Comprehensive Inference & Regression Review

Clase 13, 14.310x: Data Analysis for Social Scientists

B.Sc. Cristian Lazo Quispe

BREIT (Brescia Institute of Technology) · Aporta, Grupo Breca

SDS Programme (Statistics and Data Science, MITx) · **BREIT** (Brescia Institute of Technology), an Aporta initiative –Grupo Breca– in partnership with MIT IDSS.

clazoq@uni.pe · linkedin.com/in/cristian-lazo-quispe · github.com/CristianLazoQuispe

How does today work?

The format

This is the second half of our review. Class 12 covered **probability** (Classes 1–5); today we cover **inference, causality and regression** (Classes 6–11): confidence intervals, p-values, randomized experiments and the ATE, difference-in-differences, OLS, dummy variables, omitted variable bias and instrumental variables. The main block is **15 problems in one single numbering**, ramping up: Easy (1–4) $\rightarrow$ Intermediate (5–12) $\rightarrow$ Hard (13–15). Even the Hard ones just *apply* a formula we have already used, no derivations. **Four of them are worked in R on a `.csv` file**: see the next slide. Separately, **5 bonus problems** (B1–B5) if we finish early.

What you do

Each slide shows a problem and a blank space: **work it out** there (or on paper) before we move on. No hints, no answer key on these slides.

What I do

For each problem I will read through the **theory used**, the **thought process**, and the **step-by-step solution** out loud, right after we attempt it together.

What we will cover

1. Easy problems, 1 to 4
2. Intermediate problems, 5 to 12
3. Hard problems, 13 to 15
4. Bonus problems, 1 to 5

Which problems need the laptop

Four problems read a `.csv` file in R

They are **not** a separate block at the end: each one sits right after the pen-and-paper problem it extends. Open RStudio in `clase13/practica_R/`.

#	Tier	File it reads	Skeleton script	What you compute
2	Easy	`data/rct.csv`	`p02_ate_lm.R`	ATE and `lm(outcome ~ treatment)`
6	Intermediate	`data/rct.csv`	`p06_permutation.R`	p-value by `t.test` and by permutation
8	Intermediate	`data/did.csv`	`p08_did.R`	DiD from a 2×2 table and from `lm`
11	Intermediate	`data/wages.csv`	`p11_dummies_ovb.R`	dummies, short vs. long, OVB

The other eleven problems in the main block are pen and paper.

How to work them

Try it **from scratch** with just the slide. Stuck? Open the skeleton `pNN_*.R`, same code with blanks (`____`) to fill in. Afterwards, the full answers are in `pNN_*_sol.R`.

Contents

Problem 1 · Easy · Treatment, control and the ATE (Class 8)

Problem

A regional health office ran a **randomized** chlorination pilot in 10 villages. $W_i = 1$ means the village received the treatment, $W_i = 0$ means it did not. The outcome Y_i is the percentage of households with safe drinking water at the end of the pilot.

Village	1	2	3	4	5	6	7	8	9	10
W_i	1	1	1	1	1	0	0	0	0	0
Y_i	62	55	71	58	64	50	44	53	47	51

Question. (a) Compute `treatment_avg` and `control_avg`. (b) Report the estimated **ATE**. (c) Village 3 has $Y_3 = 71$. What is $Y_3(1)$, and what is $Y_3(0)$? (d) Why does random assignment let us read the ATE as *causal*?

Your work

Problem 2 · Easy · Read a CSV, compute the ATE, fit `lm` (Classes 6–8)

Problem

`data/rct.csv`: 300 households in a nutrition-counselling pilot, with the visits assigned **at random**.

`id`	household identifier
`treatment`	1 = received counselling, 0 = control
`baseline`	dietary-diversity score *before*
`outcome`	dietary-diversity score *after*

Find. (a) `treatment_avg`, `control_avg` and the **ATE**. (b) Check the randomization worked, using `baseline`. (c) Fit `lm(outcome ~ treatment)`: what are the intercept and the slope, and why is the slope the ATE? (d) Now fit `lm(outcome ~ treatment + baseline)`: what changes, the estimate or the standard error? (e) Interpret the ATE in one sentence.

