Automated Instrumental Variables

Supervised by professor Russel Steele

Mariel Yacolca Maguiña

Summer 2023

I. Introduction

The research project titled Automated Instrumental Variables took place during the summer of 2023 under the supervision of Professor Russel Steele. The goal of the project was to replicate the results from Yuan et al. (2022). In their paper, Yuan et al. put implement an algorithm that automatically produces instrumental variables from the decomposition of other observed variables.

Part II. of this report provides the theoretical background necessary to understand what is an instrumental variable and why we need it. Part III. explains the methodology of the author to construct the instrumental variables as well as the methods used to test their effectiveness. Part IV. explains the process we followed to attempt to replicate the results as well as the resulting measures of the effectiveness of the instrumental variables. Finally, Part V., provides a discussion of the results and possible caveats.

II. Literature Review

In this section, we will introduce the background information necessary to understand causal inference and why auto-IV.

Causal inference is the process of determining the existence and extent of the causal effect of treatment A on outcome Y. Consider a situation in which we are investigating whether there exists a causal effect of a new medication on reducing mortality. In this case, the medication is the treatment variable A, and mortality or not is the outcome Y. We will continue expanding on this example as we introduce the nuances of causal inference.

For individual i, there exists a causal effect of treatment A on outcome Y if and only if the outcome changes as the value of treatment changes. That is, for a binary treatment, $Y^{a=1} \neq Y^{a=0}$. In our example, this means that the effect of taking the medication is different from the effect of not taking it. Note that variables $Y^{a=1}$ and $Y^{a=0}$ are referred to as "counterfactual outcomes" since they represent situations that may not actually occur.

The definition of causal effect can be expanded from an individual to an entire population. There exists an average causal effect of treatment A on outcome Y if $E[Y^{a=1}] \neq E[Y^{a=0}]$. That is, the expected effect of taking the medication is different from the expected effect of not taking it. Note that throughout the rest of this report, we will refer to the average causal effect as just the causal effect.

In the real world, we cannot test the effects of the treatment easily because we cannot observe the outcomes of all possible treatments for the population. Instead, we can observe the actual outcome of the treatment that each individual actually received. In mathematical notation, we can observe $E[Y|A = 1]$ and $E[Y|A = 0]$ but only one of $Y^{a=1}$ and $Y^{a=0}$. Since we are unable to observe the counterfactual outcomes, strictly speaking, we can only draw conclusions about association but not causation.

In order to overcome the limitation described above, researchers consider marginal randomized experiments. In a randomized experiment, the sample that receives the treatment is chosen at random. As a result, we can assume marginal exchangeability. That is, under the same treatment $A = a$, the probability of outcome $Y = y$ is the same for both groups. This result also holds for the entire population.

Now, consider a situation in which we have two kinds of patients, those who are slightly sick and those in critical condition. This factor, let's call it L can affect the effectiveness of treatment A. In order to solve this, a researcher conducts a conditionally randomized experiment. That is, the researcher separates the patients based on whether they are slightly sick $L = 0$ or they are in critical condition $L = 1$. Then, a marginal randomized experiment is conducted for each group. As a result, given $L = l$, the results satisfy exchangeability.

Unfortunately, randomized experiments are not typically feasible in real life. Consider an observational study where the treatment A has worrying side effects so doctors only prescribe it to patients in critical

condition. In this scenario, we would observe an association between taking the medication and mortality. Note however that an association does not imply a causal effect.

In order to illustrate relationships between variables, we can use causal diagrams. The notation is simple: an arrow indicates a causal relation and a box around a variable indicates conditioning. Consider the following diagrams:

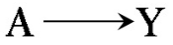

Figure 1: Simple causal relationship.

In Figure 1, we observe the simple causal diagram. This arrow indicates that medication A has a causal effect on mortality Y.

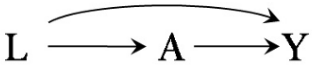

Figure 2: Presence of confounding variable.

In Figure 2, we observe that medication has a causal effect on mortality, but also that the state of the patient (critical or slightly sick) affects their risk of mortality and their chances of receiving the medication.

Variable L is called a confounder. The presence of this variable eliminates the possibility of marginal exchangeability, but conditional (on L) exchangeability is still possible. Consequently, in the presence of a confounder, information about L is required for the process of identification. Note that identification refers to the process of identifying the causal effect.

Now, to address the issue of confounding, researchers have developed several methods, one being Instrumental Variables (IV). An Instrumental Variable Z is such that it satisfies the three following conditions:

(i) Relevance: Z is associated with A

(ii) Exclusion: Z does not affect Y except through its effect on A

(iii) Unconfounded Condition: Z and Y do not share causes

In order to better illustrate how an IV works, let's go back to the example. Imagine we are trying to investigate the effects of certain medication A on mortality Y, but we notice a confounding variable, whether the patient is in critical condition or they are slightly sick. In order to address this situation, the researchers use the instrumental variable availability of a pharmacy near the patient's house. This satisfies the conditions of an instrumental variable because

(i) The availability of a pharmacy is associated with whether the patient takes the medicine.

