Introduction to Propensity Score Matching: A Review and Illustration

Shenyang Guo, Ph.D.
School of Social Work
University of North Carolina at Chapel Hill
January 28, 2005

For Workshop Conducted at the School of Social Work, University of Illinois – Urbana-Champaign

NSCAW data used to illustrate PSM were collected under funding by the Administration on Children, Youth, and Families of the U.S. Department of Health and Human Services. Findings do not represent the official position or policies of the U.S. DHHS. PSM analyses were funded by the Robert Wood Johnson Foundation Substance Abuse Policy Research Program, and by the Children's Bureau's research grant. **Results are preliminary and not quotable.** Contact information: sguo@email.unc.edu

Outline

Day 1

- Overview:
  - Why PSM?
  - History and development of PSM
  - Counterfactual framework
  - The fundamental assumption
- General procedure
- Software packages
- Review & illustration of the basic methods developed by Rosenbaum and Rubin

Outline (continued)

- Review and illustration of Heckman's difference-in-differences method
  - Problems with the Rosenbaum & Rubin's method
  - Difference-in-differences method
  - Nonparametric regression
  - Bootstrapping

Day 2

- Practical issues, concerns, and strategies
- Questions and discussions

PSM References

Check website:

http://sswnt5.sowo.unc.edu/VRC/Lectures/index.htm

(Link to file “Day1b.doc”)

Why PSM? (1)

Need 1: Analyze causal effects of treatment from observational data

- Observational data - those that are not generated by mechanisms of randomized experiments, such as surveys, administrative records, and census data.
- To analyze such data, an ordinary least square (OLS) regression model using a dichotomous indicator of treatment does not work, because in such model the error term is correlated with explanatory variable.

Why PSM? (2)

$$Y_i = \alpha + \tau W_i + X_i^{'} \beta + \varepsilon_i$$

The independent variable w is usually correlated with the error term ε. The consequence is inconsistent and biased estimate about the treatment effect τ.

Why PSM? (3)

Need 2: Removing Selection Bias in Program Evaluation

- Fisher's randomization idea.
- Whether social behavioral research can really accomplish randomized assignment of treatment?
- Consider $E(Y_1|W=1) - E(Y_0|W=0)$. Add and subtract $E(Y_0|W=1)$, we have

  $\{E(Y_1|W=1) - E(Y_0|W=1)\} + \{E(Y_0|W=1) - E(Y_0|W=0)\}$

  Crucial: $E(Y_0|W=1) \neq E(Y_0|W=0)$
- The debate among education researchers: the impact of Catholic schools vis-à-vis public schools on learning. The Catholic school effect is the strongest among those Catholic students who are less likely to attend Catholic schools (Morgan, 2001).

Why PSM? (4)

Heckman & Smith (1995) Four Important Questions:

- What are the effects of factors such as subsidies, advertising, local labor markets, family income, race, and sex on program application decision?
- What are the effects of bureaucratic performance standards, local labor markets and individual characteristics on administrative decisions to accept applicants and place them in specific programs?
- What are the effects of family background, subsidies and local market conditions on decisions to drop out from a program and on the length of time taken to complete a program?
- What are the costs of various alternative treatments?

History and Development of PSM

- The landmark paper: Rosenbaum & Rubin (1983).
- Heckman's early work in the late 1970s on selection bias and his closely related work on dummy endogenous variables (Heckman, 1978) address the same issue of estimating treatment effects when assignment is nonrandom.
- Heckman's work on the dummy endogenous variable problem and the selection model can be understood as a generalization of the propensity-score approach (Winship & Morgan, 1999).
- In the 1990s, Heckman and his colleagues developed difference-in-differences approach, which is a significant contribution to PSM. In economics, the DID approach and its related techniques are more generally called nonexperimental evaluation, or econometrics of matching.

The Counterfactual Framework

- **Counterfactual**: what would have happened to the treated subjects, had they not received treatment?
- The key assumption of **the counterfactual framework** is that individuals selected into treatment and nontreatment groups have potential outcomes in both states: the one in which they are observed and the one in which they are not observed (Winship & Morgan, 1999).
- For the treated group, we have observed mean outcome under the condition of treatment $E(Y_1|W=1)$ and unobserved mean outcome under the condition of nontreatment $E(Y_0|W=1)$. Similarly, for the nontreated group we have both observed mean $E(Y_0|W=0)$ and unobserved mean $E(Y_1|W=0)$.

The Counterfactual Framework (Continued)

- Under this framework, an evaluation of

$$E(Y_1|W=1) - E(Y_0|W=0)$$

can be thought as an effort that uses $E(Y_0|W=0)$ to estimate the counterfactual $E(Y_0|W=1)$. The central interest of the evaluation is not in $E(Y_0|W=0)$, but in $E(Y_0|W=1)$.

- The real debate about the classical experimental approach centers on the question: whether $E(Y_0|W=0)$ really represents $E(Y_0|W=1)$?

Fundamental Assumption

- **Rosenbaum & Rubin (1983)**

$$(Y_0, Y_1) \perp W \mid X.$$

- Different versions: “unconfoundedness” & “ignorable treatment assignment” (Rosenbaum & Robin, 1983), “selection on observables” (Barnow, Cain, & Goldberger, 1980), “conditional independence” (Lechner 1999, 2002), and “exogeneity” (Imbens, 2004)

General Procedure

Run Logistic Regression:

• Dependent variable: Y=1, if participate; Y = 0, otherwise.

