University of Oxford: MPhil in Politics

Causal Inference: Problem Set 1

1090063

Contents

1 Problem 1

We will examine how randomized experiments work by creating an imaginary experiment. Use the dataset `a` from the file called `experiment.Rda`. For each individual unit (i) in our sample, the dataset contains the potential outcome under control (Y_i^0 or $Y_0 i$) and the potential outcome under treatment (Y_i^1 or $Y_1 i$) in the columns `a$y0` and `a$y1`, respectively. This is a purely hypothetical scenario. In reality, we never observe potential outcomes under both treatment and control for the same units: we can only observe one of them (the fundamental problem of causal inference). By creating a randomized experiment with this dataset, we'll demonstrate how experiments overcome this fundamental problem.

1.1 Question 1

Find the "true" Average Treatment Effect across all units. **[5 points]**

The true Average Treatment Effect (ATE) is the value obtained if we were able to observe both the potential outcomes (treatment and control) for each individual. The true ATE is `20.995` and is calculated by taking the mean difference between the potential outcomes under treatment ($Y_1 i$) and control ($Y_0 i$) for each unit. However, this is a purely hypothetical scenario as we can never observe both potential outcomes for the same unit in reality.

1.2 Question 2

Next, we'll implement a randomized experiment on this sample of 100 units. We randomly assign half of the units to treatment and half to control by creating a new variable indicating treatment status (D_i).

Conduct a test to assess whether the treatment and control groups have the same average potential outcomes under control. Has randomization succeeded in creating treatment and control groups with equivalent potential outcomes under control? Why? **[5 points]**

To assess whether the randomisation succeeded in creating treatment and control groups with equivalent potential outcomes under control, we compare whether the average potential outcomes under control for the treatment and control groups are statistically different. The mean potential outcomes under control for the treatment group is `78.649` and for the control group is `80.909`. A t-test comparing the average potential outcomes under control for the treatment and control groups gives a p-value of `0.106`.

The null hypothesis is that the average potential outcomes under control are equal between the treatment and control groups. Given that `p(0.106) > 0.05`, we cannot reject this null hypothesis as the treatment and control groups are not statistically different from one another. Therefore, randomisation has succeeded in creating treatment and control groups with equivalent potential outcomes under control.

1.3 Question 3

Estimate the Average Treatment Effect based on your experiment. How similar is it to the "true" Average Treatment Effect? Explain. [**10 points**]

Since we have now randomised the distribution of the treatment and control groups, we can calculate the estimated Average Treatment Effect (ATE) to give the expected difference in outcomes between the treated and the 'comparable control'.

The estimated ATE is `19.695`. This is calculated by taking the mean difference between the potential outcomes under treatment ($Y_1 i$) of the treatment group and the potential outcomes under control ($Y_0 i$) of the control group. The estimated ATE is lower than the true ATE `19.695 < 20.995` by `6.194`% because the treatment and control groups are not perfectly balanced. However, the estimated ATE is still close to the true ATE because the randomisation has created treatment and control groups with equivalent potential outcomes under control, as shown in `1.2`.

1.4 Question 4

Now, let's see how the experimental procedure performs over repeated randomizations.

What is the average estimated ATE across your 10,000 experiments? Does this suggest that your estimator is unbiased? Why? [**10 points**]

The mean estimated ATE across the 10,000 experiments is `20.989`. This suggests that the estimator is unbiased because the average estimated ATE across the repeated randomisations is very close to the true ATE, `20.995`, a difference of `0.007`. The repeated randomisations improve our mean estimate of the ATE significantly from the single randomisation sample done in `1.3`.

1.5 Question 5

Repeat Task (4), calculating the mean difference in potential outcomes under control (a$y0) between the treatment and control groups instead of the ATE. What is the mean difference from your 10,000 experiments? What does this signify? [**10 points**]

The mean difference in potential outcomes under control between the treatment and control groups across the 10,000 experiments is `0.004`. This signifies that the randomisation has succeeded in creating treatment and control groups with equivalent potential outcomes under control. The mean difference in potential outcomes under control is close to zero, indicating that the treatment and control groups are not statistically different from one another. This is consistent with the results from `1.2`.

