A Measure of Robustness to Misspecification

Susan Athey* Guido W. Imbens†

December 2014

*Graduate School of Business, Stanford University, and NBER. Electronic correspondence: athey@stanford.edu.

†Graduate School of Business, Stanford University, and NBER. Electronic correspondence: imbens@stanford.edu.

1 Introduction

Suppose a researcher wants to predict the causal effect of a new policy or treatment, based on a data set containing information on some characteristics and outcomes for a large set of units. The researcher might specify a statistical model relating the outcomes, treatments, and characteristics, estimate that model, and use the results to predict the effect of the policy or treatment. The researcher would typically then report the point estimate of the policy effect as well as a measure of the uncertainty surrounding that point estimate. For example, if the model was estimated by maximum likelihood methods, the uncertainty might be based on the information matrix, measuring the curvature of the likelihood function at its maximum. Such measures of uncertainty are conditional on the model specification. Often, however, there is also substantial uncertainty regarding the model specification. There are typically multiple decisions going into the specification of the model that are relatively arbitrary. Different choices for those specification decisions would have left to different point estimates of the effect of the policy. Previously Leamer (1982, 1983) and White (2000) have raised concerns about the fact that the uncertainty in the estimates due to such decisions is not reflected in the standard errors. In practice researchers might attempt to address this issue by reporting estimates for a handful of different specifications to assess the sensitivity of the estimates to the specification of the model. In this paper we propose a more systematic approach to assessing the sensitivity of point estimates to model specification. We propose a way of defining a finite set of models that includes the original model as a special case. Each member of the set allows for estimation of the effect of the policy. We then calculate the variance of the point estimate of the effect of the policy over this class of models. The hope is to provide the reader of an empirical study with a sense of the robustness of the policy predictions, beyond the information contained in the point estimate and the asymptotic standard error.

2 A Simple Example

We use two data sets originally constructed by Lalonde (1986), and subsequently widely used in the evaluation literature (e.g., Dehejia and Wahba, 1999), to illustrate the basic ideas. The versions of these data sets we use are posted on the the website of Rajeev Dehejia. The first data set, which we call the "experimental Lalonde data," contains information on men participating in an experimental job training program. The focus is on estimating the average effect of the program on subsequent earnings. In this data set there are 185 individuals in the treatment group and 260 in the control group. The second data set, which we call the "non-experimental Lalonde data", replaces the individuals in the control group of the experimental Lalonde data set with observations on men from a non-experimental comparison group drawn from the Current Population Survey (CPS). The individuals in the CPS are substantially different from those in the experiment, which creates severe challenges for comparisons between the two groups. The interest is

in the causal effect of the training program on earnings in 1978, measured in thousands of earnings. In this data set there are 15,992 individuals in the control group, and again 185 individuals in the treatment group.

As the base model consider a linear regression with outcome Y_i (earnings in 1978), an indicator for participation in the training program, X_i, an intercept, and ten characteristics of the individual, denoted by Z_i, including age, indicators for being African-American, or Hispanic, an indicator for being married, years of education, having a degree, earnings in 1974 and 1975, and indicators for earnings in 1974 and 1975 being zero:

$$\mathbb{E}[Y_i|X_i, Z_i] = \alpha + \theta_B \cdot X_i + \gamma' Z_i. \tag{2.1}$$

the coefficient on the program, θ_B (where the subscript "B" stands for base model), is the object of interest. Let $\hat{\theta}_B$ be the least squares estimator for θ_B, and similarly for $\hat{\alpha}$ and $\hat{\gamma}$. Table 1 reports the results for the two data sets.

Table 1: Experimental and Non-experimental Lalonde Data: The Base Model

	Exper		Non-exper	
Variable	est.	(s.e.)	est.	(s.e.)
`treatment`	1.67	(0.67)	1.07	(0.63)
`intercept`	0.26	(3.65)	5.78	(0.44)
`age`	0.05	(0.04)	-0.09	(0.01)
`black`	-2.04	(1.02)	-0.81	(0.20)
`hispanic`	0.40	(0.20)	0.17	(0.03)
`married`	0.43	(1.40)	-0.23	(0.22)
`education`	-0.15	(0.85)	0.15	(0.14)
`nodegree`	-0.02	(1.04)	0.34	(0.18)
`re74`/1000	0.12	(0.13)	0.29	(0.02)
`re75`/1000	0.02	(0.14)	0.44	(0.02)
`u74`	1.38	(1.55)	0.35	(0.23)
`u75`	-1.07	(1.41)	-1.62	(0.25)

