Abstract Title Page
Not included in page count.

Title: Omitted variable sensitivity analysis with the Annotated Love Plot

Authors and Affiliations:
1. Hansen, Ben B.
Associate Professor, Statistics Department, University of Michigan, Ann Arbor.
ben.b.hansen@umich.edu
2. Fredrickson, Mark M.,
PhD candidate, Political Science Department, University of Illinois, Champaign-Urbana.
markmfredrickson@gmail.com

Abstract Body

Background / Context:

Description of prior research and its intellectual context.

Beginning with Cornfield *et al* (1959), statisticians analyzing quasi-experiments have in various ways quantified how much a hypothetical omitted variable (OV) could perturb impact estimates, and corresponding tests and confidence intervals, as a function of the OV's relationships with measured variables (Rosenbaum and Rubin, 1983a, Rosenbaum, 1987, Marcus, 1997, Copas and Eguchi, 2001). In contrast with related econometric approaches aiming to place absolute bounds on the perturbation caused by the OV (e.g., Manski 1995), these methods do not aim to "correct" for the omission of the variable so much as to inform deliberations about the trustworthiness of study findings. The related development of Frank *et al* (2013) makes OVs less central conceptually, but it also seeks to inform speculation about bias rather than to bound it. The best known of these sensitivity analysis methods are Rosenbaum's (2002b, 2010), which work by bounding differences between propensity scores that do and do not incorporate the hypothetical OV.

The main sensitivity parameter of the Rosenbaum method is "Rosenbaum's gamma," which quantifies the confounding potential of an OV in terms of that variable's contribution to a subject's conditional log odds of falling in the treatment group. While natural in the context of propensity score analysis, the parameter is potentially confusing to audiences unaccustomed to logarithms, odds ratios or formal conditional probability. Frank (2000) and Imbens (2003) somewhat broadened the scope of the conversation by phrasing their sensitivity analyses in terms of correlations and partial R^2 coefficients, respectively, measures that are more broadly intelligible. Hosman et al. (2010) also used partial R^2, but only for the relationship of the hypothetical OV to the outcome, terming this characteristic of the variable its "proportionate reduction in unexplained variation." These re-parametrizations do advance the cause of interpretability, but surely they remain opaque to lay audiences.

Sensitivity analyses are more vivid when accompanied with meaningful anchors for parameters determining the impact of the OV's omission. Altonji et al (2003) hypothesize an OV with associations to the treatment and outcome variables comparable to the treatment and outcome associations of all included variables taken together, and Oster (2013) finds this approach to have worked well in a comparison of certain observational studies from public health with related experiments. More closely related to current research are the proposals of Angrist and Imbens (2002), Imbens (2003), Hosman *et al.* (2010) and Hsu and Small (2013), who calibrate the treatment and outcome associations of an OV with reference to treatment and outcome associations of the included variables taken one at a time, rather than jointly. Each of these developments extracts reference values for the OV's relevant relationships with included variables from the same sample and variables as are used to estimate program effects, an approach we might call "intrinsic" calibration. If similar associations among comparable variables within well-studied samples, even unrelated samples, were available, they would carry the advantage of being less sensitive to idiosyncracies of the quasiexperiment under study; but we are not aware of methodological literature exploring such "extrinsic" calibrations.

Covariate balance plots - "Love plots," after the biostatistician (Thomas Love) who invented them - are used after propensity score matching (or weighting) to convey visually the differences

between intervention groups and their matched (weighted) controls in terms of *measured* baseline variables. . One at a time, the Love plot takes these baseline measures and treats them as if they were outcomes, calculating means of participant/non-participant differences and expressing them as effect sizes. Typically this generates large "participation effects" when differences are considered without reference to matches, but when only matched individuals are compared the resulting "effects" are uniformly small. Since the differences are in terms of pre-intervention variables, they of course estimate not effects of the program but rather biases, or components of bias attributable to measured variables, that would be incurred by in the same way averaging and differencing outcomes to estimate program effects. We are not aware of work connecting covariate balance plots to issues of unmeasured confounding.

Method/ Objective / Research Question / Focus of Study:

Description of the focus of the research.

The goal of this research is to make sensitivity analysis accessible not only to empirical researchers but also to the various stakeholders for whom educational evaluations are conducted. To do this it derives anchors for the OV-program participation association intrinsically, using the Love plot to present a wide range of possible values in a compact yet intuitive manner. At the same time it calibrates anchor points for the OV-outcome association extrinsically, drawing on published descriptions of nationally representative samples.

Setting:

Description of the research location.

