ECMA 31360, PSet 3: Causal Inference with Observational Data

Melissa Tartari, University of Chicago

References: You can access the .TEX version of this file at `Canvas/Files/PSets/PSet3/PSet3-clean.tex` and the .Rmd version of the answer template file at `Canvas/Files/PSets/PSet3/PSet3_Solutions_Template.Rmd`.

Part 1: Intent-to-Treat Versus Average Treatment Effect (10 p)

Objective: Understand the difference between the Intent-to-Treat Effect (ITT) and the Average Treatment Effect (ATE).

Background: So far, you have worked with the experimental data collected in conjunction with the NSW application. In the NSW RCT, treated individuals were offered the opportunity to participate in the NSW employment *cum* training intervention. They could choose to accept or decline the offer. Control individuals were not offered the opportunity, and therefore, by construction, none participated. When an intervention consists of an offer—of a product, program, or policy—the literature refers to the average treatment effect of the offer as the Intent-to-Treat Effect (ITT). To be clear, the ITT is a proper average treatment effect (i.e., the mean of individual treatment effects), but the terminology distinguishes it from the average treatment effect of *undertaking* the program. Thus, in the NSW context, the average treatment effect of being offered the program is often called the ITT of the NSW intervention, while the average treatment effect of actually undertaking the program is the ATE of the NSW intervention. Intuitively, the two average treatment effects differ only if some treated individuals opt not to take up the offer. Later in the course, we will introduce an expanded RCM framework that explicitly separates treatment *assignment* (offer vs. no offer) from treatment *status* (participation vs. non-participation). Using that framework, we will study identification and estimation of the average treatment effect of undertaking treatment in: (a) experiments where the experimenter offers treatment but cannot compel take-up,[1] and (b) non-experimental settings where policymakers offer treatment but cannot compel take-up. For now, we limit ourselves to a pencil-and-paper exercise that does not require the expanded RCM model but still fleshes out the distinction between ITT and ATE.

Q1. (10 p) Consider the following setup.

- Z_i denotes a binary variable equal to 1 if individual i is offered employment *cum* training, and 0 otherwise.
- D_i denotes a binary variable equal to 1 if individual i undertakes employment *cum* training, and 0 otherwise.

For instance, an individual who is offered employment *cum* training but does not take it up has $(Z_i, D_i) = (1, 0)$, while an individual who both receives the offer and enrolls has $(Z_i, D_i) = (1, 1)$.

Assume:

- Individuals not offered employment *cum* training cannot take it up.
- Individuals offered employment *cum* training flip a fair coin: if it lands Heads, they enroll in the NSW program.
- The offer itself does not *per se* affect future earnings.[2]
- The offer of employment *cum* training may or may not be randomly assigned (as in the factual NSW intervention).

Each individual has two pairs of potential outcomes:

- One pair corresponds to the **o**ffer dimension: $(Y_i^o(1), Y_i^o(0))$ denote potential 1978 earnings with and without the offer of employment *cum* training.
- The other pair corresponds to the actual treatment: $(Y_i(1), Y_i(0))$ denote potential 1978 earnings with and without participation in employment *cum* training.

Define two causal estimands:[3]

$$ATE^o := \mathbb{E}[Y_i^o(1) - Y_i^o(0)], \qquad ATE := \mathbb{E}[Y_i(1) - Y_i(0)].$$

[1] We do so using a different RCT because the NSW data do not record take-up decisions.
[2] This assumption rules out, for example, the case in which merely receiving the offer boosts optimism and job search effort.
[3] In **PSet 2** you estimated ATE^o using the NSW experimental data.

Answer the following questions:

(a) (2 p) Describe ATE^o and explain why it is commonly referred to as the ITT of the NSW intervention.

(b) (1 p) Express $(Y_i^o(1), Y_i^o(0))$ analytically as functions of $(Y_i(1), Y_i(0))$.

(c) (1 p) Derive the analytical relationship between ATE^o and ATE, and explain it in words.

(d) (2 p) Modify the setting as follows: the coin is unbalanced, landing Heads (take-up) with probability 20%. Agree or disagree with the following statement and justify your answer: *The mean impact of undergoing the NSW program is five times the mean impact of being offered participation in the NSW program.*

(e) (2 p) Modify the setting as follows: there is no coin flip. Instead, an individual with an offer participates if and only if their potential earnings under the NSW program are at least as high as their potential earnings without it. Derive the analytical relationship between ATE^o and ATE, and explain it in words.

(f) (2 p) Agree or disagree with the following statement and justify your answer: *Given a randomly assigned offer of treatment, the ITT differs from the ATE only when individuals self-select into treatment based on factors that also influence their potential outcomes.*

Part 2: Describe the NSW Pseudo-Observational Datasets (10 p)

Background: In **Pset 2**, you estimated the ATE of the NSW *offer* on 1978 earnings using the NSW experimental data. Because treatment is URA, this is also the ATT of the NSW *offer*. The Difference-in-Means (DM) estimate indicated that the NSW *offer* increased 1978 earnings by approximately \$1,794 per unit. In real-world applications, however, variation in the treatment is often *observational* rather than experimental. As discussed in class, a common—though stringent—assumption researchers make when working with observational data is that within narrowly defined subpopulations (i.e., units sharing the same values of observed pre-treatment variables, or OPVs), treatment can be treated *as if* randomly assigned. In practice, this assumption motivates the use of various forms of regression adjustment when estimating treatment effects without random assignment. Naturally, we would like to know how well these observational estimators perform. The difficulty, of course, is that researchers typically do not know the true effect and thus cannot compare their estimates to the truth. In a seminal 1986 paper, the economist Robert LaLonde[4] proposed a clever way to approximate such a comparison. In this question, you will follow in his footsteps by using two pseudo-observational datasets constructed by Dehejia and Wahba.[5] Their datasets mimic the type of observational data one might use to estimate the impact of the NSW *offer* in the absence of experimental variation. Specifically, Dehejia and Wahba combined the treated sample from the NSW experiment with samples of individuals drawn from two large national surveys.

Q2. (2 p) Consider the files `nswcps.csv` and `nswpsid.csv`. Each file contains a dataset that merges two samples:

(1) The treated sample from Dehejia and Wahba's NSW data (i.e., 185 men offered the NSW program), and

(2) A sample drawn from a large survey providing comparison (non-treated) individuals:

(2a) In `nswcps.csv`, the survey is the Current Population Survey (CPS).

