ECMA 31360, PSet 2: Causal Inference with Data from a Randomized Controlled Trial

Melissa Tartari, University of Chicago

Part 1: Testing Balance in Observed Predetermined Variables (50 p)

Learning Objectives: Learn how to test balance in observed predetermined variables (OPVs) using data from a randomized controlled trial (RCT) with unconditional random assignment (URA) via a variety of approaches.

Q1. (15p) Read and understand `Canvas/Files/Articles-Companions/Companion-to-NSW-Application.pdf`.

Q2. (23p) Use the NSW experimental dataset (see **PSet1**). Implement Procedure 4 to test balance with respect to the 10 OPVs. Specifically:

(a) (10p) Estimate the parameters of the relevant SUR system: $\boldsymbol{\pi} = \{\pi_{0,j}, \pi_{1,j}\}_{j=1}^{10}$.

(b) (1p) Compare estimates and their SEs to those obtained in **PSet1** via OLS equation-by-equation. Briefly elaborate.

(c) (3p) Expression (1) represents the top-left 4×4 block of the estimated variance-covariance matrix for the SUR estimator $\hat{\boldsymbol{\pi}}$. Why is it the case that (i) $\widehat{Cov}[\hat{\pi}_{0,1}, \hat{\pi}_{1,1}] = -\widehat{Var}[\hat{\pi}_{0,1}]$; (ii) $\widehat{Cov}[\hat{\pi}_{0,1}, \hat{\pi}_{1,2}] = -\widehat{Cov}[\hat{\pi}_{0,1}, \hat{\pi}_{0,2}]$; (iii) $\widehat{Cov}[\hat{\pi}_{0,1}, \hat{\pi}_{1,2}] = \widehat{Cov}[\hat{\pi}_{1,1}, \hat{\pi}_{0,2}]$?

$$\text{Top-left } 4 \times 4 \text{ block} = \begin{bmatrix} \widehat{Var}[\hat{\pi}_{0,1}] & \widehat{Cov}[\hat{\pi}_{0,1}, \hat{\pi}_{1,1}] & \widehat{Cov}[\hat{\pi}_{0,1}, \hat{\pi}_{0,2}] & \widehat{Cov}[\hat{\pi}_{0,1}, \hat{\pi}_{1,2}] \\ \widehat{Cov}[\hat{\pi}_{1,1}, \hat{\pi}_{0,1}] & \widehat{Var}[\hat{\pi}_{1,1}] & \widehat{Cov}[\hat{\pi}_{1,1}, \hat{\pi}_{0,2}] & \widehat{Cov}[\hat{\pi}_{1,1}, \hat{\pi}_{1,2}] \\ \widehat{Cov}[\hat{\pi}_{0,2}, \hat{\pi}_{0,1}] & \widehat{Cov}[\hat{\pi}_{0,2}, \hat{\pi}_{1,1}] & \widehat{Var}[\hat{\pi}_{0,2}] & \widehat{Cov}[\hat{\pi}_{0,2}, \hat{\pi}_{1,2}] \\ \widehat{Cov}[\hat{\pi}_{1,2}, \hat{\pi}_{0,1}] & \widehat{Cov}[\hat{\pi}_{1,2}, \hat{\pi}_{1,1}] & \widehat{Cov}[\hat{\pi}_{1,2}, \hat{\pi}_{0,2}] & \widehat{Var}[\hat{\pi}_{1,2}] \end{bmatrix}. \quad (1)$$

(d) (4p) Manually (still in R) test the joint hypothesis that the coefficients of `treat` are zero in all the equations of the system.

(e) (4p) Automate the test run in **Q2d**.

(f) (1p) What decision do you reach when $\alpha = 5\%$?

Q3. (3p) Implement Procedure 5 (Hotelling's T2 test). What decision do you reach when $\alpha = 5\%$?

Q4. (4p) Implement Procedure 6 (Composite Balance Score). What decision do you reach when $\alpha = 5\%$?

Q5. (5p) Implement Procedure 7 (Bonferroni) and Procedure 8 (Bonferroni-Holm). Comment on your findings when $\alpha = 5\%$.

Part 2: Estimating the Effect of the NSW Intervention (50 p)

Learning Objectives: Learn how to estimate the average treatment effect (ATE) of receiving a training offer on post-intervention earnings (`re78`) using the approaches listed in Table 1.

Spec	Estimator	Description
1	DM	Estimate by OLS the parameters of a simple linear regression model (SLRM)
2	Adjusted DM	Add to spec. 1 OPVs `nodegree` and `edu` in linear form
3	Adjusted DM	Add to spec. 2 the other 8 OPVs in linear form
4	Adjusted DM	Add to spec. 3 the interaction between `age` (in deviation from its sample mean) and `treat`

Table 1: Specifications used to estimate the ATE of the offer of training. "DM" stands for Difference-in-Means.

Q6. (3p) Implement Specification 1 and manually (still in R) verify that the OLS estimator of the slope coefficient equals the difference in average earnings in 1978 between the treated and control subsamples. Describe your findings.