•Choose appropriate conditioning (instrumental) variables.

• Obtain propensity score: predicted probability (p) or log[(1-p)/p].

Either

- 1-to-1 or 1-to-n match and then stratification (subclassification)
- Kernel or local linear weight match and then estimate Difference-in-differences (Heckman)

Or

1-to-1 or 1-to-n Match

- Nearest neighbor matching
- Caliper matching
- Mahalanobis
- Mahalanobis with propensity score added

Multivariate analysis based on new sample

Nearest Neighbor and Caliper Matching

- Nearest neighbor: $C(P_i) = \min_j | P_i - P_j |, \quad j \in I_0$

  The nonparticipant with the value of Pj that is closest to Pi is selected as the match.
- Caliper: A variation of nearest neighbor: A match for person i is selected only if

  where ε is a pre-specified toler $| P_i - P_j | < \varepsilon, \quad j \in I_0$

  Recommended caliper size: .25σp
- 1-to-1 Nearest neighbor within caliper (The is a common practice)
- 1-to-n Nearest neighbor within caliper

Mahalanobis Metric Matching: (with or without replacement)

- Mahalanobis without p-score: Randomly ordering subjects, calculate the distance between the first participant and all nonparticipants. The distance, d(i,j) can be defined by the Mahalanobis distance:

$$d(i, j) = (u - v)^T C^{-1} (u - v)$$

where u and v are values of the matching variables for participant i and nonparticipant j, and C is the sample covariance matrix of the matching variables from the full set of nonparticipants.

- Mahalanobis metric matching with p-score added (to u and v).
- Nearest available Mahalandobis metric matching within calipers defined by the propensity score (need your own programming).

Stratification (Subclassification)

Matching and bivariate analysis are combined into one procedure (no step-3 multivariate analysis):

- Group sample into five categories based on propensity score (quintiles).
- Within each quintile, calculate mean outcome for treated and nontreated groups.
- Estimate the mean difference (average treatment effects) for the whole sample (i.e., all five groups) and variance using the following equations:

$$\hat{\delta} = \sum_{k=1}^{K} \frac{n_k}{N} \left[\bar{Y}_{0k} - \bar{Y}_{1k} \right],$$

$$Var(\hat{\delta}) = \sum_{k=1}^{K} \left(\frac{n_k}{N}\right)^2 Var[\bar{Y}_{0k} - \bar{Y}_{1k}]$$

Multivariate Analysis at Step-3

We could perform any kind of multivariate analysis we originally wished to perform on the unmatched data. These analyses may include:

- multiple regression
- generalized linear model
- survival analysis
- structural equation modeling with multiple-group comparison, and
- hierarchical linear modeling (HLM)

As usual, we use a dichotomous variable indicating treatment versus control in these models.

Very Useful Tutorial for Rosenbaum & Rubin's Matching Methods

D'Agostino, R.B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine 17*, 2265-2281.

Software Packages

- There is currently no commercial software package that offers formal procedure for PSM. In SAS, Lori Parsons developed several Macros (e.g., the GREEDY macro does nearest neighbor within caliper matching). In SPSS, Dr. John Painter of Jordan Institute developed a SPSS macro to do similar works as GREEDY (http://sswnt5.sowo.unc.edu/VRC/Lectures/index.htm).
- We have investigated several computing packages and found that PSMATCH2 (developed by Edwin Leuven and Barbara Sianesi [2003], as a user-supplied routine in STATA) is the most comprehensive package that allows users to fulfill most tasks for propensity score matching, and the routine is being continuously improved and updated.

Demonstration of Running STATA/PSMATCH2:

Part 1. Rosenbaum & Rubin's Methods (Link to file "Day1c.doc")

