14.310x – Data Analysis for Social Scientists

MicroMasters en Statistics and Data Science (SDS) – MITx

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

Clase 13

Comprehensive Inference & Regression Review

Solucionario (Read-Aloud Solutions)

Módulos oficiales del MicroMaster:
M6–M10 – Assessing & Deriving Estimators, Confidence Intervals & Hypothesis Testing, Causality & Randomized Experiments, Single & Multivariate Linear Models, Practical Regression & Omitted Variable Bias, Endogeneity & Instrumental Variables

B.Sc. Cristian Lazo Quispe
clazoq@uni.pe
linkedin.com/in/cristian-lazo-quispe · github.com/CristianLazoQuispe

Material del *Advanced Program in Data Science & Global Skills*, diseñado y desarrollado por **BREIT (Brescia Institute of Technology)**, iniciativa de **Aporta** —plataforma de innovación e impacto social del **Grupo Breca**— en alianza con el **MIT IDSS** (Institute for Data, Systems, and Society).

Basado en MIT 14.310x (E. Duflo & S. Ellison) y en diversas fuentes abiertas (ver bibliografía al final), bajo licencia CC BY-NC-SA 4.0.

18 de agosto de 2026

How to use this document

This is the **read-aloud solutions** companion to the Clase 13 question deck (`slides_clase13.tex`). Each problem below follows the same three-part structure: **theory used** (the tool being invoked), **thought process** (how to recognize which tool applies, before any computation), and a **step-by-step solution** with a boxed final answer.

The main block is **15 problems** in one single numbering, in increasing order of difficulty: ■ Easy: 1–4 ■ Intermediate: 5–12 ■ Hard: 13–15. Separately, **5 bonus problems** (B1–B5) are held in reserve in case we finish early.

Even the Hard ones are *applications* of a formula we have already used earlier in the session, not derivations: logs with an interaction, the Wald ratio from four numbers, and a regression table read row by row.

Four of them are worked in R on a `.csv` file, and they are not a separate block: they sit inside the Easy and Intermediate sections, right after the pen-and-paper problem they extend. This is the format the MicroMasters exam uses, which is why they are rated Easy and Intermediate.

#	Tier	File read	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 with `factor()`, short vs. long, OVB

The CSV files live in `clase13/data/` and are generated by `clase13/figs/gen_data.R` with fixed seeds. The intended flow is the one we have used since Class 10: **try it from scratch** with only the statement on the slide; **if you get stuck**, open the skeleton `clase13/practica_R/pNN_*.R`, which has the same code with blanks (`____`) to fill in; **afterwards**, check against this document and the runnable `pNN_*_sol.R`. Run everything from inside `clase13/practica_R/`, since the scripts read `../data/`.

Together with Clase 12 (which reviewed the probability half of the course, Classes 1–5), this session closes the loop: today covers **inference, causality and regression**, Classes 6–11.

Easy problems (warm-up)

Problem 2 is worked in R on a `.csv` file; Problems 1, 3 and 4 are pen and paper.

1. Problem 1 · Chlorination Pilot: Treatment, Control and the ATE

Problem 1 – Easy

Class 8 – Randomized experiments, ATE

Setup. Ten villages, randomly split: five treated ($W_i = 1$), five control ($W_i = 0$). Outcome Y_i = percentage of households with safe water. Treated: $62, 55, 71, 58, 64$. Control: $50, 44, 53, 47, 51$.

Theory used

Each unit has two **potential outcomes**: $Y_i(1)$ if treated and $Y_i(0)$ if not. Only one is ever observed, that is the *fundamental problem of causal inference*. The estimand is the average treatment effect

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)].$$

When treatment is assigned **at random**, $(Y_i(1), Y_i(0)) \perp\!\!\!\perp W_i$, and therefore

$$\widehat{\text{ATE}} = \bar{Y}_{\text{treat}} - \bar{Y}_{\text{control}}$$

is unbiased for the ATE.

Thought process

The word *randomized* is the whole problem. Without it, the difference in means would only be a description, treated and control villages could differ in wealth, altitude, prior infrastructure. With it, the two groups are *exchangeable* in expectation, so the difference in means estimates a causal quantity. The arithmetic is trivial; the licence to call it causal is what we are really practising.

Step-by-step solution

(a) Group averages.

$$\texttt{treatment\_avg} = \frac{62 + 55 + 71 + 58 + 64}{5} = \frac{310}{5} = 62,$$

$$\texttt{control\_avg} = \frac{50 + 44 + 53 + 47 + 51}{5} = \frac{245}{5} = 49.$$

(b) ATE.

$$\widehat{\text{ATE}} = 62 - 49 = \boxed{13 \text{ percentage points}}.$$

On average the chlorination program raised the share of households with safe water by 13 percentage points.

(c) Potential outcomes for village 3. Village 3 was treated, so what we see is the treated potential outcome: $Y_3(1) = \boxed{71}$. Its untreated potential outcome $Y_3(0)$ is **unobserved**, it is the counterfactual, the missing half of the pair. We can never fill it in for village 3; the ATE recovers the *average* of those differences, never any individual one.

(d) Why causal. Because assignment was random, W_i carries no information about a village's characteristics. So the control group's average, 49, is a valid stand-in for what the treated villages

would have averaged without the program. That substitution is exactly what breaks down in observational data, where the comparison group differs systematically.
Sanity check: every treated value exceeds every control value except the pair 55 vs. 53, consistent with a solid positive effect, and a warning that with $n = 5$ per arm we should still ask for a p-value (Problems 5 and 6 do exactly that for a comparison of this kind).

2. Problem 2 · Read `rct.csv`, compute the ATE and fit `lm`

Problem 2 – Easy

Classes 6–8 – `data/rct.csv`, 300 households

Setup. `rct.csv` has 300 households from a nutrition-counselling program assigned at random. Columns: `id`, `treatment` (0/1), `baseline` (dietary-diversity score before), `outcome` (score after). Compute the ATE two ways and check the randomization worked.

Theory used

Same logic as Problem 1, now on a file. Two facts do all the work here: with random assignment the difference in group means *is* the ATE, and regressing the outcome on a single 0/1 dummy returns exactly that difference, because a line through two groups can only pass through their two means.

Thought process

Balance checks reverse the usual logic of testing. Everywhere else a large p-value is disappointing. Here it is reassuring, because it means treatment and control are indistinguishable on a variable measured *before* the program, which is what a working randomization should produce.

Step-by-step solution

R code

```
data <- read.csv("../data/rct.csv")
head(data)

treatment_avg <- mean(data$outcome[data$treatment == 1])
control_avg   <- mean(data$outcome[data$treatment == 0])
ate           <- treatment_avg - control_avg

t.test(baseline ~ treatment, data = data)     # balance check (pre-treatment)

m1 <- lm(outcome ~ treatment, data = data)              # the ATE as a slope
m2 <- lm(outcome ~ treatment + baseline, data = data)   # add the covariate
summary(m1); summary(m2)
```

(a) Group means and the ATE.

$$\texttt{treatment\_avg} = 51.1927, \quad \texttt{control\_avg} = 49.1927, \quad \widehat{\text{ATE}} = \boxed{2.0000 \text{ points}}.$$

(b) Balance check. On `baseline`: means 41.85 (treated) versus 41.74 (control), $t = -0.13$, $p = 0.896$. A large p-value here is **good news**, since the two groups are indistinguishable before the program.
(c) The same number from `lm`. `lm(outcome ~ treatment)` gives

	Estimate	Std. Error	t
(Intercept)	49.1927	0.6101	80.63
treatment	**2.0000**	0.8628	2.32

The slope is exactly the ATE, and the intercept is exactly the control mean. It **must** come out that way: with a single 0/1 regressor the fitted line passes through the two group means, so the intercept is the control mean and the slope is whatever gets you to the treated mean ($49.1927 + 2.0000 = 51.1927$). "Regression on a dummy" and "difference in means" are the same computation written twice.
(d) Adding the covariate. `lm(outcome ~ treatment + baseline)` gives a treatment coefficient of 1.9365 with SE = 0.7140. The estimate barely moves, because it was already unbiased, but the standard error falls from 0.8628 to 0.7140. That is the reason to control for pre-treatment variables in a randomized experiment: **precision, not bias**. Note also that `baseline` would be a legitimate control in any case, since it was measured before treatment and so cannot be a bad control.
It is worth seeing what that precision buys: the p-value falls from 0.0211 to 0.0071. Same effect, same data, but with the covariate the result also clears the 1 percent bar. Precision changes conclusions.
(e) Interpretation. The counselling program raised the dietary-diversity score by 2.0 points on average. Because assignment was random, that is a causal statement, not just an association.