2 Problem 2

Past research suggests that ballot secrecy influences turnout. A recent field experiment sent emails to a random group of non-voters around the 2014 election in Mississippi, reminding them that their vote was secret.

For this exercise, we wish to establish if the results of this experiment hold when we focus on the female subsample (instead of the full sample).

2.1 Question 1

Confirm that the randomization process was successful by making sure that women in treatment and control groups are similar in all relevant aspects, e.g., age, ethnicity, non-voting habits. Show your results either using a figure or by producing a publishable table. **[10 points]**

To test whether the women in treatment and control groups are similar in all relevant aspects, a balance test across all relevant observable covariates is conducted to see whether the groups were statistically significantly different from one another. Table 1 shows the results of comparing the characteristics of women in the treatment and control groups. The null hypothesis is that the average values of the variables are equal between the treatment and control groups. The p-values are all greater than 0.05 indicating we cannot reject the null hypothesis. This suggests that the randomisation process was successful in creating treatment and control groups with similar characteristics as there are no statistically significant differences between groups. Although differences between observed variables across groups is found, there could still be selection bias from unobserved variables; however, this is unlikely if randomisation was successful.

Table 1: Balance Table of Female Covariates

Variable	Control	Treatment	p-value	Signif.
d_age	40.48 (11.13)	40.38 (11.35)	0.721	-
d_race_blk	0.76 (0.43)	0.76 (0.43)	0.478	-
d_race_hsp	0.01 (0.10)	0.01 (0.10)	0.783	-
d_race_oth	0.01 (0.10)	0.01 (0.10)	0.781	-
never_voted	0.93 (0.25)	0.93 (0.25)	0.492	-
vote_year	147.20 (524.00)	141.68 (514.85)	0.659	-

Note:
* p <0.1, ** p <0.05, *** p <0.01. The table shows the means and standard deviations of the covariates for the treatment and control groups. The p-values are from t-tests comparing the means of the covariates between the treatment and control groups.

2.2 Question 2

Estimate the Average Treatment Effect and test whether the effects you calculated are robust to the inclusion of covariates. Report all results in a single table. Is it necessary to include covariates when calculating the ATE? **[10 points]**

Table 2: Estimated ATE for Mail Treatment (With and Without Covariates)

	ATE without Covariates	ATE with Covariates
Mail Treatment	−0.005	−0.005*
	(0.004)	(0.003)
Age		0.000
		(0.000)
Black		−0.016***
		(0.003)
Hispanic		−0.010
		(0.013)
Other Race		−0.011
		(0.013)
Never Voted		0.361***
		(0.008)
Years Since Last Vote		0.000***
		(0.000)
Num. Obvs	7969	7969
R-squared	0.000	0.563
Adj. R-squared	0.000	0.563

* p <0.1, ** p <0.05, *** p <0.01
Note: Standard errors are in parentheses. Both models are estimated using OLS regressions.

Although randomisation appears to have been successful, we still want to test for the effects of covariates on the estimated Average Treatment Effect (ATE). Firstly, a simple model is estimated for the ATE without covariates. Secondly, to estimate the ATE and test whether effects are robust to the inclusion of covariates, a regression model is compared to ATE without the inclusion of `d_age`, `d_race_blk`, `d_race_hsp`, `d_race_oth`, `never_voted`, and `vote_year`. Results of these two models are shown in Table 2.

The ATE without covariates is `-0.005`. This tells us the difference in probability that someone voted between the treatment and control groups. Therefore, as the `vote_2014` is binary, those who received mail reminding them their vote was a secret saw a decreased probability of voting of `-0.532` percentage points compared to the control group,a very small amount. The t-test for the ATE without covariates gives a p-value of `0.203` meaning the results are also not statistically significant. Whilst ballot secrecy may theoretically influence voting turnout, the effects seen are neither large nor statistically significant.