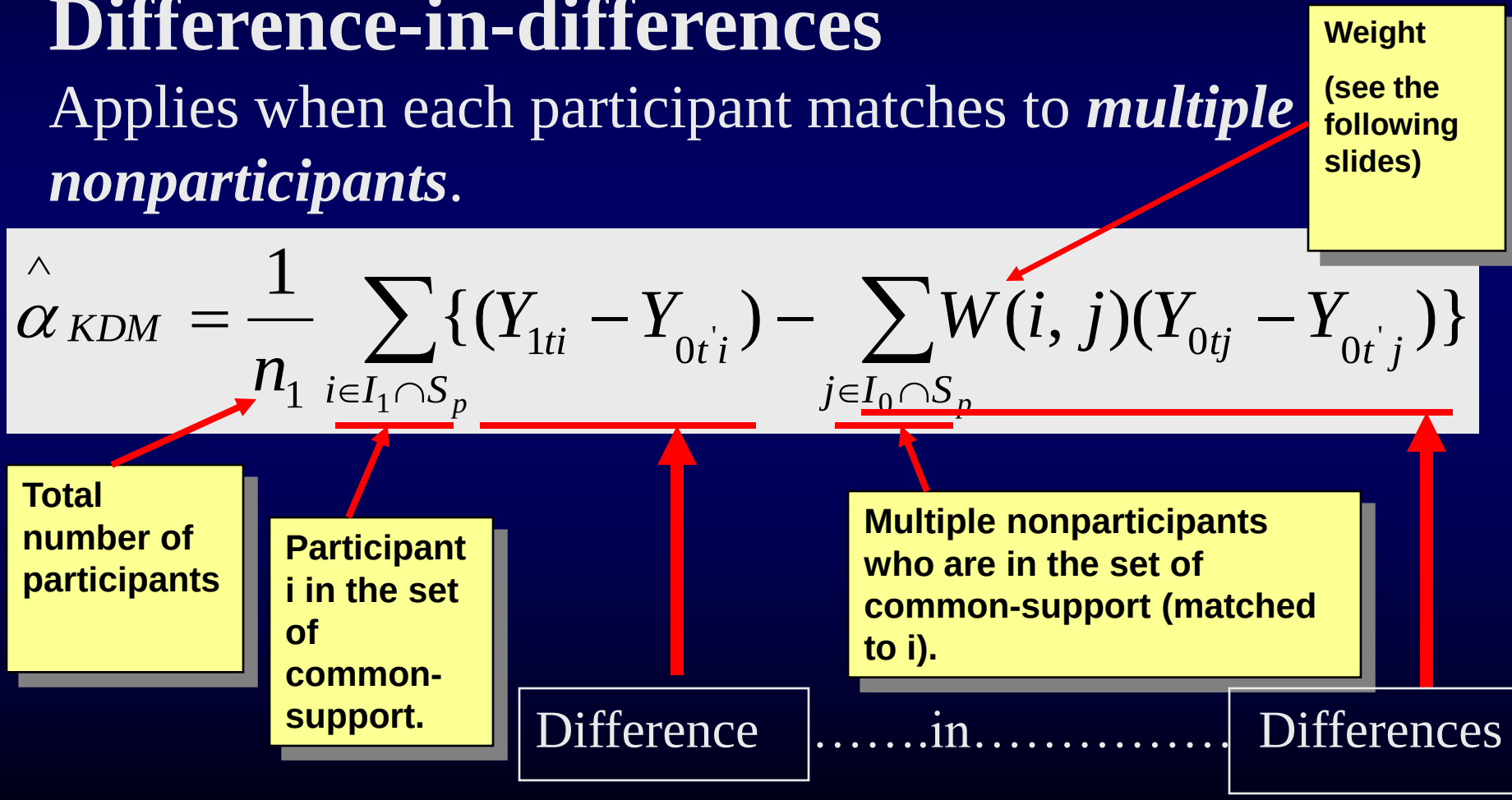
Problems with the Conventional (Prior to Heckman's DID) Approaches

- Equal weight is given to each nonparticipant, though within caliper, in constructing the counterfactual mean.
- Loss of sample cases due to 1-to-1 match. What does the resample represent? External validity.
- It's a dilemma between inexact match and incomplete match: while trying to maximize exact matches, cases may be excluded due to incomplete matching; while trying to maximize cases, inexact matching may result.

Heckman's Difference-in-Differences Matching Estimator (1)

Difference-in-differences

Applies when each participant matches to ***multiple nonparticipants***.



Heckman's Difference-in-Differences Matching Estimator (2)

Weights W(i.,j) (distance between i and j) can be determined by using one of two methods:

1. Kernel matching:

$$W(i,j) = \frac{G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n}\right)}$$

where G(.) is a kernel function and α_n is a bandwidth parameter.

Heckman's Difference-in-Differences Matching Estimator (3)