3. Problem 3 · Food Spending: Standard Error and Confidence Interval

Problem 3 – Easy

Class 6 – Estimation, confidence intervals

Setup. Random sample of $n = 64$ families, $\bar{X} = 480$ soles, $s = 96$ soles. Find the standard error and a 95 percent confidence interval for the population mean, and interpret it correctly.

Theory used

The **standard error** measures how much the *sample mean* bounces around from sample to sample:

$$\mathrm{SE}(\bar{X}) = \frac{s}{\sqrt{n}}.$$

By the Central Limit Theorem, $\bar{X}$ is approximately normal for large n, so a 95 percent confidence interval is

$$\bar{X} \pm 1.96\,\mathrm{SE}(\bar{X}).$$

Note the $\sqrt{n}$: quadrupling the sample halves the standard error.

Thought process

Two distinct quantities are easy to confuse. The **standard deviation** $s = 96$ describes how spread out *families* are. The **standard error** 12 describes how spread out the *average of 64 families* is, eight times smaller, because averaging cancels noise. The confidence interval is built from the second, never the first.

Step-by-step solution

(a) Standard error.

$$\text{SE} = \frac{96}{\sqrt{64}} = \frac{96}{8} = \boxed{12 \text{ soles}}.$$

(b) Confidence interval.

$$480 \pm 1.96(12) = 480 \pm 23.52 \implies \boxed{[456.48,\ 503.52] \text{ soles}}.$$

(Using $t_{63} = 1.998$ instead of 1.96 gives $[456.02, 503.98]$, with $n = 64$ the two are practically identical, which is the CLT doing its job.)

(c) The interpretation. The colleague is **wrong**, and this is the single most common error in the course. The population mean μ is a fixed unknown number; it does not have a probability distribution. It is either inside $[456.48, 503.52]$ or it is not, there is no 95 percent about it. What is random is the **interval**, because it is built from our sample. The correct statement is about the *procedure*: if we repeated this survey many times and built an interval each time this way, about 95 percent of those intervals would contain the true μ. Shorthand: *"we are 95 percent confident"* refers to the reliability of the method, not to a probability attached to μ.

Sanity check: the interval is ± 23.52 wide on a mean of 480, roughly 5 percent either side, a sensible precision for $n = 64$.

4. Problem 4 · Reading a Simple Regression Line

Problem 4 – Easy

Class 9 – Simple linear regression

Setup. $\widehat{\text{wage}} = 6.40 + 1.35\,\text{schooling}$, $R^2 = 0.28$, $n = 500$. Wage in soles per hour, schooling in completed years.

Theory used

In $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$:

- $\hat{\beta}_0$ is the fitted value at $X = 0$;
- $\hat{\beta}_1$ is the change in $\hat{Y}$ for a **one-unit** increase in X, in units of Y per unit of X;
- $R^2 = \text{SSE/SST}$ is the share of the variation in Y that the model reproduces, a measure of *fit*, never of *causality*.

Thought process

Every interpretation question is answered by asking "what are the units?". Wage is soles per hour, schooling is years, so the slope must be *soles per hour, per year of schooling*. Attaching units to a coefficient before saying anything about it prevents most interpretation mistakes.

Step-by-step solution

(a) Intercept. 6.40 is the predicted hourly wage of a worker with **zero** years of schooling. Whether it means anything depends on whether $X = 0$ is inside the data: if nobody in the sample has zero schooling, the intercept is just the anchor that positions the line, extrapolated beyond the data, and should not be given a substantive reading.
(b) Slope. One additional year of schooling is associated with

$$\boxed{\text{1.35 soles per hour more, on average}}.$$

Say "associated with", not "causes", see (e).
(c) R^2. Schooling accounts for 28 percent of the variation in wages across workers; the remaining 72 percent comes from everything else (experience, sector, region, ability, luck). A low R^2 does not make the slope wrong; it says wages are driven by much more than schooling. Note also $\sqrt{0.28} = 0.529$, which in a simple regression is the correlation between schooling and wage.
(d) Prediction at 11 years.

$$\widehat{\text{wage}} = 6.40 + 1.35(11) = 6.40 + 14.85 = \boxed{\text{21.25 soles per hour}}.$$

(e) Causal? No. Schooling is chosen, not assigned. People with more schooling also tend to have more ability, better family connections and better health, all of which raise wages on their own. Those omitted factors are bundled into the 1.35. This is exactly the omitted variable bias of Problems 10 and 11, and the reason Problem 14 reaches for an instrumental variable.

Intermediate problems

Problems 6, 8 and 11 are worked in R on a `.csv` file; the other five are pen and paper.

5. Problem 5 · Tutoring Program: The p-value of an Experiment

Problem 5 – Intermediate

Classes 6–7 – Two-sample hypothesis test

Setup. Randomized tutoring program. Treatment: $n_1 = 50$, $\bar{X}_1 = 78.4$, $s_1 = 12.0$. Control: $n_2 = 50$, $\bar{X}_2 = 73.0$, $s_2 = 11.0$.

Theory used

For the difference of two independent sample means,

$$\mathrm{SE}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \qquad t = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\mathrm{SE}}.$$

Variances add, standard deviations do **not**. The p-value is the probability, *if H_0 were true*, of a statistic at least as extreme as the one observed. Reject H_0 when $p < \alpha$.

Thought process

Two independent groups, means and standard deviations given, no pairing: that is a two-sample t-test. The tempting error is to add the standard deviations $(12 + 11)$ or to average them; the correct move is to add the *squared* standard errors of each mean. And because the question asks whether the program helped *or hurt*, the test is two-sided.

Step-by-step solution

(a) Hypotheses. With μ_1, μ_2 the population mean scores,

$$H_0: \ \mu_1 - \mu_2 = 0 \quad \text{(no effect)}, \qquad H_1: \ \mu_1 - \mu_2 \neq 0 \quad \text{(some effect, either direction)}.$$

(b) Standard error.

$$\mathrm{SE} = \sqrt{\frac{12.0^2}{50} + \frac{11.0^2}{50}} = \sqrt{\frac{144}{50} + \frac{121}{50}} = \sqrt{2.88 + 2.42} = \sqrt{5.30} = \boxed{2.302}.$$

(c) Test statistic. The observed difference is $78.4 - 73.0 = 5.4$, so

$$t = \frac{5.4}{2.302} = \boxed{2.346}.$$

Where does $p = 0.021$ actually come from? The statement handed it to us, but you have to be able to produce it. The p-value is the area in *both* tails of a t distribution beyond ± 2.346:

$$p = 2\,\mathbb{P}(T_{97} > 2.346).$$

Three ways to evaluate it, in order of how you would actually do it.

1. In R. The natural instinct is to reach for `t.test`, and that is right *when you have the data*. Here we do not: the problem gives only n, the means and the standard deviations, and `t.test` wants the two vectors of observations. So there are two versions.

(i) With only the summary, use `pt` on the statistic. One line:

From the summary statistics

```
t  <- 5.4 / sqrt(12^2/50 + 11^2/50)          # 2.3456
df <- (12^2/50 + 11^2/50)^2 /                # Welch degrees of freedom
      ((12^2/50)^2/49 + (11^2/50)^2/49)      # 97.27

2 * pt(-abs(t), df)                          # 0.02103   <- the p-value
qt(0.975, df)                                # 1.9847    <- 5 pct critical value
```

Read `pt` as "area to the LEFT under the t curve". We want both tails, so we take the area left of $-|t|$ and double it. The companion `qt` goes the other way: give it a probability, it returns the cutoff. Same pair for the normal, `pnorm` and `qnorm`. Memorise that four-way grid and you can do any z or t test in R without looking anything up.

