Main points from:

Applied Econometric Policy Evaluation

Louise Otte Arildsen
Martin Nørgaard Petersen
University of Copenhagen

Spring 2021

These notes cover the research designs covered in Applied Econometric Policy Evaluation as taught by Søren Leth-Petersen and Daniel le Maire in spring 2021. It further contains non-exhaustive lists of things to consider when writing and conducting an econometric analysis. It lends from Angrist and Pischke: Mostly Harmless Econometrics (2009).

[Update 2023]: If you want to get a better conceptual understanding of causality in economics, consider reading "Potential Outcome and Directed Acyclic Graph Approaches to Causality"by Guido Imbens (accessible on JSTOR). Imbens introduces the Potential Outcome (PO) approach to causality which is the one treated by Angrist and Pischke (and Søren and Daniel) and compares it to the Directed Acyclic Graph approach. He argues for the favourability of the PO-approach and makes direct reference to Mostly Harmless Econometrics. It is quite a long read, but I came to understand the purpose of the course much better after having read the paper.

1 Descriptive Analysis

Any analysis should be initiated with a descriptive analysis, see suggested implementation in the `STATA`-code.

Commenting on descriptive statistics The descriptive analysis should include comments on the following:

- How many observations are in the data set?
- Is the data cross-sectional, time-series or panel data?
- Are there any missing data?
- What are average, minimum and maximum values for each variable?

Also the following considerations should be taken into account:

- If creating new variables later in the regression analysis, consider including them in the descriptive statistics.
- With panel data, consider a graph showing important variables as functions of time.
- With cross-sectional data, consider a histogram.

See section 8 for a table template.

2 Regression Discontinuity Design

Some rules are arbitrary and therefore provides possibilities for natural experiments. Regression discontinuity (RD) research design exploits precise knowledge of the rules determining treatment. RD comes in three variations: sharp, fuzzy and kink regression discontinuity design. Interestingly, at no value of the running variable will we observe treated and untreated observations and results are thus dependent on our willingness to extrapolate across covariate values.

Sharp Regression Discontinuity Sharp RD is used when treatment status is a deterministic function of x_i, where x_0 is a cutoff. Hence treatment D can be written as an indicator function $D_i = \mathbb{1}(x_i > x_0)$. Further, it is discontinuous as regardless of how close one approaches x_0, treatment is unchanged until $x_i = x_0$. We may write the general case as:

$$Y_i = \beta_0 + f(x_i) + \rho D_i + \eta_i, \tag{2.1}$$

where Y_i is the outcome variable and f is a continous function close to x_i.

Fuzzy Regression Discontinuity The fuzzy version exploits discontinuities in the *probability* or expected value of treatment conditional on x_i. Fuzzy RD leads to an IV-type of setup where the discontinuities become an instrumental variable for treatment status. Let T_i indicate the point where $E(D_i|x_i)$ is discontinuous. The first stage is then:

$$D_i = \gamma_0 + g(x_i) + \pi T_i + \xi_{1i} \tag{2.2}$$

And the fuzzy RD is then

$$Y_i = \beta_0 + \rho D_i + \xi_{2i} \tag{2.3}$$

Which may be expanded to:

$$Y_i = \mu + \rho g(x_i) + \rho\pi T_i + \eta_i, \tag{2.4}$$

given $\mu = \beta_0 + \rho\gamma_0$ and $\eta_i = \xi_{1i} + \xi_{2i}$.

Regression Kink Design While sharp and fuzzy RD research designs exploits discontinuities in the assignment rule, regression kink design exploits that the rule determining treatment causes a kink in the relationship between the variable and the underlying treatment variable (i.e. a discontinuity in the first derivative of the running variable).

As in a sharp regression discontinuity design, regression kink design estimates quantities "close" to a cutoff. Instead of estimating a shift in the intercept, we are interested in estimating the change in slope. The regression kink design examines the slope of the relationship between the outcome of interest and the treatment variable at the exact location of the kink in the policy formula.

