Econometric Methods for Treatment Evaluation
ECON6320 Semester 2 2017
September 25-27, 2017

Pravin K. Trivedi
University of Queensland
School of Economics
p.trivedi@uq.edu.au

September 2017

Introduction

Impact of interventions

- Interested in measuring the impact of an actual or hypothetical intervention in the context of an econometric model
- Interventions may be external (exogenous) or self-selected (endogenous)
- Variable of interest is called outcome.
- Variable of intervention is called treatment
- Both outcome and treatment can be discrete or continuous.
- Discrete means categorical, i.e. binary or multi-valued
- Continuous means measured on a continuous scale, like income

Examples: external interventions

- Does textbook subsidy improve learning?
- Do good lecture notes improve student grades?
- Do teacher incentives reduce absenteeism?
- Do minimum wage laws reduce employment?
- Does class size affect student performance? How much? (Maimonides rule works?)

Examples: endogenous interventions

- How much does an additional year of education add to earnings?
- How much does more comprehensive health insurance increase health expenditure?
- How much does a newly adopted technology affect productivity?
- Main complication of endogenous treatment is selection
- Total effect of intervention depends upon pure treatment effect and selection effect
- Total TE = Pure TE + Selection effect ; goal is decomposition

Why we need a treatment evaluation framework

- Understanding the impact of interventions is central to policy making and evaluation.
- Treatment effect (TE) is a measure of the impact of an intervention; impact is defined by reference to a chosen benchmark
- TE is calculated by comparing two outcomes, at least one of which is hypothetical, i.e. unobserved or unobservable.
- Econometricians treat TE as a causal parameter in a cause-effect framework (J. Pearl disagrees)
- We need a set of relationships and assumptions (econometric framework) for deciding whether the causal parameter of interest is in principle identifiable.
- Given identifiability we need an estimation procedure to estimate TE. To address these questions econometrically we need a framework of relationships which involve causal parameters.

Why regression is not a causal relationship

- Consider $y = \beta x + u$; y denotes health status, x denotes smoking intensity
- Write $u = y - \beta x = y - E[y|x]$; what is the interpretation of β?
  - gradient of $E[y|x]$; (calculus)
  - a parameter of the joint distribution of (y, x); (statistical - justifies regressing x on y also)
  - marginal effect of a unit change in x on $E[y|x]$; (calculus -says nothing about causality)
  - does β predict the effect of change in x on y? (depends upon exogeneity of x)
  - is β a causal parameter in the sense of measuring the the average impact of an exogenously administered change in x on y? (closer to causal)
  - what is the interpretation of the OLS estimate of β? (causal? statistical? calculus?)

What do the textbooks say?

- Chen and Pearl review 6 textbooks and argue ...."... textbooks provide weak or misleading discussion of causality"
- Chris.Auld.com blog reviews additional ten texts and with a couple of exceptions reaches a similar evaluation
- Standard textbook interpretations are confused
- Issue cannot be settled until we provide more details of the framework and available data
- Need to know the status of x

Standard econometric approach

Structure consists of (in matrix and vector notation**)**

1. variables $\mathbf{W}$ ("data" matrix) partitioned as $\left[\mathbf{Y}_{endog} \;\; \mathbf{Z}_{exog}\right]$;
2. a joint multivariate probability distribution of $\mathbf{W}$, $f(\mathbf{W})$; $f_J(\mathbf{Y},\mathbf{Z}|\boldsymbol{\theta}) = f_C(\mathbf{Y}|\mathbf{Z},\boldsymbol{\theta}_1)\times\; f_M(\mathbf{Z}|\boldsymbol{\theta}_2)$;
3. a priori $\mathbf{W}$ ordered according to hypothetical cause and effect relationships with specified a priori restrictions on the model;
4. specification of functional forms and the restrictions on the parameters of the model.

- models can be expressed as structural ("behavioral" or "autonomous") equations or reduced form ("derived") equations.
- a treatment variable $D \in \mathbf{Y}$ or $D \in \mathbf{Z}$, but often no treatment assignment rule
- key parameter is **marginal effect** of a change in a variable on another variable.
- If the perturbation comes from another endogenous variable, ME is computed using a structural equation; otherwise we use the reduced form.

Objectives of econometric model

1. Data description and summary
2. Conditional prediction and policy analysis, prospective and retrospective
   – Simulation of counter-factual scenarios
   – Analysis of interventions
3. Estimation of causal ("structural", "key") parameters
4. Empirical confirmation or refutation of hypotheses.

   – very highly **structured potentially large models**

   – **reduced form** studies which aim to uncover correlations and associations

- Impact of policy may vary across impacted population because of differential responses
- Interested in the distribution of impacts, not necessarily just the average impact

Does TE differ from traditional econometric modeling?

- Traditionally, regression methodology was the cornerstone of modeling; marginal effects are of special interest.
- Distinction made between "structural" and "reduced form" parameters; "structural" = "causal"? or portal to "causal"?
- TE or ME treated as a causal parameter which could be recovered in a structural regression model .
- Experimental framework with treatment, controls, and potential outcomes not explicitly used.
- Inference about causality is probabilistic and implied; example: demand and price
- Identification and estimation of causal ("structural") parameters of interest because they are **invariant**
- However, simulations and comparison of generated "scenarios" with benchmarks widely used.
- Methodology did not give a special status to **covariate balance** or the TE parameter.

Neyman-Fisher-Rubin Framework

- Recognition that not all parameters are "causal", and that causal interpretation requires a particular framework has led to development of alternative frameworks, called Potential Outcome Model (POM), derived from statistical experimental literature.
- How to get a causal parameter estimate?
- Some argue that only an experiment can settle the issue
- POM framework, originally due to Neyman and Fisher, but expanded by Rubin, is a response.
- N-F-R introduce the idea of **counter-factual causality**

Distinctive features of modern TE analysis

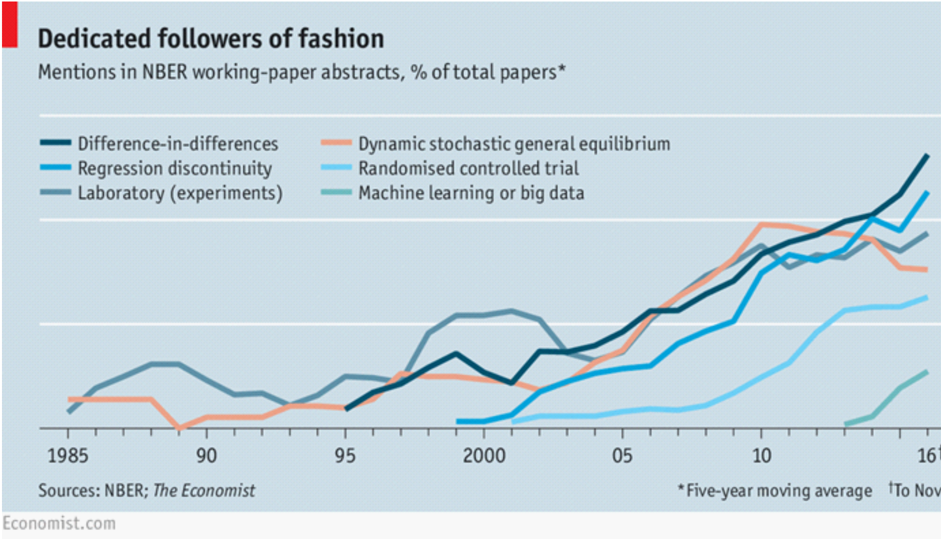
1. Causal inference requires **counterfactuals** generated by explicitly stated models of outcomes
2. Because inference on causality with **observational data** is in principle impossible. .
3. Interpretation of a causal parameter is based on a comparison of **potential outcomes** associated with levels of intervention
4. Potential outcome is a function of treatment and controls but focus is on **causal parameter(s)** associated with **treatment** (intervention)
5. Interventions are defined by **treatment assignment rule**

Distinctive features of modern TE analysis (contd.)

6. Identification of TE parameter is sought under weak functional form restrictions
7. New analytical tools designed explicitly for TE (matching, RCT, RDD)
8. Much attention is paid to the data used for identifying causal parameters - "not all data are necessarily relevant"
9. Borrows terminology and framework from experimental statistical literature
10. Mechanism by which causal intervention occurs is usually not spelt out fully - "blackbox feature"

Angrist-Pischke View

- AP claim in *Mastering 'Metrics* (2015) that "five most valuable econometric methods ["the Furious Five"] are"
  - random assignment (RCT)
  - regression (RA)
  - [matching methods!]
  - differences-in-differences (DiD)
  - instrumental variables (IV)
  - regression discontinuity design (RDD)
- These topics - mostly regression based - constitute the core of this course



Dedicated followers of fashion

Mentions in NBER working-paper abstracts, % of total papers*

Sources: NBER; *The Economist*

*Five-year moving average †To Nov

Economist.com

Course Description

- We survey number of identification and estimation strategies and their critiques.
- These widely used methods, are related to other established regression-based approaches.
- The goal of the course is to explain the logic, strength, and limitations of these methods.
- All methods we cover are potentially subject to criticism when applied incorrectly.

Plan of lectures

1. Preliminaries and overview
2. Rubin-Fisher-Neyman potential outcome model
3. Randomized and quasi-randomized trials
4. Regression adjustment
5. Matching methods
6. Natural experiments and event analysis; differences-in-differences
7. Endogenous treatment effects
8. Instrumental variable approach
9. Selection models
10. Regression discontinuity design
11. TE in general settings
12. Stata applications

Recommended references

Books

Angrist, J. D. & Pischke, Jö.-S., Mastering ‘metrics: The path from cause to effect, Princeton University Press, 2014

Cerulli, G., Econometric Evaluation of Socio-Economic Programs, Springer, 2013.

Imbens, G. W. & Rubin, D. B., Causal inference in statistics, social, and biomedical sciences, Cambridge University Press, 2015

Lee, M.-J. , Micro-econometrics for policy, program, and treatment effects, Oxford University Press Oxford, 2005

Glennerster, R. & Takavarasha, K. Running randomized evaluations: A practical guide Princeton University Press, 2013

Stata 15 Manual: STATA TREATMENT EFFECTS REFERENCE MANUAL: POTENTIAL OUTCOMES/COUNTERFACTUAL OUTCOMES

Survey articles, text books, critiques, reflections

S. Athey and G W. Imbens, The State of Applied Econometrics: Causality and Policy Evaluation, The Journal of Economic Perspectives, Vol. 31, No. 2 (Spring 2017), pp. 3-32

Deaton, A. Instruments, randomization, and learning about development J. of Economic Literature, 2010, 48, 424-455

Imbens, G. W. & Wooldridge, J. M. Recent developments in the econometrics of program evaluation J. of Economic Literature, 2009, 47, 5-86

Lee, David S., and Thomas Lemieux. Regression discontinuity designs in economics. J. of Economic Literature 48.2 (2010): 281-355.

Imbens, Guido W., and Thomas Lemieux. "Regression discontinuity designs: a guide to practice." J. of Econometrics 142.2 (2008): 615-635.

Imbens, G. W. Matching methods in practice: Three examples J. of Human Resources, 2015, 50, 373-419

Several graduate level texts, e.g. Cameron and Trivedi's Microeconometrics: Methods and Applications (chapter 25) and Wooldridge's Econometric Analysis of Cross Section and Panel Data (chapter 21), provide chapter-length treatment of treatment evaluation.

Established approaches of TE

Causal relationships

The central issue concerns the impact of (policy) intervention ("treatment"), endogenous or exogenous, on an outcome variable of interest.

- What alternative frameworks are available for analyzing the interventions?
- Was there an impact? Who was impacted?
- What impact-related parameter can we identify?
- What are the obstacles to identification of the treatment effect?
- What are the limitations of regression-based approach?

POM framework

Experimental approach to causation

- Neyman (1923 (Polish), 1990 (English)) put forward the idea that causal effects are comparisons of potential outcomes.
- Neyman's example: **potential** yield of i^{th} variety on k^{th} plot, U_{ik},
- Experimental design research of R A Fisher reinforced the concept of treatment assignment
- Assumes stable unit treatment value (SUTVA). and something like random assignment; means i's outcome depend only on i's treatment
  - No peer group effects exist/allowed; partial equilibrium approach
- Initially the concept of POM used mainly in the experimental setting.

Fisher-Rubin causal model

- Formal statement of treatment assignment (randomized vs. nonrandomized assignment) mechanism introduced by RAFisher (1925)
- Extension of the causal parameter concept to nonrandomized observational settings due to Rubin (1974, 1975, 1978)
    - Very relevant to econometric model with endog treatment, e.g. $y_2 = f_1(y_1, x),\ y_1 = f_2(z, x)$
- A comparison between hypothetical outcomes under different treatments can be made irrespective of the assignment mechanism. (Rubin)
- Connection between randomized treatment and potential outcomes was initially present in SEM in econometrics but weakened later (according to Imbens and Rubin, chapter 2)

Newer TE models (1)

- Emphasis is on a small number of "causal" parameters, sometimes just one.
- Distinction is between outcome and treatment variables. Other variables are just controls.
- Standard model has just two levels of treatment, $D = 0$ or $D = 1(binary\ treatment)$.
- Multi-level treatment set is $D_1, D_2, D_3, ..., D_m$ where treatment may be ordered or not.
- Continuous treatment variable can be accommodated.

Newer TE models (2)

- Parameters of special interest are

$ATE = \ E(\text{outcome}|\text{treatment, controls}) - E(\text{outcome}|\text{no treatment, controls})$ over the entire population

$ATET = \ E(\text{outcome}|\text{treatment, controls}) - E(\text{outcome}|\text{no treatment,controls})$ over the treated population

- The canonical version consists of just one or two equations, one so-called **"structural" or "causal" equation** which included a treatment variable (D) and the other a reduced form equation interpreted as **treatment assignment rule**.
- Which has more policy relevance: ATE or ATET?
- Standard notation: $y_0$refers to outcome w/o treatment $(D = 0)$, y_1 refers to outcome under treatment, $D = 1$
- Central question: under what assumptions is the causal parameter identified, and then consistently estimatable?

Restrictions on treatment assignment

Restriction on ass.	Econometric interpretation	Comment
1. Individualistic	i's TA prob. does not depend on x_j	
2. Probabilistic	$\forall\ i, \quad 0 < \Pr\left(D = 1 \textit{ or } 0 \middle| x, y_0, y_1\right) < 1$	All can receive
	$\forall$ possible $\left(x, y_0, y_1\right)$	treatment
3. Unconfounded	Zero dependence of assignment	CI assmption
	on potential outcome	exogeneity

Definition: Classical RCT

- Classical randomized control trials (RCT or CRCT) satisfies all three restrictions.
- In CRCT researcher knows and controls functional forms of assignment mechanism.
- In CRCT assignment mechanism is not confounded .
- Treatment assignment and subsequent outcome are conditionally (on controls) independent

$\Rightarrow$ TE is identified and estimation straight-forward.

Definition: Observational study

- In observational studies, exact assignment probabilities unknown.

& may have information about the assignment mechanism but not its functional form.

- Treatment assignment may be unconfounded, but treatment receipt may be confounded (e.g. selection)

$\Rightarrow$ Cond. independence of outcome fails and TE less straightforward to estimate.

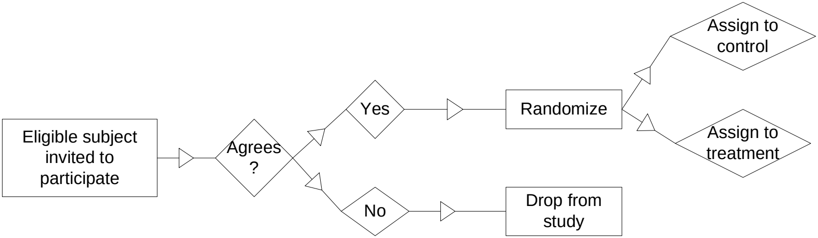
- **Regular assignment** mechanism is individualistic, probabilistic, unconfounded, but $\Pr\left(D = 1 \textit{ or } 0 | x, y_0, y_1\right)$ has unknown functional form.
- **Irregular assignment** mechanism may require a different approach, e.g. differences-in-differences.

Application with random assignment

Potential Outcome Model framework (1)

- Treatment evaluation concerned with measuring the impact of (broadly defined) interventions on outcomes
- Leading case: Continuous outcome variable, say y; treatment variable is discrete, D, where $D = 1$ if treatment is chosen or applied and $D = 0$ otherwise.
- If intensity of treatment can vary, use the term **multiple treatments;** the choice of a benchmark for comparisons is more flexible.

Simple RCT design



- RCT ⇒ Treatment is randomly assigned and hence independent of potential outcome.
- RCT precludes any selection effect which would imply $E[y, D] \neq 0$
- In observational data selection bias cannot be avoided in general.

Potential Outcome Model framework (2)

- Leading case: observe $(y_i, \mathbf{x}_i, D_i)$, $i = 1, ..., N$. x is in general a vector of pretreatment variables.
- Interested in the impact of a hypothetical change in D on y, holding $\mathbf{x}$ constant. Main feature of the so-called **potential outcome model (POM).**
- A key point: causal statements require both a factual and a counterfactual which in general is unobservable; ("missing data problem"), controls provide the counterfactuals
- Counterfactual scenario can be generated using POM.
- Target parameters are ATE and ATET.

Assumptions for TE

- SUTVA ⇒ Potential outcomes of unit i does not vary with treatments applied to other units. Same version of treatment applies to everyone.
- Individual outcome has idiosyncratic (unpredictable) i.i.d. component
- No individual is simultaneously observed in both states. Causal inference carried out in terms of **counterfactuals**.
- Triplet $(y_1, y_{0,} D)$ is the basis of treatment evaluation. Outcome under treatment $= y_1$, outcome for non-treated $= y_0$.
- How the outcome of an average untreated individual would change if such a person were to receive the treatment?
- Assume (1) randomized assignment of treatment and (2) any one that is assigned treatment gets it, and anyone that is not does not get it.
- **Estimates of treatment effects?**
  $ATE = \ E[y|D=1] - \ E[y|D=0]$. Each RHS term can be estimated as a sample average

Issues with POM

- RCTs provide a framework for simultaneously generating treated and counterfactuals.
- RCT can potentially remove the selection bias and TE easy to estimate.
- Inference based on observational data is more common.
- Observational data are commonly subject to (1) self-selection bias, (2) problem of finding relevant control group.
- But RCT is/may be possible only under exceptional conditions.

– Social experiments (expensive, ethically questionable at times and difficult to implement)

– Natural experiments

Toolkit for RCT

RCT Toolkit

0) Duflo, Glennerster and Kremer have a sort of manual for implementing RCT. See also the book
Glennerster, R. & Takavarasha, K. *Running randomized evaluations: A practical guide* Princeton University Press, 2013
1) Rationale for use of randomization: (a) remove selection bias and (b) combat publication biases.
2) How to incorporate randomization in a research design.
3) Design issues: sample size, stratification, level of randomization, and data collection methods.
4) How to allow for departure from perfect randomization.

Selection bias

(1) Randomization resolves the selection bias problem.
The regression version of the ATE expression is obtained from

$$y_i = \alpha + \beta D_i + [\gamma x_i] + \varepsilon_i$$

The treatment effect is the OLS coefficient of $\beta =$ $E[y_i|D_i = 1, [x_i]] - \ E[y|D = 0, [x_i]]$ as $E[\varepsilon_i|D_i, [x_i]] = 0$, i.e. D_i is uncorrelated with ε_i.,the unconfoundedness assumption.

(2) Other methods to control for one variety of selection bias

a. Controlling for selection-on-observables bias by including x on which y also depends.

$$y_i = \alpha + \beta D_i + \gamma x_i + \varepsilon_i$$

(3) Want $cov(x, D) = 0$ to gain efficiency and to avoid confounding.

Types of selection bias - selection on observables

- Given $y_i = \alpha + \beta D_i + \gamma x_i + \varepsilon_i$, and $cov(\varepsilon, D) \neq 0$ means that assignment is correlated with outcome.
- Then β (group mean difference) is not a consistent estimate of ATE.
- Suppose, however, that the assumption $cov(\varepsilon, D|\mathbf{z}) = 0$ where $\mathbf{z}$ is a vector of exogenous variables so that $cov(\varepsilon, \mathbf{z}) = 0$

i.e. conditional on **z**, treatment assignment and treatment outcome are no correlated.

- Implies we can assume random assignment if we can control for **z**.
- Then a consistent estimate of ATE (β) is obtained from the regression $y_i = \alpha + \beta D_i + \gamma \mathbf{z}_i + \varepsilon_i$
- This is the case of **selection on observables**
- Requires knowledge of functional form linking y and observable **z**

Types of selection bias - selection on unobservables

- Given the regression $y_i = \alpha + \beta D_i + \gamma \mathbf{x}_i + \varepsilon_i$; if $cov(\varepsilon, D|\mathbf{x}) \neq 0$, then again treatment assignment and outcome are correlated.
- Referred to as (i) selection on unobservables model, or (ii) endogenous dummy variable model.
- OLS estimator of β is inconsistent.
- Consistent estimation methods include

- (i) MLE based on a two-equation model of outcome and treatment assignment (STATA 15's `erm` command),
- (ii) instrumental variable method based on untestable hypothesis about assignment mechanism (STATA's `ivregress` command)

- Options require functional form assumptions

Publication bias

(1) Definition of publication bias
Publication bias occurs when editors, reviewers, researchers, or policymakers have a preference for results that are statistically significant or support a certain view, ⇒ selective suppression of negative results.
(2) RCT prevents manipulation of the experiment to produce biased results
(3) RCT solves the publication bias
a. If RCT is correctly implemented, there can be no question that the results give us the impact of the particular intervention that was tested.
b. In randomized evaluation the treatment and comparison groups are determined before a researcher knows how these choices will affect the results, limiting room for ex post discretion, which is called "cherry-picking".
c. Randomized evaluations can also partially overcome the file drawer and journal publication bias.

RCT design considerations

Sample and power considerations

In RCT implementation, the goal is to ensure statistically significant estimate of ATE, sufficiently high power against alternatives.

Table: Components of size and power analysis

Description	Symbol
significance level (type 1 error probability)	α
type 2 error probability	β
power	$1-\beta=\pi$
total sample size	N
treated sample size; control sample size	N_1; N_0
treatment group (mean, variance)	(μ_1, σ_1^2)
control group (mean, variance)	(μ_0, σ_0^2)
treatment effect size	$\delta=\mu_1-\mu_0$

Required sample size in RCT

- To run RCT we have to determine N_1, N_0
- Null hypothesis of zero treatment effect: $H_0 : \mu_1 = \mu_0$
- Two-sided paired z-test or t-test can be used to test the null of zero treatment effect.
- Cannot simultaneously determine both α,and β. But prefer a smaller type II error.
- In practice, require a test of minimum desired power $(1 - \beta)$ given a specified minimum detectable average difference δ between treated and untreated groups.
- δ,effect size, depends upon the scale of measurement, so we work with a standardized value, δ/σ
- We use the following to solve for required sample sizes after choosing other values.

Tests and power equation for two-sided t-test

$$z = \frac{(\overline{y}_1 - \overline{y}_0) - (\mu_1 - \mu_0)}{\sqrt{\sigma_1^2/N_1 + \sigma_0^2/N_0}} \sim N(0,1)$$

$$t = \frac{(\overline{y}_1 - \overline{y}_0) - (\mu_1 - \mu_0)}{\sqrt{s_1^2/N_1 + s_0^2/N_0}} \sim t(\nu)$$

$$\pi = \Phi(\delta/\sigma_D - z_{1-\alpha/2}) + \Phi(-\delta/\sigma_D - z_{1-\alpha/2})$$

where $\overline{y}_1 = N_1^{-1}\sum_i y_{1i}$, $\overline{y}_0 = N_0^{-1}\sum_i y_{0i}$, $s_1^2 = \sum (y_{1i} - \overline{y}_1)^2/N_1$ and $s_0^2 = \sum (y_{0i} - \overline{y}_0)^2/N_0$, $\sigma_D = \sqrt{(\sigma_1^2/N_1 + \sigma_0^2/N_0)}$.

- z-test can be used when variances are known (tricky issue in practice).
- Given unknown/ unequal variances, the t-statistic test has an approximate Student's $t-$distribution with (in general non-integer) d. of f. ν obtained using so-called Satterthwaite's formula.

Power of a one-sided paired t-test

Type I error – rejecting H_0 when it is true; Type II error – failing to reject H_0 when it is false.

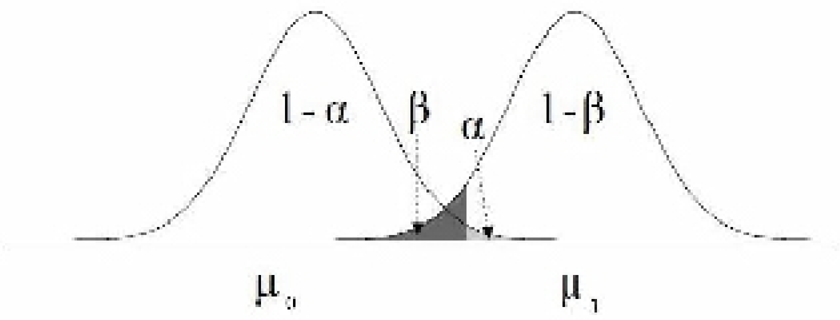
The probability of a Type II error depends on unknown population parameter so can only be computed for given values.

Consider the power of a one sided test for $\alpha = .05$, in a large sample where the test statistic has normal distribution

Then the power of the test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$ is given by

$$\pi(\delta) = 1 - \Phi\left[1.64 - \frac{\delta}{\widehat{\sigma}_D / \sqrt{N}}\right]$$

, where $\widehat{\sigma}_D$ is the standard error of the mean difference $(\overline{y}_1 - \overline{y}_0)$



1 - α

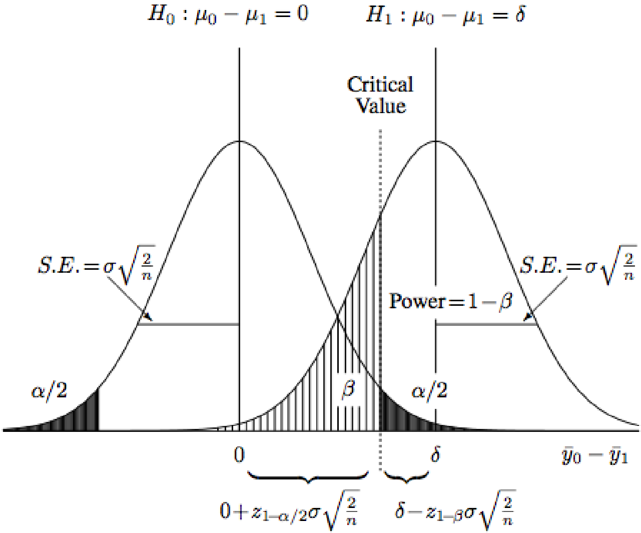
β

α

1 - β

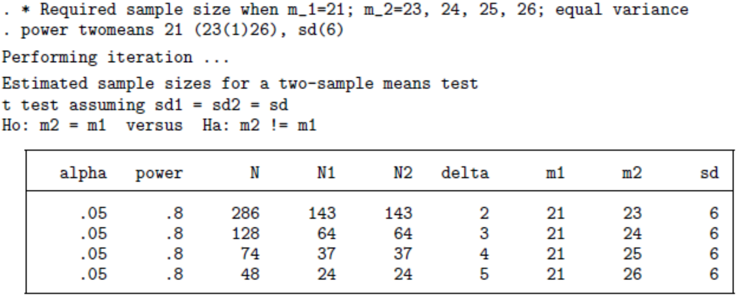
μ₀

μ₁


$H_0 : \mu_0 - \mu_1 = 0$
$H_1 : \mu_0 - \mu_1 = \delta$
Critical Value
$S.E. = \sigma\sqrt{\frac{2}{n}}$
$S.E. = \sigma\sqrt{\frac{2}{n}}$
Power $= 1 - \beta$
$\alpha/2$
β
$\alpha/2$
0
δ
$\bar{y}_0 - \bar{y}_1$
$0 + z_{1-\alpha/2}\sigma\sqrt{\frac{2}{n}}$
$\delta - z_{1-\beta}\sigma\sqrt{\frac{2}{n}}$

Example of sample size calculation

Assuming **equal** sample sizes, and **given** desired δ, π and α the power equation can be solved iteratively for the required sample size; or in the case of unequal samples, can solve for required N_1 given N_0, and vice versa. Standard practice sets $\alpha = .05$, and $\pi = 0.8$ or 0.9. In Stata this computation is done using the command `power twomeans`.



```
. * Required sample size when m_1=21; m_2=23, 24, 25, 26; equal variance
. power twomeans 21 (23(1)26), sd(6)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

  ┌──────────────────────────────────────────────────────────────────────┐
  │  alpha   power        N       N1       N2   delta       m1       m2       sd │
  ├──────────────────────────────────────────────────────────────────────┤
  │    .05      .8      286      143      143       2       21       23        6 │
  │    .05      .8      128       64       64       3       21       24        6 │
  │    .05      .8       74       37       37       4       21       25        6 │
  │    .05      .8       48       24       24       5       21       26        6 │
  └──────────────────────────────────────────────────────────────────────┘
```

Remarks

- Power increases with sample size (N), treatment effect size δ, and decreases with variance of outcome σ^2.
- Power equation can be solved for (equal) sample size if we fix π, δ, σ^2
- Variance parameters are generally unknown and may require a **pilot** to determine starting values.
- More illustrations are given in the C-T draft chapters

Variants of RCT

Required TE using a regression based test

- Two-sample t- or z-test is equivalent to testing the significance of regression coefficient β in the regression

$$y_i = \alpha + \beta D_i + \varepsilon_i$$

- Assume that a proportion P of the sample is treated, and $\varepsilon_i \sim iid(0, \sigma^2)$, then

$$var(\widehat{\beta}) = \sigma^2 / N(P(1-P))$$

- The power of a test for a true effect size of $\beta \neq 0$ and significance level of α, is the probability of rejecting the null hypothesis.

To achieve power π, we need

$$\widehat{\beta} > (t_{(1-\pi)} + t_{\alpha}) std.err.(\widehat{\beta})$$

Minimum TE

The minimum detectable effect size for a given power (δ), significance level (α), sample size (N), and portion of sample being treated (P) is

$$MDE = \delta = (t_{(1-\pi)} + t_{\alpha})\sqrt{\sigma^2 / N(P(1-P))}$$

Remark: There is a trade-off between power and size.
Equal division between treatment and comparison group is optimal, because the MDE is minimized at $P = 0.5$.

Stratified randomization

- Balancing samples is important because it limits the range of alternative explanation of the data and paradoxes.
- In a randomized experiment, controlling for other covariates won't affect the consistency of $\widehat{\beta}$, but it can reduce its variance.
- Hence including valid regressors (variables that impact outcome) in the regression will increase power.
- And stratifying (or blocking) ex ante is more efficient than controlling ex post, since it ensures an equal proportion of treated and untreated units within each block and therefore minimizes variance.
- An extreme version of blocked design is the pairwise matched design where pairs of units are constituted (for example, twins), and in each pair, one unit is randomly assigned to the treatment and one unit is randomly assigned to the control.

Level of randomization

- Usually the researcher can choose the level of randomization: the individual or the group level.
- Factors need to be considered:

(1) Budget. The larger the groups that are randomized, the larger the total sample size needed to achieve a given power. This makes individual-level randomization attractive.
(2) Spillovers from treatment to comparison groups can bias the estimation of treatment effects, especially for randomization at individual level.
(3) Randomization at the group level may be much easier from the implementation point of view.
(4) Randomizing strata will generally lead to correlated or clustered observations. Variance calculations of the treatment effect should adjust for clustering.

Re-randomization

- Characteristics **x** may be poorly balanced across treatment and control groups
- Startification, blocking and matching methods can be used to improve balance (see under matching)
- Simultaneous stratification in multiple dimensions can be difficult and may reduce sample size
- Recommendation is that if sample is unbalanced, re-randomize until balance achieved.
- Re-randomization could reduce the robustness of conclusions and increase the cost of RCT.
- How many randomizations is enough?

Data collection

(1) Practical application of size/power calculations must overcome the difficulty that it requires as inputs parameter values which are typically unknown. Use results from previous studies or conduct baseline surveys or **pilot trials.**

• Baseline survey generates control variables that will reduce the variability in final outcome and therefore reduces sample size requirements.

• Make it possible to examine the interactions between initial conditions and the impact of the program.

(2) Using administrative data (collected by the implementing organization as part of their normal functioning) could reduce the cost.

(3) Assumption that treatment randomization is across individuals when in practice randomization often takes place across strata first and then across individuals within the strata.

Pros and cons of RCT

See Deaton, JEL 2010, for a strong critique of RCT

P.1: Improves efficiency in principle. See example below.

P.2: Smaller sample size required, hence lower cost of RCT

P.3: Improves understanding of causal mechanism (Duflo, AEA Ely Lecture 2017)

C.1: Although by randomization $x \perp D$, by chance corr(x, D) may arise. (Hawthorne effect)

C.2: If x correlated or connected with past outcomes, and randomization is not perfect then corr$(x, u) \neq 0$

If x is to be included then ensure that it is pre-treatment value and decision to include made before the RCT is run.

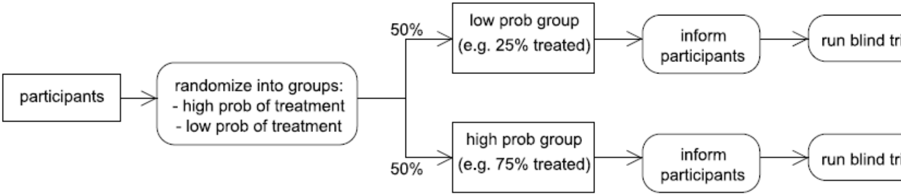
C.3: Treated and untreated groups may not be balanced in respect of covariates x.

C.4: Optimal experimental design for a randomized trial may depend upon expectations of participants.

C.5: Questionable external validity.

Two-stage randomization design dependent on behavior (Chassang et al, 2015)

Figure 2: A Two-by-Two Blind Trial.



participants

randomize into groups:
- high prob of treatment
- low prob of treatment

50%

low prob group
(e.g. 25% treated)

inform participants

run blind tri

50%

high prob group
(e.g. 75% treated)

inform participants

run blind tri

Notes: The figure shows the two stages of randomization, with participants first allocated to either a hig or low-probability treatment group, then informed of this probability (thus generating the correspondi placebo effect), and then receiving either treatment or non-treatment in a standard, blinded manner. Sour Chassang et al. (2015).

Is RCT a "gold standard" for estimation of causal parameters?

- Has a black box character if RCT design not based on understanding of why the treatment works; e.g. treated mosquito nets
- If mechanism linking treatment and outcome not very clearly established, may not be able to say more than "it works" or "it worked"
- RCT may not be feasible for ethical reasons.
- Treatment effect may depend on behavior/expectations of the treated.
- Adaptive behavior on the part of subjects means the untreated group may find substitutes for treatment and contaminate the sample.
- External validity may be questionable if outcome heavily dependent on special features of the RCT environment.
- External validity doubtful if mechanism of treatment effect not understood.