Q7. (4p) In light of the imbalance in OPVs documented in Part 1, you may consider assuming that the NSW treatment is assigned randomly conditional on having a degree. Estimate a fully saturated specification in `nodegree`, that is, regress $\texttt{re78}_i$ on $(\texttt{nodegree}_i, (1-\texttt{nodegree}_i), D_i \times \texttt{nodegree}_i, D_i \times (1-\texttt{nodegree}_i))$ (don't include a constant) with parameters $(\alpha_1, \alpha_2, \rho_1, \rho_2)$. What is the estimate of the $CATE$ for individuals with and without a degree? What is the estimate of ATE?

Q8. (2p) In **Q7** you took an extra step to get an estimate of ATE. Run a specification that directly delivers such estimate.

Q9. (22p) In the NSW application, the only OPV with significant imbalance is the binary 0/1 variable `nodegree`. You may, however, encounter situations in which the OPV with significant imbalance takes many values — perhaps even too many (relative to sample size) to feasibly implement a fully saturated specification. More generally, you may encounter several imbalanced OPVs.[1] In such situations, a common shortcut is to include (some of the) OPVs as regression covariates *without* interacting them with the treatment indicator. When the OPVs are discrete, this amounts to including group indicators for the values of the OPVs, a specification that in econometrics is referred to as a model with group fixed effects.

Claim 1 considers a simplified setting in which there is a single OPV taking M distinct values and characterizes the implications of using this fixed-effects shortcut instead of employing a fully saturated specification. Prove the claim (20p) and describe in plain English what you have learned from it (2p).

Claim 1 *Consider the RCM framework and assume that CIA and COC hold with reference to the OPV x_i, which takes values in $\{a_1, \ldots, a_M\}$. Let s_m denote the share of the population with $x_i = a_m$. You have data for a random sample $\{(Y_i, x_i, D_i) \mid i = 1, \ldots, N\}$. Let $\hat{s}_m$ denote the fraction of sample units with $x_i = a_m$. Let $\bar{Y}_{1,m}$ denote the average of Y_i in the treated subsample with $x_i = a_m$ and, similarly, $\bar{Y}_{0,m}$ denote the average of Y_i in the control subsample with $x_i = a_m$. Estimate by OLS the parameters of the following two regression specifications:*

$$\textit{Regress} \quad Y_i \quad \textit{on} \quad \left(\{1[x_i = a_m]\}_{m=1}^M, \{D_i 1[x_i = a_m]\}_{m=1}^M\right) \quad \textit{with parameters} \quad (\alpha_1, \ldots, \alpha_M, \rho_1, \ldots, \rho_M); \tag{2}$$

$$\textit{Regress} \quad Y_i \quad \textit{on} \quad \left(\{1[x_i = a_m]\}_{m=1}^M, D_i\right) \quad \textit{with parameters} \quad (\theta_1, \ldots, \theta_M, \rho). \tag{3}$$

Define the estimator of ATE based on specification (2) (i.e., the fully saturated specification) as:

$$\widehat{ATE} := \sum_{m=1}^{M} \hat{s}_m \hat{\rho}_m. \tag{4}$$

Then:

(a) (6p) The OLS estimator of ρ in (the "shortcut") specification (3) has the form:

$$\hat{\rho} = \sum_{m=1}^{M} \hat{\omega}_m \left(\bar{Y}_{1,m} - \bar{Y}_{0,m}\right), \tag{5}$$

with:

$$\hat{\omega}_m = \frac{\hat{s}_m \widehat{Var}[D_i \mid x_i = a_m]}{\sum_{j=1}^{M} \hat{s}_j \widehat{Var}[D_i \mid x_i = a_j]}. \tag{6}$$

(b) (2p) $\hat{\rho}$ in (5) equals $\widehat{ATE}$ in (2) only when either of two conditions hold:

i. (1p) the shares of treated units is identical across subsamples defined based on the value of x_i;

ii. (1p) the estimates of CATE are identical across subsamples defined based on the value of x_i.

(c) (8p) $\hat{\rho}$ is a consistent estimator of ATE only when treatment effects are homogeneous in x_i or the assignment probabilities do not depend on x_i.

*(d) (6p) If the sample is evenly split across values of x_i, then $\widehat{ATE}$ gives the same weight to each estimate of CATE (1p). If **also** $Var[Y_i(0) \mid x_i = a_m] = Var[Y_i(1) \mid x_i = a_m]$ $\forall m$ (i.e., TEs are homoskedastic within each sub-population) — denote this common value by σ_m^2, then $\hat{\rho}$ gives more weight to the most precisely estimated CATEs (5p).*

[1] Or, when using observational data, you may be willing to assume that CIA and COC hold only conditionally on many OPVs and/or OPVs that take many values.

Q10. (14p) Specification 2 through Specification 4 are examples of the regression adjustment approach. For example, you may consider implementing Specification 2 in light of the imbalance in educational attainment documented in Part 1: standard models of human capital formation and wage determination imply that educational attainment is one of the determinants of human capital hence of labor market productivity, which in turn is one of the determinants of wages hence earnings. Imbalance in OPVs flags the possible presence of observable confounders. Regression adjustment is a parsimonious and parametric approach that attempts to account for the presence of observable confounders.