The second model accounts for the covariates which were shown to be similar across treatment and control groups in `2.1`. The inclusion of covariates does not change the estimated ATE significantly, with the ATE with covariates being `-0.005`. The p-value for the ATE with covariates is `0.096`, showing signs of weak significance, unlike when estimated without covariates. The inclusion of covariates controlling for black voters and those who have never voted before are the most significant. The model with covariates also has a higher R-squared value of `(0.563 > 0.000)` compared to the model without covariates. This suggests

that the inclusion of covariates improves the model's explanatory power, but whilst the estimated ATE's are similar and neither is significantly robust, the inclusion of covariates is not necessary when calculating the ATE in this case, but should be included to ensure a more robust model.

3 Problem 3

A key problem incumbents encounter in civil wars is a lack of information to combat insurgency. Insurgents exploit information asymmetries at the local level to hide and become a difficult target for incumbents. In the absence of such information, incumbents often resort to indiscriminate violence, via large-scale reprisals against entire villages suspected to host insurgents. One example of indiscriminate violence is Aerial bombardment. Due to the nature of insurgency, bombing frequently occurs in and around settled areas, and leading to many civilian casualties.

Using data from the Vietnam War, Kocher, Pepinsky and Kalyvas (2011) examined the effect of bombings on Viet Cong support. In particular, they looked at the impact of the September 1969 bombings on hamlet control in December 1969.

The data comes from various sources. The United States compiled a gazetteer of South Vietnamese hamlets, identified their geographic coordinates, and conducted a census. District Senior Advisors (DSAs, army officers ranking major or above) were assigned to complete detailed questionnaires for every village and hamlet in their zones of operation. Some of these questionnaires were compiled monthly, others quarterly. DSAs were detached from U.S. units to live and work in the districts they rated. The RVN (Republic of Viet Nam) had 261 districts with a median area of 377 kilometers squared, or about one-fourth the size of the median U.S. county. There was a median of 36 hamlets per district in 1969. In this exercise, we will explore how matching models work, using the dataset in `Vietnam_matching.dta`. We will apply the matching procedure to estimate the effect of experiencing a bombing in September 1969 on insurgency control in December 1969.

3.1 Question 1

Which covariates should you use in the matching procedure and why? [**10 points**]

When looking at the effects of bombings on hamlet control, we want to estimate the average causal (treatment) effect of the bombing in September 1969 on the control of hamlets by the Viet Cong in December 1969. To estimate this average causal effect, we should compare the hamlets that were bombed in September 1969 to those that were not; however, as the bombings were not randomised, we need to identify suitable control hamlets where covariates are similar to those that were bombed. We can do this through a matching procedure to find hamlets which are as similar to one another as possible, other than whether they were bombed or not. Before selecting a matching technique to identify hamlets, a suitable set of covariates should be selected for use in the matching procedure.

The covariates selected to be included in the matching analysis are those which are heterogeneous across the treated and control hamlets. These covariates should be observable and relevant to the treatment assignment. The covariates selected for this analysis are `std`, `score`, `lnhpop`, `mod2a_1ajul`, and `mod2a_1admn`. These

covariates are selected as a balance test shows them to be statistically significantly different between the treatment and control groups, shown by a p-value `<0.05`.

Table 3: Balance Table of Matching Covariates

Covariate	Control	Treatment	p-value
std	5.28 (11.89)	6.52 (13.29)	0.008
ln_dist	11.07 (0.87)	11.06 (0.75)	0.734
score	-0.08 (0.63)	-0.34 (0.55)	<0.001
lnhpop	6.77 (1.01)	6.57 (0.96)	<0.001
mod2a_1ajul	2.75 (1.04)	3.58 (1.07)	<0.001
mod2a_1admn	2.80 (0.83)	3.29 (0.66)	<0.001

Note:
The table shows the means and standard deviations of the covariates for the treatment and control groups. The p-values are from t-tests comparing the means of the covariates between the treatment and control groups.

