

# regression

adam okulicz-kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Tuesday 29<sup>th</sup> October, 2024 15:56

## outline

intuition of inference (inferential statistics)

multivariate ols: intuition

wages example

interpretation and practice

violations (Wheelan, ch12)

## ps4, ps5

- we'll spend some time on ps4 and ps5
- there are some general comments that apply to most of you
- it's important we all understand them (and those i gave you individually) as they are critical for us to move forward towards the final project

## ps4/ps5

- overwhelmed, all over the place—normal!! again 3 bulletpoints res\_des.pdf; and build on some published study, just add little from yourself
- get your hands dirty with data! enough of plans and outlines, just do it!
- that's the point: if you keep on just planing and outlining, you'll keep on going in circles and confusing and overwhelming yourself
- or if you just do literature review, and no study, then just say it, and just do it as well; comprehensive pls 50+
- be clear about what YOU are doing, not about what we remotely know about the topic, what other did
- background info is cool, but cut to the chase asap!

## ps4/ps5

- lit review is always critical!
- sometimes it is all that you'll do in this class
- need to be comprehensive, ideally 50+ studies
- again, need to synthesize/criticize, tell a story, have a value added from YOU; not just summarize
- refer to <http://theaok.github.io/generic/howToGoogSch.html>
- the goal of the lit rev is not just to get to know
- it is to build foundation for your study, to find out the gap, that your study will fill
- again, be rather modest, take a little step ahead, not save the world

- ideally find a study or few studies you really like and just replicate adding little twist from yourself, maybe just for locality, maybe just update with recent data, focus and elaborate on specific angle
- again, try to find a lit review published—saves time, does much of work for you in one paper!

## ps4/ps5

- many people talk about experiments that are not!! need random assignment!! (and it needs to be feasible/ethical)
- intervention or treatment without random assign fine, can still do before/after but don't call it experiment!!
- experiment is a very specific design of random assignment to treatment
- not a synonym for any study or research as in colloquial everyday language
- lets discuss, give me several examples
- right, so nobody will conduct experiment (IRB, time consuming, etc), but you can plan one for future

## ps4/ps5

- it always helps to define precisely your X, Y, U/A !!
- internal and external validities—specifically about causality and generalizability—should have been more specific and answer the question more directly
- external validity: need to say if sample was random!
- internal validity: discuss some threats
  - really need experiment or at least a quasi experiment
- don't say increased, large etc—use numbers, esp graphs, be specific!
- INUS again: first be clear  $X \rightarrow Y$  !, and then how is X: I,N,U,S (spell out!)—someone give a good example?
- in general PS5 didn't go well, let's pull it up and have several people discuss their answers



# outline

intuition of inference (inferential statistics)

multivariate ols: intuition

wages example

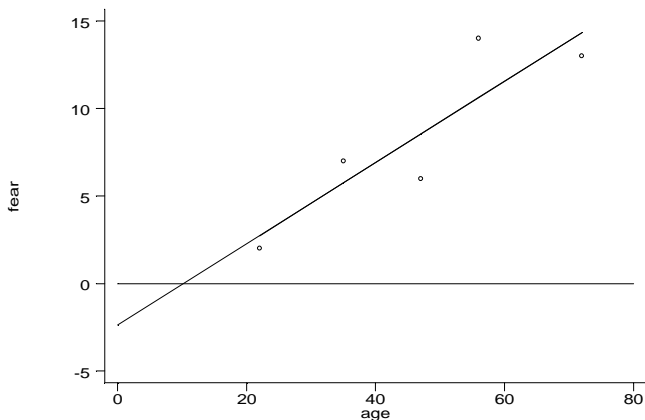
interpretation and practice

violations (Wheelan, ch12)

## finding answers

- got hypothesis?
- now it's time to analyze the data (or critique research)
- that's inference: drawing conclusions (making inferences) from data
- this is what we want to know after all!
- just use regression and “control” for other variables [elaborate later]
- we have research questions, turn them into hypotheses
  - (a brief clear testable statement)
- say have a survey measuring people's fear of crime (0-15)
  - H1: fear of crime increases with age

## example: age and fear



●  $\hat{Y}_i = \hat{\beta}_1 + \beta_2 X_i = -2.365 + .232X_i$

○ eg pre fear at 40yo

## examples

- the regression advantage: use multiple vars at once
- see some of the useful things you can predict
- <http://ianayres.yale.edu/prediction-tools>
- eg divorce probability
  - eg life expectancy <https://www.blueprintincome.com/tools/life-expectancy-calculator-how-long-will-i-live/>

# outline

intuition of inference (inferential statistics)

**multivariate ols: intuition**

wages example

interpretation and practice

violations (Wheelan, ch12)

## multivariate OLS

- multiple (multivariate) reg: most common tool in soc sci
- it finds effect of a var of interest (X) on the dependent var (Y) **controlling/holding constant other vars**
- a stat trick that makes it as if sample equal on all Xs controlled for; imitates experimental setting (randomization)
- again, in experiment you randomize into treatment and control groups so that both groups are on average the same and then we apply treatment (e.g. drug) to treatment group and see if had effect as compared to control group

## multivariate OLS

- most of the time cannot do experiment:
  - can't tell some people to smoke and some not
  - can't give college to some and not others
- but can use regression!
- eg: study effect of education ( $X$ ) on income ( $Y$ )
  - but it may not be the same for males and females?
  - just control for gender in regression
- and the effect is as if everybody had the same gender!

