Université Paris-Saclay
Licence 3 of Economics – Impact Evaluation*

April 2024

# 1 Randomized Controlled Trials (RCT) and Propensity Score Matching (PSM)

## 1.1 context

**Randomized experiments cannot always be implemented**. Thus, estimating effects from microeconomic policy, such as the effect of job training program on earnings, must deal with **observational studies and data**. Several techniques have been developped, among which the **matching method** using propensity score. The matching technique is a **systematic method for creating comparison groups with similar covariate values than the treated individual**s, which can then be used to average the differences with the treated individuals and identify a causal effect.

One way to assess the ability of such an econometric method to solve the evaluation problem is to try the method in a situation where an experimental estimate is available. This is the strategy in LaLonde's classic 1986 paper and in the papers that reanalyze his data: Dehejia and Wahba (1999), Becker and Ichino (2002), Smith and Todd (2005).

LaLonde (1986) is the first to examine a randomized experiment (the National Supported Work Demonstration, NSW) in order to assess the performance of several evaluation methods. He obtains **an unbiased estimate of the training effect from the randomized experiment**, and then compares the experimental result to those obtained from a range of parametric selection models applied to the NSW observations that received training and a set of comparison observations constructed from population survey data sets (CPS and PSID). Following papers have reanalyzed Lalonde's results, using more recent estimation methods, including matching.

The goal is to analyze **the performance of regression** and **matching estimators** using observational study, replicating some of the results of these studies.

## 1.2 Data description

We here use a subset of the data used in LaLonde (1986), constructed by Dehejia and Wahba (1999). These data are available online: `http://www.nber.org/~rdehejia/nswdata2.html`

There are two types of samples:

- One sample of experimental data, which is a subset of Lalonde's data, and they are obtained from an experimental setting (see below for details on the experiment)

*Contact: yao.kpegli@ens-paris-saclay.fr

- Two samples of observational data, which form the comparison groups. These two samples are constructed from two surveys: the Population Survey of Income Dynamics (PSID), and the Current Population Survey (CPS).

The experimental data come from a randomized experiment from the National Supported Work (NSW) Demonstration. This program was a transitional, subsidized work experience program that provided trainees with work in a sheltered training environment and then assisted them in finding regular jobs. To participate in NSW, potential participants had to satisfy a set of eligibility criteria that were intended to identify individuals with significant barriers to employment. The main criteria were: (1) the person must have been currently unemployed (defined as having worked no more than 40 hours in the 4 weeks preceding the time of selection into the program), and (2) the person must have spent no more than 3 months on one regular job of at least 20 hours per week during the preceding 6 months. Actual treatment among participants was randomized. However, as a result of these criteria as well as of self-selection into the program, persons who participated in NSW are likely to differ in many ways from the general U.S. population.

Candidates eligible for the NSW program were randomized into treatment between March 1975 and July 1977. Treated individuals in our dataset concerned individuals assigned to treatment in 1976, so that retrospective earnings information from the experiment include calendar 1974 and 1975 earnings (variables RE74 and RE75). Thus, the pretreatment variables are earnings in 1974 and 1975. Postintervention data includes calendar 1978 earnings.

Variables, for both types of data, are (from left to right) treatment indicator (1 if treated, 0 if not treated), age, education, Black (1 if black, 0 otherwise), Hispanic (1 if Hispanic, 0 otherwise), married (1 if married, 0 otherwise), nodegree (1 if no degree, 0 otherwise), RE74 (earnings in 1974), RE75 (earnings in 1975), and RE78 (earnings in 1978), and dummy variables to identify data origin with exp (1 if experimental data, 0 otherwise) and cps (1 if CPS observational data and 0 otherwise).

1. We want to verify that the treatment and control groups have the same distribution of variables **before the program intervention**. Replicate the Table 1 of means and standard deviation in Dehejia and Wahba (1999) p.1055. What can you conclude from this table?

## 1.3 Econometric analysis and replication

### 1.3.1 Standard estimation of treatment effect

2. We are now interested in replicating the comparison of earnings and estimated training effects of the NSW program in Table 2 of Dehejia and Wahba (1999). We are first interested in the estimation of the effect of training on earnings using the experimental observations, which correspond to the first raw called "NSW" for the sets of columns B. and C. in their table.
   (a) From the experimental observations, what is the most simple unbiased estimator of the average treatment effect? Write down the linear model which allows to recover

this estimator. Then, estimate it with the experimental subset of data. How do you interprete your result?