3.2 Question 2

Choose a matching estimator – briefly describe your choice. Assess balance in pre-treatment covariates between treated and control units, before and after matching. What is the measure you used for assessing balance? [**10 points**]

Before estimating the causal effect of bombings on hamlet control, we need to choose a matching estimator to identify suitable control hamlets. The matching estimator chosen should be the one which ensures that the control hamlets are as similar as possible to the treated hamlets, other than whether they were bombed or not. The matching estimator should also ensure that the control hamlets are balanced across all covariates used in the matching procedure so that we have a balanced control between treated and control groups. There are many matching estimators to choose from, for example: exact matching, nearest neighbour matching, propensity score matching, genetic matching, normalised Euclidean distance matching, and Mahalanobis distance matching. With these possible methods, we want to ensure that we keep as many observations as possible, and assess the bias-variance trade-off to ensure accurate and consistent results. The measure for assessing balance is the standardised mean difference, which is a measure of balance between the treated and control groups, with a value of 0 indicating perfect balance.

The most strict matching estimator of exact matches returns `No exact matches found`, an expected outcome due to the nature of continuous covariates in the dataset. Instead, a number of alternative estimators are considered to determine the best matching estimator for the dataset. Table 3 highlights different matching models which are considered. Models have been considered with and without replacement of observations. The intuition behind comparing these models is the expectation that replacement will help to find better matches, especially given the continuous nature of the covariates; however, this may come with the trade-off of higher variance. The table shows the number of matched observations, the Mean Absolute Standardised Difference (MASD), the Median Standardised Mean Difference (MSMD), and the Maximum Standardised

Mean Difference (MaxSMD) for each matching estimator. Lower values of MASD, MSMD, and MaxSMD indicate better balance between the treated and control groups.

Based on these results, the Mahalanobis distance matching estimator with replacement has the lowest MASD value of `0.011`. Given this result, an additional model where the ratio of control to treated units is 2 is also considered to try and increase the number of matched observations. This model has a slightly higher MASD value of `0.016`, but has a much higher number of matched observations of `1763.000` compared to the next best model matching `1428.000`. The Mahalanobis distance matching estimator with replacement and a many-to-one matching approach is therefore chosen as the best matching estimator for the dataset as it best optimises the balance between treated and control groups and maximises the number of matched observations, helping to reduce bias and increase the precision of the estimated causal effect.

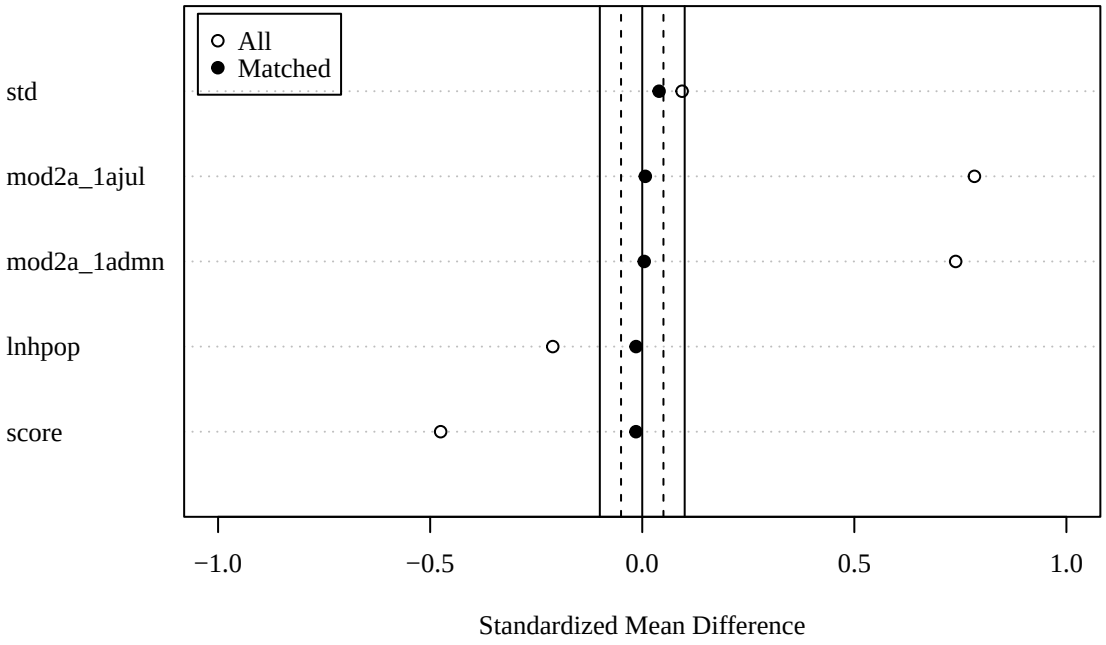
This model's balance is then asssessed across covariates using the Mean Standardised Difference before and after matching, shown in Figure 1. We see from this figure that all post-matching covariates fall within the accepted range of balance of `<0.1` which implies that as a result of this matching process, the groups are comparable on average, with respect to these observable factors.

