Lab 7 : Confirmatory Data Analysis: AB Testing

Chun-Liang Wu

Introduction

The report aims to understand the performance of a residential water conservation program in two cities, Hoboken, NJ, and Weehawken, NJ. The report will use AB testing to check the effect of the conservation program on two cities. The dataset in each city has the treatment and control groups. The report will calculate the average treatment effect (ATE) by two methods, the average and the regression model. Furthermore, the report will also divide the people in the dataset into different group (e.g. low/high income, low/large member in house, and small\large lawn) to further understand the effect of the conservation program on different groups of people. In the dataset of Weehawken, the report will also recommend possible methods to solve the problem of the minor effect of the conservation program.

Dataset

There are two datasets, Hoboken and Research (i.e. Weehawken). The Hoboken dataset consists of 200 samples and 6 features. The features include water consumption, income, number of household members, lawn size, owned or rented household and whether enrolled in the program. The Researcher dataset consists of 200 samples and 4 features. The features include water consumption, number of household members, owned or rented household and whether enrolled in the program. The Researcher doesn’t have income and lawn size features.

In addition, there are no nans in two datasets. The report won’t do any further preprocessing steps.

Analysis

The report is divided into 2 parts, Hoboken and Research. For Hoboken, the report will explore effects of the different combinations of features on the water conservation, including low/high income, low/large member in house, and small/large lawn. For Research, the report will discuss the reasons for no water consumption reduction from deploying the program.

- Hoboken

The water consumption of the average of treatment group is 295.2. The water consumption of the average of control group is 343.4. The ATE via the difference of two averages is -48.2.

For the ATE via regression model, the report utilizes AIC selection to eliminate the uncorrelated variables to best explain the TE. The AIC selection eliminates Owned feature. Table 1 shows the result of the regression model. The ATE is -37.0 for the regression model. And Members feature show a very strong positive correlation with the water consumption, indicating that as the number of the members in each household increases, the water consumption will increase.

Table 1. Table of regression between water consumption vs four variables

Variable	Coefficient	P value	Adjusted R-squared	F-statistic
Income	1.3	0.00	0.85	271.8
Lawn	0.8	0.00		
Members	19.0	0.00		
ATE	-37.0	0.00		

The difference of ATE between two methods (average and regression model) is 11.2. The difference shows that the inclusion of more variables in the ATE analysis can help explain the difference of the water consumption in treatment and control groups. In this case, the difference of Income, Lawn and Members variables in treatment and control groups accounts for 11.2 difference in water consumption. Therefore, the effect of the program will be a reduction of 37 in water consumption.

- Income

The report divides the Hoboken dataset into low income (income < 61.3) and high income (income > 61.3) group. 61.3 is the mean of the income. The low-income group has 127 samples. The high-income group has 73 samples.

Table 2 shows the ATE of the two groups via the average method. The ATE of low-come group is 37.1. The ATE of high-income group is 62.4. The ATE for high-income group is higher than the ATE of low-income group in term of the average method. The difference of ATE is 25.3. In addition, the report found out that the lawn size of high-income group is larger than the lawn size of low-income group by 50. The number can explain why average of the water consumption of the high-income group is so high.

Table 2. Table of ATE of water consumption between low-income and high-income

	Low Income	High Income
With program	278.2	326.7
Without program	315.3	389.1
ATE	37.1	62.4

For the ATE via regression model, Table 3 shows the result of the regression model. The ATE of the regression model is -35.8 for low-income group, and -38.9 for high-income group. The difference between two ATE is about 3.1.