(b) In the Table 2 of Dehejia and Wahba (1999), we note that in addition to the simple treatment estimation (columns called "Unadjusted"), the authors also compute estimated treatment effect while controlling for age, age squared, years of schooling, high school droupout status and race (columns called "Adjusted"). What is the purpose of controlling for other covariates in our experimental setting? Reproduce these two results.

(c) Finally, the authors add other variables which are pretreatment earnings (earnings in 1975 for the columns "Unrestricted differences in differences: quasi difference in earnings growth 1975–1978", and earnings in 1974 for all columns of panel C.). Replicate their results using these variables.

(d) What do you conclude on your results when you add pretreatment observables relative to your first simple estimate in question a?

3. We are now interested in assessing whether the regression method applied on observational data (no random assignment of treatment) can replicate the treatment impact. We separately use the PSID and CPS samples as comparison groups of the treated individuals (and thus exclude the non-treated experimental data).

(a) Using the observational data as a comparison group, what can we expect on the simple estimator of the average treatment effect while knowing the results of the data analysis in question 1? What about when accounting for pretreatment covariates?

(b) Perform the same regressions as in question 2 while using first the CPS data as a control group, and second the PSID data as a control group. What do you conclude on the estimation of the treatment impact while using the observational samples as comparison groups?

### 1.3.2 Estimating the treatment effect using the propensity score

The propensity score can be an intermediary tool to compare the average differences by matching experimental treated individuals with similar non-experimental individuals. The intuition of the propensity score is that whereas in the standard regressions we are trying to condition on $X_i$ (intuitively, to find observations with similar covariates), we are here trying to condition just on the propensity score because observations (experimental and non-experimental) with the same propensity score have the same distribution of the full vector of covariates $X_i$ (see class Session materials on matching and article of Dehejia and Wahba (1999) for further demonstration). Therefore, if we condition our regressions on the propensity score (that is by comparing individuals with similar propensity score), the distribution of pretreatment regressors will be the same accross the treatment and comparison group. Each individual has the same probability of assignment to treatment, as in a randomized experiment.

The estimation is done in two steps. In the first step, the propensity score is estimated **on the experimental treated units** and **on the PSID non-experimental sample**.

Second, given the estimated propensity score, **we estimate a regression of the effect of job training on earnings for individuals with similar estimated propensity score**. Finding the individuals with similar estimated propensity score, that is, *the matching method*, can be done with varying techniques. The two techniques developed in Dehejia and Wahba (1999) are the stratification strategy and the nearest-neighbor matching strategy.

4. First, we estimate the propensity score using the sample of the **experimental treated observations** and **the PSID observations** (thus discard CPS observations and experimental control observations). One issue is what functional form of the preintervention variables to include in the propensity score model. We thus want to test several functional forms for the propensity score and select the specification which leads to distributions of covariates for observations with similar propensity score that are the same accross the treatment and the comparison groups.
   (a) Generate the following variables: a dummy variable denominated `U74` which is equal to 1 if the individual is unemployed in 1974 (identified by 1974 earnings equal to 0) and 0 otherwise; a dummy variable denominated `U75` which is equal to 1 if the individual is unemployed in 1975 (identified by 1975 earnings equal to 0) and 0 otherwise; a dummy variable denominated `blackU74` which is equal to 1 if the individual is black *and* unemployed in 1974 and 0 otherwise; the squared education denominated `ed2`; and the 1974 and 1975 squared earnings denominated `re742` and `re752`.
   (b) We now try a first specification of propensity score. Estimate a **logit** specification of the propensity score (the probability of being treated) as a function of `age age2 ed ed2 black hisp married re75 re74 re742 re752 blackU74`. Then produce an histogram to compare the predicted propensity score between treated and non-treated observations. What do you conclude on your propensity score specification?
   (c) There are STATA packages which estimate the propensity score, and then can perform a matching strategy. We here produce a matching strategy which stratifies the data by propensity score, which can be opposed to a nearest-neighbor matching strategy. Observations with estimated propensity score which are in the same range are then grouped into blocks (strata) and we check whether we succeed in having similar distributions of covariates within each blocks for treated and comparison observations. Use the `findit` command to look for the `st0026_2` package developped in Becker and Ichino (2002) which test the balancing property for each block and each covariate and install it. Then, use the `pscore` command using the same covariates as for the preceding `logit` command, and with the pscore(mypscore) blockid(myblock) `logit numblo(5) level(0.005)` and `detail` in the options of the command. Explain how the algorithm processes and discuss the results obtained on your STATA output.

