regression

adam okulicz-kozaryn
adam.okulicz.kozaryn@gmail.com

this version: Wednesday 11$^{\text{th}}$ November, 2020   16:01

outline

2016 elections bonus: wrong prediction

intuition of inference (inferential statistics)

multivariate ols: intuition

wages example

interpretation and practice

violations (Wheelan, ch12)

ps4/ps5

- it always helps do define precisely your X, Y, U/A !!
- 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− >Y !, and then how is X: I,N,U,S (spell out!)–someone give a good example?

ps4/ps5

- many people talk about experiments that are not!! need random assignment!! (and it needs to be feasable/ethical)
  - intervention or treatment without random assign fine, can still do before/after but don't call it experiment!!

outline

2016 elections bonus: wrong prediction

intuition of inference (inferential statistics)

multivariate ols: intuition

wages example

interpretation and practice

violations (Wheelan, ch12)

wrong prediction! also see Wheelan (2013, ch10)

- concepts from this class help explain

- many states were wrongly predicted!

- remember no prediction argues 100% certainty

○ eg what people say in polls must be accompanied by
 actual action (voting)

○ majority voicing support for candidate in polls does not
 cause candidate to win

○ mere INUS condition, eg "shy Trump"

○ and Trump voters may not want to talk to pollsters at all

○ not only candidate support counts; also propensity to vote!

- discussion:

  http://www.economist.com/blogs/economist-explains/2016/11/economist-explains-3

not only stats failed with election prediction

- way of running it failed too!
- Hilary was data driven
○ doing what models tell her to do
- Donald had a gut feeling
○ doing what intuition tells him what to do
- and often gut feeling wins!
○ unknown unknowns, INUS condition, etc
○ extreme difficulty to argue anything without experiment
- data and statistics usually help, but are not everything!

outline

finding answers

- got hypotheis?

- now it's time to analyze the data or critique research

- that's inference: drawing conclusions (making inferences)
  from data

- just use regression and "control" for other variables
  [elaborate later]

- this is what we want to know after all!

- We have research questions, turn them into hypotheses

  ○ (a brief clear testable statement)

  ○ eg research Q: soc sci PhD good for career ?

  ○ H1: Soc sci PhD graduates make more than non-PhD

- exciting to know it, right ?

example: age and fear

- Say we have a survey measuring people's fear of crime (0-15)
- H1: fear of crime increases with age

-
○

$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i = -2.365 + .232X_i$
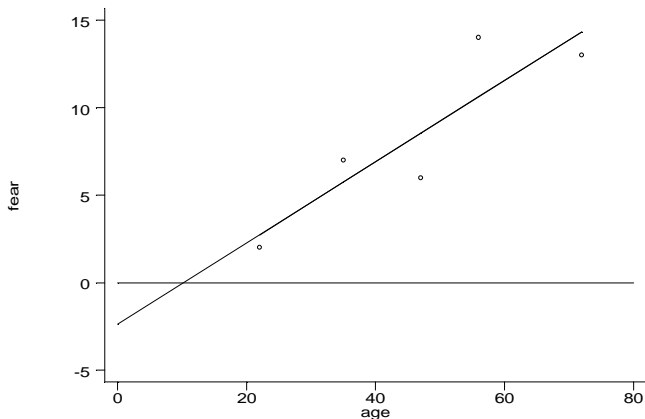
and why regression is better

- the regression advantage:

  look at many variables at the same time
- when just compare 2 means, the problem is that you are

  often comparing apples to oranges
- ○ say you compare income for males and females
- ○ but you need to take into account that females have kids...
- ○ females are discriminated, etc
- regression will take care of that–keep that in mind

examples

- see some of the useful things you can predict
- http://ianayres.yale.edu/prediction-tools

  eg life expectancy http://www.northwesternmutual.com/

  learning-center/the-longevity-game.aspx

outline

multivariate OLS

- multiple (multivariate) regression is the most common tool in social science

- it finds effect of a variable of interest (X) on the dependent variable (Y) controlling/holding constant other vars

- it's a statistical trick that makes sample equal on all characteristics that we control for and 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 is a function of X (same thing)
- $Y = f(X_1, X_2, ..., X_n, u)$
- in soc sci **always** many Xs

