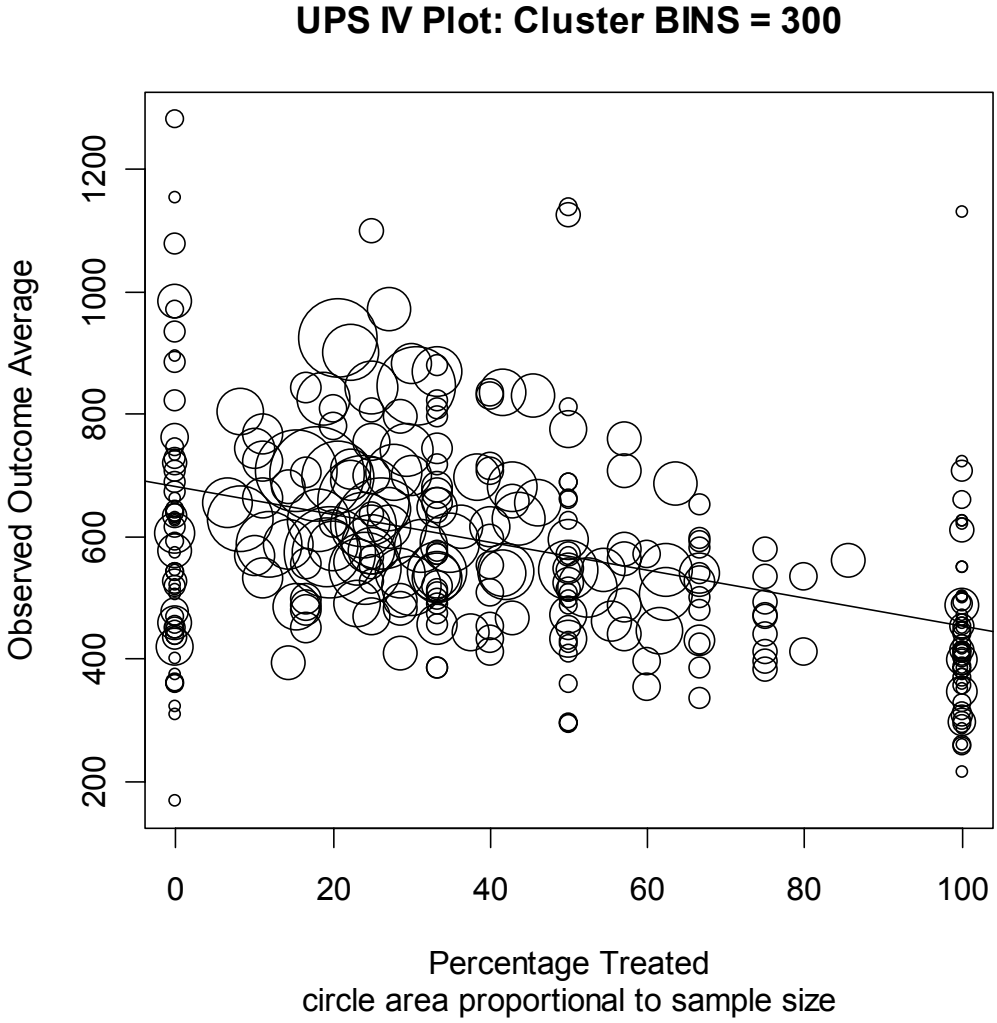
Figure 3.9: Third "Mean / Propensity" plot for "Instrumental Variable Adjustment" Analysis



Summary for Case Study Three

Table 3.3, below, summaries all of the alternative analyses for case study three that are displayed above.

Table 3.3: Sensitivity Analysis Summary

Method	Effect	Mean	Standard Deviation	Sample Size
Unadjusted Difference	Southerner minus Non-Southerner Hourly Wage (¢)	–136.7	9.42	3,010
Unadjusted Difference	Southerner minus Non-Southerner Hourly Wage (¢)	–85.60	12.00	2,040
Conventional PS Binning	Average (¢)	–41.49	29.75	2,040
PS Bin Bootstrapping	Average (¢)	–41.36	11.39	2,000
UPS with 30 Cluster-Bins (2 pure)	Unweighted (¢)	–39.36	69.39	?
	Weighted (¢)	–51.83	10.81	?
UPS with 90 Cluster-Bins (14 pure)	Unweighted (¢)	–39.99	98.18	?
	Weighted (¢)	–61.72	9.89	?
IV with 30 Cluster-Bins (2 pure)	IV Slope (¢)	–375.60	103.90	?
IV with 90 Cluster-Bins (14 pure)	IV Slope (¢)	–259.94	56.03	?
IV with 300 Cluster-Bins (94 pure)	IV Slope (¢)	–226.17	29.69	?