5. We now perform the OLS regression after restricting the NSW-treated / PSID-comparison sample using propensity score.
   (a) First estimate the propensity score using a Logit specification and where you add

control for `re75`, `age`, `age2`, `ed`, `black`, `hisp`, `nodeg` and `married`. Keep the observations of predicted values of propensity score which verify $0.1 < \hat{p}(X) < 0.9$.

(b) Finally, run the same previous regressions as those in questions 2 but on these matched observations. Compare your results, with the one obtained in the previous questions.

(c) What do you suggest to improve the matching strategy in this last approach?

# References


Becker, S. O. and Ichino, A. (2002). Estimation of Average Treatment Effects Based on Propensity Scores. *The Stata Journal*, 2(4):358–377.

Dehejia, R. H. and Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, 94(448):1053–1062. Publisher: Taylor & Francis.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, 76(4):604–620.

Smith, J. A. and Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1):305–353.


# 2 Difference-in-Differences (DiD)

This session is an introduction to the difference in differences (DiD) estimation strategy based on the paper: "*Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*", by David Card and Alan B. Krueger, American Economic Review (1994). We are interested in estimating the effect of an increase in minimum wage on the demand for low skilled labor.

## 2.1 Presentation of the research question and the data

Before moving on to the practice, we address some preliminary questions on the research question and on the research design to identify a causal effect of a variation in minimum wage on the demand for labor.

1. What does economic theory predict on the effect on an increase in minimum wage on employment, in perfectly competitive markets?

2. We want to estimate the effect empirically. We observe that minimum wage regulations vary across States in the US (some States have higher minimum wage regulations than others). We have a database with wage and employment for firms in different States.

(a) Write the regression of employment on minimum wage.

(b) Do you think that this simple regression might give us the causal effect of the minimum wage on employment? Explain.

In order to assess the impact of minimum wage laws on employment, David Card and Alan Krueger exploited the exogenous policy change in minimum wage that occurred in April 1992 in New Jersey (NJ): the hourly minimum wage was raised from 4.25 to 5.05 dollars in NJ but not in nearby States. Card and Krueger collected data at fast-food stores in NJ, before (in February 1992) and after the policy change (in November 1992) and also collected data at fast food stores in a nearby State, Pennsylvania (PA), where the minimum wage was not changed.

3. Consider the policy change:

   (a) What is the interest of looking at a policy change in order to estimate the effect of minimum wage on employment?

   (b) What is the interest to look at fast food stores?

   (c) What is the interest of having stores in different US states? Compare the characteristics of the stores by states and by time period.

   (d) We need to create unique variables measuring full-time equivalent employment and starting wage for the two time periods. Generate a variable denominated `fte` which measures the full-time equivalent employment in the store (including managers and the number of part-time employees count for half of the full-time employees), and another variable denominated `starting_wage`. *Note: be careful not to forget that the same variables observed in February 1992 and in November 1992 are stored in different variables.*

   (e) How can you check in the data that the policy has indeed been implemented in NJ stores and not in PA stores?

4. We now investigate the variation in full time employment in New Jersey stores. Compute the difference in average full-time employment employment in NJ stores before and after the policy change. Do you think this calculation provides an unbiased estimate of the effect the increase in minimum wage on employment? Why?

## 2.2 The Difference-in-Difference estimation strategy

Card and Krueger used their data set to compute differences-in-differences (DD) estimates of the effects of the New Jersey minimum wage increase. That is, they compared the change in employment in New Jersey to the change in employment in Pennsylvania around the time New Jersey raised its minimum wage.

5. Card and Krueger apply the following strategy: they compute the difference in average full-time employment before and after the policy change in NJ and Pennsylvania and calculate the difference in these average differences.

(a) Compute this "difference-in-differences". How do you interpret this number? Under which hypothesis this strategy can estimate the causal effect of the minimum wage on employment?

(b) Can you compute the "counterfactual" average FTE employment in NJ stores after April 1992 if the policy had not been implemented?

6. Let's perform the same analysis in a regression:

   (a) Write the corresponding regression model that would give you the same DiD estimator of the effect of the policy change, and perform the regression. Show how the results change when you include additional controls for: location within State, chain ownership, type of chain (i.e. KFC, Wendys, Roys or BK).