2.1 Interpretation of parameters

In Regression Kink Design we estimate an equation on the form:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2(x - x_0) + \beta_3 D_i(x - x_0) + \beta_4 Z_i + u_i \tag{2.5}$$

Here β_0 is a constant term. β_1 is only significant in regression discountinuity design. β_2 measures the slope to the left of the cutoff and $\beta_2 + \beta_3$ measures the slope to the right of the cutoff (as $D_i = 0$ if $x < x_0$). When doing regression kink design, we are mainly interested in estimating β_3 – that is the change in slope. Note that the interpretation is the same for RD beside that $\beta_2 = \beta_3 = 0$

Estimates In order to be able to interpret results we have to divide the estimated change in slope over the kink with the change in slope in the benefit schedule. This gives the treatment effect.

2.2 Specification test and critical assumptions

Under Regression Discontinuity Design we need to make the following critical assumptions.

2.2.1 Critical assumptions

- Treatment is deterministic.
- Expected outcome is a continuous function of covariates, i.e. only treatment is a discontinuous function of x_i, so thus there is no 'bunching'. See below.
- Treatment may be described by a functional form (as we're extrapolating past the discontinuity).

2.2.2 Testing for manipulation of the running variable (bunching)

We must check for bunching of the running variable. This is done to make sure that individuals with interest in treatment do not manipulate x_i close to the cutoff. We may do so either graphically or analytically.

Graphically A simple graphical inspection of bunching around the cutoff can be conducted using a **histogram** depicting the running variable around the cutoff value.

McCrary Test Formal test for manipulation of the running variable. Is implemented in `STATA` with the `DCdensity` package. The McCrary test uses local linear regression on both sides of the cutoff (and thus requires large mass points close to the boundary) and returns a **discontinuity estimate** and a corresponding standard error. If the discontinuity estimate is significant the null hypothesis of no bunching is rejected and the identifying assumption for the regression discontinuity design is not met. In this case one cannot state the existence of a causal effect.

2.2.3 Discontinuities in background characteristics

There should not be kinks in the control variables at the cutoff values, as it would muddy the analysis of the variable of interest at the cutoff. We may test if there are any significant change in slope (i.e. if β_3 is significant in equation 2.5) using different specifications. We prefer the specification at which we find no significant change in slope around the cutoff.

Graphically We may use the `STATA` command `binscatter` graphically to show a given covariate as a function of the running variable. We should expect no significant jumps around the cutoffs.

RD regressions We may also choose to formally test for discontinuities by regressing the covariate on the running variable, including the treatment variable dummy. If the coefficient on the treatment variable D is significant we should worry about discontinuities.

If we find significant results from the regressions, we might examine if these persist when decreasing the bandwidth. Optimally we seek to decrease the bandwidth as long as results remain significant in order to avoid extrapolation. If the discontinuities disappear when decreasing the bandwidth – but the results remain significant – we might conclude that in fact what might seem as being discontinuities is extrapolation.

2.2.4 Nonlinearity

One may mistakenly interpret nonlinearity for discontinuities or kinks.

Pseudo cutoffs or permutation test This technique relies on estimating the equation with arbitrary values for the cutoff.

1. Divide the sample in two at the cutoff and define two pseudo cutoffs as the medians of each sub-sample as suggested by Imbens and Lemieux (2008).

2. Run RD regression separately for different pseudo cutoffs around the true cutoff value. Expect to get smaller and smaller (or less significant) estimates the further away from 0 the pseudo cutoff is.

   The latter may be conducted in `STATA` using the `rdpermute` The test compares the estimated effect at the true cutoff point to the estimated effect at pseudo cutoff points. It tests for the sharp null hypothesis of no effect of the policy on the outcome. The `rdpermute` will output a p-value that the cutoff is random. In case this i below the significant level (5 pct.) we reject the null that there is no discontinuity.

   Dependent on the research design, the test may be conducted for sharp or fuzzy RD or regression kink design.

