters ($\mu, \sigma, \beta, \bar{x}, s, n, \alpha$)
H: Hypotheses
A: Assumptions
N: Name the test
T: Test Statistic
O: Outcome
M: Make Your Conclusion

→ Lets redo that test using the mnemonic.

→ Hypothesis Testing is one way to decide if our data says something statistically significant.
  ↳ confidence intervals are another

Hypothesis Tests Answer: "Does this particular number constructed from this particular sample indicate statistic significance at the x% significance level?"

Confidence Intervals Answer: "Suppose I have a sample of n observations, what range of numbers would indicate statistical significance at the x% significance level."

(We almost always use 5%)

Ingredients for a Confidence Interval:

- sample statistic
- critical value (based on α)
- Std. error

construct the CI. Does my null hypothesis fall in it?

$$\mu = \bar{x} \pm t_\alpha \left(s/\sqrt{n} \right)$$

sample statistic | critical value | sta. error

$$(\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) \pm t_\alpha s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

"I am $(1-\alpha)\%$ confident / There is a $(1-\alpha)\%$ chance that the true parameter lies in this interval."

(*) Let's build a 95% Confidence Interva for the SAT Math Scores. (Together)

Question: What range of numbers can I be 95% sure contains the true mean SAT Math score for SMC students?"

$$\mu = \bar{x} \pm t_{0.025} \, s/\sqrt{n}$$

$$= 620 \pm (2.13)(15/\sqrt{16})$$

$$= 620 \pm (2.13)(15/4)$$

$$= 620 \pm 7.99$$

$$= [612.01, 627.99]$$

Is 610 in that interval? No.

↳ Reject the null at 5%.

(*) Build a 95% CI for the difference in book prices using the big data set. (Independent?)

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm t_{.025} \, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$[\quad] = (14.81 - 15.64) \pm 1.96(0.04)$$

$$= -0.83 \pm 0.078$$

$$= [-0.908, -0.752]$$

* Does the confidence interval contain 0? No

  → Reject the null

* These are all ways to answer "Does this data tell us something or not?"

  ↳ The rest of the class is concerned with defining the "something."

* Build a 95% Confidence Interval for the true difference in prices

Math Tutoring Center

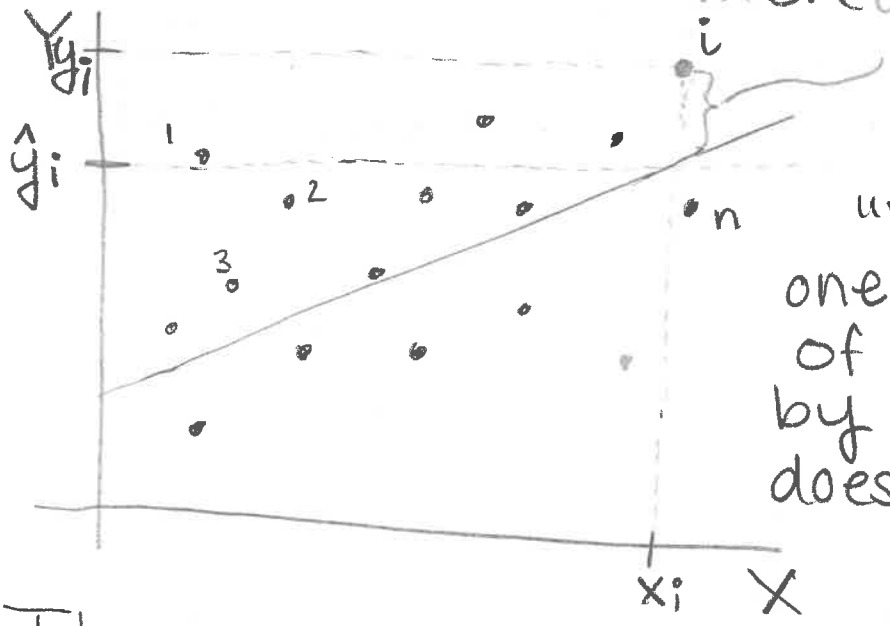
Fitting a Line to Data

One Sided Formula Sheet

* Deal with things early (phys. & mental)
* Mostly focused on the statistics of a single variable.
  - ↳ But most of our questions are about at least two variables.

→ First, plot your data.

* "When one variable increases, by how much does the other increase or decrease?"



$y_i - \hat{y}_i$

"When I get one more year of education, by how much does my starting salary increase?"

→ The most common way to condense this information in an informative way is to fit a line through it.

* Why a line?
  - easy interpretation ($\frac{\text{rise}}{\text{run}}$)
  - not a lot to be gained by making it more complicated.

(*) Today: the mechanics of doing a good job fitting a line through sampled data

Next Class: using that line to say things about the population.

(*) You will need:

- an independent variable (x) (years of educ
- a dependent variable (Y) (income
- a calculator (optional)

→ There are lots of ways to pick a line (once there's more than 2 points)

$Y = a + bX$

→ Definitely want to minimize some concept of "distance" between the line & each point.

→ The one we usually use is the sum of the squared distance between the predicted y-value ($\hat{Y}_i$) and the actual y-value (Y_i) for each x-observation (X_i).

the y-val on the line

→ each dot is a observation
$i = \{1, \ldots, n\}$

(*) We're looking for a line

$$\hat{Y}_i = a + bX$$

that minimizes the sum of the squared differences between the y_i's & $\hat{y}_i$'s for each x_i.

That is

$$\min_{\{a,b\}} \sum_{x_i} (Y_i - \hat{Y}_i)^2.$$

→ We're going to actually do this, but let's rewrite something's first.

→ there are going to be a lot of $Y - \bar{Y}$ & $X - \bar{X}$, so

- $y = Y - \bar{Y}$
- $x = X - \bar{X}$
- $\hat{Y}_i = a + b X_i + b\bar{X} - b\bar{X}$

$$= \underbrace{(a + b\bar{X})}_{\tilde{A}} + bX_i - b\bar{X}$$

($a + b\bar{X}$ doesn't depend on i)

$$= \tilde{A} + b(X_i - \bar{X}) = \tilde{A} + bx_i$$

→ so now the task is

$$\min_{\{a,b\}} \sum_i \left(Y_i - \tilde{A} - bx_i\right)^2$$

(*) Calculus Reminder: To minimize or maximize something, you take the derivative & set it = 0.

→ Don't be intimidated by the $\sum_i$. (You can always write it out)

$$S = (Y_1 - \tilde{A} - bx_1)^2 + (Y_2 - \tilde{A} - bx_2)^2 + \ldots + (Y_n - \tilde{A} - bx_n$$

To find $\tilde{A}$ (d a): chain rule!

$$\frac{\partial S}{\partial \tilde{A}} = \sum_i 2(Y_i - \tilde{A} - bx_i)^1(-1) = 0$$

$$\sum_i (Y_i - \tilde{A} - bx_i) = 0$$

$$\sum_i Y_i - \sum_i \tilde{A} - b\sum_i x_i = 0$$

① ②

Show that $\sum(x_i - \bar{x}) = 0$

① $\sum_{i=1}^{n} \tilde{A} = \underbrace{\tilde{A} + \tilde{A} + \ldots + \tilde{A}}_{n} = n\tilde{A}$

② $b\sum_{i=1}^{n} x_i = b\left[\sum_{i=1}^{n}(x_i - \bar{x})\right] = b\left[n\bar{x} - n\bar{x}\right] = 0$

$$\sum_i Y_i - n\tilde{A} - 0 = 0$$

$$\sum_i Y_i = n\tilde{A}$$

$$\frac{\sum_i Y_i}{n} = \tilde{A}$$

$$\bar{Y} = \tilde{A} = a + b\bar{X}$$

$$\boxed{\bar{Y} - b\bar{X} = a}$$

need this guy now.

To find b:

$$\frac{\partial S}{\partial b}: \sum_i 2(Y_i - \tilde{A} - bx_i)'(\not x_i) = 0 \quad (\text{div. by } -2)$$

$$\sum_i (x_i Y_i - x_i \tilde{A} - b x_i^2) = 0$$

$$\sum_i x_i Y_i - \tilde{A}\sum_i x_i - b\sum_i x_i^2 = 0$$

$$\sum_i x_i = 0$$

$$\sum_i x_i Y_i - b\sum_i x_i^2 = 0$$

$$\sum x_i Y_i = b\sum x_i^2$$

$$\boxed{\frac{\sum_i x_i Y_i}{\sum x_i^2} = b}$$

no you can't combine the Σ's.

$$b = \frac{\sum_i x_i Y_i}{\sum_i x_i^2} \qquad a = \bar{Y} - \left(\frac{\sum_i x_i Y_i}{\sum_i x_i^2}\right)\bar{X}$$

→ these are them! These are the linear regression coefficients.

↳ we have a lot to do to show that they're good for anything (they are).

↳ but for now we have them

(*) What the slope coefficient does and doesn't mean:

→ in an ideal world we would randomly assign X's (we would randomly assign levels of education & then measure income)

↳ but most of the time (especially in economics) that is not at all possible

- if by some miracle you have randomly assigned data

  b = the increase in Y caused by a 1 unit increase in X.

- if on the other hand if you do not have randomly assigned dream data

b = the increase in Y associated with a one unit increase in X.

→ it includes the confounding factors associated with X (family background, grades, study habits, etc.).

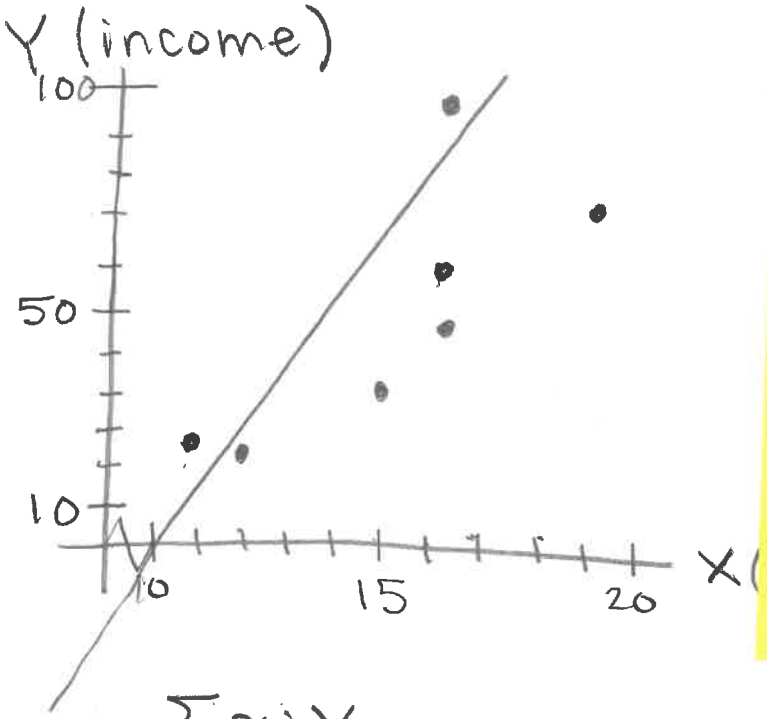
(*) Example: ① Calculate the regression line for the following data. (& plot it) ② Interpret the slope coefficient. ③ What are some potentially confounding variables?

x_i (yrs. of ed.)	Y_i (income in thous.)	$x_i - \bar{x}$	x_i^2 $(x-\bar{x})^2$	$x_i Y_i$ $(x-\bar{x})Y$
11	26	-3.86	14.90	-100.36
12	25	-2.86	8.18	-71.5
15	32	0.14	0.02	4.48
16	65	1.14	1.30	74.1
16	45	1.14	1.30	51.3
16	100	1.14	1.30	114
18	75	3.14	9.86	235.5
$\Sigma =$ 104	368		36.86	307.52

$\bar{x} = 14.86$

$\bar{Y} = 52.57$

(1)



Minimizing the squared errors used to seem like magic. How does this eq do that?? Not magic. Calculus.

$$b = \frac{\sum_i x_i Y_i}{\sum(x_i^2)} = \frac{307.52}{36.86} = 8.34$$

Always write the formula!

$$a = \bar{Y} - b\bar{X} = 52.57 - (8.34)14.86$$
$$= 52.57 - 123.93$$
$$= -71.36$$

$$\hat{Y} = -71.36 + 8.34X$$

$0 = -71.36 + 8.34x$
$71.36 = 8.34x$
$8.56 = x$

(2) There is a $8,340 increase in income associated with one extra year of schooling.

(3)
- Whether you finish a degree
- school rank
- mother/father income

(*) Estimating Y's.

→ once you have an equation for Y in terms of X, you can use it to estimate Y for different X's.

Ex: Using the regression line we estimated, calculate your expected income for the number of years of education you have.

Mine is:

$$\hat{Y}_{\text{Prof.Luedtke}} = -71.36 + 8.34(21)$$

(175.14)

$$= 103.78$$

The Simple Regression Model

→ Last time we focused on the mechanics of creating a regression line from sample data

↳ today we're going to write down a mathematical model of the economy that allows us to use that to make statistical statements about the real world.

(*) Assumptions

→ all models of the world come with assumptions (otherwise you'd just be like "sometimes stuff happens" & go get a coffee)

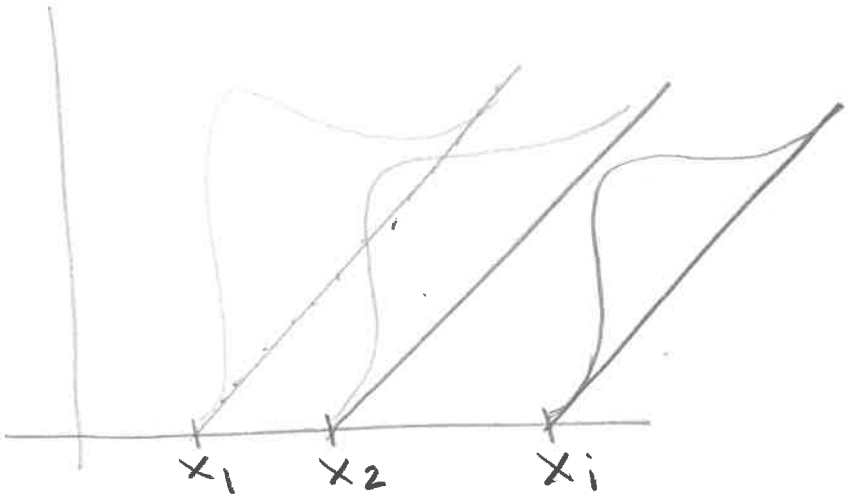
→ my advisor used to tell us: "All models are wrong. Some models are useful."

↳ Economists (or anyone doing mathematical modeling) always have to balance

- too many assumptions → not realistic
- too few assumptions → can't say anyth. useful

Simple Regression Model Assumptions

(*) The distributions of every $Y_i | x_i$ all have the same variance



→ for every x, there's a distribution of Y_i's possible (due to the randomness of life)

↳ these assumptions describe how we think about those distrib. & that randomness.

1. Homogeneous Variance:

→ all of those distributions have the same variance.

$$V[Y_1|X_1] = V[Y_2|X_2] = \ldots = V[Y_n|X_n] = \sigma^2$$

$$V[Y_i|X_i] = \sigma^2 \quad \forall i$$

2. Linearity

→ the expected value/mean of each distribution is a linear function of X.

↳ it's, infact, the same linear function of x.

$$E[Y_1|X_1] = \underline{\underline{\alpha}} + \underbrace{\beta}X_1$$

$$E[Y_2|X_2] = \underline{\underline{\alpha}} + \underbrace{\beta}X_2$$

(review subscripts)

$$E[Y_n|X_n] = \underline{\underline{\alpha}} + \underbrace{\beta}X_n$$

$$\boxed{E[Y_i|X_i] = \underline{\underline{\alpha}} + \underbrace{\beta}X_i \quad \forall_i}$$

3. Independence

→ Each Y_i is statistically independent from every other Y_i.

↳ Knowing Y_i gives you no information about Y_j.

→ these can all be stated in one sentence:

The random variables, $Y_1, Y_2, \ldots, Y_n$, are independently distributed with mean $\alpha + \beta X_i$ and variance σ^2.

→ most of the time we talk about the distribution of the error term rather than Y.

$$E[Y_i|X_i] = \alpha + \beta X_i$$

$$Y_i = \alpha + \beta X_i + u_i$$

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

$$V[u_i|X_i] = \sigma^2$$

show that if ther

→ these are equivalent.

