Applying Machine Learning to
'The Development Impact of a Best Practice Seasonal Worker Policy
by John Gibson', *University of Waikato,* David McKenzie, *World Bank*

Team: Tiffany Cheng, David Kastelman, Yue Li, Abuzar Royesh

06/07/2019

Introduction (Research Questions)

In *The Development Impact of a Best Practice Seasonal Worker Policy*, Josh Gibson and David McKenzie find that a program that allowed individuals from Tonga and Vanuatu to work seasonally in New Zealand had a large, significant impact on household income for the seasonal workers (on the order of 30%).[1] Their primary methods were regression based, namely a difference-in-difference (DD) regression and a fixed effect regression. The authors did create propensity scores, but used them just for filtering data before running their regression:[2] meaning they used the scores neither for weighting nor matching. Given the lack of a more comprehensive use of propensity scores, the authors' findings are largely dependent on their assumed models being accurate, although they point out that they get similar results for both their regression specifications. Moreover, in defending their use of the DD model, the authors note "we have a plausible reason why some households participated in the [program] and other households with similar characteristics did not - there was excess demand for [the program], and so not all households who wanted to participate were able to."

Based on the above, the research questions for our final project are twofold: first, can we not only replicate the authors' encouraging findings but also make them doubly robust to increase our confidence in the existence of the ATE? Secondly, given the surplus of interest in the program, can we say something about impact heterogeneity in a way that would be useful for policy formation?

Data Set

We used the data set published on the World Bank's data catalog as being associated with Gibson and McKenzie's paper (i.e., *TongaReplicationData.dta*). A summary of the data set and additional details are in the Appendix B under the "Data Description" heading. Note that because of the data limitations discussed in Appendix B section II, we focus on the fixed effect regression rather than the difference-in-difference regression in our replication and extensions.

Methodology

Replicating ATE and Making Estimates Doubly Robust

[1] In fact, the authors suggest the program "has indeed had positive development impacts, which dwarf those of other popular development interventions."

[2] The authors simply filtered observations to those that had a propensity score between .1 and .9 and then ran their regressions regression. See Table 1 in Gibson and McKenzie.

We start our analysis by replicating the authors' finding of a large ATE using a fixed-effects regression in table 2 in Appendix A. As noted previously, all of the authors' findings depend on the assumptions of their models, with the assumptions detailed in footnote 3.[3] In an effort to overcome this limitation, we turn to making the estimate "doubly robust." We give a detailed description of the advantages of the doubly robust estimate, and why it provides additional evidence, in Appendix B, under the "Doubly Robust Advantages" heading.

In validating the existence of an ATE, we make use of machine learning in two ways. First, we estimate our propensity score using a random forest to mitigate the possibility of incorrectly estimated weights. Secondly, after first using our propensity scores and our base fixed-effect regression in a doubly robust estimate, we formulate a second AIPW estimate using the causal forest package. This estimate replaces our fixed-effect regression with a random forest to predict semi-annual income, which loosens the parametric requirements of the regression model to further validate the existence of the ATE.

In order to create our propensity scores, we filtered the data for the 1st wave, which included 439 respondents. We defined 397 variables in the dataset (all variables excluding those on income, wave, and RSE information) as covariates (X), semi-annual per capita income (pa'anga) as the dependent variable (Y), and whether someone was in treatment or control in the 4th wave as treatment variable (W). The formula for the doubly robust estimator is in Appendix A, Figure 1. We used the *causal_forest(X, Y, W)* command from the grf package to implement AIPW and form confidence intervals.

Further Data cleaning for Causal Forests and finding AIPW ATE estimate

Since our dataset involved following a group of people over time, we modified the dataset to account for dependent data (multiple observations from a households over time) and the fact that people enter the treatment group at different points of the study. First, we defined the control group as people who never underwent the seasonal workers' program throughout the study and any waves prior to joining the season workers' program for households that did participate in the program. This dataset is used later to find the heterogeneous treatment effects.