(ii) The availability of a pharmacy does not affect mortality, except through its effect on whether the patient takes the medicine.

(iii) The availability of a pharmacy and mortality do not share causes.

The basic idea of the IV is that this variable contains information about the treatment that is free of influence from confounders. Therefore, we can approximate the effect of the confounderless treatment on the outcome. In Figure 3, U represents all confounders and Z is the instrumental variable that satisfies the three conditions outlined above.

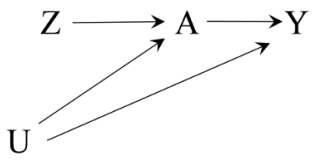

Figure 3: Instrumental Variable.

In order to use the IV method in regression analysis, we use 2 Stage Least Squares (2SLS). The procedure is the following:

1. Fit the treatment model $A_i = \alpha + \beta_0 Z_i + \beta_1 U_1 + ... + \beta_k U_k \ + \varepsilon_i$.
2. Calculate the fitted values $\hat{A}_i$.
3. Fit the outcome model $Y_i = \alpha + \beta_0 \hat{A}_i + \beta_1 U_1 + ... + \beta_k U_k \ + \varepsilon_i$

The issue with the 2SLS method and other methods that follow a similar approach is that they require a valid IV. However, because the exclusion assumption is not empirically verifiable, it becomes very difficult to distinguish between IVs and confounders.

The goal of the paper "Auto IV: Counterfactual Prediction via Automatic Instrumental Variable Decomposition" by Yuan et al. is to address the difficulty of finding a valid IV. They do so by implementing the auto-IV algorithm. In the next section, I will expand on their methodology.

III. Methodology

Yuan et al. in "Auto IV: Counterfactual Prediction via Automatic Instrumental Variable Decomposition" develop a data-driven approach to "automatically" obtain valid IVs.

Assume the following relationship between the outcome and the treatment:

$$Y = g(X) + e \tag{1}$$

where e is the error term such that $E[e] = 0$ and $Var[e] \in \mathbb{R}$. This term contains unobserved confounders which affect both X and Y. Therefore, e is correlated with X and $E[e|X] \neq E[X] = 0$. Therefore,

$$E[Y|X] = E[g(X)|X] + E[e|X] \tag{2}$$
$$\Rightarrow E[Y|X] \neq E[g(X)|X] \tag{3}$$
$$\Rightarrow E[Y|X] \neq g(X) \tag{4}$$

As a result, we cannot estimate the causal relationship $g(X)$ between X and Y by directly estimating $E[X|Y]$ due to the confounding. To solve this issue, we introduce IVs.

The data that is available to researchers is the treatment X, the outcome Y, and other observed variables V. From V, researchers aim to find representations of IVs Z and confounders C based on their relationship with X and Y. To model the representations, V is input into representation networks that will find representations ϕ^C and ϕ^Z for C and Z respectively.

Then, researchers ensure that the assumptions of relevance and exclusion of the IV are satisfied by maximizing the mutual information between ϕ^Z and X, and minimizing the mutual information between ϕ^Z and Y. To find the representation ϕ^C of C, the mutual information between C and X and Y is maximized.

After finding the IV representations, researchers test their prediction accuracy by using IV-based methods. The four experiments that were conducted in this research project were:

1. Direct Neural Network: Regress Y onto X without
2. 2SLS: Two-stage least squares with linear models.
3. 2SLS (poly): Two-stage polynomial least squares with ridge.
4. 2SLS (NN): Two-stage least squares with neural networks structure.

The data-generating process in this experiment is the following:

$$Y = g(X) + e + \sigma \tag{5}$$
$$X = Z_1 + e + \gamma \tag{6}$$
$$Z \; Unif([-3, 3]^2) \tag{7}$$
$$V = [Z, \sigma, \gamma] \tag{8}$$
$$e \; N(0, 1) \tag{9}$$
$$\gamma, \sigma \; N(0, 0, 1) \tag{10}$$

where Z are the true valid IVs, σ, γ are noise, variables V are the IV candidates, e is the unobserved error, and g is the true response function which is chosen as follows:

$$g(X) = \begin{cases} 0 & \text{if } X \geq 0, \\ 1 & \text{if } X < 0. \end{cases} \tag{11}$$
$$g(X) = -X \tag{12}$$
$$g(X) = -0.1X^2 - 0.4X \tag{13}$$
$$g(X) = 0.05X^3 + 0.1X^2 - 0.8X \tag{14}$$
$$g(X) = |X| \tag{15}$$

In the following section, we will provide further details about the experiments that were conducted using the IV-representations for each of the experiments and response functions.

IV. Simulations

Click here to visit the AutoIV GitHub repository

The goal of this project is to replicate the results of the paper. The first step was to attempt to run the code that the authors of the paper shared on GitHub. However, due to an update to the package tensorflow, the code was not longer able to run. To solve this issue, we attempted to translate the code into the newer version and install a virtual computer. Nevertheless, we ended up manually updating the code by researching the newer versions of the required functions that were causing errors.