(2b) In `nswpsid.csv`, the survey is the Panel Study of Income Dynamics (PSID).

The survey samples serve as a comparison group—that is, individuals who (presumably) did not receive the NSW *offer*.[6] Specifically:

- The PSID sample (denoted **PSID-1**) consists of 2,490 male household heads under age 55 who are not retired.
- The CPS sample (denoted **CPS-1**) consists of 15,992 male household heads under age 55 who are not retired.

Hence: `nswpsid.csv` contains the NSW-treated individuals and the PSID comparison sample; `nswcps.csv` contains the NSW-treated individuals and the CPS comparison sample. In both files, the treatment indicator variable `treat` equals 1 for NSW-treated individuals and 0 for comparison-sample individuals. Fill columns [5] and [6] of **Table 1** using, respectively, the data in `nswpsid.csv` and `nswcps.csv`.

Q3. (4 p) Briefly comment on the completed **Table 1**.

[4] Robert J. LaLonde (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, pp. 604–620. https://www.jstor.org/stable/1806062.

[5] Dehejia, R. H., and S. W. Wahba (1999). "Causal Effects in Non-Experimental Studies: Reevaluating the Evaluation of Training Programs." *JASA*, pp. 1053–1062. Dehejia and Wahba (2002). "Propensity-Score Matching Methods for Non-Experimental Causal Studies." *Review of Economics and Statistics*, pp. 151–161.

[6] When working with observational data, the untreated sample is more properly referred to as a comparison group. In practice, however, the terms control and comparison are often used interchangeably, regardless of whether treatment assignment is random.

Q4. (4 p) Comparing non-experimental estimates (based on observational data) to experimental benchmarks (based on RCTs) requires access to both survey and experimental data for the same population and period, and ideally to programs available in both contexts. This situation is rarely achievable: programs evaluated via RCTs are often unavailable to the general population, and RCT samples frequently differ from those in large-scale surveys designed to represent broader populations. Based on these challenges, explain the logic of LaLonde's method for mimicking observational data—later adopted by Dehejia and Wahba. A single concise paragraph suffices.

Variable	Definition	NSW		PSID-1	CPS-1
		Treated	Control	Control	Control
[1]	[2]	[3]	[4]	[5]	[6]
`age`	Age (in years)	25.82	25.05		
`edu`	Education (in years)	10.35	10.09		
`nodegree`	= 1 if education < 12, 0 otherwise	0.71	0.83		
`black`	= 1 if Black, 0 otherwise	0.84	0.83		
`hisp`	= 1 if Hispanic, 0 otherwise	0.06	0.11		
`married`	= 1 if married, 0 otherwise	0.19	0.15		
`u74`	= 1 if unemployed in 1974, 0 otherwise	0.71	0.75		
`u75`	= 1 if unemployed in 1975, 0 otherwise	0.60	0.68		
`re74`	Real earnings in 1974 (1982 $)	2,096	2,107		
`re75`	Real earnings in 1975 (1982 $)	1,532	1,267		
`re78`	Real earnings in 1978 (1982 $)	6,349	4,555		
`treat`	= 1 if received the NSW *offer*, 0 otherwise	1	0	0	0
Sample Size		185	260	2,490	15,992

Table 1: Sample averages for the NSW sample (treated and control groups), PSID-1 sample, and CPS-1 sample.

Part 3: Target Estimand — ATE versus ATT of the NSW *offer* (10p bonus)

Q5. In **PSet 2**, when you used the NSW experimental data, you estimated the Average Treatment Effect (ATE) of the NSW *offer* of training and employment on 1978 earnings. By contrast, in **Part 4** of this problem set, when using the pseudo-observational dataset `nswpsid.csv`, you are asked to estimate the Average Treatment Effect on the Treated (ATT) of the same *offer*, using methods that adjust for observable confounders. Explain why, in the pseudo-observational setting, the relevant target estimand is the ATT rather than the ATE. In your answer, address the following:

(a) (2p) What population the pseudo-observational dataset `nswpsid.csv` represents.

(b) (2p) Which group in this dataset corresponds to the treated group from the NSW experiment.

(c) (4p) Why estimating the ATE using this dataset would not yield a quantity comparable to the experimental ATE.

(d) (2p) State—in words and notation—the causal quantity that regression-adjusted estimates from the pseudo-observational data aim to recover, and explain why it is directly comparable to the experimental benchmark.

Part 4: Regression-Based Estimation of Treatment Effects with Pseudo-Observational Data (80 p)

Learning Objectives: Using the pseudo-observational dataset `nswpsid.csv`, you estimate the ATT of the NSW *offer* through regression-based approaches under two specifications of the conditional expectation function (CEF): (1) linear in parameters; and (2) partially linear in parameters. In doing so, you replicate (in spirit) LaLonde's approach and ask: How close are the ATT estimates based on the pseudo-observational data to the estimates obtained using the experimental data? Put differently, do the specifications considered—through their control for OPVs —resolve the confounding problem that arises because several OPVs are highly imbalanced and many are likely determinants (or functions of determinants) of the outcome variable?

Q6. (9 p) As a benchmark, obtain the Difference-in-Means (DM) estimator of the effect of the NSW *offer* by running:

$$\text{Regress } \texttt{re78}_i \text{ on } (1, D_i) \text{ with parameters } (\alpha, \rho), \tag{1}$$

where subscript i indexes individuals ($i = 1, \ldots, 2675$); $\texttt{re78}_i$ is the variable `re78`; and D_i is the treatment indicator treat.

(a) (1 p) Estimate ρ by OLS and compute the standard error under the assumption of homoskedasticity.

(b) (3 p) Compute heteroskedasticity-robust standard errors. (If necessary, review **Topic 1**: Heteroskedasticity-Robust SEs.)

(c) (5 p) Explain why the DM estimator above may not be "credible" (i.e., consistent) for the ATT of the NSW *offer*.

Q7. (8p) Understand and prove **Claim 1**.