Local Linear Regression We may estimate local average in near proximity to the cutoff x_0 to reduce the likelihood of mistaking nonlinearity for discontinuities. Using non-parametric estimation reduces our dependence on correct functional form and the assumption of constant treatment effects. However it requires a high masspoint close to x_0 and hence may be unsuited if the running variable is discrete. We use local linear regression to reduce bias, that would otherwise occur close to endpoints or cutoffs.

We may implement the local linear regression using OLS and a kernel putting less weight on observations further from the cutoff. In `STATA` we may use `reg` with a triangular kernel.

When choosing the bandwidth of the kernel we face a regular bias-variance tradeoff. We might thus consider different bandwidth (or Silvermann's rule of thumb).

2.2.5 Discrete running variables

If the running variable is discrete, we are unable to compute local averages arbitrarily close to the boundary. Within a narrow interval we may use OLS and disregard the misspecification bias when approximating a linear CEF. However on wider intervals we cannot disregard this factor. Further heterogeneity robust standard errors will turn out too small – rather we should check clustered standard errors, clustered on the running variable and choose the largest to be conservative.

Donut RD When the running variable is discrete we may further be unsure how to treat observations when $x_i = x_0$. As a robustness check we should consider removing observations that satisfy $x_i = x_0$ when the running variable is discrete.

2.3 Things to remember

Conducting an analysis based on Regression Discountinuity Design, we ought to remember the following

- Comment on the slope after the kink.
- Comment on any discontinuities.
- Comment on bunching of the running variable.
- When producing `binscatter` plots, comment on the size of the preliminary effect from eye-balling the diagram.
- Compare results using the `rdrobust` package.
- When conducting local linear regression, remove parametric variables (squared and cubed variables) as the point is to conduct a *non-parametric* estimation.

3 Instrumental Variables

Consider the equation:

$$Y_i = \alpha + \rho S_i + A_i'\gamma + v_i, \tag{3.1}$$

and assume that A_i' is unobserved (that is S_i is *endogenous)*. If we have access to an additional variable Z_i (which we call the instrument) for which it holds that

$$\text{cov}(A_i'\gamma + v_i,\, Z_i) = 0, \tag{3.2}$$

or in other words Z_i is *uncorrelated* with any other determinants of the dependent variables. This is called the **exclusion restriction** (we may exclude the instrument from the causal model).

We may accordingly write:

$$\rho = \frac{\text{cov}(Y_i,Z_i)}{\text{cov}(S_i,Z_i)} = \frac{\frac{\text{cov}(Y_i,Z_i)}{V(Z_i)}}{\frac{\text{cov}(S_i,Z_i)}{V(Z_i)}} = \frac{\pi_2}{\pi_1}, \tag{3.3}$$

alas the coefficient of interest ρ is given as the ratio between the **reduced form** (regression of Y_i on the instrument) and **the first stage** (regression of the causal variable of interest S_i on the instrument)

For this result to hold, the instrument Z_i must further satisfy the **relevance condition**, that is there must be a clear effect from the instrument on the causal variable of interest, i.e.

$$\text{cov}(Z_i, S_i) \neq 0 \tag{3.4}$$

3.1 Estimators

We may estimate the coefficients of interest in a different ways.

3.1.1 The Wald Estimator

The simplest IV-estimator is the Wald Estimator. Assume that our instrument Z_i is a dummy variable and that we have no other covariates in equation (3.1). The coefficient of interest ρ is thus given as

$$\rho_{\text{wald}} = \frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(S_i|Z_i = 1) - E(S_i|Z_i = 0)} \tag{3.5}$$

3.1.2 Two-Stage Least Squares (2SLS)

Most commonly we use 2SLS that inputs first-stage fitted values in the causal relation. It may be computed in two stages, hence the name, although we prefer to let `STATA` deal with it in order to obtain correct standard errors.

We note that the connection between the 2SLS estimator and the Wald estimator is grouped data, as the 2SLS estimator may be constructed as a set of Wald estimators.

