University of Oxford: MPhil in Politics

Causal Inference: Problem Set 2

1090063

Contents

1 Problem 1: Institutions and Economic Development [50 points]

The assignment is based on the famous Acemoglu, Johnson & Robinson (AJR) 2001 study on the importance of inclusive institutions for economic development. AJR argue that institutions leave a long imprint on countries' economic activity. They distinguish between inclusive and extractive institutions. The former diffuses economic returns across different strata of society, whereas the latter facilitates the appropriation of wealth by elites.

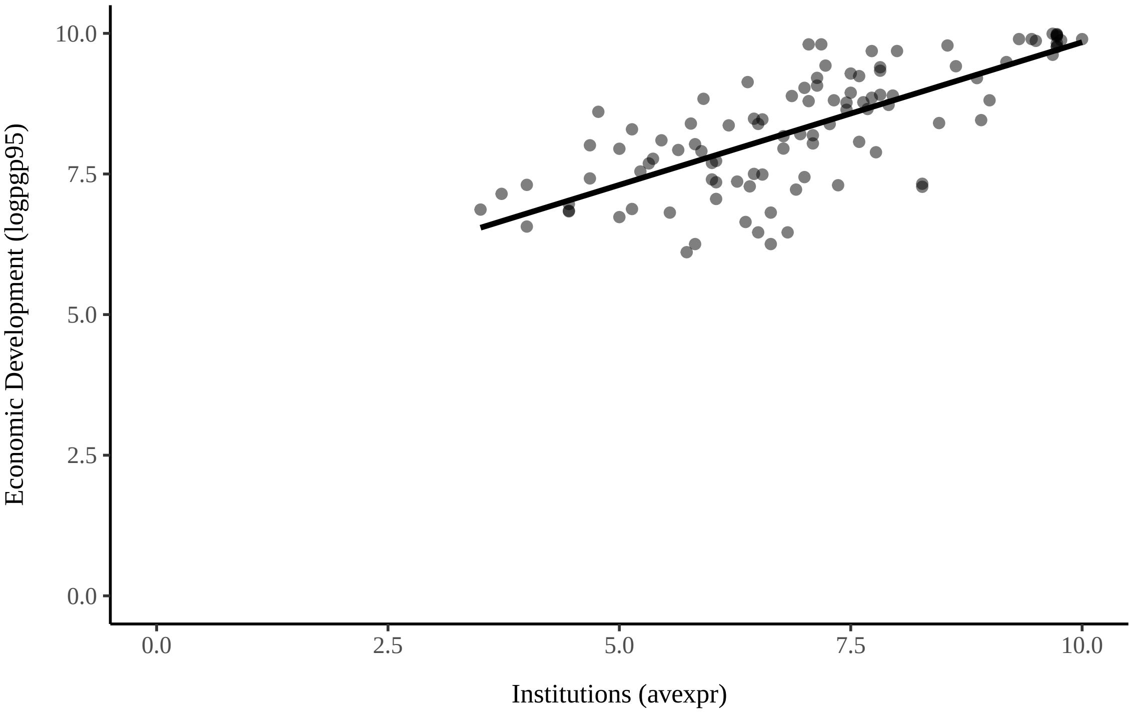
To provide evidence for the importance of institutions, they turn to the colonial structures of the 19th and early 20th centuries. Their identification strategy comes from variation in geography and climate, which determined whether colonizers would establish inclusive or extractive institutions. In those areas in which settlers encountered high mortality rates, they built extractive institutions without long-term planning. In areas with low mortality rates, they built inclusive institutions. Assuming that settler mortality rates satisfy the assumptions of an instrumental variable, this allows the authors to identify the causal effect of institutions on economic development over the long term.

1.1 Correlation

Is there a link between institutions and economic development? This is not a causal question; we are asking if there is any association between the two. Provide a scatterplot to show this is the case.

By using the AJR dataset, we can see that there is a correlation between institutions and economic development of `0.782`. The scatterplot in `Figure 1` shows the relationship between the average protection against expropriation risk between 1985-1995 `(avexpr)` and logged GDP per capita measured in 1995 `(logpgp95)`. This positive, and relatively high correlation factor shows that we would expect there to be a strong, and possibly causal relationship, between increased protection against expropriation risk (strong institutions) and economic development.

