CALIFORNIA STATE UNIVERSITY
SAN BERNARDINO
Institutional Research

Case-Control Matching with SPSS:

A Tool to Reduce Selection Bias in Common IR Studies

Allan Taing, Research Technician, ataing@csusb.edu
Tanner Carollo, Assistant Director, tcarollo@csusb.edu

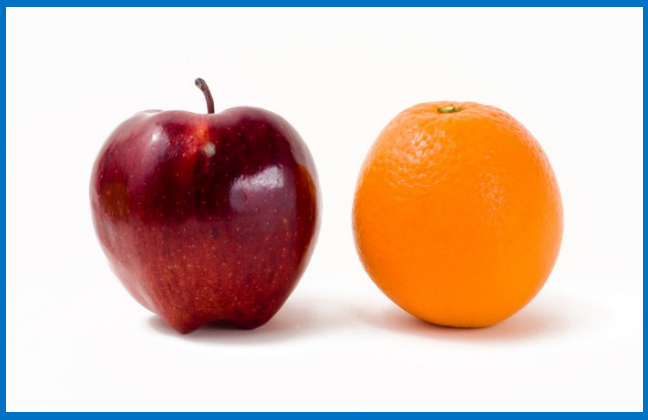
Overview of Presentation

- Research question
- Matching
- SPSS case-control matching
- Example from CSUSB
- Q-and-A

CALIFORNIA STATE UNIVERSITY
SAN BERNARDINO
Institutional Research

Research Question

- How can IR offices assess the impact of student services programs when students are not randomly selected/assigned to participate?



CALIFORNIA STATE UNIVERSITY SAN BERNARDINO Institutional Research

Theory/Rationale for Matching

- Randomized experiments as the "gold standard"
  - Shadish, Cook, & Campell (2002)
- Case-control Matching as a quasi-experimental design
  - Matching on confounding variables to account for pre-existing differences
  - Reducing selection bias
  - Improving internal validity

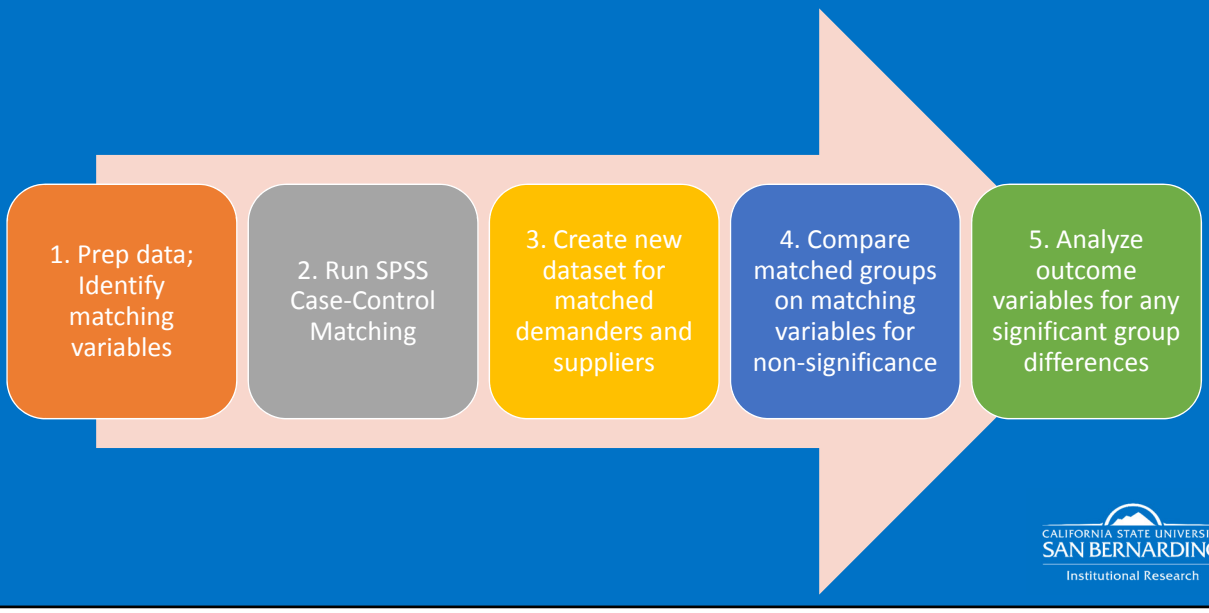
CALIFORNIA STATE UNIVERSITY SAN BERNARDINO Institutional Research

