Estimation of Causal Effects from Observational Study of Job Training Program

Dmitry Arkhangelsky and Rob Donnelly
darkhang@stanford.edu & rodonn@stanford.edu

1 Introduction

In the social sciences, researchers are often interested in measuring the effect of a treatment or intervention. For example, many labor economists are interested in measuring the effect of job training programs on increasing salaries or reducing unemployment. Measuring the "causal effect" of an intervention is relatively straight forward when the treatment is given to randomly selected individuals. In this case the difference in the average outcome of the treatment and control groups can be directly compared.

Unfortunately in many policy relevant situations economists do not have data from a randomized controlled experiment. Instead they have data on the characteristics and outcomes of a set of individuals who received the treatment, but no directly comparable control group.

We tried two approaches to estimating the effect of a job training treatment on wages. The first approach predicts counterfactual wages for the treated individuals based on models trained on a large sample from the general population. The second approach matches the treated individuals to people from the general population based on propensity scores. We find that the second approach yields more plausible estimates of the effect of job training.

2 Dataset, Features, and Preprocessing

We are using data from the National Supported Work Demonstration, which gave 12-18 months of job training to unemployed adults at 15 sites around the US. We have data on 185 men who participated in the job training, as well as 15,992 adults from the general population[1]. In order to be eligible to participate in the job training, an individual had to meet certain eligibility requirements and had to volunteer. Because of this, the pool of participants is very different from a random sample of the population[2]. Previous work in the economics literature has shown that treating the data as if it came from a randomized controlled experiment leads to very biased estimates of the effect of job training programs. In particular since the average participant in a job training program has many disadvantages in the job market relative to a randomly selected person from the whole population, naive estimates often predict that the job training actually lowered wages.

Our features include the age, years of education, race, marriage status, annual income 1 and 2 years before training program, and unemployment status 1 and 2 years before job

[1] This data is from 1972, but is still relevant since it is a commonly used baseline in the economics literature for methods of evaluating treatment effects from non-experimental data.

[2] Participants in this job training program on average have less education and are more likely to have prior criminal convictions or to be unemployed.

training. We normalized all of our variables by subtracting the mean and dividing by the standard deviation. To engineer additional features we also tried including cross products of the base features in our dataset up to degree 3.

3 Models and Results

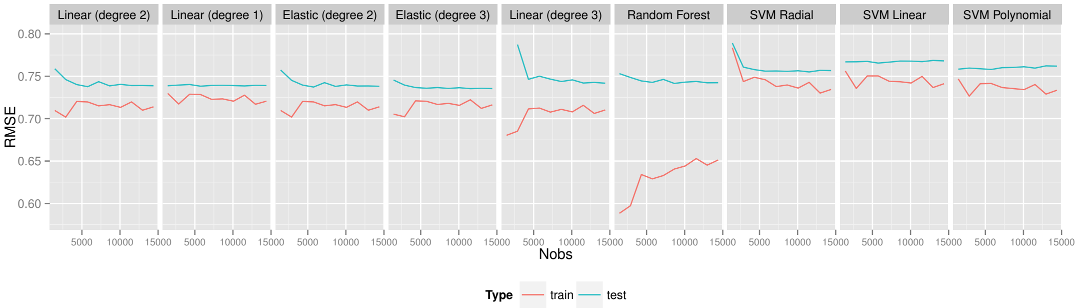
We used two distinct strategies for estimating the treatment effect. Our first approach is a more straightforward one – we use several models to predict the final wages of each individual in the large sample from the general population. We then used these models to predict wages each individual would have gotten if they hadn't received training. The difference among the individuals who received training between their actual final wage and their predicted final wage, can be used as an estimate of the change in wage caused by the job training, i.e. a casual effect.

The second approach comes from the program evaluation literature. First, we estimate the propensity score – conditional probability of being in the treatment group – and then compare the treatment outcomes between the treated and untreated individuals who have similar predicted probabilities of receiving the treatment.