Table 3. Table of regression for the two groups between water consumption vs four variables

	Low Income		High Income	
Variable	Coefficient	P value	Coefficient	P value
Income	1.5	0.07	1.9	0.00
Lawn	0.9	0.00	0.7	0.00
Members	18.3	0.00	20.3	0.00
ATE	-35.8	0.00	-38.9	0.00

The difference between two difference of ATE is about 22.2 (average: 25.3, regression: 3.1). The ATE via average method in high-income group is higher than ATE via regression in high-income group. The reason is due to the influence of the other variables. As previously mentioned, the advantage of the regression model is that the other variables can help explain the difference of water consumption. Table 4 shows how difference of mean of the three variables (Income, Lawn and Members) influence the water consumption in high-income group. The table shows that the difference of the three variables contribute to 23.2 of reduction of water consumption in high-income group. Therefore, by excluding the effect of difference in three variables, the ETA of the program for high-income and low-income groups is actually close (-35.8 vs -38.9), showing the program poses the same influence for people with high or low income.

Table 4. Table of difference of mean of three variables in high-income group

Variable	With program	Without program	Difference (without program – with program)	Coefficient from Table 3	Reduction of water consumption (difference * coefficient)
Income	70.9	74.1	3.2	1.9	6.1
Lawn	228.9	241.8	12.9	0.7	9.0
Members	2.88	3.28	0.4	20.3	8.1
Total reduction by these three variables					23.2

Since the difference of two methods is not the focus of the report, and the report has already explained why the ATE between two methods are different, therefore, in the future sections, the report will not dive into the reason behind it, rather the research will only focus on the better result of the regression model.

- Lawn

The report divides the Hoboken dataset into small lawn size (lawn size < 203.7) and large lawn size (lawn size > 203.7) group. 203.7 is the mean of the lawn size. The small-lawn group has 112 samples. The large-lawn group has 88 samples.

For the ATE via regression model, Table 5 shows the result of the regression model. The ATE of the regression model is -36.1 for small-lawn group, and -37.8 for large-lawn group. The difference between two ATE is about 3.1. In the section, the difference between two groups is also small, indicating the program poses the same influence for people with large or small lawn size.

Table 5. Table of regression for the two groups between water consumption vs four variables

	Small Lawn		Large Lawn	
Variable	Coefficient	P value	Coefficient	P value
Income	1.6	0.00	1.2	0.00
Lawn	0.8	0.00	0.8	0.00
Members	18.5	0.00	19.7	0.00
ATE	-36.1	0.00	-37.8	0.00

- Member

The report divides the Hoboken dataset into small member size (member size <= 3) and large member size (member size > 3) group. 3 is the mean of the member size. The small-member group has 121 samples. The large-member group has 79 samples.

For the ATE via regression model, Table 6 shows the result of the regression model. The ATE of the regression model is -32.6 for small-member group, and -43.2 for large-member group. The difference between two ATE is about 10.6. In the section, the difference between two groups is relatively larger than the previous two sections. It indicates that the reduction of water consumption in the family with more members is more significant than the reduction of water consumption in the family with less members. In addition, the coefficient of Members in large-member family is larger than the coefficient of Members in small-member family, showing that the number of family members in large-member family will have a larger influence on the water consumption.

Table 6. Table of regression for the two groups between water consumption vs four variables

	Small Member		Large Member	
Variable	Coefficient	P value	Coefficient	P value
Income	1.5	0.00	1.0	0.00
Lawn	0.7	0.00	0.9	0.00
Members	16.4	0.00	24.6	0.00
ATE	-32.6	0.00	-43.2	0.00

- Comparison

Fig 1 shows the ATE comparison of the variables. The large group of three variables all save more water consumption than the small group of three variables. Especially for Members, the saving in water consumption of the large group of Members is more significant than the saving in the other two large groups. In addition, the small group of Member also save the less water consumption compared to the other two small groups. The report concludes that the attribute of Member is the most sensitive in term of the water consumption saving with the deployment of the program, meaning the program can pose a significant effect on the family of large member size (people > 3) or low member size (people <= 3).

