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 typically then reports the point estimate of the policy effect as well as a measure of the uncertainty surrounding that point estimate (e.g., the standard error). Such measures of uncertainty depend on the model specification. Often, however, there is also substantial uncertainty regarding the specification of the model. There are typically elements of the specification that are relatively ad hoc, and different choices for those specification decisions would have left to different point estimates of the effect of the policy.

Previously, Leamer (1982, 1983) 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 (2000) proposes codifying the model searching process and reporting standard errors that take into account this process. However, it is not standard practice to follow these proposals. In practice researchers typically address the issue by reporting estimates for a small number of different specifications to assess the sensitivity of the estimates to model specification.

In this paper we propose a complementary but still systematic approach to assessing the sensitivity of point estimates to model specification. 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. We suggest supplementing conventional standard errors with a simple scalar measure of the sensitivity of the estimates to a range of alternative models. Each element of this set of specifications includes the base specification as a special case, but that is the only common specification for any pair of general models. We estimate each model in the set, and our proposed robustness measure is the standard deviation of the point estimate of the effect of the policy over the set of models. Each member of the set of model specifications is generated by splitting the sample into two subsamples based on covariate values, estimating the model separately for each subsample, and then combining the results to form a new estimate of the overall effect.

I. A Simple Example

We illustrate the basic ideas using two data sets originally constructed by Robert Lalonde (1986), and subsequently widely used in the evaluation literature (e.g., Rajeev Dehejia and Sadek Wahba, 1999). The first data set, 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. There are 185 individuals in the treatment group and 260 in the control group. The second data set, 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 com-
comparison group drawn from the Current Population Survey (CPS). The 15,592 individuals in the CPS are substantially different from those in the experiment, which suggests that the functional form of the model used to adjust for covariates could matter for the estimated treatment effects.

As the base model consider a linear regression with outcome $Y_i$ (earnings in 1978, in 1,000’s of dollars), an indicator for participation in the training program, $X_i$, an intercept, and ten characteristics of the individual, collectively 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:

$$E[Y_i|X_i, Z_i] = \alpha + \theta_B \cdot X_i + \gamma'Z_i.$$  

The coefficient on the program, $\theta_B$ (where “B” indicates base model), is the object of interest. Let $\hat{\theta}_B$ be the least squares estimator for $\theta_B$.

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). The estimates are similar, both in terms of the point estimate and in terms of the precision.

Clearly, however, the experimental estimate is more credible. We know a priori that, because of the randomization, there are no unobserved confounders, whereas with the non-experimental data we cannot be sure of that. The superior credibility of the experimental estimates is a reflection on substantive information about the data-generating process. There is information in the data, however, about the relative robustness of the two estimates, and that is not summarized in the combination of the point estimate and the standard error (s.e.). In this paper we propose that researchers report an additional statistic that conveys at least part of the evidence from the data about robustness.

The basic idea is that in addition to the base model we systematically explore a range of alternative specifications. Let us start by considering a single additional specification. We split the sample into two subsamples by the indicator for being unemployed in 1975, and estimate separately the same linear regression (except we omit that indicator since it is constant within each subsample). We then combine the two estimates for the effect of the training program using the relative size of the two subpopulations (35% were not employed in 1975), 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, substantially different from the original estimate of 1.07.

It is the finding 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 the estimate for the experimental data that is the focus of this paper. 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 are likely to be similar.

Now we begin to generalize the approach. The statistical model provides a parametric approximation to the conditional mean $\mu(x, w) = E[Y_i|X_i = x, Z_i = z]$. The estimand $\theta$ is the average effect of the treatment. In terms of the conditional expectation function, assuming selection on observables or unconfoundedness, the estimand is $\theta = E[\mu(1, Z_i) - \mu(0, Z_i)]$. Given an estimate of the conditional expectation function, $\hat{\mu}(x, z)$, we estimate the target $\theta$ as

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} \left( \hat{\mu}(1, Z_i) - \hat{\mu}(0, Z_i) \right).$$
The base model assumes $\mu(x, z) = \alpha + \theta \cdot x + \gamma'z$ and estimates the parameters $\beta = (\alpha, \theta, \gamma)$ by least squares, leading to a base-model estimate $\hat{\theta}_B$. Let $\mathcal{P}_B$ denote this base-model specification.

Let $\mathcal{P}$ be a set of alternative specifications, and let $\mathcal{P}$ denote a typical specification. Each specification in this set includes the base specification as a special case. We can think of the parameters of the specification $\mathcal{P}$ as the pair $(\beta, \delta)$, where $\delta = 0$ indicates the base specification. We choose the set of specifications so that for any pair $\mathcal{P}$ and $\mathcal{P}'$ the only specification they have in common is the baseline model.