# multivariate OLS

- $X \rightarrow Y$  can say that X affects Y
- $Y = f(X)$  or: Y is a function of X (same thing)
- $Y = f(X_1, X_2, \dots, X_n, u)$
- in soc sci **always** many Xs



# outline

intuition of inference (inferential statistics)

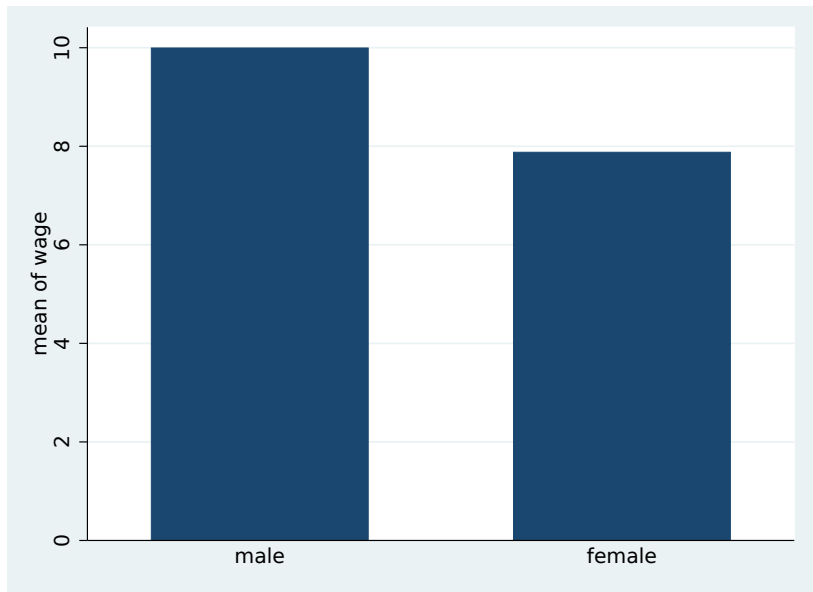
multivariate ols: intuition

wages example

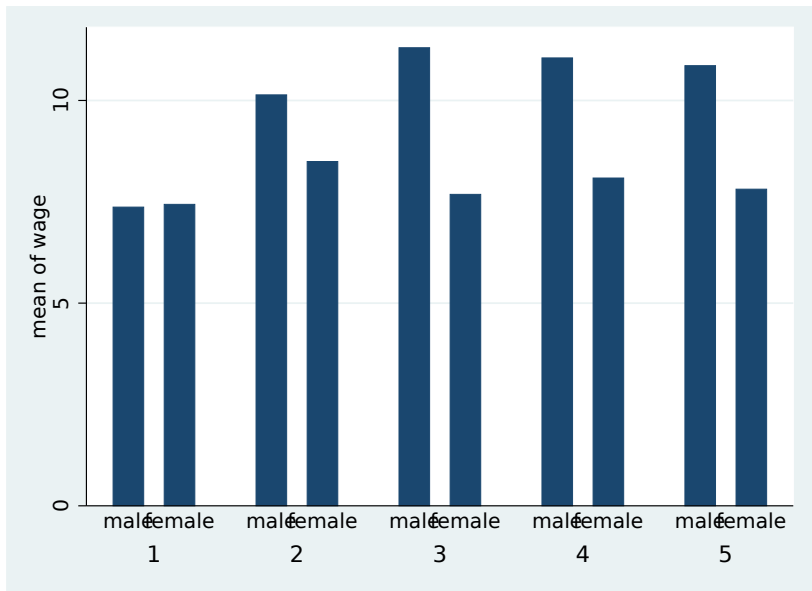
interpretation and practice

violations (Wheelan, ch12)