[3] For the difference-in-difference regression, the core assumption is that the control and treatment households' income follows the same trajectory over time except for an additive impact of the intervention. For the fixed-effect-regression, the assumption is quite similar in this case: that the treatment and control groups would perform similarly in the absence of an intervention when controlling for the households' mean income over the study period. Note the similarity in the core assumptions of both types of regressions run by the authors, making the authors reliance on model specification being accurate much less than desirable.

After doing the necessary cleaning for the causal forest, we then use the causal forest package to get a second AIPW ATE estimate, as previously described. The predicted ATE is reported in Appendix A, Table 5.

Finding Heterogeneous Treatment Effects

In order to compute heterogeneous treatment effects, if a household was treated over many consecutive waves, we kept data from the first wave they were treated and removed subsequent waves since we were not familiar with methods for measuring compounding treatment effects. We defined our outcome variable as the difference between the log of the previous wave's reported income and the log of the current wave's reported income (i.e., the percent change in income between the two waves). We chose to log-scale our outcome variable due to high variance across households in terms of income, which led to having large standard errors in our causal forest. We also removed covariates with more than 5% NA values, and we imputed missing data with the column median, since some of the methods we used did not handle NAs well. Even after applying this threshold, 103 columns remained.

Because the data set had so many covariates, we tried two methods for variable selection: lasso and causal forest variable importance. With the selected variables, we fitted several models: S-learner, T-learner, X-learner, and causal forest. We then used two methods to assess heterogeneity: test calibration and splitting by quartiles to see whether we could observe any heterogeneous effects. Further analysis investigating heterogeneous treatment effects is listed in Appendix B, part IV.

Findings

We focused on data from the island country of Tonga, and we were able to replicate the baseline means reported by Gibson and McKenzie regarding household characteristics (Appendix A, Table 1).

In the regression (Appendix A, Table 2), the mean estimate of a 214.89 impact to per-capita semi-annual income from the guest program resembles Gibson and McKenzie's ATE estimate. The units are in pa'anga, which are around 2 to 1 USD (both now and at the time of the study). Confidence intervals are also included in the table.

Validating propensity scores generated through Causal Forest:

Appendix A, Figure 2 and Table 3 show the histogram of propensity scores generated through running a Causal Forest model, which shows that while there is a bimodal distribution, we also have overlap of propensity scores away from 0 and 1. The minimum propensity score is above 0. Further, our propensity estimates have calibration that is close diagonal, suggesting a fair job of estimating the propensity score, as shown in Appendix A, Figure 3.

Using propensity scores generated through Causal Forest:

As shown in Appendix A, Table 4, outcome variable is still per-capita semi- annual income. Units are still in pa'anga. Note that the mean estimate is similar to the fixed effect regression, although with narrower confidence bounds. In Appendix A, Table 5, the units are in log semi-annual pa'anga income, so .34 is ~40% impact to per-capita income from the intervention. This once again resembles the initial estimate of Gibson and McKenzie (who saw an impact of ~30%).

Heterogeneous treatment effect outcomes

Model comparison

In Appendix A, Table 6, which compares performance between the models, we see that the T-learner has the lowest MSE. This is a surprising outcome since our data set contained a class imbalance with many more non-treated observations than treated ones. We would expect for the T-learner to perform worse than other models on the test dataset. However, overall we see that there is no significant advantage to any of the model based on the size of our standard errors. Therefore, we will run the models that test for heterogeneous effects using causal forest, since it is able to cluster at the household level and handle dependent observations, unlike the other models.

Test calibration & Quartiles

Since we had so many dimensions in our dataset, we downsized our features using two different methods: 1) running a causal forest and selecting a small subset of covariates using variable importance and 2) running a lasso. With the subset of covariates, we ran a causal forest model. The spread of CATEs are shown in Appendix A, Figure 4.

We attempted two different methods to test for heterogeneity. First, we tried test calibration from the GRF package, then we tried to split our findings into quartiles to see if there are any significant trends between CATE quartiles and AIPW.