   (b) Compare the results with a standard OLS regression of employment on wages.

7. Discuss the validity of the DiD strategy implemented. What checks can be done to test the validity of the strategy?

# References

Card, D. and Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4):772–793. Publisher: American Economic Association.

Card, D. and Krueger, A. B. (2000). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply. *American Economic Review*, 90(5):1397–1420.

# 3 Instrumental Variables (IV)

## 3.1 context

Origins of differences in economics development between countries is debated in history economics and development economics. Three main sources are usually confronted: **geography** to the extent that climate, natural resources or physical constraints influence technological diffusion and agricultural production (Bruhn and Gallego, 2011); **international trade** which favors resource and technological transfers between countries; and **institutions** as a set of rules and norms which frame individuals actions and decisions (Acemoglu et al., 2005; Rodrik et al., 2004).

The article of Acemoglu et al. (2001) investigates **the role of institutions on economic development among countries colonized by Europeans**. In their article, they measure the quality of institutions through the security of property rights. The main mechanisms they want to underline is that countries with more secure property rights will invest more in physical and human capital, and will use these factors more efficiently to achieve

a greater level of income. However, investigating the role of institutions on development is challenging, as the relationship between institutions and income is **endogeneous**. Indeed, there is potential **reverse causality** of income on institutions, that is "rich economies choose or can afford better institutions", and there are **omitted variables** as economies may differ in their institutions and their income for a variety of unobserved reasons. Thus, to estimate the impact of institutions on economic performance, we need a source of exogenous variation in institutions, which can be done using the instrumental variable approach.

The approach developed by Acemoglu et al. (2001) is the following. They present historical arguments for the present institutions to be inherited from past institutions established by colonizing countries. They show that institutions imposed by the colonial state persisted and lasted even after the independance of colonized countries. They also state that the colonization strategy and the associated established institutions depend on the feasability of settlements. In places where the environment is favorable for European settlement (disease-free and with similar climate), colonized countries were more likely to establish institutions replicating those of Europe, while in countries with low feasability of settlements colonized countries established poor protection of property rights to extract the resources from the countries.

Thus, they use settler mortality, at the time of the colonization, as an instrument for current institutions in these countries. Note that there is no selection bias here as they are interested in the effect of colonization policy conditional on being colonized, thus focusing only on the sample of countries which were colonized. In their econometric approach, they thus regress in a first stage current institutions by settler mortality rates and in a second stage they regress current performance on the predicted outcome of their first regression. This is the standard **two stage least square** (2SLS) approach.

## 3.2 Data description and preliminary analysis

Data file: *data_AJR_2001.dta*

We measure current economic performance and current quality of institutions using World Bank data observed in 1995. Our dataset contains 163 observations, among which 64 concern past-colonized countries. The variables that we are going to use are the following:

`shortnam` = three-letter country code

`euro1900` = Percentage of European settlers in the population in 1900.

`avexpr` = Risk of expropriation of private foreign investment by government, from 0 to 10, where a higher score means less risk. Mean value for all years from 1985 to 1995.

`logpgp95` = Log GDP per capita 1995, Purchasing Power Parity Basis, from World Bank

`cons1` and `cons00a` = Constraint on executive in first year of independence and in 1900. Seven-category scale, from 1 to 7, with a higher score indicating more constraints. Score of 1 indicates unlimited authority; score of 3 indicates slight to moderate limitations; score of 5 indicates substantialimitations; score of 7 indicates executive parity or subordination. Equal to 1 if country was not independent at that date.

`democ1` and `democ00a` = Democracy in first year of independence and in 1900. An 11-category scale, from 0 to 10, with a higher score indicating more democracy. Points from three dimensions: Competitiveness of Political Participation (from 1 to 3); Competitiveness of Executive Recruitment (from 1 to 2, with a bonus of 1 point if there is an election); and Constraints on Chief Executive (from 1 to 4). Equal to 1 if country not independent at that date.

`logem4` = estimated log of settlers mortality (standardized measure corresponds to the number of soldiers, missionaries, sailors who died over 1000 men strength)

`loghjypl` = Log output per worker, 1988

`baseco` = Colonial dummies. Dummy indicating whether country was a British, French, German, Spanish, Italian, Belgian, Dutch, or Portuguese colony.