SPSS Case-Control Matching: Overview

- Point-and-Click with v. 22
  - Or via syntax with Python Essentials in older versions (v. 18-21)
- "Fuzzy" Matching on matching variables
  - Researcher-defined tolerance levels/Fuzz Factor
  - Random match from eligible suppliers
- Iterative Process
- One SPSS file:
  - Demanders and Suppliers, coded 1 and 0, respectively
  - Unique ID variable for each case
  - Matching variables and outcome variables

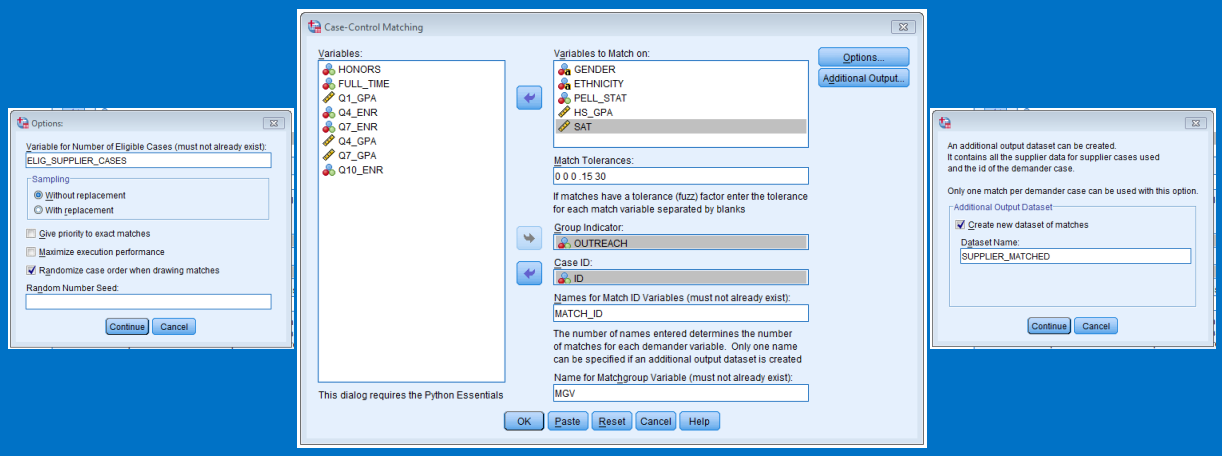
CALIFORNIA STATE UNIVERSITY SAN BERNARDINO Institutional Research

SPSS Case-Control Matching: Step-by-Step



CALIFORNIA STATE UNIVERSITY SAN BERNARDINO Institutional Research

SPSS Case-Control Matching: Demonstration



Example: EOP Matching

Cohorts	Matched Variables	EOP	Non-EOP
2008-2011	EFC	268.70	289.37
	HSGPA	3.00	3.01
	FG Status		
	No College	65%	65%
	College	28%	28%
	Unknown	7%	7%
	Gender		
	Male	31%	31%
	Female	69%	69%
	Ethnicity		
	Asian	4%	4%
	African American	13%	12%
	Hispanic	76%	78%
	White	3%	3%
	Other	4%	3%
	Comparison Variables	EOP	Non-EOP
	SAT	826.38	*865.38
	ACT	16.72	*17.85
	ELM	36.65	*39.70
	EPT	138.58	*140.50

*Sig. differences at the p<.05 level.