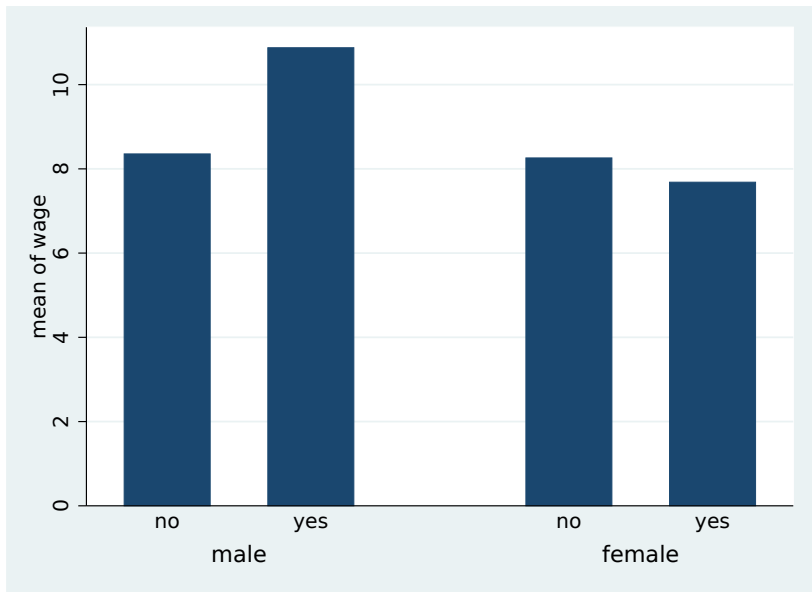## **outline**
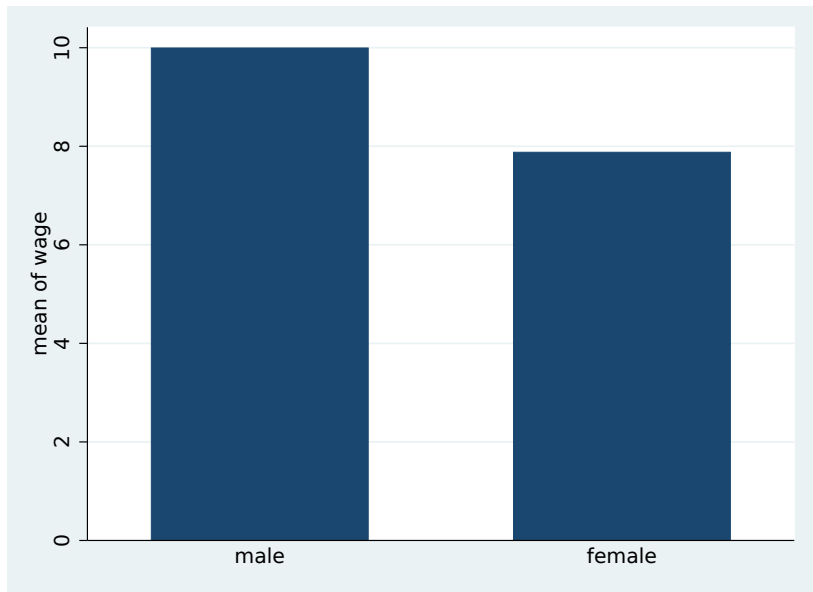
wages

wages by quintile of experience
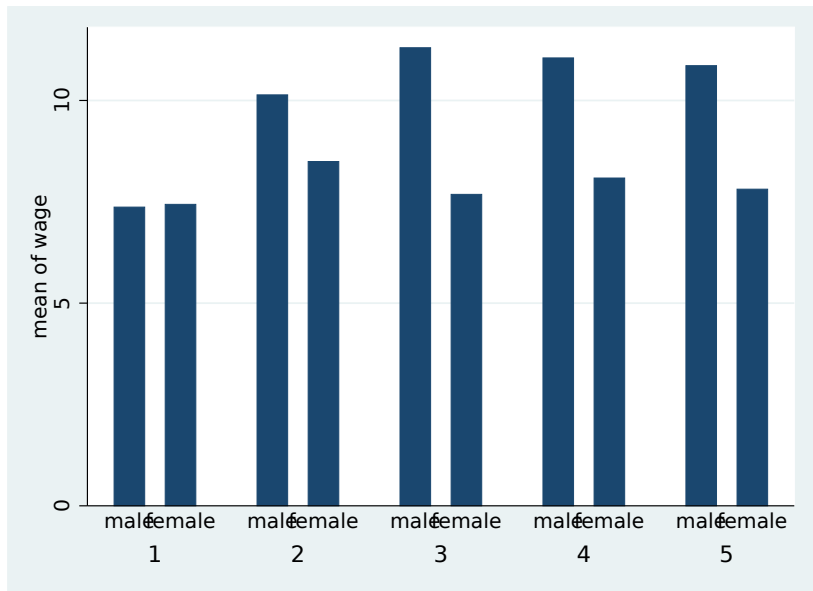
wages by marital status and experience

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
```
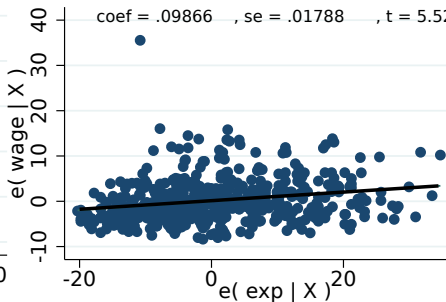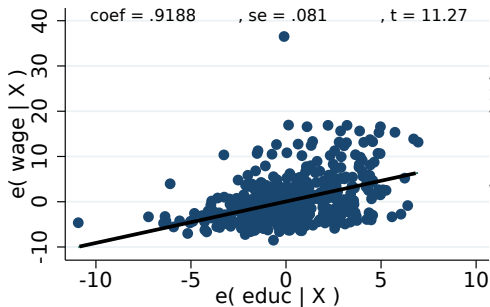
interpreting coefficients

- pretty much only one way to interpret reg correctly
- 1 unit (\$ % etc) increase in X leads to β unit (\$ % etc) increase/decrease in Y ($> 1X$: remember ceteris paribus!)
-
- and as per Wheelan ch11: focus on:
- sign
- size
- significance:
  ○ t-stat, t=coeff/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
-------------------------------------------------------
```


coef = .9188   , se = .081   , t = 11.27

coef = .09866   , se = .01788   , t = 5.52

- ext cre: why married insignificant?