3.1.3 LIML and JIVE

In the case of weak instruments limited information maximum likelihood (LIML) or Jackknife IV estimation is prefered.

3.2 Local Average Treatment Effects (LATE)

When allowing for heterogenous effects – that is abolishing our former assumption of a constant effect of treatment – we need to nuance the exclusion assumption.

Exclusion restriction The instrument only operates through a single known causal channel. That is to say, the potential outcome is only a function of S_i.

Further we need to add the following assumptions:

Independence We assume that our instrument is as good as randomly assigned. That is, it is independent of potential outcomes and treatment assignments. This assumption allows for causal interpretation of the reduced form.

Monotonicity Those effected by the instrument is affected in the same way or direction. While some may not be affected, we do not allow the instrument to both push some people into treatment and some out.

Under the above assumptions (and the first stage assumption) the Wald estimand can be interpreted as the average causal effect on the group affected by the instrument.

We need the last assumption as the net effect could in fact be due to opposite effects on defiers and compliers.

3.2.1 Compliant subpopulation

Let D_{1i} be the treatment status of individual i, when $Z_i = 1$, while D_{0i} is the treatment status when $Z_i = 0$. Then we may divide the population accordingly:

	$D_{0i} = 0$	$D_{0i} = 1$
$D_{1i} = 0$	Never-taker	Defier
$D_{1i} = 1$	Complier	Always-taker

Tabel 1: Definition of the division of population

Alas, never-takers never opt into treatment, regardless of the instrument, whereas always-takers always do so. Without the assumption of constant effects, LATE is not informative about the effect on these two groups, as they're unaffected by the instrument.

The assumption of monotonicity rules out the existence of defiers.

Average treatment effects on the treated Note that the treated consists of always-takers and compliers, which are two non-overlapping groups. The average treatment effect on the treated is a weighted average of the effect on both of these groups

Average treatment effect Likewise the non-treated consists of defiers and never-taker. The unconditional treatment effect is thus a weighted average of the effects on compliers, never-takers and always-takers. Hence, the LATE doesn't always equal the average causal effect on all the treated. This is only the case when we have instrumental variables that allow no always-takers and no never-takers.

Intention to treat (ITT) Because of noncompliance we often find that the instrument is not equal to treatment. Hence, the measure we really obtain is called the *intention to treat*, which can also have a causal interpretation (since we assumed Z_i was as good as random).

3.3 Specification checks

We may check for any association between the instrument Z_i and other characteristics that was determined before treatment and should thus not be affected by the endogenous variable S_i.

Also we may check for an association between the instrument and outcomes in samples where we know that there is no relationship between treatment and the instrument. (See Angrist and Pischke, p. 131).

4 Homogenous and heterogenous effects

With homogenenous treatment effects we assume that treatment affects the outcome variable equally regardless of covariates (i.e. age, gender, etc.). Often we may wish to examine if treatment effects are heterogenous, e.g. whether elderly people are affected differently from treatment than younger people.

In a usual regression setup we may allow for heterogenous effects by including interaction terms. If D denotes treatment, then we may include the interaction term $D(\texttt{age} - \overline{\texttt{age}})$, where $\overline{\texttt{age}} = \frac{1}{N}\sum_{i=1}^{N} \texttt{age}_i$. We subtract the mean of the controlling variable as not to affect the estimate of the treatment effect (ρ).

5 Other Research Designs

We consider the following research designs briefly.

5.1 Fixed Effect

Fixed effect research strategy requires panel data, i.e. repeated observations for the same individuals, firms or other units over time (that is data must be both time series and cross-sectional). If this is not the case, group-level omitted variable bias can be captured by group-level-fixed effects, which is known as the difference in difference research strategy (see the section below).

The key assumptions are 1) that the unobserved A_i (e.g. ability) appears without a time subscript in the linear model; meaning that it fixed over time, and 2) that the casual effect of the outcome variable is additive and constant.