*(ii) You **can** still use `t.test`, by rebuilding data with exactly that mean and standard deviation.* `scale()` forces mean 0 and sd 1, so `m + s * scale(...)` lands on the summary you were given:

Making t.test work from a summary

```
mk <- function(n, m, s) as.numeric(m + s * scale(rnorm(n)))
set.seed(1)
a <- mk(50, 78.4, 12);  b <- mk(50, 73.0, 11)
mean(a); sd(a)                        # exactly 78.4 and 12
t.test(a, b)
#   t = 2.3456,  df = 97.267,  p-value = 0.02103
#   95 percent CI: 0.8310  9.9690
```

Identical to (i), as it must be: the t test only ever looks at n, the means and the standard deviations. The trick is worth knowing for exam questions that hand you a summary table, though in practice you would just use `pt`. (The package `BSDA` has `tsum.test` for exactly this, if you would rather install it.)

And when you *do* have the raw data, as in Problems 6, 8 and 11, it really is the one-liner you expected: `t.test(treated, control)` prints t, the degrees of freedom, the p-value and the confidence interval in one go.

2. From a t table, no computer. A table does not give you p directly, it gives cutoffs, so you *bracket* the answer. At df $= 97$:

two-tailed α	0.05	**0.02**	0.01
critical value	1.985	**2.365**	2.627

Our $t = 2.346$ sits between 1.985 and 2.365, so

$$0.02 < p < 0.05,$$

and since 2.346 falls only just short of the 2 percent cutoff, p is barely above 0.02. That is enough to decide at both the 5 and the 1 percent level without ever computing 0.021.

3. Normal approximation. Using $\mathcal{N}(0,1)$ instead of t_{97} gives $2\,\mathbb{P}(Z > 2.346) = 0.019$. Close, but slightly *too small*: the t has fatter tails, so it always returns a larger p than the normal. With df ≈ 100 the gap is only 0.002; with df $= 10$ it would matter a lot.

(d) Decision at 5 percent. $p = 0.021 < 0.05$, so we **reject** H_0: the difference is statistically significant at the 5 percent level. In words, if the program truly did nothing, a gap of 5.4 points or more (in either direction) would show up in only about 2 samples in 100. That is unusual enough to doubt H_0. Equivalently, $|t| = 2.346 > 1.985$.

(e) At 1 percent. $p = 0.021 > 0.01$, so at the 1 percent level we **fail to reject** H_0. Nothing about the data changed, only how much evidence we demanded. This is worth dwelling on: significance is not a property of the effect, it is a property of the (data, threshold) pair. Reporting

the p-value itself, or the confidence interval for the difference $(5.4 \pm 1.985(2.302) = [0.83, 9.97])$, is more informative than a verdict.
Sanity check: the confidence interval excludes zero, exactly matching the 5 percent rejection, the test and the interval must always agree.

6. Problem 6 · Read `rct.csv`, get the p-value by permutation

Problem 6 – Intermediate

Classes 6–8 – `data/rct.csv`*, randomization inference*

Setup. Same file as Problem 2. The estimated ATE is 2.0 points. Now attach a p-value to it, first with `t.test` and then by **permutation** (randomization inference), and explain what each method assumes.

Theory used

The **permutation** (randomization inference) p-value works entirely inside the experiment. Under Fisher's sharp null the treatment changed nobody's outcome, so every outcome is known no matter who was treated. We can therefore reshuffle the treatment label many times, recompute the difference in means each time, and ask how often chance alone produces something as extreme as what we saw:

$$p = \frac{\#\{\,|\text{difference under a reshuffle}| \geq |\text{observed}|\,\}}{\text{number of reshuffles}}.$$

No normal distribution appears anywhere in that argument.

Thought process

The two routes to a p-value differ in what they assume. The t-test leans on the **Central Limit Theorem**: it needs the difference in sample means to be approximately normal, which is what lets it quote a tail probability from a formula. Permutation assumes **nothing** about the distribution of the outcome; its only inputs are the sharp null and the assignment mechanism we actually used. With $n = 300$ they agree. With a handful of observations instead of 300, only the permutation route would be trustworthy.

Step-by-step solution

R code

```
data <- read.csv("../data/rct.csv")
ate  <- mean(data$outcome[data$treatment == 1]) -
        mean(data$outcome[data$treatment == 0])

# Careful with the order: t.test(outcome ~ treatment) compares group 0 MINUS
# group 1 and prints a NEGATIVE t. Passing the two vectors explicitly, treated
# first, keeps every sign pointing the same way as the ATE.
treated <- data$outcome[data$treatment == 1]
control <- data$outcome[data$treatment == 0]
tt <- t.test(treated, control)                # route 1: the formula

set.seed(99)                                  # route 2: permutation
perm <- replicate(10000, {
  w <- sample(data$treatment)                 # reshuffle WHO is treated
  mean(data$outcome[w == 1]) - mean(data$outcome[w == 0])
})
sum(abs(perm) >= abs(ate))                    # how many are as extreme
mean(abs(perm) >= abs(ate))                   # the p-value

hist(perm, breaks = 40); abline(v = c(-ate, ate), col = "red", lwd = 2)
```

(a) The t-test route. Welch two-sample t-test on (`treated`, `control`):

$$t = 2.3180, \qquad \mathrm{df} = 297.08, \qquad p = \boxed{0.02113},$$

with a 95 percent confidence interval for the difference of [0.302, 3.698]. Significant at 5 percent but **not** at 1 percent, and the interval very nearly touches zero: a genuinely borderline result, which is exactly where a p-value earns its keep.

A note on signs. Writing `t.test(outcome ~ treatment)` instead would compare group 0 *minus* group 1 and print $t = -2.3180$ with the interval $[-3.698, -0.302]$: the same test, but every sign flipped relative to the ATE. Passing the two vectors with the treated group first avoids the confusion.

(b) The permutation route. Out of 10,000 reshuffles of the treatment label, **213** produced a gap at least as large as 2.0 in absolute value, so

$$p = \frac{213}{10000} = \boxed{0.0213}.$$

(c) Reading the histogram. The 10,000 reshuffled differences pile up around zero, which is exactly what the sharp null predicts: if the program did nothing, relabelling who was "treated" should produce differences that are just noise. They run from -3.26 to $+3.38$, with 95 percent of them inside $[-1.71, 1.72]$. Our observed 2.0 sits out in the right tail, beyond about 99 percent of the draws on that side, but still on the chart. That visual position *is* the p-value.

(d) Why the two agree. 0.0211 from the formula, 0.0213 from the simulation: a gap of 0.0002 is simulation noise, not disagreement. The reason shows up in the spread. The permutation distribution is centred at zero (mean 0.018) and approximately normal, with standard deviation 0.8713 against the t-test standard error of 0.8628: the same quantity measured two ways. That

closeness is precisely what the CLT promises for $n = 300$. When the CLT holds, the formula-based t-test is a shortcut for the permutation calculation, and the two must land in the same place.

(e) When they would not agree. With a small sample, a heavily skewed outcome, or a few extreme values, the normal approximation breaks down and the t-test p-value becomes unreliable, while the permutation p-value stays exactly valid, because it never needed the approximation. The cost of the permutation route is granularity: with B reshuffles the resolution is $1/B$, and in a tiny experiment the number of *possible* assignments itself puts a floor on the p-value: with 6 units split 3 and 3 there are only $\binom{6}{3} = 20$ possible assignments, so the smallest one-sided p-value the experiment can ever produce is $1/20 = 0.05$, however well the treatment works.

7. Problem 7 · Difference-in-Differences by Hand

Problem 7 – Intermediate

Class 9 – Difference-in-differences

Setup. Mean enrolment (percentage points): treated regions 61.0 before, 72.5 after; control regions 58.0 before, 64.0 after.

