✳

◉ Admin

•

•

✳

Plan For Today → Using Regression in reality

→ Less maths, more intuition

→ Going through statsmodels a bit

Multiple Linear Regression

Most basic motivation → Lot of related variables effect our outcome

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

} Go through the notation

→ What is the interpretation of β_1? [Someone nailed it last time]

→ Why do we add more variables if all we care about is β_1?

◉ Error vs residuals. Calculating the residuals.
   + Homoskedasticity

- Closed form solution for β vector $(X^TX)^{-1}X^Ty$

- Add more variables so that we can <u>control</u> for them and find the <u>partial effect of X_1</u> (β_i)

TBH → regression is all about trying to find the best estimator for β_1 (the partial effect)

RULE:
- If X_1 and X_2 correlate
- AND X_2 effects Y causally
  → Including X_2 in your regression will change the value of β_1

$Y = \beta_0 + \beta_1 X_1 + \varepsilon$
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
} Different coefficient $\textcircled{!}$

Example:

$Y_i =$ Number of ice cream sold on a beach
$T_i =$ Temperature

$Y_i = \alpha_0 + \alpha_1 T_i + \varepsilon_i$ } Estimate of α_1 ?

$P_i =$ Number of people on the beach

$Y_i = \beta_0 + \beta_1 T_i + \beta_2 P_i + \varepsilon_i$ } Estimate of β_1 and β_2 ?

- Since T_i and P_i are related <u>AND</u> P_i affects Y_i, It should be included in the model
- Might make other coefficient insignificant

② How to do regression (No hard rule – Art not a science!)

Very Naive   →   Include all regressors   (BAD!)
- multicolinearity, regressors might have no effect

Naive   →   Do lots of combinations and pick the model where R^2 is the highest   [P-Hacking]
- Might not care about R^2
- R^2 always increases!   Adjusted R^2 better

Quite Naive   →   Do lots of combinations and pick the model with the lowest p-value on the β of interest   (okay.. getting better)
But   p-hacking   [P-Hacking]

Okay   →   Define a small list of regressions you will test and do theat. Compare them.

?   →   Forwards/Backwards Step-wise with rule
- Can be okay but flaws
- Might stop too early or late

Best Way   →   [Theory] to derive things ⊕
- Education and wage example

Example:
     want to measure
① I ~~know~~ education against wage
② I know experience and gender might be important

     Reg 1: $wage_i = \beta_0 + \beta_1 Ed_i + \beta_2 X_p + \beta_3 Male_i + \varepsilon_i$

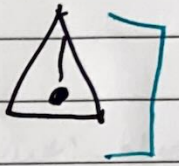③ I suspect ██ ~~sex~~ age might have an effect but I'm not sure
④ I also think type of school you went to might have an effect

→ Define more regression

→ Run _all_ simultaneously

→ Compare R^2, $\bar{R}^2$, p-values (which is the most appropriate model)?

$\Rightarrow$ How does β change.

Question? dichotomous/binary variables
How to model these

[③ Higher Order Regressors, Dummies and Interaction ⚠]

Categorical

→ See the notes a few pages ago

✱ W4-class (~~~~~)

[④ Python Notebook example ◆]

Question?

Polytomous
Variable
(k-1 dummies)

Gender + wages