Causal and Predictive Analytics – Homework 1

Individual Assignment

Individual Assignment

This assignment involves working with real-world experimental data from a landmark study on the effectiveness of online search advertising. eBay conducted this experiment to assess the return on its ads across search engines like Google, Bing, and Yahoo, focusing specifically on search queries containing the term "E-Bay."

Your performance will be evaluated as if you are a job candidate, based on the strength of your arguments, the use of data to support your conclusions, and the clarity of your writing. Do not share, copy, or discuss your written answers with other students.

Submission Checklist

- If you choose to use Generative AI, please use a single session of ChatGPT, and get a link to the chat through the "Share" button. Submit this on blackboard while submitting your assignment.
- Save a copy of your answers in case the browser you are working in closes.
- Submit your answers to the questions through blackboard

Data Dictionary

- **date**: Date of advertising
- **DMA**: Designated Market Area Code (Basically a city)
- **isTreatmentPeriod**, **isTreatmentGroup**: Dummy variables denoting whether the date belonged to the treatment period, and if the DMA belonged to the treatment group
- **revenue**: Revenue for the DMA in dollars

Analysis

The study was conducted as follows. Users were categorized by their designated market area (DMA), which is given as a categorical variable in Column 1 of the dataset. DMAs were randomly selected to be in the treatment or the control group. The variable `isTreatmentGroup` indicates that the DMA was placed in the treatment group. After a certain date, the treatment period started, and the treatment group was no longer shown search ads from eBay. The variable `isTreatmentPeriod` indicates whether the treatment period had started.

This analysis follows part of the study. The analysis uses Boolean variables, and the `read.csv`, `lm`, `summary`, `log`, `subset`, `as.Date`, and `sort` functions. As a reference, you can consult the 'Interview Case' presented in class.

```
#Set the directory
setwd("C:/Dropbox/Teaching Lectures/Assignments/Homework 1")

homeworkDB = read.csv('Homework 1 Data - 436.csv')
```

The data contains a control group, which was shown search ads throughout the data, and a treatment group, which was only shown search ads before the treatment period. We firstg run a regression that compares log(revenue) of the treatment group in the pre-treatment period and the treatment period.

```
treatOnly = subset(homeworkDB,homeworkDB$isTreatmentGroup==1)
summary(lm(log(revenue)~isTreatmentPeriod,data=treatOnly))
```

```
##
## Call:
## lm(formula = log(revenue) ~ isTreatmentPeriod, data = treatOnly)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0038 -0.7490 -0.0274  0.6929  3.8268
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       10.94865    0.01472 743.988   <2e-16 ***
## isTreatmentPeriod -0.03940    0.01987  -1.983   0.0474 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.252 on 16044 degrees of freedom
## Multiple R-squared:  0.0002451,  Adjusted R-squared:  0.0001828
## F-statistic: 3.933 on 1 and 16044 DF,  p-value: 0.04737
```

Now we will use the control group for a truly experimental approach. First, we will run a randomization check:

```
preTreatment = subset(homeworkDB,homeworkDB$isTreatmentPeriod==0)
summary(lm(log(revenue)~isTreatmentGroup,data=preTreatment))
```

```
##
## Call:
## lm(formula = log(revenue) ~ isTreatmentGroup, data = preTreatment)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9962 -0.7502 -0.0285  0.7331  3.8229
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       10.96273    0.02037 538.128   <2e-16 ***
## isTreatmentGroup -0.01408    0.02477  -0.568     0.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 10708 degrees of freedom
## Multiple R-squared:  3.017e-05,  Adjusted R-squared:  -6.322e-05
## F-statistic: 0.323 on 1 and 10708 DF,  p-value: 0.5698
```

Now, using the treatment period data, we determine the effectiveness of eBay ads. We run a regression with log(revenue) as the dependent variable, and control for whether the DMA is in the treatment group.

```
postTreatment = subset(homeworkDB,homeworkDB$isTreatmentPeriod==1)
summary(lm(log(revenue)~isTreatmentGroup,data=postTreatment))
```

```
##
## Call:
## lm(formula = log(revenue) ~ isTreatmentGroup, data = postTreatment)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0038 -0.7546 -0.0288  0.7419  3.8268
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      10.916740   0.018610 586.595   <2e-16 ***
## isTreatmentGroup -0.007494   0.022632  -0.331    0.741
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.208 on 13018 degrees of freedom
## Multiple R-squared:  8.422e-06,  Adjusted R-squared:  -6.839e-05
## F-statistic: 0.1096 on 1 and 13018 DF,  p-value: 0.7406
```

Discussion (36 marks)

Please provide written answers to each of the following questions in the blackboard link. Answers will be judged on accuracy and correct spelling/grammar. Pay close attention to what each question is asking for and the course materials. Each answer only requires a short response (Maximum of 3 sentences and 45 words). Additional words and sentences will be deleted. Please use a spelling/grammar check before you submit.

1. In Analysis A, the 'Intercept' is roughly 10.95. In managerial terms, what does this mean? (4 marks)
2. In Analysis A, the standard error of the intercept is approximately 0.014. Calculate the confidence interval provide a managerial interpretation it. (4 marks)
3. In Analysis A, the model was run without a control group. What do the resulting coefficients and associated confidence intervals say about the effectiveness of advertising? (4 marks)
4. What is the purpose of the randomization check in Analysis B? What do the results of this analysis show? (4 marks)
5. In Analysis C, the model was run with a control group. What do the resulting coefficients and associated confidence intervals say about the effectiveness of advertising? (4 marks)
6. What does the control group allow us to control for? What specific omitted variables might have caused bias in Analysis A, but wouldn't in Analysis C? (4 marks)
7. State the R-Squared value from the regression in Analysis C. Does this influence the interpretation or confidence in the estimated effectiveness of advertising? (4 marks)
8. What are some potential limitations of this experimental design, particularly with regard to external validity? How might these limitations affect the generalizability of the results to other advertising campaigns or contexts? (4 marks)
9. Based on the results from all analyses, what recommendations would you make to eBay's management regarding their online advertising strategy? Justify your recommendations based on the data and results. (4 marks)