Our overall approach requires a quasi-experimental comparison using matching or weighting for covariate adjustment. The implementation discussed here is adapted to education evaluations based on standard K-12 administrative data, using propensity-score matching for covariate adjustment and matched permutation tests for the presence of intervention effects. It is one of a number of adaptations and refinements of such methods being developed for the Evaluation Engine, a Gates-funded project to make timely quasi-experimental impact analysis available to state- and district-level decision makers.

Significance / Novelty of study:

Description of what is missing in previous work and the contribution the study makes.

By a shift of the sensitivity analysis model, and by combining generic reference values for the confounder-outcome relationship with study-specific anchors for speculation about the treatment-confounder relationship, it becomes possible to present specific, quantitative OV sensitivity analyses to lay audiences as well as to empirical researchers.

Statistical, Measurement, or Econometric Model:

Description of the proposed new methods or novel applications of existing methods.

Given an outcome Y, let the random variable Y_c represent potential response to the control condition. For non-participants ($Z=0$), Y and Y_c take the same value, whereas for the participant group ($Z=1$) Y_c is an unobserved, "counterfactual" random variable, and $Y - Y_c$ is the program effect (on a given participant). Our method applies to situations in which participants' outcomes have been found to differ significantly from those of their matched controls, and it asks whether an unobserved covariate, as opposed to a treatment effect, might explain this difference. For this purpose is assumed that there is no effect, i.e. $Y \equiv Y_c$.

Quasiexperimental analyses comparing a participant and a control group require *conditional independence* and *overlap* conditions, respectively $Y_c \perp Z|X$ and P (Z =1| X) < 1, where X denotes measured covariates (Rosenbaum and Rubin 1983b; Heckman et al 1998). In permutation-based analysis, regression techniques can be used to weaken this condition. Let $f(\cdot)$ be a guess at, estimate of or approximation to the regression of Y_c on X, or on a subset $\tilde{X}$ of X variables, and write $\epsilon_c \equiv Y_c - f(X)$. (The procedure makes no requirement that $f(\cdot)$ consistently estimate $E(Y_c|X)$, or $E(Y_c|\tilde{X})$.) Then the weakened condition is that

$$\epsilon_c \perp Z|X \text{ and } \text{P (Z =1| X)} < 1, \quad (1)$$

(Hajek et al 1999; Rosenbaum 2002a). (Our application uses as $f(\cdot)$ an empirical regression on pretest scores, this enhancing power and making available standard accommodations for measurement error in the pretest [Carroll *et al*, 2006].)

Propensity matched analysis takes place after estimating and matching on the propensity score, the conditional probability of membership in the treatment group, *Z=1*, given measured covariates, X. It adds the assumption that paired subjects have precisely the same propensity score (Rosenbaum, 2010). Like conditional independence, this model is frequently employed without being thought to be precisely true; in contrast with conditional independence, it is in various ways testable from the data (Hansen and Bowers, 2008; Heller et al., 2010).

For each subject *i* write $[i]$ for the collection of subjects matched to i, or matched to subjects matched to i. The present method entertains the possibility of (1)'s failure by relaxing it to

$$\epsilon_{cu} \perp Z|X \text{ and } \text{P (Z =1| X)} < 1, \quad (2)$$

where $\epsilon_{cui} = Y_{ci} - f(X_i) - \alpha_{[i]} - \alpha_u U = \epsilon_{ci} - g(U_i)$, with $g(\cdot)$ defined as the least squares regression of ϵ_c on matched set fixed effects and on the unmeasured variable U.

Under (2) the (weighted) mean of paired differences in Y, $\Delta[Y]$, estimates the effect of treatment on the treated, $E(Y - Y_c|Z = 1)$, with a bias that is in turn estimated by

$$\widehat{BIAS} = s_\epsilon \rho_{\epsilon u} \Delta[U/s_U], \quad (3)$$

where s_ϵ and s_U are residual standard deviations of ϵ_c and U, respectively, after centering around matched set means; $\rho_{\epsilon u}$is the partial correlation of ϵ_c and U, net of matched set fixed effects; and $\Delta[U/s_U]$ is the mean of paired differences in U/s_U.[2] If U can be assumed unrelated, within matched pairs, either to ϵ_c or to Z, then $\widehat{BIAS} = 0$ and relaxing (1) to (2) does not change p-values calculated under (1). Otherwise p-values increase, to an extent governed by $\rho_{\epsilon u}$ and $\Delta[U/s_U]$ jointly.