Upon comparing the numerical results displayed in Table 3.3 and reviewing the "visual" analyses that generated them, I conclude that the Card dataset suffers from considerable selection bias relative to determining the overall average difference in hourly wages

between Southerners and Non-Southerners. After matching wage-earners on their NEARC4, BLACK, SMSA, AGE, KWW and IQ characteristics, adjusted wage differences tend to be only half of what they would otherwise be estimated to be.

The econometric IV method using Cluster-Binning tends to give nonsensical answers for the Card data, even with as many as 300 clusters.

Conclusions

My review of recent literature on analysis of data from nonrandomized studies as well as my attempts to apply some of these new methods to example datasets have, at best, merely scratched the surface of a clearly controversial topic. Many "scientists" apparently still believe that it would be a waste of their time to analyze unplanned data because they could not conclude that they had "proven" anything. At the opposite extreme, proponents of a wide variety of new statistical/econometric approaches are apparently saying things like… "Try this! You will like it!" and "It will give answers at least as good as fill in the blank !"

While there is clearly no current consensus, I think that the "truth" (obviously) lies somewhere well within the two above sort of extreme positions. The following is a brief list of some of the main opinions I have developed:

- Scientists who find it unethical or impractical to apply many of the sound principles of "design of experiments" that would make their experimental units much more homogenous and interchangeable have come to rely much too heavily upon randomization. When the only "tool" one has is the hammer-of-randomization, every problem that one sees starts looking like some sort of "nail." In reality, the "tool" never worked nearly as well as researchers hoped it would, except in very large and truly random samples.

- The newer methods (like propensity scoring and unsupervised clustering) use subject "matching" concepts with great intuitive appeal. They obviously can be applied to data from randomized studies. Rather than simply compare the new approaches with traditional methods on actual datasets (where the "correct answer" is unknown), major research projects at academic institutions will ultimately use simulation to compare the "old" and "new" methods. Because the "old" methods make many more and much stronger assumptions that the "new" methods, the "new" methods will be declared to be much more robust than the "old" methods.

- I really enjoyed trying to understand and apply the "visualization" approaches. They strike me as being much more interesting than fitting "equations" to data embedded within some more-than-three-dimensional space, where it is impossible to "see" any lack-of-fit of the model to the data. The "R" functions I used allowed me to literally see some interesting or unusual subsets of subjects within my datasets. The (free!) software also allowed me to visually recognize homogeneity of effect signals, non-linearity in response and/or heteroskedasticity of subject-to-subject variation. I would have preferred to use even more powerful software that would have allowed me to more dynamically change the number of Cluster-Bins on my graphical displays and/or which could respond to a mouse click on a cluster by telling me all about the subjects within that cluster.

- I learned about the importance of "generalizability." Who cares about the results from a very carefully done study of some small sub-population of subjects who were offered strong incentives to even participate? How would the results from such a study help me predict what might really happen when subjects from the full target population for a treatment experience that treatment under real-world conditions? Perhaps data from nonrandomized studies are difficult to analyze, but at least the data they have clear potential to address important and practical questions.

- The Propensity Score Bin Bootstrapping approach consistently worked "best" in my three case studies. This approach not only gives essentially the "same" numerical answer (with near maximum bias removal) as the conventional, un-weighted average approach (1 subject = 1 vote), but it also produces estimates of high precision (low uncertainty), much like conventional "weighted" procedures.

- The Econometric IV methods using Cluster-Bins struck me as interesting because they not only discard information (within bin outcome differences) used by the Propensity Scoring (PS) approaches but also use information from "pure" bins that get discarded by the PS approaches! On the other hand, the Econometric IV methods using Cluster-Bins tended to give unreasonable answers in my three

Case Studies, except possibly when the number of clusters was relatively large. Unfortunately, in the limit where each subject ends up all alone in his/her own "pure" cluster, the Econometric IV approach degenerates to One-Way ANOVA by treatment group, with no adjustment whatsoever for any of the subjects' known X-characteristics.