CALIFORNIA STATE UNIVERSITY SAN BERNARDINO
Institutional Research

Example: EOP Retention Rates

Table 1. Retention Comparison
EOP vs. Non-Matched Students

Cohort	Group	Count	2nd Year Retention	3rd Year retention	4th Year Retention
	EOP	250	87%*	72%	66%
Fall 2008	All Non-EOP	1718	82%	67%	61%
	EOP	243	89%	79%*	72%
Fall 2009	All Non-EOP	1774	84%	72%	66%
	EOP	249	91%	85%*	-
Fall 2010	All Non-EOP	1524	88%	78%	-
	EOP	243	91%	-	-
Fall 2011	All Non-EOP	1888	87%		
	EOP	985	89%*	78%*	69%*
Total	All Non-EOP	6904	85%	72%	64%

Table 2. Retention Comparison
EOP vs. Matched Students

Cohort	Group	Count	2nd Year Retention	3rd Year retention	4th Year Retention
	EOP	250	87%	72%	66%
Fall 2008	Matched Non-EOP	250	81%	66%	61%
	EOP	243	89%*	79%	72%
Fall 2009	Matched Non-EOP	243	82%	72%	66%
	EOP	249	91%*	85%	-
Fall 2010	Matched Non-EOP	249	87%	78%	-
	EOP	243	91%	-	-
Fall 2011	Matched Non-EOP	243	86%		
	EOP	985	89%*	78%*	69%*
Total	Matched Non-EOP	985	84%	72%	62%

CALIFORNIA STATE UNIVERSITY SAN BERNARDINO
Institutional Research

Example: EOP GPA Comparisons

Table 3. First-Term GPA Comparison
EOP vs. Non-Matched Students

Cohort	Group	First-Term GPA
	EOP	2.74
Fall 2008	All Non-EOP	2.72
	EOP	2.70
Fall 2009	All Non-EOP	2.78
	EOP	2.87
Fall 2010	All Non-EOP	2.91
	EOP	2.89
Fall 2011	All Non-EOP	2.91
	EOP	2.80
Total	All Non-EOP	2.80

Table 4. First-Term GPA Comparison
EOP vs. Matched Students

Cohort	Group	First-Term GPA
	EOP	2.74*
Fall 2008	Matched Non-EOP	2.44
	EOP	2.70
Fall 2009	Matched Non-EOP	2.53
	EOP	2.87
Fall 2010	Matched Non-EOP	2.73
	EOP	2.89*
Fall 2011	Matched Non-EOP	2.48
	EOP	2.80*
Total	Matched Non-EOP	2.55

CALIFORNIA STATE UNIVERSITY SAN BERNARDINO
Institutional Research

Campus Impact

- EOP Director:

"For many years, our student population was being compared with other students that did not have comparable characteristics. We did not feel that the available data accurately provided a true comparison, nor the added value of our program and services provided for the population that we serve. *With the introduction of the Case Control Matching technique, our department is now able to measure and compare students with similar attributes. This allows us to truly assess the significant impact our services and interventions have on the students that participate in our program*."

CALIFORNIA STATE UNIVERSITY SAN BERNARDINO Institutional Research

Conclusion

- Case-control matching is a useful tool to *reduce* selection bias when analyzing the effectiveness of student services programs
- Deciding on matching variables and tolerance levels is *crucial*
- Check the matched groups for similarities before analyzing outcomes
- IR studies can have broad impact for campus stakeholders

CALIFORNIA STATE UNIVERSITY SAN BERNARDINO Institutional Research

Thank You!

- Questions?
- Contact us!
  - institutional_research@csusb.edu
  - 909-537-5052

CALIFORNIA STATE UNIVERSITY
SAN BERNARDINO
Institutional Research

CALIFORNIA STATE UNIVERSITY
SAN BERNARDINO
Institutional Research

CALIFORNIA STATE UNIVERSITY
SAN BERNARDINO
Institutional Research

SPSS Case-Control Matching
2014 CAIR Conference – San Diego
Allan Taing and Tanner Carollo

SPSS Case-Control Matching using point-and-click is available in SPSS 22 or higher. If you have an earlier version, you'll need to run the FUZZY matching syntax by installing Python Essentials.