Population Regression Line:

$$E[Y_i|X_i] = \alpha + \beta X_i$$

or

$$Y_i = \alpha + \beta X_i + u_i$$

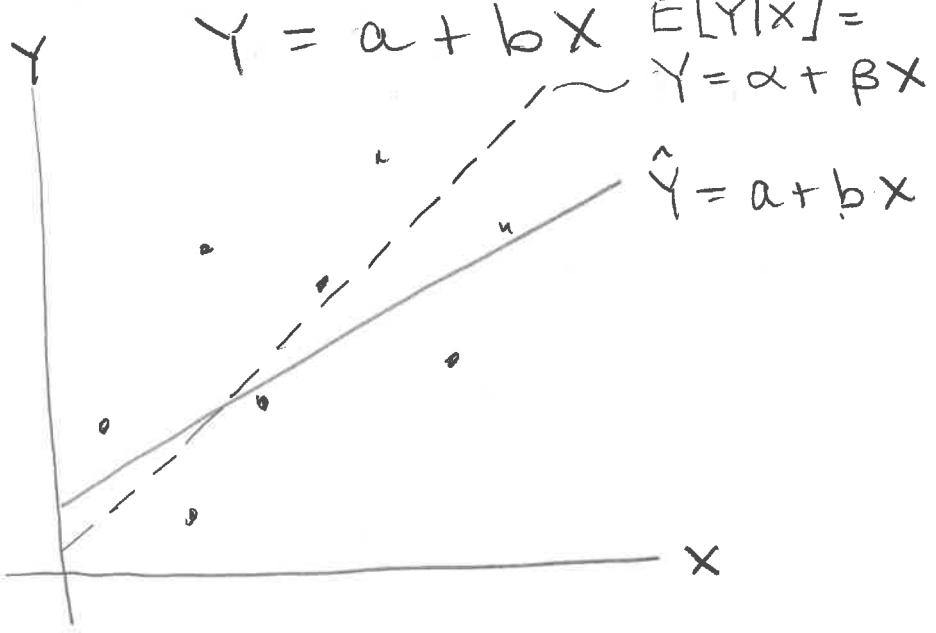
this part is the line

the true data fall around the line-

the assumpt. are about buildin the populatio line

Sample Regression Line:

$$\hat{Y} = a + bX$$



(*) Most of our interesting economic questions boil down to questions about β (not b).

→ is it > 0

→ is it < 0

→ is it just $\neq 0$ (there is <u>some</u> relationship)

<u>Bad News</u>: We really don't know what β is.

<u>Good News</u>: We definitely know what b is.

↳ We don't just know what it is, we know its distribution.

(*)

The slope estimate, b, is normally distributed with mean β and variance $\frac{\sigma^2}{\sum x^2}$.

→ this is it. This is why there is a rest of the course and we don't just pack up & go home.

(*) Estimating σ^2

→ since we almost never know σ^2

↳ we have to do what we always do: approximate the distribution using a sample estimate for s^2 and the t-distribution

$$s^2 = \frac{1}{n-2} \sum_i (Y_i - \hat{Y}_i)^2$$

↑ because we did a and b

$$SE = \frac{s}{\sqrt{\sum x^2}}$$

→ so the distribution we actually use to do anything is

t_{n-2} with mean β and variance $\frac{s^2}{\sum x^2}$.

→ now we can use all those statistical tools (hypothesis testing, confidence intervals, etc.)

if time

1. Perform a Hypothesis Test using the data from last class. Test the null hypothesis that $\beta = 0$.
2. Construct a 95% Confidence Interval for β.

(*) ~~Next time, part of class will be for a review session. → your job to bring questions~~

(*) Read Stata Tutorial

! For both of those it will ... to construct s^2 & SE:

Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	$\hat{u}_i$ $(Y_i - \hat{Y}$
26	20.38	5.62	31.58
25	28.72	−3.72	13.84
32	53.74	−21.74	472.6
65	62.08	2.92	8.53
45	62.08	−17.08	291.73
100	62.08	37.92	1437.93
75	78.76	−3.76	14.14
			2270.38

$\hat{Y} = -71.36 + 8.34X$

$\sum x^2 = 36.86$

$n - 7$

$n - 2 = 5$

$s^2 = \frac{1}{5}(2270.38)$

$= 454.08$

$$SE = \frac{21.31}{\sqrt{36.86}} = \frac{21.31}{6.07} = 3.51$$

2. Confidence Interval:

$$\beta = b \pm t^{n-2}_{.025} SE$$
$$= 8.34 \pm (2.57)(3.51)$$
$$= 8.34 \pm 9.02$$

$$[-0.68, 17.36]$$

→ Since 0 is in the confidence interval, we fail to reject the null hypothesis that $\beta = 0$.

(*) You should be comfortable performing hypothesis tests using all methods (t-score, p-value, CI) for β.

1. Hypothesis Test

(P) β = the slope coeff. of the true regression line.

b = the slope coeff. of the sample regression line

$n = 7$

$\alpha = 0.05$

(H) $H_0: \beta = 0$ $H_a: \beta \neq 0$

(A) $b \sim N(\beta, \frac{\sigma^2}{\sqrt{x^2}})$

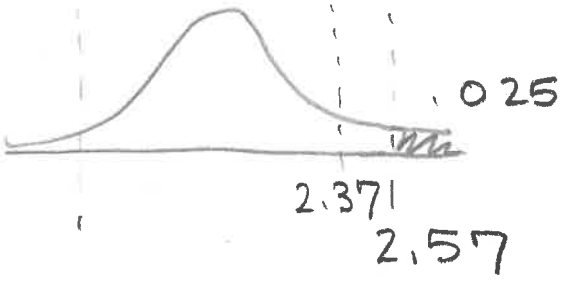
↳ use t-distribution
$df = n - 2 = 5$

(N) Two-Sided t-test

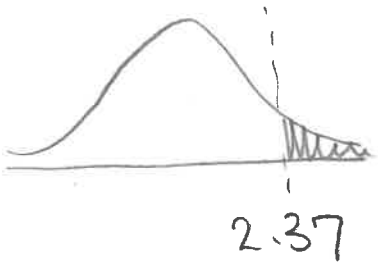
(T) $t = \frac{b - 0}{SE} = \frac{8.34}{3.51} = 2.37$

(O)

t-score:



p-value



between 2.02 & 2.57
↳ $p \in (.05, .025)$

(M) Fail to Reject null that $\beta = 0$.

Goodness-of-Fit

(*) So far, we don't have any way of saying if the regression line is any good at explaining Y.

↳ We could regress height on favorite color and get a number.

↳ But how do we know whether your favorite color actually explains any of the variation in height? (It doesn't.)

→ to fix this, we're going to calculate R^2.

(*) R^2 tells us the fraction of sample variation in Y explained by X.

→ No more.

→ No less.

(make sure you understand what things like R^2 do and do not say)

→ To calculate R^2, then, we need some measures of variation.

1. Total Sum of Squares (SST)

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

(note: different from sample variance b/c didn't divide by $n-1$)

→ This measures the total sample variation in Y. (It's our denominator.)

→ how spread out is Y? If it's super spread out, all else eq., X will do a worse job, low R^2

→ there are two things that affect Y: X & other stuff

2. Explained Sum of Squares (SSE)

$$SSE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$\bar{\hat{Y}}$

→ This is the total variation of the estimated $\hat{Y}$'s.

↳ These are the estimates of Y based entirely on X.

→ If these are also very spread out, we're doing ok.

→ a lot of statistical programs will just tell you SSR.

↳ So it's useful to write R^2 in terms of SSR.

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

We know this because

$$R^2 = \frac{SSE}{SST}$$

$$SST = SSE + SSR$$

$$SST - SSR = SSE$$

$$R^2 = \frac{(SST - SSR)}{SST} = \frac{SST}{SST} - \frac{SSR}{SST}$$

$$= 1 - \frac{SSR}{SST}.$$

(*) R^2 measures exactly what we said: the fraction of total variation in Y explained by x.

→ some very informative regressions have low R^2

→ some useless regressions have very high R^2

3. Residual Sum of Squares (SSR)

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2 \left(= \sum_{i=1}^{n} (\hat{u}_i - \bar{u}_i)^2 \right)$$

$$= \sum_{i=1}^{n} (\hat{u}_i - 0)^2$$

→ This is the sample variation in the OLS residuals $(Y_i - \hat{Y}_i)$

$$= Y_i - (a + bx_i)$$

$$\boxed{SST = SSE + SSR}$$ (I encourage you to prove this.)

(*) Now we can define R^2

$$\boxed{R^2 = \frac{SSE}{SST}} \quad \frac{\text{variation in Y expl. by x}}{\text{total var. in Y}}$$

→ sometimes it's <u>way</u> more of a pain in the buns to calculate

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2$$

↳ you have to calculate all the $\hat{y}_i$'s & subtract, & square.

→ doesn't mean it isn't a useful measure.

↳ It measures a very specific thing perfectly.

Example: Calculate the R^2 for the regression we did of income on years of schooling.

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

→ so let's be strategic about this: Which formula should we use?

↳ either way, we need SST (sadly we didn't calculate the variance of Y yet)

→ But last time, we calculated $(Y_i - \hat{Y}_i)^2$! That's $\hat{u}_i^2$!

↳ So we have SSR! = 2270.38

Y_i	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
26	-26.57	705.96
25	-27.57	760.10
32	-20.57	423.12
65	12.43	154.50
45	-7.57	57.30
100	47.43	2,249.60
75	22.43	503.10
		4853.68

$\bar{Y} = 52.57$

good way to check.
SST > SSR
SSE

don't forget the 1 −

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

$$= 1 - \frac{2270.38}{4853.68}$$

$$= 1 - 0.47 = 0.53$$

(*) 53% of the variation in income is explained by years of schooling.

Stata

→ depends on computer situation

cps78_85.dta

reg lwage educ

↑ log of wage

① Write down the regression line

$\widehat{lwage} = 0.98 + 0.069\ educ$

OR

lwage = 0.98 + 0.069 educ +

② Is the coeff. on educ stat. sig.?

3 Ways:

1. Critical Value 12.08 > 1.96
2. P-value .000
3. Confidence Interval [0.058, 0.080]

③ R^2? 0.12

12%

Multiple Regression

Simple Regression Eq: $Y = \alpha + \beta x + u$

→ probably the sketchiest assumpt. we made in the Simple Regression Model was that everything contained in u is uncorrelated with x.

$$E[u|x] = 0.$$

(*) If we're looking at a regression of wage on years of educ.

$$wage = \alpha + \beta educ + u$$

→ It's not a problem that there are othe things that affect wage

↳ we may just not be interested in them

→ What <u>is</u> a problem is if those things are correlated with education

(↓) What types of things might be in that u term?

→ things that affect wage that aren't education?

$u \longrightarrow$ { experience, gender, family income, race, marital status, ⋮ }

→ Do you find it believable that any of those things aren't correlated with years of education?

↳ I can tell you there almost no married women of color from poor backrounds in PhD programs.

(*) One of the assumptions of the Simple Regression Model fails.

↳ need a different model.

Multiple Regression Model

→ Just like the SRM but instead of just the one x_j there are several x's.

New Terminology/Conventions:

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

same ↑ → there are k explanatory variables. $(X_1, X_2, \ldots, X_k)$ ↑ same

→ α becomes β_0, just so we don't go crazy w/ our greek letters

$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ looks nice.

→ there are now k coefficients: (slope)

$\beta_1, \beta_2, \ldots, \beta_k$

↳ (k+1) total parameters:

$\beta_0 + (\beta_1, \beta_2, \ldots, \beta_k)$

→ Each of these k explanatory vars has n observations (for each observation person)

↳ Helpful to think of it like a matrix or an spreadsheet.

	educ X_1	exper. X_2	gender X_3	...	marital status X_k
person 1	10	4	m		S
person 2	•	•	•		•
person n	•	•	•		•

n x k matrix

(*) You need to be very comfortable with the regression equation & the structure of the data

↳ What's k? What's n? What does a given number in the matrix tell you?

(*) So how do we actually calculate the estimates for those β's?

↳ What we started out interested in is β_1.

↳ We may or may not be interested in β's $2 \longrightarrow k$.

↳ Need to include them to get (*) the right β_1.

(*) We're still going to use Ordinary Least Squares (OLS)

↳ minimize the sum of squared errors.

$$\min \left\{ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right\}$$

just the same

What makes up $\hat{y}_i$ is different.

$$\min_{\beta_0, \beta_1, \ldots \beta_k} \left\{ \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \ldots - \beta_k x_k \right)^2 \right.$$

→ the process is the same;

you don't need to actually do all this to tell me how you would. We'd take the derivatives, set them to 0, do algebra until our eyes bleed.

→ But now we have k+1 equations.

↳ that sounds like a job for a computer!

↳ Stata!

Command: (email link?)

use http://fmww.bc.edu/ec-p/data/wooldridge/wage1

regress wage educ exper female nonwhite married

(use same order as me)

(*) What is the regression equation?

$\hat{wage} = -1.78 + 0.58\,educ + 0.01\,exper - 2.07\,female - 0.03\,nonwhite + 0.16\,married$

(*) Interpreting the Coefficients

→ The β estimates ($\hat{\beta}$'s) give us the ceteris paribus (other things equal) effects of each X_i.

ΔX_i = the change in $X_i = X_i^a - X_i^b$

$$\Delta \hat{Y} = \hat{Y}^a - \hat{Y}^b$$

$$= (\hat{\beta}_0 + \hat{\beta}_1 X_1^a + \hat{\beta}_2 X_2^a + \ldots + \hat{\beta}_k X_k^a)$$
$$- (\hat{\beta}_0 + \hat{\beta}_1 X_1^b + \hat{\beta}_2 X_2^b + \ldots + \hat{\beta}_k X_k^b)$$

$$= \hat{\beta}_1 (X_1^a - X_1^b) + \hat{\beta}_2 (X_2^a - X_2^b) + \ldots + \hat{\beta}_k (X_k^a - X_k^b)$$

$$= \hat{\beta}_1 \Delta X_1 + \hat{\beta}_2 \Delta X_2 + \ldots + \hat{\beta}_k \Delta X_k$$

"What is the change in $\hat{Y}$ associated with a one unit change in X_1 holding other factors fixed?"

→ Other factors fixed?

$\Delta X_2 = 0$

$\Delta X_3 = 0$

one unit change in X_1?

$\Delta X_1 = 1$

→ plug those in.

$$\Delta Y = \hat{\beta}_1(1) + \hat{\beta}_2(0) + \dots + \hat{\beta}_k(0)$$

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

→ The calculus version is much easier

$$\frac{\partial \hat{Y}}{\partial X_1} = \frac{\partial(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k)}{\partial X_1}$$

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

(*) The point of MR is that even when we don't have "other factors fixed" data, we can still make "other factors fixed" conclusions

---

Stata

(*) What is the change in wage associated with one more year of education?

$$\hat{\beta}_1 = 0.58$$

$.58 an hour

(*) Run a simple regression of wage on educ.

↳ What is the effect here?

$\hat{\beta} = 0.54$

(*) Which estimate do you believe more, why?

$\hat{\beta}$, because <u>all</u> those other factors are left in u in the simple regression

↳ assumptions fail

(*) You need to be comfortable using generic variables and specific, contextual variables.

Properties of Multiple Regression

β's are true parameters, b's are estimators

(*) What makes a good estimator?

- unbiased $E[b_i] = \beta_i$
  ↳ kind of a bare minimum
- simple straightforward function of the y's.
- low variance
  ↳ not too spread out

⟶ Let's discuss the assumptions of the model & see how the estimator do.

Assumption #1: Y is a linear function of the β's

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + u$$

(Linear here means the power on each β_i is 1, as opposed to $\beta_i^2 X_i$.)

Assumption #2: The data $x_1, x_2, \ldots, x_k$ for each i are the result of a random sample.

Assumption #3: There are no perfect linear relationships between any combinations of x's.
↳ no perfect collinearity.

Assumption #4: $E[u \mid x_1, x_2, \ldots x_k] = 0$ (u: true error)

(*) Keep these last two straight.

#3: what relationships can & can't exist between the x's & each other.

#4: what relationships can & can't exist between the x's and u.

(*) Theorem: Under the assumptions #1–4, the OLS estimators are unbiased. That is,

$$E[b_0] = \beta_0$$

$$E[b_1] = \beta_1$$

$$E[b_k] = \beta_k$$

(*) What happens if the X's you use to estimate Y are different from the "true" determinents of Y?

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

true relationship

Case 1: You estimate

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3 + \hat{u}$$

(you include too many variables)

Case 2: You estimate

$$\tilde{Y} = \tilde{b}_0 + \tilde{b}_1 X_1 + \tilde{u}$$

(you include too few)

Case 1: Irrelevant Variables

We can write the true model as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

but $\beta_3 = 0$

→ all the β_i's are variable specifying numbers about the real world.

→ We could write down an equation for Y that included all the irrelevant x's & their β's that =0 but we don't because it would be silly (& impossible)

Let's check our assumptions.

#1: Linear in β's? ✓

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

#2: Random Sample? ✓

→ including X_3 has no effect on this

#3 No perfect collinearity? ✓

→ no effect

#4: $E[u|X_1, X_2, X_3] = 0$? ✓

→ Nothing about including X_3 necessitates a violation of these assumptions.

↳ the theorem holds!

↳ all the b's are unbiased!

Specifically $E[\hat{b}_3] = \beta_3 = 0$

Case 2: Omitted Variables

The real equation is

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

but you estimate

$$\tilde{Y} = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \tilde{u}$$

→ now $\tilde{u} = \beta_2 X_2 + u$

→ If X_2 is not correlated with X_1 then we're fine

$$\tilde{Y} = \tilde{\beta} + \tilde{\beta}_1 X_1 + \tilde{u}$$

satisfies the Simple Regression model assumptions & we can go about our business.

→ But we noted last time, that's usually not true (think education and family back round).

$$E[x_2|x_1] \neq 0.$$

(*) How does this violate the MR assuptions?

$$E[\tilde{u}|x_1] = E[\beta_2 x_2 + u | x_1]$$

$$= E[\beta_2 x_2 | x_1] + E[u|x_1]$$

$$\neq 0$$

→ So we can't conclude $\tilde{\beta}_1$ is unbiased.