Mean difference vs. regression adjustment

Regression adjustment approach

- Whereas test of group mean difference is easy to implement directly using standard software, there are advantages in doing so in regression framework.

This is especially the case when the regression involves multiple control variables, perhaps with nonlinearities in control variables.

- TEs from RCT can also be calculated using the **marginal effects** (ME) approach.
- ME approach is a unified approach to calculation of treatment effects in both RCT and observational data.
- Refer to the regression based approach as regression adjustment (RA).

Estimation with group mean difference and OLS

- Consider the "double regression" model

$$y_i = \beta_1 + \beta_2 D_i + \beta_3 x_i + \beta_{23} D_i x_i + u_i.$$

- This potentially allows for different response to x_i between groups. Dropping the interaction term implies β_3 does not vary between groups.
- Assume selection on observables only, and i.i.d. errors u_i'
- Assume random treatment assignment and hence $Cor(y_j D_j) = 0, \quad j = 1, 0$
- Potential outcomes (PO) $y_{i,PO}$, can be generated as predictions of the two regression models for treated (factual) and untreated (counterfactual) groups.

POM difference equals ME

- Initially ignore x_i
- Then ATE is identified with the group mean difference $E(y|D=1) - \ E(y|D=0)$ which can be calculated

with the sample group average difference, or by OLS regression of y_i on 1 and D_i.

$$\begin{array}{lcl} POM(1) & = & E[y_i|D_i = 1] = \beta_1 + \beta_2 + E[u_i|D_i = 1] = \beta_1 + \beta_2 = \mu \\ POM(0) & = & E[y_i|D_i = 0] = \beta_1 + E[u_i|D_i = 0] = \beta_1 = \mu_2 \\ POM(1) & -POM(0) = & E[y_i|D_i = 1] - E[y_i|D_i = 0] = \mu_1 - \mu_2 = \beta_2 = ATE \end{array}$$

- $\widehat{ATE} = (\widehat{\beta}_1 + \widehat{\beta}_2) - \widehat{\beta}_1 = \widehat{\beta}_2$ - also marginal effect (ME) of D is consistent.
- $\widehat{ATET} = N_1^{-1} \sum_{i=1}^{N_1} (\widehat{y}_i | D_i = 1) \ - N_1^{-1} \sum_{i=1}^{N_1} (\widehat{y}_i | D_i = 0)$ is consistent
- Including x_i and/or $D_i x_i$ means that we are controlling for covariates and allowing for interactions

Test of non-zero TE

- It follows that the null hypothesis of zero treatment effect is equivalent to the hypothesis $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$.
- A two-sided t-test is based on the Wald test statistic

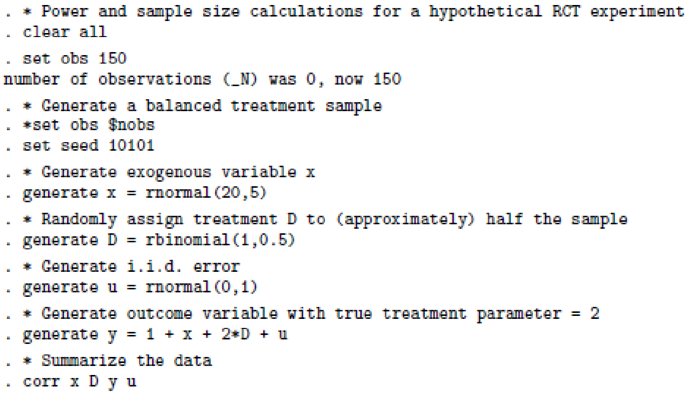
$$T = \frac{\widehat{\beta}_2}{\sqrt{\widehat{var}(\widehat{\beta}_2)}} \sim t(\nu)$$

This assumes i.i.d. errors but can be robustified against heteroskedasticity.

RA example based on simulated data

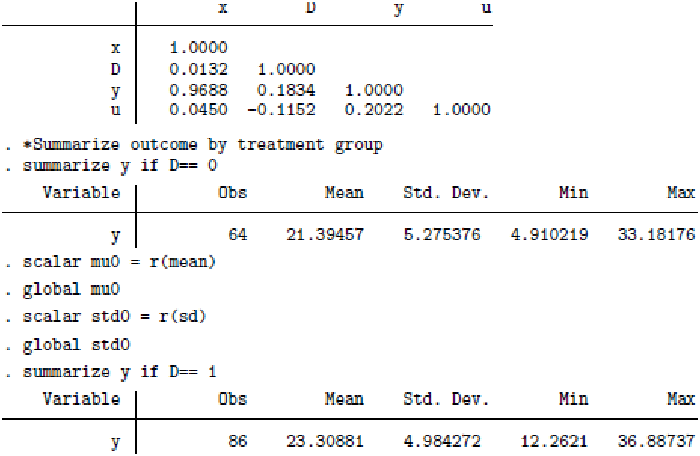
Simulation design

Generate a noisy sample with true treatment effect of 2 units in a model with significant covariate



```
. * Power and sample size calculations for a hypothetical RCT experiment
. clear all

. set obs 150
number of observations (_N) was 0, now 150

. * Generate a balanced treatment sample
. *set obs $nobs
. set seed 10101

. * Generate exogenous variable x
. generate x = rnormal(20,5)

. * Randomly assign treatment D to (approximately) half the sample
. generate D = rbinomial(1,0.5)

. * Generate i.i.d. error
. generate u = rnormal(0,1)

. * Generate outcome variable with true treatment parameter = 2
. generate y = 1 + x + 2*D + u

. * Summarize the data
. corr x D y u
```

Two-sample t-test

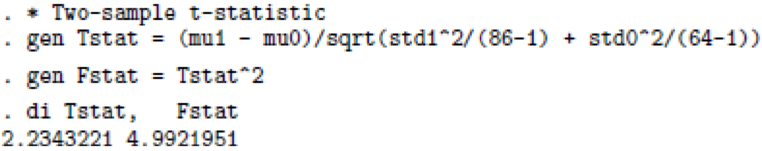


```
             |        x        D        y        u
-------------+------------------------------------
           x |   1.0000
           D |   0.0132   1.0000
           y |   0.9688   0.1834   1.0000
           u |   0.0450  -0.1152   0.2022   1.0000

. *Summarize outcome by treatment group
. summarize y if D== 0
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           y |         64    21.39457    5.275376   4.910219   33.18176
. scalar mu0 = r(mean)

. global mu0

. scalar std0 = r(sd)

. global std0

. summarize y if D== 1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           y |         86    23.30881    4.984272    12.2621   36.88737
```

Two-sample t-test has low power

- The estimated treatment effect is close to the true value of 2, but has a large standard error.

Because the outcome is noisy, there is loss of power



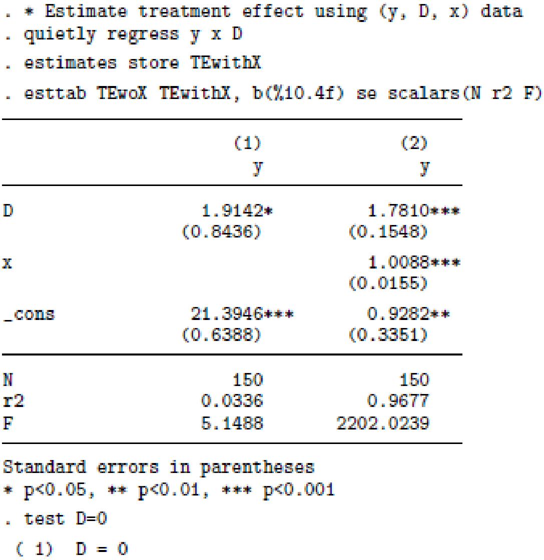
```
. * Two-sample t-statistic
. gen Tstat = (mu1 - mu0)/sqrt(std1^2/(86-1) + std0^2/(64-1))

. gen Fstat = Tstat^2

. di Tstat,   Fstat
2.2343221 4.9921951
```

- Controlling for the variation due to a relevant covariate x will improve the fit of the model and lead to a more precise estimate of the TE.

The confidence interval should be narrower.

- Controlling for x allows us to make do with a smaller sample than otherwise.

Conditioning on valid x improves precision and power

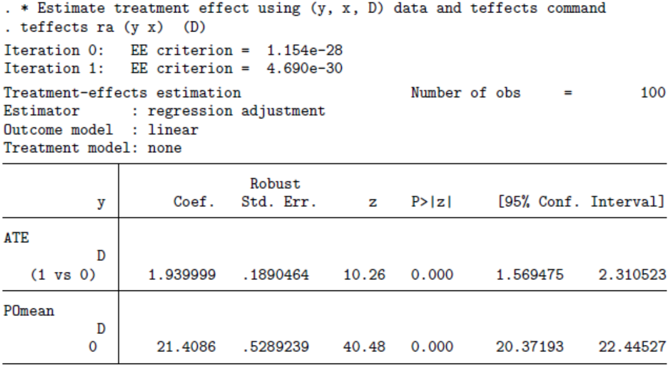


```
. * Estimate treatment effect using (y, D, x) data
. quietly regress y x D

. estimates store TEwithX

. esttab TEwoX TEwithX, b(%10.4f) se scalars(N r2 F)

--------------------------------------------
                      (1)             (2)
                        y               y
--------------------------------------------
D                  1.9142*         1.7810***
                 (0.8436)        (0.1548)

x                                  1.0088***
                                 (0.0155)

_cons             21.3946***       0.9282**
                 (0.6388)        (0.3351)
--------------------------------------------
N                     150             150
r2                 0.0336          0.9677
F                  5.1488       2202.0239
--------------------------------------------
Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

. test D=0

 ( 1)  D = 0
```

The teffects ra command

- For linear regression based inference it operates as follows.
- For a binary treatment, we run two regressions of y on x, one for the the subsample with $D = 1$, and a second for the subsample with $D = 0$.
- Each regression is used to generate predictions for the full sample. Denote these, respectively, as $\widehat{y}_1$ and $\widehat{y}_0$. These are estimates of (in principle unobservable) potential outcome means, POM.
- The ATE is the average of the difference between the two POMs.
- Example which follows shows that for this sample the difference between the two is small.
- **Explain why there is a difference at all and what it means for randomization**.

Example of teffects ra



```
. * Estimate treatment effect using (y, x, D) data and teffects command
. teffects ra (y x)  (D)
Iteration 0:   EE criterion =  1.154e-28
Iteration 1:   EE criterion =  4.690e-30
Treatment-effects estimation                    Number of obs     =        100
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
           D |
   (1 vs 0)  |   1.939999   .1890464    10.26   0.000     1.569475    2.310523
-------------+----------------------------------------------------------------
POmean       |
           D |
          0  |    21.4086   .5289239    40.48   0.000     20.37193    22.44527
------------------------------------------------------------------------------
```

TEs in nonlinear regression

How to measure treatment effects

- In linear models ATET $=$ ATE $=$ ME (marginal effect) $=$ constant
- In nonlinear models, irrespective of the assignment mechanism, ME and treatment effect are not constant but depend upon the functional form of the conditional mean.
- A marginal treatment effect measures the effect on the conditional mean of y of a change in treatment variable D.
- For continuous treatment $\text{ME}_j = \partial E[y|D = D^*]/\partial D$ where D^* is the treatment level. If $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$ and x_j is the treatment then $\text{ME}_j = \exp(\mathbf{x}^{*\prime}\boldsymbol{\beta})\beta_j$ which varies with $\mathbf{x}^*$
- For discrete treatment variable D the finite difference method yields marginal effect $\text{ME}_j = E[y|\mathbf{x} = \mathbf{x}^*, D = 1] - E[y|\mathbf{x} = \mathbf{x}^*, D = 0]$.

Marginal TE as building blocks for ATE

- In nonlinear models the marginal TE varies with the point of evaluation.
- In nonlinear models coefficients are more difficult to interpret
- Three common choices of evaluation are (1) at sample values and then average; (2) at the sample mean of the regressors; and (3) at representative values of the regressors.

AMTE=ATE	Average marginal TE	Average of ME at each treatment level
AMTET=ATET	ATE for the subpopulation receiving treatment	Average of ME for treatment receipients
MEM=ATEM	Marginal TE at mean treatment value	ME at $D = \overline{D}, x = \overline{x}$
MER=ATER	Marginal effect at a representative value	ME at $D = D^*, x = x^*$

Marginal TE for polynomial regressors

- If treatment variable appears as polynomials computing ATE becomes more complicated.
- Example: First consider a linear model that includes a cubic function in regressor D. Then $E[y|\mathbf{x}, z] = \mathbf{x}'\boldsymbol{\beta} + \alpha_1 D + \alpha_2 D^2 + \alpha_3 D^3$ and $\text{ME}_D = \alpha_1 + 2\alpha_2 D + 3\alpha_3 D^2$. How to compute the ATEM?
- Let $E[y|\mathbf{x}, z] = \exp(\mathbf{x}'\boldsymbol{\beta} + \alpha_1 D + \alpha_2 D^2 + \alpha_3 D^3)$ Then $\text{ME}_D = \ E[y|\mathbf{x}, z] \times (\alpha_1 + 2\alpha_2 D + 3\alpha_3 D^2)$. AMTE is the average of such terms evaluated for each subject.

Marginal TE for regressors in the presence of interaction terms

- Marginal treatment effects in models with interactions are more difficult to interpret and calculate.
- Stata's powerful postestimation `margins` command can be used for linear and nonlinear regression. Example will be given in the practical session.
- Main message: marginal (treatment) effects provide the basis for calculating TEs in nonlinear regression models.
- **Implication**: Having to specify a functional form to estimate TE is potentially a major limitation.

Predictive margins for estimating TEs

Predictive means and predictive margins

- Given estimated regression $\widehat{y} = \mathbf{x}'\widehat{\boldsymbol{\beta}}$, the conditional mean $E[y|\mathbf{x} = \mathbf{x}^*] = \mathbf{x}^*{\prime}\widehat{\boldsymbol{\beta}}$ is called the predictive mean (PM).
- When the dimension of **x** is high, we may want to estimate the PM at specific values of, say x^*, and then contrast these.
- Standard method is to create group-specific predictive means for group-specific contrasts, including contrasts of treated and controls, if one of the x variables is a treatment variable.
- Lane and Nelder (1982) introduced the term predictive margins to cover post estimation prediction of some variable of interest.
- The usefulness and flexibility of PM comes from the fact that it can be evaluated in a variety of ways and the result can be displayed graphically.
- Specifically PM can be used to generate TEs.

TEs using margins

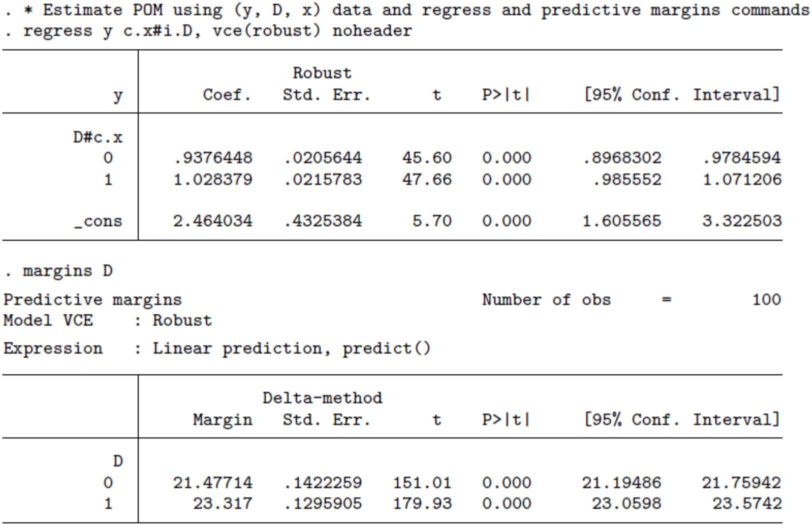
- Given estimated regression $\widehat{y} = \mathbf{x}'\widehat{\boldsymbol{\beta}} + \mathbf{w}'\widehat{\gamma}$, the conditional mean $E[y|\mathbf{x} = \mathbf{x}_k, \mathbf{w} = \mathbf{w}^*] = \mathbf{x}_k\prime\widehat{\boldsymbol{\beta}} + \mathbf{w}^{*\prime}\widehat{\gamma}$
- Different choices of $\mathbf{w}^*$ will generate different PM
- In Stata PMs can be generated postestimation using either the `predict` command or the `margins` command
- `margins` command is very flexible/powerful and can create a variety of contrasts, including PMs for treated and control groups.
- Flexibility comes from being able to use the `at` option to specify the evaluation point.
- Role of `margins plot` in displaying results

Example: TEs using margins

- We revisit the generated data set with a binary treatment variable D.previously analyzed using the `teffects ra` command
- First run the regression $y = \alpha + \beta D \cdot x + \gamma D + \varepsilon$ in which the slope parameter β_D varies according to D (i.e. model allows an interaction effect)
- Next we apply the `margins D` command to generate a table of PMs
- The difference between PMs is the estimated treatment effect.
- The approach can be applied in nonlinear models if no packaged command is available.

Example

using margins.pdf



```
. * Estimate POM using (y, D, x) data and regress and predictive margins commands
. regress y c.x#i.D, vce(robust) noheader
------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       D#c.x |
          0  |   .9376448   .0205644    45.60   0.000     .8968302    .9784594
          1  |   1.028379   .0215783    47.66   0.000      .985552    1.071206
             |
       _cons |   2.464034   .4325384     5.70   0.000     1.605565    3.322503
------------------------------------------------------------------------------

. margins D

Predictive margins                                Number of obs     =        100
Model VCE    : Robust

Expression   : Linear prediction, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |
          0  |   21.47714   .1422259   151.01   0.000     21.19486    21.75942
          1  |     23.317   .1295905   179.93   0.000      23.0598     23.5742
------------------------------------------------------------------------------
```

TE estimation using matched samples

Matching

The essential idea behind matching methods is that treatment effects can be estimated by constructing samples of treated and untreated individuals, closely matched according to specific criteria and then comparing their average outcomes. Matching methods can be applied to RCTs, experimental data, and observational data. The better the match and more balanced the sample, the less biased will be the ATE estimate.

Matching vs. Regression

However, TEs can also be estimated using regression methods applied under same assumptions as matching. So why prefer matching?

1. Regression methods rely on strong functional form assumptions.
2. Matching methods are more robust as they avoid functional form assumptions.
3. Working with a **balanced sample** is key to getting robust estimates and with matching methods this is feasible.
4. With regression methods appropriate **conditioning** is required to mitigate the biases due to unbalanced samples. Omitted variables make this difficult.
5. Regressions typically use all the available data, but a smaller trimmed balanced data set may provide "better" estimates.

Assumptions required to apply matching methods

1. Conditional independence assumption: Conditional on $\mathbf{x}$, outcomes are independent of treatment:. $y_0, y_1 \perp D|\mathbf{x}$. Also known as unconfoundedness assumption or ignorability assumption. Equivalent to regressor exogeneity and no omitted variables.
2. The overlap or matching assumption states that $0 < \Pr[D = 1|\mathbf{x}] < 1$.Means that every unit in the sample has a positive probability of receiving treatment and there are no units which are certain to be treated or to be not treated.
3. Conditional mean independence assumption states $E[y_0|D = 1, \mathbf{x}] = E[\mathbf{y}_0|D = 0, \mathbf{x}] \ = \ E[y_0|\mathbf{x}]$ which means that participation does not depend upon y_0 which should hold in a RCT.

Assumptions

- Assumption 1 is also equivalent to the no-selection-bias-on-observables assumption. That is, conditional on $\mathbf{x}$, D and y are independent.
- Any selection effects that might exist are fully captured by the regressors $\mathbf{x}$.
- Sample balance is not explicitly required as an assumption but has a role if 2 is to hold.
- OLS requires assumptions 1 and 3, but does not explicitly require 2.

Background

1. Identification of treatment effect in observational setting is difficult.
2. Randomized treatment design is a "gold standard" (may be!).
3. Comparing treated and control groups in a way that approximates or mimics randomized treatment is a goal.
4. Matching problem: How to construct suitable control groups?
5. Dehejia & Wahba (DW) study matching methods with emphasis on propensity score (PS) approach using the NSW sample.
6. DW claim: PS methods can produce treatment estimates comparable to those from randomization.

General ATET formula

- Denote the comparison group for the treated case i with characteristics $\mathbf{x}_i$ as $A_j(\mathbf{x}) = \{j | \; \mathbf{x}_j \in c(\mathbf{x}_i)\}$ where $c\left(\mathbf{x}_i\right)$ is the characteristics neighborhood of $\mathbf{x}_i$. Let N_C denote the number of cases in the comparison group and let $w(i,j)$ denote the weight given to the j^{th} case in making a comparison with i^{th} treated case, $\sum_j w(i,j) = 1$. Then a **general formula** for the matching ATET estimator is

$$\Delta^M = \frac{1}{N_T} \sum_{i \in \{D \; = \; 1\}} [y_{1,i} - \sum_j w(i,j) y_{0,j}]$$

where $0 < w(i,j) < 1$, and $\{D = 1\}$ is the set of treated individuals. Different matching estimators are generated by varying the choice of $w(i,j)$.

Simple DW(2002) matching

- Simple matching compares cells with exactly the same discrete $\mathbf{x}$
- $$\Delta^M = \sum_{k \in \{D\ =\ 1\}} w_k [\bar{y}_1 - \bar{y}_0]$$

  where $\bar{y}_1$ is the mean outcome of the treated and $\bar{y}_0$ is the mean outcome of the untreated and w_k is the weight of the k^{th} cell, i.e. the fraction of observations in cell k.
- A specific example (Dehejia and Wahba, 2002) is

  $$\frac{1}{N_T} \sum_i \left(y_i - \frac{1}{N_{C,i}} \sum_{j \in \{D\ =\ 0\}} y_j \right)$$

  where N_T is the number in the treated group $(D = 1)$ and $N_{c,i}$ is the number in the comparison group corresponding to the i^{th} observation.

General matching problem

- Mental experiment about matching methods – treated group is observed w/o a randomized trial
- Need to construct a matched control group – how to proceed?
- Consider the problem of constructing cells with matched occupants within the cell
- $\mathbf{x}_i$ $(i = 1, ...N_T)$ observed vector of characteristics of treated. Find $(y_j^c, \mathbf{x}_j^c,\ j = 1, ..., N_C)$
- For each cell with matched samples compute average cell difference in treated and untreated outcomes
- Issues: (1) $\dim(\mathbf{x})$?$-$ discrete vs. continuous x; (2) Thin cells and empty cell depending on $\dim(\mathbf{x})$; (3) One-one or one-many matching?

Other matching methods (1)

- **Nearest neighbor matching** method: Choose $A_i(\mathbf{x}) = \{j| \ \min_j \|\mathbf{x}_i - \mathbf{x}_j\|\}$ where $\|\|$ denotes the Euclidean distance between vectors. If $w(i,j) = 1$ when $j \in A_i(\mathbf{x})$, and zero otherwise, then this specification uses only one case to construct the comparison group for the treated cases.
- **Kernel matching** is non-parametric; it uses a weighted average of all individuals in the control group with weights given by

$$w(i,j) = \frac{K(\mathbf{x}_j - \mathbf{x}_i)}{\sum_{j=1}^{N_C} K(\mathbf{x}_j - \mathbf{x}_i)},$$

where K is a kernel.

PS matching

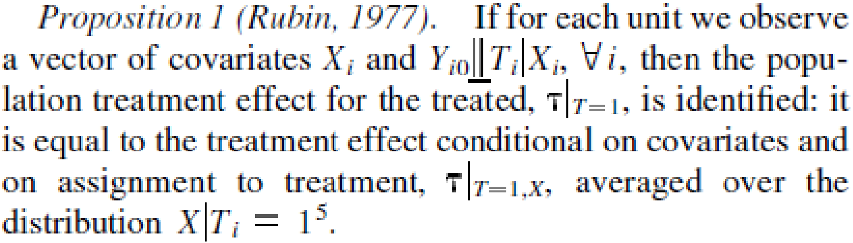
- Let $D = 1$ indicate treatment applied, $D = 0$ the opposite.
- Let $Pr[D = 1|\mathbf{x}]$ denote the conditional probability of receiving treatment, $0 < Pr[D = 1|\mathbf{x}] < 1$
- Propensity score is the estimated conditional probability : $F[D = 1|\mathbf{x}]$ where F denotes parametrically specified c.d.f.
- Standard choices of F are normal c.d.f. $\Phi(\cdot)$ and logistic c.d.f. $\Lambda(\cdot)$
- Propensity score (inexact) matching means constructing a subgroup with similar propensity score, usually based on some interval.

Advantages

1) avoid functional form assumptions for the outcome equations in estimating ATET and
2) can estimate it at specific values of $\mathbf{x}$.
Disadvantage: if $\mathbf{x}$ is high dimensional then the number of matches can become very small.
In such cases **propensity score matching** is better.

- Nearest neighbor and kernel matching can be defined in terms of propensity scores also.
- For example, for nearest neighbor matching we can define $A_i(p(\mathbf{x})) = \{p_j| \ \min_j \|p_i - p_j\|\}$.

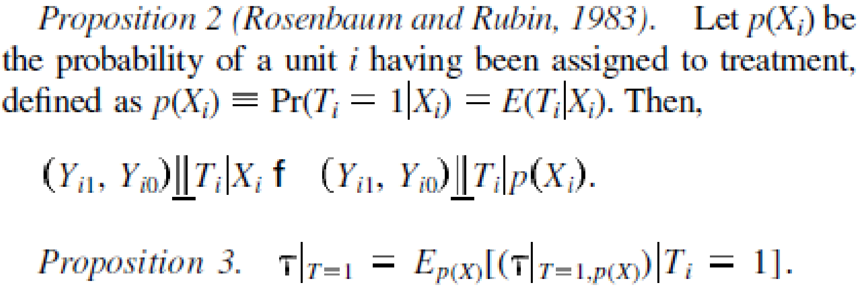
Exact matching on X



Proposition 1 (Rubin, 1977). If for each unit we observe a vector of covariates X_i and $Y_{i0} \perp\!\!\!\perp T_i | X_i$, $\forall i$, then the population treatment effect for the treated, $\tau|_{T=1}$, is identified: it is equal to the treatment effect conditional on covariates and on assignment to treatment, $\tau|_{T=1,X}$, averaged over the distribution $X|T_i = 1$[5].

- Key assumption of Proposition 1 is randomized treatment conditional on observables
- Assumes we have an exhaustive list of matching characteristics

Dimensionality reduction in matching



Proposition 2 (Rosenbaum and Rubin, 1983). Let $p(X_i)$ be the probability of a unit i having been assigned to treatment, defined as $p(X_i) \equiv \Pr(T_i = 1|X_i) = E(T_i|X_i)$. Then,

$$(Y_{i1}, Y_{i0}) \underline{\|} T_i | X_i \; \mathrm{f} \quad (Y_{i1}, Y_{i0}) \underline{\|} T_i | p(X_i).$$

Proposition 3. $\tau|_{T=1} = E_{p(X)}[(\tau|_{T=1,p(X)})|T_i = 1].$

Propensity score matching - properties

- Proposition allows us to reduce the dimension of characteristics,
- permits coexistence of continuous and discrete characteristics
- Introduces a propensity score as unidimensional matching variable
- Standard functional forms used for estimating propensity scores based on $(y_i, \mathbf{x}_i)^T$ and $(y_i, \mathbf{x}_i)^C$
- Matching according to a propensity score involves bracketing: $(y_i, \mathbf{x}_i)^{T,B}$ and $(y_i, \mathbf{x}_i)^{C,B}$
- The overlap condition for identification

TE Methods under matching

Matching methods

- We assume that the assignment mechanism is individualistic, probabilistic and satisfies uncounfoundedness.
- Functional form of the assignment mechanism is not known. Functional form of the outcome is assumed.
- The set-up assumes an active $(D = 1)$ and passive $(D = 0)$ treatment states.
- The standard regression methodology for estimating TEs is valid conditional on functional form assumptions.

PS Methods

- As the probability of receiving treatment may vary across individuals, reweighting the data is an attempt to balance the sample prior to regression.
- This is a variation on RCT but can be expected to yield consistent estimate.
- OLS less attractive because it may not be robust when the treated and untreated samples are unbalanced and do not overlap.
- Two suitable methods of matching are inverse probability weighting and propensity score matching.
- Both methods require initially a model of the conditional probability of receiving treatment.

Inverse Probability Weighting (1)

IPW addresses the problem that sampled individuals do not have the same probability of being treated.

The solution is two-fold.

First estimate the probability of receiving treatment using a logit regression,

Next weight the data before estimating the potential outcome regression(s).

Finally generate POM estimates and TEs given the regression estimates as in the RA case. .

Inverse Probability Weighting (2)

1. Estimate the conditional probability $p(D = 1|\mathbf{x}_i) = \widehat{p}(\mathbf{x}_i)$ of receiving and $(1 - \widehat{p}(\mathbf{x}_i))$ of not receiving treatment, respectively.
Essentially the estimation of **propensity score**, typically from a logit or probit regression of D_i on $\mathbf{x}_i$.
2. Assume that outcome equation has interactions. Then instead of estimating the equation $y_i = \alpha + \mathbf{x}_i' \boldsymbol{\beta} + \gamma D_i + \boldsymbol{\delta}(D_i \cdot \mathbf{x}_i)\delta + u_i$ by OLS, we estimate

$$w_i y_i = \alpha + w_i(\mathbf{x}_i' \boldsymbol{\beta} + \gamma D_i + \boldsymbol{\delta}(D_i \cdot \mathbf{x}_i)) + u_i)$$

where $w_i = 1/\widehat{p}_i(\mathbf{x}_i)$ if $D_{i=1} = 1$ and $w_i = 1/(1 - \widehat{p}_i(\mathbf{x}_i))$ if $D_i = 0$.
3. Using the resulting estimates of $(\alpha, \boldsymbol{\beta}, \gamma, \boldsymbol{\delta})$, generate POM estimates for the treated and untreated groups.

Inverse Probability Weighting (3)

4. For successful implementation want a well-fitting conditional probability model.
5. Method could be computationally unstable if have many observations with $\widehat{p}_i(\mathbf{x}_i)$ close to 0 or 1.
6. A desirable diagnostic is to check the **covariate balance** between the treated and untreated groups (before and after weighting)

Propensity score matching

- PSM is by far the most popular method of matching.
- The essential idea: construct a cell whose occupants constitute a matched control group.
- Given a matched set, average cell difference in treated and untreated outcomes can be computed.
- But the following issues have to be addressed:
    - regressors include both discrete and continuous variables;
    - some cells may be sparse or even empty;
    - whether matching should be one-to-one or one-to-many.
    - larger the number of regressors in the model, the more compelling the issues.

Propensity score matching (2)

- A successful match means that there is at least one untreated subject who matches a treated subject, i.e. a counterfactual exists.
- Apply a matching criterion - a measure of the distance between the treated and untreated subjects.
- A dimensionality problem due to large number of regressors.
- Solution: Replace the regressors by a one-dimensional function of the regressors and use the value of the function to define a match.
- PS is a natural matching criterion.

PSM details

1. Use intelligently saturated logit regression to generate PS
2. Blocking and weighting: Generate matches to obtain a balanced sample with satisfactory overlap.
    - Requires some form of bracketing (or smoothing) to create matched pairs or matched sets.
    - Stratification or interval matching divides the range of variation of the propensity score in intervals.
    - Within each interval, the treated and control units have, on the average, the same propensity score.
    - ATE is the weighted average of these average outcome differences within cells/strata/blocks.

Blocking and weighting (1)

- **Stratification or interval matching** is based on idea of dividing the range of variation of the propensity score in intervals such that within each interval, the treated and control units have, on the average, the same propensity score. ATET is the weighted average of these differences.
- Denote by b the blocks defined over intervals of propensity score. Then the treatment effect within b^{th} block is defined as

$$\text{ATET}_b^S = (N_b^T)^{-1} \sum_{i \in I(b)} Y_{1i} - (N_b^C)^{-1} \sum_{j \in I(b)} Y_{0j}$$

where $I(b)$ is the set of units in block b, N_b^T is the number of treated units in the b^{th} block, and N_0^C is the number of control units in the b^{th} block. Then the treatment effect based on stratification is defined as

$$\text{ATET}^S = \sum_{b=1}^{B} \text{ATET}_b^S \times \left[\sum_{i \in I(b)} D_i \Bigg/ \sum_{\forall i} D_i \right]$$

where the weight for each block is given by the corresponding fraction of treated units and where B is the total number of blocks.

Blocking and weighting (2)

- **Radius matching** in which $A_i(p(\mathbf{x})) = \{p_j| \; \|p_i - p_j\| < r\}$ is based on propensity scores. This means that all control cases with estimated propensity scores falling within radius r are matched to the i^{th} treated case.
- We can express ATET in terms of $p(\mathbf{x})$, assuming the overlap condition $0 < p(\mathbf{x}) < 1$.

$$\begin{aligned}\text{ATET} &= E\left[\frac{(D - p(\mathbf{x}))\, y}{\Pr[D=1]\,(1-p(\mathbf{x}))}\right], \\ \text{ATE} &= E\left[\frac{(D - p(\mathbf{x}))\, y}{p(\mathbf{x})\,(1-p(\mathbf{x}))}\right]\end{aligned}$$

  the last result being due to Dehejia (1997).
- For proof see Cameron and Trivedi (2005, ch. 25.4)

Nearest neighbor matching

- NNM is related to PSM. Previous example was radius matching.
- Create a matched set based on closeness of $(k \times 1)$ vector of regressors $\mathbf{x}_i$ to vector $\mathbf{x}_j$.
- Euclidean distance metric: $||(\mathbf{x}_i - \mathbf{x}_j)'\Omega^{-1}(\mathbf{x}_i - \mathbf{x}_j)||$ where Ω is the $k \times k$ matrix of variances and covariances of elements of x.
- Can specify the required minimum number of matches.
- To generate counterfactual take a weighted average of the outcomes in the reference group.
- If the group size is small may need to make a bias adjustment.
- In Stata the relevant commands are `teffects ipw`, `teffects psmatch`, `teffects nnmatch`
- Examples and analysis of data from a well-known RCT - Oregon Health Insurance Experiment - will be covered in the practical session

Stata's teffects commands

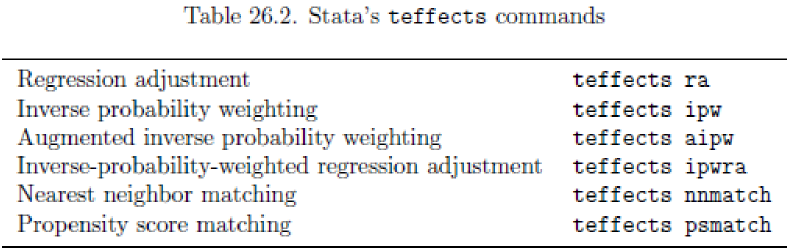


Table 26.2. Stata's teffects commands

Regression adjustment	`teffects ra`
Inverse probability weighting	`teffects ipw`
Augmented inverse probability weighting	`teffects aipw`
Inverse-probability-weighted regression adjustment	`teffects ipwra`
Nearest neighbor matching	`teffects nnmatch`
Propensity score matching	`teffects psmatch`

Differences-in-differences approach

D-i-D approach

- This approach commonly used when evaluating the impact of a shock (change) due to a "natural experiment" (NE)
- NE creates a dichotomy between "before-shock" and "after-shock" data which can be used to make inferences about the impact
- Assume that the variable of interest was moving along some time path and would have continued to do so even in absence of a shock.
- The shock acts as a shifter - the new time path shifts either up or down but otherwise remains parallel to the "before-shock" path.
- Object of interest is the estimated size of the shift.
- Observations in the pre-shock period act as control outcomes, and those after shock are treated outcomes

John Snow cholera data 1849/1854

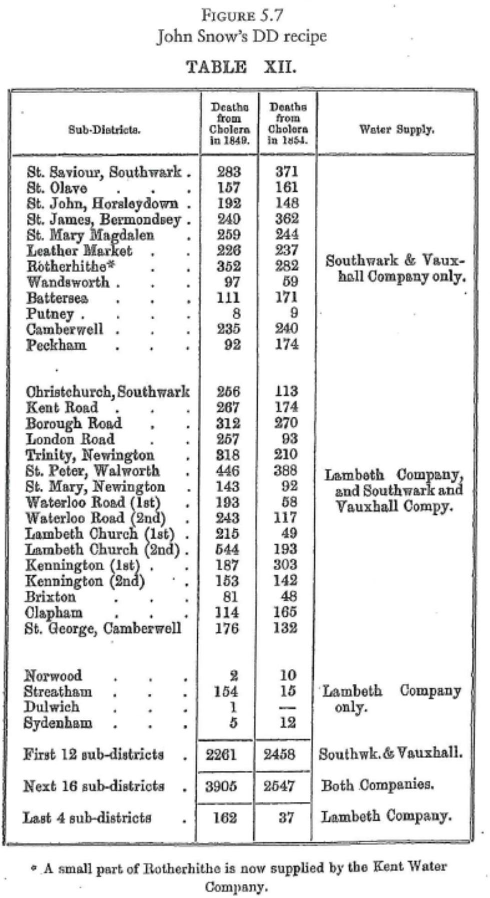


FIGURE 5.7
John Snow's DD recipe

TABLE XII.