The question is how we choose the set $\mathcal{P}$. Our approach here is to start with the set of all possible partitions of the covariate space, $W = X \times Z$ into two subsets, $W_L$ and $W_H$. With the model now allowing for separate parameters in two subsets of the covariate space, this is a large and very rich set of possible specifications that is invariant to many transformations of the covariates. Consider splitting the sample into two subsamples, based on the values of of the covariates, $X_i$ and $Z_i$. For $s \in L, H$, let $I_i = s$ if $(X_i, Z_i) \in W_s$. In the regression example we estimate the two regression models, one for each subsample,

$$Y_i = \alpha_s + \theta_s \cdot X_i + \gamma_s'Z_i,$$

for the $I_i = s$ subsample, $s \in L, H$. 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 and thus for $\theta$:

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

$$\hat{\theta}_P = \hat{T} \cdot \hat{\theta}_H + (1 - \hat{T}) \cdot \hat{\theta}_L,$$

where $\mathcal{P}$ indexes the specification, corresponding to the split of $W$ into $(W_L, W_H)$, and $\hat{T}$ is the fraction of observations such that $(x, z) \in W_H$. It is the magnitude of the difference between the original base-model estimate $\hat{\theta}_B$ and the estimate based on the split sample, $\hat{\theta}_P$, that we focus on.

The final question is how many, and which, of the possible splits of the covariate space we wish to consider. Given a sample of size $N$ there are in principle $2^N$ ways of splitting the sample. Many of these splits are very similar, and so comparing them is unlikely to generate any differences in estimates of $\theta$. Moreover, essentially random splits are unlikely to shed light on the sensitivity of the base model. The challenge is to come up with a systematic subset of the set of all sample splits that is informative, or, more generally a set of weights for this large set of possible splits. Here, as a first and simple-to-implement proposal, we split the sample by every covariate, that is, every element of $X_i$ and every element of $Z_i$. 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).

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

$$(I.2) \quad \hat{\sigma}_\theta = \sqrt{\frac{1}{\#(\mathcal{P})} \sum_{P \in \mathcal{P}} \left(\hat{\theta}_P - \hat{\theta}_B\right)^2},$$

and it is this measure of robustness we focus on. In Table I we report the estimates $\hat{\theta}_P$ 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. By this measure, the results for the experimental data are far more robust than those for the non-experimental data. We also report the test statistic for the F-test of each more general model against the null that the base model is correct, which is distributed $\chi^2(10)$. Although the question of mis-specification is
distinct from non-robustness of the treatment effect estimate, the two questions are related.

II. The General Case

Suppose we have a random sample from a large population of a pair of variables \((Y, X)\), where both \(Y\) and \(X\) may be vectors. The object of interest is a functional of the conditional distribution of \(Y\) given \(X\) and the sample values of \(X\):

\[
\theta = g(f_{Y|X}(y|x), X_1, \ldots, X_N).
\]

(II.1)

Consider a parametric model for the conditional distribution of \(Y\) given \(X\):

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

with \(\beta\) an unknown parameter. Given the maximum likelihood estimate \(\hat{\beta}\) for \(\beta\), we can estimate \(\theta\) for this base model as

\[
\hat{\theta}_B = g(f_{Y|X}(y|x; \hat{\beta}), X_1, \ldots, X_N).
\]

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

The larger set of models \(P\) is based on splits of the basic sample in terms of the covariates. Let the number of covariates be \(K_x\). For the \(k\)th covariate, and given a threshold \(c_k\), define the subsample indicators, \(I_{ikc_k} = L\) if \(X_k \leq c_k\) and \(I_{ikc_k} = H\) otherwise. Denote the values of the outcomes in the subsample with \(I_{ikc} = L\) by \(Y_{kc,L}\) and the covariates in this subsample by \(X_{kc,L}\). Denote the values of the outcomes and covariates in the subsample with \(I_{ikc} = H\) by \(Y_{kc,H}\) and \(X_{kc,H}\). We estimate the values of the parameters in the two subsamples

\[
\hat{\beta}_{k,L}(c) = \arg\max_{\beta} L(\beta|Y_{kc,L}, X_{kc,L})
\]

\[
\hat{\beta}_{k,H}(c) = \arg\max_{\beta} L(\beta|Y_{kc,H}, X_{kc,H})
\]

We choose the threshold to maximize the overall fit:

\[
\hat{c}_k = \arg\max (L(\hat{\beta}_{k,L}(c)|Y_{kc,L}, X_{kc,L}) + L(\hat{\beta}_{k,H}(c)|Y_{kc,H}, X_{kc,H})).
\]

The \(k\)th element of the set of model specifications, \(P_k\), corresponds to the conditional distribution indexed by parameters \(\beta_{k,L}\) and \(\beta_{k,H}\):

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

This leads to the estimated effect of the policy for the \(k\)th model specification \(P_k\):

(II.2)

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

The variation of these estimates around the base-model estimate is our proposed robustness metric, as in (I.2).