`lat_abst` = Absolute value of the latitude of the country (i.e., a measure of distance from the equator), scaled to take values between 0 and 1, where 0 is the equato

`rich4` = variable binaire indiquant les "néo-europes", territoires au climat tempéré colonisés par les européens

`catho80`, `muslim80` and `no_cpm80` = Percent of population that belonged to the three most widely spread religions of the world in 1980 (or for 1990-1995 for countries formed more recently). The four classifications are: Roman Catholic, Protestant, Muslim, and "other.

`sjlofr` = French legal origin dummy: Legal origin of the company law or commercial code of each country. Our base sample is all French Commercial Code or English Common Law Origin

`avelf` = Ethnolinguistic fragmentation: Average of five different indices of ethnolinguistic fragmentation

`temp` = Temperature variables: Average temperature, minimum monthly high, maximum monthly high, minimum monthly low, and maximum monthly low, all in centigrade

`humid` = Humidity variables: Morning minimum, morning maximum, afternoon minimum, and afternoon maximum, all in percent

`deslow`, `stepmid`, `desmid`, `drystep` and `drywint` = Soil quality: Dummies for steppe (low latitude), desert (low latitude), steppe (middle latitude), desert (middle latitude), dry steppe wasteland, desert dry winter, and highland

`goldm`, `iron`, `silv`, `zinc` and `oilres` = Natural resources: Percent of world gold reserves today, percent of world iron reserves today, percent of world zinc reserves today, number of minerals present in country, and oil resources (thousands of barrels per capita.

`yellow` = Yellow fever: Dummy equal to 1 if yellow fever epidemics before 1900 and 0 otherwise.

`landlock` = Dummy for landlocked: Equal to 1 if country does not adjoin the sea.

`malfal94` = Malaria in 1994: Population living where falciporum malaria is endemic (percent)

`leb95` = Life expectancy: Life expectancy at birth in 1995.

`lt100km` = Distance from the coast: Proportion of land area within 100 km of the seacoast.

1. We first want to reproduce the Table 1 of descriptive statistics in Acemoglu et al. (2001). We are here interested in whether we can observe significant differences in mean between colonized countries and the rest of the world, and between colonized countries grouped by their level of settlers mortality.
   (a) Construct a variable which identifies the quartile of settler mortality of colonized countries to which belong each country.
   (b) Compute the mean value and standard deviation for the whole sample, the subsample of colonized countries and the subsamples of colonized countries by quartiles of settler mortalities, of the variables `logpgp95`, `loghjypl`, `avexpr`, `cons1` and `cons00a`, `democ00a` and `extmort4`.
   (c) Using this table, what can you conclude on the differences of economic performances, institutions and settler mortality between countries? Does your table support the mechanisms highlighted by the authors?

## 3.3 The naive econometric analysis

2. While we know that it is likely that the estimation of a linear model of economic performance as a function of the measure of the quality of institutions will suffer from endogeneity issues, we still implement this approach as a first step. It is an interesting first step to better investigate the correlation between our variables and how it is affected by some control variables. In this question, we reproduce the OLS estimation presented in Table 2 of Acemoglu et al. (2001).

   (a) Consider the simple model where the log of income of country $i$, denoted $y_i$ (measured using the log of GDP per capita) is a linear function of a constant, the risk of expropriation $R_i$ and an error term $\varepsilon_i$, as follows:

   $$y_i = \mu + \alpha R_i + \varepsilon \tag{4}$$

   How do you interprete the coefficient $\alpha$ in Equation (4)?

   (b) Now, consider the following model:

   $$y_i = \mu + \alpha R_i + X_i^{'}\gamma + \varepsilon \tag{5}$$

   where $X_i$ is a vector of other covariates, such as geographical variables (dummy for continents and lattitude variables). How do you interprete the coefficient $\alpha$ using this second model?

   (c) Now, reproduce the OLS estimations presented the Table 2 of Acemoglu et al. (2001), and interprete your results regarding your estimated $\hat{\alpha}$ across estimations.

## 3.4 The instrumental variable approach

To instrument current institution quality, the authors use the rate of settlers of mortality at the time of colonization. We are going to show that this can be considered as a valid instrument, that is an exogeneous source of variation whose effect on current economic performance can only be assigned through past and current institutions.