Theory used

Difference-in-differences removes two confounds at once:

$$\text{DiD} = (\bar{Y}^{T}_{\text{after}} - \bar{Y}^{T}_{\text{before}}) - (\bar{Y}^{C}_{\text{after}} - \bar{Y}^{C}_{\text{before}}).$$

In regression form, with two dummies and their product,

$$Y = \beta_0 + \beta_1 \,\text{treated} + \beta_2 \,\text{post} + \beta_3(\text{treated} \times \text{post}) + u,$$

the four coefficients reproduce the four cell means exactly, and β_3 *is* the DiD.

Thought process

Ask what each naive comparison forgets. Before-and-after within the treated group forgets that *time* changes things for everyone. Treated-versus-control after the fact forgets that the groups *started* in different places. DiD is built precisely so that each comparison cancels the other's blind spot.

Step-by-step solution

(a) What is wrong with 11.5**?** That is the before-after change in treated regions only. It attributes to the program everything else that happened over the same period, economic growth, a national campaign, a demographic shift. We can see the size of that contamination directly: control regions, with no program at all, gained 6.0 points.

(b) What is wrong with 8.5**?** That is the treated-control gap after the program. It ignores that treated regions were *already* $61.0 - 58.0 = 3.0$ points ahead before anything happened. It

credits the program with a pre-existing difference.

(c) Difference-in-differences.

$$\text{DiD} = \underbrace{(72.5 - 61.0)}_{\text{treated change}=11.5} - \underbrace{(64.0 - 58.0)}_{\text{control change}=6.0} = \boxed{5.5 \text{ percentage points}}.$$

Note the two wrong answers bracket the right one: $5.5 = 11.5 - 6.0 = 8.5 - 3.0$. Subtracting the time trend from (a), or the pre-existing gap from (b), lands in the same place.

(d) The four coefficients.

$$\beta_0 = 58.0, \qquad \beta_1 = 3.0, \qquad \beta_2 = 6.0, \qquad \boxed{\beta_3 = 5.5}.$$

Reading them: β_0 is control, before. β_1 is the pre-existing gap. β_2 is the common time trend. β_3 is the extra movement in the treated group beyond that trend, the estimated program effect. *Cross-check:* the regression must return all four cell means. Treated, after $= \beta_0 + \beta_1 + \beta_2 + \beta_3 = 58.0 + 3.0 + 6.0 + 5.5 = 72.5$. ✓

8. Problem 8 · Read `did.csv`, fit the difference-in-differences

Problem 8 – Intermediate

Class 9 – `data/did.csv`, 800 observations

Setup. `did.csv` has 800 school-district observations over two periods. Columns: `region`, `treated` (0/1), `post` (0/1), `outcome` (share of pupils at grade level, percentage points). Get the DiD from the table of means and from a regression, and confirm they agree.

Theory used

The 2×2 table of means and the interaction coefficient are two views of the same object. The regression with two dummies and their product is **saturated**, four parameters for four cells, so it reproduces the table exactly. The regression is worth running anyway, because it hands us the standard error, which the table of means cannot.

Thought process

Compute the DiD by hand *first*, then run the regression. When the two numbers match to the last decimal you know the interaction is doing what you think it is doing. If they ever disagree, the model is not saturated, usually because a control was added or a term was left out.

Step-by-step solution

R code

```
data <- read.csv("../data/did.csv")
table(data$treated, data$post)                # check the four cells

cells <- with(data, tapply(outcome, list(treated, post), mean))
change_treated <- cells["1","1"] - cells["1","0"]
change_control <- cells["0","1"] - cells["0","0"]
did_hand       <- change_treated - change_control

m <- lm(outcome ~ treated + post + treated:post, data = data)
summary(m)
confint(m)["treated:post", ]
```

(a) The 2×2 table of means.

	before (`post`=0)	after (`post`=1)
control (`treated`=0)	51.7228	59.2005
treated (`treated`=1)	55.8591	69.3421

(b) DiD by hand.

$$\underbrace{(69.3421 - 55.8591)}_{13.4830} - \underbrace{(59.2005 - 51.7228)}_{7.4777} = \boxed{6.0053 \text{ percentage points}}.$$

The other grouping gives the same answer: $(69.3421 - 59.2005) - (55.8591 - 51.7228) = 10.1416 - 4.1363 = 6.0053$.

(c) DiD from the regression.

	Estimate	Std. Error	t	p
(Intercept)	51.7228	0.6284	82.30	$< 10^{-15}$
treated	4.1363	0.8932	4.63	4.3×10^{-6}
post	7.4777	0.8932	8.37	2.5×10^{-16}
treated:post	**6.0053**	1.2632	4.75	2.4×10^{-6}

The interaction is $\boxed{6.0053}$, identical to (b), as it must be.

(d) What `treated` and `post` measure. `treated` $= 4.1363$ is the *pre-existing* gap: treated districts were already ahead before the program. `post` $= 7.4777$ is the *common time trend*, how much control districts improved anyway. Neither is an effect of the program; they are precisely the two confounds that DiD subtracts off.

(e) Confidence interval.

$$\boxed{[3.53,\ 8.48] \text{ percentage points}}.$$

It excludes zero, so the effect is clearly distinguishable from nothing ($t = 4.75$, $p = 2.4 \times 10^{-6}$). But it is wide: anything from a modest 3.5 to a large 8.5 is consistent with these data. Note also that SE $= 1.2632$ is the largest in the table, because a difference of differences accumulates noise from all four cells. DiD estimates are inherently the least precise thing in a regression output.

9. Problem 9 · Dummy Variables with Three Categories

Problem 9 – Intermediate

Class 9 – Dummy variables

Setup. $\widehat{\text{wage}} = 8.20 - 1.75\,\text{Sierra} - 2.40\,\text{Selva} + 1.10\,\text{schooling}$, with three regions: Costa, Sierra, Selva.

Theory used

A categorical variable with k categories enters a regression as $k-1$ **dummies**. The omitted category is the **baseline**: it is folded into the intercept, and every dummy coefficient is a gap *relative to it*. Including all k dummies plus an intercept creates **perfect multicollinearity** (the *dummy variable trap*) because the dummies sum to 1 for every observation, exactly the intercept column.

Thought process

The reflex when reading dummy output is: *which category is missing?* That one is the reference point for everything else. A second reflex: the gap between two *non-baseline* categories is the difference of their coefficients, no new regression needed.

Step-by-step solution

(a) Baseline. Costa. It is the region with no dummy in the equation, so a Costa worker has Sierra = Selva = 0 and the intercept 8.20 already describes them.
(b) The -1.75. A Sierra worker earns on average **1.75 soles per hour less than a Costa worker with the same schooling**. The clause "with the same schooling" is essential: it is a gap after holding schooling fixed, not a raw difference in average wages.
(c) Sierra worker, 12 years.

$$8.20 - 1.75 + 1.10(12) = 8.20 - 1.75 + 13.20 = \boxed{19.65 \text{ soles per hour}}.$$

(For comparison: Costa $= 8.20 + 13.20 = 21.40$; Selva $= 8.20 - 2.40 + 13.20 = 19.00$.)
(d) Sierra − Selva gap. Subtract the two coefficients:

$$(-1.75) - (-2.40) = \boxed{0.65 \text{ soles per hour}},$$

Sierra above Selva. Confirmed by (c): $19.65 - 19.00 = 0.65$. ✓
(e) Why not all three dummies. Every worker belongs to exactly one region, so

$$\text{Costa}_i + \text{Sierra}_i + \text{Selva}_i = 1 \quad \text{for all } i,$$

which is precisely the constant column. One regressor would be an exact linear combination of the others, $X^\top X$ would be singular and $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ would not exist, infinitely many coefficient vectors would fit identically well. The fix is to drop one category (software does it silently) or to drop the intercept. With k categories you always get $k-1$ dummies.

10. Problem 10 · The Omitted Variable Bias Formula

Problem 10 – Intermediate

Class 10 – Omitted variable bias

Setup. Long model wage $= \beta_0 + \beta_1$schooling $+ \beta_2$ability $+ u$ with $\beta_1 = 1.10$, $\beta_2 = 0.85$. Ability unobserved; auxiliary slope of ability on schooling $\delta = 0.40$.