The fixed effect model:

$$Y_{it} = \alpha_i + \lambda_t + \rho D_{it} + X'_{it}\beta + \epsilon_{it} \tag{5.1}$$

$$\text{where } \epsilon_{it} \equiv Y_{0it} - E[Y_{0it} \mid A_i, x_{it}, t] \text{ and } \alpha_i \equiv \alpha + A'_i\gamma \tag{5.2}$$

5.2 Difference in difference

Difference in difference research strategy design is a version of fixed effect estimation using aggregated data.

The difference in difference model can be stated formally as

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + X'_{st} + \epsilon_{ist}$$

Where δ is the parameter of interest, D_{st} is a dummy over state and time and X is a vector of, e.g., state and time-varying covariates. The model can also include a vector X_{ist} with individual characteristics.

5.3 Propensity Score And Matching

Matching is a strategy to control for covariates through making covariate-specific treatment-control comparisons and weighted together to produce a single overall average treatment effect (ATE).

The matching estimator have causal interpretation if the conditional independence assumption (CIA) holds. The CIA asserts that conditional on characteristics X_i, selection bias disappears. Formally:

$$\{Y_{0i},Y_{1i}\} \perp D_i|X_i. \tag{5.3}$$

Intuitively, the matching estimator is a strategy that for each treated individual we seek to find an individual or a unit of treated that has the same value of covariates (or the propensity score) or at least very close.

Regression and matching Pischke and Angrist argues that regression can be seen as a particular sort of matching estimator; the difference being the weights used to combine covariate-specific events into the average treatment effect.

Matching uses the distribution of covariates among the treated to weight covariate-specific estimates whereas regression produces a variance-weighted average of these effects.

The regression estimate may be written as:

$$\delta_R = \frac{E[\sigma_D^2(X_i)\delta_X]}{E[\sigma_D^2(X_i)]}, \tag{5.4}$$

where $\sigma_D^2(X_i)$ is the conditional variance of D_i given X_i and δ_X is the average treatment effect.

This shows that the regression estimator may be written as a treatment-variance weighted average of the matching estimator δ_X.

5.3.1 Propensity score theorem

The propensity score theorem suggests that assuming that the conditional independence assumption (CIA) holds, it is sufficient to control for the *probability* of treatment $p(X_i)$. Formally,

$$\text{if } \{Y_{0i},Y_{1i}\} \perp D_i|X_i \text{ then } \{Y_{0i},Y_{1i}\} \perp D_i|p(X_i). \tag{5.5}$$

Here, the *propensity score* conditional on covariates $p(X)$ is defined as:

$$p(X) \equiv E(D \mid X) = P(D = 1 \mid X) \tag{5.6}$$

The propensity score $p(X)$ is typically estimated by a `probit` or `logit` estimator. The intuition of the propensity score is that a simple linear regression will be inconsistent if X is correlated with D. However, it is sufficient to control for $p(X)$ since it is what is correlated with treatment D.

The cost of using the propensity score method is the risk of lower asymptotic standard errors, however in finite cases this point is of little importance. Indeed, some argue that there could be gains in precision from using the propensity score in finite data sets – intuitively dropping covariates that has a statistically small

effect on the probability of treatment (propensity score) eases the statistical burden that would arise from including the covariates directly.

The propensity score method seems rather attractive in applications where it is easy to motivate. However, the difference between regression and propensity score methods is arguably small – and regression can at times be seen as a version of propensity score matching. Further methods of calculating propensity scores as well as drawing inference has not been standardised as in the case of regression.

5.4 Non-parametric Bootstrap

The bootstrap sample distribution is an alternative to the asymptotic distribution. The bootstrap distribution is used when the asymptotic distribution is hard or complicated to construct. The bootstrap sample distribution is constructed by drawing from our sample with replacement. The draws are repeated multiple times, and the result is a sample distribution which provides a good proximation for our sample distribution.

Nonparametric bootstrapping or pairs bootstrapping is the simplest way of establishing bootstrap regression estimates. Here, a pair of y_i, X_i values is drawn. For parametric bootstrapping an $\hat{\epsilon}_i$ is drawn from the sample of residuals and X_i is kept fixed, and a new value of the independent value is created.