3.1 Regression models

We tried several types of models for predicting wages: linear regression, lasso regression[3], SVM (with linear, polynomial, and radial kernels), and random forest models. We choose parameter values for each model using a grid search with 10-fold cross validation using Root Mean Squared Error (RMSE) as our selection criteria. We evaluated performance on a 10% hold-out set. The exponents following a model's name indicate the degree of the polynomial interactions of features that were included[4].

By looking at the learning curves for these models it became clear that all of the models had high bias. We attempted unsuccessfully to get access to more detailed data with a larger number of features for each individual. Adding polynomial interaction terms did not substantially reduce test error.



[3] Linear regression with L1 regularization to reduce overfitting.

[4] E.g. a 2 means that we added new features corresponding to the product of every pair of features in the base dataset.

	Training RMSE	Test RMSE	Treatment Effect
Linear^2	0.71702	0.70945	-$1787
Linear^1	0.72407	0.71039	-$1242
Lasso^2	0.72301	0.71319	-$1773
Lasso^3	0.72281	0.71324	-$1655
Linear^3	0.71309	0.71365	-$1830
Random Forest	0.65542	0.72046	-$1034
SVM (radial)	0.73545	0.72446	-$2671
SVM (linear)	0.74588	0.72951	-$2658
SVM (polynomial)	0.72340	0.73667	-$2705

Linear regression with degree two polynomial interactions gave the lowest RMSE on the hold-out set. The random forest's low training error and higher test error suggests overfitting. None of the SVM models were very successful, even after trying a wide selection of different values for the parameters. Since we are predicting a continuous outcome (salary) that is normalized to unit variance, squaring the RMSE gives a measure of the fraction of variation in salary that is unexplained by the model, roughly 50%. Since there are many determinants of future wage (the industry you work in, work ethic, etc) that we do not observe, it is not surprising that we are unable to explain all of the variation in wages.

We used these models to predict what wage the individuals in the job training would have gotten if they had not participated. The difference between predicted wage and actual wage for the individuals who received job training is the used as the estimate of the effect of the training. All of the models produced negative estimates of the effect of the job training. Measurements of the effect of job training programs that come from randomized experiments consistently find positive effects, which suggests our first approach is not producing accurate estimates of the effect of the job training.

3.2 Propensity score models

Our second approach was propensity score stratification. We first train several models to estimate a propensity score for each individual, the likelihood of he was given job training[5]. Then, we cluster observations with similar estimated propensity score and estimate average effect within each group. This allows us to compare the treated individuals with individuals from the general population who had a similar likelihood of being offered the treatment.

We use several conceptually distinct models for estimating propensity scores: logistic regression, lasso, decision tree, boosted tree and random forest. Again, tuning parameters were chosen by 10-fold cross-validation. As an accuracy measure we use the squared difference between our continuous estimate of the propensity score and the binary treatment indicator. It's worth emphasizing that perfectly predicting who received training would not advantageous here, since then the model would suggest that no person from the general population is comparable to anyone in the treated population. Despite this conceptual strangeness, propensity score matching has a substantial theoretical grounding (Caliendo, 2008)

[5] I.e., we are predicting a continuous estimate of the binary outcome, ”Training/No Training”

Training and test errors for all procedures that we used are presented in the following table. Since less than 1% of our data received the job training treatment, we also report conditional errors – average error for treated observations.

	Train Error	Test Error	Cond. Train Error	Cond. Test Error	Balance	# Groups	Treatment Effect
Logistic[1]	0.0077	0.0078	0.5112	0.5252	0.87	6	$1855
Logistic[2]	0.0058	0.0074	0.3475	0.4790	1.54	8	$1069
Lasso[3]	0.0067	0.0075	0.4210	0.4903	0.84	6	$1417
Tree	0.0066	0.0096	0.4236	0.6068	1.72	7	$136
Boosted Tree	0.0091	0.0100	0.6930	0.7317	1.49	5	$1075
Random Forest	0.0044	0.0077	0.4166	0.5358	0.56	4	$1778

Due to the unbalanced distribution of the outcome variable, it's unsurprising that the overall accuracy is much higher than the accuracy on the treated individuals. Overall both measures of test error seem to agree on which models were more accurate. Logistic regression and Lasso seemed to outperform the tree based estimates. However, it's unclear whether the method with the lowest test error is the best in terms of estimating the overall treatment effect.