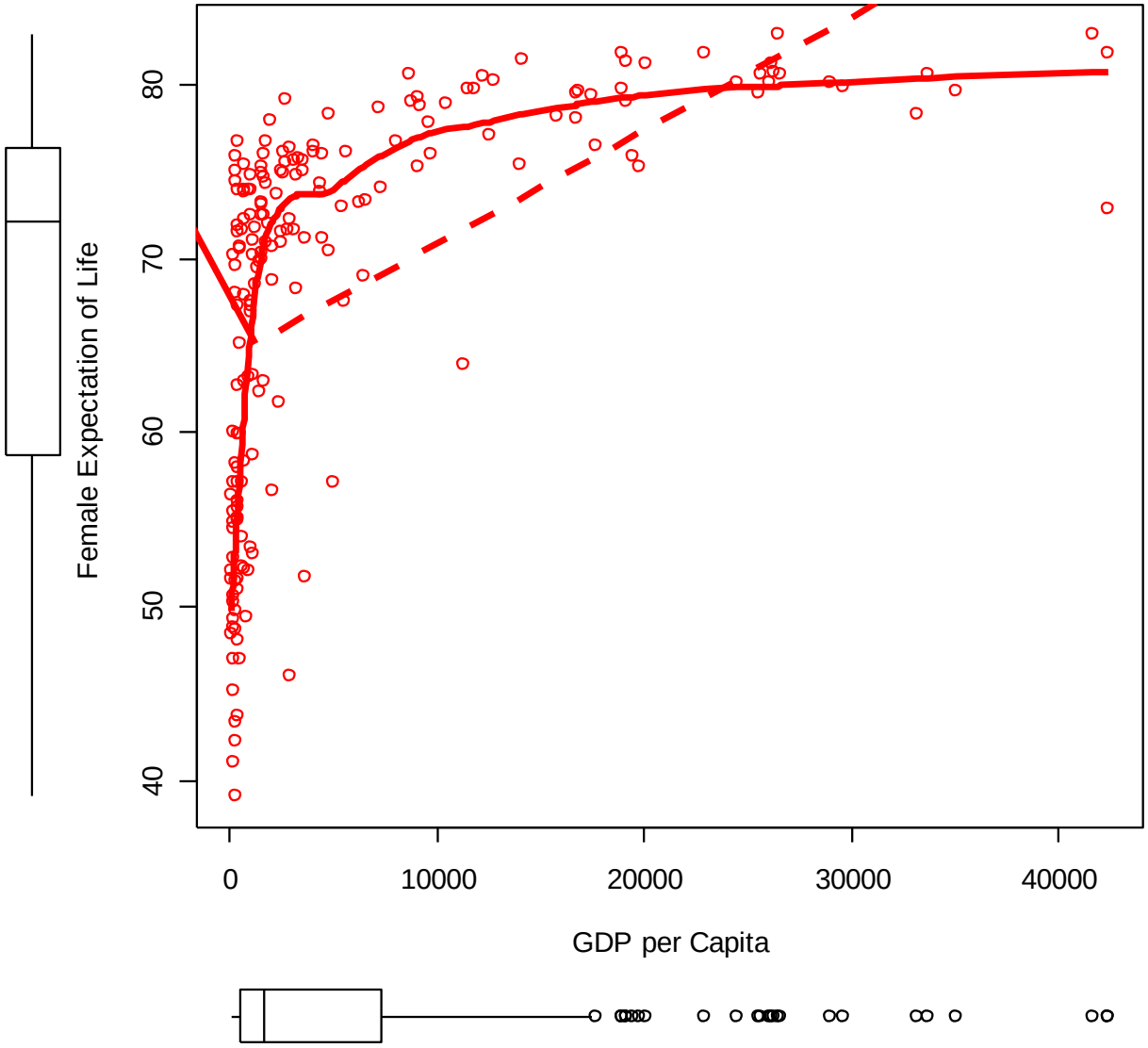
2. Local linear weighting function (***lowess***):

$$W(i,j)=\frac{G_{ij}\sum_{k\in I_0}G_{ik}\left(P_k-P_i\right)^2-\left[G_{ij}\left(P_j-P_i\right)\right]\left[\sum_{k\in I_0}G_{ik}\left(P_k-P_i\right)\right]}{\sum_{j\in I_0}G_{ij}\sum_{k\in I_0}G_{ij}\left(P_k-P_i\right)^2-\left(\sum_{k\in I_0}G_{ik}\left(P_k-P_i\right)\right)^2}$$

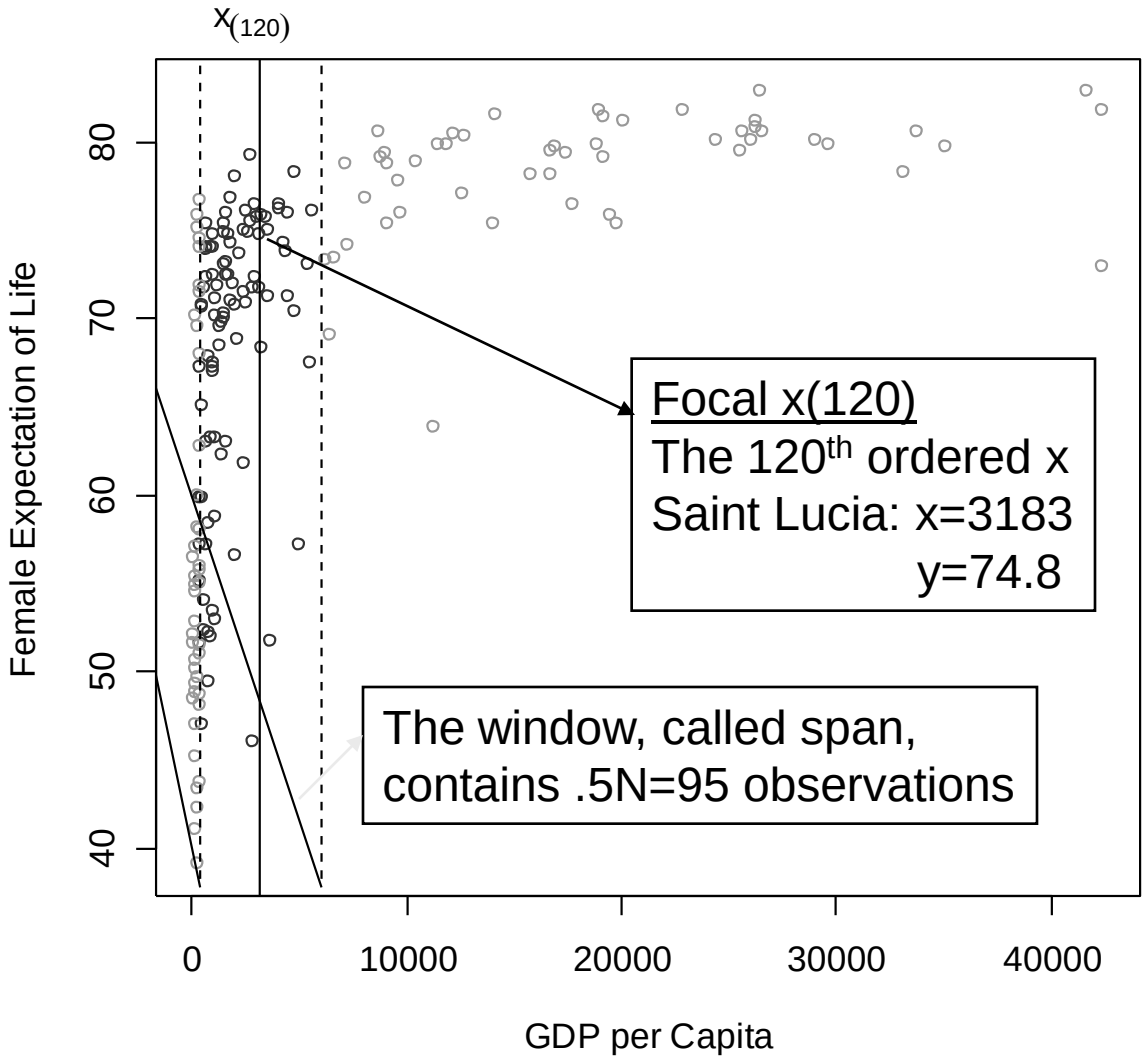
A Review of Nonparametric Regression (Curve Smoothing Estimators)