- Researchers need to have software that helps them perform "sensitivity analyses" that reveal which alternative analysis **of their data** is most reasonable. Who cares which method is "generally" recommended and accepted by self-proclaimed experts if that favored approach gives a misleading answer on **your dataset**?

References

Angrist JD, Imbens GW, Dubin DB. "Identification of Causal Effects Using Instrumental Variables." ***J Amer Stat Assoc*** 1996; 91: 444-472.

Ash AS, Schwartz M, Lezzoni, LI. "Risk Adjustment for Measuring Patient Outcomes." **Cyber Seminar.** 2002. http://www.academyhealth.org/cyberseminars/archives/speakers.htm

Barlow HB. "Unsupervised learning." ***Neural Computation*** 1989; 1: 295-311.

Becker RA, Chambers JM, Wilks AR. ***The New S Language***. Chapman & Hall, New York. 1988.

Becker SO, Ichino A. "Estimation of Average Treatment Effects based on Propensity Scores." **University of Munich Website**. 2002. http://www.sobecker.de/pscore.html

Berndt ER. ***The Practice of Econometrics.*** NY: Addison-Wesley. 1991.

Braitman LE, Rosenbaum PR. "Rare Outcomes, Common Treatments: Analytic Strategies Using Propensity Scores." ***Ann Internal Med*** 2002; 137: 693-695.

Chambers JM, Hastie TJ, eds. ***Statistical Models in S***. Chapman & Hall, New York. 1992.

Cochran WG. "The effectiveness of adjustment by subclassification in removing bias in observational studies." ***Biometrics*** 1968; 24: 205-213.

Concato J, Shah N, Horowitz R. "Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs." ***New Eng J Med*** 2000; 342: 1887-1892.

Crown WE, Obenchain RL, Engelhart L, Lair TJ, Buesching DP, Croghan TW. "The application of sample selection models in evaluating treatment effects: the case for examining the effects of antidepressant medication." ***Statistics in Medicine*** 1998; 17: 1943-1958.

Czajka, Hirabayashi, Little, Rubin DB. "Projecting from Advance Data using Propensity Modeling: An Application to Income and Tax Statistics." ***J Bus & Econ Stat*** 1992; 10: 117-131.

D'Agostino RB Jr. "Tutorial in Biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group." ***Statistics in Medicine*** 1998; 17, 2265-2281.

Efron B, Gong G. "A leisurely look at the bootstrap, jackknife and cross-validation." ***The American Statistician*** 1983; 37: 36-48.

Efron B, Tibshirani RJ. **"**Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy." ***Statistical Science*** 1986; 1: 54-77.

Efron B, Tibshirani RJ. ***An Introduction to the Bootstrap***. New York: Chapman and Hall. 1993.

Everitt BS. The ***Cambridge Dictionary of Statistics***. NY: The Cambridge University Press. 1998: pgs. 268 and 266.

Glick H. "Identifying the Impact of a Treatment on Outcomes and Cost using Observational Data: Overcoming Selection Bias." Tutorial for the **36th Annual Drug Information Association Meeting**. 2000. http://www.uphs.upenn.edu/dgimhsr/IVLEC11A.PDF

Gu XS, Rosenbaum PR. "Comparison of multivariate matching methods: Structures, distances and algorithms." ***J Comp & Graph Stat*** 1993; 2: 405-420.

Heckman JJ. "Sample Selection Bias as a Specification Error." ***Econometrica*** 1979; 47:153-161.

"How to Pull 99.5% of your Hat out of a Rabbit" ***The Economist (US)*** 1988; 309: 97.

Ihaka R, Gentleman R. "R: A language for data analysis and graphics." ***J Comp & Graph Stat*** 1996; 5(3): 299-314. http://www.r-project.org

Kaufman L, Rousseeuw PJ. ***Finding Groups in Data. An Introduction to Cluster Analysis.*** New York: John Wiley and Sons. 1990.

Kereiakes DJ, Obenchain RL, Barber BL, Smith A, McDonald M, Broderick TM, Runyon JP, Shimshak TM, Schneider JF, Hattemer CH, Roth EM, Whang DD, Cocks DL, Abbottsmith CW. "Abciximab provides cost effective survival advantage in high volume interventional practice." ***Am Heart J*** 2000; 140: 603-610.