Theory used

Omitting a relevant regressor X_2 from a regression of Y on X_1 leaves the short-regression slope

$$\mathbb{E}[\tilde{\beta}_1] = \beta_1 + \beta_2\,\delta,$$

where β_2 is the effect of the omitted variable on Y and δ is the slope of the **auxiliary regression** of the omitted variable on the included one. The **bias** is the product $\beta_2\delta$, so its *sign* is the product of two signs:

	$\delta > 0$	$\delta < 0$
$\beta_2 > 0$	upward	downward
$\beta_2 < 0$	downward	upward

Thought process

The formula requires **two** things to be true at once for bias to appear: the omitted variable must matter for Y ($\beta_2 \neq 0$) *and* it must be correlated with the included regressor ($\delta \neq 0$). If either is zero, the product is zero and omitting the variable costs us nothing in bias. That is why randomization is such a powerful device: it forces $\delta = 0$ for every possible confounder at once.

Step-by-step solution

(a) The formula.

$$\mathbb{E}[\tilde{\beta}_1] = \beta_1 + \beta_2\,\delta.$$

(b) Numerical value.

$$\mathbb{E}[\tilde{\beta}_1] = 1.10 + (0.85)(0.40) = 1.10 + 0.34 = \boxed{1.44}.$$

(c) Size and direction.

$$\text{bias} = \beta_2\delta = \boxed{+0.34, \text{ an upward bias}}.$$

The short regression overstates the return to schooling by 0.34 soles per hour per year, about 31 percent too large relative to the true 1.10.

(d) Why, in words. The short regression can only see schooling, so whenever schooling goes up in the data, ability tends to go up with it ($\delta = 0.40 > 0$), and ability independently pushes wages up ($\beta_2 = 0.85 > 0$). The regression has no way to tell the two apart, so it hands *all* of the resulting wage increase to schooling. Schooling gets paid twice: once for itself, and once for the ability travelling alongside it.

Practical note: in reality ability is unobserved, so we can never compute 0.34. But we can usually reason about the two *signs*, and that alone tells us the direction, which is often enough to know whether an estimate is an upper or a lower bound. Problem 11 does exactly that on real data.

11. Problem 11 · Read `wages.csv`, dummies and omitted variable bias

Problem 11 – Intermediate

Classes 9–10 – `data/wages.csv`, 400 workers

Setup. `wages.csv` has 400 workers with `wage`, `schooling`, `experience`, `region` (Costa / Sierra / Selva) and `ability`. Ability *is* observed here, which is exactly what lets us see a bias that is normally invisible.

Theory used

Two tools in one file. `factor(region)` generates $k-1$ dummies and drops the alphabetically first level as the baseline. And the OVB identity

$$\hat{\beta}_{\text{short}} = \hat{\beta}_{\text{long}} + \hat{\beta}_{\text{ability}} \cdot \hat{\delta}$$

holds **exactly** in any finite sample, where $\hat{\delta}$ comes from the auxiliary regression of the omitted variable on the included ones.

Thought process

The reason this one is worth doing on real data is that the OVB identity is easy to nod along to and hard to believe until you watch the two numbers agree to four decimals. It is not an approximation and not an empirical regularity, it is algebra.

Step-by-step solution

R code

```
data <- read.csv("../data/wages.csv")
table(data$region)

m_dum   <- lm(wage ~ schooling + experience + factor(region), data = data)
m_short <- lm(wage ~ schooling + experience, data = data)
m_long  <- lm(wage ~ schooling + experience + ability, data = data)
m_aux   <- lm(ability ~ schooling + experience, data = data)

anova(m_short, m_dum)                                    # do regions matter?
delta <- coef(m_aux)["schooling"]
b2    <- coef(m_long)["ability"]
coef(m_short)["schooling"] - coef(m_long)["schooling"]   # actual bias
b2 * delta                                               # the formula
```

(a) Dummy regression.

	Estimate	Std. Error	t
(Intercept)	3.7027	0.7473	4.95
schooling	1.4095	0.0587	24.01
experience	0.2651	0.0271	9.78
factor(region)Selva	−2.7791	0.3475	−8.00
factor(region)Sierra	−1.6172	0.3069	−5.27

Costa is the baseline, the category R does not print, dropped because it comes first alphabetically. Each region coefficient is a gap relative to Costa, holding schooling and experience fixed: Sierra pays 1.62 soles per hour less, Selva 2.78 less.

The regions matter *jointly*: `anova` comparing the model with and without them gives $F(2, 395) = 34.65$, $p = 1.4 \times 10^{-14}$. That is the right test for "does region matter at all", not three separate t-tests.

(b) Costa − Selva gap. $\boxed{2.78 \text{ soles per hour}}$, read straight off the Selva coefficient with the sign flipped. Checking with predicted wages at 12 years of schooling and 10 of experience: Costa 23.267, Sierra 21.650, Selva 20.488, and $23.267 - 20.488 = 2.779$. ✓(The Sierra to Selva gap needs a subtraction: $-1.6172 - (-2.7791) = 1.1619$.)

(c) Short versus long.

$$\hat{\beta}_{\text{short}} = 1.4480, \qquad \hat{\beta}_{\text{long}} = 1.1407, \qquad \text{bias} = \boxed{0.3073}.$$

The schooling coefficient **falls** by about 0.31 soles per year once ability is included.

(d) The auxiliary regression and the identity. $\hat{\delta} = 0.2611$ (slope of ability on schooling) and $\hat{\beta}_{\text{ability}} = 1.1770$, so

$$\hat{\beta}_{\text{ability}} \cdot \hat{\delta} = 1.1770 \times 0.2611 = \boxed{0.3073},$$

matching the actual bias to four decimals. They are not approximately equal, they are **algebraically identical**. Running the short regression is the same computation as running the long one and folding the omitted variable back in through the auxiliary regression.

(e) Over- or under-stating? Over-stating. Signs: $\hat{\beta}_{\text{ability}} = +1.1770$ (more ability, higher wage) and $\hat{\delta} = +0.2611$ (more schooling goes with more ability, indeed $\text{cor}(\text{schooling}, \text{ability}) = 0.615$). Positive times positive is an **upward** bias: with ability omitted, schooling is paid twice, once for itself and once for the ability travelling with it.
The practical point: in real data `ability` is unobserved, so the long regression and $\hat{\delta}$ are unavailable. But the two *signs* can still be argued from theory, and that alone tells us 1.4480 is an **upper bound** on the causal return to schooling. Signing the bias is the everyday payoff of the OVB formula.

12. Problem 12 · Reading a Full Regression Table

Problem 12 – Intermediate

Classes 9–10 – Multiple regression, inference

Setup. $n = 240$; wage in soles per hour. Estimates (standard errors): intercept 4.812 (1.205), schooling 1.104 (0.212), experience 0.318 (0.061), female −2.470 (0.930). $R^2 = 0.341$, $F(3, 236) = 40.72$, $t_{0.975,236} \approx 1.97$.

Theory used

For each coefficient, $t = \hat{\beta}_j / \text{SE}(\hat{\beta}_j)$ tests $H_0 : \beta_j = 0$ *holding the other regressors fixed*, and

$$\hat{\beta}_j \pm t_{0.975,\, n-k-1}\, \text{SE}(\hat{\beta}_j)$$

is its 95 percent confidence interval. The ***F*-test** instead tests the *joint* null $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, that the regressors are useless *all together*.

Thought process

"Divide the estimate by its standard error" is the single most-used operation in applied work: it converts a number in soles into a number of standard errors, which is comparable across variables and scales. Anything past roughly $|t| > 2$ is significant at 5 percent for a decent sample size, a rule of thumb worth internalizing.