Sub-Districts.	Deaths from Cholera in 1849.	Deaths from Cholera in 1854.	Water Supply.
St. Saviour, Southwark	283	371	Southwark & Vauxhall Company only.
St. Olave	157	161	
St. John, Horsleydown	192	148	
St. James, Bermondsey	240	362	
St. Mary Magdalen	259	244	
Leather Market	226	237	
Rotherhithe*	352	282	
Wandsworth	97	59	
Battersea	111	171	
Putney	8	9	
Camberwell	235	240	
Peckham	92	174	
Christchurch, Southwark	256	113	Lambeth Company, and Southwark and Vauxhall Compy.
Kent Road	267	174	
Borough Road	312	270	
London Road	257	93	
Trinity, Newington	318	210	
St. Peter, Walworth	446	388	
St. Mary, Newington	143	92	
Waterloo Road (1st)	193	58	
Waterloo Road (2nd)	243	117	
Lambeth Church (1st)	215	49	
Lambeth Church (2nd)	544	193	
Kennington (1st)	187	303	
Kennington (2nd)	153	142	
Brixton	81	48	
Clapham	114	165	
St. George, Camberwell	176	132	
Norwood	2	10	Lambeth Company only.
Streatham	154	15	
Dulwich	1	—	
Sydenham	5	12	
First 12 sub-districts	2261	2458	Southwk. & Vauxhall.
Next 16 sub-districts	3905	2547	Both Companies.
Last 4 sub-districts	162	37	Lambeth Company.

* A small part of Rotherhithe is now supplied by the Kent Water Company.

D-i-D transformation (1)

- For i^{th} treated case the change in the outcome is $[y_{ia} - y_{ib} | D_{ia} = 1]$
- for the untreated group the change is $[y_{ia} - y_{ib} | D_{ia} = 0]$.
- Then the difference is $[y_{ia} - y_{ib} | D_{ia} = 1] - [y_{ia} - y_{ib} | D_{ia} = 0]$ where subscripts a and b denote "after" and "before"
- (1) $y_{it,b} = \phi_i + \delta_t + \varepsilon_{it}$; (2) $y_{it,a} = y_{it,b} + \alpha + \varepsilon_{it} \equiv \phi_i + \delta_t + \alpha D_{it} + \varepsilon_{it}$
- Then $E[y_{ia} - y_{ib} | D_{ia} = 1] - E[y_{ia} - y_{ib} | D_{ia} = 0] = \alpha$: the ATE.

D-i-D transformation (2)

- The underlying assumption is that there is a separable trend path $\{\delta_t\}$ that is common to both treated and control groups.
- Any time-invariant factors would be eliminated by the differencing transformation.
- If there are other time-varying factors, then D-i-D will end up with a regression, not a constant. Example which follows shows this.
- The data framework may be complicated, e.g. a panel consisting of clusters with multiple treatments. Then group effects and time-effects will need to be added, as in panel data models.

D-i-D transformation (3)

- Regression adjustment is an alternative to taking differences. Replace ϕ_i by $\mathbf{x}_i'\boldsymbol{\beta} + \gamma y_{ib}$ to obtain

$$\begin{aligned} y_{ia,0} &= \mathbf{x}_i'\boldsymbol{\beta} + \gamma y_{ib} + \delta_a + \varepsilon_{ia,0} \\ y_{ia,1} &= \mathbf{x}_i'\boldsymbol{\beta} + \gamma y_{ib} + \delta_a + \alpha D_{ia} + \varepsilon_{ia,1}. \end{aligned}$$

- Estimate α by regressing $y_{ia,1}$ on a constant, $y_{ia,0}$, $\mathbf{x}_i$ and D_{ia}.
- No assumption like support or overlap condition is required
- Transformation is also applied to the error term which induces serial correlation (MA-1) - a problem if time series is long
- Using default estimator of variance matrix will overstate the precision of the estimator.
- Instead should use a robust sandwich variance estimator.

Bertrand, M., E Duflo, and S. Mullainathan. "How much should we trust differences-in-differences estimates?." QJE 119.1 (2004): 249-275.

DID assumptions

- α is a causal parameter because after controlling for $\mathbf{x}$, and y_b TE completely accounts for the posttreatment difference between the treated and control groups. Further, the fixed effect is given a linear functional form.
- Assumes addition of pre-treatment data is feasible
- But a matching strategy can be based on weaker assumptions.
- By assumption the same drift term both before and after.
- No heterogeneity in response

Consequences of differencing

DID in a nonlinear model

- Consider the probit model of outcome, which is nonlinear. T denotes time period, say 0 or 1; G denotes group, say 0 or 1.
- $T \times G = 1$ for the treated group
- untreated group $E[y^0|T, G, \mathbf{x}] = \Phi[\beta_T T + \beta_G G + \mathbf{x}'\boldsymbol{\beta}_x]$
- treated group $E[y^1|T, G, \mathbf{x}] = \Phi[\beta_T T + \beta_G G + \alpha(T \times G) + \mathbf{x}'\boldsymbol{\beta}_x]$

$$\tau = \Phi[\beta_T + \beta_G + \alpha + \mathbf{x}'\boldsymbol{\beta}_x] - \Phi[\beta_T + \beta_G + \mathbf{x}'\boldsymbol{\beta}_x]$$

which measures the change in the probability due to the treatment.

- In principle this result applies to any case with a nonlinear strictly monotonic transformation, e..g., quantile.
- DID now better labelled as change-in-change, CIC.

Ref: Athey and Imbens, Econometrica, 2006, 431-497

Example 1: Impact of training on wages (Dehejia and Wahba)

Application of D-i-D

Example 1: Effect of Training on Earnings

- The National Supported Work (NSW) demonstration project, conducted in the 1970's, measured the impact of training on earnings by a randomized experiment with a treatment group and a control group.
- The effect of training could then be measured by direct comparison of sample means.
- Comparison of the treated with the nontreated must then control for differences in observed characteristics, and possibly in unobserved characteristics.
- Lalonde (1986) contrasted outcomes for the NSW treated group with those for control groups drawn from two national surveys. He concluded that the observational methods were unreliable.
- Dehejia and Wahba (1999; 2002) reanalyzed a subset of the Lalonde data using alternative matching methods that they argued led to conclusions closer to those from experimental data.

Dehejia and Wahba Data

- Treated sample is one of 185 males who received training during 1976-77.
- Control group: 2,490 male household heads under the age of 55 who are not retired, drawn from the Panel Survey of Income Dynamics (PSID).
- Another comparison group is from the CPS.
- Dehejia and Wahba (1999) call these two samples the RE74 subsample (of the NSW treated) and the PSID-1 sample (of nontreated).

Summary statistics

Variable	Definition	Treated	PSID Control
AGE	age in years	25.82	34.85
EDUC	education in years	10.35	12.12
NODEGREE	1 if EDUC $<$ 12	0.71	0.31
BLACK	1 if race is black	0.84	0.25
HISP	1 if Hispanic	0.06	0.03
MARR	1 if married	0.19	0.87
U74	1 if unemployed in 1974	0.60	0.10
U75	1 if unemployed in 1975	0.71	0.09
RE74	real earnings in 1974 (in 1982 $)	2,096	19,429
RE75	real earnings in 1975 (in 1982 $)	1,532	19,063
RE78	real earnings in 1978 (in 1982 $)	6,349	21,554
D	1 if received training (treatment)	1.00	0.00
Sample size		185	2,490

Comparison of treated and control groups

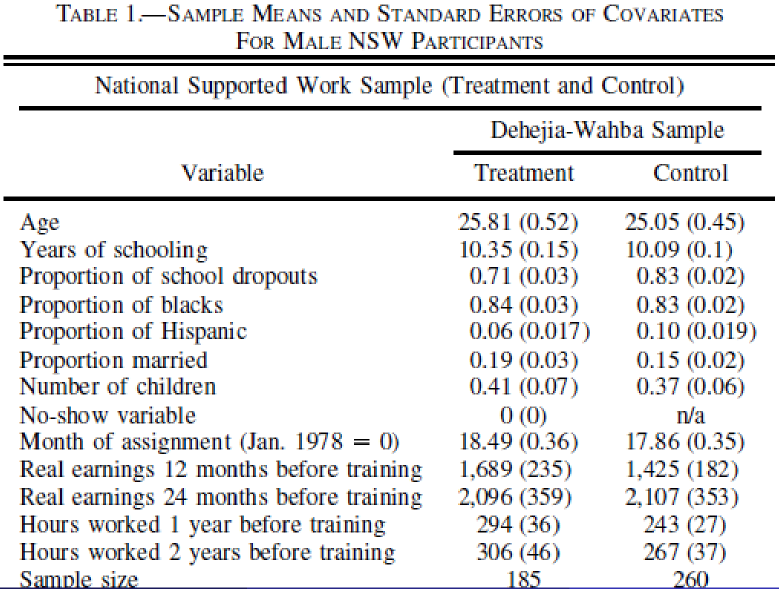


TABLE 1.—SAMPLE MEANS AND STANDARD ERRORS OF COVARIATES FOR MALE NSW PARTICIPANTS

National Supported Work Sample (Treatment and Control)

	Dehejia-Wahba Sample	
Variable	Treatment	Control
Age	25.81 (0.52)	25.05 (0.45)
Years of schooling	10.35 (0.15)	10.09 (0.1)
Proportion of school dropouts	0.71 (0.03)	0.83 (0.02)
Proportion of blacks	0.84 (0.03)	0.83 (0.02)
Proportion of Hispanic	0.06 (0.017)	0.10 (0.019)
Proportion married	0.19 (0.03)	0.15 (0.02)
Number of children	0.41 (0.07)	0.37 (0.06)
No-show variable	0 (0)	n/a
Month of assignment (Jan. 1978 = 0)	18.49 (0.36)	17.86 (0.35)
Real earnings 12 months before training	1,689 (235)	1,425 (182)
Real earnings 24 months before training	2,096 (359)	2,107 (353)
Hours worked 1 year before training	294 (36)	243 (27)
Hours worked 2 years before training	306 (46)	267 (37)
Sample size	185	260

Comparisons

- Treated group differs considerably from the control group.
- Disproportionately black (84 percent) with less than high school degree (71 percent) and unemployed in the pre-treatment year 1975 (71 percent). Estimates of the effect of training should control for these differences.
- The outcome of interest is post-treatment earnings, RE78.
- One possible measure = mean difference in RE78 between treated and control individuals, leading to estimate $\$6,349 - \$21,554 = -\$15,205$. This is called a **treatment-control comparison** estimator.
- It can equivalently be computed as the coefficient of the treatment indicator D in OLS regression of RE78 on an intercept and D using a combined treatment-control sample.

Comparisons (2)

- Treatment estimate is misleading; mostly reflects the difference in the types of individuals in the two samples
- To control for this difference include pre-treatment characteristics as regressors, and estimate by OLS

$$\mathrm{RE78}_i = \mathbf{x}_i'\boldsymbol{\beta} + \alpha D_i + u_i, \quad i = 1, ..., 2675.$$

- Leads to much smaller estimated treatment effect $\widehat{\alpha} = \$218$ when, following Dehejia and Wahba, the regressors $\mathbf{x}$ are specified to be an intercept, AGE, AGESQ, EDUC, NODEGREE, BLACK, HISP, RE74 and RE75.
- This approach is called the **control function** or **regression adjusted** estimator.

Estimated effects

Method	Definition	Estimate	St. Error
Treatment-control comparison	$\overline{\text{RE78}}_{D=1} - \overline{\text{RE78}}_{D=0}$	$-15{,}205$	656
Control function estimator	$\widehat{\alpha}$ from OLS regression 1	218	768
Before-after comparison	$\overline{\text{RE78}}_{D=1} - \overline{\text{RE75}}_{D=1}$	4,817	625
Differences-in-differences	$\widehat{\alpha}$ from OLS regression 2	2,326	749
Propensity score	See text	994	—

Note: Standard errors for first four estimates are computed using heteroskedastic-consistent standard errors from the appropriate OLS regression.

Differences-in-Differences

- **Before-after (BA) comparison** looks at the difference between post-treatment earnings RE78 and pre-treatment earnings RE75. Using mean earnings for the treated group this yields estimate $\$6,349 - \$1,532 = \$4,817$.
- This estimate may be misleading as it reflects all changes over this time period, such as an improved economy, and not just training.
- **Difference-in-differences (DID) estimator** additionally calculates a similar quantity for the control group, $\$21,554 - \$19,063 = \$2,491$, and uses this as a measure of non-treatment related changes over time in earnings, so that the change over time solely due to treatment is $\$4,817 - \$2,491 = \$2,326$.

Differences-in-Differences (2)

- DID estimator equivalent to the estimate of α in the regression

  $$\mathrm{RE}_{it} = \phi + \delta \mathrm{D78}_{it} + \gamma \alpha \mathrm{D}_{it} + \alpha \mathrm{D78}_{it} * \mathrm{D}_{it} + u_i, \quad i = 1, ..., 2675, \;\; t = 75,$$

  Here $\mathrm{RE}_{i,75}$ denotes earnings in the pre-treatment period and $\mathrm{RE}_{i,78}$ denotes earnings in the post-treatment period, so the regression is one with $5,350$ earnings observations.

- Indicator variable $\mathrm{D78}_{it}$ equals one in the post-treatment period, the indicator variable D_{it} equals one if the individual is in the treated sample, and the interaction term $\mathrm{D78}_{it} * D_{it}$ equals one for treated individuals in the post-treatment period.
- Intercept ϕ can be replaced by $\mathbf{x}_{it}'\boldsymbol{\beta}$. This makes no difference in this example where regressors are time-invariant so that $\mathbf{x}_{it} = \mathbf{x}_i$.

Simple Propensity Score Estimate

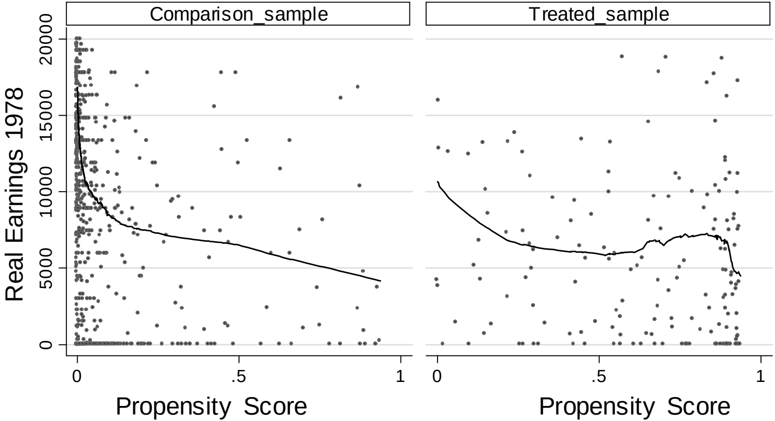
- A third approach compares the outcome RE78 uses a better counterfactual.
- Generated by specifying a regression model. For example, the regression specifies $E[\text{RE78}|\mathbf{x}]$ to equal $\mathbf{x}'\boldsymbol{\beta}+\boldsymbol{\alpha}$, if treated, with counterfactual $\mathbf{x}'\boldsymbol{\beta}$, if not treated. This places restrictions on both the effect of regressors $\mathbf{x}$ and on the effect of treatment which, conditional on $\mathbf{x}$, is assumed to be constant across individuals.
- **Match on the propensity score**, defined as the conditional probability of treatment $\Pr[D=1|\mathbf{x}]$. For this example we estimate using only 1975 data the logit model

$$\Pr[D_i = 1|\mathbf{x}_i] = \Lambda(\mathbf{x}_i'\boldsymbol{\beta}), \quad i = 1, ..., 2675, \tag{1}$$

where $\Lambda(z) = e^z/(1+e^z)$, and following Dehejia and Wahba (1999) the regressors chosen are AGE, AGESQ, EDUC, EDUCSQ, NODEGREE, BLACK, HISP, MARR, RE74, RE75, RE74SQ, RE75SQ, U74*BLACK.

PS Graph-NSW Data

Post-treatment Earnings against Propensity Score



[redacted]

Graphs by Treatment Status

PS Graph

- Treatment effect is estimated as the difference between a given treated individual $(D = 1)$ and control sample individual $(D = 0)$ with the same (predicted) propensity score.
- Each panel includes a fitted nonparametric regression of RE78 on the propensity score.
- Treatment effect is generally less than one thousand dollars, though is large and positive for propensity score around 0.80.

PS Estimators

- Many ways to compare individuals with similar propensity score and then averaging over all treated individuals.
- A simple strategy is to stratify data by propensity score, denoted $p(\mathbf{x})$, and let the counterfactual be the within-strata average of RE78 for the control group. For example, if a treated observation has propensity score $p(\mathbf{x}) =$0.35 then the counterfactual may be the average of $p(\mathbf{x})$ for control group observations with $0.30 < p(\mathbf{x}) \leq 0.40$.
- Total effect is then $\sum_s w_s(\overline{\text{RE78}}_{s,D=1} - \overline{\text{RE78}}_{s,D=0})$, where $\overline{\text{RE78}}_{s,D=1}$ and $\overline{\text{RE78}}_{s,D=0}$ denote, respectively, the strata s averages of RE78 for the treated and control observations, and the weights w_s equal the fraction of treated observations in each strata.

Stratification matching

- Stratification matching : use ten equally-spaced strata with $0.0 < p(\mathbf{x}) \leq 0.1$, $0.1 < p(\mathbf{x}) \leq 0.2$ and so on. Restrict this procedure to cases where the propensity scores for the treated and control samples overlap. Here the propensity score ranges from 0.0005 to 0.9420 for the treated sample and from 0.0000 to 0.9371, leading to dropping of $1,423$ control group individuals and 8 treated individuals. The resulting estimated total effect is \$995.

Comparison with the CPS control group

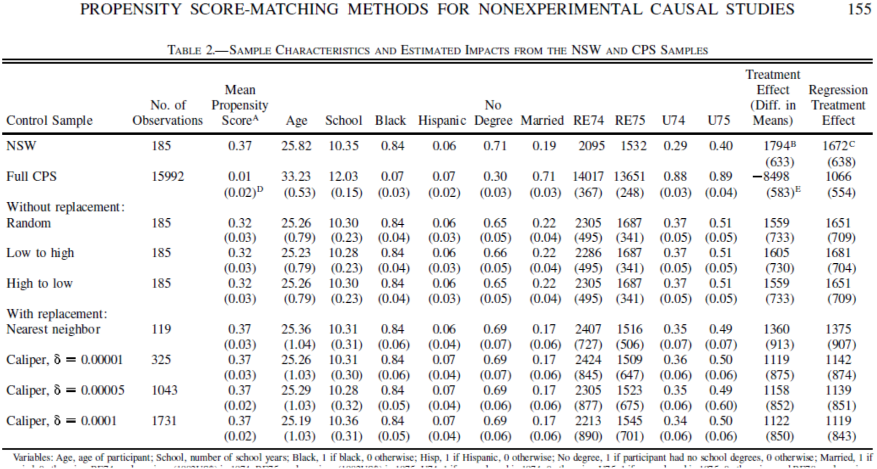


TABLE 2.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND CPS SAMPLES

Control Sample	No. of Observations	Mean Propensity Score^A	Age	School	Black	Hispanic	No Degree	Married	RE74	RE75	U74	U75	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794^B	1672^C
													(633)	(638)
Full CPS	15992	0.01	33.23	12.03	0.07	0.07	0.30	0.71	14017	13651	0.88	0.89	−8498	1066
		(0.02)^D	(0.53)	(0.15)	(0.03)	(0.02)	(0.03)	(0.03)	(367)	(248)	(0.03)	(0.04)	(583)^E	(554)
Without replacement:														
Random	185	0.32	25.26	10.30	0.84	0.06	0.65	0.22	2305	1687	0.37	0.51	1559	1651
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)
Low to high	185	0.32	25.23	10.28	0.84	0.06	0.66	0.22	2286	1687	0.37	0.51	1605	1681
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(730)	(704)
High to low	185	0.32	25.26	10.30	0.84	0.06	0.65	0.22	2305	1687	0.37	0.51	1559	1651
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)
With replacement:														
Nearest neighbor	119	0.37	25.36	10.31	0.84	0.06	0.69	0.17	2407	1516	0.35	0.49	1360	1375
		(0.03)	(1.04)	(0.31)	(0.06)	(0.04)	(0.07)	(0.06)	(727)	(506)	(0.07)	(0.07)	(913)	(907)
Caliper, $\delta = 0.00001$	325	0.37	25.26	10.31	0.84	0.07	0.69	0.17	2424	1509	0.36	0.50	1119	1142
		(0.03)	(1.03)	(0.30)	(0.06)	(0.04)	(0.07)	(0.06)	(845)	(647)	(0.06)	(0.06)	(875)	(874)
Caliper, $\delta = 0.00005$	1043	0.37	25.29	10.28	0.84	0.07	0.69	0.17	2305	1523	0.35	0.49	1158	1139
		(0.02)	(1.03)	(0.32)	(0.05)	(0.04)	(0.06)	(0.06)	(877)	(675)	(0.06)	(0.60)	(852)	(851)
Caliper, $\delta = 0.0001$	1731	0.37	25.19	10.36	0.84	0.07	0.69	0.17	2213	1545	0.34	0.50	1122	1119
		(0.02)	(1.03)	(0.31)	(0.05)	(0.04)	(0.06)	(0.06)	(890)	(701)	(0.06)	(0.06)	(850)	(843)

Variables: Age, age of participant; School, number of school years; Black, 1 if black, 0 otherwise; Hisp, 1 if Hispanic, 0 otherwise; No degree, 1 if participant had no school degrees, 0 otherwise; Married, 1 if

Comparison with the PSID control group

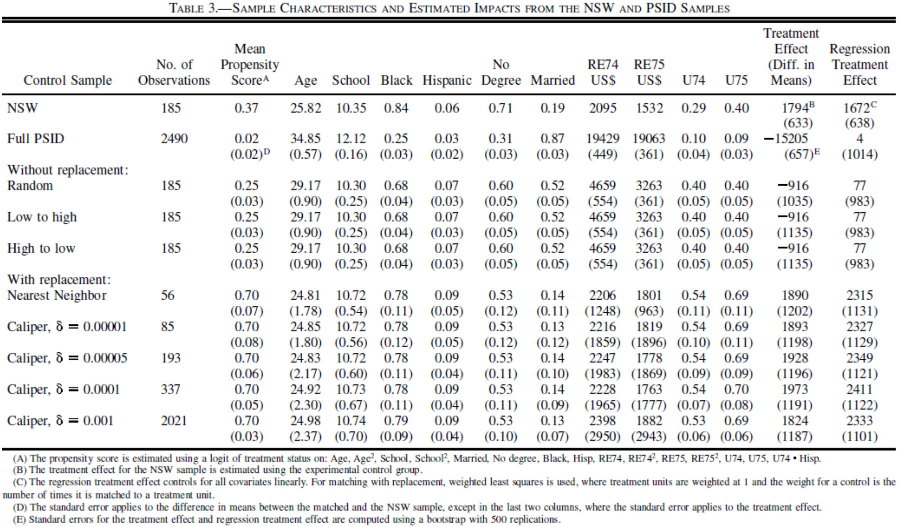


TABLE 3.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND PSID SAMPLES

Control Sample	No. of Observations	Mean Propensity Score[A]	Age	School	Black	Hispanic	No Degree	Married	RE74 US$	RE75 US$	U74	U75	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794[B]	1672[C]
													(633)	(638)
Full PSID	2490	0.02	34.85	12.12	0.25	0.03	0.31	0.87	19429	19063	0.10	0.09	−15205	4
		(0.02)[D]	(0.57)	(0.16)	(0.03)	(0.02)	(0.03)	(0.03)	(449)	(361)	(0.04)	(0.03)	(657)[E]	(1014)
Without replacement:														
Random	185	0.25	29.17	10.30	0.68	0.07	0.60	0.52	4659	3263	0.40	0.40	−916	77
		(0.03)	(0.90)	(0.25)	(0.04)	(0.03)	(0.05)	(0.05)	(554)	(361)	(0.05)	(0.05)	(1035)	(983)
Low to high	185	0.25	29.17	10.30	0.68	0.07	0.60	0.52	4659	3263	0.40	0.40	−916	77
		(0.03)	(0.90)	(0.25)	(0.04)	(0.03)	(0.05)	(0.05)	(554)	(361)	(0.05)	(0.05)	(1135)	(983)
High to low	185	0.25	29.17	10.30	0.68	0.07	0.60	0.52	4659	3263	0.40	0.40	−916	77
		(0.03)	(0.90)	(0.25)	(0.04)	(0.03)	(0.05)	(0.05)	(554)	(361)	(0.05)	(0.05)	(1135)	(983)
With replacement:														
Nearest Neighbor	56	0.70	24.81	10.72	0.78	0.09	0.53	0.14	2206	1801	0.54	0.69	1890	2315
		(0.07)	(1.78)	(0.54)	(0.11)	(0.05)	(0.12)	(0.11)	(1248)	(963)	(0.11)	(0.11)	(1202)	(1131)
Caliper, $\delta = 0.00001$	85	0.70	24.85	10.72	0.78	0.09	0.53	0.13	2216	1819	0.54	0.69	1893	2327
		(0.08)	(1.80)	(0.56)	(0.12)	(0.05)	(0.12)	(0.12)	(1859)	(1896)	(0.10)	(0.11)	(1198)	(1129)
Caliper, $\delta = 0.00005$	193	0.70	24.83	10.72	0.78	0.09	0.53	0.14	2247	1778	0.54	0.69	1928	2349
		(0.06)	(2.17)	(0.60)	(0.11)	(0.04)	(0.11)	(0.10)	(1983)	(1869)	(0.09)	(0.09)	(1196)	(1121)
Caliper, $\delta = 0.0001$	337	0.70	24.92	10.73	0.78	0.09	0.53	0.14	2228	1763	0.54	0.70	1973	2411
		(0.05)	(2.30)	(0.67)	(0.11)	(0.04)	(0.11)	(0.09)	(1965)	(1777)	(0.07)	(0.08)	(1191)	(1122)
Caliper, $\delta = 0.001$	2021	0.70	24.98	10.74	0.79	0.09	0.53	0.13	2398	1882	0.53	0.69	1824	2333
		(0.03)	(2.37)	(0.70)	(0.09)	(0.04)	(0.10)	(0.07)	(2950)	(2943)	(0.06)	(0.06)	(1187)	(1101)

(A) The propensity score is estimated using a logit of treatment status on: Age, Age^2, School, School^2, Married, No degree, Black, Hisp, RE74, RE74^2, RE75, RE75^2, U74, U75, U74 ▪ Hisp.

(B) The treatment effect for the NSW sample is estimated using the experimental control group.

(C) The regression treatment effect controls for all covariates linearly. For matching with replacement, weighted least squares is used, where treatment units are weighted at 1 and the weight for a control is the number of times it is matched to a treatment unit.

(D) The standard error applies to the difference in means between the matched and the NSW sample, except in the last two columns, where the standard error applies to the treatment effect.

(E) Standard errors for the treatment effect and regression treatment effect are computed using a bootstrap with 500 replications.

Matching Using Propensity Scores

- Fitted Propensity Score: Obtained using two different logit specifications, from DW (1999) and DW (2002) respectively.

Matching Algorithms and Balancing: DW algorithm for matching propensity scores.

- Start with a parsimonious logit model to estimate $p(\mathbf{x})$.

1. Sort data according to $\widehat{p}(\mathbf{x})$. Initially a rough grid with equal ranges may be used. The sample observations are stratified such that within a stratum the $\widehat{p}(\mathbf{x})$ for treated and control units are close.
2. Within each stratum test for the equality of means between treated and control units for each covariate. Regressors are balanced if there is no statistically significant difference.
3. If, for some stratum, there is no balance, then for the **unbalanced stratum** use a finer grid to achieve balance.
4. If there are many unbalanced strata, then the original logit model is reestimated with an improved specification that includes interaction and higher order terms among the regressors.

Matched PS

Minimum $\widehat{p}(x)$	Treated	Untreated	Total
0.000364	9	960	969
0.10	10	56	66
0.20	14	33	47
0.40	24	22	46
0.60	33	7	40
0.80	95	8	103
Total	185	1086	1271

Note: From the second row, for example, the propensity score lies between 0.10 and 0.20 for 10 treated and 56 untreated individuals.

NSW - CPS comparison

Figure 1.—Histogram of Estimated Propensity Score, NSW and CPS

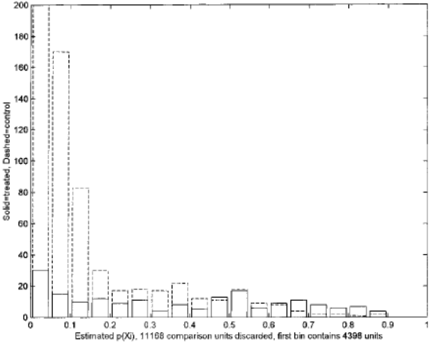
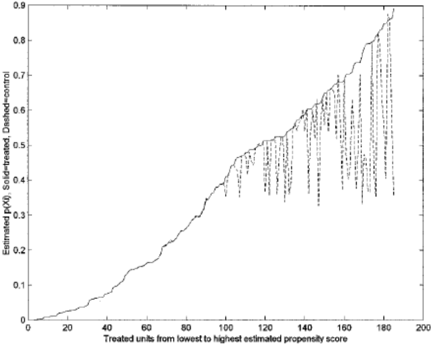


Figure 3.—Propensity Score for Treated and Matched Comparison Units, Random Without Replacement



NSW - PSID Comparison

Figure 2.—Histogram of Estimated Propensity Score, NSW and PSID

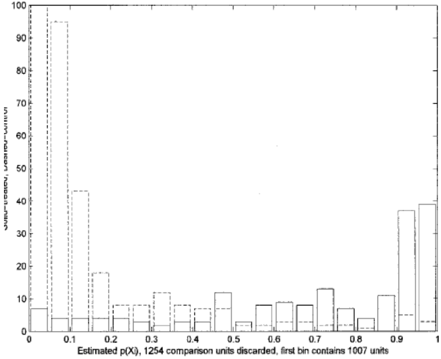


Figure 4.—Propensity Score for Treated and Matched Comparison Units, Lowest to Highest

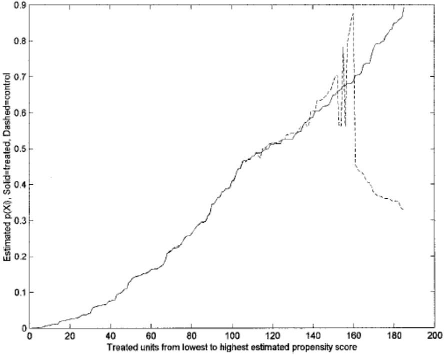


Figure 5.—Propensity Score for Treated and Matched Comparison Units, Highest to Lowest

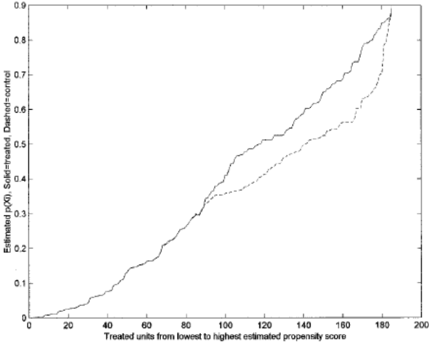
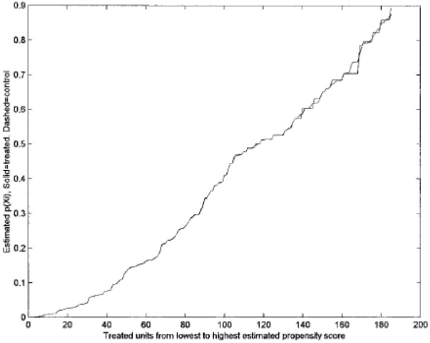


Figure 6.—Propensity Score for Treated and Matched Comparison Units, Nearest Match



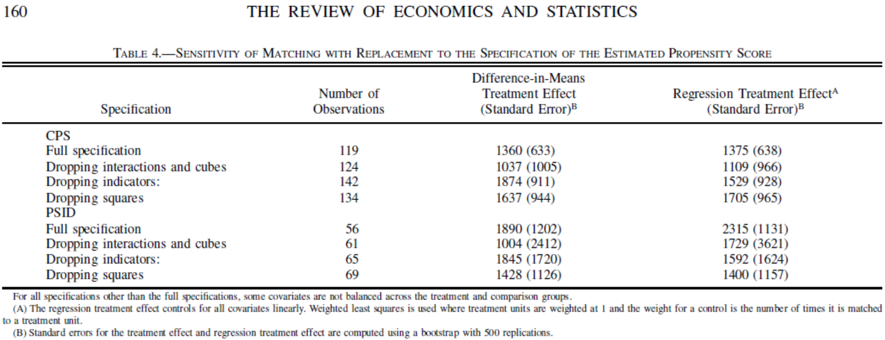
Matched PS (II)

- Above PS computation has been restricted to the common support region by testing the balancing property using those observations whose propensity scores lie in the intersection of the supports of the propensity score of the treated and the control units.
- This restriction reduces the original sample significantly. The size of the control group drops down from 2490 units to 1086 for the DW (2002) specification.
- Results differ from DW (2002) because the latter exclude control units from NSW-PSID composite samples not on the basis of common support region but on the basis of whether the estimated propensity score of a sample unit is less than the minimum of the estimated propensity score for the treated units.