Figure 1: Correlation Scatterplot: Institutions and Economic Development



Notes: The correlation factor is calculated with the ‘pairwise complete.obs‘ argument to handle missing values.

1.2 Causal relationship

Is this relationship causal? How do mortality rates help in answering this question?

We are interested in the relationship between institutions and economic development. This can be modelled by the regression specification shown in (1).

$$\texttt{logpgp95}_i = \beta_0 + \beta_1 \texttt{avexpr}_i + \varepsilon_i \tag{1}$$

where:

- $\texttt{logpgp95}_i$ is the log GDP per capita in 1995 for country i,
- $\texttt{avexpr}_i$ is the average protection against expropriation risk between 1985–1995 for country i,
- β_0 is the intercept,
- β_1 captures the effect of institutions on economic development,
- ε_i is the error term.

For the relationship shown in `1.1` to be causal, a number of conditions must hold. Most importantly, the independent variable $\texttt{avexpr}_i$ should not be correlated with the error term ε_i. $\texttt{avexpr}_i$ should therefore be exogenous and isolated from any unobserved confounding variables to ensure conditional independence. However, this is unlikely to be the case. For example, cultural norms of trust and co-operation can influence the quality and strength of institutions, as may colonial and legal legacies. Moreover, the dataset does not include possible confounders such as education which should be included and controlled for in the model as education can be a determinant of both institutions and economic development. Consequently, ε_i is likely to be correlated with $\texttt{avexpr}_i$ and the model is likely to suffer from omitted variable bias, as well as from the reverse causality of richer countries affording to build better institutions. This means that the estimated coefficient $\hat{\beta_1}$ will be biased and inconsistent if an OLS regression were used, and therefore the relationship shown in `1.1` is not causal.:

$$\mathbb{E}[\hat{\beta}_{OLS}] \neq \beta_{\text{true}} \tag{2}$$

To address this issue, AJR use settler mortality rates as an instrument for institutions. The idea is that settler mortality rates are correlated with the quality of institutions, but not with the error term ε_i. This means that settler mortality rates can be used to isolate the effect of institutions on economic development, ensuring the exogeneity of the independent variable. The authors argue that settler mortality rates are a valid instrument because they are determined by geographical and climatic factors, which are not correlated with the error term. This means that settler mortality rates can be used to identify the causal effect of institutions on economic development, with the expectation being that high mortality rates are correlated with extractive institutions, and lower mortality rates result in inclusive institutions.

1.3 ITT Estimation

Estimate the ITT and interpret it.

The ITT estimates the causal effect of the treatment assignment of our instrument, **`logem4`** on the outcome of logged GDP per capita, **`logpgp95`**. When doing an instrumental variable causal analysis, we initially assume that the instrument is randomly assigned to those who are treated. However, regressing the instrument **`logem4`** on the relevant geographical covariates, we find that there is a statistically significant relationship between counties in **`africa`** and based on a country's latitude **`(lat_abst)`** and the instrument. This suggests that the instrument is not randomly assigned to those who are treated. We should therefore control for these covariates to ensures that the randomisation assumptions is more likely to hold.

Taking these significant covariates into account, the ITT estimate is **`-0.337`** and is statistically significant. As we have estimated a log-log model, this means that a 1% increase in settler mortality is associated with a **`-0.337`**% decrease in GDP per capita. Once we scale the ITT estimate, we find that the ITT estimate is **`-0.305`**. This means that a one standard deviation increase in settler mortality rates is associated with a **`-0.305`** standard deviation decrease in log GDP per capita in 1995 which is a moderate effect size. These two ITT estimates both suggest that higher settler mortality rates are associated with lower economic development, which is consistent with the idea that high mortality rates lead to extractive institutions and lower economic development.

Table 1: Randomisation Check of Instrument

	Additional Covariates
Countries in Asia	0.137
	(0.303)
Countries in Neo-Europe	−0.832
	(0.568)
Country Latitude	−3.050***
	(0.883)
Countries in Africa	1.169***
	(0.248)
Num. Obvs	87
R-squared	0.461
Adj. R-squared	0.435

* p <0.1, ** p <0.05, *** p <0.01
Note: Standard errors are in parentheses.

Table 2: ITT Estimates

	ITT Model inc. Covariates
Log of Settler Mortality	−0.337***
	(0.080)
Countries in Asia	1.573**
	(0.651)
Log Output per Worker	−0.639***
	(0.191)
Num. Obvs	81
R-squared	0.588
Adj. R-squared	0.572

* p <0.1, ** p <0.05, *** p <0.01
Note: Standard errors are in parentheses.