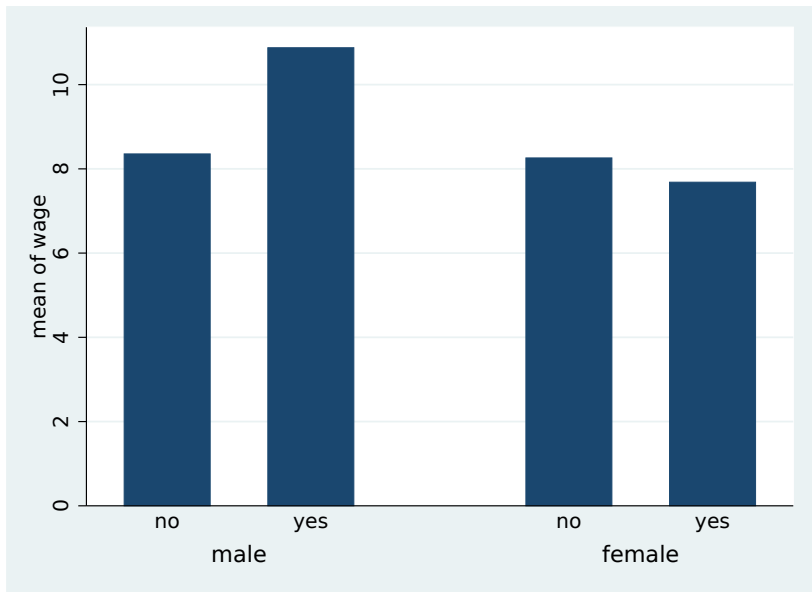
## wages (never do reg w/o des sta)



## wages by quintile of experience



## wages by marital status and gender



## descriptive stats



Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
wage	534	9.02	5.1	1	44.5
educ	534	13.01	2.6	2	18
exp	534	17.82	12.3	0	55

		wage	educ	exp
-----+-----				
wage		1.00		
educ		0.38	1.00	
exp		0.08	-0.35	1.00

## interpreting coefficients

- pretty much only one way to interpret reg correctly
- 1 unit (\$ % etc) increase in  $X$  leads to  $\beta$  unit (\$ % etc) increase/decrease in  $Y$  ( $> 1X$ : remember ceteris paribus!)
- 
- and as per Wheelan ch11: focus on:
- sign
- size
- significance:
  - t-stat,  $t = \text{coeff}/\text{se}$ , sig if  $|t| > 2$
  - $p$  is prob of getting this result or larger if no assoc (Wheelan p198), sig if  $p < .05$
  - $95\% CI = \pm 2 * se$

## multivariate ols



wage		Coef.	Std. Err.	t	P> t
-----+-----					
educ		.9188352	.081526	11.27	0.000
exp		.0986602	.0178812	5.52	0.000
married		.5704847	.4357421	1.31	0.191
_cons		-5.07037	1.224631	-4.14	0.000
-----					

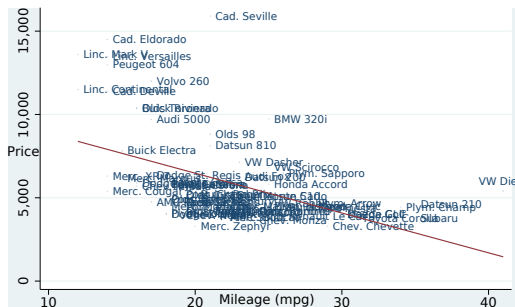
## now let's turn to cars!

- let's say we want to explain price with mpg and weight
  -
- research Q: fuel efficient cars don't have to cost a fortune
- hypothesis: the higher the mpg, the lower the price
- 
- but the problem with fuel efficient cars is that they are tiny
  - and cannot really use them for much



# interpret: $\beta$ , p, t, CI; predict price for 10mpg

price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
mpg	-238.8	53	-4.50	0.000	-344, -133
_cons	11253	1170	9.61	0.000	8919, 13587



## interpret: $\beta$ , p, t, CI; predict price for 10mpg

•

price		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----						
mpg		-49.5186	.15	-0.57	0.567	-221, 122
weight		1.746	.64	2.72	0.008	.46, 3
_cons		1946	3597	0.54	0.590	-5226, 9118

## predicted values (p200 Wheelan, 2013)

- $\text{weight} = -118 + 4.3 * (\text{height in}) + .12 * (\text{age}) - 4.8 * (\text{female})$
- 53yo female who is 5'5:
  - $-118 + (4.3 * 65) + (.12 * 53) - (4.8 * 1) = 163$
- 35yo male who is 6'3:
  - $-118 + (4.3 * 75) + (.12 * 35) - (4.8 * 0) = 209$
- remember life expectancy game? same thing!!
  - <https://www.blueprintincome.com/tools/life-expectancy-calculator-how-long-will-i-live/>
- banks, insurance companies, etc
  - use models like this to predict whether you'll repay loan
  - and hence how risky you are, and whether you should get one

## a “complete” explanation

- $wage = f(\text{native ability, education, family background, age, gender, race, height, weight, strength, attitudes, neighborhood influences, family connections, interactions of the above, chance encounters, ...})$
- multiple regression will tell you the effect of one variable while controlling for the effect of other variables (again, as if everybody was the same on other vars)
- $wage_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_n X_{ni} + u_i$

# outline

intuition of inference (inferential statistics)

multivariate ols: intuition

wages example

interpretation and practice

violations (Wheelan, ch12)

## practice regressions interpretations

- Happy Tourists, Unhappy Locals <http://link.springer.com/article/10.1007/s11205-016-1436-9>

## ps6: flip the class!

- was it difficult?
- any challenges?
- need to cover anything about regression again?