To cluster observation into several groups we took observed treatment status and ran a simple decision tree using estimated propensity score as the only covariate. This procedure results in groups of people with similar propensity score. Another possible procedure is to cluster observations using some unsupervised algorithm like K-means. However, one important advantage of using trees is that the number of classes is selected automatically.

For each model of propensity score we report the average treatment effect, assessed balance and the number of classes. Formally, we compute the following statistics:

$$\text{Average Treatment Effect} = \frac{\sum_{i=1}^{K} N_{1i}\left(\overline{Y}_{i1} - \overline{Y}_{i0}\right)}{\sum_{i=1}^{K} N_{1i}} \tag{1}$$

here i is a generic class, N_{1i} is the number of treated individuals in this class, $\overline{Y}_{ik}$ is the average outcome in i-th class, for k group (either treatment or control).

To assess balance we compute the following normalized sum of squares:

$$\text{Balance} = \sum_{j=1}^{m} \frac{\sum_{i=1}^{K} \frac{\left(\overline{X}_{ji1} - \overline{X}_{ij0}\right)^2}{\hat{\mathbb{V}}[X_{jk1}] + \hat{\mathbb{V}}[X_{jk0}]}}{K} \tag{2}$$

where $\hat{\mathbb{V}}[X_{jik}]$ is the estimated variance of the j-th covariate in the i-th class and k-th group. A lower value corresponds to more similarity between the treated and untreated individuals within a group, which is desirable.

The best model in terms of balance is the random forest model. There was considerable variation in the estimated treatment effect between different models, but all of the models predict positive effects on wages from the job training and the magnitudes are plausible and increase in salary of roughly \$1000 for a group of individuals whose previous salaries averaged \$6349[6].

4 Conclusions and Future Research

When researchers have used randomized controlled experiments to evaluate similar job training programs they generally find the training leads to a wage increase of \$800 to \$1600. With this as a baseline, the first approach was wildly inaccurate since every model predicted a substantial negative effect. This approach might be more successful on a dataset with more features, since all of the models tried had high bias. Having more information about each individual might allow us to generate models that can explain more of the variation in wages. If we were able to create highly accurate predictions of each individuals wage, then we would be able to generate accurate predictions of what each individual in the job training program would have earned if he didn't participate. We would then have an accurate estimate of the effect of the job training even without running a randomized controlled experiment.

In contrast the propensity score based estimates all had the correct sign. With the exception of the simple tree model, all of the estimate treatment effects are consistent with the \$800 to \$1600 range found in experiments of similar programs. The high variation suggests that only limited confidence can be had in the estimates from any single propensity score model. Running several different models, as we have done here, may give researchers a sense of the general range of outcomes they should expect if the program were re run as a true experiment.

For future research in this area, we are also interested in testing out unsupervised learning algorithms as mechanisms for clustering individuals into discrete groups and then comparing the treatment effects within each group. Approaches like this might also allow researchers to know not just the average affect of the treatment across the whole population, but also have predictions of which subsets of the population the treatment worked especially well on.

5 References

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys.*

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 604-620.

Smith, Jeffrey A., and Petra E. Todd. ”Reconciling conflicting evidence on the performance of propensity-score matching methods.” *American Economic Review* (2001): 112-118.

[6] This data is from 1978, so adjusting for inflation, this corresponds to increasing an increase of \$3,500 over a salary of \$23,000 in 2014 dollars.