3. We first investigate the relationship between settlers mortality and current economic performance. Briefly recall what is the serie of mechanisms between these two variables. Then, reproduce the Figure 1 in Acemoglu et al. (2001).

   Equation (5) describes the relationship between current institutions and log GDP. In addition we have

   $$R_i = \lambda_R + \beta_R C_i + X_i^{'}\gamma_R + \nu_{Ri} \tag{7}$$
   $$C_i = \lambda_C + \beta_C S_i + X_i^{'}\gamma_C + \nu_{Ci} \tag{8}$$
   $$S_i = \lambda_S + \beta_S m_i + X_i^{'}\gamma_S + \nu_{Si} \tag{9}$$

   where $C$ is the measure of early institutions, $S$ is the measure of European settlements in the colony (fraction of the population with European descent in 1900), and $m$ is the logarithm of mortality rates faced by settlers. $X$ is a vector of covariates that affect all variables.

4. To document our IV strategy we consider the models in Equations (7), (8) and (9).
   (a) Estimate the models in Equations (7), (8) and (9) on the sample of colonized countries and using latitude as a control variable.
   (b) How do the results can be interpreted with regards to the hypothesized serie of mechanisms linking settlers mortality and current institutions?

5. We first reproduce the two stage least square (2-SLS) approach "by hand". Protection against expropriation variable, $R_i$, is treated as endogenous, and modeled as

$$R_i = \zeta + \beta m_i + X_i^{'}\delta + \nu_i \tag{10}$$

   We then obtain predicted values of the protection against expropriation risks variable $\hat{R}_i$ from the estimation of the model in Equation (10). These are the variations of institutions assigned through the variations of the settlers mortality and partialling out any other endogeneous variations in institutions. We then incorporate this predicted value in replacement of $R_i$ in the model of Equation (5), which is now an exogeneous regressor.
   (a) Estimate the first stage of the 2-SLS approach, that is the estimation of the model in Equation (10) where you control for the lattitude of country $i$. Store your predicted values in a new variable.
   (b) Estimate the second stage of the 2-SLS approach. Interprete your estimation of the $\hat{\alpha}$ coefficient. How does it compare to the estimation in the naive approach in question 2.(c)?
   (c) Why does this 2-SLS approach "by-hand" is not valid?

6. STATA has a valid built-in command to implement IV regression using the 2-SLS approach. The command is `ivreg` where you need to precise the instrumental variable of your endogeneous variable. That is, if you want to regress economic performance on current institutions by instrumenting the latter with the settlers mortality the syntax is: `ivreg logpgp95 (avexpr=logem4)`. Replicate the estimation done in question 5 using this STATA command. What do you observe?

## 3.5 Validity of the approach

### 3.5.1 The exclusion restriction

The validity of the 2SLS results using the IV approach depends on the assumption that settler mortality in the past has no direct effect on current economic performance. This is also called the exclusion restriction. In other words, conditional on the controls included in the regression, the mortality rates of European settlers more than 100 years ago must have no effect on GDP per capita today, other than their effect through institutional development. Thus, the mortality of settlers must be excluded from the causal model of interest.

7. We check the exclusion restriction assumption of our instrument in the following questions.
   (a) What variables (available in our dataset) can plausibly be correlated with both settler mortality and economic outcome?
   (b) Add these variables as control variables in your 2SLS estimation done in question 6, and check whether the addition of these variables affects your estimates.

#### 3.5.2 The effect of outliers

8. One may argue that results are driven by some countries with very specific institutions and economic performance in the sample, not corresponding to the mechanisms at hand for the the majority of the sample. Indeed, richer colonized countries, such as the United States, Australia or New Zealand, have very different characteristics to the other countries in the sample. Replicate the estimation done in question 6 when excluding these countries (identified with the dummy variable `rich4`)). How does it affect your results?

## References


Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review*, 91(5):1369–1401.

Acemoglu, D., Johnson, S., and Robinson, J. A. (2005). Chapter 6 Institutions as a Fundamental Cause of Long-Run Growth. In Aghion, P. and Durlauf, S. N., editors, *Handbook of Economic Growth*, volume 1, pages 385–472. Elsevier.

Bruhn, M. and Gallego, F. A. (2011). Good, Bad, and Ugly Colonial Activities: Do They Matter for Economic Development? *The Review of Economics and Statistics*, 94(2):433–461.