For the experimental data set the estimate of the effect of the program is 1.67 (s.e. 0.67) and for the non-experimental data set the estimate is 1.07 (s.e. 0.63). These estimates are fairly similar, both in terms of the point estimate and in terms of the precision. Clearly, however, the experimental estimate is more credible. Part of this is because we know a priori that, because of the actual randomization, there are no unobserved confounders, whereas with the non-experimental data we cannot be sure

of that. This knowledge about the superior credibility of the experimental estimates cannot be inferred from the data, it is a reflection on substantive information. There is information in the data, however, that is informative about the relative credibility of the two estimates, and that is not summarized in the combination of the point estimate and the standard error. What we are proposing in this paper is that researchers report an additional statistic, beyond the point estimate and the standard error, that conveys at least part of the evidence from the data that the point estimate from the non-experimental data concerning the efficacy of the training program is not as compelling as that from the experimental data.

The basic idea is that in addition to the base model we systematically explore a range of alternative specifications. The question is how to choose the set of alternative specifications to consider. Let us start by considering a single additional specification. For example, we can split the sample into two subsamples by the value of `u75`, the indicator for being unemployed in 1975. In the experimental sample there are 156 individuals employed in 1975 (82 in the control group and 74 in the treatment group), and 289 individuals unemployed in 1975 (178 in the control group and 156 in the treatment group). We estimate separately the same linear regression as in (2.1) for the two subsamples, with the only difference that we leave out the regressor corresponding to the the indicator for the individual being unemployed in 1975 because it is constant within the two subsamples. This leaves us with two estimates for the effect of the training program. We then combine them using the relative size of the two subpopulations (0.35 for the `u75=0` subsample and 0.65 for the `u75=1`), leading to a point estimate of 1.71 for the more general model, very close to the estimate from the base model, which is 1.67. We can do the same for the non-experimental sample. In that case we obtain an estimate of -0.83 for the general model that allows for a full set of interactions with `u75`, substantially different from the original estimate of 1.07. It is the fact that the estimate of the same substantive parameter, the average causal effect of the program on the outcome, is more sensitive to this change in the specification for the non-experimental data than for the experimental data that we focus on. Note that this sensitivity is distinct from, although related to, the question whether the based model is correctly specified. It may well be that the additional parameters in the more general model are different from zero, and found to be so in terms of statistical significance, without the estimate of the average causal effect being different according to the two specifications. The questions are related though in the sense that if the model is correctly specified, the parameter estimates should be similar. In this case if we perform an F-test to test the more general model against the base model, we find an chi-squared statistic equal to 7.0 for the experimental data, and 81.9 for the non-experimental data, which under the null hypothesis of zero parameters for the general model should have a chi-squared distribution with degrees of freedom equal to ten.

Of course exploring alternative specifications is common in empirical work. Typically researchers report estimates based on a preferred specification and in addition results for some alternative specifications that they consider particularly salient. Here we fo-

cus on a more systematic approach to reporting robustness to alternative specifications. Rather than focusing on the actual point estimates for a small number of alternative specifications, we focus on the variation over a larger number of alternative specifications. Although there will remain arbitrary decisions in this process, the aim is to reduce the arbitrariness in exploring a very small number of alternative specifications.

First let us illustrate this in the context of this linear model for the Lalonde data. To do so we re-interpret the model and the estimand. We can think of the statistical model as providing a parametric approximation to the conditional mean

$$\mu(x, w) = \mathbb{E}[Y_i | X_i = x, Z_i = z].$$

The estimand θ, can be thought of as the average effect of the treatment. In terms of the conditional expectation function, assuming selection on observables or unconfoundedness, we can express the estimand as

$$\theta = \mathbb{E}\left[\mu(1, Z_i) - \mu(0, Z_i)\right].$$

Given an estimate of the conditional expectation function, $\hat{\mu}(x, z)$, we can estimate the target θ as

