CEE322: Data Analytics for Urban Systems

Lab 7: Confirmatory Data Analysis: AB Testing

Author: Davyd Tamrazov

Due date: Nov 09, 2020

Stanford University

1 Introduction

The following report aims to evaluate the impact that the residential water conservation program has had in the cities of Hoboken, NJ, and Weehawken, NJ, to help United Water in understanding the success of the deployed program. This is done by analyzing the household data collected by the United Water containing specific household descriptors and the indicator of participation in the program. In particular, AB testing is performed on the treatment / control groups corresponding to the households enrolled and not enrolled in the water conservation program respectively. The average treatment effect is calculated directly and with the linear regression models for each city, with all of the results validated and compared. Additionally, the effect of the program is compared between different groups of people.

2 Dataset

2.1 Dataset 1: Hoboken

The first dataset describes 200 households with their water usage, income, lawn size, and the number of household members. Additionally, indicators for owned properties as well as whether the household is enrolled in the water conservation program are specified in each entry. The summary of the variables in this dataset is provided in Table 1. Dataset appears to have an equal number of people in the program and not in the program, so no pre-processing was performed in order to avoid introducing any bias.

Water Usage (gal)		*Enrollment*		*Income ($1000)*		*Lawn Size (ft^2)*		*Household Members*		*Ownership*	
Min	175.8	No	100	Min	50.04	Min	73.37	Min	1.00	Rented	106
25%	269.6	Yes	100	25%	53.48	25%	168.7	25%	2.00	Owned	94
Median	311.2			Median	57.67	Median	197.6	Median	3.00		
Mean	319.3			Mean	61.27	Mean	203.7	Mean	3.04		
75%	357.8			75%	65.45	75%	233.2	75%	4.00		
Max	583.9			Max	112.6	Max	414.1	Max	6.00		

Table 1: *Summary of the variables in the Dataset 1*

2.2 Dataset 2: Weehawken

This dataset contains another set of 200 households in Weehawken, NJ, but described only with the total water consumption and number of household members as well as indicators of ownership and program enrollment. The summary of the variables is illustrated in Table 2. Similarly to the previous dataset, no pre-processing was performed in order to avoid introducing any bias.

Water Usage (gal)		*Program Enrollment*		*Household Members*		*Ownership*	
Min	177.4	No	100	Min	1.00	Rented	92
25%	280.2	Yes	100	25%	2.00	Owned	108
Median	320.9			Median	3.00		
Mean	321.5			Mean	3.12		
75%	360.3			75%	4.00		
Max	474.5			Max	6.00		

Table 2: *Summary of the variables in the Dataset 2*

3 Analysis

3.1 Evaluation for Hoboken, NJ

Initially, it is important validate randomization of the treatment and control groups used for the AB testing in order to eliminate selection bias. This can be done by analyzing distribution of continuous characteristics of each group, ensuring that the values of the feature in each group follow a similar distribution. This is visualized with the histograms and the Q-Q plots with the normal distribution in Figure 1.

