econ2300

INTRODUCTORY ECONOMETRICS — SUMMARY

Topic 1: Intro + Reviewing what we know

Key takeaways:

- correlation ≠ causation
- covariance = $\text{cov}(x,z)$
  $= E(X-u_x)(Z-u_z)$
  $= \delta_{xz}$
- $\bar{Y}$ = sample average
- correlation = $\text{corr}(x,z)$
  $= \dfrac{\text{cov}(x,z)}{\text{var}(x)\,\text{var}(z)}$
  $= \dfrac{\delta_{xz}}{\delta_x \delta_z}$
  $\{-1 \leq r \leq 1\} \rightarrow = r_{xz}$
- Law of large numbers: $\bar{Y} \xrightarrow{p} \mu_Y$ increasing sample size (i.e. increasing the sample size makes $\bar{Y}$ approach population average)
- Central Limit Theorem (CLT): Increasing sample size will make $\bar{Y}$ (sample mean) more normally distributed.

Causal effects + statistical concepts

Economics suggests relationships, but never quantitative magnitudes of causal effects $\longrightarrow$ ∴ in this course we want to find out

Causal effect: changing 'a' causes ___ change in 'b'

- we could use experiments to measure quantitatively
  BUT they are hardly ever feasible (expensive, time consuming etc.)
- so we use observational data

Using observational data to estimate causal effects has difficulties:

- confounding effects
- simultaneous causality
- correlation ≠ causation

} we will unpack these later

Discrete: Probability of a specific value of Y (e.g. $\Pr(Y=650)$)
Continuous: Probability of Y being between values (e.g. $\Pr(640 \leq Y \leq 660)$)

Why use $\bar{Y}$ to estimate μ_Y?

- $\bar{Y}$ is least squares estimator of μ_Y*
- $\bar{Y}$ is unbiased: $E(\bar{Y}) = \mu_Y$
- $\bar{Y}$ is consistent: $\bar{Y} \xrightarrow{p} \mu_Y$
- $\bar{Y}$ has smaller δ than all other linear unbiased estimators

* Least squares estimation creates a linear regression to fit a set of data points such that the sum of the squared residuals (i.e. distance in terms of y between the data points and regression line) is minimised

Estimation, Hypothesis tests & Confidence intervals

Statistical inference: using observational data to draw conclusions about the population.
↳ the first step to obtaining inferences about the population is <u>estimation</u>

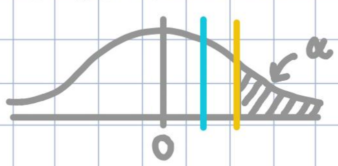
Estimation: * Estimators include values such as $\bar{y}$ (sample mean), Y_1 (1st observation), median, mode etc.
↳ We usually use $\bar{y}$ * using OLS to estimate slope

Hypothesis Testing:
- null hypothesis (H_0) always includes equals ($=, \leq, \geq$)
- alternative hypothesis (H_1) is an inequality ($\neq, >, <$)

Steps: ① HYPOTHESIS

2 sided	1 sided >	1 sided <
$H_0: E(Y) = \mu_Y$	$H_0: E(Y) \leq \mu_Y$	$H_0: E(Y) \geq \mu_Y$
$H_1: E(Y) \neq \mu_Y$	$H_1: E(Y) > \mu_Y$	$H_1: E(Y) < \mu_Y$

② DECISION RULE Reject H_0 if.... $p < \alpha$ (p-value method)
|test statistic| > |critical value| (t-test, z test)



t crit for α
t calc

* 1 sided > *
(in this case, t calc < tcrit ∴ don't reject H_0)

③ TEST STATISTIC (calculate)

$$\text{test statistic (e.g. t calc)} = \frac{\text{estimate} - \text{hypothesized value}}{\text{std. error of estimate}}$$

④ DECISION Either reject or fail to reject H_0
↳ Find critical value from table of critical values

⑤ CONCLUSION "At the ____ level of significance (LOS) there is [sufficient / insufficient] evidence to conclude that [statement about H_1]"

Confidence Interval: A 95% confidence interval for μ_Y is an interval that contains the true value of μ_Y in 95% of repeated samples.
↳ i.e... set of values of μ_Y NOT rejected by a hypothesis test with a 5% LOS.