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} \Big(\hat{\mu}(1, Z_i) - \hat{\mu}(0, Z_i) \Big).$$

The base model assumes the parametric structure

$$\mu(x, z) = \alpha + \theta \cdot x + \gamma' z,$$

and estimates the parameters of this specification by least squares, leading to a base-model estimate $\hat{\theta}_B$. Now consider splitting the sample into two subsamples, based on the values of X_i and Z_i. Let W be the support of (X_i, Z_i), and let W_L and $\mathbb{W}_H$ denote the partition of W that generates the two subsamples, so that $W = W_L \cup W_H$ and let $I_i \in \{L, H\}$ be an indicator for the subsample,

$$I_i = \begin{cases} L & \text{if } (X_i, Z_i) \in W_L, \\ H & \text{if } (X_i, Z_i) \in W_H. \end{cases}$$

Now we estimate the two regression models, one for each subsample,

$$Y_i = \alpha_L + \theta_L \cdot X_i + \gamma_L' Z_i,$$

for the $I_i = L$ subsample, and

$$Y_i = \alpha_H + \theta_H \cdot X_i + \gamma_H' Z_i,$$

for the $I_i = H$ subsample, using least squares. Note that not all the parameters in these regressions are necessarily identified, but that does not generate any problems because we

are primarily interested in the predictions which are identified. The estimated parameters define a new estimate for the conditional expectation,

$$\hat{\mu}(x,z) = \mathbf{1}_{(x,z)\in W_L} \cdot \Big(\hat{\alpha}_L + \hat{\theta}_L \cdot x + \hat{\gamma}_L' z\Big) + \mathbf{1}_{(x,z)\in W_H} \cdot \Big(\hat{\alpha}_H + \hat{\theta}_H \cdot x + \hat{\gamma}_H' z\Big),$$

and thus an estimate for θ:

$$\hat{\theta}_s = \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{\mu}(1,Z_i) - \hat{\mu}(0,Z_i)\Big) = \overline{I}\cdot\hat{\theta}_H + (1-\overline{I})\cdot\hat{\theta}_L,$$

where s indexes the split of W into (W_L, W_H). It is the magnitude of the difference between the original base-model estimate $\hat{\theta}_B$ and the estimate based on the split sample, $\hat{\theta}_s$, that we focus on. If for a typical split this difference $\hat{\theta}_s - \hat{\theta}_B$ is small in magnitude, we consider the base-model estimate $\hat{\theta}_B$ robust to alternative specifications, and if not, we consider it sensitive to the specification. The final question is how to split the sample. We propose to split the sample by every covariate, that is, every element of X_i (one in our application) and every element of Z_i (ten in our application). In each case if the covariate takes on more than two values we split at the value that generates the biggest increase in the model fit (the biggest decrease in the sum of squared residuals), following the literature on regression trees (Breiman, Friedman, and Stone, 1983).

The case where we split on the value of X_i is somewhat special. In that case there is no estimate of θ for each of the two subsamples, because the treatment does not vary within the subsamples. Rather, the more general model implies a model for the conditional expectation, which then lets us infer the value for the average treatment effect.

Given the estimates $\hat{\theta}_s$ for the eleven choices of the sample splits, we calculate the standard deviation as

$$\hat{\sigma}_\theta = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\Big(\hat{\theta}_s - \hat{\theta}_B\Big)^2}, \tag{2.2}$$

and it is this measure of robustness we focus on. In Table 2 we report the estimates $\hat{\theta}_s$ for each of the sample splits, and the resulting standard deviation, for both the experimental and non-experimental Lalonde data. We find that for the experimental data $\hat{\sigma}_\theta = 0.13$, about 20% of the standard error of $\hat{\theta}_B$, whereas for the non-experimental data it is 2.13, 338% of the standard error of $\hat{\theta}_B$ for the non-experimental data. It is clear that the results for the experimental data are far more robust than those for the non-experimental data.