Step-by-step solution

(a) t for *female*.

$$t = \frac{-2.470}{0.930} = \boxed{-2.656}.$$

(b) Significant? $|-2.656| = 2.656 > 1.97$, so **yes**, reject H_0 at the 5 percent level (the exact two-sided p-value is 0.0084). It is also significant at 1 percent, since $0.0084 < 0.01$.
(c) Confidence interval for *schooling*.

$$1.104 \pm 1.97(0.212) = 1.104 \pm 0.418 \implies \boxed{[0.686,\ 1.522]}.$$

The interval excludes zero, matching $t = 1.104/0.212 = 5.21$. Its width tells us the data are consistent with a return anywhere between roughly 0.7 and 1.5 soles per year, statistically clear, but not precisely pinned down.

(d) Interpreting *female*. Women earn on average **2.47 soles per hour less than men with the same schooling and the same experience**. Note what this is *not*: it is not the raw gender wage gap (that would come from a regression with no controls), and it is not evidence of discrimination on its own, because occupation, hours and sector are all still omitted.
(e) t-tests versus the F-test. They answer different questions:

- each t-test asks whether *one* regressor adds explanatory power **given that the others are already in the model**;
- the F-test asks whether the regressors are jointly useless.

Here $F(3, 236) = 40.72$ is enormous ($p \approx 3 \times 10^{-21}$), so the model as a whole clearly explains wages. The two can disagree, and instructively so: with highly correlated regressors every individual t can be insignificant while F is large, because the data cannot say *which* of them matters, only that they matter together.
Sanity check: F can be recovered from R^2: $\dfrac{R^2/k}{(1-R^2)/(n-k-1)} = \dfrac{0.341/3}{0.659/236} = 40.7$. ✓

Hard problems (challenge)

Problems 13 to 15, pen and paper. They are the most demanding of the session, but each one is still an application of a formula we have already used, not a derivation.

13. Problem 13 · Logs and an Interaction

Problem 13 – Hard

Class 10 – Functional form, interactions

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

Theory used

In a **log-linear** model, a coefficient b on X means a one-unit increase in X raises Y by approximately $100b$ *percent*, a *semi-elasticity*. The approximation comes from $\log(1+r) \approx r$ for small r; the exact percentage change is

$$100(e^{\,b} - 1),$$

and the gap between the two grows fast once $|b|$ passes about 0.2.
An **interaction** $X \times D$ lets the slope on X differ by group: the effect of X is b_X when $D = 0$ and $b_X + b_{XD}$ when $D = 1$.

Thought process

With an interaction in the model, no coefficient can be read on its own any more. 0.062 is not "the" return to schooling, it is the return *for rural workers*; 0.145 is not "the" urban premium, it is the premium *at schooling* $= 0$. The habit to build: after reading a coefficient, ask "for

whom, and at what value of the other variable?"

Step-by-step solution

(a) Why logs. Wages are strongly right-skewed and effects are naturally proportional, "10 percent more" is a more meaningful statement than "2 soles more" across workers earning 5 and 50 soles. Taking logs turns the coefficient into a percentage change, comparable across the whole distribution.

(b) Return to a year of schooling.

$$\text{rural (urban} = 0) : 0.062 \;\Rightarrow\; \boxed{\approx 6.2 \text{ percent}},$$

$$\text{urban} : 0.062 + 0.021 = 0.083 \;\Rightarrow\; \boxed{\approx 8.3 \text{ percent}}.$$

Schooling pays about 2 percentage points more per year in cities.

(c) The 0.145**.** It is the urban premium **for a worker with zero years of schooling**, the only point where the interaction term vanishes. It is *not* the average urban premium, and quoting it as such is the classic mistake with interacted models.

(d) Urban premium at 10 years of schooling.

$$0.145 + 0.021(10) = 0.145 + 0.21 = 0.355 \;\Rightarrow\; \boxed{\approx 35.5 \text{ percent (approximation)}}.$$

The premium grows with schooling: cities reward education more.

(e) Where the shortcut hurts. Compare $100b$ with the exact $100(e^b - 1)$:

	b	$100b$	$100(e^b - 1)$
rural return	0.062	6.2	6.40
urban return	0.083	8.3	8.65
urban premium, 10 yrs	0.355	35.5	**42.62**

For the two small coefficients the error is a fraction of a percentage point, negligible. For $b = 0.355$ the shortcut understates the effect by 7 percentage points, roughly a fifth of the answer. **Rule:** the approximation is fine below about 0.1; past 0.2, use the exact formula.

14. Problem 14 · Vouchers by Lottery: the Wald Estimator

Problem 14 – Hard

Class 11 – Instrumental variables, Wald estimator

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

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

An OLS regression of Y on D gives 9.50. The formula is given on the slide.

Theory used

With a binary instrument Z, the **Wald estimator** is a ratio of two differences we can read straight off the table:

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

First stage = how much the instrument moves the treatment. **Reduced form** = how much the instrument moves the outcome.

Thought process

Read the ratio as *undoing a dilution*. The lottery did not put everybody into a private school: it moved attendance by 45 percentage points, not 100. So its effect on scores (2.70) is a watered-down version of the effect of attendance itself. Dividing by the first stage scales it back up.

Everything here is arithmetic on four numbers. The only thing to be careful about is which difference goes on top: the **outcome** difference is the numerator.

Step-by-step solution

(a) First stage.

$$\mathbb{E}[D \mid Z=1] - \mathbb{E}[D \mid Z=0] = 0.62 - 0.17 = \boxed{0.45}.$$

Winning the lottery raised the probability of attending a private school by 45 percentage points.

(b) Reduced form.

$$\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0] = 46.80 - 44.10 = \boxed{2.70}.$$

Winning the lottery raised test scores by 2.70 points.

(c) Wald estimate.

$$\hat{\beta}_{\text{IV}} = \frac{2.70}{0.45} = \boxed{6.00 \text{ test points}}.$$

Attending a private school raises the test score by about 6 points.

(d) Why is OLS (9.50) bigger? Because attendance is *chosen*, not assigned. Families who send children to private schools tend to have higher income and more educated parents, and those things raise test scores on their own. OLS credits private schooling with the whole gap, family background included. That is **selection bias**, the same idea as the omitted variable bias of Problem 10, and here it inflates the estimate from 6.00 to 9.50.

The lottery fixes it: it was random, so winners and losers are comparable in every respect. Using only the variation in attendance that came from the lottery strips the family background out.

Sanity check: multiply back. first stage $\times\ \hat{\beta}_{\text{IV}} = 0.45 \times 6.00 = 2.70 =$ reduced form. ✓

One line of vocabulary for later: because the lottery only changed the behaviour of the families who enrolled *because* they won, the 6.00 is an effect for that group, called the **LATE** (Local Average Treatment Effect). The 17 percent who would have attended anyway and the 38 percent who never attend contribute nothing to it.

15. Problem 15 · Capstone: Reading a Policy Evaluation

Problem 15 – Hard (capstone)

Classes 6–11 – Inference, interactions, causality

Setup. 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.

Theory used

Three tools, all already used earlier today:

- $t = \hat{\beta}/\text{SE}$, and $\hat{\beta} \pm 1.96\,\text{SE}$ for the interval (Problems 3 and 12);
- with $\log(Y)$ on the left, a coefficient b is roughly a $100b$ percent change (Problem 13);
- with an interaction, the effect for the group with the dummy $= 1$ is **main effect + interaction** (Problem 13).

Thought process

Nothing here is new; the work is knowing *which* of the five rows to use for each question. Part (c) is the one people trip on: for a female-headed household the program effect is not 0.118 and not 0.087, it is the two added together. And part (e) is the habit worth keeping: ask, row by row, *where did this variation come from?*

Step-by-step solution

(a) Did the program work?

$$t = \frac{0.118}{0.036} = \boxed{3.278}.$$

Since $3.278 > 1.96$, **yes**: reject $H_0 : \beta_{\text{treated}} = 0$ at the 5 percent level ($p \approx 0.001$). The transfer raised income for male-headed households by about 0.118 log points, that is roughly 12 percent.

(b) Confidence interval for *treated*.

$$0.118 \pm 1.96(0.036) = 0.118 \pm 0.071 \implies \boxed{[0.047,\ 0.189]}.$$

Roughly 5 to 19 percent. Positive throughout, so the effect is clearly there, but the range is wide: significance settles the *sign*, not the *size*.