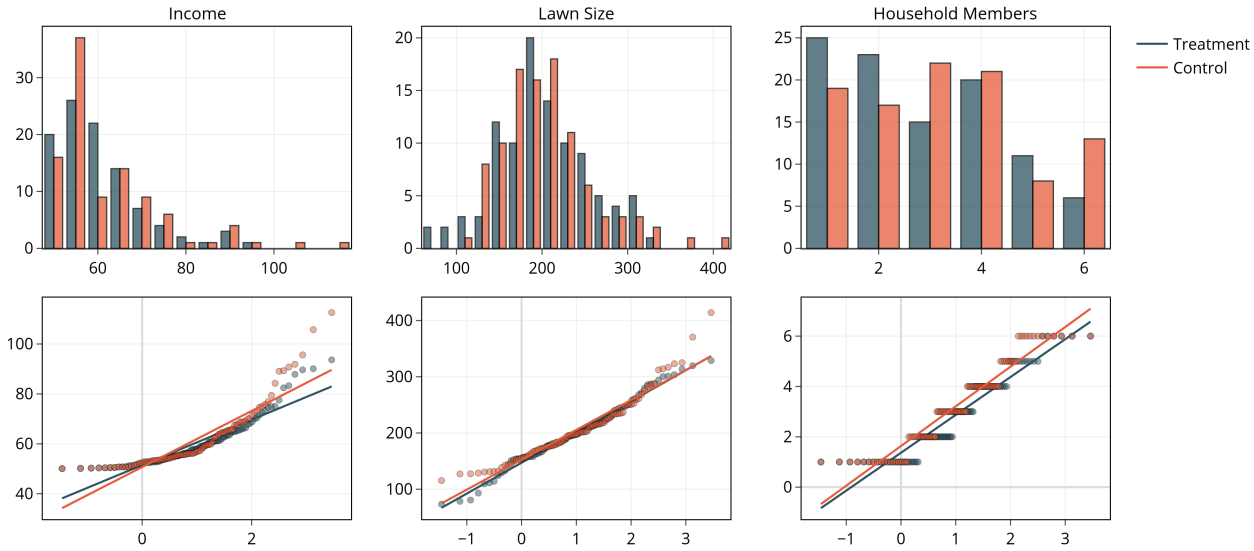

Figure 1: *Distributions of continuous variables in each group*

Here, it may be observed that distribution of income in treatment and control groups follows a very similar distribution (not normal distribution as shown on the Q-Q plot). Similarly, the number of

household members is a discrete value distribution, with both groups being having approximately the same scatter of household sizes. Conversely, lawn size does appear to follow some form of the normal distribution with the parameters of the distribution being very similar in each group. The categorical characteristics of the household that could not be graphically displayed, namely, ownership of the household have a slightly larger deviation in each group. As such, the treatment group has 53 owned household, while the control group has 41 owned household. Nevertheless, given a small size of each group, these deviations can be considered acceptable and the split of the treatment and control groups appropriate.

The AB testing can now be performed on the dataset. To start with, average treatment effect (ATE) can be calculated two ways: by directly comparing the expected water consumption in each group; and by comparing coefficients corresponding to the group indicator in a linear regression model. As such, ATE calculated directly results in:

$$\text{ATE} = \mathbb{E}\left(\text{Water Consumtion} \mid \text{Treatment}\right) - \mathbb{E}\left(\text{Water Consumption} \mid \text{Control}\right) = \mathbf{-48.16\ gal}$$

From the above results, it can be seen that based on the direct ATE, program appears to reduce average household water consumption by 48.2 gallons. However, such estimate does not account for the slight variations in the household characteristics that may affect water consumption too, as two groups may have different water consumption prior to the start of the program.

The impact of other characteristics can be accounted for by setting up a linear regression model, where the true effect of the program can be determined. The following hypothesis test is set up for each linear regression model:

H_0 : Program has no effect on the water consumption;

H_A : Program has a statistically significant effect on the water consuption

Significance level: 0.05 (two-tailed)

A number of linear regression models were generated using different sets of characteristics, each described and validated in Table 3 below. Note that ownership characteristic of the household appeared to have no effect on the water consumption and is, thus, excluded from all of the models.

	Model 1		*Model 2*		*Model 3*		*Model 4*	
	Estimate	P-value	Estimate	P-value	Estimate	P-value	Estimate	P-value
(Intercept)	89.92	0.000	48.41	0.010	77.92	0.000	35.89	0.003
Treatment Group	**-43.22**	0.000	**-37.23**	0.000	**-37.85**	0.000	**-36.99**	0.000
Income	1.714	0.000	3.907	0.000	–	–	1.305	0.000
Lawn Size	0.716	0.000	–	–	0.982	0.000	0.806	0.000
Household Members	–	–	16.48	0.000	19.80	0.000	19.01	0.000
Adjusted R-squared	0.668		0.632		0.822		0.849	
F-statistic	134.3		114.9		307.6		271.8	
Residuals: 25%	-27.68		-26.25		-19.33		-18.16	
Residuals: Median	-2.012		2.396		-1.086		0.624	
Residuals: 75%	28.92		26.18		21.33		19.79	