(*) We need one more assumption (about the variance of u) to conclude these estimators for β_i's are the best ones to use.

Assumption #5: The error term, u, has the same variance for any values of the x's.

↳ homoskedasticity

⟶ So whether educ = 10 or 18, u has the same variance

$$V[u|18, f, mar.] = \sigma^2$$

$$V[u|10, m, single] = \sigma^2$$

Gauss-Markov Theorem:

Under Assumptions #1-5, the estimato $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, computed using OLS, for $\beta_0, \beta_1, \ldots, \beta_k$ are the Best Linear Unbiased Estimators.

Best: Lowest Variance

Linear: Can be written as a linear functior of y's

Unbiased: $E[\text{estimators}] = \underset{\text{values}}{\text{true}}$

(*) What the thm says: if we compute estimates for β_i's (call them $\hat{\beta}_i$) using OLS, they will be "linear" estimator that are unbiased. Among "linear" & unbiased estimators for β_i's, these have the smallest variance.

(*) What it's good for: the method for computing our estimates for the β coefficients is not only practical (it's "linear"), and reasonable (unbiased), but it's the least noisy (lowest variance) one we might reasonably use.

Avoiding Symbol Overload

(*) A lot of moving parts:

- variables (Y)
- data (Y_i)
- parameters (β)
- estimators (b's)

(*) The "True" / Theoretical Model (variables, parameters)

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

no hats.

↳ hats mean estimators

(variables, estimators)

(*) The equation we estimate

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{u}$$

hats mean estimate or estimator (squiggles too)

(*) The estimation method we use:
(estimators, data)

OLS

↳ produces specific estimators using data (an actual numbers)

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$ (simple reg version)

→ write down the wage, educ, etc. model out in this way.

- the true model
- the equation we estimated
- the actual estimates of the $\hat{\beta}_i$'s.

Multiple Regression: Inference

→ "Inference" just refers to formal hypothesis testing

↳ what can we actually infer from what we have?

(*) Typically questions about β's

- Is $\beta_i = 0$? (today)
- Are $\beta_1 = \beta_2 = \beta_3 = 0$ together? (next time

→ Need probability distribution of β's to do this

↳ the distribution of the β's depends on the distribution of $Y|X$ ("Linear")

↳ which depends on the distribution of u

Assumption #6: The population error, u, is independent of the explanatory var $x_1, \ldots, x_k$, and is distributed $N(0, \sigma_u^2)$.

Theorem: Under assumptions #1–6, conditional on X's,

$$b_j \sim N(\beta_j, \sigma^2_{\beta_j})$$ almost always use a $\hat{\sigma}^2_{b_j}$.

→ the variance of $\hat{\beta}_j$ is a complicated formula (described in the book)

$$\sigma^2_{\beta_j} = \frac{\sigma^2_u}{SST_j(1-R_j^2)}$$

σ^2_u: u's variance

SST_j: total variance in x_j

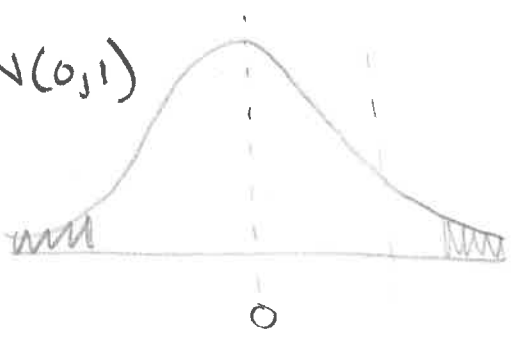
R_j^2: R^2 of a regression of x_j on all the other X's.

↳ so all of that is insane which is why we ask Stata to do it.

What the theorem says: under the assumptions we've outlined, the $\hat{\beta}$'s we estimate have this specific distribution.

What it's good for. We can use our normal distr hyp. testing tools to say things about the β's!

$$\frac{\hat{\beta}_j - \beta_j}{\sigma^2_{\hat{\beta}_j}/n} \sim N(0,1)$$



→ just one more step to take this from a math problem to an implementable process.

* As (almost) always: we don't actually know $\sigma^2_{\hat{\beta}_j}$.

↳ we don't know σ^2_u

↳ we assumed it into existence (it's probably a reasonable assumption that $u \sim N(0, \sigma^2_u)$ but we certainly don't know what σ^2_u is)

$$V[u] = E[(u - E[u])^2]$$

definition of var

$$= E[(u-0)^2] = E[u^2]$$

→ closer! but we don't observe u either

$$\frac{1}{\square} \sum_{i=1}^{n} \hat{u}_i^2$$

$n - (k+1)$

↑ num obs

↑ num parameters to estimate

$$\hat{\sigma}^2_{\hat{\beta}_j} = \frac{\frac{1}{n-(k+1)} \sum_{i=1}^{n} \hat{u}_i^2}{SST_j (1 - R_j^2)}$$

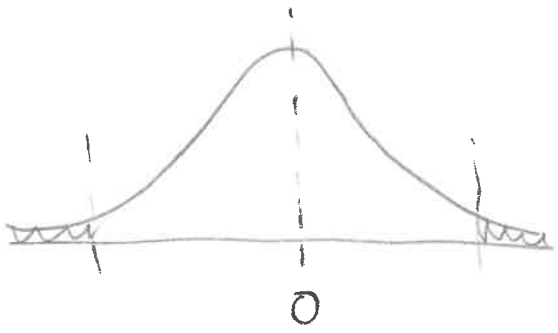
→ so if you were on a desert island and someone asked you for an estimator for the variance of β_j, that's what you would do.

→ For now: Stata.

→ But now we can actually do some hypothesis testing
  ↳ all the ingredients for a t-test.

* We've come a long way from our $\bar{x}$, std. dev. hyp test days
→ the picture & the process are the same.

$$\frac{b_j - \beta_j}{\sqrt{\hat{\sigma}^2_{b_j}}} \sim t_{n-(k+1)}$$



(*) Today: test one β at a time

↳ next time: relationships between β's.

(*) Hypothesis Testing of a single β_i:

Question: Does variable X_i have a statistically significant effect on the dependent variable, Y?

Test: Is b_i (statistically sign.) $\neq 0$?

⟶ can do one-sided tests but usually just do two-sided

(Typically use $\alpha = 0.05$)

3 Different Ways:

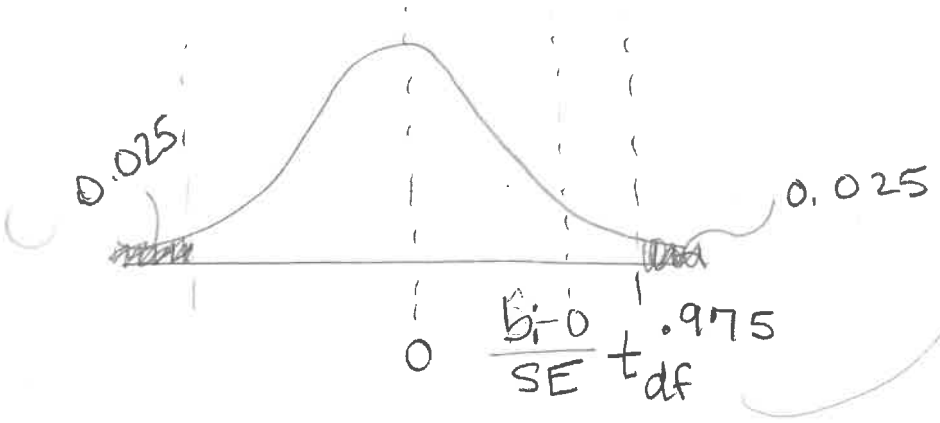
1. Critical Value
2. P-Values
3. Confidence Intervals

$$\frac{b_i - 0}{SE(\hat{\beta}_i)}$$

$H_0: \beta_i = c$

$H_a: \beta_i \neq c$

1. Critical Value

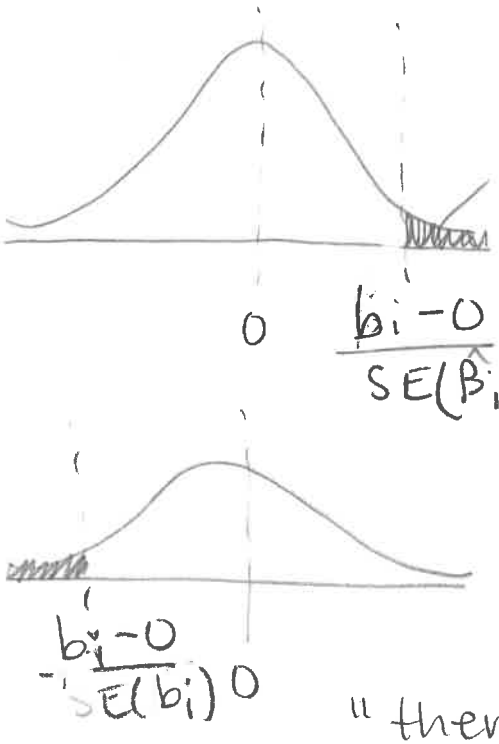


look up .975 in the t_{df} Critical Value

if $\frac{b_i - 0}{SE} < t$ fail to reject

$> t$ reject

2. P-Values



Calculate the p-value
↳ area in the tail (whichever tail)

if p-value > 0.025
↳ fail to reject

if < .025
reject

"there is a p% chance we would see this data & the null ($\hat{\beta}_i = 0$) is true."

3. Confidence Interval

→ calculate the 95% confidence interval
↳ "I am 95% confident that the true value of β_i is contained in this interval."

$[b_i^{LB}, b_i^{UB}]$

if the null value ($\beta_i = 0$) in [ ] → fail to reject
not in [ ] → reject

(*) Keeping all this null, reject, fail to reject stuff straight.

- The null hyp is that $\hat{\beta}_i = 0$.
  - ↳ that var. x_i is not a statistically sign. determinant of Y.

- If we fail to reject the null.
  - ↳ can't conclude $\hat{\beta}_i \neq 0$.
  - ↳ can't conclude x_i a sign. det. of Y

- If we reject the null
  - ↳ reject $\hat{\beta}_i = 0$
  - ↳ can conclude x_i a sign. det of Y

Stata!

⟶ load wage data

⟶ run regression from last time

(*) Is education a statistically significant determinant of wage?

$H_0: \hat{\beta}_1 = 0$

$H_a: \hat{\beta}_1 \neq 0$

1. Critical Value

df?

$df = n - (k+1) = 526 - (6) = 520$

$t^{.975}_{520} = 1.96$

$$\text{test value} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{0.58}{0.052} = 11.15$$

(difference from rounding)

test value $> t^{.975}_{520}$ ⟶ reject the null

2. P-Value

$p\text{-value}_1 = 0.000$

p-value < 0.025 ⟶ reject the null

3. Confidence Interval

[0.48, 0.68]

null value (0) not in interval

↳ reject the null

Conclusion: Reject the null $\beta_1 = 0$, conclude education is a statist. significant determ. of wage.

(*) Are there any variables that we included that aren't sign. at the (two-sided) 5% (2.5%) level?

non-white
married

(*) Are there any that are close?

married
the constant

→ The rule I use is: if you have to mess with the α or the test to get significance, you probably can't conclude it's significant.

Mult. Reg.: Inference (Cont'd.)

→ Last time: individual coefficients statistically signif.

→ This time: relationships between and among coefficients.

Question: Which plays a bigger determ. of your wage: level of education or level of experience?

Variable	$\hat{\beta}$	
const	−1.78	$\hat{\beta}_0$
educ	0.58	$\hat{\beta}_1$
exper.	0.056	$\hat{\beta}_2$
female	−2.07	$\hat{\beta}_3$
nonwhite	−0.025	$\hat{\beta}_4$
married	0.066	$\hat{\beta}_5$

→ The coeff. on education is larger than experience

↳ but just like we can't conclude married is sign. even though the coeff. > 0, we can't conclude this w/o formally testing it.

(*) What are we asking?

"Is $\hat{\beta}_1 \neq \hat{\beta}_2$, statistically significant

don't assume the direction

$H_0: \hat{\beta}_1 = \hat{\beta}_2 \longrightarrow \hat{\beta}_1 - \hat{\beta}_2 = 0$

$H_a: \beta_1 \neq \beta_2 \longrightarrow \hat{\beta}_1 - \hat{\beta}_2 \neq 0$

↑ know the distr. ↑ know the distr.

⟶ Need distribution of the random variable $b_1 - b_2$.

$$\hat{\beta}_1 - \hat{\beta}_2 \sim N(\beta_1 - \beta_2, ?)$$

$$V[\hat{\beta}_1 - \hat{\beta}_2] ?$$

$$= \underbrace{V[\hat{\beta}_1]}_{\text{got this}} + \underbrace{V[\hat{\beta}_2]}_{\text{got this}} - \underbrace{Cov[\hat{\beta}_1, \hat{\beta}_2]}_{\text{a huge mess}}$$

Options:

- calculate it analytically (calculus)
- do some algebraic gymnastics w/ the regression (in the book)
- * • ask the nice computer

→ cue up our regression from last time

(*) the command for these super complicated fancy tests is: test.

↳ uses whatever regression you ran last.

↳ name the β's using their associated variables.

Command: test educ (==) exper

(>, <, ==)

we'll talk F about the F part in a minute.

→ it lists the null:

educ − exper = 0

→ gives the p-value (& the test statistic)

Prob > F = 0.000

→ Since p-value less than 0.025 reject null!

Conclusion: $\hat{\beta}_1$ is statistically significantly greater than $\hat{\beta}_2$

↳ Years of education is a stronger determinant of your wage than years of experience.

(*) You can do even crazier equations than $\hat{\beta}_1 - \hat{\beta}_2 = 0$.

↳ If you wanted to know "Is $2\hat{\beta}_1 + 3\hat{\beta}_2 + \hat{\beta}_4 = \hat{\beta}_3$?"

test 2*educ + 3*exper + nonwhite == female

(*) We know how to test if $\hat{\beta}_1$ is signif. and we know how to test if $\hat{\beta}_2$ is signif. but how do we test if they're jointly significant?

Question: If we control for education & experience, does your race, gender, and marital status play a (statist. sign.) role in determining your wage?

Or: Are $\hat{\beta}_3$, $\hat{\beta}_4$, and $\hat{\beta}_5$ jointly equal to zero?

(*) How is this diff. from testing each, one at a time?

→ When we test each individually We're testing $\hat{\beta}_3 = 0$ or $\hat{\beta}_4 = 0$ or $\hat{\beta}_5 = 0$.

→ This is testing $\hat{\beta}_3 = 0$ and $\hat{\beta}_4 = 0$ and $\hat{\beta}_5 = 0$

(*) We're asking if this estimated equation (Unrestricted Model)

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{u}$$

is statistically different from this estimated equation

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_2 + \tilde{u} \quad \text{(Restricted Model)}$$

$$(= \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_2 + \tilde{\beta}_3 X_3 + \tilde{\beta}_4 X_4 + \tilde{\beta}_5 X_5 + \tilde{u}$$

where $\tilde{\beta}_3$, $\tilde{\beta}_4$ and $\tilde{\beta}_5$ are restricted to 0)

The null hypothesis is

H_0: $\hat{\beta}_3 = 0$ and $\hat{\beta}_4 = 0$ and $\hat{\beta}_5 = 0$

$\hat{\beta}_3 = \hat{\beta}_4 = \hat{\beta}_5 = 0$

H_a: H_0 is not true

At least one of $(\hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)$ are not $= 0$.

(*) The Restricted Model is the one where we impose the restrictions from the null hypothesis.

(*) The Unrestricted Model has no such restrictions.

(*) Each variable set to 0 is a restricti

$q \equiv$ the number of restrictions

$= 3$ in this case

(also called numerator df)

⟶ We can actually run this test using the R^2's from the two model estimations.

$$\frac{(R^2_{UR} - R^2_R)/q}{(1 - R^2_{UR})/n-(k+1)} \sim F_{q,n-(k+1)}$$

$F_{1,n-(k+1)}$ ↳ $t_{n-(k+1)}$ ↳ that's why last one said

⊛ Using Stata:

regress wage .... married

Unrestricted $R^2 = 0.3158$

$n-(k+1) = 520$

$F_{3,520} = 2.60$

regress wage educ exper

Restricted $R^2 = 0.2252$

$q = 3$

$$F \text{ statistic} = \frac{(0.3158 - 0.2252)/3}{(1-0.3158)/520}$$

$$= \frac{0.0906/3}{0.6842/520}$$

$$= \frac{0.0906}{3} \cdot \frac{520}{0.6842} = \frac{47.112}{2.0526}$$

$$= 22.95$$

$22.95 > 2.60$

reject null

(*) Can also use "test"

Command: test (female = 0)
(nonwhite = 0) (married = 0)

⟶ don't forget that it uses the most recently estimated model
↳ reestimate UR model

P-value = 0.000
Reject null

Conclusion: $\hat{\beta}_3$, $\hat{\beta}_4$, and $\hat{\beta}_5$ are not all jointly 0.
↳ education and experience aren't the only thing that determine wage.

Functional Form

(*) We're pretty much in
doing linear regression
variables.