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
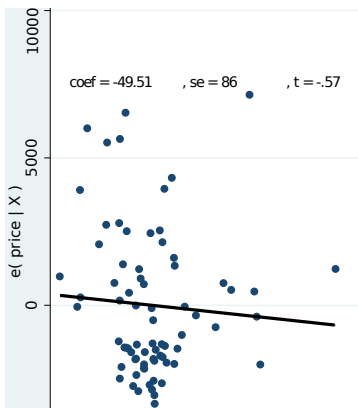○ and cannot really use them for much

interpret: β, 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
-------------------------------------------------------------------
```


-
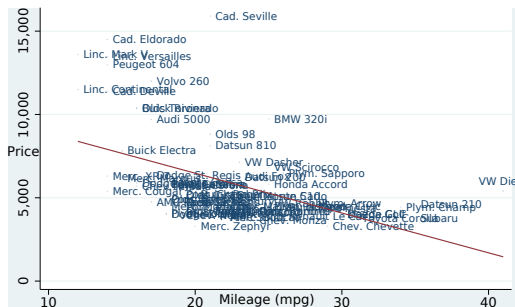
interpret: β, p, t, CI; predict price for 10mpg

```
-------------------------------------------------------------------
 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
-------------------------------------------------------------------
```


coef = -49.51   , se = 86   , t = -.57

coef = 1.7465   , se = .64   , t = 2.72
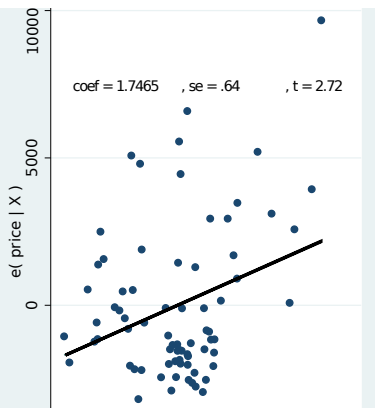
predicted values (p200 Wheelan, 2013)

- weight=-118+4.3*(height in)+.12*(age)-4.8*(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.northwesternmutual.com/learning-center/
  tools/the-longevity-game
- 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(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)
- $wage_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_n X_{ni} + u_i$

outline

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

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, corr$\neq$causation, etc
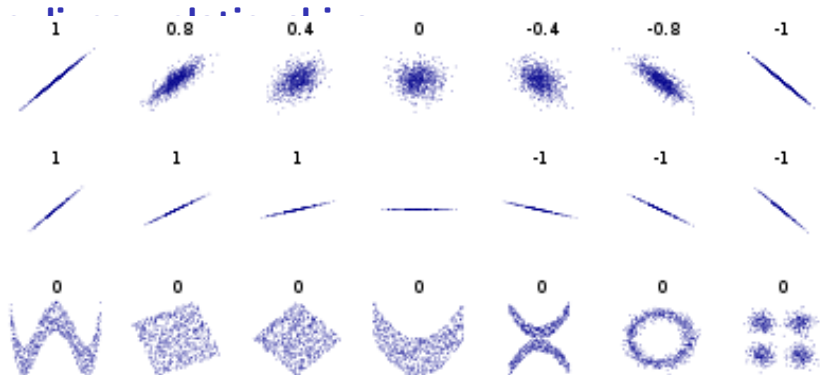
-
- like corr, won't detect nonlinear relationships!
- ○ example of nonlinear rel? extra credit!
- ○ [need quadratics, logs, etc] BUT!:
- just break it up into subsets/subsamples! dig deeper!
- ○ say for males and females separately
- ○ say for low and hi val separately

reverse causality (p216 Wheelan, 2013)

- more lessons—— >bad golf, or
- bad golf—— >more lessons
- solution:
○ lag variable: bad golf last month—— >more lessons now
○ use exogenous shock–remember from res_des.pdf:
○ terrorist attack—— >policing—— >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!
- draw say 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:
  //www.westmont.edu/~phunter/ma5/excel/regression.html
- http://www3.wabash.edu/econometrics/
  EconometricsBook/Basic%20Tools/ExcelAddIns/
  OLSRegression.htm
- http://finance.wharton.upenn.edu/~bodnarg/courses/
  readings/regression
- python
- http://www.learndatasci.com/
  predicting-housing-prices-linear-regression-using-python
- https://stackoverflow.com/questions/19991445/
  run-an-ols-regression-with-pandas-data-frame

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.