Table 4: Comparison of Matching Estimators

Model	Matched	MASD	MSMD	MaxSMD
Nearest Neighbour (No Replacement)	1428	0.0268	0.0158	0.0596
Nearest Neighbour (Replacement)	1318	0.0323	0.0278	0.0798
Mahalanobis (No Replacement)	1428	0.0270	0.0293	0.0391
Mahalanobis (Replacement)	1289	0.0108	0.0061	0.0297
Mahalanobis (Replacement, Ratio = 2)	1763	0.0162	0.0146	0.0397
Propensity Score Matching (No Replacement)	1428	0.0268	0.0158	0.0596
Propensity Score Matching (Replacement)	1318	0.0323	0.0278	0.0798
Euclidean (No Replacement)	1428	0.0279	0.0236	0.0551
Euclidean (Replacement)	1300	0.0213	0.0197	0.0446

Note:
MASD = Mean Absolute Standardized Difference, MSMD = Median Standardized Mean Difference, MaxSMD = Maximum Standardized Mean Difference. Genetic matching has been discarded due to higher imbalance and heavier computational load. The binary variable "bombed_969_bin" is the treatment variable used to balance the dataset.

Figure 1: Standardised Mean Differences Before and After Matching



3.3 Question 3

Estimate the causal effect of interest with matching. Are the results from your matching analysis different from using a simple OLS regression without covariates? If there is a difference: how do you explain it? [**10 points**]

When using matching techniques to combat issues of non-randomisation in the receipt of a treatment, we estimate the Average Treatment Effect on the Treated (ATT) to determine the causal effect of the treatment on the outcome. After using the binary treatment variable `bombed_969_bin` to balance the dataset, we can estimate the causal effect of bombings on hamlet control in December 1969 (`mod2a_1adec`), using the continuous variable `bombed_969`, using the covariates `std`, `ln_dist`, `score`, `lnhpop`, and `mod2a_1admn`.

Two ATT models are estimated. The first model (`ATT (OLS)`) uses Ordinary Least Squares; however, this requires a strong assumption that the scaling of the outcome variable is linear and continuous. Instead, the 5-point scale of `mod2a_1adec` is ordinal in nature such that we cannot assume the differences between categories is equal. Therefore, the second model (`ATT (Ordered Logistic)`) uses an ordered logistic model to estimate the causal effect of bombings on hamlet control in December 1969. Finally, a third model (`OLS (No Covariates)`) is a simple OLS regression without covariates. The results are presented in Table 4.

The OLS coefficient shows us that for a one-unit increase in bombings per hamlet in September 1969, the average control of hamlets by the Viet Cong in December 1969 increases on the 1-5 scale by **0.093**. This compares to the even smaller effect of **0.017** in the ATT model using OLS. The ordered logistic model shows that the odds of a hamlet being controlled by the Viet Cong in December 1969 is **1.055** times higher if it was bombed in September 1969.

Table 5: Comparison of Causal Effect Estimates from Matching and OLS Regression

	ATT (OLS)	ATT (Ordered Logistic)	OLS (No Covariates)
(Intercept)	1.196***		2.679***
	(0.163)		(0.011)
bombed_969	0.017***	0.054***	0.093***
	(0.004)	(0.013)	(0.005)
std	0.003*	0.009*	
	(0.001)	(0.004)	
score	−0.008	−0.078	
	(0.036)	(0.101)	
lnhpop	−0.074***	−0.252***	
	(0.020)	(0.056)	
mod2a_1ajul	0.608***	1.708***	
	(0.022)	(0.075)	
mod2a_1admn	0.130***	0.414***	
	(0.034)	(0.095)	
1|2		−0.008	
		(0.480)	
2|3		3.740***	
		(0.458)	
3|4		6.347***	
		(0.476)	
4|5		8.736***	
		(0.498)	
Num.Obs.	1763	1763	8941
R2	0.525		0.032
F	322.945		294.255
RMSE	0.68	3.24	1.00

* p <0.05, ** p <0.01, *** p <0.001
Note: Standard errors are in parentheses. Both ATT models use matched data to estimate the causal effect of “bombed_969“ on “mod2a_1adec“.