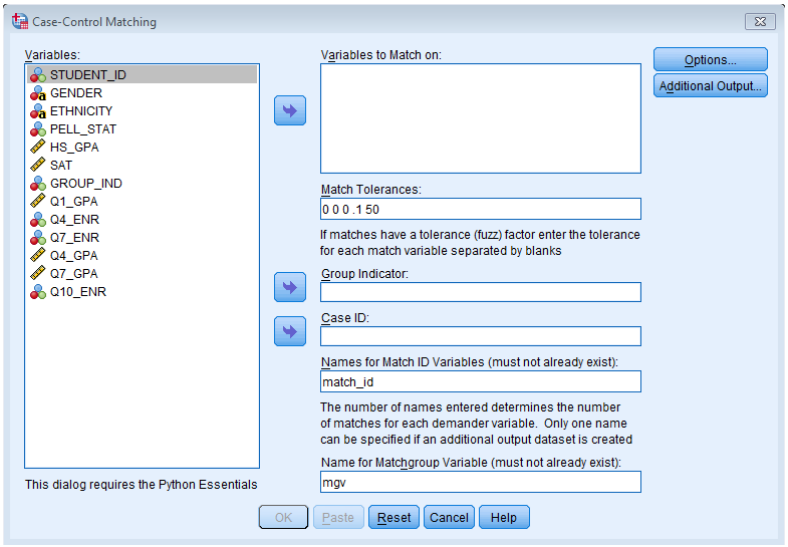
STEP 1: PREP YOUR DATA

Create a single SPSS file with treatment/participant cases (demanders), and control/non-participant cases (suppliers). Each case needs to have a unique ID variable. In addition to the matching and outcome variables, create a binary group indicator variable to distinguish demanders (1) and suppliers (0).