From the results in Appendix A, Table 7 and Table 8, we see that both causal forest models (one with lasso variable selection and the other with causal forest variable selection) show no significance in `*differential.forest.prediction*`, which measures how the CATE predictions covary with true CATE. The corresponding p-value acts as a test for heterogeneity. If the coefficient is significantly greater than zero, we can reject the null hypothesis of no heterogeneity, whereas if the coefficient is smaller than 0, it is not meaningful and should not be interpreted. Since the coefficients from both of our models are negative, we fail to reject the null hypothesis of no heterogeneity. This does not mean that heterogeneity does not exist; we simply do not know for certain if it exists.

We investigate this question further by separating households into quartiles based on CATE and checking the differences in AIPW across the quartiles for both causal forest models. Both fail to show any visible trend regarding AIPW across the quartiles, as shown in Appendix A, Table 9 and Table 10.

We also ran the Wald Test, which tests the null hypothesis that average CATE is the same across quartiles. Using this test, we find a p-value of 0.24 for the quartiles created using lasso for variable selection, which does not allow us to reject the null hypothesis. Similarly, running the Wald test on the quartiles created using causal forest for variables selection yielded an insignificant p-value of 0.17. Therefore, we are unable to show that the seasonal workers' program had heterogeneous treatment effects. Further analysis into partial dependence and policy evaluations are discussed in the Appendix B, under "Further Analyses on Heterogeneity."

Conclusion

Expanding on Gibson and McKenzie's original paper, we sought to answer two questions. First, can we make the demonstrated effect doubly robust to further validate its existence, with the added benefit of increasing the efficiency of our estimate. Secondly, given excess demand for the program, can we say anything about a heterogeneous treatment impact in order to provide insights for designing a more optimal policy.

On the first question, we were in fact able to make estimated average treatment effect doubly robust. We implemented AIPW estimates two different ways, finding a significant effect both times. First we used a random forest to estimate propensity scores to make an AIPW estimate together with the original fixed effects regression. Next we loosened parametric assumptions by constructing a second AIPW estimate where we replaced the fixed effect regression with a random forest to predict semi-annual income. Whereas the fixed effect regression had a confidence interval for semi-annual income of between 47 and 392 pa'ang, the AIPW estimate using the fixed-effect regression had a confidence interval between 111 and 308, showing the increased efficiency of the estimate.

On the second question, we did not find treatment heterogeneity, as validated by both the test calibration function in the `grf` package and by separating households into quartiles based on their

CATE and comparing differences in AIPW across the quartiles. For the latter, we also ran the Wald Test, which did not reject the null hypothesis that average CATE is the same across quartiles.

Appendix A: Figures and Tables

Table 1: Replication of Household Characteristics

Variable	RSE Households	Non-RSE Households
Household Size	5.7	4.82
Number of Males 18 to 50	1.5	1.25
Share of Males 18 to 50s that:		
Are Literate in English	0.92	0.85
Have more than 10 years of schooling	0.46	0.49
Have very good self-reported health	0.68	0.6
Drank alcohol in last month	0.42	0.39
Household Durable Assets Index	0.07	-0.06
Number of Relatives in NZ	5.41	4.8
Number of Cattle	0.45	0.47
Number of Pigs	5.57	5.49
Number of Chickens	5.11	5.12
Proportion with income per capita below US$1 per day	0.19	0.12
Proportion with income per capita below US$2 per day	0.49	0.36
Share of adults who previously have worked or studied in NZ	0.38	0.2
Had a male aged 18 to 50 work for pay in early 2007	0.21	0.27
Mean days hard labor in past week males 18 to 50	4.56	3.97
Have a traditional-style dwelling	0.15	0.13
Located on Tongatapu or Efate	0.81	0.8
Mean change in weekly wage income 2006 to 2007 (pa'anga)	8.91	14.67
Semi-annual per capita consumption (pa'anga)	828.76	1184.32
Semi-annual per capita income (pa'anga)	978.5	1342.01

Table 2: Fixed Effect Regression

Call: felm(formula = pcy ~ rse_fixed.x | rse_id + wave, data = ps_df)

Residuals:

Min	1st Q	Median	3rd Q	Max
-2597.0	-294.7	-35.9	245.3	5558.9

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
rse_fixed.x	214.89	85.43	2.515	0.0121 *