3 The General Case

Suppose we have a random sample from a large population of a pair of variables (Y_i, X_i), where both Y_i and X_i may be vectors. The object of interest is a functional of the

Table 2: Variation of $\hat{\theta}$ over Model Specifications for the Lalonde Data

	Exper		Non-exper	
Variable	$\hat{\theta}_B$	s.e.	$\hat{\theta}_B$	s.e.
`Base Model`	1.67	(0.67)	1.07	(0.63)
$\hat{\sigma}_\theta$		[0.13]		[2.13]
Split on	est.	$\mathcal{X}^2(10)$	est.	$\mathcal{X}^2(10)$
`treatment`	1.58	22.8	-4.26	45.0
`age`	1.55	10.1	1.97	144.5
`black`	1.71	11.4	1.38	26.2
`hispanic`	1.61	7.2	1.54	59.7
`married`	1.87	11.0	1.06	10.4
`education`	1.77	14.7	1.25	74.5
`nodegree`	1.33	18.6	1.77	46.1
`re74`	1.64	11.2	0.58	71.0
`re75`	1.63	8.9	-0.88	94.1
`u74`	1.64	11.2	-2.44	88.6
`u75`	1.71	7.0	-0.83	81.9

conditional distribution of Y_i given X_i and the sample values of X_i:

$$\theta = g(f_{Y|X}(\cdot|\cdot), X_1, \ldots, X_N). \tag{3.1}$$

Consider a parametric model for the conditional distribution of Y_i given X_i:

$$f_{Y|X}(y, x) = f_{Y,X}(y|x; \beta),$$

with β an unknown parameter. Given the maximum likelihood estimate $\hat{\beta}$ for β,

$$\hat{\beta} = \arg\max_{\beta} L(\beta|\mathbf{Y}, \mathbf{X}),$$

we can estimate we can estimate θ for this base model as

$$\hat{\theta}_B = g(f_{Y|X}(\cdot|\cdot; \hat{\beta}), X_1, \ldots, X_N).$$

Now we propose evaluating $\hat{\theta}$ for a discrete set of models that generalizes the base model.

A key feature is that the estimand is well defined for each member of the larger set of models. The larger class of models is based on splits of the basic sample in terms of the covariates. Let the number of covariates be K_X. Given a threshold c, and given the kth covariate, define the subsample indicators,

$$I_{ikc} = \begin{cases} L & \text{if} X_{ik} \leq c \\ H & \text{if} X_{ik} > c. \end{cases}$$

Denote the values of the outcomes in the subsample with $I_{ikc} = L$ by $\mathbf{Y}_{kc,L}$, and the covariates in this subsample by $\mathbf{X}_{kc,L}$. Denote the values of the outcomes and covariates in the subsample with $I_{ikc} = H$ by $\mathbf{Y}_{kc,H}$ and $\mathbf{X}_{kc,H}$. We estimate the values of the parameters in the two subsamples

$$\hat{\beta}_{k,L}(c) = \arg\max_{\beta} L(\beta|\mathbf{Y}_{kc,L}, \mathbf{X}_{kc,L}) \qquad \text{and} \quad \hat{\beta}_{k,H}(c) = \arg\max_{\beta} L(\beta|\mathbf{Y}_{kc,H}, \mathbf{X}_{kc,H})$$

We choose the threshold to maximize the overall fit in a tree-type fashion:

$$\hat{c}_k = \arg\max \Big(L(\hat{\beta}_{k,L}(c)|\mathbf{Y}_{kc,L}, \mathbf{X}_{kc,L}) + L(\hat{\beta}_{k,H}(c)|\mathbf{Y}_{kc,H}, \mathbf{X}_{kc,H}) \Big).$$

The conditional distribution of Y given X is now

$$\hat{f}_{Y|X}(y|x) = f_{Y|X}(y|x; \beta_{k,L}(\hat{c}_k)) \cdot \mathbf{1}_{\{x_k \leq \hat{c}_k\}} + f_{Y|X}(y|x; \beta_{k,H}(\hat{c}_k)) \cdot \mathbf{1}_{\{x_k > \hat{c}_k\}}.$$

This leads to the estimated effect of the policy:

$$\hat{\theta}_k = g\left(f_{Y|X}(\cdot|\cdot; \hat{\beta}_{k,L}(\hat{c}_k)) \cdot \mathbf{1}_{\{x_k \leq \hat{c}_k\}} + f_{Y|X}(\cdot|\cdot; \hat{\beta}_{k,H}(\hat{c}_k)) \cdot \mathbf{1}_{\{x_k > \hat{c}_k\}}, X_1, \ldots, X_N \right). \quad (3.2)$$