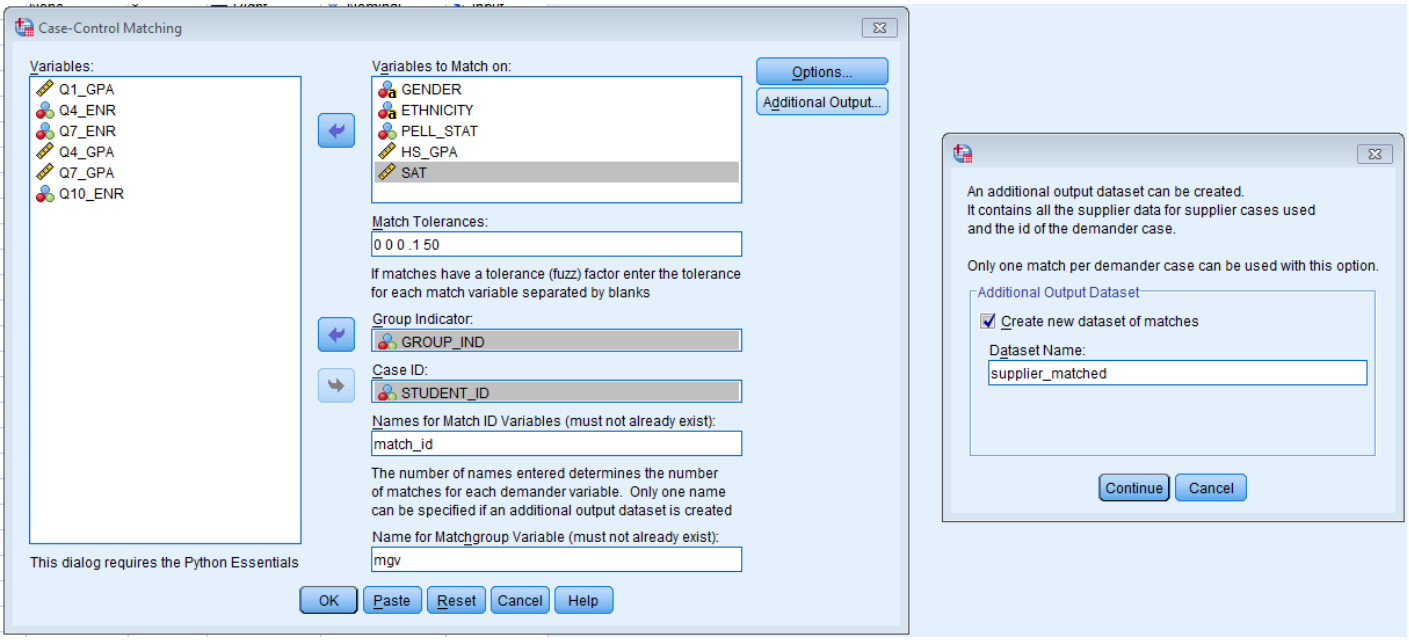
STEP 2: RUN CASE-CONTROL MATCHING

Open the Case-Control Matching dialog box from the Menu Bar [Data > Case-Control Matching]

- Use the arrow button to move the variables you would like to match on from the *Variables* box to the *Variables to Match On* box [i.e., GENDER, ETHNICITY, PELL_STAT, HS_GPA, SAT].
- Set your tolerances/fuzz factor in the *Match Tolerances* box. You'll need to identify a tolerance value for each matching variable in order listed in the *Variables to Match On* box above and separated by a single blank space.
    - String values must have a tolerance value of 0.
    - A tolerance value of 0 for continuous variables would result in exact matches on those variables. In this example, the tolerance level is set to .1 for GPA, and 50 for SAT.
- Move the demander/supplier indicator variable to the *Group Indicator* box [i.e., GROUP_IND].
- Move the unique ID variable to the *Case ID* box [i.e., STUDENT_ID].
- Create a variable name in the *Names for Match ID Variables* box.
    - If a demander case is matched with a supplier case, the supplier case's Case ID will populate this variable in the demander case [i.e., match_id].
- Create a variable name in the *Matchgroup Variable* box.
    - This is used by SPSS when identifying matches; you will not be using this variable [i.e. mgv].



- Next, click on the *Additional Output* button
- Click on the checkbox to *Create new dataset of matches*, and type in a name for the new dataset.
    - SPSS will create a new dataset for the supplier cases that have an identified match to the demander cases (you can name it supplier_matched).
- Click "Continue" in the *Additional Output*, then click "OK" in the Case-Control Matching dialog box to run the program.



Case-control matching is an iterative process; you may have to run this a few times while adjusting your *Match Tolerances*, or fuzz factor, to obtain an acceptable sample size. Once you have obtained an acceptable number of matches, you can move to the next steps. There are no rules for an acceptable number of matches, but the match tolerance ranges should be justifiable on theoretical/practical grounds.

STEP 3: BUILD MATCHED DATASET

Create a dataset with only the matched demander cases and supplier cases.

- In your original dataset, go to the Select Cases dialog box, and under "If condition is satisfied" select cases [IF match_id NE 0], and for output, "Copy the selected cases to a new dataset" (you can name it demander_matched).
- In the new dataset, merge in the supplier_matched dataset [Data > Merge Files > Add Cases]. This merged dataset will now have all of the matched demander cases and matched supplier cases.

STEP 4: CHECK MATCHING VARIABLES

Compare the matched demander cases and matched supplier cases on the matching variables

- For categorical variables, run frequencies to ensure that the distributions are equal
- For continuous variables, run t-tests to ensure that there is non-significance. You want to ensure that the two groups do not significantly differ on these matching variables.
- If the groups significantly differ, go back to Step 2 and Step 3 to set stricter tolerance levels and create a smaller comparison group until continuous matching variables do not significantly differ.

STEP 5: RUN ANALYSES ON OUTCOME VARIABLES

Now you can run significance tests on your outcome variables (tests of proportions for retention rates, test of mean differences for GPA, etc.) to see if the matched treatment/participant group significantly differs from the matched control/non-participant group.