1.4 LATE Estimation

Estimate the LATE, using both a Wald estimator and a 2SLS estimator. Interpret your findings.

The Local Average Treatment Effect (LATE) estimates the causal effect of the treatment on the treated, which in this case is the effect of how extractive institutions are on economic development. The LATE is estimated using two methods: the Wald estimator and the 2SLS estimator, with both methods accounting for the covariates of `lat_abst` and `africa`.

The Wald estimator is calculated as the ratio of the previously calculated ITT `(ITT_Y)` and the ITT of the treatment calculated as the first stage least squares regression of the treatment on the instrument `(ITT_D)`. The Wald estimator suggests that a 1% increase in average protection against expropriation risk `avexpr` is associated with a `0.956`% increase in GDP per capita. This is a statistically significant result, but the standard errors are calculated using a delta method which can be more imprecise than the in-built standard errors in the 2SLS estimator.

On the other hand, when calculating the LATE using the 2SLS estimator, we calculate this using the `ivreg` function which correctly accounts for bias when calculating the standard errors. Using this method, we find that a 1% increase in average protection against expropriation risk `avexpr` is associated with a `0.880`% increase in GDP per capita. This is a slightly smaller effect size than the Wald estimator, but it has a greater and more precise statistical significance. The LATE estimates suggest that institutions have a strong positive effect on economic development, such that higher average protection against expropriation risk is associated with higher GDP per capita. This is consistent with the idea that inclusive institutions lead to better economic outcomes.

Table 3: LATE Estimates

Method	Estimate	Std. Error	P-value	Significance
Wald Estimator	0.956	0.477	0.045	**
2SLS Estimator	0.880	0.295	0.004	***

Note:
* p <0.1, ** p <0.05, *** p <0.01. Standard errors are in parentheses.

1.5 IV Assumptions

Assess the plausibility of the IV assumptions in this setting. For each of the assumptions relevant first stage, monotonocity, independence, and eclusion restriction below, explain (in words) what it means substantively in the context of this study and provide a statistical test or verbal argument assessing its plausibility.

Relevant First Stage

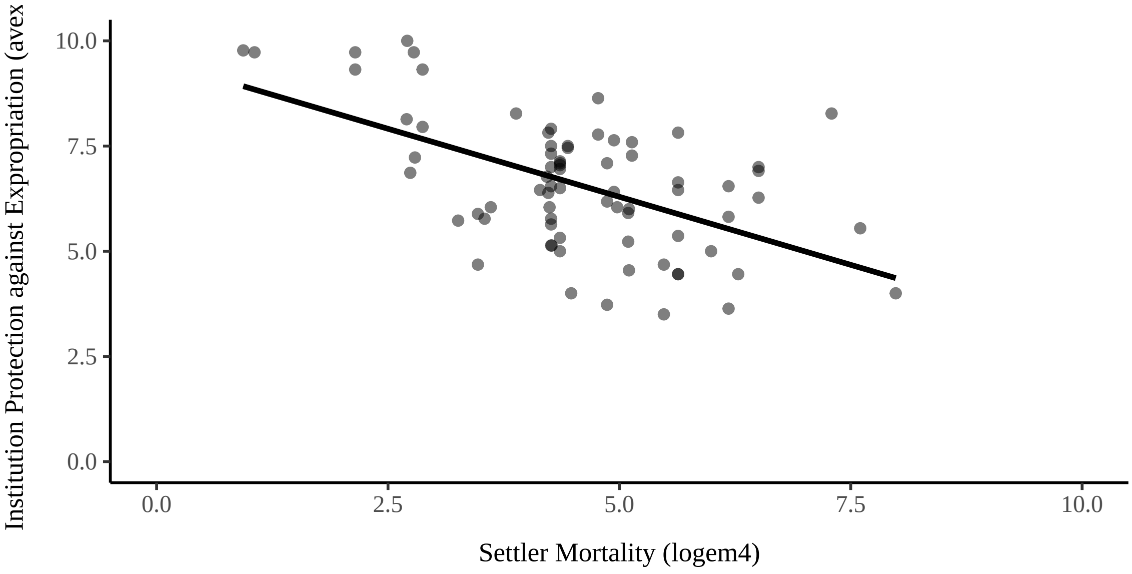
The relevant first stage assumption states that the instrument must be correlated with the treatment. In this case, settler mortality rates must be correlated with the average protection against expropriation risk `(avexpr)`. This is a necessary condition for the instrument to be valid. We have shown this when estimating the ITT `(ITT_D)` which gave a statistically significant result of `-0.353`. This suggests that settler mortality rates are correlated with the average protection against expropriation risk, and therefore the relevant first stage assumption holds.