Your work

Problem 3 · Easy · Standard error and confidence interval (Class 6)

Problem

A household survey draws a random sample of $n = 64$ families in a district. Monthly food spending has sample mean $\bar{X} = 480$ soles and sample standard deviation $s = 96$ soles.

Question. (a) Compute the **standard error** of $\bar{X}$. (b) Build a 95 percent confidence interval for the population mean (use 1.96). (c) A colleague says: "there is a 95 percent probability that the true mean lies between those two numbers." Is that the right reading? If not, state the correct one.

Your work

Problem 4 · Easy · Reading a simple regression (Class 9)

Problem

Using $n = 500$ workers, a labour economist estimates by OLS

$$\widehat{\text{wage}} = 6.40 + 1.35\,\text{schooling}, \qquad R^2 = 0.28,$$

where *wage* is in soles per hour and *schooling* is completed years of education.

Question. (a) Interpret the intercept 6.40, and say whether it is meaningful here. (b) Interpret the slope 1.35, **with units**. (c) What does $R^2 = 0.28$ tell us? (d) Predict the wage of a worker with 11 years of schooling. (e) Is 1.35 the causal return to a year of schooling? Why or why not?

Your work

Contents

Problem 5 · Intermediate · The p-value of an experiment (Classes 6–7)

Problem

A tutoring program was randomly assigned to students. End-of-year test scores:

	n	mean	s
Treatment	50	78.4	12.0
Control	50	73.0	11.0

Question. (a) State H_0 and H_1 for "the program had no effect". (b) Compute the standard error of the difference in means. (c) Compute the t statistic. (d) With roughly 97 degrees of freedom the two-sided p-value is 0.021. What do you conclude at the 5 percent level? (e) Would you conclude anything different at the 1 percent level?

Your work

Problem 6 · Intermediate · Read a CSV, p-value by permutation (Classes 6–8)

Problem

Same file as Problem 2, `data/rct.csv`. The estimated ATE is 2.0 points. Now attach a p-value to it, **two different ways**.

Find. (a) The p-value from a two-sample `t.test` (pass the treated and control vectors, treated first, so the signs match the ATE). (b) The p-value by **permutation**: reshuffle the treatment label with `sample()` 10,000 times, recompute the difference in means each time, and count how many are at least as extreme as the one observed. (c) Plot the 10,000 reshuffled differences and mark the observed value: what is the histogram showing? (d) Why do the two p-values agree here? (e) Name a situation where they would **not**, and say which one you would trust.

Your work

Problem 7 · Intermediate · Difference-in-differences by hand (Class 9)

Problem

A conditional-transfer program was launched in some regions and not in others. Mean secondary-school enrolment (percentage points):

	Before	After
Treated regions	61.0	72.5
Control regions	58.0	64.0

Question. (a) A colleague reports the effect as $72.5 - 61.0 = 11.5$. What is wrong with that? (b) Another reports $72.5 - 64.0 = 8.5$. What is wrong with *that*? (c) Compute the **difference-in-differences**. (d) In the regression $Y = \beta_0 + \beta_1\,\text{treated} + \beta_2\,\text{post} + \beta_3(\text{treated} \times \text{post}) + u$, give the numerical value of all four coefficients.

Your work

Problem 8 · Intermediate · Read a CSV, fit the difference-in-differences (Class 9)

Problem

`data/did.csv`: 800 school-district observations, two periods.

`region`	Norte / Centro / Sur
`treated`	1 = district got teacher training
`post`	1 = after the program started
`outcome`	share of pupils at grade level (pp)

Find. (a) The 2×2 table of cell means (`tapply`). (b) The DiD **by hand** from those four numbers. (c) The DiD from `lm(outcome ~ treated + post + treated:post)`, and confirm it is the same number. (d) What `treated` and `post` measure separately. (e) The 95 percent confidence interval for the DiD, and whether the effect is distinguishable from zero.

Your work

Problem 9 · Intermediate · Dummy variables with three categories (Class 9)

Problem

A wage regression on Peruvian data, with region entered as dummies:

$$\widehat{\text{wage}} = 8.20 - 1.75\,\text{Sierra} - 2.40\,\text{Selva} + 1.10\,\text{schooling}.$$

Sierra and *Selva* are 0/1 indicators; the three regions are Costa, Sierra and Selva.