Residual standard error: 708.4 on 942 degrees of freedom	
Multiple R-squared(full model): 0.639	Adjusted R-squared: 0.5167
Multiple R-squared(proj model): 0.006672	Adjusted R-squared: -0.3297
F-statistic(full model):5.227 on 319 and 942 DF	p-value: < 2.2e-16
F-statistic(proj model): 6.327 on 1 and 942 DF	p-value: 0.01206

ATE	C.I. Lower Bound	C.I. Higher Bound
214.89	46.51	391.76

Table 3: Summary of Propensity Scores

Min	1st Quartile	Median	Mean	3rd quartile	Max
0.03812	0.08438	0.26186	0.42495	0.83160	0.91439

Figure 1: Doubly Robust formula

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\Bigg(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + \frac{W_i}{\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(1)}(X_i)\right) - \frac{1-W_i}{1-\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(0)}(X_i)\right)\Bigg).$$

Figure 2: Histogram of Propensity Scores

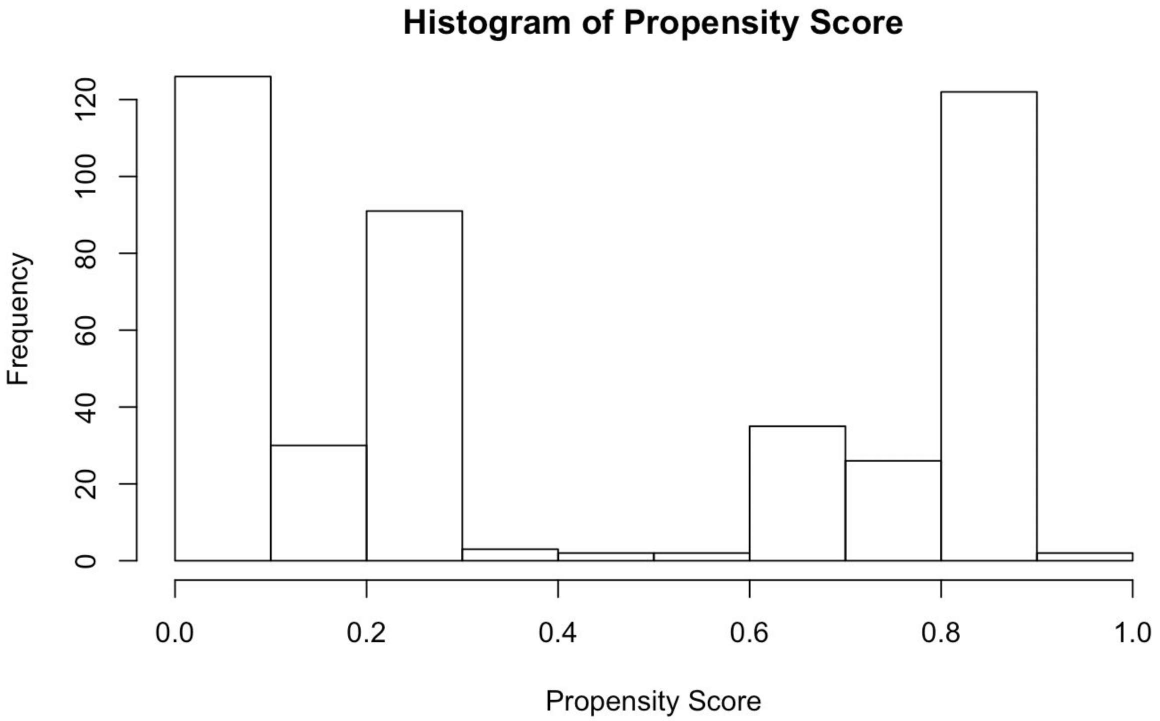

Figure 3: Calibrating Propensity Estimates

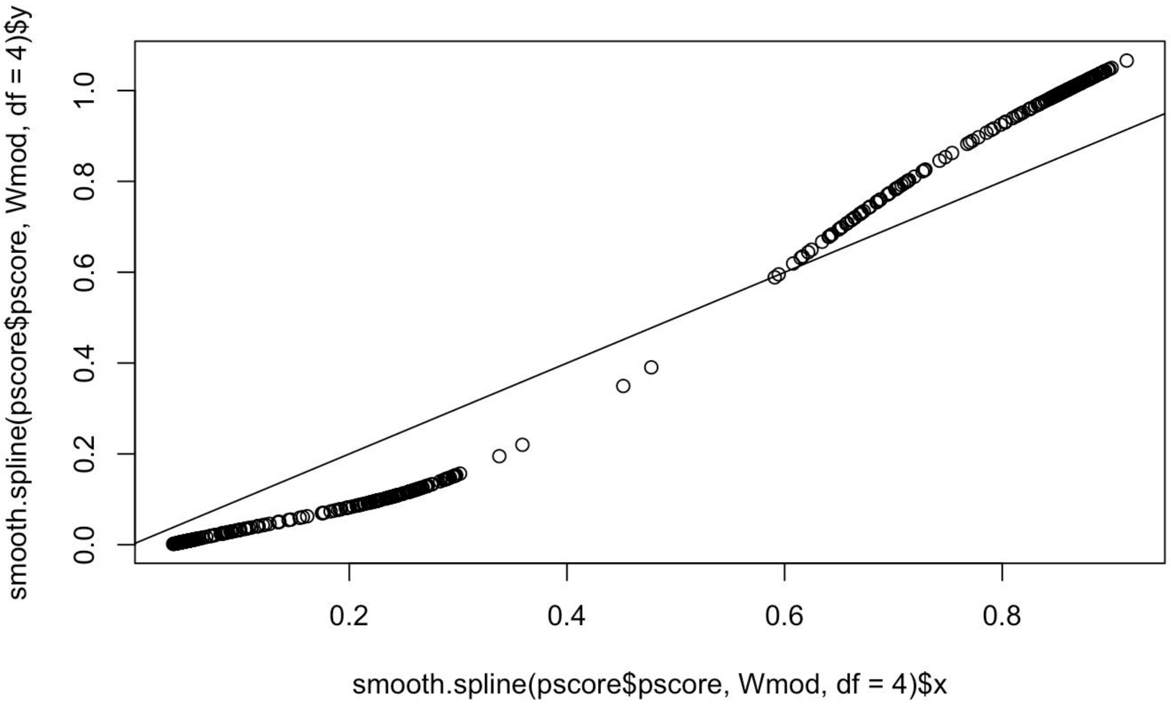

Table 4: Estimating AIPW ATE Using Random Forest Propensity Scores w/ Fixed Effect Regression

ATE	C.I. Lower Bound	C.I. Higher Bound
209.51	111.01	308.01

Table 5: Estimating AIPW ATE Using Causal Forest for Both Model and Propensity Scores

ATE	C.I. Lower Bound	C.I. Higher Bound
.34	.21	.47

Table 6: Comparing Different Models

Model	MSE	Standard Error
Sample ATE	8.78	2.50
S-learner	9.16	2.64
T-learner	8.91	2.57
X-learner	8.94	2.57
Causal forest	8.99	2.59

Figure 4: Histogram of CATEs after variable selection using causal forest and lasso