III. 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. Imbens, Rubin and Sacerdote (2001) use data from lottery winners to estimate 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 related to prior earnings as well as demographics. Finally consider a regression relating the logarithm of weekly earnings to years of education, using data from the National Longitudinal Study of Youth (NLSY), with covariates experience, experience-squared, mother’s education, father’s education, IQ, and a test score called “knowledge of the world of work.”

In each case we report the estimate and standard error from the basic model, and also calculate the standard deviation of the estimates based on the alternative models. Table II presents the results. For all studies the variation in estimates due to changes in the specification is substantial. More important, there is substantial variation in the ratio of the standard deviation to the stan-
Table I: Variation of $\hat{\theta}$ over Model Specifications: Lalonde Data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Experiment $\hat{\theta}_B$ (s.e.)</th>
<th>Non-experiment $\hat{\theta}_B$ (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Model</td>
<td>1.67 (0.67)</td>
<td>1.07 (0.63)</td>
</tr>
<tr>
<td>$\hat{\sigma}_0$</td>
<td>[0.13]</td>
<td>[2.13]</td>
</tr>
<tr>
<td>Split on</td>
<td>$\hat{\theta}_P$</td>
<td>$X^2(10)$</td>
</tr>
<tr>
<td>Treatment</td>
<td>1.58</td>
<td>22.8</td>
</tr>
<tr>
<td>Age</td>
<td>1.55</td>
<td>10.1</td>
</tr>
<tr>
<td>Black</td>
<td>1.71</td>
<td>11.4</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1.61</td>
<td>7.2</td>
</tr>
<tr>
<td>Married</td>
<td>1.87</td>
<td>11.0</td>
</tr>
<tr>
<td>Education</td>
<td>1.77</td>
<td>14.7</td>
</tr>
<tr>
<td>No Degree</td>
<td>1.33</td>
<td>18.6</td>
</tr>
<tr>
<td>Earn 74</td>
<td>1.64</td>
<td>11.2</td>
</tr>
<tr>
<td>Earn 75</td>
<td>1.63</td>
<td>8.9</td>
</tr>
<tr>
<td>Unempl 74</td>
<td>1.64</td>
<td>11.2</td>
</tr>
<tr>
<td>Unempl 75</td>
<td>1.71</td>
<td>7.0</td>
</tr>
</tbody>
</table>

IV. Discussion

The main limitation of our approach to robustness is that the method of selecting a set of models and the weights assigned to them is ad hoc. To evaluate the merits of our approach, we propose considering a set of desired features for generating sets of alternative models. First, the alternative models should be close to the original model and nest it as a special case. We achieve this by maintaining the original model, but allowing its parameters to vary with partitions of the covariate space. Second, it should be robust to ad hoc decisions about the functional form of covariates as well as which ones to include. Our approach uses model fit to determine how to split the sample according to a single covariate, and thus is robust to monotone transformations of each covariate, but is not, for example, robust to alternative specifications of covariate interactions or linear transformations of the vector of covariates. Future work will explore this further. Third, the set of models should be rich enough to discover important sources of non-robustness, yet each individual model should be distinct from others. This can be addressed in future work through the specification of the model generation process as well as through data-driven approaches to weighting alternative models. We thus view our approach as a first step that achieves many, but not all, of these criteria in a natural and intuitive way. Despite its limitations, our method seems to shed light on the credibility of empirical evidence in our applications.

V. References


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

<table>
<thead>
<tr>
<th></th>
<th>Experiment</th>
<th>Non-experiment</th>
<th>Lottery</th>
<th>NLSY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>1.67</td>
<td>1.07</td>
<td>-0.44</td>
<td>0.059</td>
</tr>
<tr>
<td>(s.e.)</td>
<td>(0.67)</td>
<td>(0.63)</td>
<td>0.012</td>
<td>(0.010)</td>
</tr>
<tr>
<td>$\sigma_\theta$</td>
<td>[0.13]</td>
<td>[2.13]</td>
<td>[0.10]</td>
<td>[0.004]</td>
</tr>
<tr>
<td>s.e./$\sigma_\theta$</td>
<td>0.20</td>
<td>3.38</td>
<td>0.83</td>
<td>0.40</td>
</tr>
</tbody>
</table>