$\mu: \bar{x} \pm z_{crit}(\sigma/\sqrt{n})$ $\mu: \bar{x} \pm t_{crit}(s/\sqrt{n})$ $p: \hat{p} \pm z_{crit}(\sigma_{\hat{p}}) \rightarrow \sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

"based on the sample, [popn. parameter] is estimated with [LOC]% confidence to be between ____ and ____."

Topic 2: Linear Regression with one regressor

Key Takeaways: $Y_i = \beta_0 + \beta_1 X_i + u_i$
- β_0 = intercept
- β_1 = slope
- u_i = regression error (omitted factors)

* 'hat' (^) shows it is an estimator

R^2: fraction of variance of Y explained by X (percentage, unitless)
($0 \leq R^2 \leq 1$)

Entity Demand OLS regression:

$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it}$ where $\tilde{x}$ denotes x minus avg. value of x ($\bar{x}$)
i.e. $Y - \bar{Y} = \beta_1 (X - \bar{X}) + (u - \bar{u})$

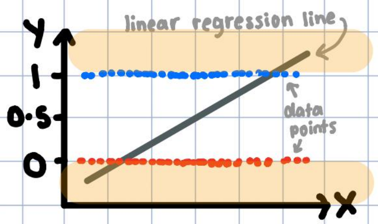
LSA #1: $E(u_{it} | X_{i1}, \ldots, X_{iT}, \alpha_i) = 0$

u_{it} has mean zero, given the entity fixed effect (α) and entire history of X's
- extension of previous LSA #1
- no omitted lagged effects
- no feedback from u to future X

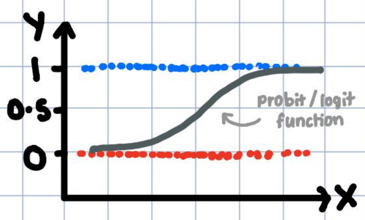
LSA #2: $(X_{i1}, \ldots, X_{iT}, u_{i1}, \ldots, u_{iT})$ are i.i.d
- satisfied if entities are randomly sampled from population using SRS
- does NOT require observations to be i.i.d. (random) for the same entity over time – that would be unrealistic!
  (e.g. high beer tax this year means probably high next year too – correlation)
  ∴ $cov(u_t, u_{t+1})$ is often plausibly non-zero

Topic 9: Regression with a binary dependant variable

Key takeaways: $Y = \beta_0 + \beta_1 X + u$ → Y = binary (i.e. = 0 or 1)
↳ can still use OLS regression as long as conditional mean independence is satisfied!



β_1 = Δ predicted probability that $Y=1$, for unit ΔX
BUT issue with using linear model is that we recieve probabilities outside the 0-100% range (can't interpret)
∴ we use probit or logit! $(0 \leq Pr(Y=1|X) \leq 1)$



Probit: $Pr(Y=1|X) = \Phi(Z)$ where... $Z = \beta_0 + \beta_1 X$
- use normal distribution Z table to find probability

Logit: $Pr(Y=1|X) = \frac{1}{1+e^{-Z}}$ ← Just plug Z in to the equation to find probability!

Measures of fit for probit + logit:
- R^2 and $\bar{R}^2$ doesn't make sense for binary Y (data points arent ON regression line)

∴ instead we use...

1. Fraction correctly predicted
   fraction of Y for which predicted prob = >50% when actual Y=1 (or <50% when Y=0)
2. Pseudo R^2
   improvement in value of log-likelihood, relative to having no X's
   want to maximise

In linear probability model, predicted value of Y is interpreted as predicted probability that Y=1...
- if LSA #1 holds, $E(Y|X) = Pr(Y=1|X) = \beta_0 + \beta_1 X$
- OLS is valid, given LSA's hold, and yields unbiased

For linear, probit or logit models... we can still use P-values in the same way
- ALSO, if chi^2 is large, it indicates statistical significance