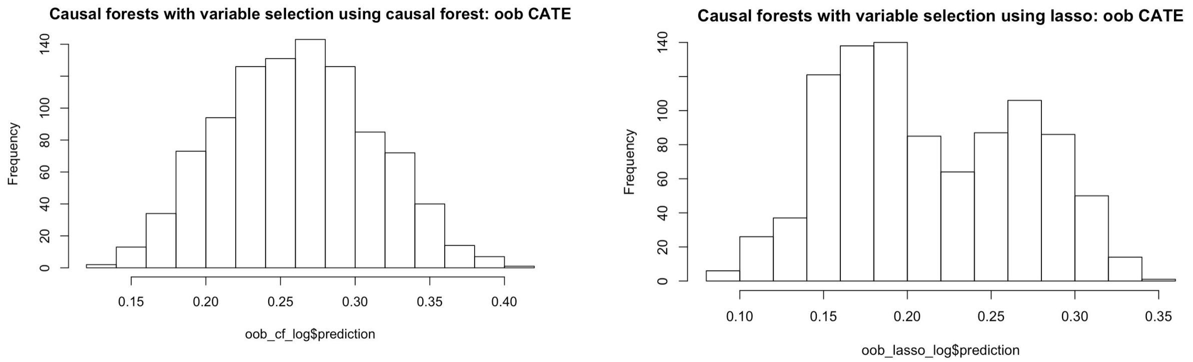

Table 7: Test calibration output after variable selection using causal forest

	Estimate	Std. Error	t value	Pr (>t)
mean.forest.prediction	1.25011	0.55439	2.2549	0.01218 *
differential.forest.prediction	-11.45570	3.22606	-3.5510	0.99980

Table 8: Test Calibration output after variable selection with lasso

	Estimate	Std. Error	t value	Pr (>t)
mean.forest.prediction	2.11015	0.63806	3.3071	0.00***
differential.forest.prediction	-10.02225	2.24570	-4.4629	0.999996

Table 8: Checking AIPW for Quartiles after variable selection using lasso

Quartiles Lasso	Mean CATE	AIPW	AIPW Standard error
1	0.15	0.46	0.20
2	0.18	-0.02	0.37
3	0.23	0.22	0.50
4	0.29	0.34	0.24

Table 9: Checking AIPW for Quartiles after variable selection using causal forest

Quartiles Causal Forest	Mean CATE	AIPW	AIPW Standard error
1	0.19	-0.17	0.65
2	0.24	0.60	0.32
3	0.28	0.07	0.22
4	0.33	0.37	0.13

Figure 5: Household Size, Total Spending, and Education Spending by CATE

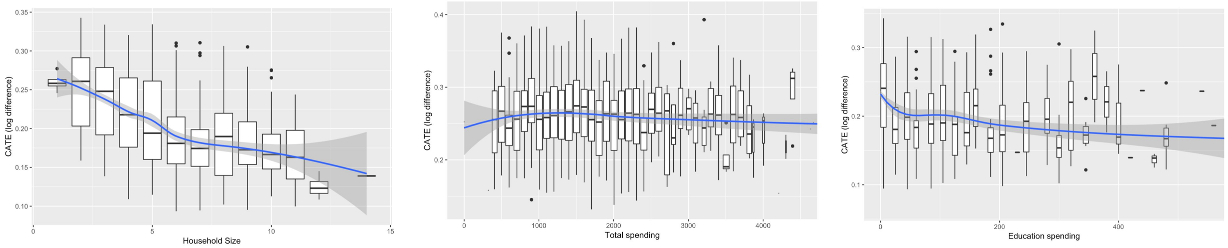

Table 11: Policy Improvements

Data	Mean	Standard Error
Train set	-0.11	0.15
Test set	-0.07	0.10

Appendix B: Additional Information

I. Data description:

The data set contains both households that participated in the RSE program as well as those that did not for comparison. It includes: 1) data from a baseline survey that was administered to each household during wave 1 before selected workers left for the program in New Zealand, and 2) data from subsequent interviews with the households at the 6, 12, and 24 month mark (i.e., waves 2 through 4). Households were interviewed on a range of topics, including financial status (total spend, education expenditures, etc.), household size, ownership of livestock, ownership of technological goods, access to transportation, access to certain types of food (e.g., certain meats, fruits, etc), and many more. There are a total of 474 covariates in the data set outside of household id, and a data dictionary containing all covariates can be found on the World Bank's page. The data set has 1,774 observations with each household being associated with between 1-4 rows depending on how many follow-up surveys they completed.

Before our analyses, we followed the same data cleaning steps as Gibson and McKenzie, which they published in a Stata file on the World Bank data catalog (i.e., *ReStat_ReplicationFile_tonga.do*). Similar to them, we created dummy variables indicating the wave that an observation came from. Also, due to differences in the surveys administered across waves, per capita income and consumption data were distributed across multiple columns, and we combined the information into one column. Finally, we ensured that migrants who were currently in New Zealand were not included in the household size count.