ATET Results (1)

Matching	#	#	ATET	Std.Err	% of
procedure	treated	control			$1794
DW (2002) specification[a]					
Nearest neighbor	185	53	2385	1209[c]	133
Radius, $r = 0.001$	54	517	-7815	1118[d]	-436
Radius, $r = 0.0001$	24	92	-9333	2282[d]	-520
Radius, $r = 0.00001$	15	19	-2200	2986[d]	-123
Stratification	185	1086	1452	1041[c]	81
Kernel	185	1058	1309	975[c]	73

Sensitivity analysis



160 THE REVIEW OF ECONOMICS AND STATISTICS

TABLE 4.—SENSITIVITY OF MATCHING WITH REPLACEMENT TO THE SPECIFICATION OF THE ESTIMATED PROPENSITY SCORE

Specification	Number of Observations	Difference-in-Means Treatment Effect (Standard Error)^B	Regression Treatment Effect^A (Standard Error)^B
CPS			
Full specification	119	1360 (633)	1375 (638)
Dropping interactions and cubes	124	1037 (1005)	1109 (966)
Dropping indicators:	142	1874 (911)	1529 (928)
Dropping squares	134	1637 (944)	1705 (965)
PSID			
Full specification	56	1890 (1202)	2315 (1131)
Dropping interactions and cubes	61	1004 (2412)	1729 (3621)
Dropping indicators:	65	1845 (1720)	1592 (1624)
Dropping squares	69	1428 (1126)	1400 (1157)

For all specifications other than the full specifications, some covariates are not balanced across the treatment and comparison groups.

(A) The regression treatment effect controls for all covariates linearly. Weighted least squares is used where treatment units are weighted at 1 and the weight for a control is the number of times it is matched to a treatment unit.

(B) Standard errors for the treatment effect and regression treatment effect are computed using a bootstrap with 500 replications.

ATET Results (2)

A selection of results for various matching methods are summarized above. The nearest neighbor estimate of ATET for the DW (2002) specification is \$2385 and for the DW (1999) specification, it is approximately at \$560. The performance of stratification and kernel matching is also mixed, ranging from \$1452 to \$2156.

The benchmark estimate of the treatment effect is \$1794; obtained by regressing $RE78$ on D for the DW (2002) version of the NSW sample of both participants and non-participants.

For the DW (2002) specification, the nearest neighbor estimator is very close to the benchmark estimate and is even better than DW (2002) in terms of reduced bias.

For stratification and kernel estimates, the bias is larger. For the radius matching estimator, this bias is worse which gives negative estimates of the treatment effect as opposed to the positive estimates that DW (2002) found using caliper matching.

Conclusions

1. PS methods approximate Lalonde's benchmark estimates well.
2. the choice of the matching algorithm becomes important after "irrelevant comparison units" have been discarded.
3. "Selection-only-on-observables" is an important and strong assumption.

Comment from Smith and Todd (JoE, 2005)

- DW found low bias in applying PSM , but some have expressed skepticism of this finding.
- Several studies of Heckman and coauthors conclude that the following conditions should hold for low bias to be achieved.
    1. Should include a rich set of variables related to program participation and labor market outcomes
    2. Comparison group should be drawn from the same local labor market as the participants
    3. Outcome variable should be measured in the same way for treated and comparison groups.
- These conditions are not met in the DW studies.
- Smith and Todd show that DW results are sensitive to their choice of subsample of LaLonde data.

OHIE background

- OHIE an important modern example of RCT or a social experiment in the tradition of the famous RHIE.
- Background: See Finkelstein et al. (2012) and Baicker et al. (2013).
- Here we provide only the essential details for interpreting the application that follows.
- At the time of the experiment the Oregon Health Program (OHP) was separated into two components
- OHP Plus, which served the categorically eligible Medicaid population, and
- (OHP) Standard, an expansion program targeting low-income uninsured adults, ineligible for OHP Plus.

OHIE background (2)

- Due to budgetary constraints OHP Standard was wound back and closed to new applicants in 2004, leading to significant attrition over

the following four years.

- After 80 per cent decline in enrollments, in January of 2008 the state determined to expand the program by an additional 10,000 positions.
- Anticipating excess demand the OHP sought and received permission to assign **selection by lottery.**
- The RCT provides an opportunity to assess the impact of expanded health insurance coverage on a variety of health and financial outcomes within

RCT design framework.

OHIE background (3)

- Lottery enrollment (February to March 2008) some 90,000 individuals
- Over the next 6 months the government conducted eight waves of lottery draws resulting in some 35,000 individuals being offered the opportunity to apply for OHP coverage.
- Opportunity to apply was extended to all members of the selected individual's household, thus selection was random conditional on the number of household members in the lottery list.
- Approximately 35,000 individuals from 30,000 unique households were selected, and of those approximately 30 per

cent were eligible and enrolled by the given deadlines.

OHIE background (4)

- Following the treatment, researchers tracked lottery participant outcomes over the next 12 months with three mail surveys.
- Examples consider the third of these mail surveys, which was undertaken in seven mail out waves approximately 12 months after treatment (July and August 2009).
- Nearly all individuals selected in the lottery as well as an approximately equal number of non-selected individuals were mailed questionnaires regarding health care needs, experiences and costs over the previous 6 months.
- Following an intensive follow-up protocol undertaken on a subset of non-responders, the researchers achieved an estimated response rate of approximately 50 per cent.

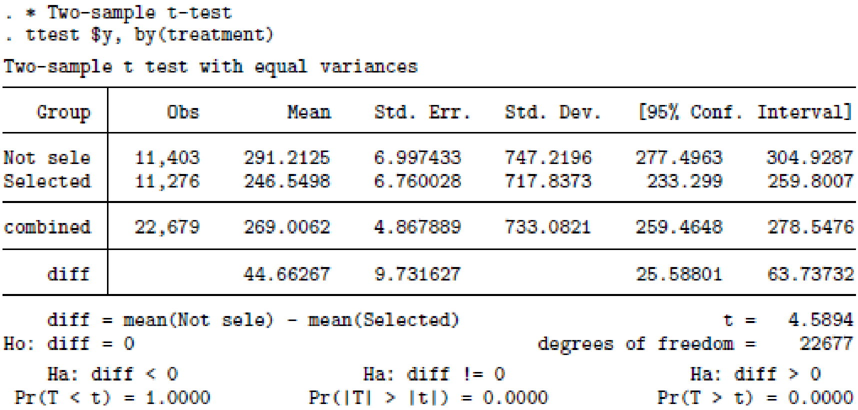
OHIE background (5)

- Examples measure the impact of expanded health coverage on out of pocket medical expensitures over a 12 month period.
- Include indicator variables capturing household size and survey wave to control for potential correlation with the probability of treatment.
- Include a set of relevant covariates to improve efficiency – smoking status, income as a percentage of the federal poverty line, education level and employment

Example 2: Oregon Health Insurance Experiment

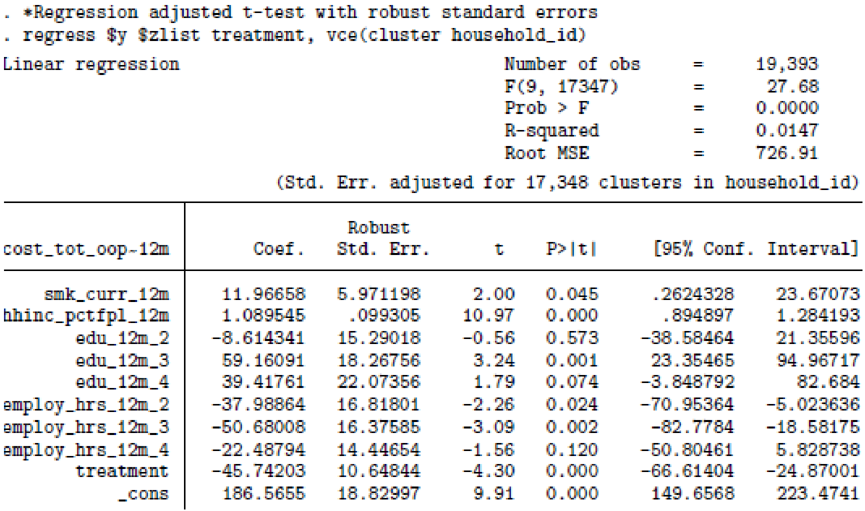
OHIE RCT Data

- The OHIE can be viewed as an important modern example of RCT or a social experiment.
- The background to this experiment has been covered in detail elsewhere.
- The size and complexity of the OHIE data set is reflected in the public use data files that we were able to access.
- In this section we will focus on the continuous variable **cost_tot_oop_12m** which measures out of pocket medical expenses last 12 months



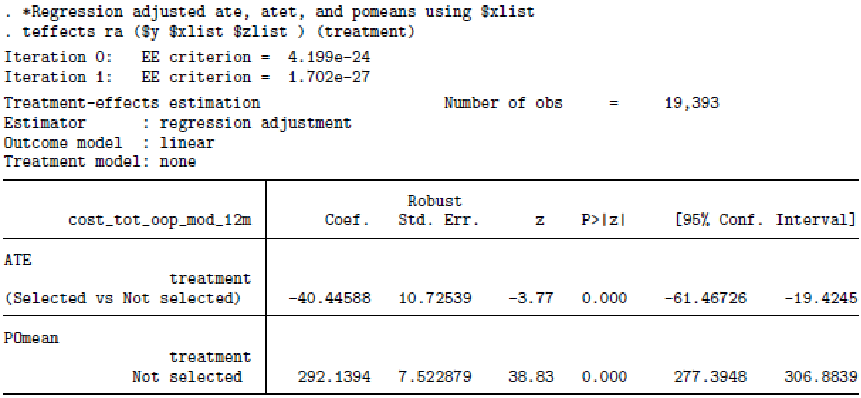
```
. * Two-sample t-test
. ttest $y, by(treatment)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Not sele |  11,403    291.2125    6.997433    747.2196    277.4963    304.9287
Selected |  11,276    246.5498    6.760028    717.8373     233.299    259.8007
---------+--------------------------------------------------------------------
combined |  22,679    269.0062    4.867889    733.0821    259.4648    278.5476
---------+--------------------------------------------------------------------
    diff |            44.66267    9.731627                25.58801    63.73732
------------------------------------------------------------------------------
    diff = mean(Not sele) - mean(Selected)                        t =   4.5894
Ho: diff = 0                                     degrees of freedom =    22677

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

OHIE RA t-test



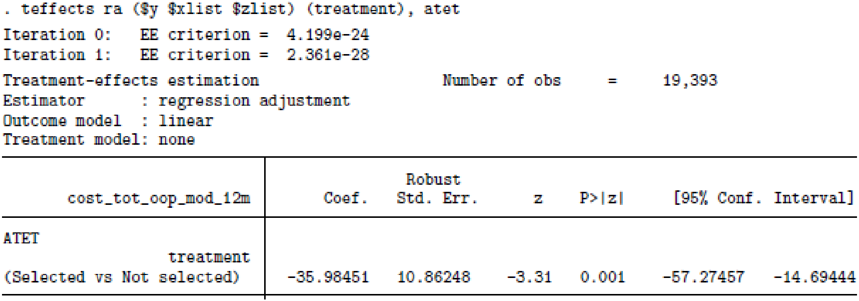
```
. *Regression adjusted t-test with robust standard errors
. regress $y $zlist treatment, vce(cluster household_id)

Linear regression                               Number of obs     =     19,393
                                                F(9, 17347)       =      27.68
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0147
                                                Root MSE          =     726.91

                          (Std. Err. adjusted for 17,348 clusters in household_id)
-----------------------------------------------------------------------------------
                |               Robust
cost_tot_oop-12m|      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+------------------------------------------------------------------
     smk_curr_12m|   11.96658   5.971198     2.00   0.045     .2624328    23.67073
hhinc_pctfpl_12m|   1.089545    .099305    10.97   0.000      .894897    1.284193
        edu_12m_2|  -8.614341   15.29018    -0.56   0.573    -38.58464    21.35596
        edu_12m_3|   59.16091   18.26756     3.24   0.001     23.35465    94.96717
        edu_12m_4|   39.41761   22.07356     1.79   0.074    -3.848792      82.684
employ_hrs_12m_2|  -37.98864   16.81801    -2.26   0.024    -70.95364   -5.023636
employ_hrs_12m_3|  -50.68008   16.37585    -3.09   0.002     -82.7784   -18.58175
employ_hrs_12m_4|  -22.48794   14.44654    -1.56   0.120    -50.80461    5.828738
        treatment|  -45.74203   10.64844    -4.30   0.000    -66.61404   -24.87001
            _cons|   186.5655   18.82997     9.91   0.000     149.6568    223.4741
```

RA ATE estimate



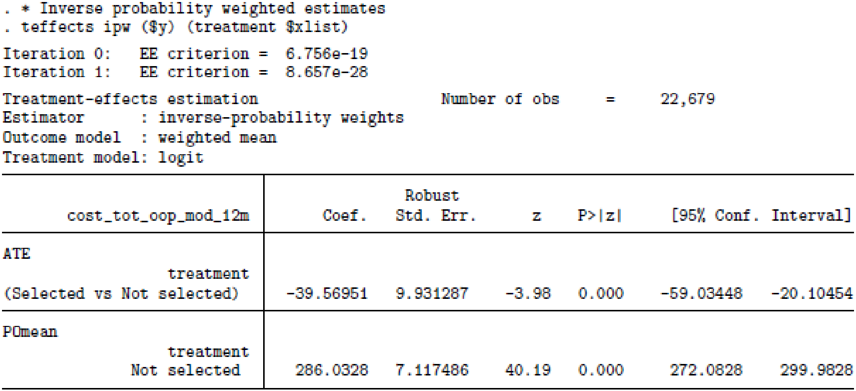
```
. *Regression adjusted ate, atet, and pomeans using $xlist
. teffects ra ($y $xlist $zlist ) (treatment)
Iteration 0:   EE criterion =  4.199e-24
Iteration 1:   EE criterion =  1.702e-27
Treatment-effects estimation                    Number of obs     =     19,393
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
------------------------------------------------------------------------------
                         |               Robust
       cost_tot_oop_mod_12m |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------+----------------------------------------------------------------
ATE                      |
               treatment |
(Selected vs Not selected) |  -40.44588   10.72539    -3.77   0.000    -61.46726    -19.4245
-------------------------+----------------------------------------------------------------
POmean                   |
               treatment |
            Not selected |   292.1394   7.522879    38.83   0.000     277.3948     306.8839
------------------------------------------------------------------------------
```

RA ATET estimate



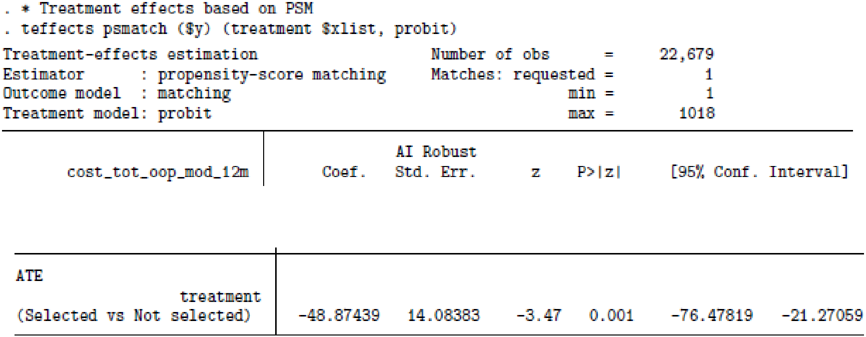
```
. teffects ra ($y $xlist $zlist) (treatment), atet

Iteration 0:   EE criterion =  4.199e-24
Iteration 1:   EE criterion =  2.361e-28

Treatment-effects estimation                   Number of obs     =     19,393
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
------------------------------------------------------------------------------------------
                          |               Robust
       cost_tot_oop_mod_12m |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------------+---------------------------------------------------------------
ATET                      |
                 treatment |
(Selected vs Not selected) |  -35.98451   10.86248    -3.31   0.001    -57.27457   -14.69444
------------------------------------------------------------------------------------------
```

IPW ATE estimate



```
. * Inverse probability weighted estimates
. teffects ipw ($y) (treatment $xlist)

Iteration 0:   EE criterion =  6.756e-19
Iteration 1:   EE criterion =  8.657e-28

Treatment-effects estimation                    Number of obs     =     22,679
Estimator      : inverse-probability weights
Outcome model  : weighted mean
Treatment model: logit
------------------------------------------------------------------------------------------
                           |               Robust
      cost_tot_oop_mod_12m |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------------+--------------------------------------------------------------
ATE                        |
                 treatment |
(Selected vs Not selected) |  -39.56951   9.931287    -3.98   0.000    -59.03448   -20.10454
---------------------------+--------------------------------------------------------------
POmean                     |
                 treatment |
              Not selected |   286.0328   7.117486    40.19   0.000     272.0828    299.9828
------------------------------------------------------------------------------------------
```

PSM ATE estimate

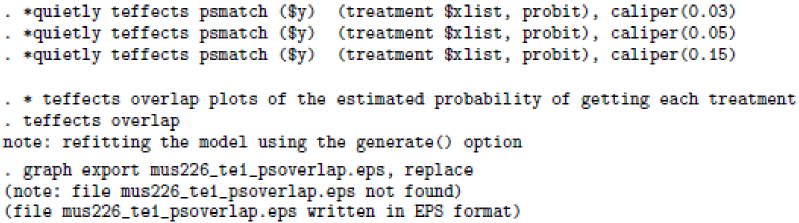


```
. * Treatment effects based on PSM
. teffects psmatch ($y) (treatment $xlist, probit)
Treatment-effects estimation                Number of obs      =     22,679
Estimator      : propensity-score matching  Matches: requested =          1
Outcome model  : matching                                 min =          1
Treatment model: probit                                   max =       1018
--------------------------------------------------------------------------------
                        |              AI Robust
     cost_tot_oop_mod_12m |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]


--------------------------------------------------------------------------------
ATE                     |
               treatment |
(Selected vs Not selected) |  -48.87439   14.08383    -3.47   0.001    -76.47819   -21.27059
--------------------------------------------------------------------------------
```

Diagnostic checks for treatment balance

- For matching methods to work satisfactorily the treated and untreated groups should be "similar".
- This should be tested. One way - graph and compare the distribution of propensity scores.
- Stata has `teffects overlap` (post-estimation) command to check this. This should be run after `teffects nnmatch` or teffects `psmatch command`
- If the match is not good, consider trimming the sample or changing the matching criterion.
- If sample sizes are similar a comparison is easier.

Stata's teffects overlap command



```
. *quietly teffects psmatch ($y)  (treatment $xlist, probit), caliper(0.03)
. *quietly teffects psmatch ($y)  (treatment $xlist, probit), caliper(0.05)
. *quietly teffects psmatch ($y)  (treatment $xlist, probit), caliper(0.15)

. * teffects overlap plots of the estimated probability of getting each treatment
. teffects overlap
note: refitting the model using the generate() option

. graph export mus226_te1_psoverlap.eps, replace
(note: file mus226_te1_psoverlap.eps not found)
(file mus226_te1_psoverlap.eps written in EPS format)
```

Output from the overlap command

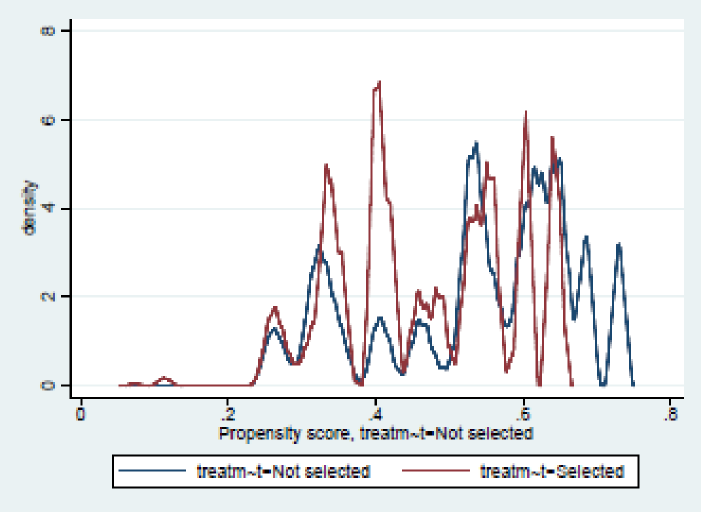


Figure 26.2. Probability overlap range

TE estimation with observational data

Linking counterfactual to linear models

- Treatment effect models are 'black box' in the sense that they don't explain how the causal chain goes from treatment to the response.
- Effectiveness of a treatment is closely connected to the *mechanism by which the treatment is delivered.* Absent a specification of that mechanism, more tenuous is the causal inference derived from it. Examples: Hours watching TV –> weight gain (or academic performance)
- Regression models are 'structural' in the sense that they specify how a 'third' variable comes in and links treatment and response.
- Treatment variable may be heterogeneous if a broad label is used to describe a variety of diverse treatments (distinct from heterogeneous response itself)

Structural form equations and treatment effect

- *Key assumption (invariance):* Given outcome equation ('y-eqn') and treatment equation ('D-eqn'), the y_j equation does not change when the D-equation does, otherwise the parameters in the y_j equation are useless for policy intervention on D. Consider a structural form for y_i (related to *Marschak-Lucas policy evaluation critique*)
- $$y_i = \beta_1 + \beta_2 D_i + \beta_3 x_i + u_i.$$
- With self selection into treatment we have the **structure**:
$$D_i = \alpha_x x_i + \alpha_c c_i + \varepsilon_i.$$
- After substituting the second into the first we get the **reduced form**:
$$y_i = \beta_1 + \beta_2 \beta_c c_i + (\beta_3 + \beta_2 \alpha_x) x_i + u_i + \varepsilon_i,$$
where x_i, c_i are observed and u_i, ε_i are error terms.

Structural form equations and treatment effect (2)

- If D_i is exogenously assigned then equation (2) disappears from the system. Structure (1) is unaffected by changes in (2) (by the invariance assumption) and hence the parameters in (1) are useful for policy analysis.
- Estimating the reduced form (3) is problematic if the parameters in (2) change ("Lucas critique") as then the parameter in (3) change also.

IV Estimation of LATE

Local ATE (LATE)

- Outcome is a function of observable $\mathbf{x}$ and a participation decision indicator D :

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \alpha D_i + u_i, \tag{2}$$

- Participation depends on IV z (which may be binary); interpret as treatment assignment mechanism

$$D_i^* = \gamma_0 + \gamma_1 z_i + v_i \ , \tag{3}$$

- D_i^* is a latent variable; has observable counterpart D_i generated by

$$D_i = \begin{cases} 0 \text{ if } D_i^* \leq 0 \\ 1 \text{ if } D_i^* > 0 \end{cases} . \tag{4}$$

Assumptions

- Assumptions. (1) IV z appears in the D-equation and does not appear in $\mathbf{y}$-equation. May be continuous or discrete, and in a special case is binary: exclusion restriction. Exclusion of regressors $\mathbf{x}$ from the y-equation is a simplification.
- (2) Conditional on $(\mathbf{x}, z)$ $\text{Cov}[z, v] = \text{Cov}[u, z] = \text{Cov}[\mathbf{x}, u] = 0$,

  $$\text{Cov}[D, z] \neq 0.$$

  D depends upon z in a nontrivial fashion → use the notation $D(z)$ to emphasize dependence of D on z.
- (3) No randomness of coefficients in (3)
- (4) Many studies assume a just-identified model with untestable exclusion restriction, but this is not essential.
- Under the above assumptions IV estimation of $(\boldsymbol{\beta}, \alpha)$ is consistent.
- OLS is biased because $\text{Cov}[D, u] \neq 0$

Estimation method

- What is different from standard just-identified model with only continuous endogenous variables?
  - This model has an endogenous dummy variable which could represent choice behavior → endogenous selection
  - Outcome equation could also be discrete/binary.
  - Joint conditional distribution of (y_i, D_i) is harder to specify in an unrestricted fashion ; joint normality often assumed
- In the standard case, 2sls, IV/GMM, MLE are widely-used estimators.
  - Standard problems are: weak instruments; sensitivity to different valid instruments which vary in their strength
- Ignoring discreteness of D,linear methods (as well as diagnostic tests of linear models) are often applied

LATE estimator (1)

- Let $z' = z + \delta$, $\delta \neq 0$. Then noting $E[D|\mathbf{x}, D(z)] = \Pr[D(z) = 1]$, taking expectations gives

$$\begin{aligned} E[y|\mathbf{x}, D(z)] &= \mathbf{x}'\boldsymbol{\beta} + \alpha\Pr[D(z) = 1], \\ E[y|\mathbf{x}, D(z')] &= \mathbf{x}'\boldsymbol{\beta} + \alpha\Pr[D(z') = 1], \end{aligned}$$

  where, after subtraction, we have

$$E[y|\mathbf{x}, z'] - E[y|\mathbf{x}, z] = \alpha\left[\Pr[D(z') = 1] - \Pr[D(z) = 1]\right].$$

- Solving for the **local average treatment effect** (LATE):

$$\begin{aligned} \alpha_{IV} &= \frac{E[y|\mathbf{x}, z'] - E[y|\mathbf{x}, z]}{\Pr[D(z') = 1] - \Pr[D(z) = 1]}, \\ &= \frac{E[y|z'] - E[y|z]}{\Pr[D(z') = 1] - \Pr[D(z) = 1]}. \end{aligned} \tag{5}$$

  where the second line averages over $\mathbf{x}$.

LATE estimator (2)

- This expression is well-defined if $\Pr[D(z')=1]- \ \Pr[D(z)=1] \neq 0$.
- The sample analog of this expression is the ratio of the mean difference between the treated and the nontreated divided by the change in the proportion treated due to the change in z.
- If the y-equation has no $\mathbf{x}$, then

$$\widehat{\alpha}_{IV} = \frac{\widehat{\text{cov}}[y,z]}{\widehat{\text{cov}}[D,z]}$$

$$\text{plim } \widehat{\alpha}_{IV} = \frac{E[y|z'] - E[y|z]}{\Pr[D(z')=1] - \Pr[D(z)=1]}$$

- $\widehat{\alpha}_{IV} =$ ratio of the average causal effect of z on y and the average causal effect of z on D.

Remarks and critique of the IV model

- As it stands, the IV estimator has a blackbox character. The role of $\mathbf{z}$ needs some elaboration.
- Why and how does z impact D? Which of the participants in the treatment get impacted and why?
- Is there a theory of the mechanism by which z impacts D? Is there more than one operating mechanism?
- Suppose there is more than one IV; could different IVs differ in their total impact because different subpopulations are susceptible to different IVs? What does $\widehat{\alpha}_{IV}$ measure? ATE?
- Angrist, Imbens, Rubin (AIR, JASA 1996) divide the population into ("compliers, defiers, never takers, always takers") depending upon assignment and choice of treatment.
- AIR argue that $\widehat{\alpha}_{IV}$ is a measure of treatment effect on the compliers.

Assumptions of AIR (JASA 1996)

- Stable Unit Treatment Value Assumption (SUTVA) $\rightarrow$ (No 'general equilibrium' effects or no interdependence in treatment effects.)

Definition: Causal effect of z on D for i is $D_i(1) - D_i(0)$
Definition: Causal effect of z on Y : for i is $Y_i(1, D_i(1)) - Y_i(0, D_i(0))$

- Potential outcomes: $Y[z, D_i(z)] \rightarrow [Y_i(0,0)\ ,\ Y_i(1,0),\ Y_i(0,1),\ Y_i(1,1)]$
- Potential treatments: $[D_i(z)$ $\rightarrow\ D_i(0) = 0;\ D_i(0) = 1;\ D_i(1) = 0;\ \ D_i(1) = 1;$ Potential assignments: $z\ \rightarrow z_i = 0;\ \ z_i = 1$
- Assume treatment assigned randomly so all units have same probability of assignment.
- Assume treatment effect is not zero or that $z \rightarrow D$ is a nontrivial effect.
- Assume a valid exclusion restriction so zero causal effect for never-takers and always-takers.

Classification of units

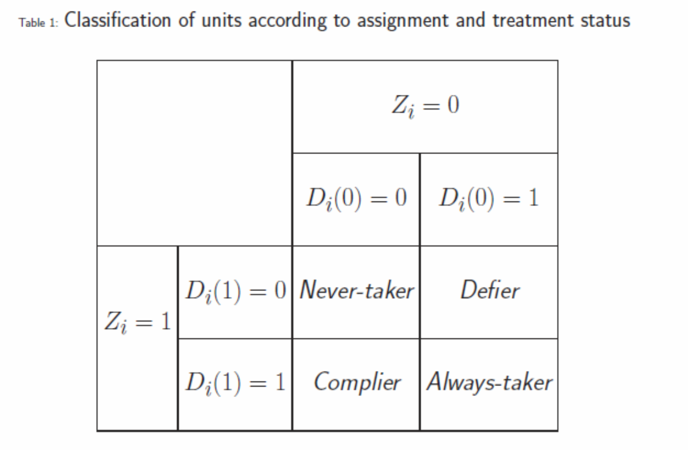


Table 1: Classification of units according to assignment and treatment status

		$Z_i = 0$	
		$D_i(0) = 0$	$D_i(0) = 1$
$Z_i = 1$	$D_i(1) = 0$	*Never-taker*	*Defier*
	$D_i(1) = 1$	*Complier*	*Always-taker*

Are above assumptions enough for identification?

- No, b/c causal parameter $E[Y_i(1, D_i(1)) - Y_i(0, D_i(0))]$ is a weighted average of the effects on compliers and defiers
- Need monotonicity assumption to rule out defiers.

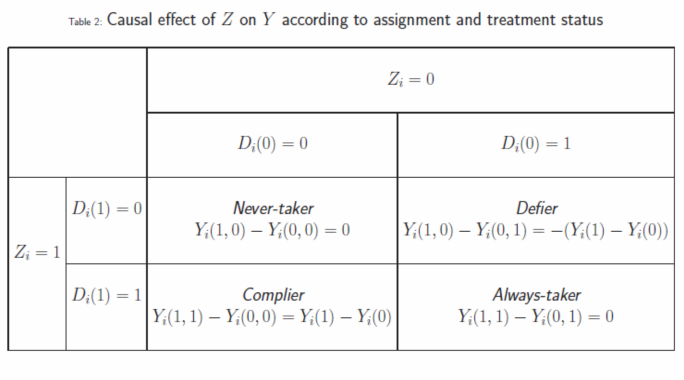


Table 2: Causal effect of Z on Y according to assignment and treatment status

		$Z_i = 0$	
		$D_i(0) = 0$	$D_i(0) = 1$
$Z_i = 1$	$D_i(1) = 0$	*Never-taker* $Y_i(1,0) - Y_i(0,0) = 0$	*Defier* $Y_i(1,0) - Y_i(0,1) = -(Y_i(1) - Y_i(0))$
	$D_i(1) = 1$	*Complier* $Y_i(1,1) - Y_i(0,0) = Y_i(1) - Y_i(0)$	*Always-taker* $Y_i(1,1) - Y_i(0,1) = 0$

Monotonicity assumption

- Above analysis applies when the treatment effect does not vary with individuals.
- If, however, the treatment effect is heterogeneous, then there is a potential for confounding the variation induced by z - - is the observed variation due to $z-$differences or $\alpha-$differences?
- Under heterogeneity the idiosyncratic component of the treatment effect,

  $$u_{i,1} = u_{i,0} + D_i(\alpha_i\left(\mathbf{x}_i\right) - \alpha\left(\mathbf{x}_i\right)),$$

  is a function of $\alpha_i\left(\mathbf{x}_i\right) - \alpha(\mathbf{x}_i)$. Then the previous assumptions are not enough to determine ATE or ATET.
- Solution: add **monotonicity assumption** as an additional identifying condition. This says that the instrument affects participation in a monotone fashion. This means that if on average participation is more likely given $Z = w$ than given $Z = z$, then anyone who would participate given $Z = z$ must also participate given $Z = w$.

Interpretation of “local”

- Because it measures the treatment effect on the “compliers” that are induced to participate in the treatment as a result of the change in z.
- LATE depends upon the particular values of z used to evaluate the treatment and on the particular instrument chosen.
- “Movers” may not be representative of the whole treated, let alone the whole population. $\Rightarrow$ LATE parameter may not be informative about the consequences of large policy changes.
- If the instrument is binary, the LATE parameter and the IV estimate are equivalent.
- Under overidentification the LATE parameter estimated for each instrument will in general differ. However, a weighted average may be constructed.

Critique of LATE

- Heckman (JEL, 2010) cites and summarizes many previous discussions of LATE. Imbens (JEL, 2010) responds.
    - (1) Unclear who the "compliers" are and whether LATE extends to the population.
    - (2) In general LATE $\neq$ ATE or ATET; if no heterogeneity then LATE $=$ ATE
    - (3) LATE is mechanical, does not make explicit the implicit economic assumptions .
- Different instruments identify different parameters and may refer to different TEs.
- Marginal treatment effect is more insightful and more fundamental than LATE
- LATE framework is too limited. Treatments may be multinomial or continuous or ordered. Greater generality is required.

IV when treatment effect is heterogeneous (1)

- Consider a linear model with an endogenous treatment variable whose coefficient is random, i.e. treatment effect is not constant across the treated.
- Suppose treatment variable y_2 is continuous. Outcome y_1 is the outcome depends on y_2 and exogenous $\mathbf{x}_i$. The model is

$$\begin{aligned}
y_{1,i} &= (\alpha + v_i) D_i + \mathbf{x}_i' \boldsymbol{\beta}_1 + \varepsilon_i \\
&= \alpha D_i + \mathbf{x}_i' \boldsymbol{\beta}_1 + \varepsilon_i + v_i y_{2i} \\
&= v_i \overline{D}_i + \alpha y_{2i} + \mathbf{x}_i' \boldsymbol{\beta}_1 + w_i; \\
D_i^* &= \gamma z_i + \mathbf{x}_i' \boldsymbol{\beta}_2 + \boldsymbol{\eta}_i, \\
w_i &= \varepsilon_i + v_i \left(D_i - \overline{D} \right)
\end{aligned}$$

IV when treatment effect is heterogeneous (2)

- Marginal (TE) response $= (a + v_i)$
- Assume $E[\varepsilon_i|\mathbf{x}_i, D_{2i}] = \ E[v_i|\mathbf{x}_i, D_i] = 0$. Then $E[\varepsilon_i + v_i D_i|\mathbf{x}_i, y_{2i}] = 0$, and $\mathrm{V}[\varepsilon_i + v_i D_i|\mathbf{x}_i, y_{2i}]$ depends upon $\mathbf{x}_i$ and hence is heteroskedastic.
- OLS estimator of $(\alpha, \boldsymbol{\beta}_1)$ is consistent but not efficient. Follows from the assumed exogeneity of y_2.
- Now y_2 is endogenous. Assume:

$$\begin{aligned} E[\boldsymbol{\varepsilon}_i|\mathbf{x}_i, z_i] &= E[\eta_i|\mathbf{x}_i, z_i] = E[v_i|\mathbf{x}_i, z_i] = 0, \\ E[\boldsymbol{\varepsilon}_i^2|\mathbf{x}_i, z_i] &= \sigma_\varepsilon^2; \ \ E[v_i^2|\mathbf{x}_i, z_i] = \sigma_v^2; \ \ E[\boldsymbol{\eta}_i^2|\mathbf{x}_i, z_i] = \sigma_\eta^2. \end{aligned}$$

- Endogeneity is introduced by permitting correlation between v and η. Specifically assume that $E[v_i|\eta] = \rho\eta_i$, which would hold if $(v \ \ \eta \ )$ were bivariate normal distributed. Under these assumptions, z is a valid instrument, and $\mathbf{x}_1$ is exogenous. The exclusion of z from the y_1 equation is an identifying restriction.

IV is consistent!

- For estimation of (**??**) use instruments $(z\ \mathbf{x})$
- However for consistent estimation need $E[w_i|\mathbf{x}_i, z_i] = 0$.

– first component of w_i, ε_i, is uncorrelated with z_i by assumption;
– second component of w_i is $v_i\left(D_i - \overline{D}\right)$ can be shown (using iterated expectations) to not affect the result that the IV estimator is consistent.

- IV estimator is consistent but not efficient because of the heteroskedastic error.

More on heterogeneity and LATE estimation

- Consider 3 cases
  - Case 1: Multiple group specific instruments $(z_{1,}..., z_m)$ impact a single endogenous treatment variable, each has different impact on the treatment variable D, but treatment effect of D on y is homogeneous across groups.
  - Case 2: Treatment effect varies across groups but is constant within each group. A single IV is available for the endogenous treatment.
  - Care 3: There is dual heterogeneity, with multiple instruments impacting treatment, and variation in treatment effect across groups.
- We apply standard LATE methodology of IV regression of y on (D, x). What parametr is identified?
  - Case 1: LATE is identified;
  - Case 2: Weighted sum of group-specific LATE parameter is identified; Case 3:
  - Case 3: Under additional assumption of independence between $(\partial D/\partial z)$ and $(\partial y/\partial D)$, a weighted average of group-speciifc TEs.
- Interpretation of LATE is harder if the details of mechanism are unclear.

IV-LATE Estimation in nonlinear models

IV and LATE estimation in nonlinear models

- Suppose conditional expectation function $E[y|x]$ is not linear as we have assumed so far.
- Specifically suppose that $y_i = E[y_i|\mathbf{x}_i] + u_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}) + u_i$ and some components of $\mathbf{x}_i$ are correlated with u_i
- The (L)ATE effect of the treatment variable x_j in this model is the AME $= N^{-1}\sum_{i=1}^{N}\beta_j \exp(\mathbf{x}_i'\boldsymbol{\beta})$
- Also suppose that we have available a sufficient number of IVs which satisfy the moment restriction $E[\mathbf{z}_i'(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})]= \mathbf{0}$

This is a nonlinear IV problem which solves :

$$\min_{\beta} Q(\boldsymbol{\beta}) = \left[\sum_{i=1}^{N}[\mathbf{z}_i'(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})]\right]\mathbf{W}\left[\sum_{i=1}^{N}[\mathbf{z}_i'(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})]\right]$$

where $\mathbf{W}$ is a weighting matrix. Details omitted here.

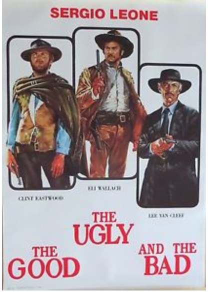
Implementing nonlinear IV or nonlinear GMM

- In Stata we can implement nonlinear GMM is several ways.
  - (1) Using evaluator moment function version of GMM- see example below
  - (2) Using a 2-step procedure based on residual augmentation
  - (3) Using a control function for endogenous regressors
- Methods (2) and (3) are two step methods in which we estimate first the treatment assignment function, use the estimates to generate a new variable(s), and then add this variable(s) to the outcome equation and estimate it on the assumption that conditional on the inclusion of generated regressors, there is no endogeneity problem.
- How to get the ATE or AME for a binary treatment variable?
  - Use bootstrap to estimate the standard errors if any 2-step procedure was used to handle endogeneity
  - Use `margins` command if it works, otherwise use Stata's postestimation `predict` command to get sample estimates of $E[y|\mathbf{x}, D = 1]$ and $E[y|\mathbf{x}, D = 0]$

NLIV estimation of over-identified model