Claim 1 (Identification and Estimation of ATT via Regression Adjustment) *Consider the RCM model with* $(Y_i(1), Y_i(0), D_i, \mathbf{x}_i) \overset{iid}{\sim} G$ *for some distribution function* G. $\mathbf{x}_i$ *denotes the vector of OPVs (or functions of OPVs) with support* $\mathcal{X}$. *Assume:*

(a) ***Conditional Overlapping Condition (COC)****:* $0 < Pr(D_i = 1|\mathbf{x}_i = \mathbf{x}) < 1 \ \forall \mathbf{x} \in \mathcal{X}$.

(b) ***Baseline Conditional Mean Independence*** *(*$CMIA_0$*):* $E[Y_i(0)|D_i = 1, \mathbf{x}_i = \mathbf{x}] = E[Y_i(0)|D_i = 0, \mathbf{x}_i = \mathbf{x}] = E[Y_i(0)|\mathbf{x}_i = \mathbf{x}] \ \forall \mathbf{x} \in \mathcal{X}$.

(c) ***Linearity of the CEFs of POs:*** $E[Y_i(0)|\mathbf{x}_i = \mathbf{x}] = \alpha_0 + \boldsymbol{\theta}_0^T \mathbf{x}$ *and* $E[Y_i(1)|D_i = d, \mathbf{x}_i = \mathbf{x}] = \alpha_1 + \gamma d + \boldsymbol{\theta}_1^T \mathbf{x}$ *with* $d \in \{0, 1\}$ *and* $\mathbf{x} \in \mathcal{X}$, *where* $(\alpha_0, \alpha_1, \gamma)$ *are scalars and* $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ *are vectors.*

(d) ***Homogeneity****:* $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1 = \boldsymbol{\theta}$.

Then,

(i) (1p) $ATT := E[Y_i(1) - Y_i(0)|D_i = 1]$ *is identified subject to assumptions* ***(a)-(b)***.

(ii) (3p) $E[Y_i|D_i = d, \mathbf{x}_i = \mathbf{x}] = \alpha + \rho d + \boldsymbol{\beta}^T \mathbf{x}$ *for some* $(\alpha, \rho, \boldsymbol{\beta})$ *subject to assumptions* ***(a)-(d)****, i.e., it is linear-in-parameters.*

Suppose that you have a random sample $\{(Y_i, D_i, \mathbf{x}_i)|i = 1, \ldots, n\}$ *where* $Y_i := Y_i(1)D_i + Y_i(0)(1 - D_i)$ *denotes the observed outcome and* $(D_i, \mathbf{x}_i)$ *are not perfectly collinear. Regress* Y_i *on* $(1, D_i, \mathbf{x}_i)$ *with coefficients* $(\alpha, \rho, \boldsymbol{\beta})$. *Let* $\hat{\rho}$ *denote the OLS estimator of* ρ.

(iii) (1p) $\hat{\rho}$ *is a consistent estimator of* ρ *subject to assumptions* ***(a)-(d)***.

(iv) (3p) $\rho = ATT$, *i.e.,* $\hat{\rho}$ *is a consistent estimator of ATT subject to assumptions* ***(a)-(d)***.

Q8. (8p) Here you obtain a Regression-Adjusted Difference-in-Means (Adj-DM) Estimator of the effect of the NSW *offer* via:

$$\text{Regress} \quad \texttt{re78}_i \quad \text{on} \quad (1, D_i, \mathbf{x}_i) \quad \text{with parameters} \quad (\alpha, \rho, \boldsymbol{\beta}), \tag{2}$$

with[7]

$$\mathbf{x}_i = (\texttt{age}_i, \texttt{agesq}_i, \texttt{edu}_i, \texttt{nodegree}_i, \texttt{black}_i, \texttt{hisp}_i, \texttt{re74}_i, \texttt{re75}_i).$$

That is, you implement Procedure A:

Procedure A to Estimate the parameters of a MLRM

(a) Assume $E[Y_i|D_i = d, \mathbf{x}_i = \mathbf{x}] = \alpha + \rho d + \boldsymbol{\beta}^T \mathbf{x}$.[a]

(b) Regress Y_i on a constant D_i and $\mathbf{x}_i$, denote the OLS estimator of the slope parameter by $\hat{\rho}_A$.

[a] **Claim 1** provides the restrictions that yield this CEF for the observed outcome.

(a) (3p) Report the estimate of ρ and the heteroskedasticity-robust SE.

(b) (5p) Why may the Adj-DM approach improve over the DM approach when our target is the ATT of the NSW *offer*?

Q9. (6p) Consider again Specification 2 estimated in **Q8**. Here you implement two alternative procedures, as detailed below, to verify the "partialling-out" interpretation of OLS coefficients in MLRM. (If necessary, review **Topic 2**: Partialling-out Interpretation of OLS).

[7] From a pedagogical standpoint, we have you use both age in linear form (`age`) and age squared (`agesq`) to underscore that the linearity that has bite is linearlity-in-parameters.

Procedure B to Estimate the parameters of a MLRM

(a) Assume $E[Y_i|D_i = d, \mathbf{x}_i = \mathbf{x}] = \alpha + \rho d + \boldsymbol{\beta}^T \mathbf{x}$.

(b) Without loss of generality: $D_i = m_0 + \mathbf{m}_1^T \mathbf{x}_i + v_i$ with $E[v_i] = 0$ and $Cov(v_i, \mathbf{x}_i) = 0$.

(c) **First Stage**: Regress D_i on a constant and $\mathbf{x}_i$ and obtain the OLS estimator $(\hat{m}_0, \hat{\mathbf{m}}_1)$ and residuals $\hat{v}_i := D_i - \hat{D}_i = D_i - (\hat{m}_0 + \hat{\mathbf{m}}_1^T \mathbf{x}_i)$.

(d) **Second Stage**: Regress Y_i on a constant and $\hat{v}_i$, denote the OLS estimator of the slope parameter by $\hat{\rho}_B$.

Procedure C to Estimate the parameters of a MLRM

(a) Assume $E[Y_i|D_i = d, \mathbf{x}_i = \mathbf{x}] = \alpha + \rho d + \boldsymbol{\beta}^T \mathbf{x}$.

(b) Without loss of generality: $D_i = m_0 + \mathbf{m}_1^T \mathbf{x}_i + v_i$ with $E[v_i] = 0$ and $Cov(v_i, \mathbf{x}_i) = 0$.