In the following list I will include a brief description of all the changes that were made to the code:

1. The function tf.set_random_seed(seed) was changed to tf.random.set_seed(seed).
2. The function tf.random_shuffle() was changed to tf.random.shuffle().
3. In TensorFlow 1, we could directly run self.x = tf.compat.v1.placeholder(tf.float32, shape=[None, self.dim_x], name='x'), but in TensorFlow 2, we must disable eager execution tf.compat.v1.disable_eager_execution() because the default is graph-based execution.
4. The change from tf.contrib.layers.xavier_initializer() to tf.keras.initializers.GlorotUniform() was made because TensorFlow 2 deprecated the tf.contrib module, recommending the use of tf.keras.

5. The line tf.layers.dropout() was changed to tf.keras.layers.Dropout().

6. The line tf.train.exponential_decay() was changed to tf.keras.optimizers.schedules.ExponentialDecay()

In the following tables, we can observe the resulting MSEs from implementing each of the experiments mentioned in the previous section for each kind of data-generating function using AutoIV and TrueIV.

Table 1: NoIV

	sin	step	abs	linear	poly2d	poly3d
VanNN	1.00 ± 0.01	1.00 ± 0.00	0.98 ± 0.05	0.72 ± 0.24	0.99 ± 0.03	0.99 ± 0.02

Table 2: AutoIV

	sin	step	abs	linear	poly2d	poly3d
2SLS	0.89 ± 0.06	0.67 ± 0.03	0.23 ± 0.03	0.03 ± 0.01	0.16 ± 0.01	0.57 ± 0.05
Poly-Ridge	0.95 ± 0.01	0.67 ± 0.03	1.00 ± 0.00	0.03 ± 0.00	0.43 ± 0.02	0.98 ± 0.01
2S-NN	1.00 ± 0.01	0.99 ± 0.10	0.96 ± 0.16	0.42 ± 0.44	0.77 ± 0.37	1.00 ± 0.03

Table 3: TrueIV

	sin	step	abs	linear	poly2d	poly3d
2SLS	2.99 ± 9.28	0.69 ± 0.04	0.23 ± 0.04	0.03 ± 0.01	0.16 ± 0.02	0.76 ± 0.37
Poly-Ridge	0.95 ± 0.01	0.68 ± 0.02	1.00 ± 0.00	0.03 ± 0.00	0.43 ± 0.02	0.98 ± 0.01
2S-NN	1.01 ± 0.02	0.96 ± 0.16	0.98 ± 0.07	0.49 ± 0.48	0.84 ± 0.28	1.00 ± 0.03

In all instances except four the VanNN experiment has worse results than the experiments that use either the AutoIV or the TrueIV. In all cases except two, the AutoIV shows better results than the TrueIV. The step function shows the worse results and the linear function show the best results. All these results are consistent with the original paper.

V. Discussion

Initially, we thought that the code provided by the author created an IV but the code to implement IV-based methods was not provided. So, we coded the experiments, but the resulting mean and standard error MSE are very different from those reported. In order to address the issue, I contacted the author of the original paper. His answer indicated that the experiments are actually performed using the code provided, but they are implemented using neural networks.

In addition to the table, there were two graphs that helped us understand the effectiveness of AutoIVs. These graphs were generated for poly3d but they can be generated in a similar manner for the rest of the functions. Figure 4 shows the X-fitted values that result from the first stage of a 2SLS. Figure shows the Y-fitted values that result from fitting the 2SLS. In both cases, the TrueIV and the AutoIV generate similar results. Finally, Figure 6 shows the model generated using AutoIV.

Now, I will proceed with some final remarks. Understanding the methodology of the paper was probably one the most difficult tasks throughout the project. In particular, we started this project thinking that

the AutoIVs were produced independently of the experiments, but this was not the case. Some possible extensions to research in this area would be applying AutoIVs to real-world data and evaluating its performance.

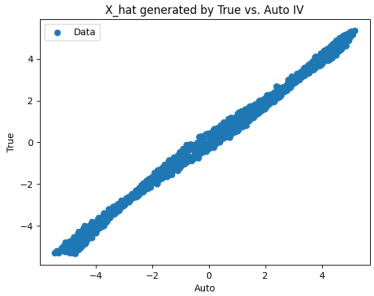

Figure 4: X fit.

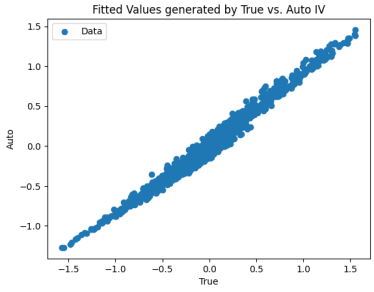

Figure 5: Y fit.

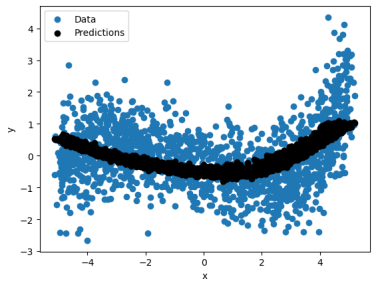

Figure 6: Model.