IDEA: ONE BIG EXAMPLE

week 4, 2024

~ Week 3 – Lesson Plan ~

① Admin:

last year's notes. Due 11:59PM

- Problem Sheet due Wednesday 5pm alongside Python notebook
- Worked solution to the problem sheet now posted
  - → Proofs not included but easy to see online
  - → Please submit your ~~question~~ answer + self-marking sheet as a pdf.
- No update on summative datasets unfortunately. will aim to share list soon. If delay continues, will share provisional list
- Admin Questions ?

② Regression Introduction:

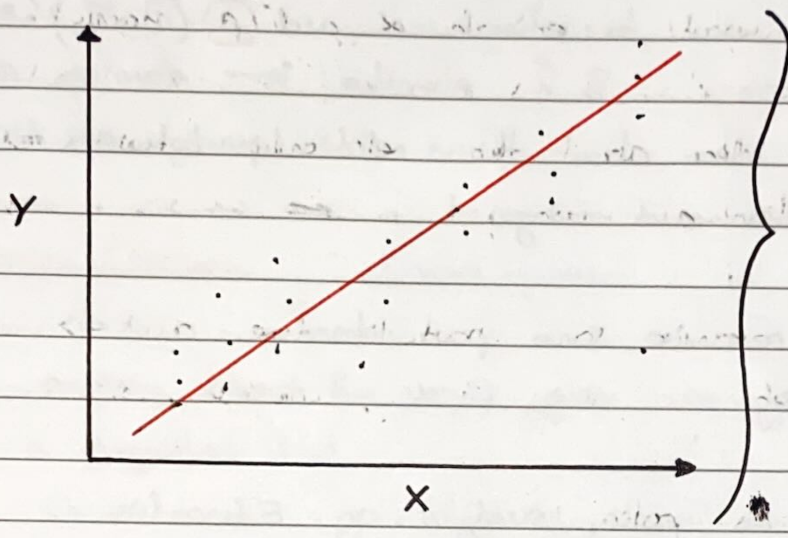
- Comment on the size of regression as a topic. HUGE
- Plan for today is to tackle a really small chunk of that by focusing on the fundamentals which are the same for all models
- Gateway to further models / topics
- Explore the fundamental topics through one big example simulating a basic research question (Mincerian regression)

③ Introduction to the example:

- General point of regression is to estimate the relationships between variables. To establish a causal relationship + example eg. weather
- To do this we impose a model on the world and use statistical techniques to estimate the parameters of that model

Example: Relationship between years of education, X, and wage aged 40, Y. (Mincerian regression – very famous). Plot it up on a graph → Famous Labour Economics Model (Min-sear-ian)

Dataset



this is your dataset!

Draw on graph

Assumed Population Model (Imposed) : $Y_i = \alpha + \beta X_i + \varepsilon_i$

{Capital Letters <=> Population variables} or $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Definition of α and β as population statistics!

Aim: Estimate that relationship

[We regression to estimate the α and β parameters!] → Drive This Message!

Definitions:

① α, β Population statistics (true given imposed model)

② $\hat{\alpha}, \hat{\beta}$ Estimated α, β. Sample statistics

① X_i, Y_i, ε_i the i^{th} example in the population

② x_i, y_i, ε_i the i^{th} example in the sample

* Dependent / independent variables

[**Aim:** As always in statistics! We our sample to try to make estimates / inference about the population]

So... we estimate $Y_i = \alpha + \beta X_i + \varepsilon_i$ with our sample

→ Meaning of α and β in terms of the population? Interpretation

→ Why do we want to estimate α and β (mainly)?

① We have no idea about the relationship between increased education and wages! ⇒ $\beta_{H_0} = 0$

② We want to reexamine some past literature and test their finding. ⇒ $\beta_{H_0} = 0.2$

③ We want to do policy analysis. eg. Education think tank suggest MASSIVE returns to education ⇒ $\beta_{H_0} = 1.0$

Goal: estimate β with our sample. Compare it to preconception

* Mainly interested in β as it is our causal effect *

④ OLS:

- Ordinary Least Squares regression (OLS) gives us a method of finding $\hat{\alpha}$ and $\hat{\beta}$ values. Under the Gauss-Markov assumption (not going to cover here) we can

  → Derive formula for $\hat{\alpha}, \hat{\beta}$

  → Derive the distribution for them!

Worth learning if you get time!

see extension

Good Extension proof

Derivation not covered here but can be done via Gauss' method or standard optimisation (in ∠N) Comment about the ∠N derivation being incomplete

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

n, 40 year old adults
x (education)
y (wages)

⑤ Estimation of $\hat{\beta}$:

→ Have your sample $\underbrace{x_1, y_1}_{①}, \underbrace{x_2, y_2}_{②}, \dots, \underbrace{x_n, y_n}_{ⓝ}$

→ Calculate $\bar{x}, \bar{y}$

→ Sub into formula → estimate $\hat{\beta}, \hat{\alpha}$. eg. $\hat{\beta} = 0.3$

(or just use software to do this!)