Monotonicity

The monotonicity assumpton in IV estimation means that the instrument moves the treatment in the same direction for all individuals, meaning there are no "defiers" who would do the opposite of what the instrument encourages. In the context of this study, we would therefore expect that for an increase in settler mortaility rates the average protection against expropriation risk would decrease for all countries. An increase in settler mortality should never lead to an increase in institutional quality for any country.

We can show this visually below in `Figure 2`. The scatterplot shows the relationship between settler mortality rates and average protection against expropriation risk. The fitted line shows a negative relationship, suggesting that higher settler mortality rates are associated with lower average protection against expropriation risk. This is consistent with the idea that higher settler mortality rates lead to extractive institutions, and therefore the monotonicity assumption holds.

Figure 2: Monotonicity Scatterplot: Settler Mortality and Institutions



Independence

The conditional independence assumption states that the instrument must be independent of the error term in the outcome equation, such that the instrument is effectively randomly assigned. There should be no omitted variables or unobserved confounders that affect both the instrument and the outcome.

$$\text{cov}(z_i, \varepsilon_i) = 0 \tag{3}$$

In this case, settler mortality rates must be independent of the error term in the regression of average protection against expropriation risk on GDP per capita. This means that settler mortality rates should not be correlated with any unobserved factors that affect economic development.

This assumption is difficult to test directly due to the unobserved nature of the error term. We have deliberately controlled for potential observerd confounders, and mortaility rates are determined by geographical and climatic factors, which are not correlated with the error term in the outcome equation. This suggests that settler mortality rates are independent of the error term, and therefore the independence assumption holds.

However, it is important to note that this assumption is not always easy to justify. For example, if there are unobserved factors that affect both settler mortality rates and economic development, such as cultural norms or historical legacies, then the independence assumption may not hold. This could lead to biased estimates of the causal effect of institutions on economic development.

Exclusion Restriction

The exclusion restriction states that the instrument must only affect the outcome through the treatment. In this case, settler mortality rates must only affect GDP per capita through their effect on average protection against expropriation risk **(avexpr)**. This means that settler mortality rates should not have a direct effect on GDP per capita, and any correlation between settler mortality rates and GDP per capita should be entirely mediated by average protection against expropriation risk.

However, as the instrument-outcome relationship is not directly observable, this assumption cannot be directly tested. The argument put forward by AJR is that conditional on the controls included in the regression, the mortality rates of settlers have no effect on GDP per capita, other than their effect through institutional development. They argue that settler mortality rates are not the result of local disease environments, which could in turn affect the economic performance of the country. Instead, settler deaths came from a lack of immunity, and therefore local people were not affected by the same diseases, so these diseases would not be the reason the countries were poor. This suggests that settler mortality rates are not directly related to GDP per capita, and therefore the exclusion restriction holds.

2 Problem 2: Centralisation and Public Investment [50 points]

This question uses data from Malesky, Nguyen, and Tran's (2014) study of the effect of centralising control of public services in Vietnam. Scholars have long speculated that de-centralised control of public services in developing countries can lead to corruption and 'capture' by local politicians who fail to use tax revenue to invest in public services. Centralisation of control may, therefore, lead to increased public investment. This is difficult to test because control of government spending rarely changes hands. But in 2009, Vietnam decided to experiment with re-centralising control of its public services in some areas but not others, placing them under central government control instead of control by local district authorities in around one-fifth of districts nationwide. The authors use a difference-indifferences framework to ask what impact this had on infrastructure spending, comparing changes in treated districts (where spending was re-centralised) to untreated districts (where it remained under local control). They have data on infrastructure spending before the change (in 2006 and 2008) and after the change (in 2010), for small areas called 'communes', which are subsets of districts.

2.1 Dataset Variables

Create a new dataset for the years 2008 and 2010 only. In this dataset, create: (i) a dummy variable called "post" equalling 1 if the year is after the treatment and (ii) a variable for the interaction between post and treatment.

