AAE637 Assignment3

Jesmyn ZHANG

March 2024

Section 1: Propensity Score Matching and Inverse Probability Weights

1.3 What's the intuition behind clustering the standard errors at the district level?

Clustering standard errors at the district level accounts for potential similarities or shared effects among schools within the same district, recognizing that observations within a cluster may not be entirely independent. The results in the image indicate that the standard errors are adjusted for clustering by district code ('dcode'), ensuring more reliable inference by acknowledging intra-district correlation.

1.5

What are the mean, standard deviation, min, and max of the propensity score?

Linear regression

Number of obs	1,823
F(3, 563)	153.38
Prob > F	0.0000
R-squared	0.3808
Root MSE	15.715

(Std. err. adjusted for 564 clusters in dcode)

math4	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
lowexpend	-4.273246	1.301817	-3.28	0.001	-6.830257	-1.716235
lunch	-.4614302	.0218218	-21.15	0.000	-.5042922	-.4185682
enroll	-.013035	.0045792	-2.85	0.005	-.0220295	-.0040406
_cons	96.3268	1.766388	54.53	0.000	92.85729	99.79632

Figure 1: OLS regression with specified controls and cluster standard errors

```
. * Summarize the propensity scores to get mean, SD, min, and max
. summarize p_scores
```

Variable	Obs	Mean	Std. dev.	Min	Max
p_scores	1,823	.2495886	.1052205	.0610491	.9003724

Figure 2: Enter Caption

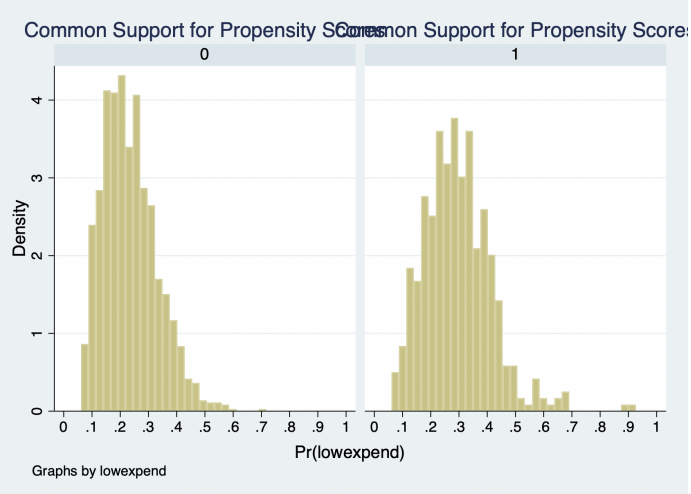

Figure 3: Enter Caption

1.6 Common Support Figure

1.11 How many schools are excluded according to the two alternative common support adjustments?

According to the output :

When using kmatch with common support adjustment (comsup), 3 observations are excluded because their propensity scores lie outside the common support.

When manually implementing common support following Crump et al. (2009), the final number of observations used in the matching is 1,820, down from the original 1,823. This indicates that 3 schools were excluded based on the manually implemented common support criteria.

In both alternative common support adjustments, a total of 3 schools are excluded from the analysis.

1.13 Interpret the coefficient on the OLS model and the final PSM model with the Crump et al. (2009) adjustment.

OLS Model: The coefficient of -4.273 from the OLS model suggests a significant negative impact of being in the low expenditure group on math scores across the whole sample. The OLS model estimates this relationship while controlling for other variables but does not account for the potential biases from unobserved confounders as effectively as the PSM model.

Final PSM Model with Crump et al. Adjustment: The ATT of -1.306 from the PSM model, after applying the Crump et al. (2009) common support adjustment, indicates a negative impact of being in the low expenditure group but is smaller in magnitude compared to the OLS estimate. This difference could be due to the PSM model's focus on comparing more similar schools (i.e., those within the common support) and potentially offering a more accurate estimate of the causal effect of expenditure on math scores for the treated schools.