## do scatterplots

- it is useful to produce a scatterplot
  - you'd see outliers
    - and whether the relationship is due to them
  - **blackboard**: relationships biased due to outliers
  - say marriage rate and divorce rate across states



## think about it

- always interpret results!
- give it some thought
- ask yourself whether results make sense and why
- think about measurement and what it means
  - eg does marriage cause divorce or sth about NV?
- and as always, remember design principles:
  - INUS condition
  - threats to validity
- and note that in addition to regression
  - it is critical to have theory/logic/mechanism
  - see Wheelan (2013, p207)

## Wheelan in ch11 mentions Whitehall studies

- fascinating stuff!
- high status causes better health!
- great book 'Status Syndrome' <http://a.co/jaUuwT7>
- say nobel prize or oscar boosts one's health and longevity
- these successful folks live longer and in better health
- than exact same people (income, lifestyle, etc) but without status

# outline

intuition of inference (inferential statistics)

multivariate ols: intuition

wages example

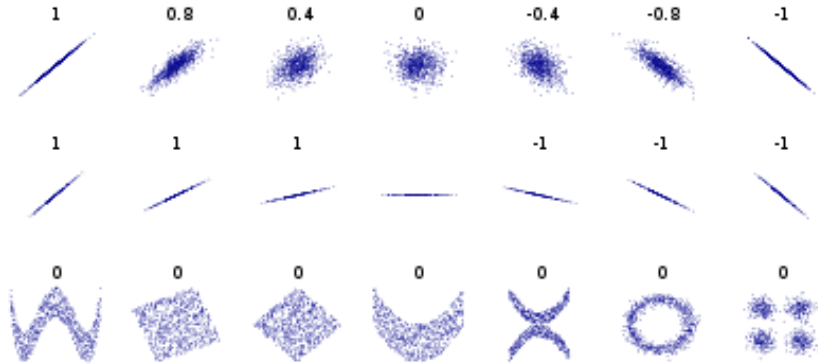
interpretation and practice

violations (Wheelan, ch12)

## do not kill people with regressions (p212 Wheelan, 2013)

- recently tens of thousands of females were killed or made sick with estrogen, because regressions showed that estrogen was good
- regression estimates are never causal by themselves!
- remember the gold standard: experiment!
- again, INUS, unknown unknowns,  $\text{corr} \neq \text{causation}$ , etc

## nonlinear relationships



- like corr, won't detect nonlinear relationships!
- example of nonlinear rel? extra credit!

## what to do about nonlinear rel?

- just break it up into subsets/subsamples! dig deeper!
  - say for males and females separately
  - say for low and hi val separately
- that's a quick way to see nonlinear relationship!
- eg it may first rise and then fall

## reverse causality (p216 Wheelan, 2013)

- more lessons—  $\rightarrow$  bad golf, or
- bad golf—  $\rightarrow$  more lessons
- solution:
  - lag variable: bad golf last month—  $\rightarrow$  more lessons now
  - use exogenous shock—remember from res\_des.pdf:
  - (terrorist attack—  $\rightarrow$ ) policing—  $\rightarrow$  crime
- or think about it! miserable people choose cities?
- then i looked at only people who were born in urban/rural

## omitted variable bias (p217 Wheelan, 2013)

- golf— > heart disease and cancer?
- control for age!
- age is killing people, not golf!



## extrapolate beyond data (p220 Wheelan, 2013)

- only interpret within range of data!
- remember regression of fear on age?
- and reg line hits y-axis at -3

## data mining (p221 Wheelan, 2013)

- if you torture your data enough, it will confess
- likewise, if you throw enough variables, you will find significant relationships
- but remember: you need theory, causal mechanism/path, story!

## run it

- excel

- <http://www3.wabash.edu/econometrics/EconometricsBook/Basic%20Tools/ExcelAddIns/OLSRegression.htm>

- python

- <http://www.learn datasci.com/predicting-housing-prices-linear-regression-using-python>
- <https://stackoverflow.com/questions/19991445/run-an-ols-regression-with-pandas-data-frame>

- how's ps6 going? anyone show what you have so far?
- troubles finding reg tables?

LEVITT, S. D. AND S. J. DUBNER (2010): Freakonomics, vol. 61, Sperling & Kupfer.

WHEELAN, C. (2013): Naked statistics: stripping the dread from the data, WW Norton & Company.