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

→ the linear part ref
on the β's.
→ the X's are just da
have pretty much an
as they satisfy the

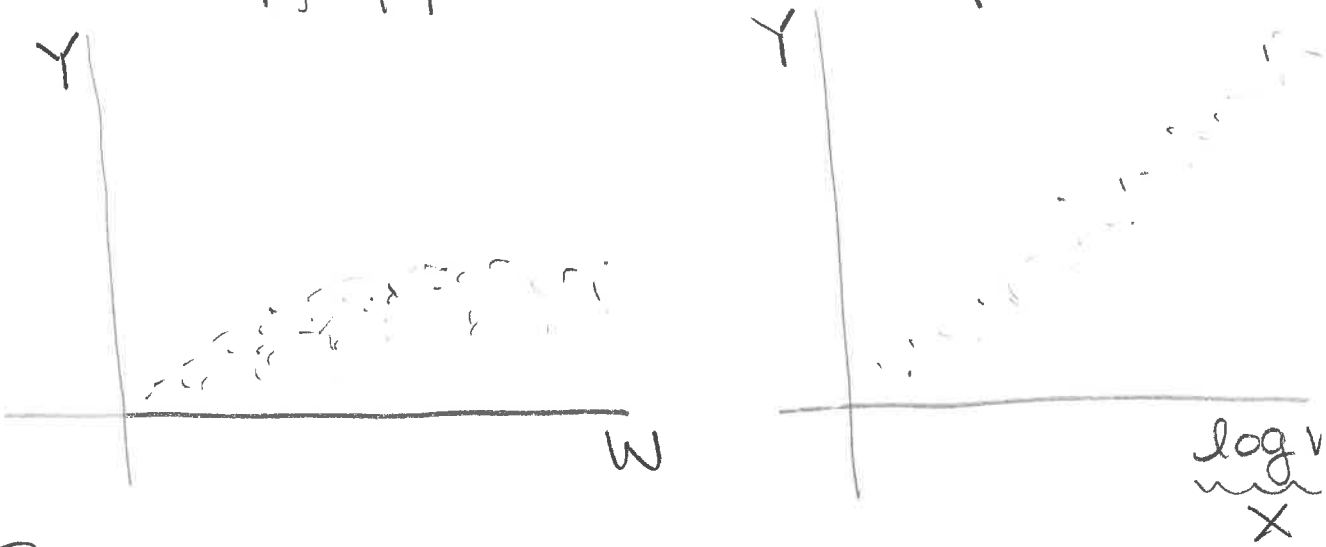
$$Y = \beta_0 + \beta_1 \underbrace{(7 W_1)}_{X_1} + \beta_2 (1$$

→ from a practical per
just mult. W_1 by 7 i
X_1 and reg Y X_1 as

(*) There are pretty much
can do to a variable
1. Scale it
→ add a constan
→ mult by a c

→ this is pretty cosmetic
↳ do it to make things easier to work with (fewer 0's, more standard interpretation, etc.)

2. Take the natural log of a variable
→ log(x)
→ We do this when a variable doesn't appear to be linearly related to Y, appears to be exponential



3. Raise it to a power
→ usually squared: x^2
→ can be other ones: x^3, $\sqrt{x}$, et
→ square variables when we want to make big numbers bigger
• age, experience
17 vs. 21 2 vs. 6

→ first one: doesn't change much, straightforward

→ 2nd & 3rd: a little trickier

(*) Scaling

Y = college GPA

W_1 = SAT score

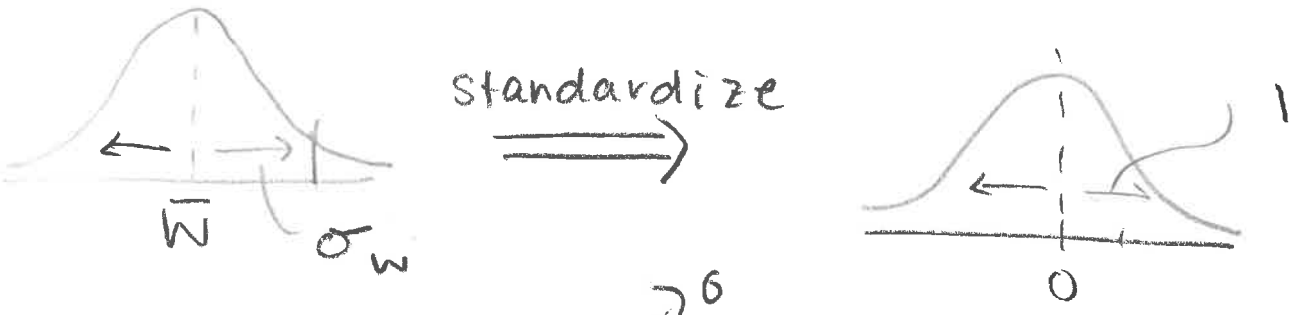
W_2 = ACT score

$$Y = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + u$$

→ standardize everything!

Standardize: subtract the mean, divide by std. deviation.

→ construct a z-score



$$\frac{Y - \bar{Y}}{\sigma_Y}, \quad \frac{W_0 - \bar{W}_0}{\sigma_0}, \quad \frac{W_1 - \bar{W}_1}{\sigma_1}, \quad \frac{W_2 - \bar{W}_2}{\sigma_2}$$

z_y z_0 z_1 z_2

$$\frac{u - \bar{u}}{\sigma_{z_u}} \quad \frac{u - 0}{\sigma}$$

→ take our equation

$$Y = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + u$$

→ and average it

$$\sum_{i=1}^{n} Y = \sum_{i=1}^{n} (\beta_0 + \beta_0 1 + \ldots + u)$$

$$\bar{Y} = \beta_0 \cdot 1 + \beta_1 \bar{W}_1 + \beta_2 \bar{W}_2 + \bar{u}$$

$\bar{u} = 0$

→ subtract $\bar{Y}$ from Y

(*) always remember your Algebra I.

→ got to do the same thing to both sides of the equation.

$$Y - \bar{Y} = \beta_0 - \beta_0 1 + \beta_1 W_1 - \beta_1 \bar{W} + \beta_2 W_2 - \beta_2 \bar{W}_2 + u - \bar{u}$$

$u - \bar{u} = 0$

$$\frac{Y - \bar{Y}}{\sigma_Y} = \frac{\beta_1 (W_1 - \bar{W}_1) + \beta_2 (W_2 - \bar{W}_2) + u}{\sigma_Y}$$

→ classic algebra trick:

mult. by $1 = \sigma_1/\sigma_1 = \sigma_2/\sigma_2$

$$\frac{Y - \bar{Y}}{\sigma_Y} = \frac{\beta_1 (W_1 - \bar{W}_1)\sigma_1/\sigma_1 + \beta_2 (W_2 - \bar{W}_2)\sigma_2/\sigma_2 + u}{\sigma_Y}$$

$$\frac{Y - \bar{Y}}{\sigma_Y} = \beta_1 \left(\frac{W_1 - \bar{W}_1}{\sigma_1}\right)\sigma_1/\sigma_Y + \beta_2 \left(\frac{W_2 - \bar{W}_2}{\sigma_2}\right)\sigma_2/\sigma_Y + u/\sigma_Y$$

$$z_y = \underbrace{\beta_1 \sigma_1/\sigma_y}_{b_1} z_1 + \underbrace{\beta_2 \sigma_2/\sigma_y}_{b_2} z_2 + \underbrace{u/\sigma_Y}_{v}$$

(z_1 ↓ x_1, z_2 ↓ x_2)

$$\hat{Y} = b_1 x_1 + b_2 x_2 + v$$

Before: β_1 told us the effect of hard SAT score on GPA level

↳ one more SAT point

Now: b_1 tells us the effect of 1 std. deviation impr. in SAT score on your GPA relative to everyone else

↳ one SAT std. dev from the mean associated with a b_1 std. dev. impr. in GPA.

Stata!

load gpa.dta

hmm. no ACT

↳ but highschoolsize

↳ use that for W_2 like we meant to all along!

1. Create the Standardized Vars.

Var	Mean	Std. Dev
cumgpa	2.08	0.9896
sat	898.91	168.19
hssize	312.77	198.78

(sum)

"summarize cumgpa sat hssize"

(*) to create a new variable use "generate" (gen)

```
gen z_gpa = (cumgpa - 2.08) / 0.9896
gen z_sat = (sat - 898.91) / 168.19
gen z_hs = (hssize - 312.77) / 198.78
```

→ now we have three new vars we can just use as usual.

↳ although we should specify no constant

```
reg z_gpa z_sat z_hs, noconstant
```

one word

$$\hat{z\_gpa} = 0.1521\ z\_sat + -0.00005\ z\_hs$$

(*) A 1 std. dev. in SAT score is assoc. with a 0.15 std. dev. increase in college GPA.

(*) Log Transformations

→ When we use log transformation to make the variables work better

↳ sometimes they make the data better satisfy the G-M Assumptions

→ They can have really nice Economic interpretations.

Y – depend.
X – ind.

→ If we regress

$$\log(Y) = \beta_0 + \underline{\beta_1} \log(x) + u$$

β_1 is the elasticity of Y with respect to X

Elasticity: the % Δ of one variable associated with a 1% Δ in another

Ex: the price elasticity of demand is the %Δ in quantity demanded associated with a 1% Δ in price.

$$E_D = \%\Delta Q / \%\Delta P$$

Ex: We have data

Y = number of houses sold in US counties

x = avg. annual salary in the counties

regress

$$\log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 \log(x)$$

$\hat{\beta}_1$ estimator for the income elasticity of demand

⟶ A 1% increase in income is associated with a $\hat{\beta}_1$% increase in houses sold.

(*) If we also included a second independent variable, like county population that isn't logged (b/c it's already linearly related with Y)

$$\widehat{\log(Y)} = \hat{\beta}_0 + \hat{\beta}_1 \log(x_1) + \hat{\beta}_2 x_2$$

→ β_2 is not an elasticity
↳ it's called a semi-elasticity

"A 1 unit increase in X_2 is associated with a $(100\,\beta_1)$ % increase/decrease in Y."

"Y then x"

Model	Dep. Var	Ind. Var	Interpret.
level level	Y	X	$\Delta y = \beta_1 \Delta x$
level log	Y	log(x)	$\Delta y = \left(\frac{\beta_1}{100}\right)\% \Delta x$
log level	log(Y)	X	$\%\Delta y = 100\,\beta_1 \Delta x$
log log	log(Y)	log(x)	$\%\Delta Y = \beta_1 \%\Delta x$

What we'll be using

→ Every Metrics textbook should have a version of this table to refer to.
↳ yours is on p. 39

(*) to predict Y when you regressed log(Y), adjust $e^{\log Y}$ by

$$a = \frac{1}{n}\sum_{i=1}^{n} e^{\hat{u}} = e^{\bar{\hat{u}}} \longrightarrow \hat{Y} = a \cdot e^{\log Y}$$

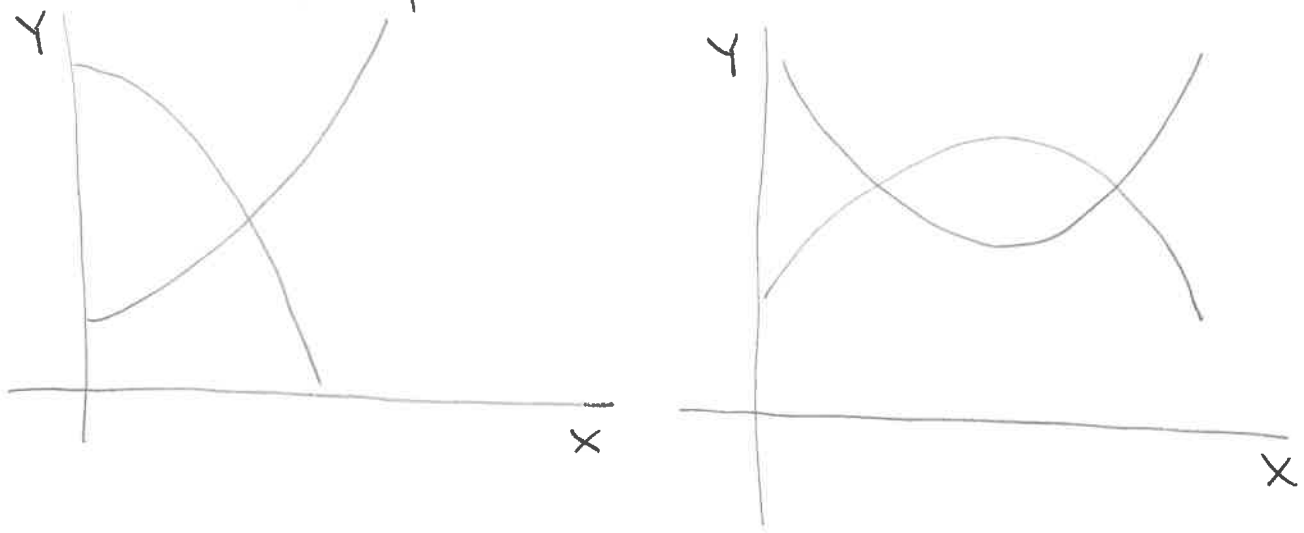
Functional Forms: Quadratics & Interactions

→ last time: first two things we can do to variables (scale them, take the log

→ today: the third thing
  - ↳ multiply variables together
  - ↳ sometimes its the same variable: X^2, X^3, etc.
  - ↳ sometimes it's different variables: $X_1 \cdot X_2$

✳ Quadratics:

→ when the variable appears non-linear, quadratic



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \underbrace{X_1^2}_{X_2} + \ldots + u$$

→ squaring things exaggarates differences

- big numbers get bigger
- small numbers get smaller
- that's why we square the errors in OLS

→ usually the squared variable is also there as a level

- doesn't violate perfect collinearity
- BUT: makes the interpretation of the effect of X on Y a little trickier

not a problem to implement just call $x_2 = x_1^2$.

- If $\Delta x_1 = 1$, $\Delta Y = \beta_1$ or $= \beta_2$ or $= \beta_1 + \beta_2$, etc.

→ I'm afraid we'll need a smidgen of calculus today

Calculus [Reminder?]

$$f(x) = 1x^{0} + 4x^{2} + 1x^{3}$$

$$\frac{df(x)}{dx} = \frac{d(x)}{dx} + \frac{d(4x^2)}{dx} + \frac{d(x^3)}{dx}$$

$$= 1x^{0} + 4(2x^{1}) + 3x^{2}$$

$$= 1 + 8x + 3x^{2}$$

partial der. ⟶ just treat other vars as cons.

$$\frac{\partial Y}{\partial X_1} = \frac{\partial(\beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \dots)}{\partial X_1}$$

$$= \frac{\partial \beta_0}{\partial X_1} + \frac{\partial \beta_1 X_1}{\partial X_1} + \frac{\partial \beta_2 X_1^2}{\partial X_1} + \dots$$

$$= 0 + \beta_1 \cdot 1 + \beta_2 (2X_1) + 0$$

$$= \beta_1 + 2\beta_2 X_1 \leftarrow \text{"Partial Effect"}$$

(*) the effect of a change in the variable X_1 depends on the particular value of X_1

Example: $\hat{Y} = 3 + 2X_1 + 4X_1^2$

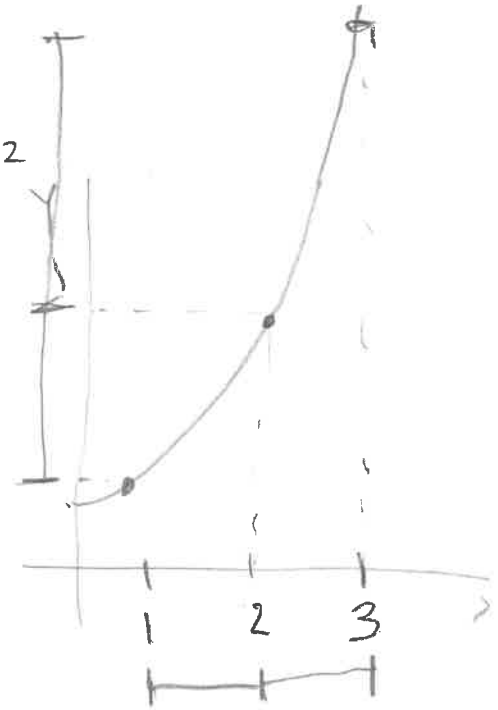
$$\frac{\partial \hat{Y}}{\partial X_1} = 2 + 8X_1$$

When $X_1 = 1$

$$\Delta Y = 2 + 8(1) = 10$$

When $X_1 = 2$

$$\Delta Y = 2 + 8(2) = 18$$

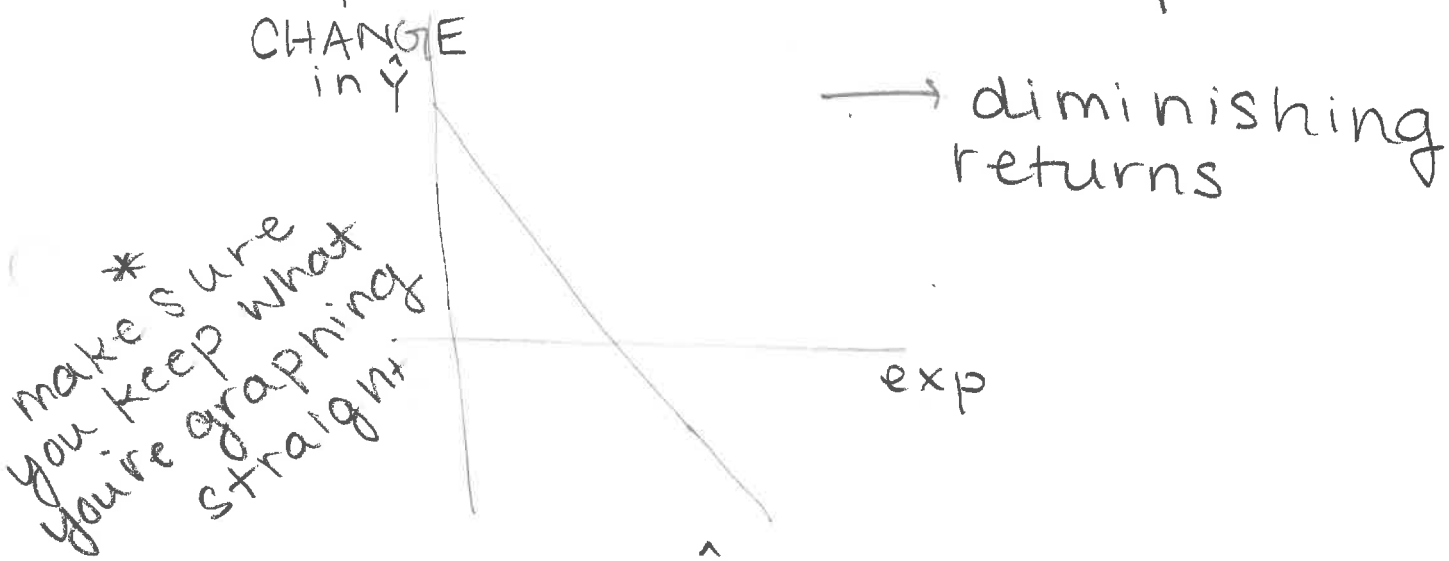


(*) What about when $\beta_1 > 0$ and $\beta_2 < 0$?
↳ what the heck does that mean?