I am grateful to John Fox, the author of the two Sage green books on nonparametric regression (2000), for his provision of the R code to produce the illustrating example.

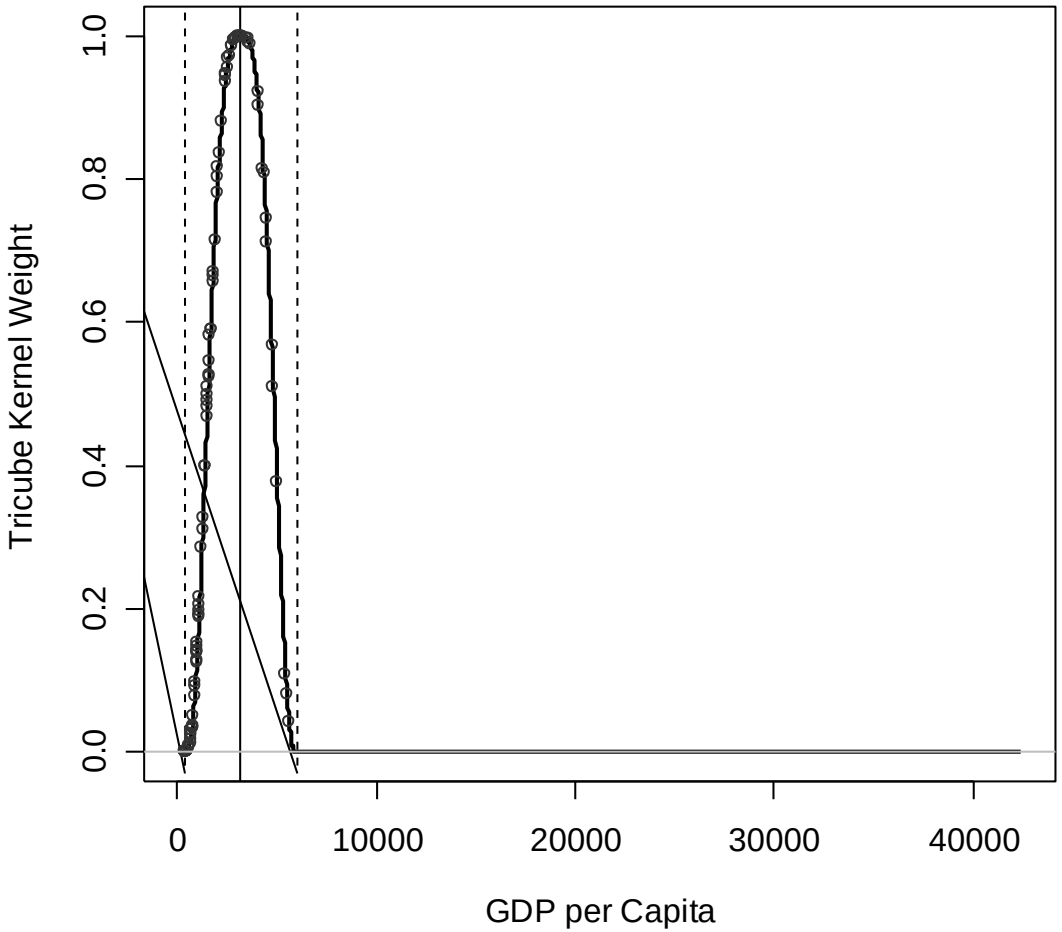
Why Nonparametric? Why Parametric Regression Doesn't Work?