A new dataset has been created with the years 2008 and 2010 only. The variable `post` equals 1 if the year is after the treatment, and the variable `treat_post` is the interaction between post and treatment, given by `treat_post = post * treatment`. This means that `treat_post` equals 1 if the year is 2010 and the commune is in a treated district, and 0 otherwise.

2.2 Difference-in-Differences

Use the dataset and variables you created in a) to calculate a difference-in-difference estimate of the causal effect of centralisation on infrastructure investment. Interpret the resulting coefficient and its statistical significance.

The difference-in-differences estimate of the causal effect of centralisation on infrastructure investment is `0.247`. This means that the average infrastructure spending in treated districts increased by `0.247` units in 2010 compared to untreated districts. This is a statistically significant result, with a p-value of `0.010`. This would suggest that re-centralising control of public services in Vietnam led to an increase in infrastructure spending.

However, this first model does not account for the possible unobserved heterogeneity of time-invarient differences between districts which may affect infrastructure spending, as well as the time-based differences across years, such as economic trends. Therefore, a fixed effects model, or first differences model, is required to control for these unobserved factors. The fixed effects model uses `factor(district)` and `factor(year)` to

Table 4: Difference-in-Differences Estimates

	Difference-in-Differences	Difference-in-Differences with FE
Post Treatment	0.219***	
	(0.050)	
Treatment Group	−0.240***	
	(0.075)	
Treatment Post Interaction	0.247*	−0.318
	(0.129)	(0.258)
Log Population Density	0.159***	0.098***
	(0.020)	(0.018)
Num. Obvs	4415	4415
R-squared	0.051	0.303
Adj. R-squared	0.050	0.188

* p <0.1, ** p <0.05, *** p <0.01
Note: Clustered standard errors are in parentheses. Both models cluster standard errors by district.

control for these unobserved factors. When accounting for this unobserved heterogeneity though, we find a statistically insignficant result of `-0.318` which is -ve and thus at odds with the non-fixed effects model and the hypothesis that re-centralisation of control of public services leads to increased infrastructure spending. This insignificant result may be due to a genuine lack of effect, or the high-within district noise such that detecting a treatment effect is harder.

2.3 Parallel Trends

Difference-in-differences estimation relies on the ‘parallel trends’ assumption. Explain what that assumption means in this study.

The parallel trends assumption states that in the absence of the treatment, the average outcome for the treated and untreated groups would have followed the same trend over time. In this study, this means that if the re-centralisation of control of public services had not occurred, the average infrastructure spending in treated districts would have followed the same trend as in untreated districts. This is shown by:

$$E[Y_i^0(1) \mid D_i = 1] \approx E[Y_i^0(0) \mid D_i = 1] + (E[Y_i^0(1) \mid D_i = 0] - E[Y_i^0(0) \mid D_i = 0]) \tag{4}$$