(c) Without loss of generality: $Y_i = l_0 + \mathbf{l}_1^T \mathbf{x}_i + \epsilon_i$ with $E[\epsilon_i] = 0$ and $Cov(\epsilon_i, \mathbf{x}_i) = 0$.

(d) **First Stage for** D_i: Regress D_i on a constant and $\mathbf{x}_i$ and obtain the OLS estimator $(\hat{m}_0, \hat{\mathbf{m}}_1)$ and residuals $\hat{v}_i := D_i - \hat{D}_i = D_i - (\hat{m}_0 + \hat{\mathbf{m}}_1^T \mathbf{x}_i)$.

(e) **First Stage for** Y_i: Regress Y_i on a constant and $\mathbf{x}_i$ and obtain the OLS estimator $(\hat{l}_0, \hat{\mathbf{l}}_1)$ and residuals $\hat{\epsilon}_i := Y_i - \hat{Y}_i = Y_i - (\hat{l}_0 + \hat{\mathbf{l}}_1^T \mathbf{x}_i)$.

(f) **Second Stage**: Regress $\hat{\epsilon}_i$ on $\hat{v}_i$, denote the OLS estimator of the slope parameter by $\hat{\rho}_C$.

(a) (2p) Implement Procedure B and verify that $\hat{\rho}_A = \hat{\rho}_B$.

(b) (2p) Implement Procedure C and verify that $\hat{\rho}_A = \hat{\rho}_C$.

(c) (2p) Use the above findings to give meaning to the expression "partialling-out" interpretation of OLS in a MLRM.

Q10. (18p) Here you estimate the effect of the NSW *offer* via Chernozhukov et al (2018)'s Double Machine Learning (DML) estimation procedure[8] based on a partially-linear regression (PLR) specification introduced by Robinson (1988)[9][10]

$$\text{Regress} \quad \texttt{re78}_i \quad \text{on} \quad (D_i, g(\mathbf{x}_i)) \quad \text{with parameters} \quad (\rho, g), \tag{3}$$

where:

$$\mathbf{x}_i \quad \text{includes all OPVs as well as a column of 1's.}$$

and g is an unknown and possibly non-linear function, i.e., a generalization of $\alpha + \boldsymbol{\beta}^T \mathbf{x}_i$ in Specification 2. That is, you implement Procedure D:

Procedure D to Estimate the parameters of a Partially Linear Regression Model

(a) Assume $E[Y_i|D_i = d, \mathbf{x}_i = \mathbf{x}] = \rho d + g(\mathbf{x})$.

(b) Let $m(\mathbf{x}) := E[D_i|\mathbf{x}_i = \mathbf{x}]$.

(c) Let $l(\mathbf{x}) := E[Y_i|\mathbf{x}_i = \mathbf{x}]$.

(d) **First Stage for** D_i: "Regress" D_i on $m(\mathbf{x}_i)$ and obtain the residuals $\hat{v}_i := D_i - \hat{m}(\mathbf{x}_i)$.

(e) **First Stage for** Y_i: "Regress" Y_i on $l(\mathbf{x}_i)$ and obtain the residuals $\hat{\epsilon}_i := Y_i - \hat{l}(\mathbf{x}_i)$.

(f) **Second Stage**: Regress $\hat{\epsilon}_i$ on $\hat{v}_i$, denote the OLS estimator of the slope parameter by $\hat{\rho}_D$.

[8] Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. (2018). "Double/Debiased Machine Learning for Treatment and Structural Parameters." The Econometrics Journal 21(1): C1–C68. https://doi.org/10.1111/ectj.12097

[9] Robinson, P. M. (1988). Root-N-consistent semi-parametric regression. Econometrica 56, 931-54. doi:10.2307/1912705

[10] This specification is called partially linear because it is linear in D_i but possibly not linear in $\mathbf{x}_i$.

(a) (2p) Why may we prefer Specification 3 over specification Specification 2?

(b) (2p) Write a sentence to describe in plain English the parallels between Procedure C for a linear-in-parameter model and Procedure D for a partially linear-in-parameter model.

(c) (1p) Install four R packages: `DoubleML`, `data.table`, `mlr3`, and `mlr3learners`.

(d) (1p) If your data is not already a `data.table` object convert it.

(e) (1p) Collect all the original OPVs in a list named, for example, `pretreat_colnames`. Note: Henceforth when we refer to these OPVs in mathematical expressions we use the notation $\mathbf{x}_i$. (Note: you don't need to add a constant, the R functions below will do that on your behalf and on the fly.)

(f) (2p) Specify data and variables for the causal model by running the script:

```
dml_data_psid <- DoubleML::DoubleMLData$new(dt,
                                            y_col = "re78",
                                            d_cols = "treat",
                                            x_cols = pretreat_colnames)
```

Look at the resulting object (print it to stout, determine its class).

(g) (1p) Add to your script `lgr::get_logger("mlr3")$set_threshold("warn")` to suppress messages from the `mlr3` package.

(h) (2p) Consider two conditional expectation functions:

$$l(\mathbf{x}) := E[\texttt{re78}_i|\mathbf{x}_i = \mathbf{x}],$$
$$m(\mathbf{x}) := E[\texttt{treat}_i|\mathbf{x}_i = \mathbf{x}].$$

Run the script:

```
# Specify a RF model as the learner model for l(x)=E[re78|X=x]
ml_l_rf <- mlr3::lrn("regr.ranger")

# Specify a RF model as the learner model for m(x)=E[treat|X=x]
ml_m_rf <- mlr3::lrn("classif.ranger")
```

The above script declares that we want to use a Random Forest (RF) approach to get $\hat{l}(\mathbf{x})$ and $\hat{m}(\mathbf{x})$ for all values $\mathbf{x}$. (Note: You do not need to know what a RFM is. It suffices for you to think of this approach as a way to flexibly estimate the form of a function of many variables.)