Table 3: *Summary of linear regression models using different sets of features*

From the above table, residuals in all models are well behaved with 25% and 75% quantiles of the distributions being similar, while centered around zero. Further, the null hypothesis can be rejected for each of the models, as the p-value of the treatment group index is statistically significant. Therefore, it can be stated with certainty that water conservation program has an effect on the water consumption of the households. It may further be observed that the magnitude of the estimate of the parameter corresponding to the treatment group index varies between models, suggesting that certain variables affect water consumption more. As such, it may be noticed that whenever the number of the household members is included in the model, its parameter estimate is positive and large, while the estimate of the treatment group index decreases. This implies that the number of household members has a significant effect on the water consumption, namely, as the number of people in the household goes up, there is a large increase in the water consumption.

Additionally, it is clear that the fit of the model considerably improves when both lawn size and the number of household members are included, underlining the importance of both variables. Similarly to the number of the household members, lawn size appears to increase water consumption as all three are statistically significant. However, the best fitting model results when income is also added to the model, meaning that all three variables influence water consumption. Estimate of the treatment group index in the best fitting model is -36.99, suggesting that the average household water consumption differs by around 37 gallons between control and treatment groups. This reduction is considerably lower that the ETA estimate directly from the expected values. The explanation for such difference arises from the previously made observations that multiple other variables are responsible for the difference in water consumption. In particular, as shown on Figure 1, there is a noticeable deviation in the household members distribution between groups and, given that this characteristic has the largest effect on the water consumption, it accounts for a sizeable portion of the difference in consumption between groups. Thus, the ETA estimate obtained with the regression identifies the true effect of the program on the households' water consumption.

Using the above conclusion, a linear model is fit to certain groups of people as outlined in the Table 4 below. Namely, high and low income are defined as those who earn more than $70k and less than $70k respectively. Additionally, another groups is defined by filtering the number of household members to 3 or more.

	High Income		*Low Income*		*3+ Members*		*1-2 Members*	
	Estimate	P-value	Estimate	P-value	Estimate	P-value	Estimate	P-value
(Intercept)	2.613	0.950	47.99	0.051	19.39	0.247	65.28	0.006
Treatment Group	**-34.52**	0.002	**-37.06**	0.000	**-42.74**	0.000	**-29.16**	0.000
Income	1.700	0.017	1.121	0.011	1.432	0.000	1.090	0.010
Lawn Size	0.789	0.000	0.803	0.000	0.811	0.000	0.780	0.000
Household Members	20.29	0.000	18.67	0.000	21.33	0.000	8.921	0.187
Adjusted R-squared	0.876		0.744		0.864		0.747	
F-statistic	57.25		121.4		182.9		62.23	
Residuals: 25%	-21.35		-17.14		-18.11		-16.02	
Residuals: Median	1.607		0.511		3.590		-1.305	
Residuals: 75%	18.60		19.10		17.09		21.65	

Table 4: *Summary of the linear regression models using different groups of households*

Similarly to the previous analysis, residuals are generally well behaved, thereby validating the models and allowing to make conclusions from the estimates. As such, the null hypothesis can be rejected for each model with the treatment group parameter estimate being statistically significant. From these results, it is evident that the effect of the water conservation program on the high income group is marginally lower than that on the low income group, by approximately 2 gallons. Thus, it can be concluded that difference in income does not have a significant effect on the program success. On the other hand, effect on the households with 3 or more members is significantly larger than on the smaller households with the difference in the consumption reduction being around 13.5 gallons. This can be justified by the fact that larger households generally use more water as can be identified by the positive coefficient for household members characteristic in each model. As a result, the overall reduction in the consumption due to the program is proportionally larger.

3.2 Evaluation for Weehawken, NJ

In this section, a similar analysis is performed on the data collected from Weehawken, NJ. As before, it is necessary to validate the treatment and control groups by considering the distribution of the continuous characteristics in each group, as shown in Figure 2. However, note that this dataset has much fewer features, namely, it includes only the number of household members and the program enrollment and ownership indicators.