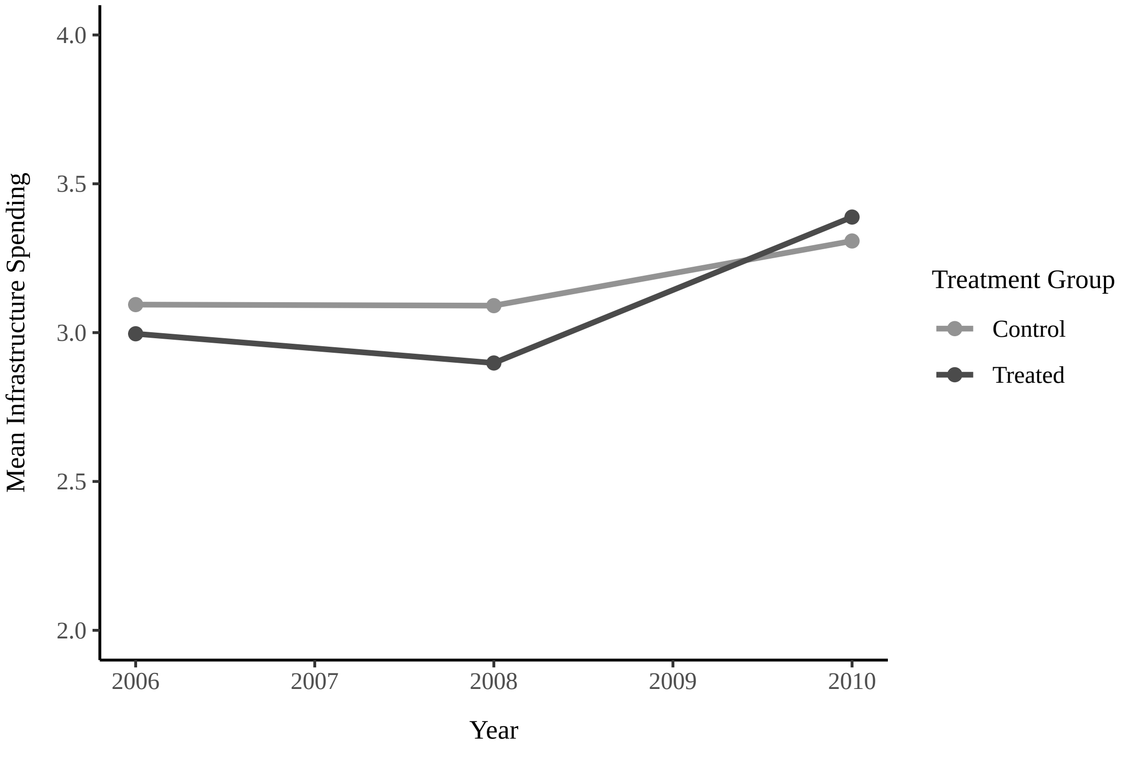
This assumption is important because it allows us to isolate the effect of the treatment on the outcome. If the parallel trends assumption holds, we can attribute any difference in infrastructure spending between treated and untreated districts after the treatment to the re-centralisation of control of public services. If the parallel trends assumption does not hold, then any difference in infrastructure spending could be due to other factors, and we would not be able to attribute it to the treatment.

2.4 Testing the Parallel Trends Assumption

Now, we'll assess the parallel trends assumption empirically. Return to the full dataset and calculate the means of the outcome variable for 2006, 2008, and 2010 for both the treated and control groups (six means in total). Plot these means separately over time for the treated and control groups. Do you think that the parallel trends assumption holds in this case?

`Figure 3` shows how the mean infrastructure spend has changed over time between 2006-2010. By looking at the trend of the data between 2006 and 2008, we can see that both the control and treatment groups were on roughly a similar trajectory, with the treated group seeing a sligh reducting in spend compared to the control group. We can assume that the similar trajectories seen between the two groups would have continued through until 2010 supporting the argument for the parallel trends assumption. As expected, parallel trend diverges in 2010, with the treated group seeing a large increase in infrastructure spending compared to the control group after re-centralisation of control of public services.

Figure 3: Mean Infrastructure Spending Over Time by Treatment Group



2.5 Placebo Difference-in-Differences

Create a new dataset for the years 2006 and 2008 only and use it to estimate a placebo difference-in-differences effect before the treatment occurred. What do you conclude about the parallel trends assumption?

Table 5: Placebo Difference-in-Differences Estimates

	Difference-in-Differences	Difference-in-Differences with FE
Post Treatment	−0.004	
	(0.031)	
Treatment Group	−0.133**	
	(0.065)	
Treatment Post Interaction	−0.088	−0.220***
	(0.092)	(0.066)
Log Population Density	0.094***	0.092***
	(0.012)	(0.012)
Num. Obvs	4498	4498
R-squared	0.016	0.016
Adj. R-squared	0.016	0.015

* p <0.1, ** p <0.05, *** p <0.01
Note: Standard errors are in parentheses. District fixed effects were excluded due to missing district values in 2006, but year fixed effects were retained to control for time-specific shocks.

The objective of the placebo difference-in-differences calculation is to estimate whether there is a difference in infrastructure spending between the treated and control groups before the treatment occurred, between 2006-2008. The placebo difference-in-differences estimate is `-0.088` for the non-fixed effects model, and `-0.220` for the fixed effects model. The fixed effects model uses `factor(year)` to control for unobserved heterogeneity as `disrict` is `NA` in 2006.

Since the more robust fixed-effects model is statistically significant, this model suggests treated areas were already on a downward trend vs control pre-treatment. Therefore, this raises concern over whether the parallel trends assumption holds. If there was a statistically significiant divergence in trends before the treatment occurred, then it may have been the case that they would not have been on the same trend between 2008-2010 either.