(i) (2p) Here you initialize and parametrize the model object (use below to perform estimation). Run the script:

```
# Set seeds for cross-fitting
set.seed(3141)

# Set the DML specification
obj_dml_plr <- DoubleML::DoubleMLPLR$new(dml_data_psid,
                                         ml_l = ml_l_rf, ml_m = ml_m_rf,
                                         n_folds = 2,
                                         score = "partialling out",
                                         apply_cross_fitting = TRUE)
```

The above script specifies:

- the data object generated in **Q10f**, namely `dml_data_psid`;
- the models for the first stage regressions picked in **Q10h**, namely `ml_l_rf` and `ml_m_rf`;
- that we want to split the sample into 2 parts (`n_folds = 2`),
- that we want to use the "partialling out" approach to estimate causal impacts (`score = "partialling out"`), and
- that we want to apply cross-fitting (`apply_cross_fitting = TRUE`).

(j) (2p) Here you fit the DML model defined in **Q10i**. Run the script:

```
obj_dml_plr$fit()
obj_dml_plr
```

At a high level the above script implements all of the following operations:

- fits the two models for the first stage selected in **Q10h**, i.e., gets $\hat{l}(\cdot)$ and $\hat{m}(\cdot)$.
- gets residuals, i.e., $\hat{\epsilon}_i = \texttt{re78}_i - \hat{l}(\mathbf{x}_i)$ and $\hat{v}_i = \texttt{treat}_i - \hat{m}(\mathbf{x}_i)$

- implements the second stage, i.e., regresses $\hat{\epsilon}_i$ on $\hat{v}_i$ to obtain the DML estimate of ρ in Specification 3 .

Because you specified `n_folds = 2` and requested `apply_cross_fitting = TRUE` in **Q10i** the 2-stage estimation procedure proceed as follows. First the entire data is split into two sub-samples, call them S_1 and S_2 (hence the term "2 folds"). Sample S_1 is used to fit the 1st stage models. These fitted models are used to compute residuals in sample S_2 and these residuals are used to fit the 2nd stage model using only data in sample S_2. Denote the resulting estimate $\hat{\rho}_{1,2}$. Then the samples are swapped (hence the term "cross fitting").[11] That is, sample S_2 is used to fit the 1st stage models. Sample S_1 is used to fit the 2nd stage model. Denote the resulting estimate $\hat{\rho}_{2,1}$. The DML estimate is the average of $\hat{\rho}_{1,2}$ and $\hat{\rho}_{2,1}$.

(k) (2p) Take a look at the output, i.e., at `obj_dml_plr`. What is the DML estimate of the ATT of the NSW *offer* and SE?

Q11. (10p) Compare the estimates of ATT of the NSW *offer* based on the pseudo-observational data and Specification 1 , Specification 2 , and Specification 3 , to those based on experimental NSW data and the DM estimator (from **PSet 2**). In his 1986 article LaLonde concluded that the non-experimental methods available at the time could not systematically replicate experimental benchmarks, casting doubt on the credibility of these methods. How do you feel about this conclusion in light of your findings? Weave all of this into a couple of well-written paragraphs.

Q12. (21p) Read Imbens and Xu's 2024 article "LaLonde (1986) after Nearly Four Decades: Lessons Learned" through Section 4.1 included (`Canvas/Files/Articles/Imbens-Xu-Lessons-Learned-LaLonde-2024.pdf`). How has the conclusion you reached in **Q11** changed after reading this article (if at all)? Write 3 paragraphs that summarize the role of overlap (and trimming) for the ability of non-experimental methods to replicate experimental benchmarks.

Guidance

Before answering any of the questions that use the NSW data, divide `re74` and `re75` by $1,000$; this avoids working with OPVs with vastly different scales. You did this in PSet 2 too.

Q2. **Notes**: You want to limit attention to observations with `treat=0`. You filled **Table 1**'s columns 3 and 4 in PSet 2.

Q3. **Hint**: How do the PSID-1 and CPS-1 samples compare to the NSW-treated sample? Are the PSID-1 and CPS-1 samples "good" control groups? Does what you see raise concerns about using the combined NSW-treated and PSID-1 samples (or NSW-treated and CPS-1 samples) to estimate the effect of the NSW *offer* ? If so, why?

Q4. **Hint:** Both PSID and CPS include information on whether an individual was employed and whether they enrolled in a training course during the previous 12 months. Thus, LaLonde (or Dehajia and Wahba) could have exploited exclusively observational variation in whether an individual is employed and enrolled in a training program. Why do you think that they chose not to follow this approach to create observational data to study the performance of non-experimental methods?

Q6. Here is some guidance:

(a) **Programming Guidance:** Use `stats::lm( )` which by defaults returns estimates of the variance-covariance matrix of the OLS estimator based on the assumption of homoskedasticity. Say that your linear model is `m1 <- lm(re78 ~ treat, data = df)`. There are many ways to access the SE of estimator $\hat{\rho}$ computed under homoskedasticity. (1) Use `summary(m1)$coefficients["treat", c("Estimate", "Std. Error")]`, or (2) `sqrt(diag(vcov(m1)))`. Alternatively, (3) grab all SEs by using `lmtest::coeftest(m1, vcov. = vcov(m1))`. This script runs t-tests for each of the coefficients using the variance-covariance matrix that you supply via the argument `vcov.`, which in this case is the variance-covariance matrix estimated assuming homoskedasticity. The appeal of approach (3) is that you may supply alternative estimates of the variance-covariance matrix (you will do this below). Package `lmtest` allows you to perform z and t tests on estimated coefficients from, among others, method `lm( )`. It returns a coefficient matrix with columns containing the estimates, associated SEs, test statistics, and p-values.

(b) **Programming Guidance:** There are several R packages to estimate the variance-covariance matrix of $(\hat{\alpha}, \hat{\rho})$ under general heteroskedasticity. Here are two ways:

- Use `sandwich::vcovHC(m1, type = ''HC0")` from package `sandwich`.
- Use `car::hccm(m1, type = ''hc0")` from package `car`.

In both cases, the argument `type = "hc0"` (or `"HC0"`) tells R that you want to use the variance covariance matrix estimated using White's (1980) estimator, often referred to as HCE (heteroscedasticity-consistent estimator). Display robust SEs by typing, e.g., `lmtest::coeftest(m1, vcov. = sandwich::vcovHC(m1, type = "HC0"))`.

[11] Cross-fitting is implemented to eliminate the bias from overfitting resulting from the fact that the two conditional mean functions $l(\cdot)$ and $m(\cdot)$ are estimated via ML models, in our case the RF models specified in **Q10h**.