```
. * Command gmm for one-step GMM (nonlinear IV) of overidentified Poisson model
. gmm (docvis - exp({xb:private medicaid age age2 educyr actlim totchr}+{b0})), ///
>  instruments(income ssiratio medicaid age age2 educyr actlim totchr) onestep nolog
Final GMM criterion Q(b) =  .0495772

GMM estimation

Number of parameters =   8
Number of moments    =   9
Initial weight matrix: Unadjusted                 Number of obs   =      3,677

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     private |   .5920142   .3397345     1.74   0.081    -.0738533    1.257882
    medicaid |   .3186685   .1909951     1.67   0.095    -.0556751     .693012
         age |   .3323179   .0705348     4.71   0.000     .1940723    .4705634
        age2 |   -.002176   .0004643    -4.69   0.000     -.003086   -.001266
      educyr |   .0190887   .0092216     2.07   0.038     .0010147    .0371626
      actlim |   .2084978   .0433758     4.81   0.000     .1234828    .2935128
      totchr |    .241843   .0129869    18.62   0.000     .2163892    .2672968
-------------+----------------------------------------------------------------
         /b0 |  -11.86323   2.732711    -4.34   0.000    -17.21924   -6.507211
------------------------------------------------------------------------------
Instruments for equation 1: income ssiratio medicaid age age2 educyr actlim totchr _cons
```

Varieties of IV-based TE analysis



1. The assignment (selection) equation based on well-argued theory or evidence of how the mechanism works. Causal parameter is then plausible and may be generalizable.

2. Argument supporting the assignment mechanism is essentially a blackbox in which case we may be restricted to a conclusion like "it worked" or "it works" but external validity is questionable. 3. A statistically valid instrument arrives by a drone from somewhere, role in the assignment mechanism unknown, then accept or reject?

Regression Discontinuity Design

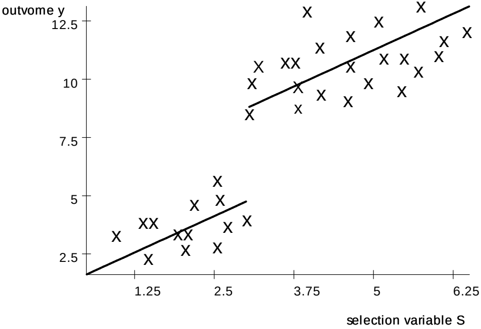
Two influential case studies

David S. Lee. (2008) Randomized experiments from non-random selection in U.S. House elections, *Journal of Econometrics,* 142(2), 675-697.
W. van der Klaauw, (2002) Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach , *International Economic Review*, 43(4), 2002, 1249-87

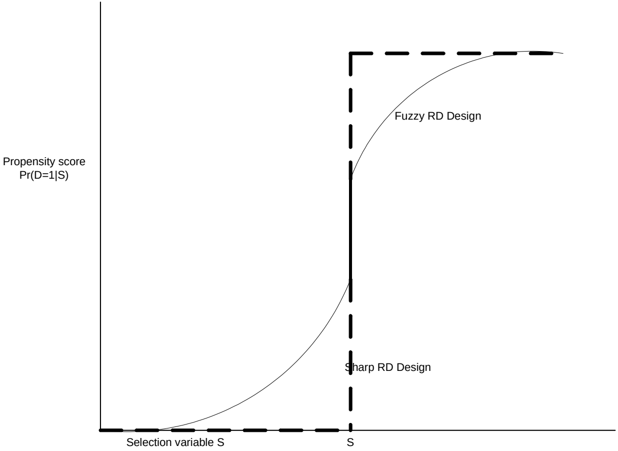
RDD Design

- Consider a framework for evaluating causal effects of interventions in which **assignment to a treatment is determined at least partly by the value of an observed covariate lying on either side of a fixed threshold.**
- Crossing the threshold ⇒ treatment will be assigned. Not crossing the threshold ⇒ treatment not assigned.
- Example: Offer of financial support in college depends upon satisfactory academic performance measured by a threshold value.
- RDD designs were first introduced in the evaluation literature by Thistlewaite and Campbell [1960. Regression-discontinuity analysis: an alternative to the ex-post Facto experiment. *Journal of Educational Psychology* 51, 309–317]
- RD may be sharp (single-valued threshold → SRD) or fuzzy (a band → FRD)

- Researchers interested in the causal effect of a binary intervention (as in the RCM)
- Distinguish between two designs, the Sharp and the Fuzzy (SRD and FRD) designs. D_i is a **deterministic function** of one of the covariates, the forcing (or treatment-determining) variable $X : D_i = 1\{X_i \geq c\}$.
- No idiosyncratic element in selection for treatment.



Fuzzy RDD



- $\{X_i \geq c\} \Rightarrow i \in$ treatment group (where participation is mandatory), and all units with $\{X_i < c\} \Rightarrow i \in$ the control group (whose members are not eligible for the treatment).

Using a control function

- Suppose the outcome equation is

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \alpha D_i + u_i, \tag{6}$$

In the sharp RD design

$$E[u|c, D] = E[u|c], \tag{7}$$

where u denotes the error on the outcome equation, because c is the only systematic determinant of D, c will capture any correlation between D and u.

- If $D = 1|X_i \geq c]$,dependence between D_i and u_i would make OLS inconsistent
- Treatment can be estimated by specifying and including the conditional mean function $E\left[u|D, c\right]$ as a "control function" in the outcome equation. Thus

$$y_i = \beta + \alpha D_i + k\left(x_i\right) + \varepsilon_i, \tag{8}$$

where $\varepsilon_i = y_i - \ E\left[y_i|D_i, x_i\right]$. If $k(x)$ is correctly specified, the

Identification Assumptions

- ATE estimated by comparing average y value of those just above and those just below the cutoff.
- In this RD design,

$$\lim_{x \downarrow c} E[y|x] - \lim_{x \uparrow c} E[y|x] = \alpha + \lim_{x \downarrow c} E[u|x] - \lim_{x \uparrow Sc} E[u|x]. \tag{9}$$

- Formally assume that without treatment, individuals in a small interval around c would have similar average outcomes

Assumption A1. The conditional mean function $E[u|x]$ is continuous at c.

$$\lim_{x \downarrow c} E[y|x] - \lim_{x \uparrow c} E[y|x] \tag{10}$$

Assumption A2. The mean treatment effect function $E[\alpha_i|x]$ is right continuous at c.

$$y_i = \beta + \alpha W_i + k(x_i) + \varepsilon_i, \tag{11}$$

where $\varepsilon_i = y_i - \ E[y_i|D_i, x_i]$

SRD (1)

- Under randomization $(Y(1),\ Y(0)) \perp D$
- To identify the average causal effect of the treatment, the SRD design looks at the discontinuity in the conditional expectation of the outcome, i.e.

$$\tau_{SRD} = E[Y_i(1) - Y_i(0)|X_i = c]$$

- Justification? Appeal to **smoothness** and **continuity of distribution function assumption.**
- Conditional regression of $Y(1)$ and $Y(0)$ on $X = x$ are continuous in x. Then the estimand is the difference of two regression functions at a point.

$$\tau_{SRD} = \lim_{x \downarrow c} E[Y_i(1)|X = x] - \lim_{x \uparrow c}[Y_i(0)|X_i = x]$$

- Compare mean outcomes for treated and untreated *at the margin*. Identify the intervention effect locally at the threshold for selection.

SRD (2)

- Because of stochastic independence assumption $(Y(1),\ Y(0)) \perp D|X = c$, SRD is interpreted as a quasi-experimental design. Similar to an exclusion restriction.
- The local dispersion in outcomes at c is purely random and equivalent to variation under random assignment.
- For a sufficiently large sample $\tau = \lim_{x\downarrow c} E[Y_i(1)|X = x] - \lim_{x\uparrow c}[Y_i(0)|X_i = x$ can be estimated using only data in the neighborhood of c.
- If sample not large enough, assume parametric form for the regression function (usually two polynomial) away from discontinuity.

SRD (3)

- Example: Party affiliation and voting behavior of congressman. Here $x = 50\%$ of the votes. Districts in which the Democratic vote is just less than 50% are very similar to the districts in which the vote exceeds 50%. Yet party affiliation (here the causal variable) leads to a sharp difference in voting behavior at 50%, i.e. a case of SRD.
- In the FRD design the probability of receiving the treatment need not change from zero to one at the threshold. Design allows for a smaller jump in the probability of assignment (propensity score) to the treatment at the threshold:

FRD

- FRD allows for random unobserved component in the treatment assignment; for identification require additional assumption.
- Treatment assignment may depend upon both observables and unobservables (to the econometrician)
- Assume: $Y(1), Y(0), D(c)$ stochastically independent of X in the neighborhood of c.
- In FRD design the probability of receiving the treatment need not change from zero to one at the threshold. May have a smaller jump in the probability of treatment assignment at the threshold:

$$\begin{aligned} \tau_{FRD} &= \frac{\lim_{x \downharpoonright c} E[Y_i|X = x] - \lim_{x \upharpoonright c}[Y_i|X_i = x]}{\lim_{x \downharpoonright c} E[D_i|X = x] - \lim_{x \upharpoonright c}[D_i|X_i = x]} \\ &= \frac{\Delta \text{ in outcome}}{\Delta \text{ in assignment probability}} \end{aligned}$$

where $\lim_{x \downharpoonright c} E[D_i|X = x] - \lim_{x \upharpoonright c}[D_i|X_i = x] \neq 0$ because of the known discontinuity at c.

FRD with heterogeneous response

- In the case of **heterogeneous treatment responses** we need additional assumptions.

Assumption A2*. The average treatment effect function $E[\alpha_i|x]$ is continuous at c.
Assumption A3. D_i is independent of α_i conditional on x near $x = c$.

$$y_i = \beta + \alpha E[D_i|x_i] + k\left(x_i\right) + \varepsilon_i, \qquad (12)$$

where $\varepsilon_i = y_i - \ E\left[y_i|D_i, x_i\right]$ and $k(x_i)$ is a specification of $E[u_i|x_i]$.

Monotonicity assumption

- **Assumption A4:** $D_i(x)$ is non-increasing in x at $x = c$.
- Assumptions of an FRD analysis ⇒ comparing treated and control units with $X_i = c$, is likely to be the wrong approach.
- Categorize heterogeneous responses into: (1) compliers; (ii) defiers; (iii) never-takers, and (iv) always-takers
- Why? Because the treated units with $X_i = c$ is heterogeneous with both *compliers* and *always-takers*, and control units at $X_i = c$ consist only of *never-takers*. More on this when we cover the **Local ATE** in the IV case.
- Comparing these different types of units has no causal interpretation under the FRD assumptions.

Other assumptions

- Fuzzy **nonparametric** regression discontinuity: D determined by x and ε.
- SRD rare in practice because treatment assignment usually involves multiple (not just one) decisions.
- To deal with fuzzy RDD, need to assume

  i) Selection is on observable at $x \approx \tau$

  ii) The propensity score for receiving treatment has a break at τ

Key theoretical and conceptual contributions

1. the interpretation of estimates for fuzzy regression discontinuity (FRD) designs allowing for general heterogeneity of treatment effects (Hahn et al., 2001, HTV from hereon),
2. adaptive estimation methods (Sun, 2005),
3. specific methods for choosing bandwidths (Ludwig and Miller, 2005), and
4. various tests for discontinuities in means and distributions of non-affected variables (Lee, 2007; McCrary, 2007)

Interpretation and connections with TET estimation

External validity

- $D = 1$ if $x > c$, otherwise 0, where τ is a known threshold. Although the T and C groups are not comparable in most x values, they are comparable on a small neighborhood of $x = \tau$. The treatment D is imposed on the individuals by a law or rule depending on x or ε.
- In time series RDD is equivalent to the before-after (BA) design.
- Comparison between DD design vs. RDD and BA.
- DD has the advantage that there the control group is subject to a time effect but not the treatment effect; but in BA and RDD, everybody potentially gets the treatment.

- SRD/FRD designs at best provide estimates of the average effect for a subpopulation
- FRD design restricts the relevant subpopulation even further to that of compliers at this value of X.

Graphical Analysis

Graphical analyses: RD designs suggests that the effect of the treatment of interest can be measured by the extent of the discontinuity in the expected value of the outcome at a particular point.

Figure IIa: Candidate's Probability of Winning Election t+1, by Margin of Victory in Election t: local averages and parametric fit

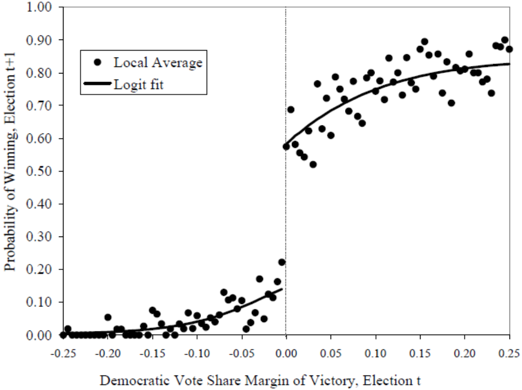
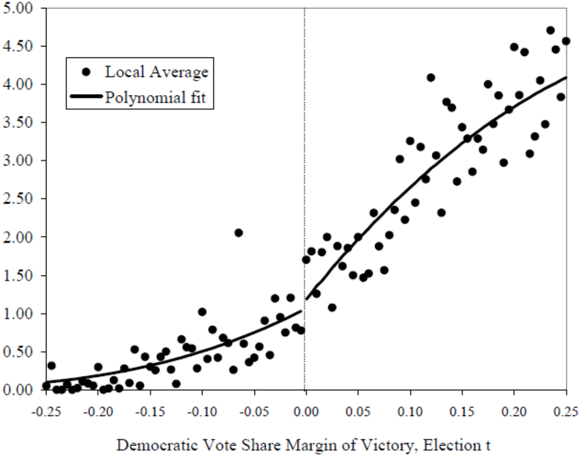


Figure IIb: Candidate's Accumulated Number of Past Election Victories, by Margin of Victory in Election t: local averages and

Graphical Analysis 2

Democratic Vote Share Margin of Victory, Election t

Figure IIb: Candidate's Accumulated Number of Past Election Victories, by Margin of Victory in Election t: local averages and parametric fit



Parametric estimation of RD model

- RDD can be implemented by parametric methods.
- Using parametric regression model to extrapolate counterfactual is an option

$$y_i = \beta_1 + \beta_w D_i + \beta_x x_i + u_i, \quad E(u) = 0; \;\; E(u|x) = 0$$

where $\beta_D \neq 0$ implies a break or discontinuity in the regression function as x increases.

- Parametric RDD is heavily model-dependent, so the possibility of misspecification is a serious problem. This motivates nonparametric RDD.

Parametric estimation under endogenous selection

- If $\mathrm{Cov}[D,u] \neq 0$, OLS regression will produce a biased estimate of α.
- Consider

$$y_i = \beta + \alpha E[D_i|x_i] + k(x_i) + \varepsilon_i, \tag{13}$$

  where $\varepsilon_i = y_i - \ E[y_i|x_i]$ and $k(x_i)$ is a specification of $E[u_i|S_i]$.
- Stage 1: Propensity score function for a fuzzy RD design as

$$E[D_i|x_i] = f(x_i) + \gamma 1[x_i \geq c] \tag{14}$$

  where continuous function of x, $f(x_i)$, is continuous at c. By specifying the functional form of f (or by estimating f semi- or nonparametrically) we can estimate γ, the discontinuity at c.

Stage 2: The control function-augmented outcome equation is then estimated with D_i replaced by the first stage estimate of $E[D_i|x_i] = \Pr[D_i = 1|x_i]$; this estimate will be discontinuous in x whereas the included control function for $k(x)$ would be continuous in x at c. Correct specification of $f(S_i)$ and $k(S_i) \Rightarrow$ consistency of two-stage procedure.

Nonparametric estimation

- Nonparametric regression can estimate the treatment effect in both the SRD and FRD designs.

$$y_{ji} = \beta_d j + g(x_i) + u_{ji}$$

  where $g(\cdot)$ is an unknown function continuous at the point of discontinuity;

- Leads to the interpretation of **borderline randomization**. β_d is the treatment effect for the subpopulation $x \approx \tau$.
- Two unusual features in estimation:

  (1) we need the value of the regression function at a single point, and (2) that single point is a boundary point.

Nonparametric estimation 2

- Then the usual nonparametric kernel regression does not work very well.
- Local linear regression (Fan and Gijbels, 1996) is more relevant.
- Instead of locally fitting a constant function, we can fit linear regression functions to the observations within a distance *h* **on either side of the discontinuity point.**
- In the FRD design, the treatment effect is a ratio of two differences. So local linear regression should be used for both differences.

LATE - RDD - Matching

- Identification assumptions of FRD and LATE are very similar
- SRD/FRD cannot identify the treatment effects on individuals far from the discontinuity threshold. Hence similar to LATE
- Matching excludes selection on unobservables. Inference based on balanced samples using observables
- Combining RE and matching can help deal with endogenous or self-selected treatrment (GMM, control functions, IV)

Examples of RDD

- See Lee & Lemeiux (JEL, 2010) for coverage of applications in education, labor, political economy, crime, environment, health etc

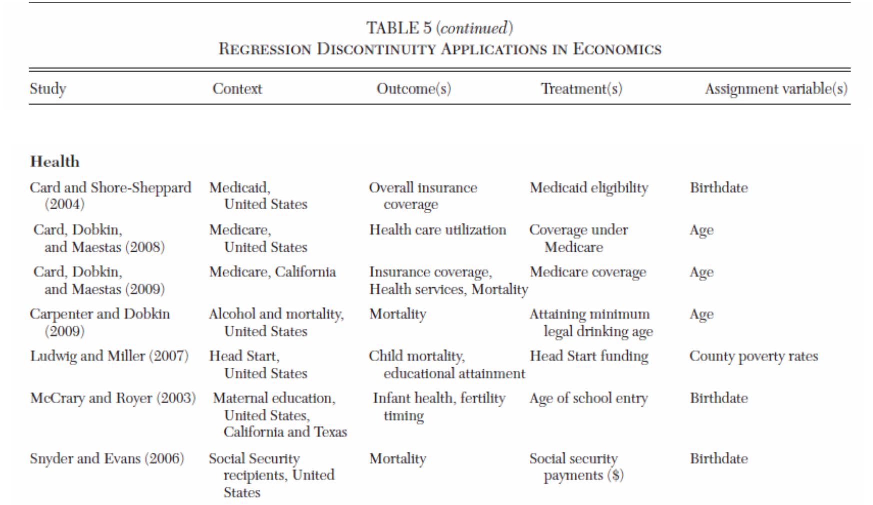


TABLE 5 (*continued*)
REGRESSION DISCONTINUITY APPLICATIONS IN ECONOMICS

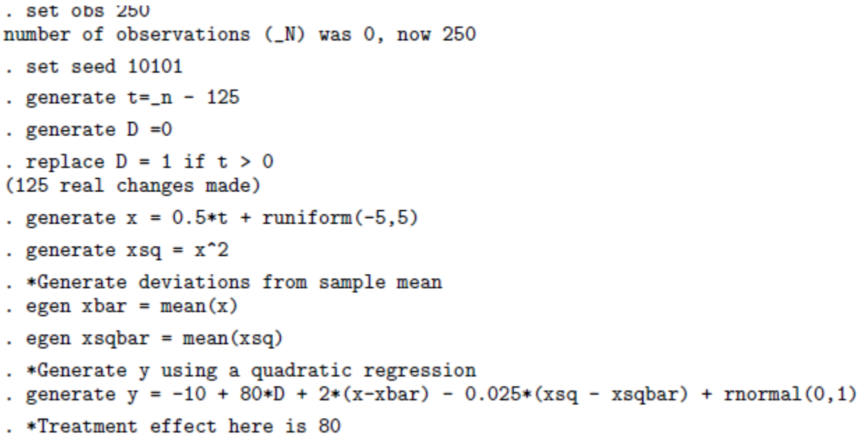
Study	Context	Outcome(s)	Treatment(s)	Assignment variable(s)
Health				
Card and Shore-Sheppard (2004)	Medicaid, United States	Overall insurance coverage	Medicaid eligibility	Birthdate
Card, Dobkin, and Maestas (2008)	Medicare, United States	Health care utilization	Coverage under Medicare	Age
Card, Dobkin, and Maestas (2009)	Medicare, California	Insurance coverage, Health services, Mortality	Medicare coverage	Age
Carpenter and Dobkin (2009)	Alcohol and mortality, United States	Mortality	Attaining minimum legal drinking age	Age
Ludwig and Miller (2007)	Head Start, United States	Child mortality, educational attainment	Head Start funding	County poverty rates
McCrary and Royer (2003)	Maternal education, United States, California and Texas	Infant health, fertility timing	Age of school entry	Birthdate
Snyder and Evans (2006)	Social Security recipients, United States	Mortality	Social security payments ($)	Birthdate

Implementing RD estimation in Stata

Implementing RD

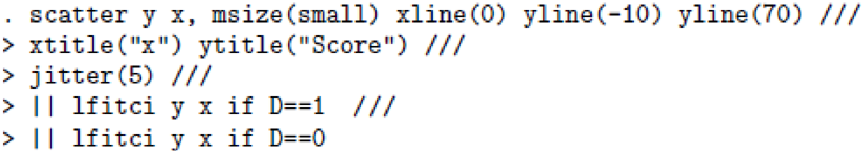
- Two main user-provided packages are `rdrobust` (Calanico et al: Stata Journal, 2014, 16(2)) `and rd` (Nichols, Stata Journal 2007, 7(4))
- rdrobust is more up to date and complete; includes commands for point and interval estimation, bandwidth and window selection and data plots.
- Standard data plot commands in Stata include twoway scatters; another useful one i user-provided `cmogram`
- Graphical plots can be very suggestive and have a significant role in implementing rd analysis.
- rd analysis can also be implemented in a parametric setting so standard estimation commands have a role also

Simulating RD data using a parametric model



```
. set obs 250
number of observations (_N) was 0, now 250

. set seed 10101

. generate t=_n - 125

. generate D =0

. replace D = 1 if t > 0
(125 real changes made)

. generate x = 0.5*t + runiform(-5,5)

. generate xsq = x^2

. *Generate deviations from sample mean
. egen xbar = mean(x)

. egen xsqbar = mean(xsq)

. *Generate y using a quadratic regression
. generate y = -10 + 80*D + 2*(x-xbar) - 0.025*(xsq - xsqbar) + rnormal(0,1)

. *Treatment effect here is 80
```

Code for a scatter plot



```
. scatter y x, msize(small) xline(0) yline(-10) yline(70) ///
> xtitle("x") ytitle("Score") ///
> jitter(5) ///
> || lfitci y x if D==1  ///
> || lfitci y x if D==0
```

RD plot with linear fitted lines

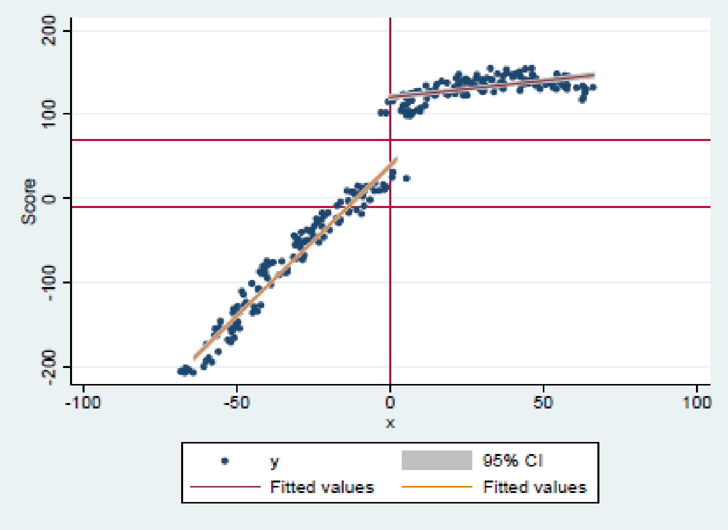
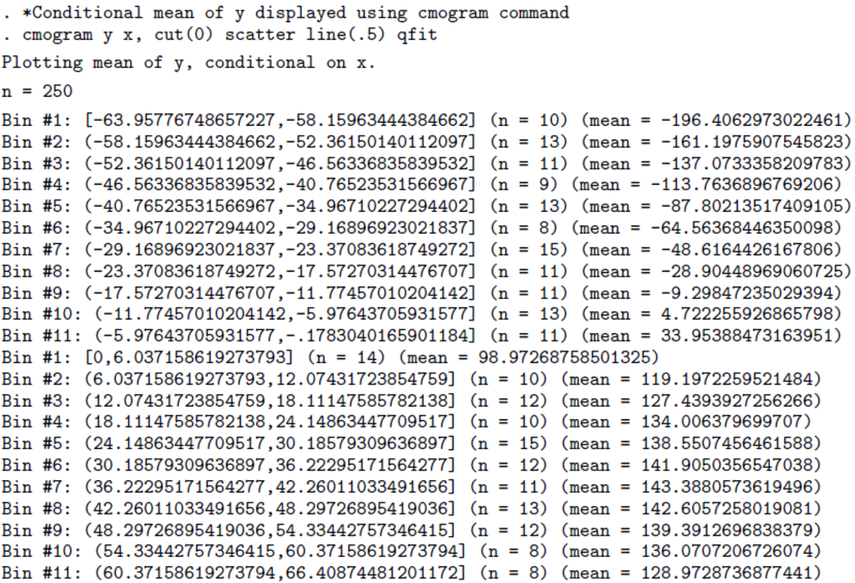


Figure 27.1. Regression discontinuity

Binned smoothed data



```
. *Conditional mean of y displayed using cmogram command
. cmogram y x, cut(0) scatter line(.5) qfit
Plotting mean of y, conditional on x.
n = 250
Bin #1: [-63.9577674865722 7,-58.15963444384662] (n = 10) (mean = -196.4062973022461)
Bin #2: (-58.15963444384662,-52.36150140112097] (n = 13) (mean = -161.1975907545823)
Bin #3: (-52.36150140112097,-46.56336835839532] (n = 11) (mean = -137.0733358209783)
Bin #4: (-46.56336835839532,-40.76523531566967] (n = 9) (mean = -113.7636896769206)
Bin #5: (-40.76523531566967,-34.96710227294402] (n = 13) (mean = -87.80213517409105)
Bin #6: (-34.96710227294402,-29.16896923021837] (n = 8) (mean = -64.56368446350098)
Bin #7: (-29.16896923021837,-23.37083618749272] (n = 15) (mean = -48.6164426167806)
Bin #8: (-23.37083618749272,-17.57270314476707] (n = 11) (mean = -28.90448969060725)
Bin #9: (-17.57270314476707,-11.77457010204142] (n = 11) (mean = -9.29847235029394)
Bin #10: (-11.77457010204142,-5.97643705931577] (n = 13) (mean = 4.72225926865798)
Bin #11: (-5.97643705931577,-.1783040165901184] (n = 11) (mean = 33.95388473163951)
Bin #1: [0,6.037158619273793] (n = 14) (mean = 98.97268758501325)
Bin #2: (6.037158619273793,12.07431723854759] (n = 10) (mean = 119.1972259521484)
Bin #3: (12.07431723854759,18.11147585782138] (n = 12) (mean = 127.4393927256266)
Bin #4: (18.11147585782138,24.14863447709517] (n = 10) (mean = 134.006379699707)
Bin #5: (24.14863447709517,30.18579309636897] (n = 15) (mean = 138.5507456461588)
Bin #6: (30.18579309636897,36.22295171564277] (n = 12) (mean = 141.9050356547038)
Bin #7: (36.22295171564277,42.26011033491656] (n = 11) (mean = 143.3880573619496)
Bin #8: (42.26011033491656,48.29726895419036] (n = 13) (mean = 142.6057258019081)
Bin #9: (48.29726895419036,54.33442757346415] (n = 12) (mean = 139.3912696838379)
Bin #10: (54.33442757346415,60.37158619273794] (n = 8) (mean = 136.0707206726074)
Bin #11: (60.37158619273794,66.40874481201172] (n = 8) (mean = 128.9728736877441)
```

cmogram plot with np regression

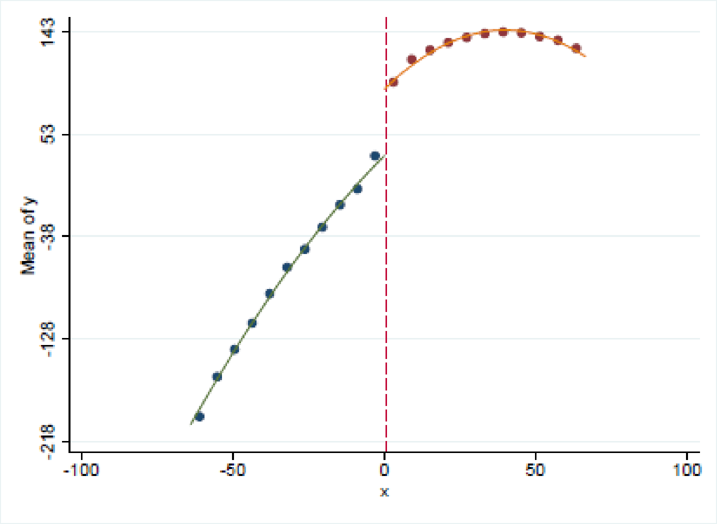


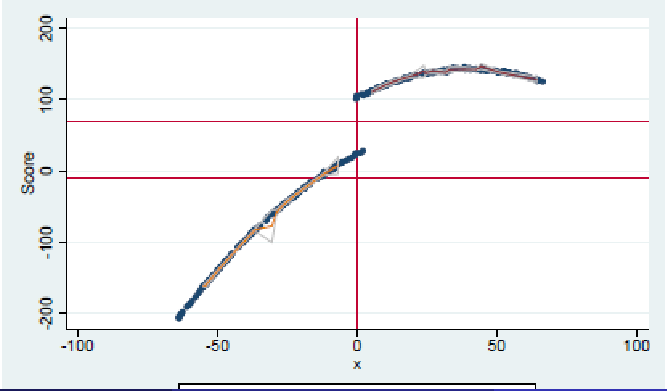
Figure 27.2. Conditional mean of y given x using bins

rd plot with lpolyci superimposed

scatter

$_4$*9.pdf*