⑥ Making Claims / Inference

Causal inference!

- Can we say anything about the world given we got $\hat{\beta} = 0.3$? $\boxed{\text{No!}}$
- Need to do a hypothesis test

Logic of hypothesis testing [Important]

① Start with an assumption about the world, $\beta = 0$

② Given $\beta = 0$, what is the likelihood of getting my result by chance? ∴ Calculate prob

③ If the probability is sufficiently small (<0.05, <0.01) conclude prior belief was wrong

⚠ You never conclude that your new estimate is correct, Only that the old belief is likely to be wrong (and maybe
- where the true value might lie ⚠

- So, we want to work out the probability of getting $\hat{\beta} = 0.3$ given our prior belief (β_{H_0}) that $\beta = 0$.
- To do this we need a distribution
- $\hat{\alpha}, \hat{\beta}$ depend on our sample so have distribution (recall last lecture). The CLT means that we even know the distribution!

$$\boxed{\hat{\beta} \approx N(\beta, \mathrm{var}(\hat{\beta}))}, \quad \boxed{\mathrm{var}(\hat{\beta}) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}}$$

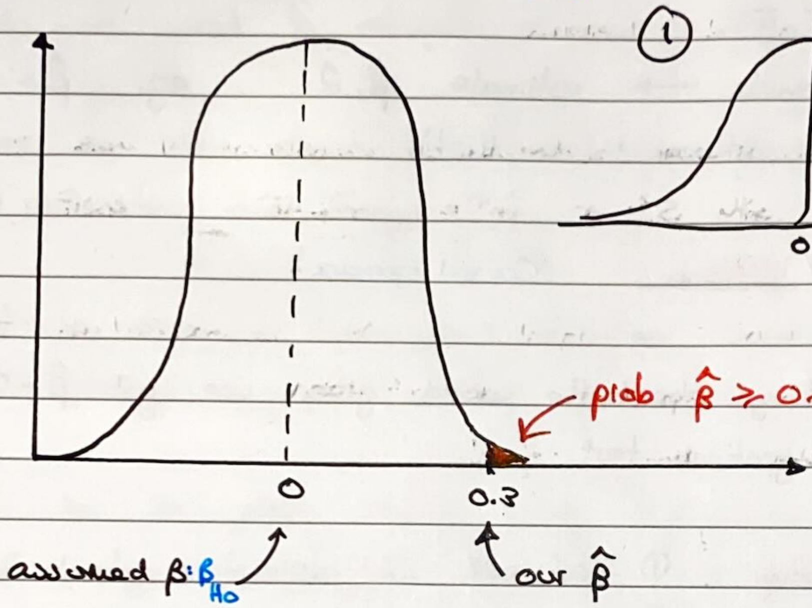
true β

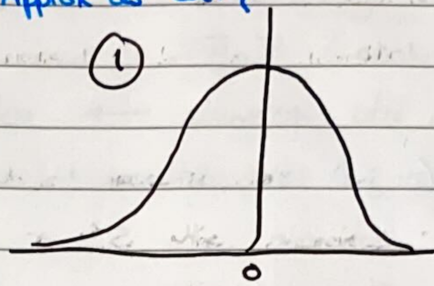
ε / y ← both!

σ (variance of ε), population

So, given our assumption β_{H_0} : $\beta = 0$, we can write

1. $\hat{\beta} \,\tilde{\approx}\, N(0, Var(\hat{\beta}))$ {Approx as CLT}





→ work out the probability via normal distribution knowledge (p-value)

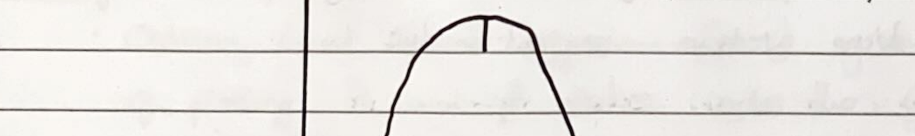
→ If prob < 0.05 ... $\boxed{\text{reject } \beta_0}$

↳ $\hat{\beta} \,\tilde{\approx}\, N(\beta_{H_0}, Var(\hat{\beta}))$

$$\frac{\hat{\beta} - \beta_{H_0}}{SE(\hat{\beta}_j)} \,\tilde{\approx}\, N(0,1)$$

(assuming large sample)

2. Alternatively, $\beta_{H_0} = 0.2$



⇒ Not sufficiently unlikely to reject 0.2

• Explains why we don't conclude