(c) **Hint**: From **PSet1** we know that $\hat{\rho} = \overline{re78}_1 - \overline{re78}_0$, i.e., the DM estimator. Your answer must address: (a) What would the DM estimate say if regarded as a credible estimate of the NSW *offer*? (b) Under what formal conditions would the population counterpart of the DM estimator identify ATT? (c) Are these conditions plausible in light of the evidence you have collected so far?

Q7. **Note**: You may quote and use results proven in past PSets AS-IS. The reason why we ask you to prove this claim is because you will want to use it when answering subsequent questions.

Q8. Here is some guidance:

(a) **Programming Guidance:** Add column `agesq` (`age` squared) to your data frame using, e.g., `dplyr::mutate( )`.

(b) **Hint**: Think about what you learnt from **Claim 1**.

Q9. **Note #1**: We ask you to review the partialling-out interpretation of OLS because you will leverage it in subsequent questions. You already did this in **PSet 2**, so this is a bit of a repetition. **Note #2**: In both Procedures there are statements that read "without loss of generality etc. etc." The theory underpinning those statements is this claim:

Claim 2 (Linear Decomposition) *Let X and Z be two random variables with finite mean. Z has also finite variance. Then, there exist two scalars (π_0, π_1) and a random variable V such that:*

$$X = \pi_0 + \pi_1 Z + V, \tag{4}$$

where V has two properties: (a) $E[V] = 0$, and (b) $Cov(Z, V) = 0$. (This decomposition extends to the case when $\mathbf{Z}$ is a random vector.)

Here is some guidance:

(a) **Programming Guidance:** If you run `s1 <- lm(treat ~ x1 + x2, data = dt)`, retrieve the residuals as `s1$residuals`.

Q10. Here is some guidance:

(a) **Pedagogy:** In ECMA 31360, we ask you to implement Double Machine Learning (DML) as a black box: we use the method in practice without having covered the full theory that explains why it can deliver consistent estimators under CIA+COC. This is intentional. The goal of this exercise is not for you to master the DML theory, but to take away a few big ideas and practical lessons:

i. **Flexibility about functional form.** In many empirical settings, we would like to avoid strong functional form assumptions (FFAs). DML is one way to remain more agnostic about the form of key conditional expectation functions— $\mathbb{E}[Y_i(0) \mid D_i = d, \mathbf{x}_i = \mathbf{x}]$ and $\mathbb{E}[Y_i(1) \mid D_i = d, \mathbf{x}_i = \mathbf{x}]$ —while still targeting a causal parameter (here, the ATT) under CIA+COC.

ii. **Why this can be valuable.** FFAs are rarely justified by economic theory; in practice they are often chosen for convenience. DML highlights that, at least in principle, we can let data-driven methods learn flexible relationships between potential outcomes/treatment and covariates, rather than committing to a specific linear specification.

iii. **Costs and trade-offs.** Flexibility is not free. DML typically performs well only with large samples, and it requires researcher choices (e.g., choice of learner, tuning, and cross-fitting settings). So judgment does not disappear—it simply shifts from choosing a functional form to choosing and validating a learning procedure.

If you would like to see the theoretical foundations and the conditions under which DML works (and why cross-fitting matters), consider courses such as ECMA 31350 and ECMA 31380.

(d) **Programming Guidance:** Assuming that your data frame is called `df`, use `dt <- data.table::as.data.table(df)`. `data.table` is an extension of `data.frame` and allows for fast manipulation of very large data.

Q12. **Note:** The article mentions an object called propensity score. You have already encountered it, though not under this name. Consider the assignment probability $p(\mathbf{x}) := \Pr(D_i = 1 \mid \mathbf{x}_i = \mathbf{x})$. It is a function from the support of $\mathbf{x}_i$, denoted $\mathcal{X}$, (domain) to the unit interval $[0, 1]$ (range). We call this function the propensity score function because it yields the probability ("propensity") that a unit with OPVs taking value $\mathbf{x}$ has to be assigned to be treated. Let $p_i := p(\mathbf{x}_i) := \Pr(D_i = 1 \mid \mathbf{x}_i)$, then p_i is a random variable (because it is a function of the RVs $\mathbf{x}_i$) whose population distribution f_p is induced by the population distribution $f_{\mathbf{x}}$ via the assignment process.

Review of Selected Topics in Econometrics

TOPIC 1. **Heteroskedasticity-Robust Standard Errors**.

Let's return to introductory econometrics. Consider an outcome variable Y_i and a collection of explanatory variables (or covariates) $\mathbf{x}_i$. Assume that $(Y_i, \mathbf{x}_i)$ are independent across units but not necessarily identically distributed. To accommodate the possibility of unit-specific distributions, we write

$$Y_i \mid \mathbf{x}_i = \mathbf{x} \overset{independent}{\sim} f^{(i)}_{Y|\mathbf{x}},$$

with $f^{(i)}_{Y|\mathbf{x}} \neq f^{(i')}_{Y|\mathbf{x}}$ for some (i, i'). You already know that in econometrics, $\mathbb{E}_i[Y_i \mid \mathbf{x}_i = \mathbf{x}]$ is called the conditional expectation function (CEF).[12] Let $\mathcal{I}$ denote the index set—that is, the collection of labels for each unit in the conceptual reference population (possibly infinite). Let $\mathcal{X}$ be the set of possible values of the covariates. In full generality, the CEF is a function from $\mathcal{I} \times \mathcal{X}$ to the real line, allowing the conditional expectation to vary across both units and covariate values. Econometric analysis typically assumes that the CEF is the same for all units sharing the same value of the covariates,[13] and we then write this common CEF simply as $m(\mathbf{x}) := \mathbb{E}[Y_i \mid \mathbf{x}_i = \mathbf{x}]$. Under this assumption, the CEF is a real-valued function from $\mathcal{X}$ to the real line.