Example: $\widehat{wage} = 3.73 + .298\,exp - .0061\,exp^2$

Effect of x on $\hat{Y}$?

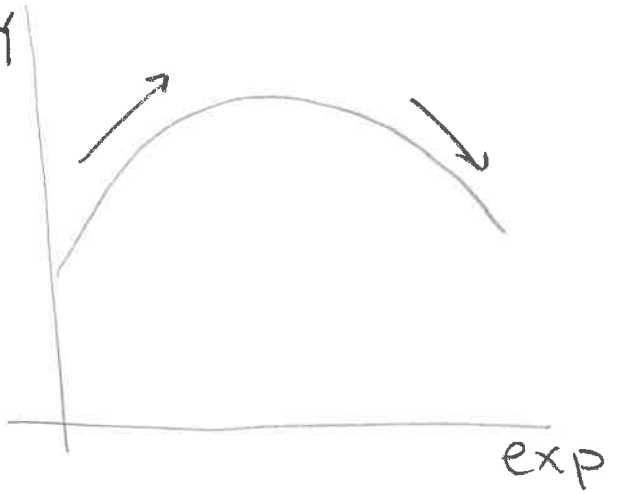
$$\frac{\partial \hat{Y}}{\partial exp} = 0.298 - 0.0122\,exp$$



→ diminishing returns

* make sure you keep what you're graphing straight

→ the change to $\hat{Y}$ is positive for small values of exp but eventually it's negative



(economically relevant)

→ A natural question is: what's the turning point?

↳ thankfully our high school algebra classes gave us a formula for this!

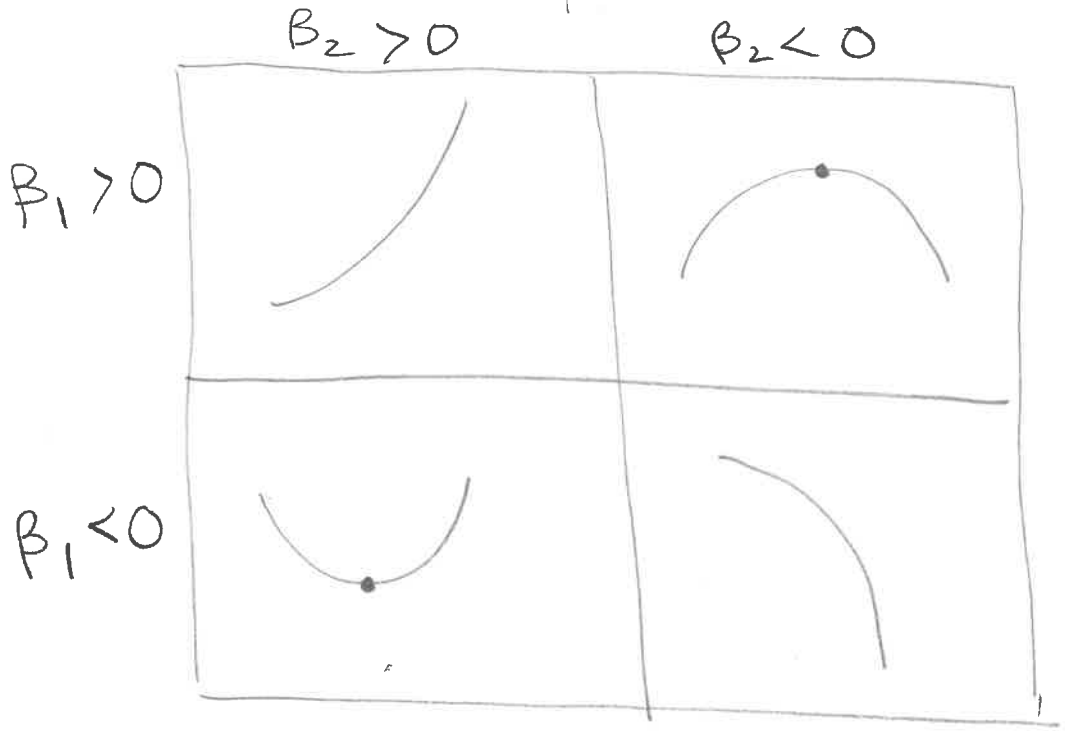
$$x^* = \left| \frac{\hat{\beta}_1}{2\hat{\beta}_2} \right|$$

x^*: turning point x

the turning point tells us when it stops being good & starts being bad (or vice versa)

Turning Point Experience Level =

$$\left| \frac{0.298}{-0.0122} \right| = |-24.4| = \underline{24.4}$$



(*) Interaction Terms:

→ Sometimes the effect of a variable depends on the value of another variable

→ If you've ever watched Flip or Flop (or really any HGTV show) you know that the effect of number of bedrooms has diminishing returns (x^2) but also depends on the square footage.

$$\hat{price} = \hat{\beta}_0 + \hat{\beta}_1 sqft + \hat{\beta}_2 rooms + \hat{\beta}_3 rooms^2 + \hat{\beta}_4 sqft \cdot rooms$$

→ Now we need to be a little more specific about partial derivatives.

→ when you have a function with multiple variables:

$$f(x, w, z) = x^2 + zx + 2wz + 4w^2x$$

and you just want the derivative (chang with respect to one variable

$$\frac{\partial f(x, w, z)}{\partial w}$$ "the partial derivative of f with respect to w

→ hard to say, easy to do

→ just treat all the variables that aren't that variable like constants

$$\frac{\partial f(x,w,z)}{\partial w} = 0 + 0 + 2(1)z + 4(2w)x$$
$$= 2z + 8wx$$

1. What is the effect on price associated with sqft?

$$\frac{\partial \hat{price}}{\partial sqft} = 0 + \hat{\beta}_1 + 0 + 0 + \hat{\beta}_4 rooms$$
$$= \hat{\beta}_1 + \hat{\beta}_4 rooms$$

→ the effect of square feet depends on how many bedrooms there are.

2. What is the effect on price associated with rooms?

$$\frac{\partial \hat{price}}{\partial rooms} = 0 + 0 + \hat{\beta}_2 + 2\hat{\beta}_3 rooms + \hat{\beta}_4 sqft$$
$$= \hat{\beta}_2 + 2\hat{\beta}_3 rooms + \hat{\beta}_4 sqft.$$

→ the effect of the number of bedrooms depends on both the number of rooms and the square feet.

→ it's all well and good to have an equation describing the effect associated with a variable

↳ But we would really prefer to have a single number

↳ "dollars per sq ft"

⊛ Average Partial Effect

$$\text{Partial Effect (Rooms)} = \hat{\beta}_2 + 2\hat{\beta}_3 \text{rooms} + \hat{\beta}_4 \text{sqft}$$

→ average across all values

$$\frac{1}{n}\sum_{i=1}^{n} PE(\text{rooms})_i = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_2 + 2\hat{\beta}_3 \text{rooms} + \hat{\beta}_4 \text{Sqft}\right)$$

$$= \hat{\beta}_2 + 2\hat{\beta}_3 \frac{1}{n}\sum_{i=1}^{n} \text{rooms}_i + \hat{\beta}_4 \frac{1}{n}\sum_{i=1}^{n} \text{sqft}_i$$

$$= \hat{\beta}_2 + 2\hat{\beta}_3 \overline{\text{rooms}} + \hat{\beta}_4 \overline{\text{sqft}}$$

→ definition: $APE = \overline{PE}$ computed by plugging in $\bar{x}$'s.

→ "on average, an increase of one bedroom is associated with an increased (APE) dollars."

Stata!

- load hprice data
- sum price area rooms

Var	Mean	Sta. Dev.
price	96100.66	43223.73
area	2106.73	694.96
rooms	6.59	.9012

↑

- generate rooms_sq = rooms * rooms
- generate rooms_area = rooms * area
- regress price area rooms rooms_sq rooms_area

$$APE(rooms) = \hat{\beta}_2 + \hat{\beta}_3 \overline{rooms} + \hat{\beta}_4 \overline{area}$$
$$= 54,013.65 + (-4209.14)(6.5$$
$$+ 4.62(2106.73)$$
$$= 54013.65 - 27,738.23 + 9,825.49$$
$$= 36,100.91$$

"On average, one more bedroom is assoc. with $36,100.91 in price increase."

Binary Variables

(*) So far: quantitative data
(measure a quantity)

all qualitative
- mostly continuous:
  - wage
  - sq. feet
  - police per-capita
- some discrete
  - education
  - bedrooms
  - num. children

(*) A lot of interesting & important questions involve qualitative data

binary
- gender
- marital status

categorical
- region of the US (N, S, E, W)

→ this ~~chapter~~ section addresses how to use this type of information correctly.

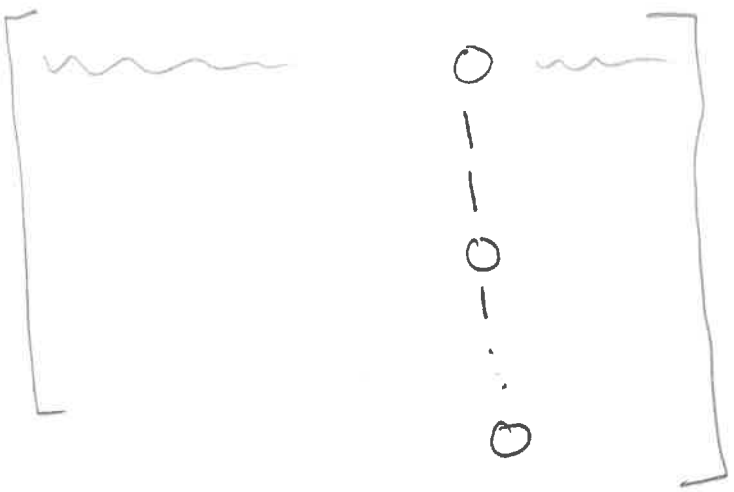
(*) Binary Variables

→ take on the value 0 or 1

→ what's a 0 and what's a 1 is up to you.

Words of Caution:

1. Be consistent
   - with what we do in class
   - across your work
2. Name your variable according to what's = 1
   - female = 1 if female
     0 if male
     (not "gender")
   - married = 1 if married
     0 if not married
   - always tell people what's 1 and what's 0.
3. Never include both (for example) male and female
   - perfect collinearity.



Example:

Model: (just to differeniate it)

$$Wage = \beta_0 + \delta_1 male + \beta_1 educ + u$$

male = 1 if male
0 if female

$$E[wage \mid male, educ] = \beta_0 + \delta_1 male + \beta_1 educ$$

↑ plug in values

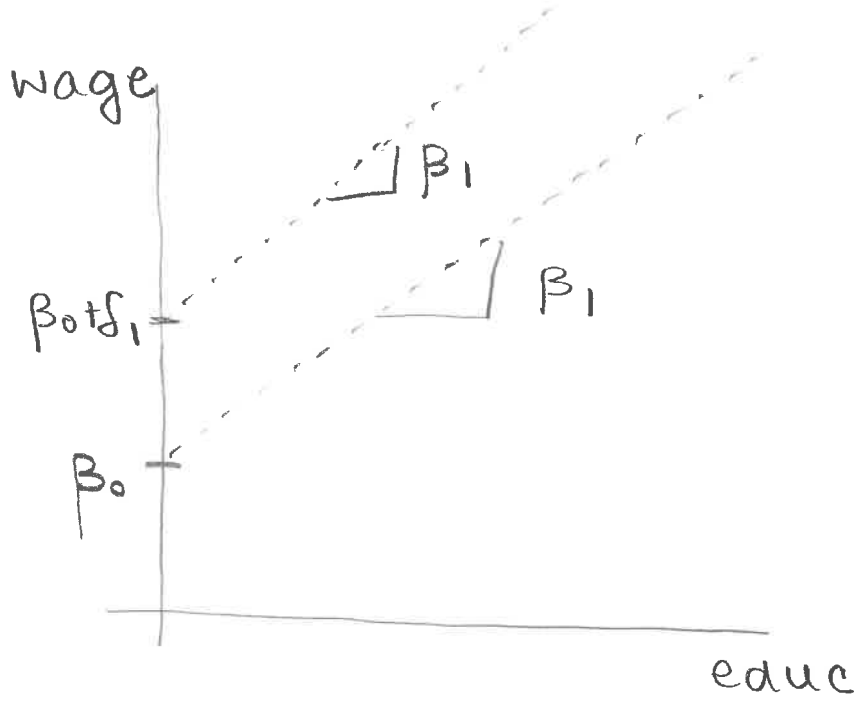
$$E[wage \mid 1, educ] = \beta_0 + \delta_1 (1) + \beta_1 educ$$

$$= \beta_0 + \delta_1 + \beta_1 educ$$

$$E[wage \mid 0, educ] = \beta_0 + \delta_1 (0) + \beta_1 educ$$

$$= \beta_0 + \beta_1 educ$$

→ graph them!



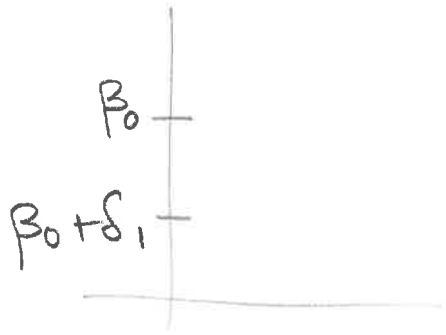
→ assumes returns to educ don't depend on gender

→ δ_1 is the difference in wage due to being male ($\delta_1 > 0$)

→ make sure you keep your definitio straight

↳ if we had used "female"

$\delta_1 < 0$



→ the variable that is a 0 is called the base group or benchmark group.

↳ what you're making comparison to.

* Categorical & Ordinal variables

→ it's straightforward to extend the binary variable process to cover multiple categories.

Example: Regions of the Country

$$\text{region} = \begin{cases} 1 = \text{North} \\ 2 = \text{South} \\ 3 = \text{East} \\ 4 = \text{West} \end{cases}$$

→ comes in two flavors: categorical & ordinal

ordinal: the order matters

$$\text{credit rating} = \begin{cases} 1 = D \\ 2 = C \\ 3 = B \\ 4 = A \end{cases}$$

→ categorical: order doesn't matter

can switch

region = { 1 = N, 2 = E, 3 = S, 4 = W }

→ ordinal: can say something like "higher is better"
  ↳ direction means something.

(*) To actually implement this it's best to basically make a binary variable for every (−1) category.

→ could make credit rating one variable, but that would assume constant partial effects: going from D → C is the same as B → A

→ making region its own variable makes no sense.
  ↳ because order doesn't matter.

north = { 1 north, 0 other } east = { 1 east, 0 other }

south = { 1 south, 0 other }

(*) Always leave out one category
↳ perfect collinearity

$CR\_D = \begin{cases} 1 & \text{if D} \\ 0 & \text{if other} \end{cases}$ $CR\_C = \begin{cases} 1 & \text{if C} \\ 0 & \text{if other} \end{cases}$

$CR\_B = \begin{cases} 1 & \text{if B} \\ 0 & \text{if other} \end{cases}$

(leave out A)

⟶ the one you leave out is the base case

⟶ Note: sometimes it's not practical to create separate dummies
Example: school rank
⟶ can just treat as qualitative
⟶ break into percentiles, etc.

Stata!