(c) Effect for a female-headed household. Add the main effect and the interaction:

$$0.118 + 0.087 = \boxed{0.205 \text{ log points}}, \text{ that is about } \boxed{21 \text{ to } 23 \text{ percent}}.$$

(The quick reading is $100 \times 0.205 = 20.5$ percent; the exact value is $100(e^{0.205} - 1) = 22.8$ percent. At this size the two start to separate, as Problem 13 warned.)

(d) Is the interaction significant?

$$t = \frac{0.087}{0.058} = 1.500 < 1.96 \implies \boxed{\text{no}}.$$

The point estimate suggests the program helps female-headed households more, but the data cannot distinguish 0.087 from zero. Say it carefully: *we did not find evidence of a difference*, which is not the same as *we found evidence of no difference*. Interactions are always estimated less precisely than main effects, so a study powered to detect an average effect is usually underpowered for subgroups.

(e) Which coefficients are causal? Only **treated** (and its interaction). The transfer was handed out *at random*, so it is uncorrelated with everything we did not measure and the omitted variable bias term $\beta_2\delta$ is zero.

Schooling is not causal. Nobody randomized education. The 0.049 still carries ability, family background and health, all correlated with schooling and with income, so it is biased **upward**, exactly the structure of Problem 10. The same applies to *female_head*.

That is the message of the whole session: randomization buys a causal reading for the *one* variable that was randomized. Every other row of the table is observational and has to be read with the OVB formula in hand.

Bonus problems (if time allows)

Held in reserve in case we finish early.

16. Bonus 1 · Setting Up a Test About Commute Times

Bonus 1 – Easy

Class 6 – Hypotheses, tails, Type I error

Setup. Claimed mean commute 45 minutes. Sample: $n = 100$, $\bar{X} = 47.8$, $s = 14$.

Theory used

H_0 always states the *status quo* as an equality; H_1 is what we would need evidence to believe. With n large,

$$z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \dot{\sim} \mathcal{N}(0,1) \quad \text{under } H_0.$$

A **Type I error** is rejecting a true H_0; its probability is the significance level α we choose.

Thought process

Choose the tails *before* looking at the data. The claim is "the average is 45", and a commute that is surprisingly *short* would contradict it just as much as one that is surprisingly long, so

the test is two-sided. Switching to one-sided after seeing $\bar{X} > 45$ halves the p-value dishonestly.

Step-by-step solution

(a) Hypotheses. $H_0 : \mu = 45$ versus $H_1 : \mu \neq 45$.
(b) Standard error and statistic.

$$\text{SE} = \frac{14}{\sqrt{100}} = 1.4, \qquad z = \frac{47.8 - 45}{1.4} = \frac{2.8}{1.4} = \boxed{2.00}.$$

(c) Decision at 5 percent. $p = 0.0455 < 0.05$, so **reject** H_0: the evidence contradicts the ministry's claim at the 5 percent level. It is a close call, $|z| = 2.00$ against a critical value of 1.96. Note that using t_{99} instead of the normal gives $p = 0.0482$: still a rejection, but narrower still.
(d) One-sided. For $H_1 : \mu > 45$ the p-value is half the two-sided one, $\boxed{0.0228}$, since the normal is symmetric and all the rejection region sits in one tail.
(e) Type I error, in context. Concluding that the true average commute differs from 45 minutes *when it really is 45*. Choosing $\alpha = 0.05$ means accepting a 5 percent chance of making exactly that mistake, of publicly contradicting a correct ministry claim on the strength of an unlucky sample.

17. Bonus 2 · Why Trainees Earn Less

Bonus 2 – Intermediate

Class 8 – Selection bias decomposition

Setup. Participants in a voluntary job-training program earn 400 soles *less* than non-participants. An experimental study finds the ATT is +600 soles.

Theory used

The naive comparison always splits into two pieces:

$$\underbrace{\mathbb{E}[Y \mid D=1] - \mathbb{E}[Y \mid D=0]}_{\text{observed difference}} = \underbrace{\mathbb{E}[Y(1) - Y(0) \mid D=1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y(0) \mid D=1] - \mathbb{E}[Y(0) \mid D=0]}_{\text{selection bias}}.$$

The selection bias compares the two groups on the outcome they would have had **without** treatment. Randomization sets it to zero.

Thought process

When a comparison has the "wrong" sign, the decomposition is the diagnostic tool. It forces the question: what would these two groups have looked like *anyway*? Here the answer is obvious once asked, people sign up for job training because they are struggling in the labour market.

Step-by-step solution

(a) Identifying the terms. Observed difference $= -400$; ATT $= +600$; selection bias is the unknown.
(b) Solving.

$$-400 = 600 + \text{selection bias} \implies \text{selection bias} = \boxed{-1000 \text{ soles}}.$$

(c) What it says. Even **without** any training, participants would have earned 1,000 soles less than non-participants. They are a negatively-selected group: people enrol precisely because they are unemployed, have obsolete skills, or have weak networks, exactly the characteristics that depress earnings.
That single number resolves the paradox. The program genuinely *helps* its participants, by 600 soles. But it starts them from 1,000 soles behind, so after training they are still 400 behind. The naive comparison sees only the final gap and concludes the program is harmful, a conclusion that is off by 1,000 soles and gets the sign backwards.
(d) If it had been randomized. Selection bias would be zero, so the observed difference would have equalled the treatment effect: +600 soles (and with randomization the ATT and ATE coincide). This is why voluntary-enrolment programs are so hard to evaluate, and why the same decomposition drives the IV argument in Problem 14, where OLS overstated the voucher effect because selection ran the *other* way.

18. Bonus 3 · A Quadratic Experience Profile

Bonus 3 – Intermediate

Class 10 – Functional form, turning point

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

Theory used

With $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ the effect of X is no longer constant:

$$\frac{\partial \hat{Y}}{\partial X} = \hat{\beta}_1 + 2\hat{\beta}_2 X.$$

Setting it to zero gives the **turning point** $X^* = -\hat{\beta}_1/(2\hat{\beta}_2)$: a maximum when $\hat{\beta}_2 < 0$, a minimum when $\hat{\beta}_2 > 0$.

Thought process

Once X and X^2 are both in the model they can never be moved independently, you cannot change experience without changing its square. So neither coefficient means anything alone; only the derivative, evaluated at a specific value, has an interpretation.

Step-by-step solution

(a) Why the squared term. It allows **diminishing returns**. The first years on the job teach a lot; the twentieth adds less than the second; and late in a career wages may plateau or fall. A straight line cannot express any of that, so it would force a constant return at every level of experience.

(b) Marginal effect.

$$\frac{\partial \widehat{\text{wage}}}{\partial\, \text{exp}} = 0.680 - 2(0.017)\,\text{exp} = \boxed{0.680 - 0.034\,\text{exp}}.$$

(c) Evaluated at two points.

$$\text{exp} = 5 : 0.680 - 0.034(5) = \boxed{+0.51 \text{ soles per year}},$$

$$\text{exp} = 25 : 0.680 - 0.034(25) = \boxed{-0.17 \text{ soles per year}}.$$

Early in a career an extra year adds about half a sol per hour; at 25 years it *subtracts* about 0.17.

(d) Turning point.

$$\text{exp}^* = \frac{0.680}{2(0.017)} = \frac{0.680}{0.034} = \boxed{20 \text{ years}}.$$

Wages peak at 20 years of experience (the profile adds 6.80 soles at the peak) and decline slowly after. Worth a reality check: is 20 years plausible, and are there enough workers past that point in the sample to identify the downturn? If almost nobody has 30 years of experience, the falling arm of the parabola is extrapolation, not evidence.

(e) Can 0.680 be read alone? No. It is the marginal effect at exp = 0 only, the return in a worker's very first year. Quoting it as "the return to experience" would overstate the effect for everyone else in the sample and would be flatly wrong past 20 years, where the true effect is negative. With a quadratic, always report the derivative at a value people care about, or report the turning point.