Question. (a) Which region is the **baseline**, and how do you know? (b) Interpret -1.75 in one sentence. (c) Predict the wage of a Sierra worker with 12 years of schooling. (d) What is the estimated Sierra – Selva wage gap? (e) Why can we **not** include all three region dummies together with the intercept?

Your work

Problem 10 · Intermediate · The omitted variable bias formula (Class 10)

Problem

The **long** (correctly specified) model for hourly wages is

$$\text{wage} = \beta_0 + \beta_1\,\text{schooling} + \beta_2\,\text{ability} + u,$$

with $\beta_1 = 1.10$ and $\beta_2 = 0.85$. Ability is **not observed**, so a researcher runs the **short** regression of wage on schooling alone. In the auxiliary regression of ability on schooling the slope is $\delta = 0.40$.

Question. (a) Write the OVB formula for $\mathbb{E}[\tilde{\beta}_1]$. (b) Compute its numerical value. (c) How big is the bias, and is it **upward** or **downward**? (d) Explain in plain words why the short regression gets it wrong in that direction.

Your work

Problem 11 · Intermediate · Read a CSV, dummies and omitted variable bias (Classes 9–10)

Problem

`data/wages.csv`: 400 workers, with `wage`, `schooling`, `experience`, `region` (Costa / Sierra / Selva) and `ability`. Here `ability` *is* observed, which is exactly what lets us *see* the bias that normally hides.

Find. (a) Run `lm(wage ~ schooling + experience + factor(region))`: which region is the **baseline**, and what does each region coefficient mean? (b) The predicted Costa to Selva gap. (c) The **short** regression (omitting `ability`) and the **long** one: how much does the schooling coefficient move? (d) The **auxiliary** regression `lm(ability ~ schooling + experience)` to get δ, and verify bias $= \hat{\beta}_{\textbf{ability}} \cdot \hat{\delta}$. (e) Is the short regression over- or under-stating the return to schooling?

Your work

Problem 12 · Intermediate · Reading a full regression table (Classes 9–10)

Problem

OLS on $n = 240$ workers; dependent variable is the hourly wage in soles.

	Estimate	Std. Error
(Intercept)	4.812	1.205
schooling	1.104	0.212
experience	0.318	0.061
female	−2.470	0.930

$R^2 = 0.341$, $\quad F(3, 236) = 40.72$. Use $t_{0.975,236} \approx 1.97$.

Question. (a) Compute the t statistic for *female*. (b) Is it significant at 5 percent? (c) Build a 95 percent confidence interval for the *schooling* coefficient. (d) Interpret the *female* coefficient. (e) The individual t-tests and the F-test answer **different** questions, which is which?

Your work

Contents

Problem 13 · Hard · Logs and an interaction (Class 10)

Problem

With *urban* a 0/1 indicator, OLS gives

$$\widehat{\log(\text{wage})} = 1.82 + 0.062\,\text{schooling} + 0.145\,\text{urban} + 0.021\,(\text{schooling} \times \text{urban}).$$

Question. (a) Why is the dependent variable in logs, what does a coefficient mean now? (b) Give the approximate return to one more year of schooling for a **rural** worker, and for an **urban** worker. (c) Interpret the 0.145 carefully: for *whom* is it the urban premium? (d) What is the urban premium for a worker with 10 years of schooling? (e) The exact percentage change is $100(e^{b} - 1)$. For which of your answers does the shortcut $100b$ start to hurt?

Your work

Problem 14 · Hard · Vouchers by lottery: the Wald estimator (Class 11)

Problem

A city offered private-school vouchers by **lottery**. $Z = 1$ if a family won, $D = 1$ if the child attended a private school, Y = test score.

	$Z = 1$ (won)	$Z = 0$ (lost)
$\mathbb{E}[D \mid Z]$, share attending	0.62	0.17
$\mathbb{E}[Y \mid Z]$, mean score	46.80	44.10

OLS of Y on D gives 9.50. You may use

$$\hat{\beta}_{\text{IV}} = \frac{\text{reduced form}}{\text{first stage}} = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[D \mid Z = 1] - \mathbb{E}[D \mid Z = 0]}.$$

Question. (a) Compute the **first stage**. (b) Compute the **reduced form**. (c) Compute the **Wald estimate**. (d) Why is OLS (9.50) so much bigger, in one sentence?