- load data from spreadsheet
- regress
  price male_author

(*) Interpret Coefficient

- make genre dummies

sci_fi = 1 scifi
0 other

⋮

regress price male_author
sci_fi ___ ___ ___ ___

(*) Interpret coefficients

Heteroskedasticity

(*) The assumption that took our OLS estimates from unbiased to BLUE was homoskedasticity

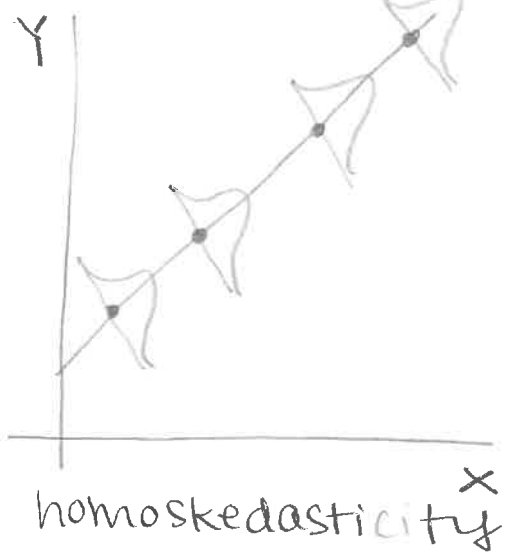
$$V[u \mid x = x_i] = \sigma \quad \forall x_i$$

⟶ the variance of u doesn't change with values of x.

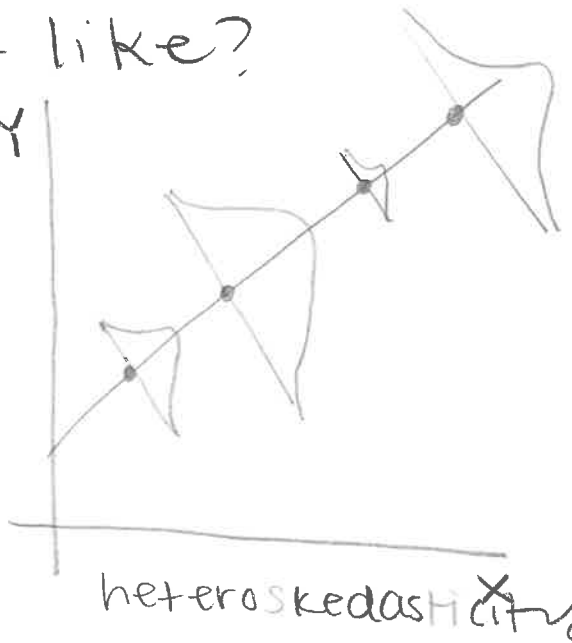
⟶ what happens when that assumptio doesn't hold?

- What does that look like?
- what problems does it cause?
- how do we test for it?
- how do we fix the problems?

(*) What does it look like?

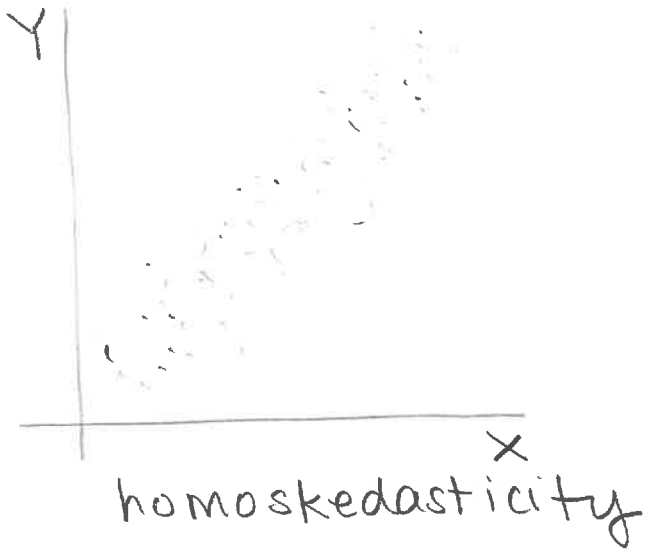


homoskedasticity

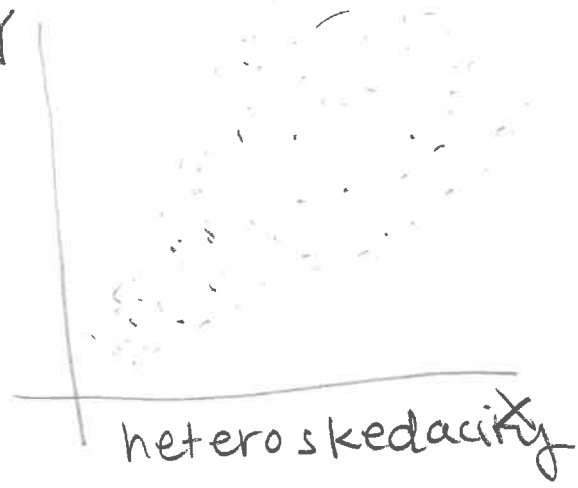


heteroskedasticity

$$V[u \mid X = x_i] = \sigma_{(i)}$$



homoskedasticity



heteroskedacity

(*) What problems does it create?

Reminder: Homoskedasticity was not an assumption needed for unbiasedness
↳ heteroskedasticity can't lead to bias

The Problem:

→ need an estimate for the variance of the β's to test for significance
↳ no $V[\hat{\beta}]$ → no distribution
↳ no t-tests

- the estimate of $V[\hat{\beta}]$ requires an estimate for $V[u]$
  - ↳ can't assume $V[u]$
    - ↳ can't assume $V[\hat{\beta}]$
      - ↳ can't test for $\hat{\beta}$ significance

⟶ so you'll have the right β (it's unbiased) but you can't say whether it's statistically $\neq 0$.

(*) How do we test for it?

Null Hyp: Everything is homoskedastic and grand.

$$V[u|x=x_i] = \sigma^2 \leftarrow \text{a constant}$$

⟶ We'll do some statistics and either reject the null (heteroskedastic) or fail to reject.

What is the definition of variance?

$$V[u] = E\left[(u - E[u])^2\right]$$

$E[u] = 0$

$$= E\left[(u-0)^2\right]$$

$$V[u|x] = E[u^2|x]$$

$$V[u|x] = \sigma^2 \text{ (under null)}$$

$$E[u^2|x] = \sigma^2 \text{ (under null)}$$

→ under the null hypothesis, there's no relationship between u^2 & the x's.

→ Let's say we do our normal regression

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{u}$$

if we take these & square them

$$\hat{u}^2 = \hat{\delta}_0 + \hat{\delta}_1 x_1 + \hat{\delta}_2 x + \hat{v}$$

→ if our null hyp is true, $\hat{\delta}_1 = \hat{\delta}_2 = 0$

→ we know how to test that!

↳ just an F-test

H_0: $\hat{\delta}_1 = \hat{\delta}_2 = 0$

H_a: H_0 is not true

Bruesch-Pagan Test Outline:

1. Regress Y on x's.
2. Calculate the residuals, $\hat{u}$
3. Square them
4. Regress $\hat{u}^2$'s on x's
5. Jointly test x coefficients = 0
   - if reject null: heterosk.
   - if fail to reject: homosk.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \text{ on } \begin{bmatrix} X_{11} & \cdots & X_{k1} \\ X_{12} & \cdots & X_{k2} \\ \vdots & & \vdots \\ X_{1n} & \cdots & X_{kn} \end{bmatrix}$$

$$\Downarrow$$

$$\begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix} \rightarrow \begin{bmatrix} \hat{u}_1^2 \\ \hat{u}_2^2 \\ \vdots \\ \hat{u}_n^2 \end{bmatrix} \text{ on } \begin{bmatrix} \sim \\ \sim \\ \sim \end{bmatrix}$$

→ it's so cool!
↳ We made this data! But it can tell us something meaningful about the world.

(*) How do we fix the problems it causes?

→ the problem is that we have no estimate for the variance / std. error of the $\hat{\beta}$'s.

In the simple regression case

$$Y = \alpha + \beta X + u$$

$$V[\hat{\beta}] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma_i^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

(we can do a multiple regression version but it's notationally crazy)

→ if there's homoskedasticity

$$\sigma_i^2 = \sigma^2 \quad \forall i$$

↳ doesn't vary with i

$$V[\hat{\beta}] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = \frac{\sigma^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

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

σ^2 ← estimate

$\sum_{i=1}^{n}(x_i - \bar{x})^2$ ← calculate

done!

→ but if heteroskedastic can't pull that σ^2 out

$$V[\hat{\beta}] = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma_i^2}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}$$

→ but we can estimate this in pretty good way

↳ use $\hat{u}_i^2$ as an estimator for σ_i^2

→ so we can just use:

$$V[\hat{\beta}] = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\hat{u}_i^2}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}$$

(*) "Pretty Good" in what way?

- this method is valid "as the sample size tends to infinity"
  → for large samples this is good.
- if no heteroskedasticity, the regular t-tests are valid for any sample size.

→ heteroskedasticity is not a huge problem
  ↳ check for it.
  ↳ if present, use $\hat{u}$ ("robust standard errors")
  ↳ if not, don't

→ it's better if it isn't there but if it is, hopefully you have large enough data sets that it isn't a problem.

Stata!

load cps data

scatter lwage exper
  → eh hard to say
scatter lwage educ
  → yep. heterosk.

reg lwage educ exper expersq

hettest
  → reject null

reg lwage educ exper expersq, robust

Binary Variables: Interaction Terms

→ Before: used interaction terms to allow for differential effects

sqft • bedrooms

→ Binary Variables can also have differential effects

Example: the effect of being married on your wage may be different if you're male or female

Example: the effect of years of education may be different if you're male or female.

2 Types:

1. Binary - Binary interactions
   female - married

2. Binary - Qualitative interactions
   female - education

→ using two binary variables allow for four different intercepts

female = 1 female
0 male

married = 1 married
0 not married

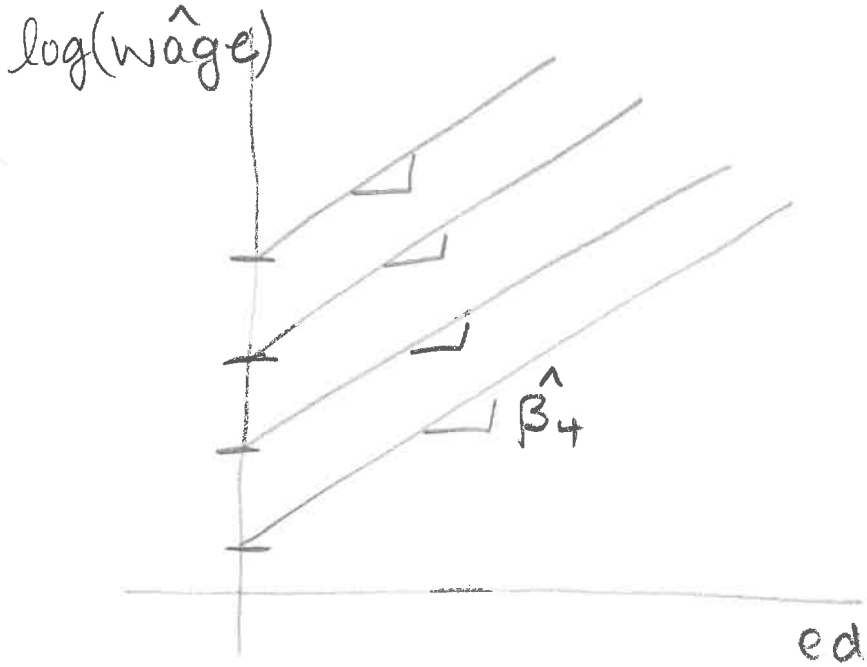
What's the base case?
male and not married

Four Cases:

married \ female	1	0
1	married women	married men
0	single women	single men ~ base case

→ each of these have a different intercept

$$\widehat{\log(wage)} = \hat{\beta}_0 + \hat{\beta}_1 female + \hat{\beta}_2 married + \hat{\beta}_3 femal \cdot married + \hat{\beta}_4 educ$$



→ We'll calculate these things in a moment.

→ but this still assumes the returns to education are the same for everyone ($\hat{\beta}_4$)

↳ to allow for differential returns (different slopes)

female · educ

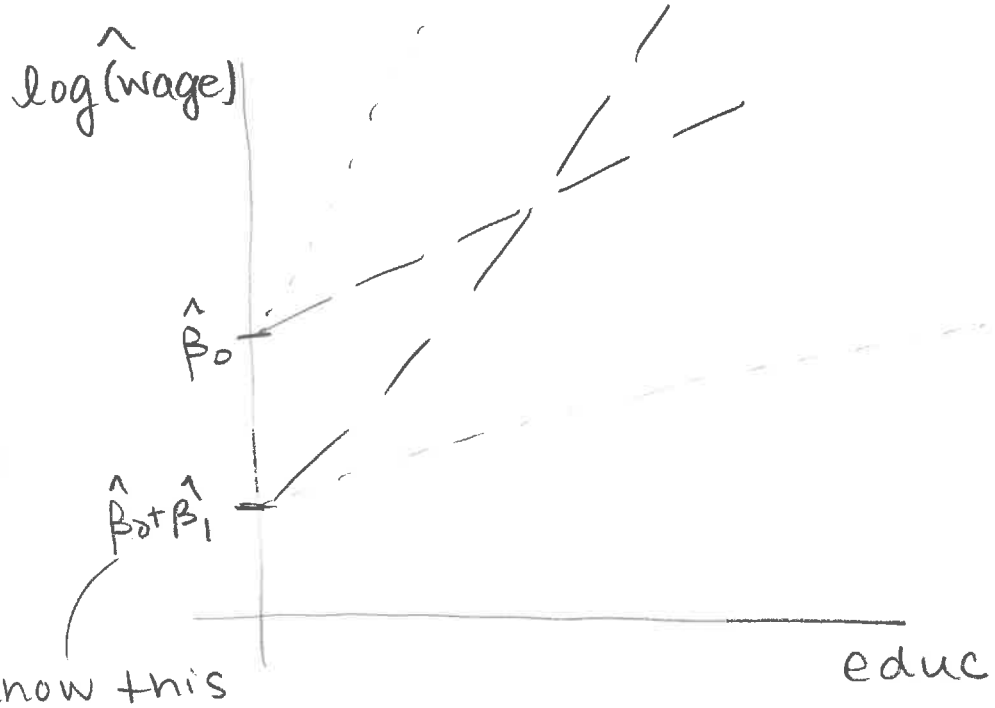
$$\log(\hat{w}age) = \hat{\beta}_0 + \hat{\beta}_1 female + \hat{\beta}_2 educ + \hat{\beta}_3 female \cdot educ$$

$$E[\log(\hat{w}age) \mid female = 0] = \hat{\beta}_0 + 0 + \hat{\beta}_2 educ + \hat{\beta}_3 \cdot 0 \cdot educ = \hat{\beta}_0 + \hat{\beta}_2 educ$$

base case relationship

$$E[\log(wage) \mid female=1] = \hat{\beta}_0 + \hat{\beta}_1(1) + \hat{\beta}_2 educ + \hat{\beta}_3(1)educ$$

$$= \underbrace{(\hat{\beta}_0 + \hat{\beta}_1)}_{\text{female intercept}} + \underbrace{(\hat{\beta}_2 + \hat{\beta}_3)}_{\text{female slope}} educ$$



know this intercept is lower from before

$\hat{\beta}_3 < 0$

→ female has flatter slope

→ returns $(\hat{\beta}_2 + \hat{\beta}_3)$ are smaller

$\hat{\beta}_3 > 0$

→ female has steeper slope

→ returns $(\hat{\beta}_2 + \hat{\beta}_3)$ larger

(*) Joint Testing

Question: Is there any effect on wage from being a female?

→ need to test $\hat{\beta}_1$ & $\hat{\beta}_3$ jointly

H_0: $\hat{\beta}_1 = \hat{\beta}_3 = 0$

H_a: H_0 is not true
(at least 1 is not $= 0$)

(could throw married in too

→ unrestricted model

$$\widehat{\log(wage)} = \hat{\beta}_0 + \hat{\beta}_1 female + \hat{\beta}_2 educ + \hat{\beta}_3 female \cdot educ$$

→ restricted model:

$$\widehat{\log(wage)} = \hat{\beta}_0 + \hat{\beta}_2 educ$$

→ 2 restrictions ($q = 2$)

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

Stata!