```
. scatter y x, msize(small) xline(0) yline(-10) yline(70) ///
> xtitle("x") ytitle("Score") ///
> || lpolyci y x if D==1, bw(0.05) deg(2) n(250) fcolor(none) /
> || lpolyci y x if D==0, bw(0.05) deg(2) n(250) fcolor(none)
```



Eyeball test of y/x discontinuity

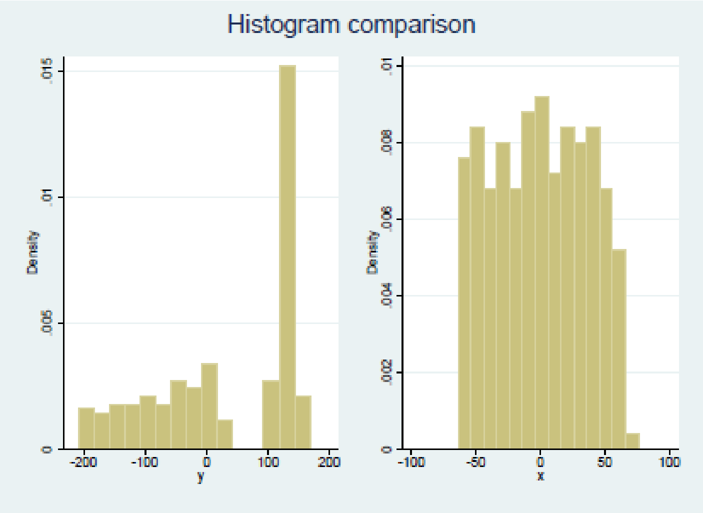


Figure 27.4. Histogram comparison

Stata's rdrobust command

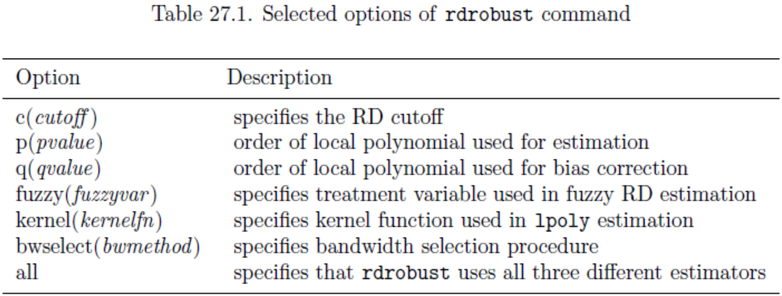


Table 27.1. Selected options of `rdrobust` command

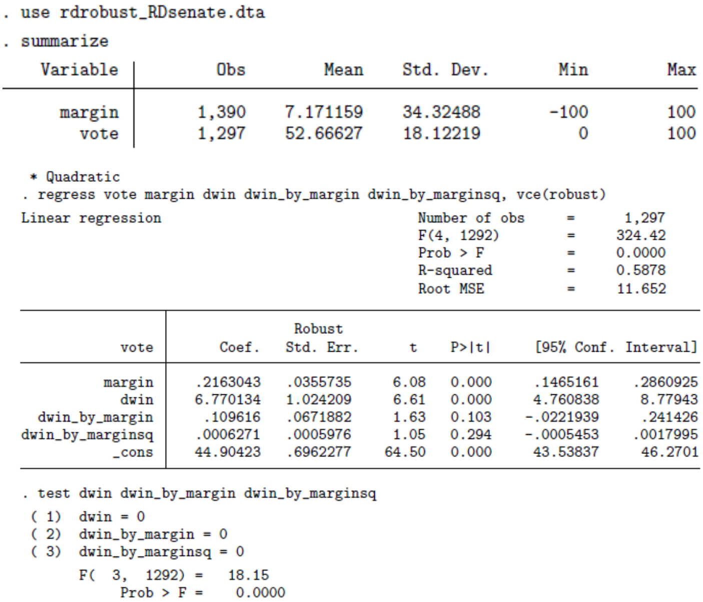
Option	Description
c(*cutoff*)	specifies the RD cutoff
p(*pvalue*)	order of local polynomial used for estimation
q(*qvalue*)	order of local polynomial used for bias correction
fuzzy(*fuzzyvar*)	specifies treatment variable used in fuzzy RD estimation
kernel(*kernelfn*)	specifies kernel function used in `lpoly` estimation
bwselect(*bwmethod*)	specifies bandwidth selection procedure
all	specifies that `rdrobust` uses all three different estimators

Empirical Application of RD

Incumbent advantage example

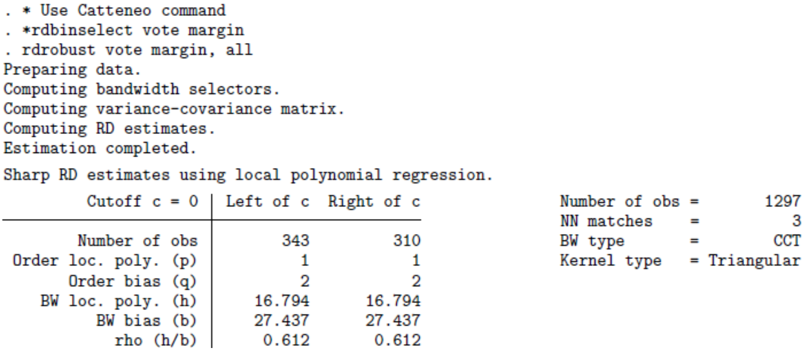
- In this well-known application to data from the US senate elections the interest is in discontinuity in relation between percentage of **vote** (y) at $t+1$ given the **margin** achieved at t. Does the incumbent advantage in polling jump at the point of discontinuity?

Parametric test of srd using quadratic regression



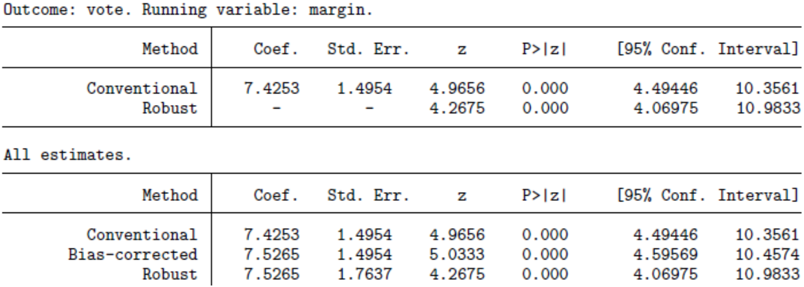
```
. use rdrobust_RDsenate.dta

. summarize

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      margin |      1,390    7.171159    34.32488       -100        100
        vote |      1,297    52.66627    18.12219          0        100

  * Quadratic
 . regress vote margin dwin dwin_by_margin dwin_by_marginsq, vce(robust)
 Linear regression                               Number of obs     =      1,297
                                                 F(4, 1292)        =     324.42
                                                 Prob > F          =     0.0000
                                                 R-squared         =     0.5878
                                                 Root MSE          =     11.652

 ----------------------------------------------------------------------------------
                  |               Robust
             vote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
 -----------------+----------------------------------------------------------------
           margin |   .2163043   .0355735     6.08   0.000     .1465161    .2860925
             dwin |   6.770134   1.024209     6.61   0.000     4.760838     8.77943
   dwin_by_margin |    .109616   .0671882     1.63   0.103    -.0221939     .241426
 dwin_by_marginsq |   .0006271   .0005976     1.05   0.294    -.0005453    .0017995
            _cons |   44.90423   .6962277    64.50   0.000     43.53837     46.2701
 ----------------------------------------------------------------------------------

 . test dwin dwin_by_margin dwin_by_marginsq

  ( 1)  dwin = 0
  ( 2)  dwin_by_margin = 0
  ( 3)  dwin_by_marginsq = 0

        F(  3,  1292) =   18.15
             Prob > F =    0.0000
```

rdrobust output 1 under srd



```
. * Use Catteneo command
. *rdbinselect vote margin
. rdrobust vote margin, all
Preparing data.
Computing bandwidth selectors.
Computing variance-covariance matrix.
Computing RD estimates.
Estimation completed.

Sharp RD estimates using local polynomial regression.
        Cutoff c = 0 | Left of c  Right of c            Number of obs =       1297
    -----------------+-----------------------           NN matches    =          3
       Number of obs |       343         310            BW type       =        CCT
  Order loc. poly. (p) |       1           1            Kernel type   = Triangular
      Order bias (q) |         2           2
   BW loc. poly. (h) |    16.794      16.794
         BW bias (b) |    27.437      27.437
           rho (h/b) |     0.612       0.612
```

rdrobust output under srd 2



```
Outcome: vote. Running variable: margin.
--------------------------------------------------------------------------------
          Method |    Coef.    Std. Err.      z      P>|z|     [95% Conf. Interval]
--------------------------------------------------------------------------------
    Conventional |   7.4253     1.4954     4.9656    0.000     4.49446     10.3561
          Robust |     -          -        4.2675    0.000     4.06975     10.9833
--------------------------------------------------------------------------------

All estimates.
--------------------------------------------------------------------------------
          Method |    Coef.    Std. Err.      z      P>|z|     [95% Conf. Interval]
--------------------------------------------------------------------------------
    Conventional |   7.4253     1.4954     4.9656    0.000     4.49446     10.3561
  Bias-corrected |   7.5265     1.4954     5.0333    0.000     4.59569     10.4574
          Robust |   7.5265     1.7637     4.2675    0.000     4.06975     10.9833
```

rdrobust under frd set-up

- How to generate frd sample? How in practice does frd data get generated?
- One way is that treatment assignment depends on both the running variable x and another unobserved missing variable z.
- Reconsider the dgp used under srd.
- With the cut-off now obscured we expect RD estimate under srd assumptions to be possibly biased
- A numerical example based on generated data illustrates that the estimated treatment effect depends upon choice of local polynomial and bandwidth selected.

Example of frd data generation

```
. set obs 250
number of observations (_N) was 0, now 250

. set seed 10101

. generate t=_n - 125.5

. generate z = runiform(-3,2)

. generate x = 0.5*t + runiform(-5,5)

. generate D = 0

. * Add an extra condition for treatment assignment
. * Treatment assignment also depends upon z
. replace D = 1 if z>0 & x >0
(46 real changes made)

. generate xsq = x^2

. *Generate deviations from sample mean
. egen xbar = mean(x)

. egen xsqbar = mean(xsq)

. egen Dbar = mean(D)

. generate u = runiform(-50,50)

. *Generate y using a quadratic conditional regression
```

Estimation under srd assumptions applied to frd data

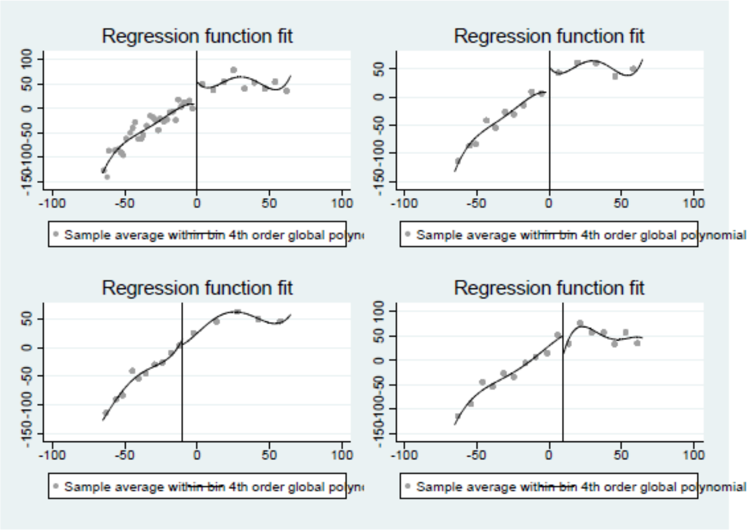
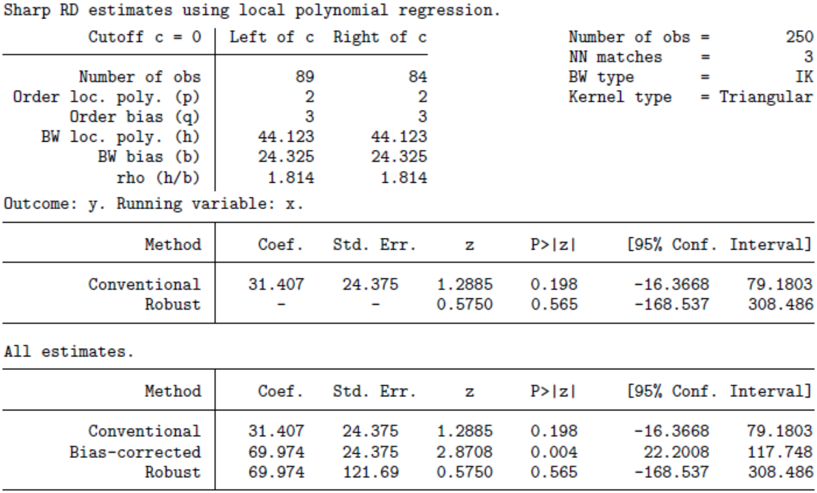


Figure 27.10. RD plots with correctly and incorrectly specified cutoff

srd estimate based on frd data



```
Sharp RD estimates using local polynomial regression.
        Cutoff c = 0 | Left of c  Right of c            Number of obs =        250
---------------------+----------------------            NN matches    =          3
                     |                                  BW type       =         IK
       Number of obs |        89          84            Kernel type   = Triangular
  Order loc. poly. (p) |         2           2
      Order bias (q) |         3           3
   BW loc. poly. (h) |    44.123      44.123
         BW bias (b) |    24.325      24.325
           rho (h/b) |     1.814       1.814
Outcome: y. Running variable: x.
---------------------------------------------------------------------------------
              Method |   Coef.    Std. Err.      z      P>|z|    [95% Conf. Interval]
---------------------+-----------------------------------------------------------
        Conventional |  31.407     24.375     1.2885    0.198    -16.3668    79.1803
              Robust |     -          -       0.5750    0.565    -168.537    308.486
---------------------------------------------------------------------------------

All estimates.
---------------------------------------------------------------------------------
              Method |   Coef.    Std. Err.      z      P>|z|    [95% Conf. Interval]
---------------------+-----------------------------------------------------------
        Conventional |  31.407     24.375     1.2885    0.198    -16.3668    79.1803
      Bias-corrected |  69.974     24.375     2.8708    0.004     22.2008    117.748
              Robust |  69.974     121.69     0.5750    0.565    -168.537    308.486
---------------------------------------------------------------------------------
```

Power and sample size determination in RDD (1)

- Parametric RDD model: $y_i = c + \beta_D D_i + \gamma g(x_i) + \varepsilon_i$
- Parameter of interest is β_D.Suppose we apply OLS.

$$var_{RDD}(\widehat{\beta}_D) = \frac{\sigma_y^2(1 - R^2)}{Np(1-p)(1-r_{D,x}^2)}$$

where R^2 =regression R^2, p =proportion of treated; $r_{D,x}^2$ = squared correlation between D and x, σ_y^2 = variance of outcome.

Then (Lee, Hyunshik, and Tom Munk. "Using regression discontinuity design for program evaluation."
Proceedings of the 2008 Joint Statistical Meeting. American Statistical Association, 2008))

Power and sample size determination in RDD (2)

- In designing RDD we are dealing with a design question which is essentially the same as in an RCT
- Hence sample size can be determined using the same tools as in RCT.
- We require α (significance level), $1-\beta$ (desired power), and desired minimum detectable standardised treatment effect size, $\delta = \beta_D/\sigma$
- Then the required sample size is determined by

$$N^* = \frac{(1-R^2)(z_{1-\alpha} - z_\beta)^2}{\delta^2 p(1-p)(1-r^2_{D,x})}$$

Power and sample size determination in RDD (3)

- In a large sample the power is given by

$$1-\beta = 1-\Pr\left[Z < \left(z_{1-\alpha} - \delta\sqrt{\frac{Np(1-p)(1-r^2_{D,x})}{(1-R^2)}}\right)\right]$$

- RDD is less efficient than RCT; Variance of β_D under RCT is given by

$$var_{RCT}\left(\widehat{\beta}_D\right) = \frac{\sigma^2_y(1-R^2)}{Np(1-p)}$$

so the relative efficiency of RCT is $RE = 1/(1-r^2_{D,x})$

Multi-level treatment with many counterfactuals

Multilevel treatment effects

- Most of preceding discussion has focussed on a binary set-up with just one level of treatment which is received or not.
- In practice treatments are often multi-valued
- Each subject receives one of several mutual exclusive treatments.
- Multilevel treatments may be exogenous or self-selected and endogenous.
- For example in the context of health insurance a purchaser of an insurance policy may choose between policies with different levels of generosity and coverage which then would affect the use of medical services.
- We consider extending the analytical methods considered so far to such settings.

ML treatments - features

- Multilevel treatment models may have ordered or unordered treatments.
- In either case there will be pairwise comparisons between either adjacent or far apart levels of treatment
- Implication - multilevel treatments involve many more parameters, more counterfactuals, and require additional computations to support pairwise comparisons
- However, even in a multivalued treatment case, binary methodology can be used for specific pairwise comparisons,
- In general a multivalued treatment effects framework will deliver greater efficiency; Cattaneo(2010} establishes

consistency and normality of a class of ATE estimators

Assumptions in ML TE estimation

- As before assume conditional independence (selection on observables) and exclude endogenous treatment.
- Then results on regression adjustment and propensity scores extend to multivalued treatments.
- Stata's `teffects` commands {`teffects ra`}, {`teffects ipw`}, {`teffects ipwra`}, `teffects aipw`

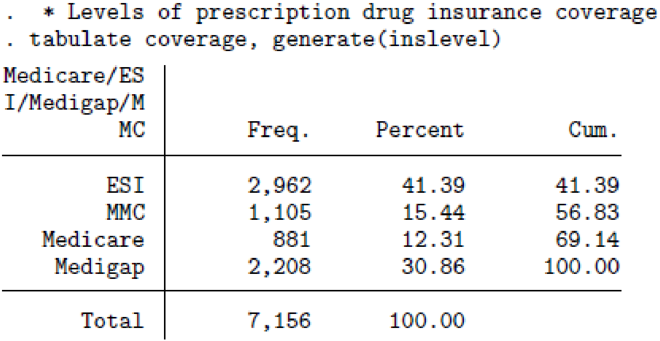
extend to multivalued treatments and will be illustrated in the next section.

- Commands {`teffects psmatch`}`,` {`teffects nnm`} do not.

Example

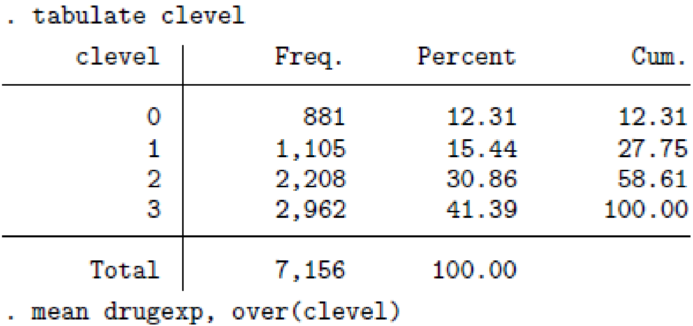
- We use data on prescription drug expenditures of the elderly Medicare population in the USA in 2003 and 2004 derived from Medicare Current Beneficiary Survey. See (Li & Trivedi, HEC, 2014) for details.
- This is subset of the data used pertaining to the elderly seeking prescription drug coverage through privately obtained access to various sources
- Includes employer-sponsored plans {ESI}, {Medigap} plans and Medicare managed care plans {MMC}.
- Sample includes individuals with prescription drug insurance from these three sources and we add a fourth comparison group of {Medicare} elderly without such coverage.
- The objective is to estimate treatment effects of the three levels of insurance {`inslevel`} which is treated as exogenous (!)

Frequency distribution across 3 plans with drug coverage and one without (RA)



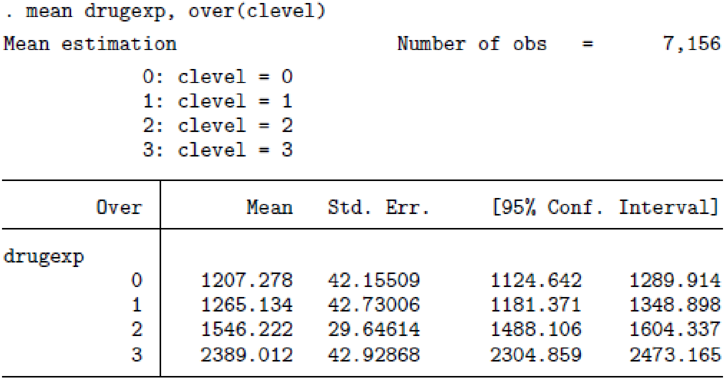
```
.  * Levels of prescription drug insurance coverage
. tabulate coverage, generate(inslevel)

Medicare/ES |
I/Medigap/M |
        MMC |      Freq.     Percent        Cum.
------------+-----------------------------------
        ESI |      2,962       41.39       41.39
        MMC |      1,105       15.44       56.83
   Medicare |        881       12.31       69.14
    Medigap |      2,208       30.86      100.00
------------+-----------------------------------
      Total |      7,156      100.00
```

Ordered frequency distribution across 3 plans with drug coverage and one without



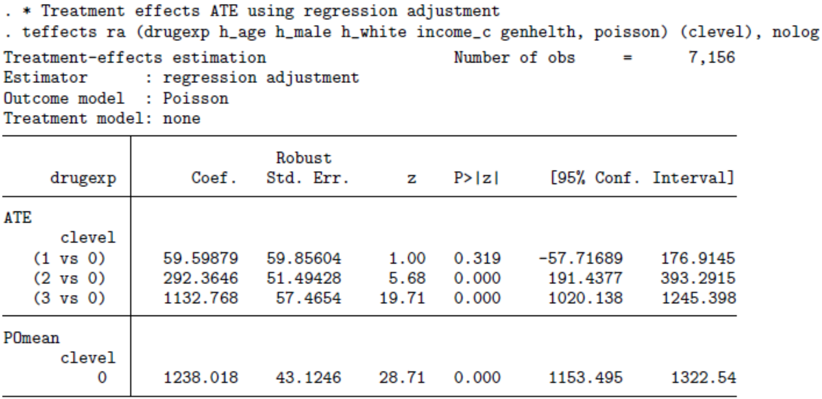
```
. tabulate clevel

     clevel |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        881       12.31       12.31
          1 |      1,105       15.44       27.75
          2 |      2,208       30.86       58.61
          3 |      2,962       41.39      100.00
------------+-----------------------------------
      Total |      7,156      100.00

. mean drugexp, over(clevel)
```

Annual expenditure on prescription drugs by level of coverage



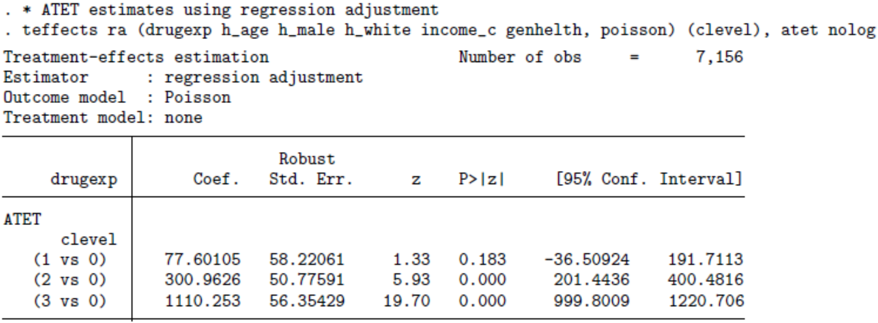
```
. mean drugexp, over(clevel)

Mean estimation                     Number of obs     =      7,156

            0: clevel = 0
            1: clevel = 1
            2: clevel = 2
            3: clevel = 3

--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
drugexp      |
           0 |   1207.278   42.15509      1124.642    1289.914
           1 |   1265.134   42.73006      1181.371    1348.898
           2 |   1546.222   29.64614      1488.106    1604.337
           3 |   2389.012   42.92868      2304.859    2473.165
--------------------------------------------------------------
```

ATE of insurance levels 1-3 relative to no insurance (level 0)



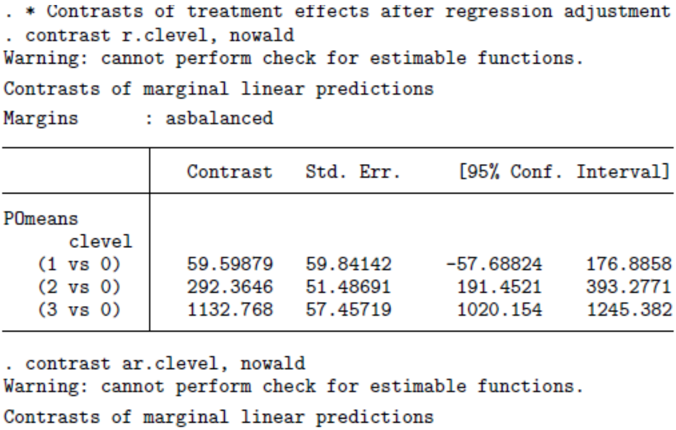
```
. * Treatment effects ATE using regression adjustment
. teffects ra (drugexp h_age h_male h_white income_c genhelth, poisson) (clevel), nolog
Treatment-effects estimation                    Number of obs     =      7,156
Estimator      : regression adjustment
Outcome model  : Poisson
Treatment model: none
------------------------------------------------------------------------------
             |               Robust
     drugexp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
      clevel |
   (1 vs 0)  |   59.59879   59.85604     1.00   0.319    -57.71689    176.9145
   (2 vs 0)  |   292.3646   51.49428     5.68   0.000     191.4377    393.2915
   (3 vs 0)  |   1132.768    57.4654    19.71   0.000     1020.138    1245.398
-------------+----------------------------------------------------------------
POmean       |
      clevel |
          0  |   1238.018    43.1246    28.71   0.000     1153.495     1322.54
------------------------------------------------------------------------------
```

ATET of insurance levels 1-3 relative to no insurance (level 0)



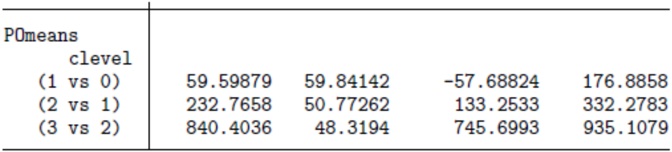
```
. * ATET estimates using regression adjustment
. teffects ra (drugexp h_age h_male h_white income_c genhelth, poisson) (clevel), atet nolog

Treatment-effects estimation                    Number of obs     =      7,156
Estimator      : regression adjustment
Outcome model  : Poisson
Treatment model: none

------------------------------------------------------------------------------
             |               Robust
     drugexp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATET         |
      clevel |
   (1 vs 0)  |   77.60105   58.22061     1.33   0.183    -36.50924    191.7113
   (2 vs 0)  |   300.9626   50.77591     5.93   0.000     201.4436    400.4816
   (3 vs 0)  |   1110.253   56.35429    19.70   0.000     999.8009    1220.706
------------------------------------------------------------------------------
```

Continuation of previous table



```
. * Contrasts of treatment effects after regression adjustment
. contrast r.clevel, nowald
Warning: cannot perform check for estimable functions.

Contrasts of marginal linear predictions

Margins      : asbalanced

------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
POmeans      |
      clevel |
   (1 vs 0)  |   59.59879   59.84142    -57.68824    176.8858
   (2 vs 0)  |   292.3646   51.48691     191.4521    393.2771
   (3 vs 0)  |   1132.768   57.45719     1020.154    1245.382
------------------------------------------------------------

. contrast ar.clevel, nowald
Warning: cannot perform check for estimable functions.

Contrasts of marginal linear predictions
```

ATE of insurance levels 1-3 relative to adjacent level



```
-------------+------------------------------------------------------
POmeans      |
      clevel |
    (1 vs 0) |   59.59879   59.84142     -57.68824    176.8858
    (2 vs 1) |   232.7658   50.77262      133.2533    332.2783
    (3 vs 2) |   840.4036    48.3194      745.6993    935.1079
-------------+------------------------------------------------------
```

Other estimators

This exercise could also be implemented using teffects aipw command; for details see chapters.

Maximum likelihood estimation of TEs for models with endogenous treatments

Estimation of a canonical 2-equation model with endogenous treatment

- Begin with the canonical two-equation model presented during the coverage of LATE.
- This model has one "structural" equation for a continuous outcome variable y and one reduced form equation for a binary treatment variable, D.
- Estimation of this model was done using 2SLS or IV estimation, under the assumption that a valid IV is available and that there is an exclusion restriction for identification.
- Under certain additional assumptions this model can be extended and estimated by maximum likelihood.

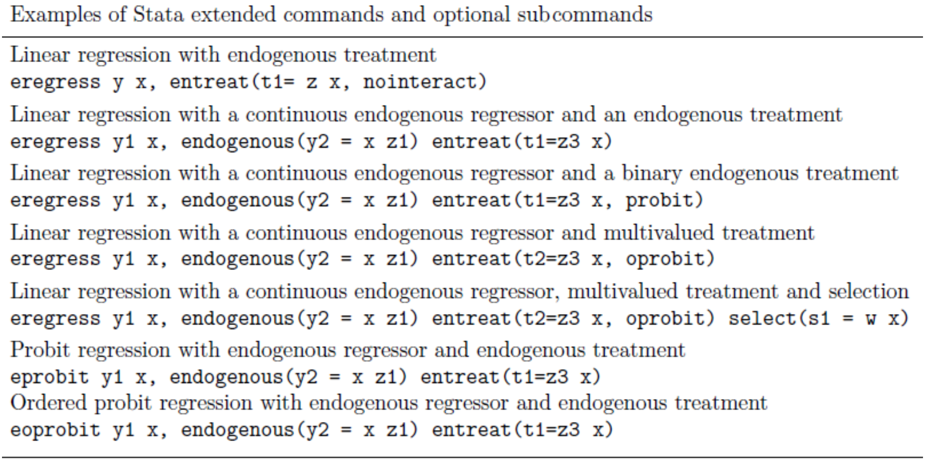
Assumptions

1. Key assumption is that of joint multivariate normal distribution of errors
2. Second key assumption is that the endogeneity structure is recursive, not simultaneous. No feedback from outcome to the treatment is allowed.
3. Many empirically interesting extensions are now estimable and described in the table below.
4. Binary or continuous treatment variable, including multivalued treatment, is allowed.
5. Additional endogenous variables may enter the outcome equation.
6. A specific form of selection is allowed also.

Scope of Stata's new eregress command

1. Expanded scope of the `eregress` command is described in the table below.
2. Linear and nonlinear joint normal models supported include LIML, probit, ordered probit, tobit.
3. Sub-command `endogenous` is an option for specifying reduced form for additional endogenous regressors.
4. Equation for endogenous treatment is specified using `entreat` subcommand
5. Ignoring endogeneity of treatment variable often amounts to neglect of selection bias; i.e. then the estimated treatment effect includes the selection component. i.e. $\widehat{ATE} = \text{Pure TE} + \text{Selection effect}$

Stata's extended TE commands for endogenous treatments



Examples of Stata extended commands and optional subcommands

Linear regression with endogenous treatment

```
eregress y x, entreat(t1= z x, nointeract)
```

Linear regression with a continuous endogenous regressor and an endogenous treatment

```
eregress y1 x, endogenous(y2 = x z1) entreat(t1=z3 x)
```

Linear regression with a continuous endogenous regressor and a binary endogenous treatment

```
eregress y1 x, endogenous(y2 = x z1) entreat(t1=z3 x, probit)
```

Linear regression with a continuous endogenous regressor and multivalued treatment

```
eregress y1 x, endogenous(y2 = x z1) entreat(t2=z3 x, oprobit)
```

Linear regression with a continuous endogenous regressor, multivalued treatment and selection

```
eregress y1 x, endogenous(y2 = x z1) entreat(t2=z3 x, oprobit) select(s1 = w x)
```

Probit regression with endogenous regressor and endogenous treatment

```
eprobit y1 x, endogenous(y2 = x z1) entreat(t1=z3 x)
```

Ordered probit regression with endogenous regressor and endogenous treatment

```
eoprobit y1 x, endogenous(y2 = x z1) entreat(t1=z3 x)
```

Computational detail

- Joint **normality** and **recursive structure** assumptions allow factorization of the likelihood function into a conditional part and a marginal part.
- Conditional part is the structural outcome equation, the marginal part is the one or more of reduced form equations for endogenous regressors.
- Computation is now feasible using the algorithm proposed by Roodman, D.. "Estimating fully observed recursive mixed-process models with cmp." (2009).
- Post-estimation Stata's `margins` command or the `estat teffects` command can be used to estimate ATE.

Interpreting ATE with endogenous treatment

- When the regressor is endogenous, it includes the effect of individual specific random error term,
- Hence treatment effect measures the total effect of the regressor and the effect of the idiosyncratic error. $ATE = E[y|\mathbf{x},\eta]$
- Averaging over all impacted observations then implies averaging over the random component also.
- The resulting ATE is then called either the average structural mean (ASM) or, in a binary outcome model, the average structural probability (ASP).
- In a linear model the implication is not consequential, but in a nonlinear model, computation involves numerical integeration to average over the error term.

Application of eregress to endogenous ordered probit treatment variable

An empirical application of `eregress` to an ordered probit model with endogenous multinomial treatment is given in the course packet.

The variable of interest is log prescription drug expenditure; endogenous treatment variable is an ordered variable for category of private insurance

There are four categories, the base category provides no coverage.

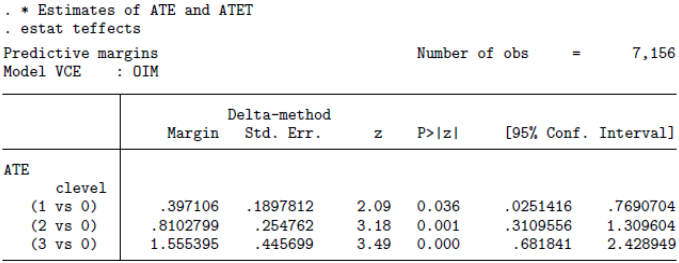
Multinomial probit model is used as reduced form for insurance choice.

The treatment effect is measured as percentage increase over the base category.

Estimated ATE for the three insurance levels are shown in the table below.

They are estimated to be 39.8, 81.0, and 155.5 per cent higher, respectively.

ATE for the three insurance levels



```
. * Estimates of ATE and ATET
. estat teffects

Predictive margins                              Number of obs     =      7,156
Model VCE    : OIM