Furthermore, bootstrapping is used to refine asymptotic inference.

6 Standard Errors

We use robust standard errors as standard. We should discuss the use of standard errors, when discussing estimates.

6.1 Clustering

In cross section analysis we assume that data are independent, i.e. each observation is uncorrelated with the observation before and after. This is however often an unlikely scenario. Often we find that observations are correlated within a group or *cluster*. We call this the clustering problem or Moulton problem. Further, we might also face the issue of serial correlation in time series data.

For this reason we may use clustered standard errors on the running variable to correct for the bias of the standard errors. We do not cover the derivation of clustered standard errors, however note that with too few clusters the standard errors are downward biased.

Therefore Angrist and Pischke suggest at least 42 cluster as a rule of thumb (42 is the answer to life the universe and everything. If in doubt google it.). To be conservative we ought to check using both clustered and non-clustered robust standard errors, and apply the largest.

6.1.1 On a technical note

The two following pieces of STATA Code conducts the same estimation; regression with fixed time and entity effects clustering on entity.

```
reg lnths D lnemp i.year i.state, cluster(state)
xtreg lnths D lnemp i.year, fe cluster(state)
```

When running the regression without clustered SEs the results are identical, however when clustering there might be a need for conducting an adjustment of degrees of freedom (use option `dfadj`) available in the `xtreg` packages. In large-samples the '$N - K$'-degrees of freedom adjustment should be minor and negligible. Only in small samples should it be necessary to adjust for degrees of freedom. However, the argument for clustered standard errors indeed requires large samples, so the use of degrees of freedom adjustment for clustered standard errors is borderline skizofrenic.

7 General things to remember

When conducting econometric analysis, the following list covers important points. Clearly, it is not exhaustive.

- Always use robust standard errors as standard.
- Compare robust and clustered standard errors against against each other if relevant. Choose the largest for conservative reasons.
- When using panel data it is relevant to use clustered standard errors.
- Always comment on controls.
- If in doubt whether to include or leave out controls – do both and evaluate the results.
- Consider if there is a formal way to test for graphical evidence?
- Be aware of the dummy trap.
- When leaving out controls (for space considerations) remove those that are insignificant first.
- Comment on effects close to an eventual cutoff.
- Check if cutoff values have been defined correctly. I.e. $z > 30$ does *not* include 30, whereas `inrange(`z`,30,80)` does.
- Before handing over an econometric analysis with `STATA`-code, run the whole code from start to finish in order to receive a readable log-file.
- Regression discontinuity design may be constructed also as a difference in difference design.

,

7.0.1 Examples of specification tests

Additionally, one ought to consider the following specification tests.

- Are covariates balanced?
- Overlap in the distribution of treated and untreated for different levels of treatment probabilites.
- Test for no differences in pre-treatment outcomes between treated and non-treated.

8 Tables

The summary of the `STATA` output may be inputted in the following table: b

8.1 Estimation results

Estimation of different models may be inputted in tables on the following format. Further, encapsule the table in `\begin{landscape} ... \end{landscape}` to rotate the table. The `lscape`-package is already included if using the `packages.sty` from norgaardpetersen.dk.

	Model name 1	**Model name 2**
	b/se	*b/se*
D_{sub}^{super}	-67.715	
	(57.157)	
var. 2	264.826***	
	(9.275)	
var. 3	-53.267***	
	(9.783)	
var. 4	0.025***	0.038***
	(0.006)	(0.007)
var. 5	-0.790	-1.237
	(0.819)	(1.178)
var. 6	-1.948	7.658
	(16.465)	(18.808)
var. 7	-39.756	30.705
	(38.251)	(48.787)
var. 8		-31.920
		(128.529)
var. 9		209.337***
		(5.652)
var. 10		-74.439***
		(12.261)
Constant	6603.566***	11537.938***
	(109.029)	(223.992)