- load cps75_85 data

① Female & Married

- create female· married
- regress lwage female married female_married educ

$$\widehat{lwage} = 0.87 - 0.15\,female + 0.25\,marri - 0.16\,female\_married + 0.073\,educ$$

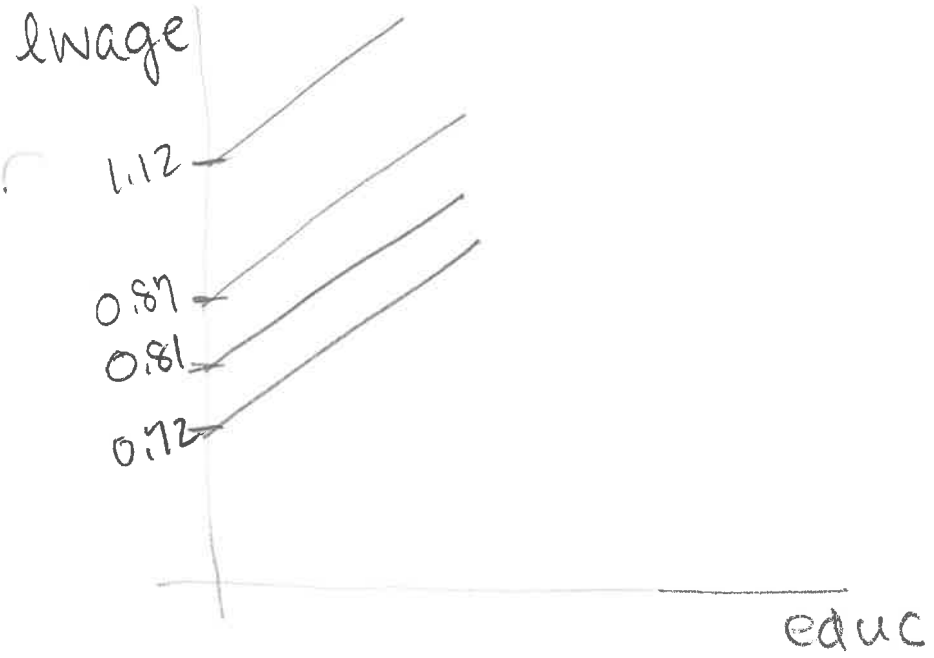
- graph

$$E[lwage \mid female=0, married=0] = 0.87 + 0.073\,educ$$

$$E[lwage \mid 0,1] = \underset{1.12}{0.87} + 0.25(1) + 0.073\,educ$$

$$E[lwage \mid 1,0] = \underset{0.72}{0.87 - 0.15(1)} + 0.073edu$$

$$E[lwage \mid 1,1] = \underset{0.81}{0.87 - 0.15(1) + 0.25(1) - 0.16(1)} + 0.073educ$$



(2) Female & Education

- gen female · educ

- reg lwage female educ female_edu

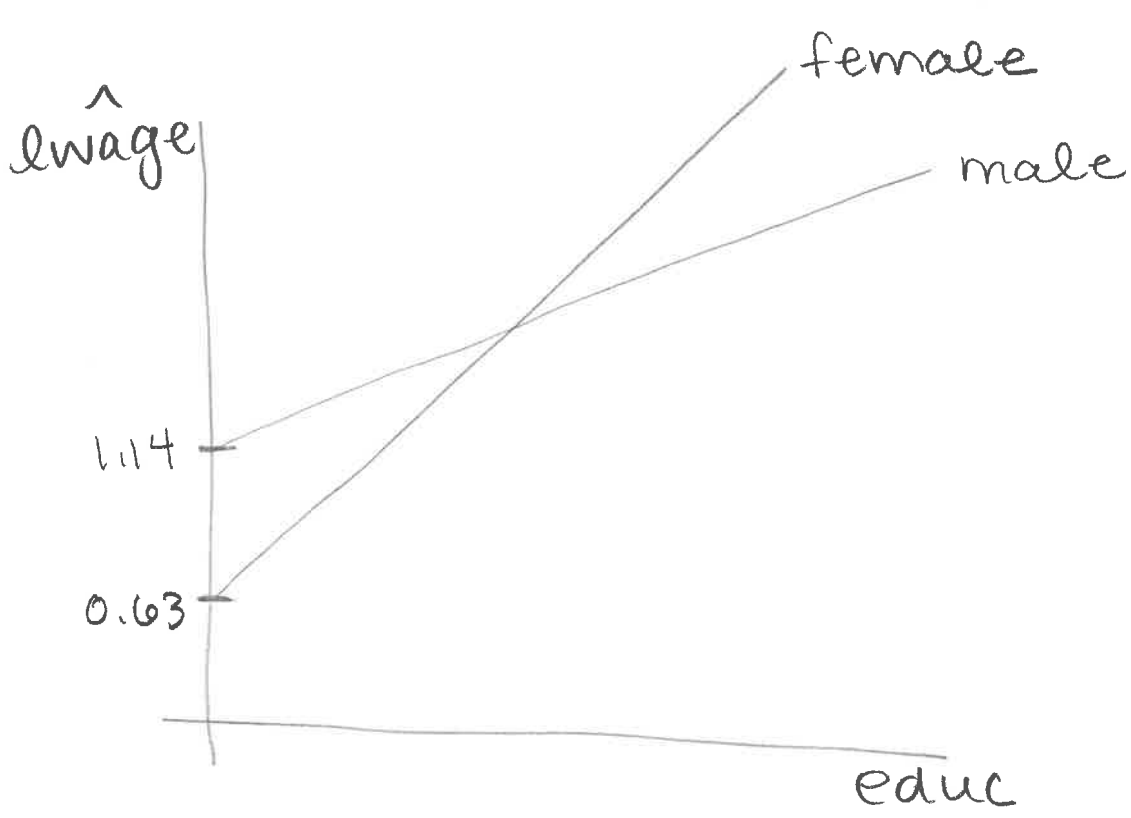
$$\widehat{lwage} = 1.14 - 0.51\,female + 0.065\,educ + 0.018\,female\_educ$$

- graph

$$E[lwage \mid female=0] = 1.14 + 0.065\,educ$$

$$E[lwage \mid female=1] = 1.14 - 0.51 + 0.065\,educ + 0.018\,\underset{(1)}{\text{female}}\_\text{edu}$$

$$= 0.63 + 0.083\,educ$$



→ men have a lower intercept (w/ no education, men's wages are higher)

→ women have a steeper slope (an increase of education of one year earns women a larger increase in wage)

- test

Does being a woman affect your wage?

$H_0: \hat{\beta}_1 = \hat{\beta}_3 = 0$

H_a: H_0 is not true

test (female = 0)(female_educ = 0)

$F(2, 1080) = 41.44$

p-value = 0.000

reject the null that both are 0.

Time Series Econometrics

(*) This last section of the class is meant to expose you to some of the different genres of econometrics (&techniques)

- time series (today)
- panel data
- differences in differences
- big data

→ each of these could be their own course

↳ our goal is to get an introduction to each

↳ if you have to use any of these in the future...

1) you know you need to do something different from the standard
2) you know where to start

→ Up First: Time Series Econometrics

(*) Time series Econometrics is just econometrics applied to time series data

↳ data that occurs across time.

Examples:

- annual GDP
- quarterly sales
- quarterly unemployment
- weekly bond rates

→ annual, quarterly, weekly, daily

→ you should always specify the periodicity of your data

→ typically, time series data are used a lot in macro

↳ plenty of exceptions: home prices, sales, etc.

→ graphing time series:

- x: time
- y: data

↳ comes with an automatic x-var built in.

(*) Notation

→ the value of y at time t may depend on the value of x at time t

$$y_t = \alpha + \beta x_t$$

- house prices today depend on interest rates today
- inflation this year depends on unemployment this year

→ the value of y at time t may depend on the value of x at time $t-1$ $\quad y_t = \alpha + \beta x_{t-1}$

- stock prices today depend on consumer sentiment yesterday
- tax revenue this year depends on employment last year.

→ the value of y at time t may depend on the value of $\underline{y}$ at time $t-1$

$$y_t = \alpha + \beta y_{t-1}$$

- aggregate output this year depends on aggregate output this year
- stock prices today depend on stock prices yesterday.

→ When we use a previous time period of a variable, we call it a lag of that variable.

- y_{t-1}, x_{t-1} : first lag of y, x
- x_{t-2} : 2nd lag
- ⋮
- x_{t-h} : h^{th} lag

("lag" looks a lot like "log", make sure you read it right)

time period	y_t	y_{t-1}	x_t	x_{t-1}
0	5	—	20	—
1	9	5	17	20
2	7	9	35	17
3	2	7	42	35
4	4	2	57	42
5	9	4	19	57

"the value of y_{t-1} at $t=2$ is 9" but it isn't useful because it doesn't match with anything

→ need a value for all vars to include in regression

└ can be data here

→ Note: by including 1 lag, we lose one row of data

↳ h lags → h rows → num obs decreases by h

(*) When you include a lag of y, it's likely the errors are correlated
↳ this is called serial correlation and it has an entire chapter to itself.
↳ we're going to focus on the lags of other variables

$$y_t = \beta_0 + \delta_0 x_t + \delta_1 x_{t-1} + \ldots + \delta_h x_{t-h} + u$$

"Finite Distributed Lag (FDL) Model of Order h"

(*) Interpretation

δ_0: the immediate change in y from a 1 unit change in x

δ_1: the change in y from a 1 unit change in x_{t-1}
↳ the effect on y now from a change yesterday

δ_h: the effect on y now from a 1 unit change in x h periods ago

(*) Trends & Seasonality

→ many variables contain a time trend
  ↳ they're growing (or shrinking) over time

→ if we just regress one on the other, we may find fale relationships because they're both moving together in time.

→ include t (omitted var bias otherwise)

$$y_t = \beta_0 + \delta_0 x_t + \delta_1 x_{t-1} + \beta_1 t + u$$

→ the coefficient on t describes the general time trend

β_1: the change in y associated with a 1unit increase in time
  ↳ time moving forward by 1 period

→ some data vary with certain types of times

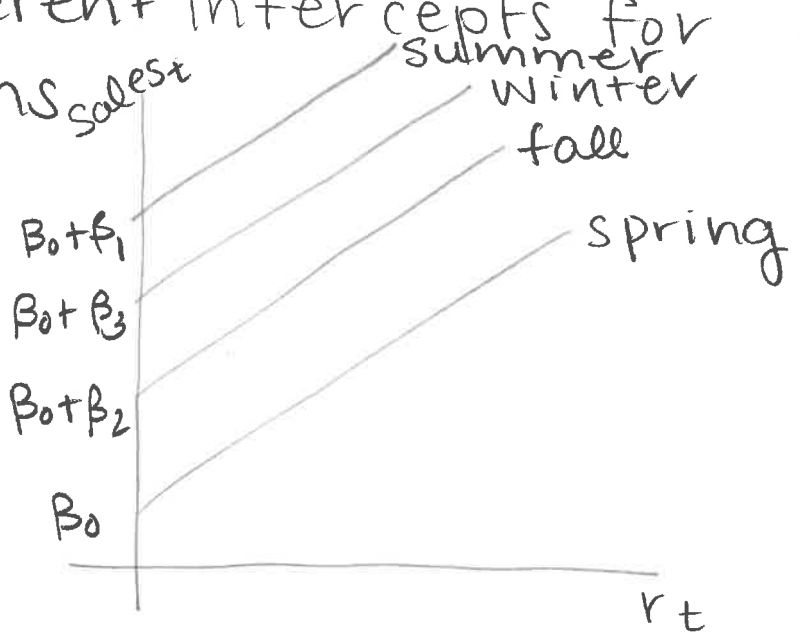
- house sales ($ rentals as many of us know) are much higher in the summer than in other seasons.
- gym memberships in January
- output growth during wartime years

→ We deal with this by including dummies for (each − 1) "season" (month, year, etc.)

$$house\_sales_t = \beta_0 + \delta_0 r_t + \beta_1 summer + \beta_2 fall + \beta_3 winter + u$$

(Always remember to omit one category.)

→ creates different intercepts for different seasons



Stata!

(*) What is the effect of tax exemptions for children on the US fertility rate?

fertility rate: births per 1000 women

→ when you get more money back on your taxes per child, how does it affect the number of children you have?

load fertil3.dta
(*) tell stata this is a time series
sum gfr pe year

```
tsset year
```

(1) FDL Order 2

$$gfr_t = \beta_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + u_t$$

→ you can generate the lag variables directly with gen.

↳ but you can just type L. before a variable to lag it

reg gfr pe L.pe L2.pe

→ slightly different output

L.pe: the effect on fertility today of the pe increasing last year

→ when the pe increases by 1 a year ago, the number of children decreases by 0.02

test joint sign. of all pe's
↳ not jointly significant

(*) Trends & Seasonality
- year
- WWII
- the advent of the pill

$$gfr_t = \beta_0 + \delta_0 pe_t + \beta_1 WW2_t + \beta_2 pill + \beta_3 t + u_t$$

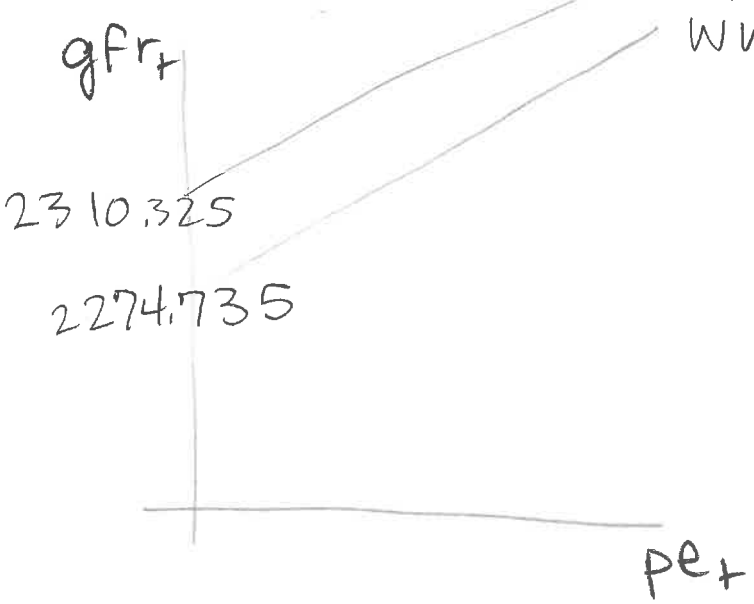
(if time, drop year → pill is sig)
↳ time trend might have been the real driver

(*) What kind of time trend?

→ negative

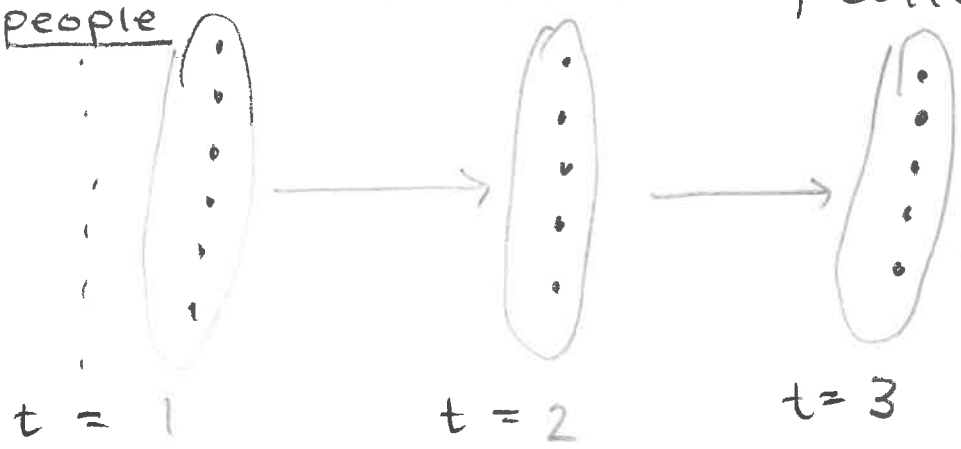
→ on average, every year the number of births per 1000 women decreases by 1.14 children

(*) Graph the effect of pe_t on gfr_t, compare WWII not WWII



Difference in Differences ("Diff-in-Diff")

(*) The next two classes deal with data on a cross section, collected over time



→ allows us to track changes over time

→ particularly useful for policy evaluation.

Examples:

- Did a work training program lead to a decrease in unemployment?
- Does subsidizing health insurance improve health?
- Did the introduction of the birth control pill decrease the fertility rate?

(*) There are 2 ways to collect this data:

1. Sample the same people from a population across multiple time periods

I'm going to say people (cities, schools, etc.)

→ "Panel Data" (makes a panel)

person	t=1	t=2
1	7	9
2	2	9
3	10	10
4	u	u
5	u	u

(Next Class

→ [class example]

2. Sample a different random sample of people from a population in each time period

→ "Pooled Cross Section" (You pool the data)

t=1

7
2 10
9 14
11

t=2

10
9 2
13 7

(Today)

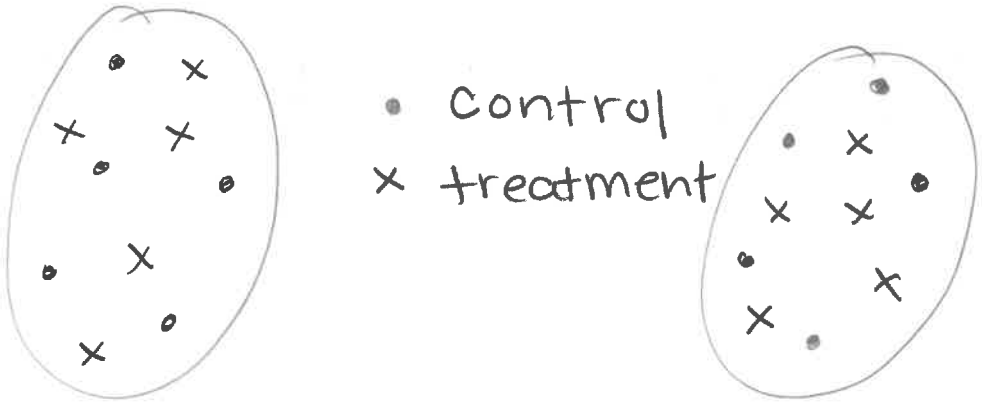
→ [class example]

Difference-in-Differences uses a pooled cross sectio

(*) Ingredients for a Diff-in-Diff Estimation to analyze a policy or change.

- 2 Time Periods
  - → before & after the policy applies
  - → Why needed: there are other things that change over time besides the policy and don't want to eroneously attribute that to the policy.

- 2 Different Groups of [People]
  - → one to which the policy applies ("Treatment Group")
  - → one to which the policy does not apply ("Control Group")
  - → Why needed: there may be things the people in the treatment (or control) have in common
    - ↳ don't want to eroneously capture that either.