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
      clevel |
   (1 vs 0)  |    .397106   .1897812     2.09   0.036     .0251416    .7690704
   (2 vs 0)  |   .8102799    .254762     3.18   0.001     .3109556    1.309604
   (3 vs 0)  |   1.555395    .445699     3.49   0.000      .681841    2.428949
------------------------------------------------------------------------------
```

Preview of selected topics not covered in the course (1)

- Extensions to linear and nonlinear panel data models
- Heterogeneity in responses characterized in a flexible manner
  - Random effect models allowing for heterogenous response parameters (Li & Trivedi, HEC, 2014)
  - Finite mixture models (Munkin & Trivedi, HEC, 2012)
  - Dirichlet mixture models (Hu, Munkin, Trivedi, JAE, 2015)
- Studying impact on a the distribution of outcomes, not just the means
  - Medicare Plan D choices and impact on prescription drug expenditure (Li & Trivedi, HEC, 2014)
  - Bayesian approaches
- Interdependence between hierarchical treatments

Li-Trivedi example of multi-level counterfactuals

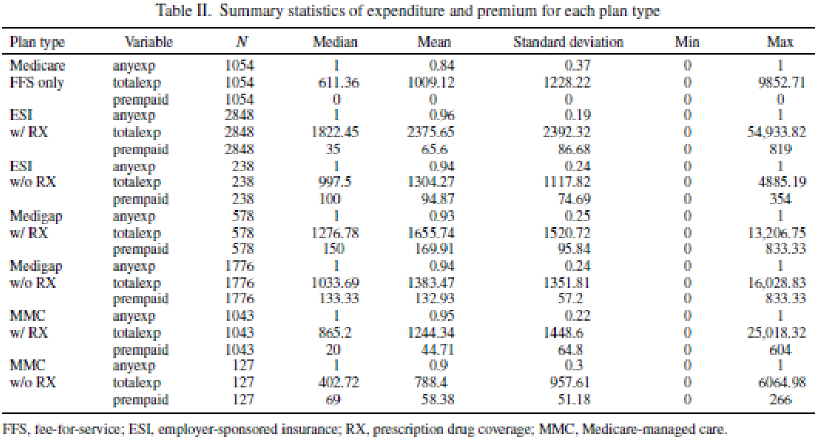


Table II. Summary statistics of expenditure and premium for each plan type

Plan type	Variable	*N*	Median	Mean	Standard deviation	Min	Max
Medicare	anyexp	1054	1	0.84	0.37	0	1
FFS only	totalexp	1054	611.36	1009.12	1228.22	0	9852.71
	prempaid	1054	0	0	0	0	0
ESI	anyexp	2848	1	0.96	0.19	0	1
w/ RX	totalexp	2848	1822.45	2375.65	2392.32	0	54,933.82
	prempaid	2848	35	65.6	86.68	0	819
ESI	anyexp	238	1	0.94	0.24	0	1
w/o RX	totalexp	238	997.5	1304.27	1117.82	0	4885.19
	prempaid	238	100	94.87	74.69	0	354
Medigap	anyexp	578	1	0.93	0.25	0	1
w/ RX	totalexp	578	1276.78	1655.74	1520.72	0	13,206.75
	prempaid	578	150	169.91	95.84	0	833.33
Medigap	anyexp	1776	1	0.94	0.24	0	1
w/o RX	totalexp	1776	1033.69	1383.47	1351.81	0	16,028.83
	prempaid	1776	133.33	132.93	57.2	0	833.33
MMC	anyexp	1043	1	0.95	0.22	0	1
w/ RX	totalexp	1043	865.2	1244.34	1448.6	0	25,018.32
	prempaid	1043	20	44.71	64.8	0	604
MMC	anyexp	127	1	0.9	0.3	0	1
w/o RX	totalexp	127	402.72	788.4	957.61	0	6064.98
	prempaid	127	69	58.38	51.18	0	266

FFS, fee-for-service; ESI, employer-sponsored insurance; RX, prescription drug coverage; MMC, Medicare-managed care.

Average vs. median ATET

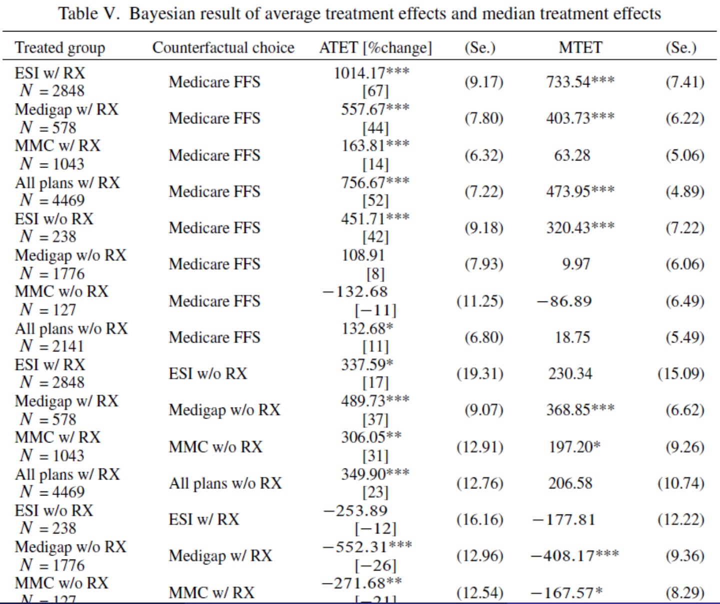
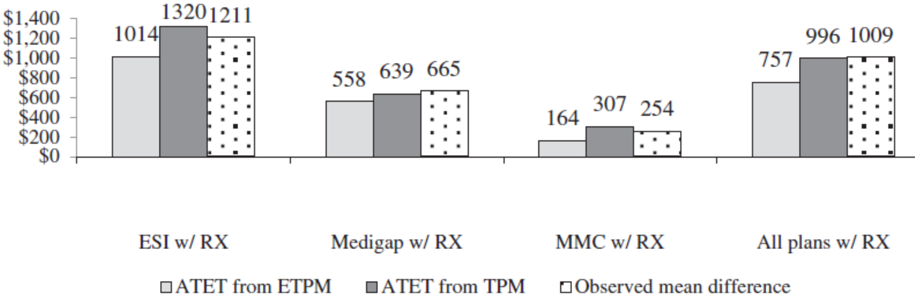


Table V. Bayesian result of average treatment effects and median treatment effects

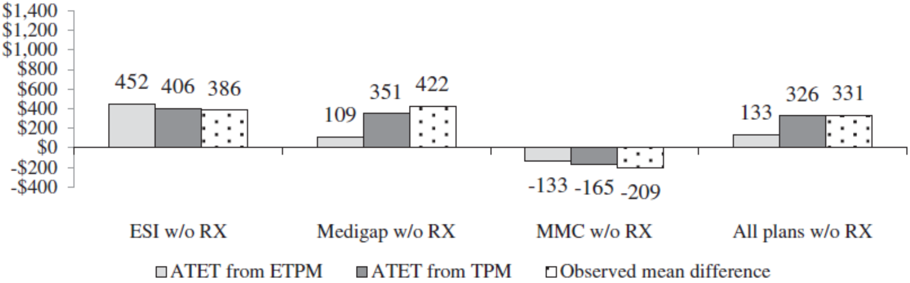
Treated group	Counterfactual choice	ATET [%change]	(Se.)	MTET	(Se.)
ESI w/ RX $N = 2848$	Medicare FFS	1014.17*** [67]	(9.17)	733.54***	(7.41)
Medigap w/ RX $N = 578$	Medicare FFS	557.67*** [44]	(7.80)	403.73***	(6.22)
MMC w/ RX $N = 1043$	Medicare FFS	163.81*** [14]	(6.32)	63.28	(5.06)
All plans w/ RX $N = 4469$	Medicare FFS	756.67*** [52]	(7.22)	473.95***	(4.89)
ESI w/o RX $N = 238$	Medicare FFS	451.71*** [42]	(9.18)	320.43***	(7.22)
Medigap w/o RX $N = 1776$	Medicare FFS	108.91 [8]	(7.93)	9.97	(6.06)
MMC w/o RX $N = 127$	Medicare FFS	−132.68 [−11]	(11.25)	−86.89	(6.49)
All plans w/o RX $N = 2141$	Medicare FFS	132.68* [11]	(6.80)	18.75	(5.49)
ESI w/ RX $N = 2848$	ESI w/o RX	337.59* [17]	(19.31)	230.34	(15.09)
Medigap w/ RX $N = 578$	Medigap w/o RX	489.73*** [37]	(9.07)	368.85***	(6.62)
MMC w/ RX $N = 1043$	MMC w/o RX	306.05** [31]	(12.91)	197.20*	(9.26)
All plans w/ RX $N = 4469$	All plans w/o RX	349.90*** [23]	(12.76)	206.58	(10.74)
ESI w/o RX $N = 238$	ESI w/ RX	−253.89 [−12]	(16.16)	−177.81	(12.22)
Medigap w/o RX $N = 1776$	Medigap w/ RX	−552.31*** [−26]	(12.96)	−408.17***	(9.36)
MMC w/o RX $N = 127$	MMC w/ RX	−271.68** [−21]	(12.54)	−167.57*	(8.29)

Treatment effects of drug plans with different counterfactuals (1)



(a) Treated group: plan with drug coverage; Counterfactual choice: Medicare FFS

Treatment effects of drug plans with different counterfactuals (2)



(b) Treated group: plan without drug coverage; Counterfactual choice: Medicare FFS

Brief remarks about some under-researched topics

Idiosyncratic list of under-researched topics

1. Estimating the distribution of TEs

   Motivation: Policy interventions often impact the full distribution of outcomes, distribution is nonsymmetric, so focus on ATE insufficient

2. TEs in panel data framework

   Motivation: Interventions have short-term and long-term impacts.

3. Joint treatment of multi-valued treatments and multi-valued outcomes

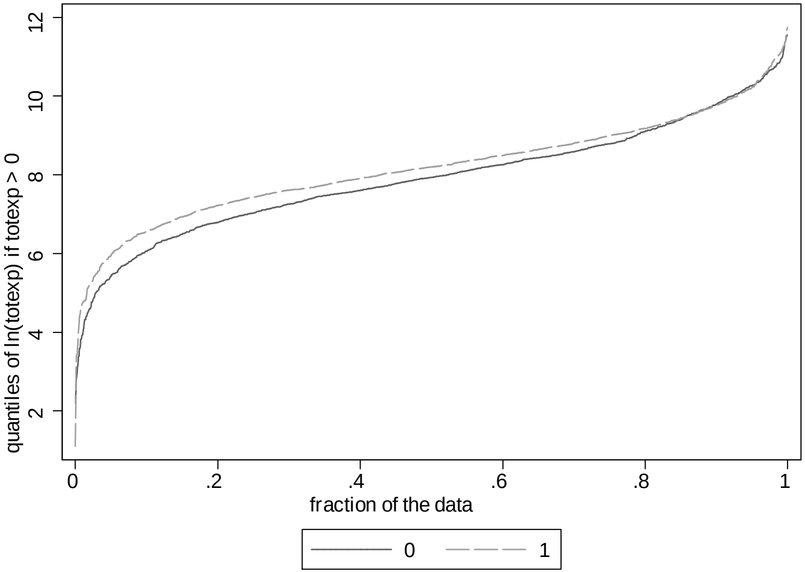
   Motivation: Interventions may target several outcomes and use several interventions simultaneosuly.

Modeling the distribution of outcomes

Three methods are available and well-established

1. Quantile regression for continuous outcomes
    1. continuous/discrete exogenous intervention (qreg)
    2. binary endogenous treatment (ivqte)
2. Bayesian modeling
    1. Posterior distribution of any function of parameters and variables for exogenous or endogenous interventions
3. Regression adjustment
    1. Postestimation **predictive margins** generating conditional distributions for specified configurations of exogenous variables

CQR approach explained



CQR approach explained (2)

- Treatment shifts the distribution of outcomes (vertical axis) as do other covariates.
- The shift can be vertical upwards or downwards, or different at different quantiles (horizontal axis)
- At each quantile of interest we can estimate a conditional quantile function with treatment variable and conditioning covariates.

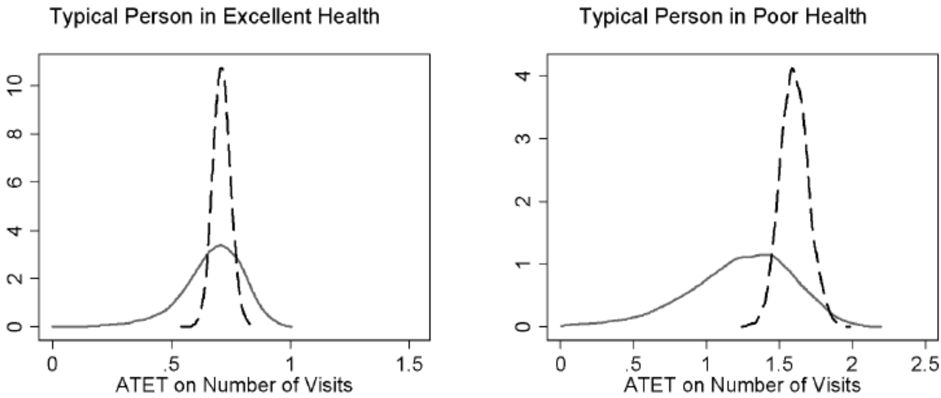
This can be done with a single- or multivalued intervention

- Potential outcomes can be generated at each quantile of interest

Bayesian posterior distribution of ATET, ATE

"We estimate the **posterior distributions of the ATE and LATE**, parameters for a synthetic “representative” individual who is a white 40-year-old male with 13 years of education, with very good health, and without any injuries, physical limitation, or chronic conditions, who is married and has three family members, including himself, annual income of $32,000, living in a metropolitan area in the South, observed in year 2001, and whose employment is with a firm of 146 employees."
From Munkin & Trivedi (*JBES, 2006*)

Posterior distribution of ATET, ATE : Example



From:From Munkin & Trivedi (*JBES, 2006*)

Post. distribution ATET and ATE

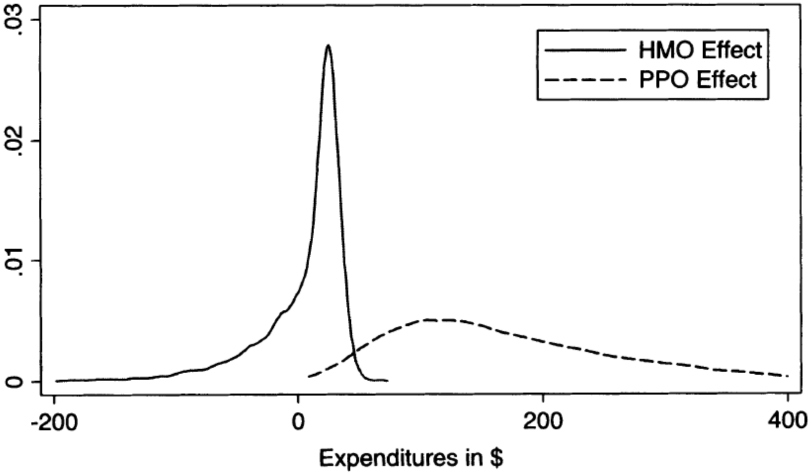


Figure 2. Density of treatment effects for ambulatory expenditures

From: Deb, Munkin, Trivedi: JAE-2006

TE in panel framework

- Panel data allows us to study treatment effects in a dynamic framework allowing for unobserved heterogeneity as well.
- For continuous outcomes and linear models, RA is the obvious approach as it can simultaneously deal with several complications
- For nonlinear models (binary or count outcomes, interval regression, survival models) random effects framework is easier to handle
- Main limitation is that regressor balance is hard to maintain.

MV-outcomes and interventions

- This topic is at an embyonic stage.
- Usual practice is to study a single outcome at any time, ignoring possible dependence between outcomes.
- The standard assumption rules out dependence between outcomes of different subjects.
- For identification may need to rule out also dependence between different outcomes also

Appendix A: Binary Outcome Models and Propensity Scores

Introduction

- Discrete choice or qualitative response models are for y that takes only a finite number of discrete values.
- Here we consider binary outcome models where only two values are taken, 0 and 1.
- Particularly logit and probit models, which are **nonlinear models**.
- Regression models for binary outcomes are constructed to model the conditional probability of a binary discrete outcome
- Examples: Whether to buy a new car; whether to vote for a particular party; whether to choose a particular course
- Topic well covered in most texts.

Outline

- General properties of binary outcome models
- Probit, logit, LPM and OLS models.
- Latent variable formulations, especially random utility model.

Simple binary outcome model

- The coin toss example of introductory statistics.
- Let p denote the probability of a head $(y = 1)$ on one coin toss.
- Then $\Pr[y = 1] = p$ and $\Pr[y = 0] = 1 - p$.
- For N tosses y_i is the i^{th} of N independent realizations of head or tail.
- The MLE for p is the sample mean $\overline{y}$,
  i.e. the proportion of tosses that are heads.

Bernoulli distribution

- $\Pr[y=1]=p$ and $\Pr[y=0]=1-p$.
- Compact expression for density

$$f(y)=p^y(1-p)^{1-y}.$$

- This is Bernoulli density which is the binomial with one trial per observation.
- Moments
  - $E[y]=1\times p+0\times(1-p)=p$
  - $\mathrm{V}[y]=(1-p)\times p+(0-p)\times(1-p)=p(1-p)$.
- Note that p can be interpreted as $E[y]$ or as $\Pr[y=1]$.

Examples

- Economic applications are
  - labor supply: $y = 1$ if work and $y = 0$ otherwise
  - insurance status: $y = 1$ if have private health insurance, $y = 0$ otherwise
- Assumption of independent trials may be reasonable.
- Assuming a constant probability p for each trial is not and expected to depend of an individual's characteristics.
- Extend the Bernoulli model so p_i may be a function of regressors $\mathbf{x}_i$.

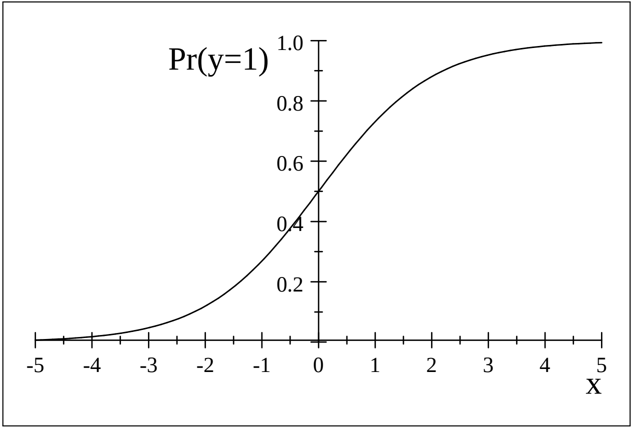
Binary outcome models

- Regression model formed by parameterizing p_i to depend on regressors $\mathbf{x}_i$ and parameters $\boldsymbol{\beta}$.
- Usually specify single-index model

$$\mathsf{E}[y_i|\mathbf{x}_i] = p_i = F(\mathbf{x}_i'\boldsymbol{\beta}).$$

- Usually chose $F(\cdot)$ to be a cumulative distribution function (cdf).
- Then $0 \leq F(\cdot) \leq 1 \Rightarrow 0 \leq p \leq 1$.
  - logistic cdf gives logit model.
  - standard normal cdf gives probit model.

Show $\frac{\exp(x)}{1+\exp(x)}, -\infty < x < +\infty$



Plot of the logistic function

Propensity score

- Given functional form F, and fitted model $F(\mathbf{x}'\widehat{\beta})$, for each $i = 1, ..., N$, we can generate postestimation fitted values

$$\widehat{p}_i(y_i = 1|\mathbf{x}_i) = F(\mathbf{x}_i'\widehat{\beta}).$$

- These fitted values are conditional (fitted) probabilities $E[y_i = 1|\mathbf{x}_i]$. They are referred to as **propensity scores**.
- In the matching literature, the propensity score is a scalar measure of similarity and is an alternative to matching based on the vector $\mathbf{x}$.
- Low values of PS indicate that $y = 1$ is unlikely to be on=bserved.

Maximum likelihood (MLE)

- Density

$$f(y) = p^y(1-p)^{1-y}, \quad p = F(\mathbf{x}'\boldsymbol{\beta})$$
$$\Rightarrow \ln f(y) = y \ln F(\mathbf{x}'\boldsymbol{\beta}) + (1-y)\ln(1-F(\mathbf{x}'\boldsymbol{\beta}))$$

- Log-likelihood function is

$$\mathcal{L}(\boldsymbol{\beta}) = \textstyle\sum_{i=1}^{N} \left\{ y_i \ln F(\mathbf{x}_i'\boldsymbol{\beta}) + (1-y_i)\ln(1-F(\mathbf{x}_i'\boldsymbol{\beta})) \right\}.$$

- Let $F'(z) = \partial F(z)/\partial z$. MLE solves

$$\textstyle\sum_{i=1}^{N} \left\{ \dfrac{y_i}{F(\mathbf{x}_i'\boldsymbol{\beta})} F'(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i + \dfrac{1-y_i}{1-F(\mathbf{x}_i'\boldsymbol{\beta})} F'(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i \right\} = \mathbf{0}.$$

Asy. Distribution OF MLE

- The MLE f.o.c. simplify to

$$\begin{aligned}\textstyle\sum_{i=1}^{N} \frac{y_i - F(\mathbf{x}_i'\boldsymbol{\beta})}{F(\mathbf{x}_i'\boldsymbol{\beta})(1-F(\mathbf{x}_i'\boldsymbol{\beta}))} F'(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i &= \mathbf{0} \\ \textstyle\sum_{i=1}^{N} \left[\frac{y_i - F(\mathbf{x}_i'\boldsymbol{\beta})}{\sqrt{F(\mathbf{x}_i'\boldsymbol{\beta})(1-F(\mathbf{x}_i'\boldsymbol{\beta}))}}\right] \frac{F'(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i}{\sqrt{F(\mathbf{x}_i'\boldsymbol{\beta})(1-F(\mathbf{x}_i'\boldsymbol{\beta}))}} &= \mathbf{0}.\end{aligned}$$

- General ML result if density correctly specified
- For binary outcome MLE

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{ML}} &\overset{a}{\sim} \text{N}\left[\boldsymbol{\beta}_0, \left(-\text{E}[\partial^2\mathcal{L}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}']\big|_{\boldsymbol{\beta}_0}\right)^{-1}\right] \\ &\overset{a}{\sim} \text{N}\left[\boldsymbol{\beta}_0, \left(\textstyle\sum_{i=1}^{N} \frac{1}{F(\mathbf{x}_i'\boldsymbol{\beta}_0)(1-F(\mathbf{x}_i'\boldsymbol{\beta}_0))} F'(\mathbf{x}_i'\boldsymbol{\beta}_0)^2\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right]\end{aligned}$$

Misspecification

- For binary data the dgp density is always Bernoulli as

$$\begin{aligned} \Pr[y=1] &= p \\ \Rightarrow \Pr[y=0] &= 1-\Pr[y=1] = 1-F(\mathbf{x}'\boldsymbol{\beta}). \end{aligned}$$

- Therefore only possible misspecification of dgp is if $p \neq F(\mathbf{x}'\boldsymbol{\beta})$.
- Clearly inconsistent estimator if $p \neq F(\mathbf{x}'\boldsymbol{\beta})$ as then

$$\mathrm{E}[y_i - F(\mathbf{x}_i'\boldsymbol{\beta})] \neq 0$$

leading to left-hand side of f.o.c. not having expected value **0**.

Weighted NLS interpretation

- Since

$$\begin{aligned} \mathrm{E}[y|\mathbf{x}] &= F(\mathbf{x}'\boldsymbol{\beta}) \\ \mathrm{V}[y|\mathbf{x}] &= F(\mathbf{x}'\boldsymbol{\beta})(1-F(\mathbf{x}'\boldsymbol{\beta})) \\ \partial\mathrm{E}[y|\mathbf{x}]/\partial\beta &= F'(\mathbf{x}'\boldsymbol{\beta})\mathbf{x} \end{aligned}$$

  the MLE first-order conditions imply

$$\textstyle\sum_{i=1}^{N} \frac{y_i - \mathrm{E}[y_i|\mathbf{x}_i]}{\mathrm{V}[y_i|\mathbf{x}_i]} \frac{\partial\mathrm{E}[y_i|\mathbf{x}_i]}{\partial\boldsymbol{\beta}} = \mathbf{0}.$$

- Residuals are orthogonal to regressors upon weighting to adjust for heteroskedasticity. i.e. nonlinear WLS .

Logit model

- Logit is a widely used functional form, especially in biometrics
- Computationally convenient.
- The logit model specifies

$$p = \Lambda(\mathbf{x}'\boldsymbol{\beta}) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}'\boldsymbol{\beta}}},$$

- $\Lambda(z) = e^z / (1 + e^z) = 1/(1 + e^{-z})$ is the logistic cdf.
- The derivative $\Lambda'(z) = \Lambda(z)(1 - \Lambda(z))$ is the logistic density.
- For this reason also called logistic regression model .

Logit MLE

- The logit ML conditions simplify to

$$\sum_{i=1}^{N}(y_i - \Lambda(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0}.$$

- F.O.C. are nonlinear in parameters
- Notice that there is no "error term" in the binary model.
- The logit MLE has distribution

$$\widehat{\boldsymbol{\beta}}_{\text{Logit}} \overset{a}{\sim} \mathsf{N}\left[\boldsymbol{\beta}_0, \left(\sum_{i=1}^{N} \Lambda(\mathbf{x}_i'\boldsymbol{\beta}_0)(1 - \Lambda(\mathbf{x}_i'\boldsymbol{\beta}_0))\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right].$$

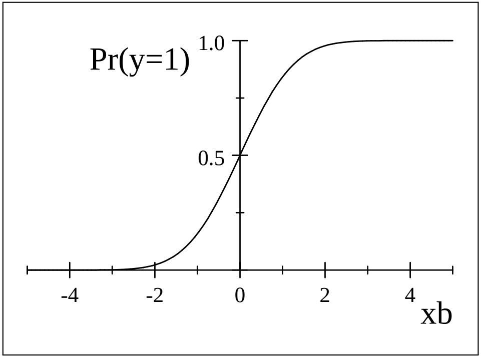
Probit model

- The probit model specifies

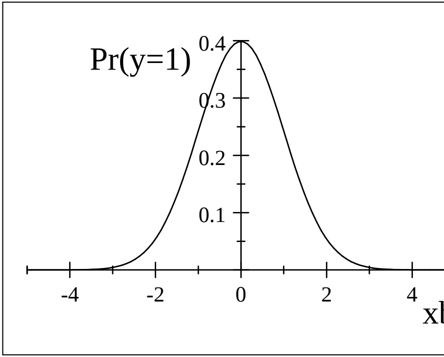
$$p = \Phi(\mathbf{x}'\boldsymbol{\beta}).$$

- $\Phi(z) = \int_{-\infty}^{z} \phi(s)ds = \int_{-\infty}^{z} (1/\sqrt{2\pi}) \exp(-s^2/2) ds$ is the c.d.f. of the standard normal.
- The derivative $\Phi'(z) = \phi(z) = (1/\sqrt{2\pi}) \exp(-z^2/2)$ is the standard normal p.d.f.
- The f.o.c.do not simplify, unlike logit case.
- The probit MLE has distribution

$$\widehat{\boldsymbol{\beta}}_{\text{Probit}} \overset{a}{\sim} \mathsf{N}\left[\boldsymbol{\beta}_0, \left(\sum_{i=1}^{N} \frac{\phi(\mathbf{x}_i'\boldsymbol{\beta}_0)^2}{\Phi(\mathbf{x}_i'\boldsymbol{\beta}_0)(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}_0))}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right].$$



Standard normal c.d.f.



Standard normal p.d.f.

Linear probability model (LPM)

- The LPM specifies
$$p = \mathbf{x}'\boldsymbol{\beta}.$$
- The LPM MLE f.o.c. conditions are
$$\sum\nolimits_{i=1}^{N} \frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\mathbf{x}_i'\boldsymbol{\beta}(1 - \mathbf{x}_i'\boldsymbol{\beta})}\mathbf{x}_i = \mathbf{0},$$
- The LPM model has the obvious weakness of permitting probabilities outside the (0, 1) interval.
- Furthermore, the MLE estimator can be numerically unstable if $\mathbf{x}_i'\boldsymbol{\beta}$ close to 0 or 1.

OLS

- The LPM is more simply estimated by OLS , which also specifies $\mathrm{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$.
- The LPM OLS f.o.c. conditions are

$$\sum_{i=1}^{N}(y_i - \mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i = \mathbf{0},$$

- Allow for the intrinsic heteroskedasticity of binary data

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} \overset{a}{\sim} \mathrm{N}\left[\boldsymbol{\beta}_0, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right]$$

where for Ω use

$$\widehat{\Omega} = \mathrm{Diag}[(y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}})^2]$$

How to interpret coefficients (1)

- Coefficients in different models are not directly comparable **due to different scaling**.
- Instead compare across models effect of a one unit change in regressors on $\mathrm{P}[y=1|\mathbf{x}]=\mathrm{E}[y=1|\mathbf{x}]$.
- Now
$$\begin{aligned}\mathrm{E}[y|\mathbf{x}] &= F(\mathbf{x}'\boldsymbol{\beta}) \\ \partial\mathrm{E}[y|\mathbf{x}]/\partial\mathbf{x} &= F'(\mathbf{x}'\boldsymbol{\beta})\times\boldsymbol{\beta}\end{aligned}$$
  where $F'(z)=\partial F(z)/\partial z$.
- Thus the effect depends on the functional form of F and the evaluation point $\mathbf{x}$, in addition to parameter $\boldsymbol{\beta}$.

How to interpret coefficients (2)

- This suggests for slope parameters the rule of thumb

$$\begin{aligned} \widehat{\beta}_{\text{Logit}} &\simeq 4\widehat{\beta}_{\text{OLS}} \\ \widehat{\beta}_{\text{Probit}} &\simeq 2.5\widehat{\beta}_{\text{OLS}} \\ \widehat{\beta}_{\text{Logit}} &\simeq 1.6\widehat{\beta}_{\text{Probit}}. \end{aligned}$$

- This works quite well, for $0.1 \leq F(\mathbf{x}'\beta) \leq 0.9$.
- Better to compare marginal effects, not coefficients.

Odds ratio calculations for logit (1)

- For the logit model

$$\begin{aligned} p &= \exp(\mathbf{x}'\boldsymbol{\beta})/(1+\exp(\mathbf{x}'\boldsymbol{\beta}) \\ \Rightarrow \tfrac{p}{1-p} &= \exp(\mathbf{x}'\boldsymbol{\beta}) \\ \Rightarrow \ln \tfrac{p}{1-p} &= \mathbf{x}'\boldsymbol{\beta}. \end{aligned}$$

- $p/(1-p)$ is the odds ratio which measures the probability that $y=1$ relative to the probability that $y=0$.
- E.g. Pharmaceutical drug study where $y=1$ denotes survival and $y=0$ denotes death. An odds ratio of 2 means that the odds of survival are twice those of death.

Odds ratio calculations for logit (2)

- Statistical analyses and packages offer the option of printing the odds ration $p/(1-p) = \exp(\mathbf{x}'\boldsymbol{\beta})$.
- Suppose the j^{th} regressor increases by one unit.
  Then $\mathbf{x}'\boldsymbol{\beta}$ increases to $\mathbf{x}'\boldsymbol{\beta}+\beta_j$.
  And $\exp(\mathbf{x}'\boldsymbol{\beta})$ increases to $\exp(\mathbf{x}'\boldsymbol{\beta}+\beta_j) = \exp(\mathbf{x}'\boldsymbol{\beta}) \times \exp\left(\beta_j\right)$.
- Thus the odds ratio has increased by a multiple $\exp\left(\beta_j\right)$.
- E.g. a logit slope parameter of 0.1 means that a one unit change in the regressor increases the odds ratio by a multiple $\exp(0.1) \simeq 0.105$. The relative probability of $y = 1$ has increased by 10.5 percent.
- This interpretation widely used in applied biostatistics.

Semi-elasticity interpretation

- For economists it is more natural to interpret β_j as a semi-elasticity for the odds ratio, since $\ln p/(1-p) = \mathbf{x}'\boldsymbol{\beta}$.
- Then a logit slope parameter of 0.1 means that a one unit change in the regressor increases the odds ratio by a multiple 0.1.
- This coincides exactly with the interpretation used in statistics for very small of β_j, since then $\exp(\beta_j) = \beta_j$.

Which functional form?

- Which F – logit, probit or linear probability?
- Theoretically it depends on the data generating process (dgp).
- Unlike other applications of ML there is no problem in specifying the distribution – the only possible distribution for a $(0, 1)$ variable is the Bernoulli.
- The problem lies in specifying a functional form for the parameter of this distribution.
- If the dgp has $p_i = \Lambda(\mathbf{x}_i' \boldsymbol{\beta}_0)$ then a logit model should be used, and estimators based on other models such as probit are potentially inconsistent.
- Similar conclusions hold if instead for the dgp has $p_i = \Phi(\mathbf{x}_i' \boldsymbol{\beta}_0)$ or $p_i = \mathbf{x}_i' \boldsymbol{\beta}_0$.

Why logit?

- Logit model is the binary model used by statisticians :
    - F.o.c. and asymptotic distribution are relatively simple.
    - Logit model corresponds ises the canonical link function for the binomial, a generalized linear model.
    - Coefficients can be interpreted in terms of the log-odds ratio.
    - Easy generalization to multinomial logit.
    - A discriminant analysis interpretation can be given.

Why Probit?

- The probit model is often used by economists .
  - It is motivated by a latent normal random variable.
  - So ties in with tobit models and multinomial probit.
- Empirically, either logit and probit can be used
  - little difference between results from probit and logit analysis, once rescale parameter estimates.
  - Greatest difference is in prediction of probabilities close to 0 or 1.

Why LPM?

- The LPM should not be used as probabilities outside the (0, 1) interval and be numerically unstable.
- Nonetheless OLS can be useful for preliminary data analysis.
- Very widely used in the context of endogenous binary variable
- In practice standard errors of slope coefficients are often quite similar across logit, probit and OLS (even using the incorrect $s^2(\mathbf{X}'\mathbf{X})^{-1}$ in the case of OLS).
- Final results should, however, use probit or logit.

Measuring the fit of the model

- Several measures of model adequacy have been proposed.
- Many are very specific to binary outcome models.
- There is no single best measure . See Amemiya (1981) and Maddala (1983).
- Approaches:
  - R-squared measures.
  - Compare $\widehat{y}$ with y.
  - Compare predicted $\widehat{\Pr}[y=1]$ with actual $\Pr[y=1]$.

Pseudo-R-squared

- There are many R-squareds for binary models as R^2 in linear model has many interpretations.
- McFadden proposed two. We favor McFadden (1974)

$$R^2 = 1 - \frac{\mathcal{L}_{fit}}{\mathcal{L}_0},$$

  where

  - $\mathcal{L}_{fit}$ = log-likelihood in the fitted model
  - $\mathcal{L}_0$ is the log-likelihood in the intercept-only model.

- This R^2 should be only used for discrete choice models.

- In other nonlinear models instead use

$$R^2 = 1 - (\mathcal{L}_{\max} - \mathcal{L}_{fit})/(\mathcal{L}_{\max} - \mathcal{L}_0),$$

  where $\mathcal{L}_{\max}$ is the maximum possible value of the log-likelihood.
- For binary outcome models $\mathcal{L}_{\max} = 0$.
- For some other models $\mathcal{L}_{\max}$ can be unbounded restricting use of this.

Predicting y=1

- Many measures compare predicted $\widehat{y}$ with y.
  - The problem is in defining a rule for when $\widehat{y} = 1$.
  - Obvious is $\widehat{y} = 1$ when $\widehat{p} = F(\mathbf{x}'\widehat{\beta}) > 0.5$.
  - But this can e.g. yield $\widehat{y} = 0$ all the time if most of the sample has $y = 0$.

Predicting Pr[y=1]

- Can compare predicted $\widehat{\Pr}[\mathbf{y} = \mathbf{1}]$ with $\Pr[\mathbf{y} = \mathbf{1}]$.
- But testing whether on on average the predicted probabilities equal the sample frequencies is not helpful over the entire sample, since for the logit model with an intercept the f.o.c. imply $\sum_{i=1}^{N} y_i - \Lambda(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, so that $\sum_{i=1}^{N} \widehat{p}_i = \bar{y}$.
- Useful for subsamples.

Random utility models (1)

- In the random utility formulation a consumer selects the choice with highest utility .
- The discrete variable y
  - takes value 1 if choice 1 has higher utility
  - takes value 0 if choice 0 has higher utility.
- The random utility model specifies the utilities of alternatives 0 and 1 to be

$$\begin{array}{c} U_0 = \mu_0 + \varepsilon_0 = \mathbf{x}'\boldsymbol{\beta}_0 + \varepsilon_0 \\ U_1 = \mu_1 + \varepsilon_1 = \mathbf{x}'\boldsymbol{\beta}_1 + \varepsilon_1 \end{array}$$

where

  - μ_0 and μ_1 are deterministic components of utility, whose dependence on regressors is detailed below.
  - ε_0 and ε_1 are random components of utility.

Random utility models (2)

- The alternative with highest utility is chosen. So the observed choice is