In clearing up confusion, for example, we noticed that households that had participated in the program in wave 2 or wave 3 were often coded as not having participated in the program in wave 4. We only saw this miscoding issue in wave 4 and it did not impact all households. We hypothesize that additional data cleaning was done on the data used by Gibson and McKenzie compared to the data set that we were able to download, which may have impacted our findings.

II. Data limitations:

One major limitation that we encountered had to do with the reliability of our data set. When we re-ran Gibson and McKenzie's code in Stata with no revisions, we found discrepancies between the output and those reported in their paper, which seem to stem from the coding of the core intervention variable, "RSEworker." We tried to create our own correctly coded RSEworker variable based on description from the paper. In doing so, we were able to closely match the estimated impact from their fixed-effect regression, but not from their difference-in-difference regression. For the duration of the paper, then, we chose to run further analysis starting from the fixed-effect-regression, but of course if we were to publish, we'd need to contact the authors, figure out the clear data discrepancies, and validate that we are correctly interpreting and using the dataset.

III. Doubly Robust Advantages

A common idea in causal inference is to use propensity scores to match or weight observations prior to analysis. While regression analysis will more often than not suffer some from omitted variable bias, propensity scores overcome omitted variable bias to a certain degree by matching observations prior to analysis so that factors correlated with the covariates used to create the propensity scores are balanced between control and treatment groups. The basic idea is to "simulate" a randomized control trial by matching (or weighting) observations to create similar treatment and control groups. This is contrat to trying to control for differences during analysis by adding factors to a regression. Of course if there are latent variables driving differences between treatment and control groups, leading to a failure in "unconfoundedness," propensity score weighting will fail to uncover an unbiased ATE, but this limitation is shared by regression modeling. One limitation unique to propensity scores is that if the propensity estimates are themselves biased or there is not good overlap between treatment and control groups, new bias can be introduced into the ATE estimate.

Given the strengths and weaknesses of regression analysis and propensity score weighting, the typical best practice in causal inference is actually to combine the two approaches through an approach called augmented inverse propensity weighting (AIPW). The basic idea is to apply weighting to the residuals of a model rather than to the model itself. The result is that if either the model specification is accurate or propensity scores have been accurately estimated with good overlap, unbiased (and also efficient!) ATE estimates can be uncovered.

IV. Further Analyses on Heterogeneity

Partial Dependence

We then wanted to explore whether any covariates had direct relationships to the CATE values to see who might benefit the most from the program.

We created partial dependence plots of household size, total spend, and educational spend against the CATE values (of difference in log-scale income), shown in *Figure 5*. We hypothesized that total spend and educational spend (the middle and right most plot) would have some correlation with CATE, although there does not seem to be any major heterogeneity. In contrast, when we looked at household size (the left-most plot), we found a downward trend between CATE and household size, signaling that the larger the household, the less benefit the household received from the seasonal workers' program.

Policy Evaluation

Finally, in order to understand whether this policy would actually lead to any benefit for workers, we tested the improvement of the policies, as shown in *Table 11*. We see from the table above that the policy proposal yielded no significant improvement in the outcome for RSE workers, since the confidence interval contains zero. While the policy did not lead to significant improvements, this does not necessarily indicate that there were no benefits from the program for season workers.

References:

Mckenzie, David, and John Gibson. "The Development Impact of a Best Practice Seasonal Worker Policy." *Policy Research Working Papers*, 2010. doi:10.1596/1813-9450-5488.

Athey, Susan, and Stefan Wager and Nicolaj Nørgaard Mühlbach. "Exploring Causal Inference in Experimental and Observational Studies - Part 1" Apri, 2019.

Susan Athey, Stefan Wager, Vitor Hadad, Sylvia Klosin, Nicolaj Muhelbach, Xinkun Nie, Matt Schaelling. “Estimation of Heterogeneous Treatment Effects and Optimal Treatment Policies.” May, 2019.