Whilst all models suggest a statistically significant effect of bombing increasing insurgent control of hamlets, the OLS (No Covariates) model is likely to be biased due to the non-randomised nature of the treatment and the lack of control for confounders. The ATT models are more robust as they use matching to balance the dataset across covariates, reducing bias and increasing the precision of the estimated causal effect. The ordered logistic model is likely to be the most accurate as it accounts for the ordinal nature of the outcome variable. Therefore, the difference seen between OLS (No Covariates) and the ATT (Ordered Logistic) model is due to the lack of control for confounders in the OLS model, leading to a downwardly biased estimate of the causal effect of bombings on hamlet control in December 1969, as well as the assumption of linearity in the OLS model which is not appropriate for the ordinal outcome variable.

3.4 Question 4

How does the comparison between matching and OLS regression change when you include covariates (control variables)? Report the models in a single table and discuss. [**10 points**]

When including covariates in the OLS regression model, the estimated causal effect of bombings on hamlet control in December 1969 is reduced from **0.093** to **0.025**. This brings the coefficient estimate a lot closer to the **0.017** estimate from the ATT (OLS) model. This is due to now accounting for the confounding variables in the OLS model. The ATT models are still more robust as they use matching to balance the dataset across covariates, reducing bias and increasing the precision of the estimated causal effect. Although there are far fewer observations in the ATT models, both the ATT and OLS models are strongly statistically significant.

However, the OLS regression model may still not give meaningful causal estimates. Whilst the inclusion of covariates to try to satisfy the Conditional Indpendence Assumption (CIA), whereby otential outcomes for control units are the same as for treated units, when those units have the same covariate values (X_i), we cannot guarantee $\beta_{OLS} = \tau_{ATE}$ as we do not know which covariates satsify the CIA. We include all the observed covariates in the OLS model, but there may still be unobserved covariates and heterogeneity which is not accoutned for in the model, leading to omitted variable bias. The ATT models therefore remain more robust as they use matching to balance the dataset across covariates, reducing bias and increasing the precision of the estimated causal effect.

Table 6: Comparison of Causal Effect Estimates from Matching and OLS Regression

	ATT (OLS)	ATT (Ordered Logistic)	OLS (No Covariates)	OLS (With Covariates)
(Intercept)	1.196***		2.679***	0.670***
	(0.163)		(0.011)	(0.063)
bombed_969	0.017***	0.054***	0.093***	0.025***
	(0.004)	(0.013)	(0.005)	(0.004)
std	0.003*	0.009*		0.002***
	(0.001)	(0.004)		(0.001)
score	−0.008	−0.078		−0.098***
	(0.036)	(0.101)		(0.014)
lnhpop	−0.074***	−0.252***		−0.029***
	(0.020)	(0.056)		(0.008)
mod2a_1ajul	0.608***	1.708***		0.507***
	(0.022)	(0.075)		(0.011)
mod2a_1admn	0.130***	0.414***		0.276***
	(0.034)	(0.095)		(0.013)
1|2		−0.008		
		(0.480)		
2|3		3.740***		
		(0.458)		
3|4		6.347***		
		(0.476)		
4|5		8.736***		
		(0.498)		
Num.Obs.	1763	1763	8941	8941
R2	0.525		0.032	0.563
F	322.945		294.255	1917.862
RMSE	0.68	3.24	1.00	0.67

* p <0.05, ** p <0.01, *** p <0.001
Note: Standard errors are in parentheses. Both ATT models use matched data to estimate the causal effect of “bombed_969“ on “mod2a_1adec“.