We abstract generic reference values for $\rho_{\epsilon u}$ from the rich body of empirical research developed for the planning of randomized field trials. For instance, let *Y* represent students' scores on a statewide, grade-level examination, and suppose a *U* with a relationship to ϵ_c comparable in strength to pretest-posttest associations. Working with nationally representative samples, Hedges and Hedberg (2007) calculate values of a quantity, "η_W," that can be transformed onto the scale of $\rho_{\epsilon u}$. The results vary somewhat by grade and subject, but most translate to partial correlations in the neighborhood of .75. Hedges and Hedberg also study demographic variables as predictors of test outcomes, finding them to have substantially less predictive power: partial

[2] If some participants are matched to multiple controls then $\Delta[\cdot]$ downweights the contributions of their paired differences to the mean of paired differences so that the total weight of each participant's paired differences is 1.

correlations of 0.1 are typical. The study of Zhu et al (2012) assisted the planning of school-level trials by quantifying the additional contribution of classroom-level information, over and above school and student data, to the prediction of student test scores; partial correlations of roughly 0.1 and 0.75 are again entailed, the former for lower grades and the latter for high school. In this way we externally calibrate anchor points for the omitted variable-outcome relationship: Strong and weak prediction are modeled as having partial correlations with the outcome of 0.75 and 0.1, respectively.

Usefulness / Applicability of Method:

Demonstration of the usefulness of the proposed methods using hypothetical or real data.

An illustration of the proposed method is presented at the bottom of the Figure. The main, upper part of the figure compares baseline measures of participants in a hypothetical program for 11th graders to those of all 11th graders in the state (red circles) and to those of their matched counterparts (turquoise squares). This portion of the plot should be recognizable as a standard Love plot. It presents quantities $\Delta[X/s_X]$, for X equal to various baseline variables – precisely the needed material for an internal calibration of the remaining sensitivity parameter, $\Delta[U/s_U]$.

The Love plot presents a large number of possible anchors for $\Delta[U/s_U]$, the imbalance in U remaining after matching. The relevance of these anchors to each of the program effects that was found to be statistically significant is presented below the Love plot, with an indication of how large a post-matching imbalance in U would be needed to upset that finding. In the figure, for example, an apparent math effect is made insignificant either by allowances for a weak outcome predictor U with a standard bias $|\Delta[U/s_U]|$ of .17 or more, or by a strong outcome predictor U with a standard bias as small as .03, these being the positions marked on the x-axis with "m" and "M", respectively.

In this example the most sensitive of the three findings is that of a Math effect, m's appearing to the left of r's and w's in both upper and lower case, while the least sensitive is the finding for writing, the W and w appearing to the right of all others. A strong predictor of the outcome that also is as confounded with participation, in the state population, as is homelessness within the past two years, could upset the finding for math (M) but not for reading (R) or writing (W); we see this by comparing the horizontal position of the Homeless variable's red circle to those of the M, R and W.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

Like other modes of sensitivity analyses, ours requires a model, which itself may fail. Its use of generic reference values to calibrate OVs' proportionate reductions in unexplained variation is in effect an additional assumption. What is gained in return is a vast widening of the audience for sensitivity analysis. Broadening participation in this game increases the pool of information brought to bear on the inherently subjective assessments required to assess omitted variable sensitivity. This in turn may lead to better decisions regarding whether a quasi-experimental finding needs to be replicated, perhaps with a randomized controlled trial, before using it as a basis for decision making.

Appendix A

References

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, *113*(1), 151–184.

Angrist, J., & Imbens, G. (2002). Comment. *Statistical Science*, *17*(3), 304–07.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective* (Second ed., Vol. 105). Chapman and Hall/CRC.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā, Series A, Indian Journal of Statistics*, *35*, 417–446.

Copas, J. B., & Eguchi, S. (2001). Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistics Society*, *63*(4), 871–895.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., & Wynder, E. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, *22*, 173–203.

Frank, K. (2000, November). Impact of a confounding variable on a regression coefficient. *Sociological methods and research*, *29*(2), 147–194.

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? using rubinâĂŹs causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, *35*(4), 437-460.

Hájek, J., Šidák, Z., & Sen, P. K. (1999). *Theory of rank tests* (Second ed.). New York: Academic Press.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, *95*(2), 481–488.

Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, *23*(2), 219–236.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1998). Matching as an Econometric Evaluation Estimator. *Review of Economic Studies*, *65*(2), 261–294.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87.

Heller, R., Rosenbaum, P. R., & Small, D. S. (2010). Using the cross-match test to appraise covariate balance in matched pairs. *The American Statistician*, *64*(4), 299–309.

Hosman, C. A., Hansen, B. B., & Holland, P. W. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Annals of Applied Statistics*, *4*(2), 849-870. doi: 10.1214/09-AOAS315

Hsu, J. Y., & Small, D. S. (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics*. (in press)

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 126–132.

Manski, C. F. (1995). *Identification problems in the social sciences*. Harvard University Press.

Marcus, S. M. (1997). Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *Journal of Educational and Behavioral Statistics*, *22*, 193–201.