Rather than reporting the full set of estimates $\hat{\theta}_k$, we then calculate and report the variation of these estimates around the base-model estimate:

$$\hat{\sigma}_\theta = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(\hat{\theta}_k - \hat{\theta}_B \right)^2}. \quad (3.3)$$

4 Four Applications

Let us consider the measure of sensitivity in four empirical examples. The first two were introduced already, the experimental and non-experimental Lalonde data. In addition we consider lottery data analyzed in Imbens, Rubin and Sacerdote (2001). In that case the authors are interested in the marginal propensity to earn out of unearned income. They estimate this propensity using a linear model relating average yearly earnings to the annual lottery payment, using eighteen additional covariates, twelve of them related to prior earnings and the remaining six characteristics of the lottery players including education, age and sex. In the final illustration we focus on a regression relating the logarithm of weekly earnings to years of education, using data from the National Longitudinal Study of Youth. We include in the regression experience, experience-squard, mother's education, father's education, and two ability measures, iq and a test score referred to as kww (knowledge of the world of work).

In each case we estimate the estimand using the basic model, its standard error, and also calculate the standard deviation of the estimates based on the alternative models. Table 3 presents the results, including also the ratio of the standard deviation and the

standard error. We find that for all studies the variation in estimates due to changes in the specification is substantial. More important, however, is the finding that there is substantial variation in this ratio, from 0.20 to 3.38, suggesting that some studies are much more sensitive to alternative specifications than others. Reporting this sensitivity may be useful for assessing the credibility of empirical studies.

Table 3: Variation of $\hat{\theta}$ over Model Specifications

	Exper	Non-exp	IRS	NLS
Est	1.67	1.07	-0.44	0.059
(s.e.)	(0.67)	(0.63)	0.012)	(0.010)
σ_θ	[0.13]	[2.13]	[0.10]	[0.004]
ratio	0.20	3.38	0.83	0.40

5 Discussion

In most empirical work researchers present estimates based on a preferred specification, sometimes in combination with a handful alternative specifications. Such practices have long been critized as leading to estimates that may be sensitive to assumptions in a way that is not captured by the reported measures of precision. For example, Leamer (1983) in his celebrated paper "Let's Take the Con out of Econometrics," argues that the credibility of much empirical work is very low partly because of specification searches that underly the estimates for the reported models. He proposes reporting estimates based on a larger class of models for a range of prior distributions. White () proposes codifying the model searching process and reporting standard errors that take into account this process. Our proposals are complementary to those by Leamer (1982, 1986) 0and White (200). We propose considering a much larger class of models, for which choosing an appropriate class of prior distributions in the spirit of Leamer's work would be challenging. Rather than fully specifying the process of model selection as White (2000) suggests, we are comfortable with reporting conventional standard errors for the preferred model, but we suggest combining this with a simple scalar measure of the sensitivity of the estimates to a range of alternative models that nest this preferred specification.

Our measure of sensitivity to model specification is the standard deviation of the estimates over a range of alternative models. These alternative models are generated by splitting the sample based on values of all the exogenous variables in the original model, one or more at a time, with the splitting threshold determined by the fit.

Although our proposed measure of sensitivity does not capture all aspects of sensitivity that one might wish to capture, we argue, partly based on four applications, that

this measure does shed light on the credibility of empirical evidence.

The main limitation of our approach is the manner in which the set of alternative models is selected.

References

Breiman, L., J. Friedman, and C. Stone (1984), *Classification and Regression Trees*, Wadsworth.

Dehejia, R. and S. Wahba, (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, Vol. 94(448): 1053-1062.

Imbens, G., D. Rubin, and B. Sacerdote, (2001), "Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players," *American Economic Review*, Vol. Vol. 91(4): 778-794.

Lalonde, R., (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, Vol. 76(4): 604-620.

Leamer, E., (1983), "Let's Take the Con out of Econometrics," *American Economic Review*, Vol. 73(1): 31.43

Leamer, E., (1982), "Sets of Posterior Means with Bounded Variance Priors," *Econometrica*, Vol. 50(2): 725-736.

White, H., (2000), "Data Snooping: A Reality Check," *Econometrica*, Vol. 68(5): 1097-1126.