Kotz S and Johnson NL. Encyclopedia of Statistical Sciences. NY: John Wiley & Sons Publishing. 1982: 280.

McClellan M, McNeil BJ, Newhouse JP. "Does More Intensive Treatment of Myocardial Infarction in the Elderly Reduce Mortality?: Analysis Using Instrumental Variables." ***JAMA*** 1994; 272: 859-866.

Ming K, Rosenbaum PR. "A note on optimal matching with variable controls using the assignment algorithm." ***J Comp & Graph Stat*** 2001; 10: 455-463.

Murnane RJ, Newstead S, Olsen RJ. "Comparing Public and Private Schools: the Puzzling Role of Selectivity Bias." ***J Bus & Econ Stat*** 1985; 3: 23-35.

Obenchain RL. "Nearest Neighbors Analysis for PRRAP, the Probable Report Rate Analysis Plan." **Bell Laboratories** TM-79-1711-4. 1979. Holmdel, NJ.

Obenchain RL, Melfi CA. "Propensity Score and Heckman Adjustments for Treatment Selection Bias in Database Studies." **Proceedings of the Biopharmaceutical Section,** 1997; 297-306. Washington, DC: American Statistical Association.

Obenchain RL. **S-plus functions for propensity score adjustment using binning and smoothing.** 1999. http://www.math.iupui.edu/~indyasa/download.htm.

Obenchain RL. "Trees, Clustering and Smoothing in Propensity Scoring." **Sixth Great Lakes Symposium on Applied Statistics**. 2000. Kalamazoo, MI. http://www.stat.wmich.edu/GLS_conf/grt_lake.PDF

Obenchain RL. **PSbinbst.EXE: A Console Application for Propensity Score Bin BootStrapping.** 2002. http://www.math.iupui.edu/~indyasa/download.htm.

Obenchain RL. **R functions for supervised and unsupervised propensity scoring, smoothing and graphing.** 2003. http://www.math.iupui.edu/~indyasa/download.htm.

Vogt PW. ***Dictionary of Statistics and Methodology***. USA: Sage Publications. 1999: pgs. 24, 34 and 45.

Robins JM, Hernan MA, Brumback B. "Marginal Structural Models and Causal Inference in Epidemiology." ***Epidemiology*** 2000; 11: 550-560.

Rosenbaum PR, Rubin RB. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." ***Biometrika*** 1983; 70: 41-55.

Rosenbaum PR, Rubin DB. "Reducing Bias in Observational Studies Using Subclassification on a Propensity Score." ***J Amer Stat Assoc*** 1984; 79: 516-524.

Rosenbaum PR, Rubin DB. "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score." ***American Statistician*** 1985; 39: 33-38.

Rosenbaum PR. "Multivariate matching methods." In: Kotz S, Read CR, Banks D, eds. ***Encyclopedia of Statistical Sciences***, Update Volume 2. New York: J Wiley 1998: 435-438.

Rosenbaum PR. ***Observational Studies***. New York: Springer-Verlag 2002.

Rosenbaum PR. "Optimal matching in observational studies." ***J Amer Stat Assoc*** 1989; 84: 1024-1032.

Royall RM. "Ethics and Statistics in Randomized Clinical Trials." ***Statistical Science*** 1991; 6; 52-62.

Rubin DB. "Bias reduction using Mahalanobis metric matching." ***Biometrics*** 1980; 36: 293-298.

Rubin DB. "Estimating Causal Effects from Large Data Sets Using Propensity Scores." ***Ann Internal Med*** 1997; 127: 757-763.

Savitz DA. "Human Studies of Human Health Hazards: Comparisons of Epidemiology and Toxicology." ***Statistical Science*** 1988; 3: 306-313

Wald A. "The Fitting of Straight Lines if Both Variables are Subject to Error." ***Ann Math Stat*** 1940; 11: 284-300.

Wooldridge JM. ***Introductory Econometrics***. South-Western College Publishing: USA. 2000.

Zanutto BL, Hornik R, Rosenbaum PR. "Matching with Doses in an Observational Study of a Media Campaign Against Drug Abuse." ***J Amer Stat Assoc*** 2000; 96: 1245-1254.