Rodrik, D., Subramanian, A., and Trebbi, F. (2004). Institutions Rule: The Primacy of Institutions Over Geography and Integration in Economic Development. *Journal of Economic Growth*, 9(2):131–165.


# 4 Regression Discontinuity Design (RDD)

## 4.1 Context

You are asked to analyze artificially generated data inspired by an article by Kremer et al. (2009) entitled "Incentives to Learn". Using a regression-discontinuity method, you are tasked with analyzing the impact of a merit-based scholarship program for young girls in a region of western Kenya. In this region, 85% of children attend primary school, but a large number drop out as early as grades 3, 4, or 5. Admission to secondary school depends on the results

of an examination at the end of grade 5. Secondary education, although subsidized, requires significant parental contribution to costs. In this program, the top 15% of girls with the highest scores received a scholarship to cover the costs of secondary school.

Note: In the original article, the scholarship program was also randomly allocated among different schools, but we do not consider this aspect here.

The database concerns young girls living in villages where the scholarship program has been implemented. For simplicity, we will refer to the group of girls who received a scholarship as the treatment group, and the other girls as the control group.

**Key Variables**:

- momsecondary = 1 if the student's mother attended secondary school
- dadsecondary = 1 if the student's father attended secondary school
- runningwater = 1 if the parents' house has access to running water
- distance = Distance in kilometers between the parents' house and the nearest health post
- score = Result obtained by the girl on the end-of-grade 5 test
- winner = 1 if the girl received a merit scholarship
- highestgrade = the final level of education obtained by the girl

## 4.2 First part: Descriptive statistics

This section allows you to describe your sample. What is the level of education in your sample? What are the potential endogeneity issues you face?

1. Present descriptive statistics regarding the parents' level of education, access to running water (as an indicator of wealth level), and the distance to the nearest health post (as an indicator of remoteness). Discuss the results obtained.

2. Merit-based scholarships are awarded to the top 15% of students: construct a variable named "merit" taking the value 0 if the individual scored below this threshold, and 1 if above.

3. For each of the variables used in question 1, conduct a statistical difference test between the sample of girls who did not achieve a score sufficient to merit the scholarship and those who did.

4. Conduct a regression between the final level of education of the girls and the test score obtained. What is the value of the parameter obtained? What is its statistical significance? Do you think that this coefficient can be interpreted strictly causally? If not, why?

## 4.3 Second part: Graphical Analysis

In this part, you are asked to produce a series of graphs to visualize the possible existence of discontinuities.

1. Create a graph presenting, for each interval of 10 points on the test scores, the proportion of girls who actually received the scholarship. Comment on your results. In your opinion, are we more likely to find ourselves in a situation of clear discontinuity or fuzzy discontinuity?

2. Create the same graph, this time using the average final level of education obtained by the girls instead of the test score. Comment on your results.

## 4.4 Third part: Verification of Discontinuity Validity

1. Create a variable called "interval1," which equals 1 for all girls whose score is within 10 units of the minimum score to qualify for the scholarship, and 0 otherwise.

2. Repeat the series of difference tests produced in the first part, but this time only for the sample that satisfies the condition "interval1=1". Do your results differ from those obtained in the first part? What are the implications?

3. Create three additional variables "interval2," "interval3," and "interval4" where you vary the number of units between the obtained score and the merit variable level.

4. Repeat the series of difference tests and comment on your results.

## 4.5 Fourth part: Estimations

In this part, you are asked to estimate the level of the causal relationship between receiving the scholarship and the final level of education achieved.

1. Perform an OLS estimation of the final level of education on receiving the scholarship. What coefficient do you obtain? What is its statistical significance? What does this indicate about the magnitude of the effect of the scholarship on the number of years of education completed?

2. Perform the same estimation, but this time adding the distance to the test score threshold as another control variable. How does this affect the coefficient obtained for the scholarship? If you observe changes, how do you interpret them?

3. Perform the same estimation as in 2, but this time adding the test score obtained as another control variable raised to the square, then to the square and the cube. How does this affect the coefficient obtained for the scholarship? If you observe changes, how do you interpret them?

4. Conduct a two-stage estimation to assess the effect of the treatment on the number of years of study around the discontinuity.

5. Comment on your findings regarding the impact of merit scholarships on the level of secondary education attainment among young girls.

# References

Kremer, M., Miguel, E., and Thornton, R. (2009). Incentives to learn. *The Review of Economics and statistics*, 91(3):437–456.