1.16 Summarize the findings across the six most recent specifications. Be sure to consider. the differences in the estimation methods when comparing the results and standard errors

Propensity Score Matching (PSM)
Model for Propensity Score: Logistic regression was used to estimate the propensity score based on enrollment, indicating the likelihood of being in the low expenditure group.
Outcome: The PSM method's main output is the propensity scores, used to match units from the treated and control groups to estimate the Average Treatment Effect on the Treated (ATT) or Average Treatment Effect (ATE). Inverse Probability Weighting (IPW)

```
. summarize Age Male NeighborInc CityInc AgeInc Income
```

Variable	Obs	Mean	Std. dev.	Min	Max
Age	10,000	39.9146	8.918871	25	55
Male	10,000	.4966	.5000134	0	1
NeighborInc	10,000	10001.21	2504.204	-64.6232	18833.56
CityInc	10,000	5010.499	998.0277	1115.004	8949.624
AgeInc	10,000	.4003324	.1104544	.1257142	.8857564
Income	10,000	92440.65	6141.336	66419.34	114198.1

Figure 4: Enter Caption

Weights Calculation: Weights were derived from the propensity scores to account for the probability of receiving the treatment. This approach aims to create a pseudo-population where the treatment assignment is independent of observed characteristics.
Outcome: The regression of math4 on lowexpend with IPW applied shows a coefficient for lowexpend of -1.079, with a p-value of 0.730, indicating no significant effect of expenditure status on math scores after adjusting for the treatment probability.
Doubly Robust Estimation (DR)
Model for Outcome: Separate regressions for the outcome variable (math4) were conducted for the treated and control groups, including enrollment as a covariate.
Outcome: The DR approach combines the propensity score model with outcome regression models to estimate treatment effects.

Section 2: Constructing Simulated Data

2.3 Provide some intuition for why I use an exponential distribution to distribute population across cities.

Using an exponential distribution to distribute population across cities captures the real-world phenomenon of city size distribution, known as Zipf's Law, which suggests that a few cities are very large while many are much smaller. The exponential distribution, with its characteristic long tail, models this uneven distribution effectively. It allows for few large cities (high population) and many smaller ones (low population), reflecting the actual urban patterns. Additionally, it offers simplicity and flexibility in modeling, making it a practical choice for simulations and studies of urban dynamics.

2.9 Explain the intuition behind constructing income in the way that we did

In constructing the income variable for the simulation, we've aimed to mirror the multifaceted factors that affect real-world earnings. A base income level represents what individuals might expect without other influences, and we've introduced variability to reflect both systematic factors like gender income differentials and local economic conditions—evident in the neighborhood and city effects—as well as individualized factors linked to age, modeled through a Gamma distribution to represent career progression and experience. Lastly, we've included random variation to account for the unpredictable elements of personal circumstances, ensuring the simulation captures the complexity and randomness present in actual income data.

Section 3: Panel Models, Identifying Variation, and Inference

3.2 Before running this model, how should we adjust our standard errors and why?

Before running a fixed-effects model with city and age as categorical variables, we should adjust our standard errors to account for clustering within cities. This is necessary because observations within the same city are likely to be correlated, violating the OLS assumption of independent errors. Clustering standard errors at the city level will correct for this intra-city correlation and provide more accurate statistical inference. This adjustment is particularly important in panel data settings where repeated measurements could introduce serial correlation that standard OLS errors do not account for. Therefore, we use clustered standard errors to ensure our estimated coefficients are robust to such within-group correlations.

3.3 Explain the identifying variation used when estimating the model.

The identifying variation in the model comes from the changes in income within cities and age groups that are not explained by those fixed characteristics. By controlling for city and age fixed effects, the model isolates the impact of time-varying factors like the treatment and gender on income. It leverages the assumption that, within each city and age cohort, the observed differences in income can be attributed to these factors rather than to omitted, unchanging characteristics of the groups. The treatment effect is identified by the difference in income of treated versus untreated neighborhoods within the same city, assuming that treatment assignment is random or unrelated to other income determinants after controlling for city and age effects.

3.4

The observation count in e(sample) is less than the expected 10,000 because the regression model excludes observations that do not meet certain criteria. This can happen if there are missing values for any variables in the model, if the sample was restricted to a specific subgroup or time period, or if all individuals within a fixed effect category, such as a city, share the same characteristics, leaving no within-group variation to exploit. This exclusion ensures that the estimation only uses valid and relevant data, but it is crucial to understand and justify the reasons behind any such reductions in sample size to maintain the integrity of the analysis.