We can similarly define the skedastic function as the conditional variance

$$\mathrm{Var}_i[Y_i \mid \mathbf{x}_i = \mathbf{x}],$$

a function from $\mathcal{I} \times \mathcal{X}$ to the positive real line. We often use the shorthand notation

$$\sigma^2_i := \mathrm{Var}_i[Y_i \mid \mathbf{x}_i = \mathbf{x}].$$

When $\sigma^2_i = \sigma^2_{i'}$ for all (i, i'), we say that homoskedasticity is present. Conversely, when $\sigma^2_i \neq \sigma^2_{i'}$ for some (i, i'), we say that heteroskedasticity is present.[14] Heteroskedasticity is therefore one way in which the distributions of $(Y_i, \mathbf{x}_i)$ can differ across units. A particular form of heteroskedasticity arises when

$$\mathrm{Var}_i[Y_i \mid \mathbf{x}_i = \mathbf{x}] = \mathrm{Var}_{i'}[Y_{i'} \mid \mathbf{x}_{i'} = \mathbf{x}] \quad \text{for all } (i, i'),$$

but

$$\mathrm{Var}[Y_i \mid \mathbf{x}_i = \mathbf{x}] \neq \mathrm{Var}[Y_i \mid \mathbf{x}_i = \mathbf{x}'] \quad \text{for some } (\mathbf{x}, \mathbf{x}').$$

We then say there is heteroskedasticity in $\mathbf{x}_i$, because the conditional variance is the same for any two units with the same value of the explanatory variables but differs across values of $\mathbf{x}$. Formally, in this case the conditional distributions $f^{(i)}_{Y|\mathbf{x}}$ and $f^{(i')}_{Y|\mathbf{x}}$ have the same variance for each given $\mathbf{x}$, but that variance changes with $\mathbf{x}$.

Given a common CEF $m(\mathbf{x}) := \mathbb{E}[Y_i \mid \mathbf{x}_i = \mathbf{x}]$, define the regression unobservable

$$u_i := Y_i - m(\mathbf{x}_i),$$

so that, by construction, $\mathbb{E}_i[u_i \mid \mathbf{x}_i] = \mathbb{E}[u_i \mid \mathbf{x}_i] = 0$. Then, $\mathrm{Var}_i[u_i \mid \mathbf{x}_i = \mathbf{x}] = \mathrm{Var}_i[Y_i \mid \mathbf{x}_i = \mathbf{x}]$. This explains why econometric textbooks often describe homoskedasticity and heteroskedasticity as properties of the conditional distribution of u_i. Specifically, given $\sigma^2_i := \mathrm{Var}_i(u_i \mid \mathbf{x}_i = \mathbf{x})$, there is *arbitrary heteroskedasticity* when $\sigma^2_i \neq \sigma^2_{i'}$ for some (i, i'). A special case is *covariate-dependent* (or x-) heteroskedasticity,

$$\sigma^2(\mathbf{x}) := \mathrm{Var}_i(u_i \mid \mathbf{x}_i = \mathbf{x}) = \mathrm{Var}(u_i \mid \mathbf{x}_i = \mathbf{x}),$$

in which the conditional variance is the same for all units sharing $\mathbf{x}$ but may vary across values of $\mathbf{x}$. In this case, homoskedasticity obtains under the further restriction that $\sigma^2(\mathbf{x}) = \sigma^2(\mathbf{x}')$ for all $(\mathbf{x}, \mathbf{x}')$, and we typically denote this common variance as σ^2.

Why does heteroskedasticity matter?

[12] The subscript i indicates that this expectation is taken with respect to the unit-specific conditional distribution $f^{(i)}_{Y|\mathbf{x}}$.

[13] This is equivalent to assuming that while $f^{(i)}_{Y|\mathbf{x}} \neq f^{(i')}_{Y|\mathbf{x}}$ for some (i, i'), all these distributions have the same mean.

[14] Etymologically, *homoskedasticity* means "same skedastic function," while *heteroskedasticity* means "different skedastic functions."

Given a linear-in-parameters CEF $m(\mathbf{x};\boldsymbol{\beta}) := \mathbb{E}[Y_i \mid \mathbf{x}_i = \mathbf{x}] = \beta_0 + \boldsymbol{\beta}_1^T\mathbf{x}$, you know that the unbiasedness and consistency of the OLS estimator do not depend on the form of the skedastic function. Nevertheless, the form of the skedastic function is crucial for characterizing and estimating the variance of the OLS estimator, and hence for computing standard errors.[15] To see this, consider the case where there is only one covariate, so that we write $m(x;\boldsymbol{\beta}) = \beta_0 + \beta_1 x$. Hence, $Y_i = \beta_0 + \beta_1 x_i + u_i$ with $E[u_i \mid x_i = x] = 0$, and as above we let $\sigma_i^2 := \text{Var}_i[u_i \mid x_i = x]$ to be fully agnostic about the presence and form of heteroskedasticity. Suppose we have access to a random sample $\{(Y_i, x_i) \mid i = 1, \ldots, n\}$ stored in a data frame `df`. Let $\mathbf{X} = \{x_i \mid i = 1, \ldots, n\}$ and (with some abuse of notation) use $\mathbf{x}$ to denote the particular sample realization of $\mathbf{X}$.

Suppose we execute `lm(y ~ x, data=df)`, taking advantage of the function `lm` in the `stats` package. This conveniently produces the OLS estimates of $\boldsymbol{\beta}$, along with their standard errors. Let's focus in particular on the OLS estimator of the slope parameter, $\hat{\beta}_1$, and its standard error. Here is the problem: `lm` defaults to assuming homoskedasticity. Thus, the standard error that it outputs is based on the expression that $\text{Var}[\hat{\beta}_1 \mid \mathbf{X} = \mathbf{x}]$ takes under homoskedasticity. Using σ^2 to denote the common variance and $SST_x := \sum_{i=1}^n (x_i - \bar{x})^2$, this expression is:

$$\text{Var}[\hat{\beta}_1 \mid \mathbf{X} = \mathbf{x}] = \frac{\sigma^2}{SST_x}, \tag{5}$$

and the standard error returned by `lm` is:

$$se[\hat{\beta}_1 \mid \mathbf{X} = \mathbf{x}] = \sqrt{\frac{\hat{\sigma}^2}{SST_x}}, \tag{6}$$

with $\hat{\sigma}^2 := \frac{1}{n-2}\sum_{i=1}^n \hat{u}_i^2$ and $\hat{u}_i := Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$.
If homoskedasticity does not hold, then the SEs computed according to expression (6) are incorrect because the expression of the variance of the OLS estimator of the slope coefficient is:

$$\text{Var}[\hat{\beta}_1 \mid \mathbf{X} = \mathbf{x}] = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2}{SST_x^2}, \tag{7}$$

with σ_i^2 denoting the unit-specific variance. Note that this expression generalizes (5): if $\sigma_i^2 = \sigma^2$ for all i, then (7) reduces to (5). That is, this expression is correct whether or not homoskedasticity holds. Naturally, we would like to compute the SE of $\hat{\beta}_1$ based on expression (7). The obvious challenge is that to do so we need to estimate the n parameters $\{\sigma_i^2\}_{i=1}^n$. This may seem an insurmountable challenge because the data contain only one observation (Y_i, x_i) for each sample unit.