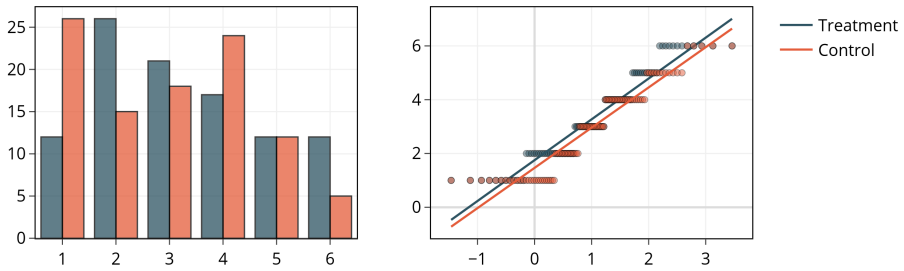

Figure 2: *Distribution of the number of household members in each group*

As the figure above illustrates, the number of people in the household is generally similar in each group, apart form a significantly large number of single-member households in the control group. Conversely, this is counterbalanced by the larger number of households with 2 people in the treatment group. Thus, since both household sizes can qualify as small households and given a small dataset size, such margin may be deemed as acceptable for the AB testing. Additionally, the number of owned households is 55 and 53 for treatment and control groups respectively, what is also consistent with the randomization required for the analysis.

Calculating the ATE directly using the water usage outputs results in +3.48 gallons. This suggests that deploying the program appears to marginally increase the average household consumption by 3.48 gallons – an opposite to the desired effect of the program. This, however, is not consistent with the results obtained in the previous analysis for the city of Hoboken and shows that there is very limited effect of the program on the water consumption in Weehawken.

Next, the ETA can be determined using linear regression. Similarly to the previous analysis, it was found that ownership status does not affect water consumption in any way, so it is excluded from

the models. Thus, this leaves only one possible model to be fit to the data shown in Table 5.

	Estimate	P-value	Adjusted R-squared	F-statistic	25%	Median	75%
(Intercept)	272.2	0.000					
Treatment Group	-1.509	0.844	0.1691	21.24	-39.53	-0.632	34.67
Household Members	16.10	0.000					

Table 5: *Summary of the linear regression model*

From the residuals shown in the above table, it is evident that the model is valid, although it does not fit the data well. Looking at the p-values of the model features confirms that the effect of the treatment group parameter is statistically insignificant (p-values is > 0.05). This suggests that the null hypothesis cannot be rejected in this case and the water conservation program appears to have no statistically significant effect on the water consumption, corroborating conclusions made from the direct ETA calculations.

The primary difference between the datasets acquired for the cities of Hoboken and Weehawken is that the latter only has two features characterizing the household. On the other hand, the randomization on the available features appears to be performed well in both cities. As was previously shown, some household characteristics have an effect on the magnitude of the difference in water consumption between treatment and control groups (see Table 4). Thus, it could be hypothesized that some unobserved features in both cities have a significant effect on the water consumption and, thus, significantly differentiate these two places. Since the testing was replicated for Weehawken city based on the features described in the Hoboken households, such unobserved characteristics could have a drastic effect on the outcome. Further, another hypothesis could be that tests were performed at different times of the year. For example, the test could have been performed in the summertime or other time when the households have a higher requirements for the water usage, such as watering the lawn. Thus, during that time, it is not feasible for the households to significantly reduce their water usage regardless of the program. Finally, it could be that the test in Weehawken wasn't performed for long enough, thereby limiting the extent to which the program can have an effect on the water consumption.

4 Conclusion

In conclusion, households from the cities of Hoboken and Weehawken were analyzed to determine the average treatment effect (ETA) of the water conservation program. ETA was calculated directly as well as using coefficients resulting from the linear regression models. It was found that in the city of Hoboken, water conservation program has had a sizeable effect, reducing the average water consumption by 37 gallons. Additionally, the number of household members was identified as another important feature accounting for a significant portion of the water consumption. As such, it was determined that the effect of this program is considerably larger in households with 3+ members compared to the smaller households (43 vs 29 gallons reduction). On the other hand, water conservation program in the city of Weehawken appeared to have no statistically significant effect. Several hypotheses were presented that could explain why the results in this city differ.