(a) (3p) Report the estimates of the ATE and test the null hypothesis H_0 that ATE is zero versus the alternative hypothesis H_1 that it is different from zero.

(b) (5p) Are there reasons to add OPVs as regression covariates when they are balanced across treatment and control groups? If so, what are they? Comment on the estimation results from Specification 2 and Specification 3.

(c) (1p) With reference to Specification 3, do you think that it is problematic to use lagged / past values of the dependent variable as regression covariates? Explain.

(d) (5p) Are there reasons to add interactions between OPVs and the treatment indicator? If so, what are they? With reference to Specification 4, test the following two hypotheses, one-at-a-time: i) the ATE is zero; ii) the effect of the offer of training does not vary by the age of the subject.

Q11. (5p) Write a short paragraph describing a plausible mechanism underlying the estimated ATE of being offered training.

Guidance

Before answering the questions that use the NSW data in a regression specification, divide `re74` and `re75` by 1,000; this avoids working with OPVs that have vastly different scales.

Q1. **Note**: In PSet1 you tested equal means for the 10 OPVs present in the NSW experimental data using equation-by-equation OLS, which is called Procedure 1 in the companion document. The companion document helps you understand that doing so is problematic because of the "multiple comparisons" or "multiple testing" problem. The same document presents alternative testing procedures that obviate the problem. Subsequent questions in this part of the PSet have you implement them using the NSW experimental data.

Q2. Here is the guidance:

(a) **Programming guidance**: Use `systemfit::systemfit()`, where `systemfit` is both the name of the R package and of the function. The vignette is here, go directly to Section 4.1 and 4.2. Implementation takes 3 steps: 1) collect the 10 formulas in a list (create each formula with `stats::formula()`); 2) pass the list to `systemfit::systemfit()`, specifying `method = ''SUR''`; 3) summarize the output using `summary( )`. Example: If your list of formulas is named `sur_system`, you type `sur_fit <- systemfit::systemfit(formula = sur_system, data = df, method = ''SUR")` then `summary(sur_fit)`.

(d) **Hint**: Using the expressions given in the companion document, compute the value of the F test statistic and compute its p-value (1p each) as well as the value of the S (aka Wald) test statistic and its p-value (1p each).

(e) **Programming guidance**: Use `car::linearHypothesis()` to compute the value of the F test statistic and its p-value (1p each) as well as the value of the S (aka Wald) test statistic and its p-value (1p each). You know that you are doing this task correctly if these values coincide with those in **Q2d**.

Q3. **Programming guidance**: You may directly leverage `DescTools::HotellingsT2Test()`, that is, you are not expected to manually implement the testing procedure by yourself.

Q4. **Hint:** Use a multi-variate linear regression model (MLRM) with **only** linear terms for the 10 OPVs. Spell out null and alternative hypotheses. Use the F-test for the overall significance of the regression, namely $F = \frac{R^2/K}{(1-R^2)/(n-(K+1))}$ with $K = 10$. **Programming guidance**: First implement the test manually (still in R). Then, confirm your results by leveraging `summary()` after estimation, e.g., `summary(lm_fit)$fstatistic` returns the test's value and degrees of freedom. The p-value for the F-test of overall significance of a regression is not directly stored in the `summary(lm_fit)` object, use `stats::pf()` to compute the test's p-value.

Q5. **Programming guidance**: For both procedures you may directly leverage `stats::p.adjust()`, that is, you are not expected to manually implement the testing procedure by yourself.

Q6. **Programming Guidance:** To test H_0 that ATE is zero in specification 1 you may apply `summary()` on the object returned by `stats::lm()` or extract such information from that object in any other way of your choosing.

Q7. **Hint:** Use the relevant specification given in `CAUS_2_RCM.pdf`.

Q8. **Hint:** Use the relevant specification given in `CAUS_2_RCM.pdf`.

Q9. **Hint**: Leverage the partialling out interpretation of OLS in a multivariate regression model to write the analytical expression of $\hat{\rho}$, go from there. Note: $Var[D_i \mid x_i = a_m] = \Pr(D_i = 1 \mid x_i = a_m)\,(1 - \Pr(D_i = 1 \mid x_i = a_m))$, hence a "natural" estimator of this variance is $\bar{D}_m(1 - \bar{D}_m)$, where $\bar{D}_m$ denotes the sample proportion of treated in subsample with $x_i = a_m$.

Q10. **Note**: These specifications are an even more restrictive version of specification (5), here the OPVs are entered linearly because they take many values, making fully saturated specifications infeasible in practice.

(a) **Hint:** Specification 3 has 11 regression covariates. Specification 4 has 12 regression covariates.

(d) **Note:** If additional assumptions are needed to carry out i) and ii), state them. **Programming Guidance**: Use `car::linearHypothesis()`.

Q11. **Hint:** Think of the possible pathways through which employment *cum* on-the-job training may cause an increase in post-intervention earnings. A well-written paragraph suffices.