The Task: Determining the Y-value for a Focal Point X(120)



Weights within the Span Can Be Determined by the Tricube Kernel Function

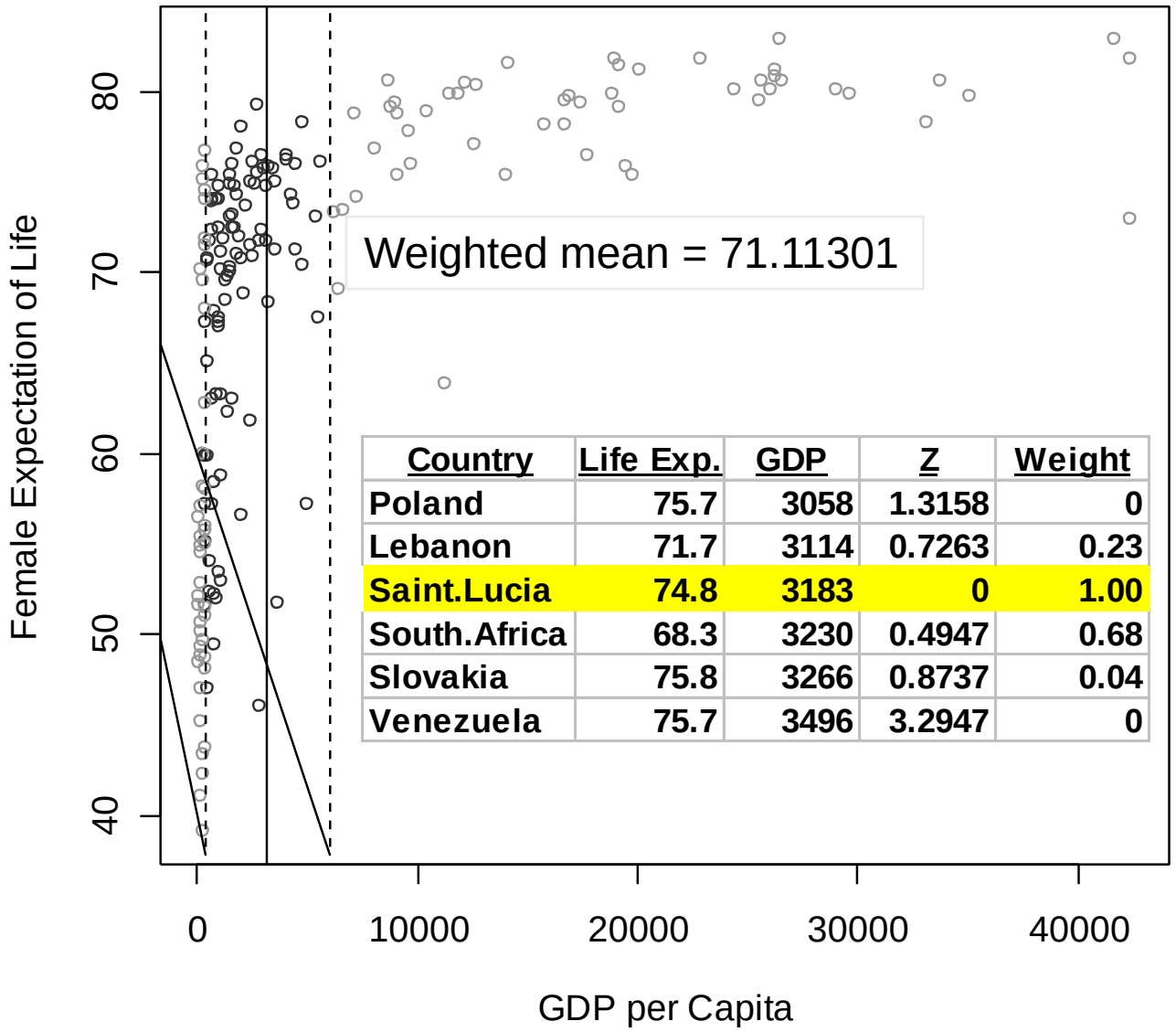


Tricube kernel weights

$$z_i = (x_i - x_0)/h$$

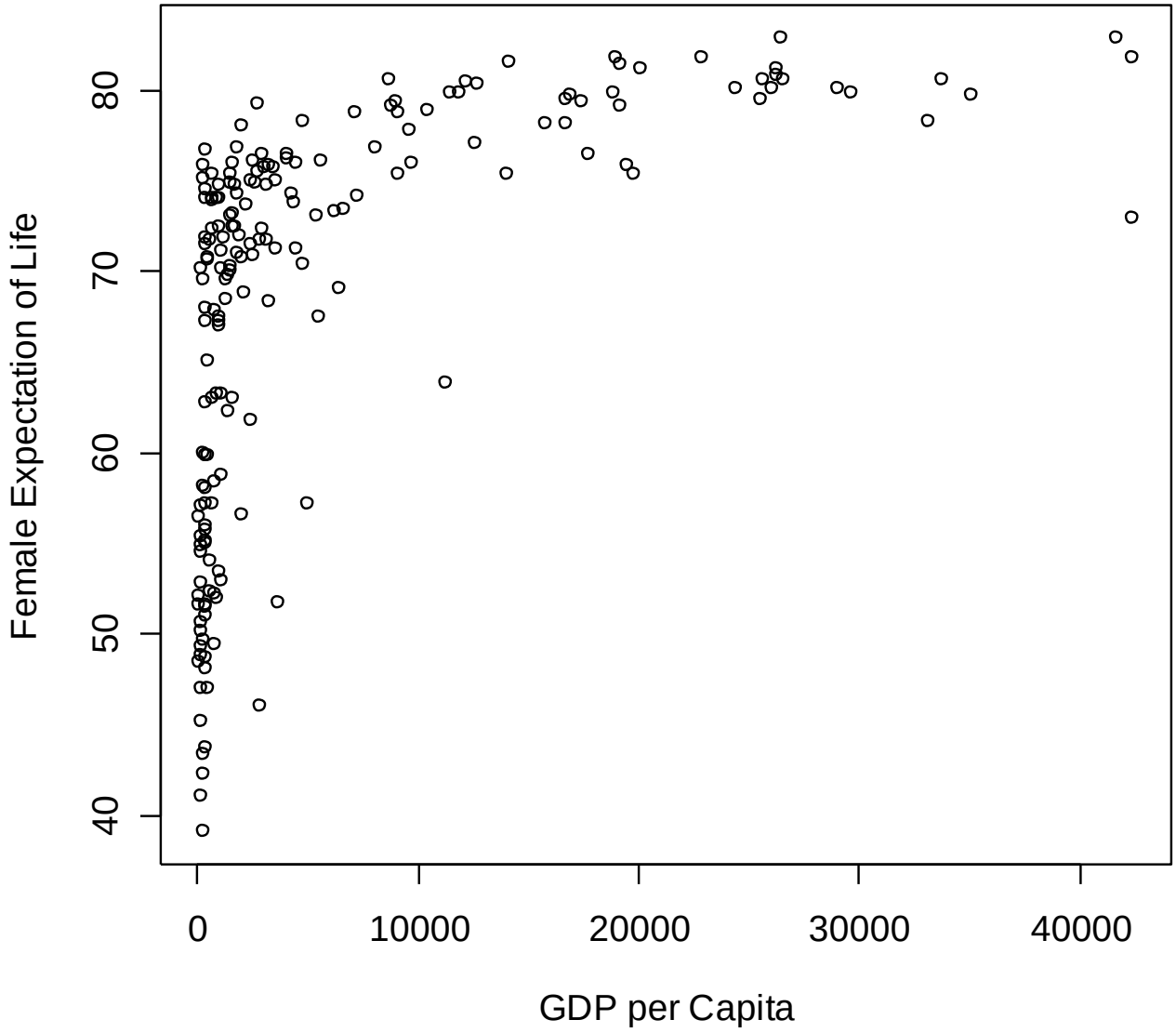
$$K_T(z) = \begin{cases} (1-|z|^3)^3 \ldots\ldots\ldots for |z|<1 \\ 0 \ldots\ldots\ldots\ldots\ldots for |z|\geq 1 \end{cases}$$

The Y-value at Focal X(120) Is a Weighted Mean



Country	Life Exp.	GDP	Z	Weight
Poland	75.7	3058	1.3158	0
Lebanon	71.7	3114	0.7263	0.23
Saint.Lucia	74.8	3183	0	1.00
South.Africa	68.3	3230	0.4947	0.68
Slovakia	75.8	3266	0.8737	0.04
Venezuela	75.7	3496	3.2947	0

The Nonparametric Regression Line Connects All 190 Averaged Y Values



Review of Kernel Functions

- Tricube is the default kernel in popular packages.
- Gaussian normal kernel:

$$K_N(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

- Epanechnikov kernel – parabolic shape with support [-1, 1]. But the kernel is not differentiable at z=±1.
- Rectangular kernel (a crude method).