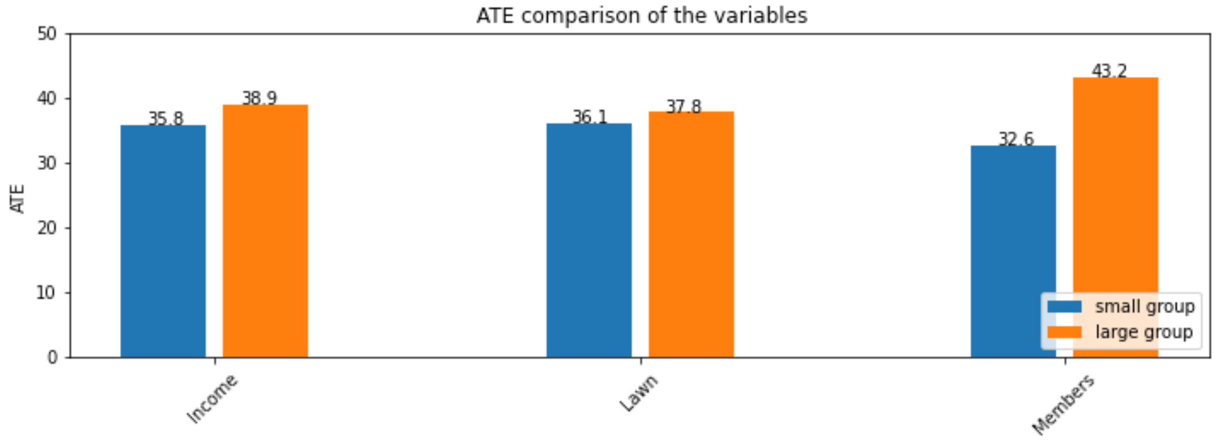

Fig 1. ATE comparison of the variables

- Research

The water consumption of the average of treatment group is 323.3. The water consumption of the average of control group is 319.8. The ATE via the difference of two averages is -3.5.

For the ATE via regression model, Table 7 shows the result of the regression model. The ATE is -1.6 for the regression model, but the p value is very high, 0.83, indicating very week relationship with the water consumption. Owned feature also shows a weak relationship with water consumption. In addition, adjusted r-squared is also small.

Table 7. Table of regression between water consumption vs four variables

Variable	Coefficient	P value	Adjusted R-squared	F-statistic
Members	16.1	0.00	0.17	14.3
Owned	5.0	0.51		
ATE	-1.6	0.83		

All the indicators show that the number of the variables in the dataset are too few to capture the complexities of the water consumption (Researcher dataset doesn't have income and lawn size features). For example, maybe the household with the deployment of the program accidentally all have the larger lawn size than the household without the deployment of the program. In this case, since the dataset doesn't include Lawn variable, the regression can't separate the larger lawn's effect out of the ATE, thus causing the weak significance of ATE with the deployment of the program. To solve this problem, the consultants of Research dataset need to enlarge the dataset via collecting more variables per sample.

Conclusion

The report aims to understand the performance of a residential water conservation program in two cities, Hoboken, NJ, and Weehawken, NJ by conducting AB testing on the two datasets, Hoboken and Research.

- Hoboken

The report first conducts AB testing on the whole Hoboken. The ATE of the program via the regression model is about -37. In next analysis, the report further separates out different groups in the dataset, low/high income, low/large member in house, and small\large lawn size, and conducts regression AB testing on each of the group. The report found out the larger group of all three variables all save more water consumption than the small group of three variables. Especially for Members variable, the water conservation of large Member saves the most consumption in all the groups, shown as Fig 1. In addition, the water conservation of small Member saves the least consumption in all the groups. Therefore, the report concludes that the attribute of Member is the most sensitive in term of the water consumption saving with the deployment of the program.

- Research

The ATE of the program in the Research is very small in both average and regression method. The report points out that it is due to the number of the variables in the dataset are too few to capture the complexities of the water consumption. The regression model can't separate out the ATE of the program if the number of the variables is not enough to explain the different attribute of each sample. Therefore, the report suggests that the Research dataset need to be expanded via considering more variables per sample.