Griffin Applied Economics Incubator 2021
The Causality Challenge

1 Context

The aim of this challenge is to examine how the ranking of a product on a webpage might affect the demand for that product. We focus on an online retail platform that puts in contact sellers and buyers.

To provide more context, we briefly describe the consumer search process here. The consumer first begins her search query on the website by specifying product attributes. In response to her query, she then sees all the products that match her request. The information recorded on the webpage includes the seller ID, the product's position in the ranking, and some of its characteristics (price, associated seller attributes, and a promotion indicator). After observing the list of products, the consumer can click on a particular one to observe more information. We measure demand for a product through the clicks for that product. Note that multiple searches made by the same consumer are not linked.

2 Goals of analysis

Your goal in the challenge is twofold:

As a preliminary goal, you are asked to predict the binary 0/1 click variable, given the set of attributes that are available. This part of the challenge is a classical prediction problem, which we will judge based on a hold-out sample.

The main goal of the challenge is to estimate the *causal effect* of a product's position on the webpage on the clicks for that product. That is, you are asked to estimate the effect on clicks of an exogenous change in a product's position.

3 Data description

The variables are given at the search impression level, with a total of $94,038$ unique search IDs. Each search ID contains up to a maximum of 33 product listings, with a total of $2,522,288$ observations.

The variables in the data set consist of three consumer attributes (`C`), six seller attributes (`S`), and seven product attributes (`P`), along with the time of the search (`date_time`), the search identifier (`searchID`), the identifier of the seller (`sellerID`), an indicator that the webpage runs a promotion for the product (`promo`), the price of the product (`price`), the position of the product on the webpage (`position`), and an indicator that the consumer selects the product (`click`). For a given

search (i.e., webpage) we observe a set of products with different seller characteristics. A list of the variables is given in Table 1.

Table 1: Variables for the Causality Challenge

`date_time`	Date and time of the search
`searchID`	ID of the search
`sellerID`	ID of seller
`promo`	Indicator that the webpage runs a promotion for the product
`price`	Price of the product
`position`	Position of the product
`click`	Click indicator
`C1`	Consumer country
`C2`	Avg rating of consumer's past purchases
`C3`	Avg price of consumer's past purchases
`S1`	Seller country
`S2`	Seller rating 1
`S3`	Seller review score given by consumers
`S4`	Indicator that the seller belongs to a brand
`S5`	Seller rating 2
`S6`	Avg previous price posted by seller
`P1 - P7`	Product attributes

4 What you need to do

1. First step: predict `click`. You may use any of the available data. We have divided the original data into two random subsamples. In the first one, `train.csv`, you have access to all the variables. In the second one, `test.csv`, you have access to all the variables except `click`, which you need to predict. We will declare the winner based on prediction performance in terms of **misclassification rate** on that test subset; that is, the fraction of observations for which your prediction differs from the 0/1 click observation (unavailable to you) in the `test` sample.

2. Second step: estimate a set of causal effects of `position` on `click`. We will assess the plausibility of your proposed estimates by comparing them with estimates based on experimental variation that we have withheld. Specifically, we will compare your estimates to the average treatment effects corresponding to changes in position estimated on the withheld experimental sample (unavailable to you). We will give equal weight to each of the 32 effects that you will report, corresponding to 32 possible changes in position (see below). In addition, we will take into account the credibility of your research approach, so you will need to explain and motivate your approach.

5 Submission guidelines

1. Answers to the first step (Prediction): Your numerical answer should be the predicted values of `click` (i.e. 0's and 1's, **do not report predicted probabilities or values other than**

0 **and** 1) for the observations in the `test` subset. Please **do not modify the ordering** of the observations. Also, you will need to provide us with any code that you used.

2. Answers to the second step (Causality): For each position from 1 through 32, you must provide an estimate of **the change in the probability of click as a result of exogenously increasing the position by one**. For example, your first entry should be a number that indicates the change in the click probability from an exogenous change in position from 1 to 2, i.e. from being placed first on the web page to being placed in second position. Your second entry should correspond to changing the position from 2 to 3, and so on. Note that we are asking for **point estimates** only, not confidence intervals. In addition, you will need to provide us with your code and a write-up justifying your approach.

3. Your submission files should include the following:

   (a) a *.csv* or *.txt* file containing your predicted values of `click`;

   (b) a script (*.do*, *.R*, or using any language of your choice) containing your code for both Prediction and Causality;

   (c) a pdf file containing your estimates and write-up for Causality.