3.5

When we estimate the cross-sectional model with the addition of neighborhood fixed effects, alongside the previously included city and age fixed effects, we're likely to encounter a significant reduction in the variation available to identify the effect of other variables, such as the treatment effect or gender effect. This happens because adding neighborhood fixed effects controls for all variation within neighborhoods, leaving only within-neighborhood variation to identify the effects of interest.

Given that our model already controls for city and age, adding neighborhood fixed effects may lead to a situation where there's little to no variation left within these neighborhoods to exploit for estimating the coefficients of other variables. This is particularly true if neighborhoods are relatively homogeneous within themselves or if many of the neighborhoods contain a small number of observations.

As a result, the model might show a high number of coefficients being omitted due to perfect collinearity, a significant loss in degrees of freedom, and potentially, the standard errors of the estimated coefficients might increase, reflecting the reduced variation available for estimation. This reduction in identifying variation could render the model less effective at isolating the impacts of variables of interest, such as the treatment effect, from the fixed effects.

0.1 3.6

Transitioning to a panel model with the data previously described, the appropriate fixed effects to include would be individual fixed effects and time fixed effects. Individual fixed effects control for all time-invariant characteristics of each individual, capturing unobserved heterogeneity that could influence income levels. Time fixed effects control for any external factors affecting all individuals equally in a given time period, such as economic conditions or policy changes.

For standard errors, the appropriate adjustment in a panel model context, especially when including both individual and time fixed effects, is to cluster the standard errors at the level of the individual. This adjustment accounts for autocorrelation within each individual's observations across time, acknowledging that observations for the same individual over different time periods are likely not independent. Clustering at the individual level ensures that the standard errors are robust to this within-individual correlation, providing more reliable statistical inference.

3.8

The estimation results indicate that the fixed-effects model has not found any variation within the individuals' observations that can be exploited to estimate the coefficients for the treatment variable ('istreatedindividual'), gender variable ('Male'), and the time fixed effect ('year 2023'). All these variables are marked as "omitted" which suggests that within each individual's set of observations, there is no variation in these variables.

This result can occur if, for instance, the treatment variable does not change over time within individuals, meaning everyone has the same treatment status in both time periods. The same logic applies to the gender variable, as one's gender does not change over time, so it cannot be estimated within an individual fixed effects model. For the time variable, it suggests that the model does not detect any change associated with moving from one time period to another once individual fixed effects are accounted for.

The constant term ('cons') is significant and represents the average log income for the base category of the categorical variables (those not included in the model) when all other variables are held at zero.

Additionally, the 'sigmau' (standard deviation of the individual-specific effect) is approximately 0.075, which suggests that there is variation between individuals in the dataset. The 'rho' value is 1, which indicates that all the variance in the model is due to the variation between individuals (ui).

This could suggest that either the treatment was uniformly applied across all observations within each panel unit (making it impossible to identify the effect within the fixed-effects framework), or that the variables of interest are constant within individuals over time and hence cannot provide within-individual variation needed for the fixed-effects estimation.

3.10

The fixed-effects regression, clustered by neighborhood, aimed to analyze the impact on log income across different years within a panel data structure. However, key findings indicate challenges within the model and data:

Year Effects: The model attempted to include year dummies but found year=2023 was omitted due to collinearity, highlighting a lack of within-individual temporal variation that could be captured by the fixed-effects model.

Constant Term: The model estimated a constant term for log income at 11.46632, but standard errors, t-statistics, and confidence intervals could not be computed or were not provided, indicating potential issues with model specification or data variability.

Variance Components: The reported sigmau value was 0.07506938, suggesting some variation across individuals. However, sigmae was reported as 0, and the rho value at 1 indicated that all the variance in log income is attributed to individual-specific effects, suggesting no observed within-group variance.

These results suggest that the fixed-effects model, while controlling for individual heterogeneity and clustering by neighborhood, could not identify additional effects of time or other covariates within the structure of this dataset. The absence of estimated effects for the included time dummies and the peculiarities in the variance components call for a careful reassessment of the model's appropriateness and the data's capability to inform on within-individual changes over time.