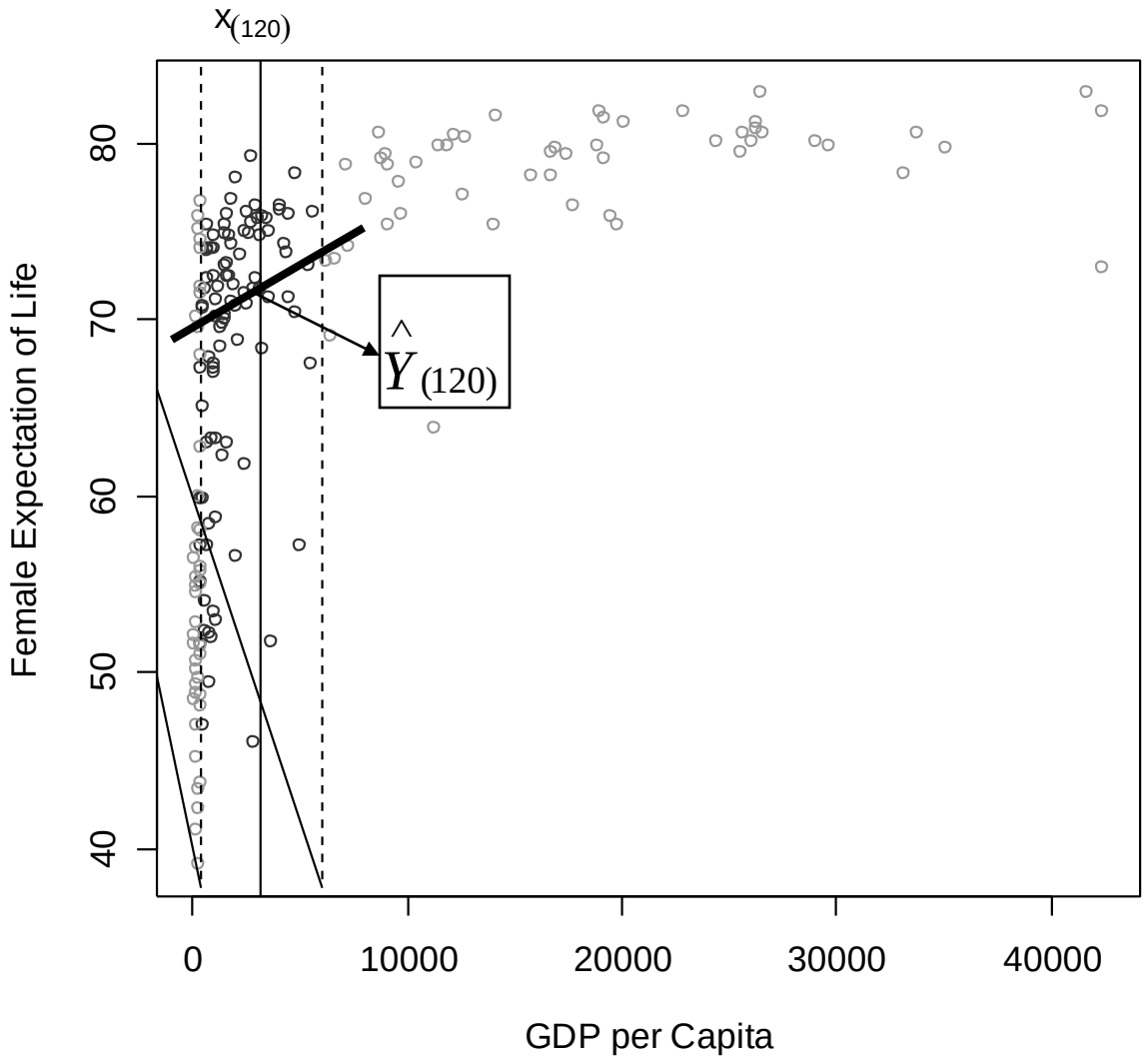
Local Linear Regression (Also known as lowess or loess)

- A more sophisticated way to calculate the Y values. Instead of constructing weighted average, it aims to construct a smooth local linear regression with estimated β_0 and β_1 that minimizes:

$$\sum_{1}^{n}[Y_i - \beta_0 - \beta_1(x_i - x_0)]^2 K(\frac{x_i - x_0}{h})$$

where K(.) is a kernel function, typically tricube.

The Local Average Now Is Predicted by a Regression Line, Instead of a Line Parallel to the X-axis.



Asymptotic Properties of lowess

- Fan (1992, 1993) demonstrated advantages of lowess over more standard kernel estimators. He proved that lowess has nice sampling properties and high minimax efficiency.
- In Heckman's works prior to 1997, he and his co-authors used the kernel weights. But since 1997 they have used lowess.
- In practice it's fairly complicated to program the asymptotic properties. No software packages provide estimation of the S.E. for lowess. In practice, one uses S.E. estimated by **bootstrapping**.

Bootstrap Statistics Inference (1)

- It allows the user to make inferences without making strong distributional assumptions and without the need for analytic formulas for the sampling distribution's parameters.
- **Basic idea:** treat the sample as if it is the population, and apply Monte Carlo sampling to generate an empirical estimate of the statistic's sampling distribution. This is done by drawing a large number of "resamples" of size n from this original sample randomly **with replacement**.
- A closely related idea is **the Jackknife**: "drop one out". That is, it systematically drops out subsets of the data one at a time and assesses the variation in the sampling distribution of the statistics of interest.

Bootstrap Statistics Inference (2)

- After obtaining estimated standard error (i.e., the standard deviation of the sampling distribution), one can calculate 95 % confidence interval using one of the following three methods:
  - Normal approximation method
  - Percentile method
  - Bias-corrected (BC) method
- The BC method is popular.

Finite-Sample Properties of lowess

The finite-sample properties of lowess have been examined just recently (Frolich, 2004). Two practical implications:

1. Choose optimal bandwidth value.
2. Trimming (i.e., discarding the nonparametric regression results in regions where the propensity scores for the nontreated cases are sparse) may not be the best response to the variance problems. Sensitivity analysis testing different trimming schemes.

Heckman's Contributions to PSM

- Unlike traditional matching, DID uses propensity scores differentially to calculate weighted mean of counterfactuals. A creative way to use information from multiple matches.
- DID uses longitudinal data (i.e., outcome before and after intervention).
- By doing this, the estimator is more robust: it eliminates temporarily-invariant sources of bias that may arise, when program participants and nonparticipants are geographically mismatched or from differences in survey questionnaire.

Demonstration of Running STATA/PSMATCH2:

Part 2. Heckman's Difference-in-differences Method

(Link to file "Day1c.doc")