19. Bonus 4 · Pooled versus Within

Bonus 4 – Hard

Class 10 – Fixed effects, Simpson's paradox

Setup. Three schools, two classes each. Class sizes and scores: A $(18, 61), (22, 58)$; B $(26, 70), (30, 67)$; C $(34, 79), (38, 76)$. Pooled slope $+0.96$.

Theory used

Fixed effects add a dummy for each group, which removes all *between-group* variation and leaves the regression to be identified by *within-group* variation only. The FE slope is the average of the within-group slopes. When between and within variation point in opposite directions, the pooled and FE estimates have opposite signs, **Simpson's paradox**.

Thought process

When a pooled result defies everything you know about a subject, look for a group structure. The question to ask is always: *is this comparison being made across groups or inside them?* Here, comparing across schools compares rich schools with poor ones; comparing inside a school holds the school constant.

Step-by-step solution

(a) Within school A. Two points, so the slope is just rise over run:

$$\frac{58-61}{22-18} = \frac{-3}{4} = \boxed{-0.75}.$$

(b) Schools B and C.

$$\text{B: } \frac{67-70}{30-26} = \frac{-3}{4} = -0.75, \qquad \text{C: } \frac{76-79}{38-34} = \frac{-3}{4} = -0.75.$$

All three are **identical** at -0.75: inside every school, the larger class scores lower.

(c) Resolving the contradiction. The two estimates answer different questions.

The pooled slope compares *across* schools, and the school averages line up as $(20, 59.5)$, $(28, 68.5)$, $(36, 77.5)$, strongly increasing. Schools with bigger classes also have higher scores, because class size here is a marker of something else: well-resourced, popular, high-demand schools run bigger classes *and* enrol better-prepared pupils. That between-school pattern is powerful enough to swamp the within-school effect and flip the pooled sign to $+0.96$.

The within-school slope compares two classes *in the same school*, holding constant everything that makes schools differ, neighbourhood, funding, teaching quality, pupil intake. That comparison is negative and identical everywhere: -0.75.

In OVB language, school quality is the omitted variable: it raises scores ($\beta_2 > 0$) and is positively correlated with class size ($\delta > 0$), so the pooled estimate is biased upward, so far upward that it changes sign.

(d) What fixed effects do. Adding a dummy per school lets each school have its own intercept, so only *within*-school variation identifies the slope. All the between-school differences, including everything unobserved about a school, as long as it does not change across its own classes, are absorbed. The FE regression recovers

$$\boxed{-0.75},$$

the common within-school slope, and it is the credible estimate of the class-size effect. This is the great strength of fixed effects: they control for unobserved group characteristics without our having to measure them. The limitation is the flip side, they can do nothing about confounders that vary *within* a school.

20. Bonus 5 · OLS by Hand

Bonus 5 – Hard

Class 9 – OLS algebra, R^2

Setup. Five districts. X = health posts per 10,000 inhabitants $= 1, 2, 3, 4, 5$; Y = prenatal-checkup coverage index $= 3, 4, 8, 9, 11$.

Theory used

The OLS estimates minimising $\sum(y_i - \beta_0 - \beta_1 x_i)^2$ are

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

With $\mathrm{SST} = \sum(y_i - \bar{y})^2$ and $\mathrm{SSR} = \sum \hat{u}_i^2$,

$$R^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}}.$$

Whenever the model has an intercept, the residuals sum to zero by construction.

Thought process

Doing this once by hand pays off for the rest of the course: the slope is a *ratio of a covariance to a variance*, which is why regression on a dummy gives a difference in means (Problem 1) and why the OVB formula has the shape it does. Build the deviation table first; everything else drops out of it.

Step-by-step solution

(a) Means and sums of products. $\bar{x} = 3$, $\bar{y} = 7$.

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	3	-2	-4	8	4
2	4	-1	-3	3	1
3	8	0	1	0	0
4	9	1	2	2	1
5	11	2	4	8	4
				$S_{XY} = 21$	$S_{XX} = 10$

(b) Coefficients.

$$\hat{\beta}_1 = \frac{21}{10} = \boxed{2.1}, \qquad \hat{\beta}_0 = 7 - 2.1(3) = 7 - 6.3 = \boxed{0.7}.$$

So $\hat{Y} = 0.7 + 2.1X$: one more health post per 10,000 inhabitants is associated with 2.1 more points of coverage.

(c) As a ratio of covariance to variance. With $n - 1 = 4$ in both denominators,

$$\mathrm{Cov}(X, Y) = \frac{21}{4} = 5.25, \qquad \mathrm{Var}(X) = \frac{10}{4} = 2.5, \qquad \frac{5.25}{2.5} = 2.1. \ \checkmark$$

The $n - 1$ cancels, which is why S_{XY}/S_{XX} and Cov / Var give the same number.

(d) Residuals. Fitted values $\hat{y}_i = 0.7 + 2.1 x_i$:

$\hat{y}_i$	2.8	4.9	7.0	9.1	11.2
$\hat{u}_i$	0.2	-0.9	1.0	-0.1	-0.2

$$\sum \hat{u}_i = 0.2 - 0.9 + 1.0 - 0.1 - 0.2 = \boxed{0}. \checkmark$$

This is not a coincidence: it is the first-order condition for the intercept, so it holds in *every* regression with a constant term.

(e) R^2.

$$\text{SST} = 16 + 9 + 1 + 4 + 16 = 46, \qquad \text{SSR} = 0.04 + 0.81 + 1.00 + 0.01 + 0.04 = 1.90,$$

$$R^2 = 1 - \frac{1.90}{46} = \boxed{0.9587}.$$

About 96 percent of the variation in coverage is reproduced by the fitted line.

Two cross-checks. The explained sum of squares must be $\hat{\beta}_1^2 S_{XX} = (2.1)^2(10) = 44.1$, and indeed $44.1 + 1.90 = 46 = \text{SST}$. ✓And in a simple regression R^2 is the squared correlation: $\text{corr}(X, Y) = 0.9791$ and $0.9791^2 = 0.9587$. ✓

One caution: a high R^2 says the line fits, not that health posts *cause* coverage. With five districts and no controls, this is a description.

Fuentes y atribución

Material de repaso **basado en MIT 14.310x – Data Analysis for Social Scientists** (Esther Duflo & Sara Fisher Ellison), MIT OpenCourseWare, licencia **CC BY-NC-SA 4.0** (uso no comercial, con atribución, compartir bajo la misma licencia), y en las fuentes de la bibliografía. Los enunciados se basan en el material del curso y en diversas fuentes abiertas; cuando un problema se adapta de una fuente específica, se cita en el enunciado. This review session covers the inference, causality and regression core of Classes 6–11 of this course (estimators and confidence intervals, hypothesis testing and p-values, causality and randomized experiments, single and multivariate linear models, practical regression issues and omitted variable bias, and instrumental variables). It is the companion to Clase 12, which reviewed the probability core of Classes 1–5. All 20 problems and the 3 R practices were authored fresh for this session; none are copied from the course's per-class problem banks or Kahoot question sets. The datasets in `clase13/data/` are simulated with fixed seeds by `clase13/figs/gen_data.R`, and every number quoted in this document was verified by running that script and the solution scripts in `clase13/practica_R/`.

Bibliografía.

- Duflo, E. & Ellison, S. F. *14.310x: Data Analysis for Social Scientists*. MIT OpenCourseWare (slides, lecture notes y problem sets). ocw.mit.edu.
- Larsen, R. J. & Marx, M. L. *An Introduction to Mathematical Statistics and Its Applications*, 6.ª ed., Pearson, 2018.
- DeGroot, M. H. & Schervish, M. J. *Probability and Statistics*, 4.ª ed., Pearson, 2012.
- Blitzstein, J. & Hwang, J. *Introduction to Probability* (Harvard Stat 110); W. Chen, *Probability Cheatsheet*.
- Diez, D., Çetinkaya-Rundel, M. & Barr, C. *OpenIntro Statistics*. openintro.org.
- Materiales abiertos de la comunidad del curso en GitHub: `gevorg-hunanyan/probability`; `hanmochen/Probability-Theory-2020`; `mynameisjanus/14310xDataAnalysis`.

creativecommons.org/licenses/by-nc-sa/4.0