Your work

Problem 15 · Hard · Capstone: reading a policy evaluation (Classes 6–11)

Problem

A cash transfer was **randomly assigned** to 1,200 households. Dependent variable: log(income).

	Estimate	Std. Error
(Intercept)	6.900	0.081
treated	0.118	0.036
female_head	−0.145	0.041
treated × female_head	0.087	0.058
schooling	0.049	0.007

Use 1.96 for the 95 percent critical value.

Question. (a) Compute the t statistic for *treated*: did the program work? (b) Build its 95 percent confidence interval. (c) What is the total effect for a **female-headed** household? (d) Is the interaction significant? (e) The transfer was randomized, schooling was not. Which coefficients can be read as **causal**?

Your work

Contents

Bonus 1 · Easy · Setting up a test (Class 6)

Problem

A ministry claims the average commute in a city is 45 minutes. A random sample of $n = 100$ commuters gives $\bar{X} = 47.8$ minutes with $s = 14$ minutes.

Question. (a) State H_0 and H_1 for testing the claim. (b) Compute the standard error and the z statistic. (c) The two-sided p-value is 0.0455. Decide at the 5 percent level. (d) Had the alternative been one-sided ($\mu > 45$), what would the p-value be? (e) Define the **Type I error** in the language of this problem.

Your work

Bonus 2 · Intermediate · Decomposing a puzzling comparison (Class 8)

Problem

A voluntary job-training program is evaluated by comparing annual earnings of participants and non-participants. Participants earn **400 soles less** on average. Yet an experimental study of the same program finds the effect on participants (the ATT) is +600 soles.

Question. (a) Write the decomposition
observed difference = ATT + selection bias and identify each term.
(b) Compute the selection bias. (c) In words, what does a selection bias of that sign and size say about who signs up? (d) What would the observed difference have been if the program had been randomized instead?

Your work

Bonus 3 · Intermediate · A quadratic experience profile (Class 10)

Problem

A wage regression with a quadratic in experience gives

$$\widehat{\text{wage}} = 5.20 + 0.680\,\text{exp} - 0.017\,\text{exp}^2.$$

Question. (a) Why would anyone add the squared term, what shape is being allowed? (b) Write the marginal effect of one more year of experience as a function of *exp*. (c) Compute it at $\text{exp} = 5$ and at $\text{exp} = 25$. (d) Find the **turning point**. (e) Is it sensible to interpret the coefficient 0.680 on its own? Why not?

Your work

Bonus 4 · Hard · Pooled versus within (Class 10)

Problem

Test scores and class sizes in three schools, two classes each:

School	A		B		C	
Class size	18	22	26	30	34	38
Score	61	58	70	67	79	76

Pooling all six classes, OLS of score on class size gives a slope of $+0.96$.

Question. (a) Compute the slope *within* school A. (b) Do the same for B and C, what do you notice? (c) The pooled slope is positive but every within-school slope is negative. Explain the contradiction. (d) What does adding school **fixed effects** (a dummy per school) do, and which slope does it recover?

Your work

Bonus 5 · Hard · OLS by hand (Class 9)

Problem

Five districts. X = health posts per 10,000 inhabitants, Y = prenatal-checkup coverage index.

X	1	2	3	4	5
Y	3	4	8	9	11

Question. (a) Compute $\bar{X}$, $\bar{Y}$, $S_{XY} = \sum(x_i - \bar{x})(y_i - \bar{y})$ and $S_{XX} = \sum(x_i - \bar{x})^2$. (b) Compute $\hat{\beta}_1$ and $\hat{\beta}_0$. (c) Show $\hat{\beta}_1$ equals $\mathrm{Cov}(X, Y)/\mathrm{Var}(X)$. (d) Compute the five residuals and verify they sum to zero. (e) Compute R^2 from SST and SSR.

Your work

That's a wrap

Recap

15 problems, four of them worked in R on a CSV file, plus 5 bonus, across estimation & confidence intervals, hypothesis testing & p-values, randomized experiments & the ATE, difference-in-differences, OLS and its interpretation, dummy variables, omitted variable bias and instrumental variables, the inference and regression core of Classes 6–11.

Together with Class 12

Class 12 covered the probability half (Classes 1–5). With today, the whole course has been reviewed end to end.

Let's go through the solutions together.