$t=1$ (Before)

$t=2$ (After)

(*) How to Implement a Diff-in-Diff Estimation

→ create a dummy variable for the treatment group

$$treat = \begin{cases} 0 & \text{control} \\ 1 & \text{treatment} \end{cases}$$

→ create a dummy variable for one of the time periods

$$t\_2 = \begin{cases} 0 & t=1 \\ 1 & t=2 \end{cases}$$

→ interact them

$$treat\_t\_2 = treat \times t\_2$$

(*) Regress!

Because the time is capturing the policy. Treat is allowing for similarities in the groups that aren't policy related

$$Y = \beta_0 + \delta_0 t\_2 + \beta_1 treat + \delta_1 treat\_t\_2$$

β_0 = average out come for the control in 1st time period

δ_0 = change for all [people] between the two time periods

β_1 = difference between the two groups of people not due to the policy

δ_1 = the diff in diff estimator
= the change in outcome before and after the policy due to the policy

	t=1 Before	t=2 After	After-Before
Control	β_0	$\beta_0 + \delta_0$	δ_0
Treatment	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Treatment − Control	β_1	$\beta_1 + \delta_1$	δ_1

* Example: The effect of a new garage incinerator on nearby housing prices

$$near\_inc = \begin{cases} 0 & \text{if a house is } > 3 \text{ mi. from the location of the incinerator} \\ 1 & \text{if } < 3 \text{ mi} \end{cases}$$

$$y\_81 = \begin{cases} year = 1978 \\ year = 1981 \end{cases}$$

$$price = \beta_0 + \delta_0 y\_81 + \beta_1 near\_inc + \underline{\underline{\delta_1}}\, y\_81 \cdot near\_inc + u$$

δ_0 = the effect on all housing prices over time (inflation)

β_1 = the location effect not due to the incinerator (farther from most jobs)

δ_1 = the effect on housing prices near the incinerator <u>due</u> to the incinerator

Britney Example

(*) Stata!

load INJURY.dta

→ policy: the maximum amount of weekly earnings that were covered by workers comp. increased

→ if you have a low wage, you weren't hitting the old cap, this won't affect you

↳ Control

→ high wage → treatment

→ data before and after the change ("afchange")

(*) Does this policy increase the duration of workers comp claims?

gen after_high = afchange * highearn

reg [og durat] afchange highearn after_hig

$\delta_1 = 0.188$

→ duration increased by about 19% due to the policy

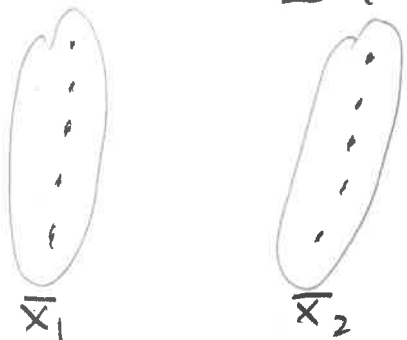
→ coeff on afchange insig.
↳ didn't affect low income workers as expected.

Panel Data: Fixed Effects Model

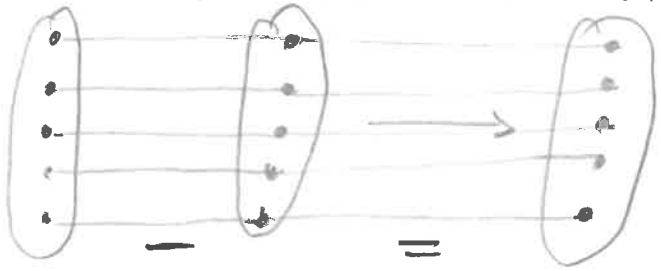
→ rather than pooled data across people, cities, schools, etc.
  ↳ track the same people, cities, schools, etc. in multiple time periods

(*) Why this is useful: contains more information.

→ Intuition: it's like when you have a paired 2-sample t-test, & you're interested if the diff between them is > 0
  ↳ if it's unpaired you just have 2 lists
    ↳ 2 means to use (2 data points)



  ↳ but if it's paired, you can create a new list



(n data points)

Organizing Panel Data:

Variable X $\quad X_{it}$

individual	$t=1$	$t=2$	$t=3$
1	X_{11}		
2	X_{21}		
3	X_{31}	X_{32}	X_{33}
4	X_{41}		
5	X_{51}		

→ this is just one variable, & look at how much space it takes up

→ normally we're working with multiple variables

indiv	t	Y	x^1	x^2
1	1	Y_{11}	x^1_{11}	
	2	Y_{12}	x^1_{12}	
	3	Y_{13}	x^1_{13}	
2	1			
	2			
	3			
3	1		x^1_{31}	
	2		x^1_{32}	
	3		x^1_{33}	
⋮				

(*) We're using this data to answer questions about how something has changed over time

↳ often before & after a policy is implemented.

→ interested in the effect of some policy (policy = $\begin{cases} 0 \\ 1 \end{cases}$ otherwise; 1 if the policy applies to the city)

on some outcome variable

→ why can't we use:

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}$$

(y_{it}: outcome; x_{it}: explanatory var)

First Problem: there are probably things that changed over time that affect the outcome that aren't the policy

↳ don't want to eroneously attribute those to the policy

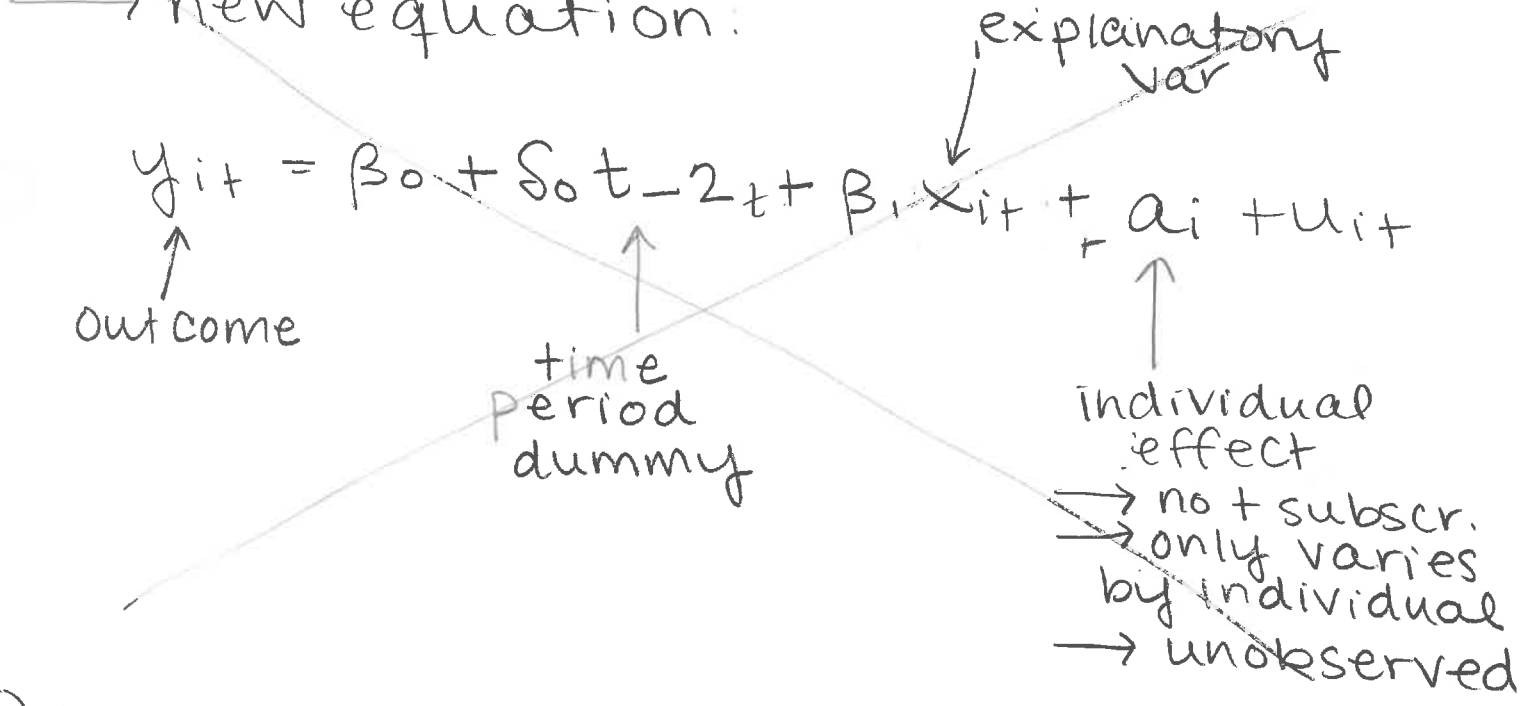
↳ need to take time into account somehow

Second Problem: there are probably people, city, school, etc. specific things that affect the outcome

↳ don't want to attribute those to the policy either.

↳ many of these are unobservable or at least unobserved in your panel

→ new equation:



(*) Dealing with the Individual Effects

→ also called the Fixed Effects (b/c they're fixed over time)

→ there are 3 ways we use to deal with a_i

1. First Differences Model
* 2. Fixed Effects Model (14-1)
3. Random Effects Model

most commonly used

→ Basically, we want a_i to go away but still be able to answer the question & take advantage of all the power of the panel data

* Fixed Effect Model

$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}$$

→ for each i, average the equation over time

num time periods

$$\sum_{t=1}^{T} y_{it} = \sum_{t=1}^{T} (\beta_0 + \beta_1 x_{it} + a_i + u_{it})$$

$$= \sum_{t=1}^{T} \beta_0 + \sum_{t=1}^{T} \beta_1 x_{it} + \sum_{t=1}^{T} a_i + \sum_{t=1}^{T} u_{it}$$

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + a_i + \bar{u}_i$$

→ subtract $\bar{y}_i$ from y_{it} (on both sides of the equation

→ "time demeaning"

$$y_{it} - \bar{y}_i = \beta_0 + \beta_1 x_{it} + \alpha_i + u_{it}$$

$$\ddot{y}_{it} - \left(\beta_0 + \beta_1 \bar{x}_i + \alpha_i + \bar{u}_i\right)$$

$$= \beta_1 \underbrace{\left(x_{it} - \bar{x}_i\right)}_{\ddot{x}_{it}} + \underbrace{u_{it} - \bar{u}_i}_{\ddot{u}_{it}}$$

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{u}_{it}$$

→ gets rid of a_i

↳ anything about the individual constant over time, is averaged out

→ but keeps all the power of the panel data

(*) Example: Does the use of a grant for worker training decrease the number of scrapped products on an assembly line?

factory	t	grant	t_2	$\overline{grant}$	$\overline{t\_2}$
1	1	0	0	0.5	0.5
	2	1	1		
2	1	0	0	0	0.5
	2	0	1		
3	1	0	0	0.5	0.5
	2	1	1		

factory	t	$\ddot{grant}$	$\ddot{t\_2}$
1	1	-0.5	-0.5
	2	0.5	0.5
2	1	0	-0.5
	2	0	0.5
3	1	-0.5	-0.5
	2	0.5	0.5

→ all the information
↳ just shifted around the time mean

Stata!

→ open JTRAIN.DTA

→ open data browser
fcode — firm
year

→ need to tell stata what the individual variable is (fcode) and the time variable (year)

state year

```
xtset fcode year
```

Y X1 X2 X3 X4

```
xtreg lscrap d88 d89 grant grant_1, fe
```

→ obtaining a grant in 1988 decreased the scrap rate in 1989 ($e^{-0.422} - 1 = -.344$) 34.4%

Introduction to Big Data

→ Big Data Econometrics is it's own course, this is merely an introduction

  ↳ goal: should you need to use Big Data tools you'll have an idea of where to start

→ courses in Computer Science & Statistics

  ↳ new Data Science major

(*) Big Data

→ it's called big data because the files/databases containing the data are prohibatively big.

  ↳ standard data-handling program (STATA, Excel, etc) would crash.

→ need special tools & software to:

- store
- access ("query")
- mess with ("computation")

(*) Storage & Accessing

→ usually done across multiple servers

→ Google, Amazon, etc. have made it possible to rent data storage space.

→ for most statistical processes you need to access many if not all values.

↳ just calculating $\bar{x}$!

- find each x_i
- query the value of x_i
- add each one
- divide by n (which is huge)

(*) Most of the time when we refer to Big Data Techniques we're referring to Machine Learning.

↳ usually used to try to predict some y as a function of some predictors (x's).

↳ in the presence of computational challenges

→ in Econ we're usually trying to detect relationships.

↳ Machine Learning has tools that can help us do that job

(*) TWO Important Differences

1. Testing & Training

→ an important difference from "Normal" Econometrics: when we have this much data we can use only some of the data to build the model and some of the data to test the model.

→ always viewed as a good idea

↳ "we didn't have enough data"

↳ now we do!

→ "cross validation"

↳ use some subset of the data to build the model (e.g. estimate the regression coefficients)

↳ use the model to guess the corresponding values for another subset (e.g. plug in the x's to get $\hat{y}$)

↳ see how far off you are ($y - \hat{y}$?)

2. Controlling Complexity

→ We intuitively know that while always including more & more explanatory variables can make the errors smaller

↳ but having too many is cumbersome / overly complex

↳ very non-rigorous way limit the x's.

→ Machine Learning goes one step further: mathematically penalizes too many x's / complexity

→ For example: Instead of minimizing the sum of squared errors

$$\min (y - \beta_0 - \beta_1 x_1 - \dots - \beta_k x_k)^2$$

minimize that plus some count of the number of nonzero β's

$$\min (y - \beta_0 + \beta_1 x_1 - \dots - \beta_k x_k + \lambda(\beta_i \neq 0))^2$$

(*) Two Particular Tools:

1. Classification Trees
2. Variable Selection

→ both concerned with determining what are the most important characteristics in determining some out come.

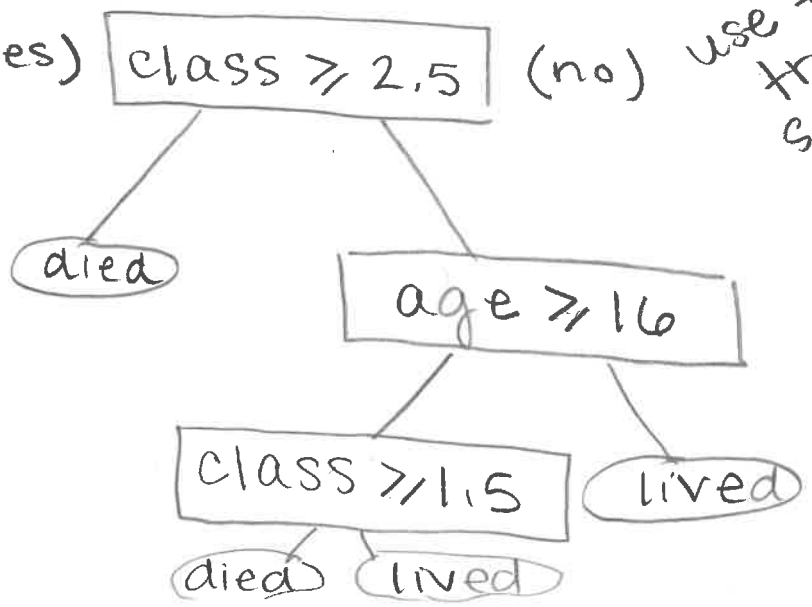
1. Classification Trees

→ binary outcomes

→ trees are a particularly efficient data structure

↳ can handle lots of data

(Dark)
Example: What role did class & age play in whether a titanic passenger survived?



→ use the training set of data to make the tree.

→ then use each person in the testing set and ask the tree for it's prediction (like plinko)

↳ compare the prediction to the real life out come.

Data

Class	Age	Prediction	Real Outcome
1	17	L	L
3	17	D	L
2	34	D	D
1	80	L	L
2	27	D	D

→ this tree classified 723 of 1046 passengers correctly

2. Variable Selection

→ regression but where too many variables are penalized

→ you have so many explanatory variables to choose from

$$\min\left[(y - \beta_0 - \beta_1 x_1 - \ldots - \beta_K x_K)^2 + \lambda \sum_{i=1}^{k} |\beta_k|^2\right]$$

punishes non-zero β's

Example: What factors are important in determining GDP?

→ lots of options

→ Sala-i-Martín data has 42 explanatory variables

↳ he enumerated all the combinations

↳ could just use Machine learning regression

- more comp. efficient
- does a good job

Most Important Factor:

1. Equipment Investment

(*) Model Uncertainty

→ we've spent a lot of time throughout this course talking about the uncertainty (error) created by sampling from a population

↳ we almost never talk about the uncertainty/error created by picking one model over another

(using age & age^2 but not age age^2 & age^3)

→ Big Data (both the data itself & the tools) make it possible to average across the predictions of many models!

Example: "Forests" of Trees.

→ create a bunch of different classification trees (a "forest")

↳ ask each for a prediction

↳ average across predictions

→ If 7 of 10 trees say a person would die on the titanic, the prediction is they would die.

(&

→ Econometricians, Computer Scientists and Statisticians alike agree that this is better than just one model

⊛ Big Data Takeaways:

→ Challenges
- needs separate software for storage, access, computation

→ Benefits
- more information
- allows us to do stuff we said we should do but needs lots of data
  - cross validation
  - complexity control
  - using multiple models