Oster, E. (2013, August). *Unobservable selection and coefficient stability: Theory and validation* (Tech. Rep. No. 19054). National Bureau of Economic Research.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, *74*, 13– 26. ((Correction: V75 p396))

Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, *17*(3), 286–327.

Rosenbaum, P. R. (2002b). *Observational studies* (Second ed.). Springer-Verlag.

Rosenbaum, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *American Statistician*, *59*(2), 147–152.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B, Methodological*, *45*, 212–218.

Rosenbaum, P. R., & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rosenbaum, P. R., & Silber, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association*, *104*(488), 1398-1405.

Zhu, P., Jacob, R., Bloom, H., & Xu, Z. (2012). Designing and analyzing studies that randomize schools to estimate intervention effects on student academic outcomes without classroom-level information. *Educational Evaluation and Policy Analysis*, *34*(1), 45–68.

Appendix B

Tables and Figures

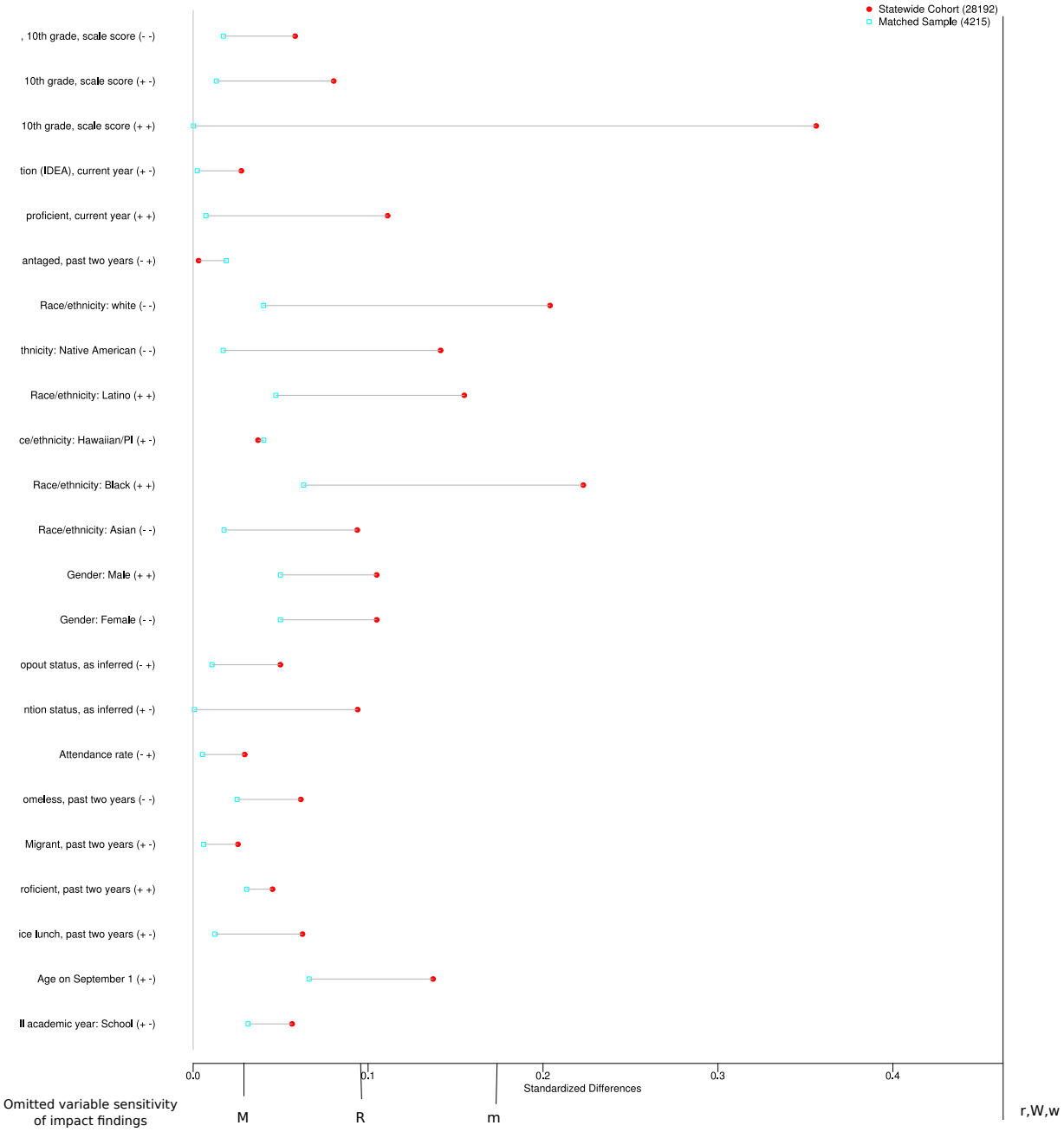

Figure 1: Annotated Love Plot