Are we at a standstill?

That is: Do we have to accept SEs computed based on an expression that is not valid under heteroskedasticity? No. In a seminal contribution, econometrician **Halbert White** (1980)[16] showed that we can use $\hat{u}_i^2$ as an estimator for σ_i^2. The resulting estimator is called White's heteroskedasticity-robust estimator of the variance of OLS. Continuing to focus on β_1, this estimator is:

$$\widehat{\text{Var}}_{White}[\hat{\beta}_1 \mid \mathbf{X} = \mathbf{x}] = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \hat{u}_i^2}{SST_x^2}. \tag{8}$$

To be sure, $\hat{u}_i^2$ is not a consistent estimator of σ_i^2. However, in a surprising result, White showed that under mild conditions $\widehat{\text{Var}}_{White}[\hat{\beta}_1^{OLS} \mid \mathbf{X} = \mathbf{x}]$ itself is consistent. Thus, we are justified in using its square root as the heteroskedasticity-robust standard error of $\hat{\beta}_1^{OLS}$.

White's estimator extends to the case of multiple explanatory variables. Consider a generic multiple linear regression model:

$$Y_i = \beta_0 + \sum_{k=1}^K \beta_k x_{k,i} + u_i, \quad \text{with } \mathbb{E}[u_i \mid \mathbf{x}_i = \mathbf{x}] = 0.$$

Suppose we have access to a sample $\{(Y_i, x_{1,i}, \ldots, x_{K,i}) \mid i = 1, \ldots, n\}$ drawn independently from the population. Let $\mathbf{X} = \{(x_{1,i}, \ldots, x_{K,i}) \mid i = 1, \ldots, n\}$, and let $\mathbf{x}$ denote its realized value. Then, White's heteroskedasticity-robust estimator of the variance of $\hat{\beta}_1$ is:

$$\widehat{\text{Var}}_{White}[\hat{\beta}_1 \mid \mathbf{X} = \mathbf{x}] = \frac{\sum_{i=1}^n \hat{r}_{1,i}^2 \hat{u}_i^2}{\left(\sum_{i=1}^n \hat{r}_{1,i}^2\right)^2}, \tag{9}$$

[15] It also matters for the efficiency properties of the OLS estimator, but that is a separate discussion.
[16] White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, 48(4): 817–38. https://www.jstor.org/stable/1912934.

where $\hat{r}_{1,i}$ are the residuals from an auxiliary regression of $x_{1,i}$ on $(1, x_{2,i}, \ldots, x_{K,i})$, and $\hat{u}_i$ are the residuals from regressing Y_i on $(1, x_{1,i}, \ldots, x_{K,i})$. This result follows from the partialling-out interpretation of OLS, discussed in the next topic.

Later econometricians have *tweaked* White's estimator and proposed alternative heteroskedasticity-robust estimators of $\text{Var}[\hat{\beta}_1 \mid \mathbf{X} = \mathbf{x}]$. You can easily implement these estimators in `R` using functions such as `sandwich::vcovHC()` or `car::hccm()`.

TOPIC 2. **"Partialling-Out" Interpretation of OLS in a MLRM (aka Frisch–Waugh–Lovell Theorem).** Simple linear-in-parameter regression models (SLRM) are of the form

$$Y_i = \alpha + \beta x_i + u_i \tag{10}$$

where x_i is a single regression covariate. MLRMs are of the form:

$$Y_i = \alpha + \beta_1 x_{1,i} + \cdots + \beta_K x_{K,i} + u_i \text{ with } K > 1. \tag{11}$$

In PSet1 you derived the form of the OLS estimator of the slope coefficient in SLRM (10), namely

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \underbrace{=}_{\text{also equivalent to}} \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{12}$$

Note that if you regress x_i on a constant, the fitted value is $\hat{x}_i = \bar{x}$, thus the regression residuals are $\hat{r}_i := x_i - \hat{x}_i = x_i - \bar{x}$. Similarly, if you regress Y_i on a constant, the fitted value is $\hat{Y}_i = \bar{Y}$, thus the regression residuals are $\hat{v}_i := Y_i - \hat{Y}_i = Y_i - \bar{Y}$. Accordingly, we can rewrite $\hat{\beta}$ in expression (12) as:

$$\hat{\beta} = \frac{\sum_{i=1}^n \hat{r}_i Y_i}{\sum_{i=1}^n \hat{r}_i^2} \underbrace{=}_{\text{also equivalent to}} \frac{\sum_{i=1}^n \hat{r}_i \hat{v}_i}{\sum_{i=1}^n \hat{r}_i^2}. \tag{13}$$

Similar steps yield a very compact representation of the OLS estimator of the slope coefficient**s** in a MLRM. For example, the OLS estimator of β_1 in MLRM (11) can be written as:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \hat{r}_{1,i} Y_i}{\sum_{i=1}^n \hat{r}_{1,i}^2} \underbrace{=}_{\text{also equivalent to}} \frac{\sum_{i=1}^n \hat{r}_{1,i} \hat{v}_{1,i}}{\sum_{i=1}^n \hat{r}_{1,i}^2}, \tag{14}$$

where $\hat{r}_{1,i}$ denotes the residuals from regressing $x_{1,i}$ on a constant and all remaining regression covariates, i.e., $\{x_{2,i}, \ldots, x_{K,i}\}$ and $\hat{v}_{1,i}$ denotes the residuals from regressing Y_i on a constant and all remaining regression covariates, i.e., $\{x_{2,i}, \ldots, x_{K,i}\}$. Similar expressions hold for $\hat{\beta}_2$, $\hat{\beta}_3$, etc.