$$\begin{aligned}\Pr[y=1] &= \Pr[U_1 > U_0] \\ &= \Pr[\mu_1+\varepsilon_1 > \mu_0+\varepsilon_0] \\ &= \Pr[\varepsilon_0-\varepsilon_1 < \mu_1-\mu_0] \\ &= F(\mu_1-\mu_0),\end{aligned}$$
where F is the cdf of $(\varepsilon_0-\varepsilon_1)$.
- Different distributions of ε_0 and ε_1 give different discrete choice models.

Random utility models (3)

- Binary probit arises if ε_0 and ε_1 are normal, as is readily seen by noting that then $(\varepsilon_0 - \varepsilon_1)$ is normally distributed, upon normalization of the variance of $(\varepsilon_0 - \varepsilon_1)$ to unity.
- Binary logit model arises if ε_0 and ε_1 are type I extreme value distributed, defined soon, as then the difference $(\varepsilon_0 - \varepsilon_1)$ can be shown to be logistic distributed.
- The random component ε in utility model is needed. Otherwise, choice would be deterministic, with alternative 1 always chosen if $\mu_1 > \mu_0$.

Appendix B: Nonparametric density and regression estimation

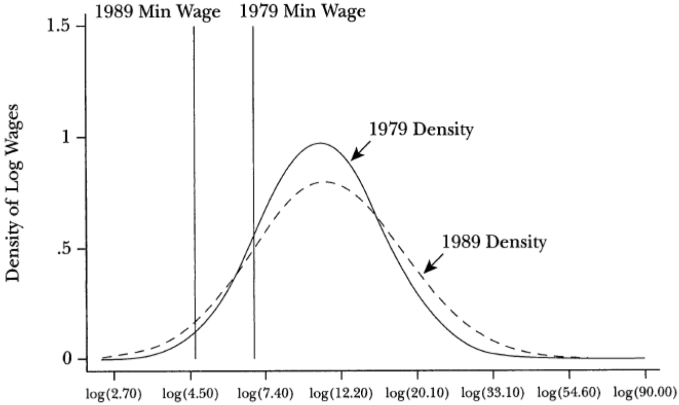
Outline

1. Motivating NP methods
2. The histogram estimator
3. Nonparametric kernel density estimator
4. Nonparametric regression estimators
5. Stata Commands

Parametric or non-P?

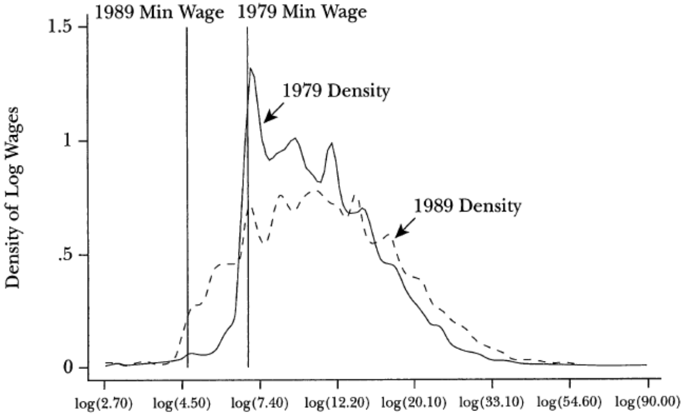
- We are often interested in looking at the key features of a distribution of some variable
- If the focus is on wage distribution, we may want to see more than just the mean
- In comparing distribution differences and changes over time, a visual tool is helpful and suggestive.

See example from DiNardo and Tobias (JEP, 2001)



Women's Wages 1979 and 1989: A Parametric View
(2000 Constant Dollars)

Why nonparametrics?



Women's Wages 1979 and 1989: A Nonparametric View
(2000 Constant Dollars)

Parametric and nonparametric regression

▶ Up to now, all regression models relied on functions (or densities) which depend on an unknown finite-dimensional parameter.

◦ A finite-dimensional parameter is an element of R^q with $q < N$.

◦ For example, linear regression models use an additive combination of the covariates (x_0).

◦ Nonlinear regression models specify a known function of (a linear index of) the covariates.

Parametric vs. nonparametric methods

- ML theory is based on a assumption about the density of the data, which depends on a finite dimensional parameter.
- If the functional form or the distributional assumption is wrong, the parameter estimators of these models are inconsistent, however.
- To circumvent this kind of misspecification problem due to assumptions about functional form, nonparametric methods can be used .
- Nonparametric density and nonparametric regression estimators are the base for most nonparametric econometric models.

Density estimation

1. Parametric density estimation:

▶ Assume a density and use estimated parameters of this density

e.g. normal density estimate: assume $y_i \sim N[\mu, \sigma^2]$ and use $N[\overline{y}, s^2]$.

▶ Nonparametric density estimate: a histogram

- ○ break data into bins and use relative frequency within each bin
- ○ Problem: a histogram is a step function, even if data are continuous

2.Smooth nonparametric density estimate: kernel density estimate.

- ○ Kernel density estimate smooths a histogram in two ways:
- ○ use overlapping bins so evaluate at many more points
- ○ use bins of greater width with most weight at the middle of the bin.

Empirical density (1)

▶ Let the data $z_1, ..., z_n$ be a sample from the distribution of a random vector Z

▶ Interested in the general problem of estimating the distribution of Z nonparametrically, i.e. without restricting it to belong to a known parametric family.

▶ First consider how to estimate nonparametrically the density function of Z.

▶ Distribution function (cdf) and the density function are equivalent ways of representing the distribution of Z, but there may be advantages in analyzing a density:

○ The graph of a density may be easier to interpret if one is interested in aspects such as symmetry or multimodality.

○ Estimates of certain population parameters, such as the mode, are more easily obtained from an estimate of the density.

Uses of NP density (2)

Use of nonparametric density estimates
Nonparametric density estimates may be used for:

○ Exploratory data analysis.

○ Estimating qualitative features of a distribution (e.g. unimodality, skewness, etc.).

○ Specification and testing of parametric models.

• If $Z = (X, Y)$, they may be used to construct a nonparametric estimate of the conditional mean function (CMF) of Y given X

Histogram (1)

- A **histogram** is a "naive" estimate of the density
- Method: Split the range of y into equally spaced intervals and calculate the fraction of the sample in each interval.
- More formally: consider estimation of the density $f(X = x_0)$ of a scalar continuous random variable X evaluated at x_0.

Since the density is the derivative of the cdf $F(X = x_0)$, i.e. $f(X = x_0) = dF(x_0)/dx$, we have

$$\begin{aligned} f(x_0) &= \lim_{h\to\infty} \frac{F(x_0+h) - F(x_0-h)}{2h} \\ &= \lim_{h\to\infty} \frac{\Pr\left[x_0 - h < x < x_0 + h\right]}{2h}. \end{aligned}$$

Histogram (2)

- For a sample $\{x_i,\ i = 1, ..., N\}$ of size N, use the estimator

$$\widehat{f}_{\mathrm{HIST}}(x_0) = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{1}(x_0 - h < x_i < x_0 + h)}{2h}, \tag{15}$$

where the indicator function

$$\mathbf{1}(A) \quad = \begin{cases} 1 \text{ if condition } A \text{ is satisfied} \\ \qquad\qquad 0 \text{ otherwise.} \end{cases}$$

- The estimator $\widehat{f}_{\mathrm{HIST}}(x_0)$ is a histogram estimate centered at x_0 with bin width $2h$,
- Since it equals the fraction of the sample that lies between $x_0 - h$ and $x_0 + h$ divided by the bin width $2h$.
- If $\widehat{f}_{\mathrm{HIST}}$ is evaluated over the range of x at equally spaced values of x each $2h$ units apart, it yields a histogram.

Histogram (3)

- The estimator $\widehat{f}_{\mathrm{HIST}}(x_0)$ gives all observations in $x_0 \pm h$ equal weight as

$$\widehat{f}_{\mathrm{HIST}}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} \frac{1}{2} \times \mathbf{1}\left(\left| \frac{x_i - x_0}{h} \right| < 1 \right). \tag{16}$$

- This "naive" density estimate is a step function, even if the underlying density is continuous.
- Smoother estimates can be obtained by using weighting functions other than the indicator function above.
- Choosing intervals of different widths can produce rather different looking figures
- See the example below.

Limitations of the histogram estimator

- Although widely used, the histogram estimator has several drawbacks

• The results depend on the choice of the range $(a_0, b_0]$.
• Given $(a_0, b_0]$, the results also depend on the number of bins J or, equivalently, on the bin width h.

-
  - For example, given the data, increasing J (reducing h) tends to give a histogram that is only informative about the location of the distinct sample points.
  - Reducing J (increasing h) eventually leads to a completely uninformative rectangle.
  - However, J may safely be increased if the sample size N also increases.

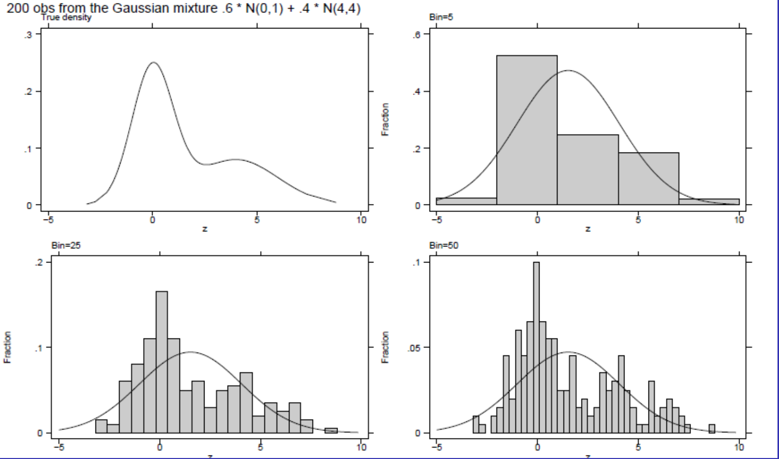
Limitations of the histogram estimator (2)

- Keeping the bin width h fixed over the range of the data can lead to loss of detail at points where the data are clustered
- If h is reduced to deal with this problem, then estimates may appear noisy where data are sparse.
- Histogram is a step function with jumps at the end of each bin, so cannot incorporate prior information on the degree of density smoothness.
- Method is problematic if we want derivatives of the density

Limitations of the histogram estimator

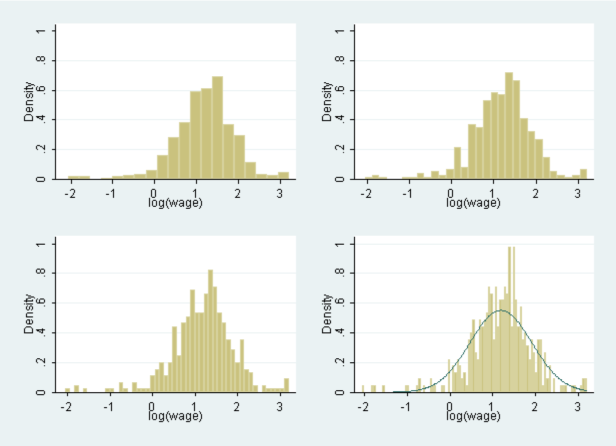
- *"The shape of the histogram can potentially be influenced by where you place the bin centers. Moreover, with a histogram, choosing the width of the bins and the location of the first bin also determines the choice of bin centers."*
  "The histogram assigns equal weight to all points falling in the bin"

Figure 1: Histogram estimates of density.



Histogram examples (2)

data



$_8$2.*pdf*

The graphs show histograms of the logarithm of wage (from the data set Mroz.dta). The histogram of the top left panel divides the data in 20 classes (the default value of the histogram command of Stata), that of the top right panel uses 30 classes, and those of the bottom left and right panels use 50 and 100 classes, respectively. For the histogram of the bottom right panel, a normal density fitted to the data is added (blue line).

The histogram method

- The histogram method is a useful tool for exploratory data analysis, but has some undesirable features,

• the need to choose a partition of the range of Z into cells,
• the density estimates of f are not smooth.
Now consider a method that tries to overcome these two problems.

Kernel density estimator

- *What is a kernel? It is merely a smoothing or weight-assigning function.*
- Consider the empirical density (2). Putting $a = z - h$ and $b = z + h$, where h is a small positive number, gives

$$\widehat{f}(z) = \frac{1}{2Nh} \sum_{i=1}^{N} \mathbf{1}\left(z - h < Z < z + h\right). \tag{17}$$

which is the fraction of sample points falling in the interval $(z - h, z + h]$ divided by the length $2h$ of the interval.

- An advantage of this method over the histogram method is that there is no need to partition the range of Z into cells.

However, $\widehat{f}(z)$ still has two drawbacks: (1) estimates depends on the constant h (2) $\widehat{f}(z)$ is a step function with jump points $z = Z_i \pm h$

- Can get smooth kernel density estimates if we modify $\widehat{f}(z)$ such that estimates of f are smooth.

Smooth kernel density estimator (1)

- Now $\widehat{f}(z)$ may also be written as

$$\widehat{f}(z) = \frac{1}{Nh} \sum_{i=1}^{N} w\left(\frac{z - Z_i}{h}\right)$$

where

$$w(u) = \begin{cases} 1/2 \text{ if } -1 < u < 1 \\ 0 \text{ otherwise.} \end{cases}$$

is a symmetric bounded non-negative function that integrates to one and corresponds to the density of a uniform distribution on the interval $[-1, 1]$.

- $\widehat{f}(z)$ is not smooth because it is a sum of step functions.
- If we replace w by a smooth function K, we get a smooth estimate of f because it is a sum of smooth functions.

Smooth kernel density estimator (2)

- The **kernel density estimator** generalizes the histogram estimate (16) by using an alternative weighting function, so
$$\widehat{f}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right). \tag{18}$$
- The weighting function $K(\cdot)$ is called a kernel function and satisfies certain restrictions.
- The kernel function $K(\cdot)$ is a piecewise continuous function, symmetric around zero, which integrates to unity, and satisfies additional boundedness conditions.
- The parameter h is a smoothing parameter called the bandwidth, and two times h is the window width.
- The density is estimated by evaluating $\widehat{f}(x_0)$ at a wider range of values of x_0 than used in forming a histogram – usually evaluation is at the sample values $x_1, ..., x_N$.
- A typical kernel density estimator proceeds by using the formula for the general kernel density where the function $K(0)$ is replaced by one of the

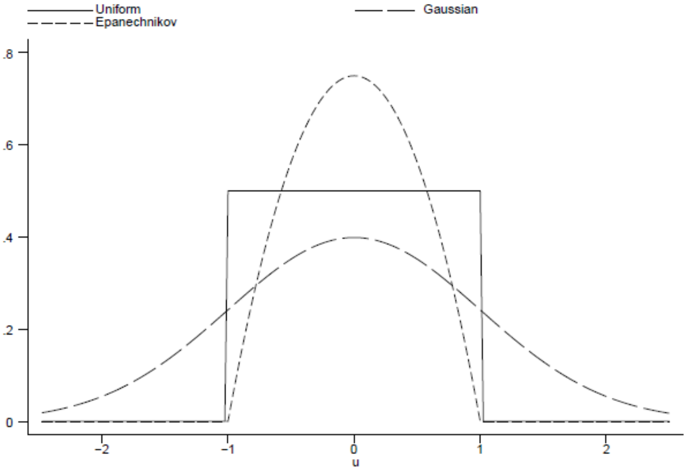
Kernel density estimator

Some commonly-used kernel functions are:

Kernel	**Kernel Function** $K(z)$	δ
Uniform (or box or rectangular)	$\frac{1}{2} \times \mathbf{1}(\lvert z \rvert < 1)$	1.3510
Triangular (or triangle)	$(1 - \lvert z \rvert) \times \mathbf{1}(\lvert z \rvert < 1)$	-
Epanechnikov (or quadratic)	$\frac{3}{4}(1 - z^2) \times \mathbf{1}(\lvert z \rvert < 1)$	1.7188
Gaussian (or normal)	$(2\pi)^{-1/2} \exp(-z^2/2)$	0.7764

- Uniform kernel uses same weights as a histogram of bin width $2h$, except that it produces a running histogram which is evaluated at a series of points x_0 rather than using fixed bins.
- Different kernels merely change the relative weights.
- Given $K(\cdot)$ and h the estimator is very simple to implement. If the kernel estimator is evaluated at r distinct values of x_0 then computation of the kernel estimator requires at most Nr operations, when the kernel has unbounded support.

Kernel weights



Gaussian kernel example

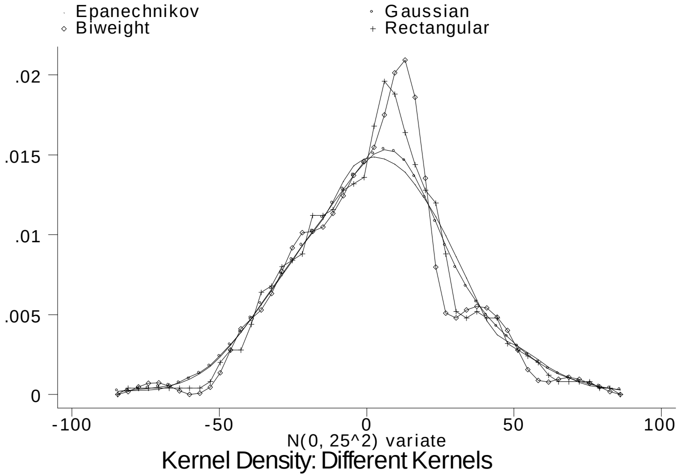
$$f(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}} \right) \exp \left(-\frac{1}{2} \left[\frac{(x_i - x_0)^2}{h} \right] \right)$$

- If we were estimating the probability density function at x_0, the most weight would go to observations at x_0.
- Why? Because the value of this kernel is maximized at $x_i = x_0$.
- Because the "support" of this kernel is the entire real line, we use all the data to estimate the probability density function at x_0.
- However, the weight we assign observations far away from x_0 with a Gaussian kernel is quite small.

Kernel Density Example

- A random sample of size 100 drawn from the $\mathcal{N}[0, 25^2]$ distribution.
- For the particular sample drawn the sample mean is 2.81 and the sample standard deviation is 25.27.
- Figure shows the effect of different kernels for given choice of bandwidth, here $h = 12.5$, ignoring possible adjustment of bandwidth for different kernels
- The Gaussian kernel gives quite similar results to the Epanechnikov.
- The other two kernels, the biweight and the rectangular are not nearly as smooth. As already noted the rectangular uses the same weight function as the histogram, and produces a running histogram.
- The variation in density estimate with kernel choice is not as great as the variation with bandwidth choice..

Kernel density example



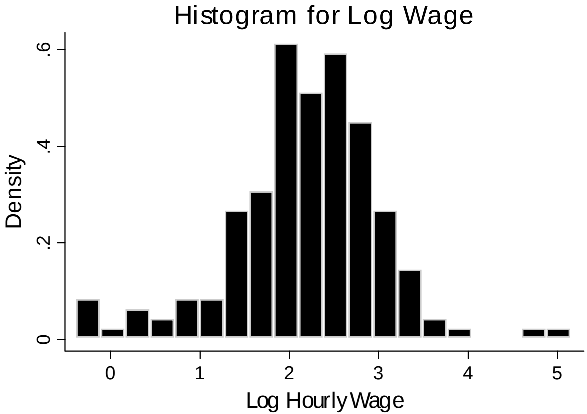
Nonparametric Density Example

- Data: hourly wage and education for 175 women aged 36 years who worked in 1993.
- Data from the Michigan Panel Survey of Income Dynamics (PSID).
- A histogram of the natural logarithm of wage.
- The bin width is chosen so that there are 30 bins, each of width about 0.20.

Nonparametric density example (2)

- This is an unusually narrow bin width for only 175 observations, but many details are lost with a larger bin width.

log-wage data possibly slightly left-skewed.



Nonparametric density example (2)

- The kernel density estimate based on the Stata `kdensity` command, which uses the Epanechnikov kernel (Stata default kernel).
- Choice of bandwidth.
  - Stata selects a default bandwidth of $h = 0.21$.
  - The kernel estimate is a weighted average of observations that have log wage within 0.21 of the log wage at the current point of evaluation, with more weight placed on data closest to the current point of evaluation.
  - Figure shows three kernel density estimates, with bandwidths of 0.07, 0.21 and 0.63, corresponding to one-third the default, the default, and three times the default bandwidth.
  - Smallest bandwidth is too small as it leads to too jagged a density estimate.
  - Largest bandwidth oversmooths the data.
  - The goldilocks choice is the default value of 0.21, which gives a smooth estimate.

Choice of bandwidth

"*Choice of bandwidth essentially involves a trade-off between bias (misreporting the shape) and variance (lack of precision) of the estimates. Intuitively, the larger the bandwidth, the "smoother" the resulting estimates (lower variance), but we may have oversmoothed the true density and thus obtained a biased estimate of that density. Note that this is a problem for histograms as well....*"

- MSE$[\widehat{\theta}]$ =E$[\widehat{\theta}-\theta]^2$ =E$[\widehat{\theta}-$ E$[\widehat{\theta}]$ $+$ E$[\widehat{\theta}]-\theta]^2$ =var$[\widehat{\theta}]+\left(bias[\widehat{\theta}]\right)^2$
- Optimality criterion balances bias and variance using a mean squared error type criterion. **At a specific point** y the MSE criterion is

$$\begin{aligned} MSE(h) &= \mathrm{E}\left(\left[\widehat{f}(y)-f(y)\right]^2\right) \\ &= \left(\mathrm{E}\left[\widehat{f}(y)-\mathrm{E}f(y)\right]\right)^2+\mathrm{var}\left[\widehat{f}(y)\right] \end{aligned}$$

Choice of bandwidth (2)

- A practitioner is interested in the **global** or **total MSE** at all values of y. The relevant measure then is mean integrated squared error (MISE) defined as

$$\begin{aligned} MISE(h) &= \mathrm{E}\left(\int \left[\widehat{f}(y) - f(y)\right]^2 dy\right) \\ &= \left(\int \mathrm{E}\left[\widehat{f}(y) - \mathrm{E}f(y)\right]^2 dy + \int \mathrm{var}\left[\widehat{f}(y)\right] dy\right) \end{aligned}$$

  where the first term corresponds to the squared bias and the second to the sampling variance.
- The optimal value of h, h_{opt}, is that which minimizes MISE, i.e. the one that provides the best trade-off between bias and variance
- But to calculate h_{opt} we need to calculate the expectations, which requires knowledge of the true distribution of y!!
- It can be shown that h_{opt} depends upon (i) the true density function and how it fluctuates, (ii) the choice of the kernel, and (iii) the sample size.

Automatic choice of bandwidth

- Silverman rule-of-thumb bandwidth:

$$\widetilde{h}_{opt} = 1.059\sigma N^{-1/5} \text{ if the reference distribution is } N[\mu, \sigma^2]$$

- If the reference distribution is not Normal, and/or there are outliers in the data, a preferred alternative is

$$\widetilde{h}_{opt} = 0.9N^{-1/5}\left(min\left\{\widehat{\sigma}, \frac{q_3 - q_1}{1.349}\right\}\right)$$

where $q_3 - q_1$ is the interquartile range (the difference between the 75th and 25th percentile) and 1.349 is the *iqr* of the standard normal.

- As is preferred, this bandwidth gets smaller as N - the number of observations- increases, but does not go to zero "too fast."

Choice of bandwidth (2)

- The optimal bandwidth varies with the kernel. The optimal kernel is the Epanechnikov, though this advantage is slight.
- While there is little loss in using the simplest kernels such as uniform and triangular, smoother kernels such as Epanechnikov and Gaussian are preferred as they lead to smoother kernel density estimates.
- Bandwidth choice is much more important and the optimal value varies with the kernel.
- In practice one uses Silverman's plug-in estimate or its variants
- These plug-in estimates for h work well in practice, especially for symmetric unimodal densities, and even if $f(x)$ is not the normal density. Usual to check for sensitivity by using variations such as twice and half the plug-in estimate.

For a discussion of cross-validation (CV) and Adaptive Methods see Ahamada and Flachaire (2010,chapter 1)

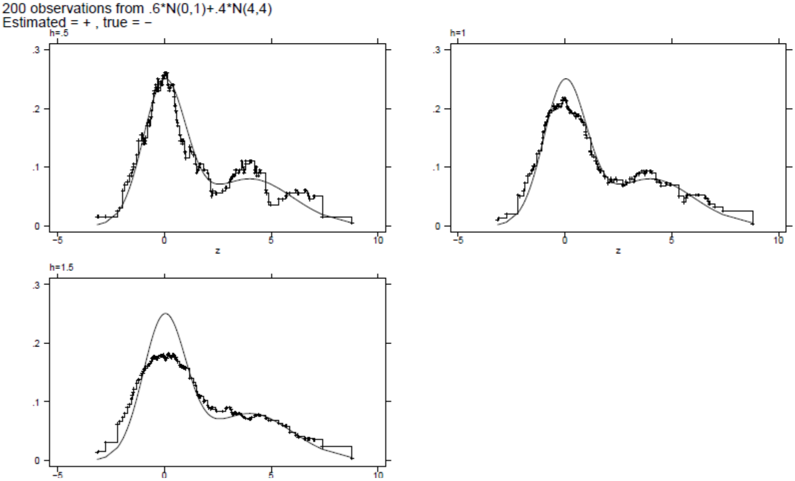
Nonparametric density example (3)

Possible uses of kernel density estimate?

1. Comparison to the normal, by superimposing a normal density with mean equal to the sample mean and variance equal to the sample variance.
2. A second possibility is to compare log wage kernel density estimates for different subgroups.

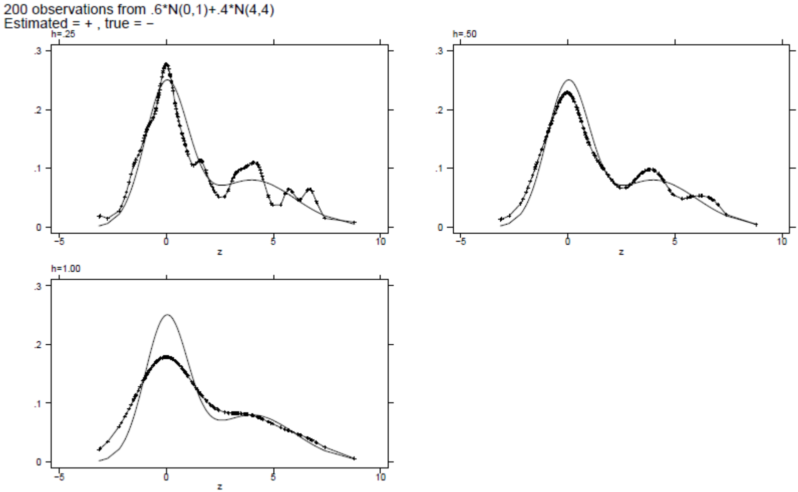
Kernel density examples

Figure 2: Uniform kernel density estimates.



A canonical example

Figure 3: Gaussian kernel density estimates.



Nonparametric regression

- Consider regression of scalar dependent variable y on a scalar regressor variable x. The regression model is

$$\begin{aligned} & y_i = m(x_i) + \varepsilon_i, \quad i = 1, ..., N, \\ & \varepsilon_i \sim \text{iid } [0, \sigma_\varepsilon^2]. \end{aligned} \tag{19}$$

- There are no functional form assumptions and no distributional assumptions
- Our task is to consider methods for estimating the function m in the regression equation
- A nonparametric method widely used is the **lowess local regression** method, a **lo**cal **we**ighted average estimator similar to kernel regression that instead uses a variable bandwidth.

Lowess regression (1)

- Formally, the local linear regression estimator of $m(x_0)$ is given as the α^*, which minimizes the weighted least squares function:

$$min_{\alpha_o,\alpha_1} \sum_i^N [(y_i - \alpha_0 - \alpha_1(x_i - x_0))^2 K((x_i - x_0)/h_n)],$$

with K denoting the kernel and h_n the bandwidth.

- A local weighted regression line at each point x is fitted using centered subsets that include the closest $0.8N$ observations, the Stata default, where N is the sample size, and the weights decline as we move away from the center point.
- Near the end-point lowest and highest values of x smaller uncentered subsets are used.

Lowess regression (2)

- The **local weighted average estimator** takes the form

$$\widehat{m}(x_0) = \sum_{i=1}^{N} w_{i0,h} y_i, \tag{20}$$

  where the weights $w_{i0,h} = w(x_i, x_0, h)$,sum to one, so $\sum_i w_{i0,h} = 1$.
- The weights are specified to be relatively large (small) for values of x_i close to (far from) x_0.
- As h becomes smaller $\widehat{m}(x_0)$ becomes less biased, as only observations close to x_0 are being used, but more variable, as fewer observations are being used. The parameter h is generic notation for a **window width parameter**, with smaller values of h leading to a smaller window with most weight placed on observations with x_i close to x_0.

OLS & Lowess

- The OLS predictor for the linear regression model is a weighted average of y_i, since some algebra yields

$$\widehat{m}_{\mathrm{OLS}}(x_0) = \sum_{i=1}^{N} \left\{ \frac{1}{N} + \frac{(x_0 - \bar{x})(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} \right\} y_i .$$

  The OLS weights, however, can actually increase as the distance between x_0 and x_i increases if, for example, $x_i > x_0 > \bar{x}$.

- **Local regression** instead uses weights that are decreasing in $|x_i - x_0|$.

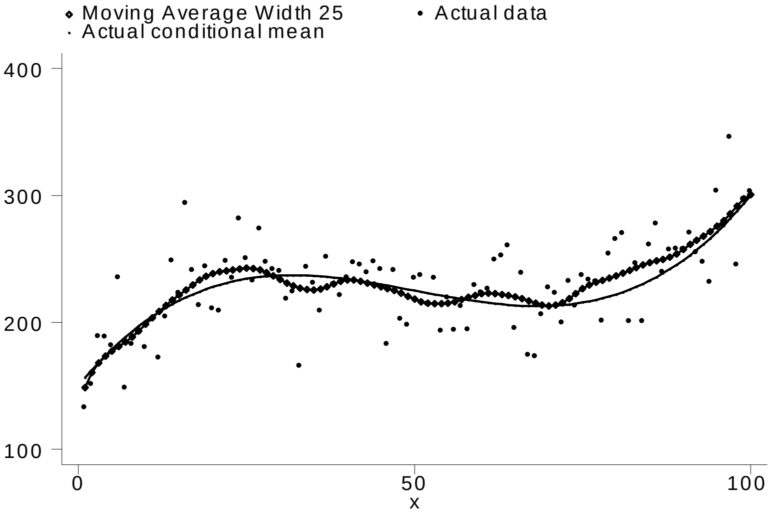
Lowess regression example

- As an illustration, consider data generated from the model

$$\begin{aligned} y_i &= 150 + 6.5x_i - 0.15x_i^2 + 0.001x_i^3 + \varepsilon_i, \quad i = 1, ..., 100, (21) \\ x_i &= i, \\ \varepsilon_i &\sim \mathcal{N}[0, 25^2]. \end{aligned}$$

- The Lowess estimator provides a smooth estimate of $m(x)$ as it uses kernel weights rather than an indicator function

Lowess regression example



Kernel Density Estimates with Bandwidth 25

Figure plots the Lowess estimate with $k = 25$. This local regression estimate is

Kernel Regression (1)

- Kernel regression is a weighted average estimator using kernel weights. Issues such as bias and choice of bandwidth presented for kernel density estimation are also relevant here.
- The goal is to estimate the regression function $m(x)$ in the model $y = m(x) + \varepsilon$
- Thus more generally we consider a kernel weighting function $K(\cdot)$.
- This yields the **kernel regression estimator**

$$\widehat{m}(x_0) \equiv \frac{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right) y_i}{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)}. \qquad (22)$$

Kernel regression (2)

- Several common kernel functions, uniform, Gaussian, Epanechnikov, quadratic and quartic, have already been given. The uniform kernel leads to the roughest estimates, the Gaussian has the computational advantage of not having to compute the indicator function $\mathbf{1}\left(|z| < 1\right)$, and as stated earlier the Epanichnikov is optimal.
- This estimator is called Nadaraya & Watson estimator.
- The kernel regression estimator is a special case of the weighted average with weights

$$w_{i0,h} = \frac{\frac{1}{Nh} K\left(\frac{x_i - x_0}{h}\right)}{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)} \tag{23}$$

which by construction sum to one.

Curse of dimensionality

- Extending the NPR to a general *k*-dimensional regression is difficult because of the "curse of dimensionality."
- CoD essentially implies that in a high dimensional regression we will encounter many empty "hyperspaces" - regions with no observations unless the sample size is very large.
- The required sample size increases exponentially with the dimension making NPR not very practicable.
- NPR can be used if one has just one or two regressors.
- Alternatively, one might choose to apply the nonparametric approach to a subset of regressors only, e.g. partial linear regression.

Summary

Major limitations for applied work

1. the lack of a commonly accepted method to choose an appropriate bandwidth
    1. typical expressions for an "optimal" bandwidth involve unknown properties of the function we are trying to estimate
2. the lack of a simple way to compute reliable standard errors.
3. Curse of dimensionality which restricts modeling to low dimensions