The Causal Impact of anti-Depressants at the Individual Level

Noah Dixon

9-4-24

Research Proposal

Question: Is it possible to infer causality, at the individual level, when administering antidepressants for depressive symptoms?

Main Idea: Ideally, to test this hypothesis, we would run a double-blind N-of-1 study with periodic, blocked randomization. However, with antidepressants, this is infeasible, due to biological concerns (e.g., high stabilization and washout periods) and ethical concerns (e.g., suicidal ideation). Fundamentally, this restricts analysis to the aggregate level – at which point, time-varying heterogeneity is obscured. Consequently, in order to evaluate the individual-level effect, we need to treat the data as being quasi-experimental. Traditionally, this is where the analysis stopped – there was no plausible way to extract this much data. However, recent technological advancements have altered the bounds of what is permissible. In Brodersen et al. (2015), for instance, they introduce a Bayesian Structural Time Series (BSTS) model to estimate the individual-level counterfactual time series had the intervention not occurred, which they subsequently apply to marketing data. Additionally, in Robins (1986), he introduces the g-method approach – an approach to individual-level causal inference when there is an abundance of longitudinal data. One serves as a methodological framework; the other as a causal framework. The idea is to "layer" one on top of the other, to construct an individual-level counterfactual that uses a conglomerate utility measure rather than a single outcome variable, and which can subsequently be used to identify the personal average treatment effect (PATE). From here – and assuming a true RCT is conducted – it may be possible to reconstruct the ATE, with individual-level heterogeneity.

Basic Framework

Motivation: At the individual level, mental health is diagnosed by weekly tracking an individuals responses to the HAM-D (Hamilton Depression Rating Scale).[1] At the population level, tests are validated using an RCT, which is subsequently estimated using an MMRM model. However, while such metrics establish average efficacy, they may not translate to *individual* responses. Fundamentally, these measures do not account for interdependencies across time, individual-level shocks, any form of biometric tracking, most forms of measurement error, and only a small subset of adverse effects. Further, the consensus in the literature is that antidepressants are only beneficial, in the long run, for individuals with severe depression (Fournier et al. 2010; Kirsch et al. 2008), with the effect sizes being statistically small (Cipriani et al. 2018) or negligible for those with small-to-mild depression. Around 11-13% of adults report using antidepressants, but only 2-3% of individuals report having severe depression (Brody & Gu 2020; Villarroel & Terlizzi 2020). Further, Wong, Motulsky, and Eguale (2016) report that over 50% of individuals may be using antidepressants for something other than depression. *And all of this assumes that an individuals utility is inextricably linked to their score on the test.* There are many reasons to believe that the individual-level data

[1] Although recently, new tests – like the MADRS (Montgomery-Asberg Depression Rating Scale), QIDS-C (Clinician-rated Quick Inventory of Depressive Symptomatology), and PHQ-9 (Patient Health Questionnaire) – have gained traction, each successive test placing more emphasis on the individual.

neglected under an MMRM model is an integral piece of the puzzle, and thus there is a necessity, in the literature and in the market, to introduce rigorous statistical methodology to deduce the causal impact of antidepressants on a *specific* individual.

Experimental Design

Outcomes: The outcomes tracked are tri-daily (morning, afternoon, evening) "mood" scores and weekly HAM-D scores, in conjunction with "richer" measures, including: sleep, step count, mood diary, adverse events, and biomarkers. More generally, if the data is saturated enough, we *may* be able to identify the PATE, absent of the ATE.

Methodology: Hinges on the methodological framework proposed in (Brodersen et al. 2015):

- **No anticipation.** This is hard to justify, in practice. An individual may alter their behavior in anticipation of the medicine. One way to obviate this is to randomize when an individual begins taking the medication. For example, the individual is told that they will begin taking the medication sometime over the course of the month, but they do not know exactly when this will occur. This is something that could reasonably be approved by IRB.

- **Structural stability of the counterfactual model.** Using a Bayesian Structural Time Series (BSTS) model helps make a sufficient case for this. In practice, it is impossible to justify. One pitfall of this model is that we can never fully recover an unbiased estimate. However, we can recover a *plausible estimate*. At the individual-level, this is probably good enough.

- **No concurrent interventions.** If an individual begins taken epileptic medication at the same time, it is impossible to extract the true effect of the antidepressant. However, it is still possible to extract the aggregate effect of the antidepressant and epileptic medication, given that the other assumptions hold.

- **Sufficient pre-intervention data.** This requires collecting a sufficient amount of data, *prior* to treatment.

Identification: The BSTS assumptions have nothing to do with identification. In order to identify, we need to layer on a causal framework. More generally, the basic causal assumptions, as stated in Hernn and Robins (2020), are:

- **Well-defined intervention and outcome**. This requires consistent timing, dosage, and measurement each period. This is automatically guaranteed under a quasi-experimental design.

- **Consistency.** Denote A_{it} as the treatment or exposure for person i at time t. Define Y_{it} as being the outcome at time t. Then, if $A_{it} = a$,

$Y_{it} = Y_{it}(a)$. In other words, there is no interference from *other* periods unless modeled.

- **Positivity.** For all relevant histories, $0 < \Pr(A_{it} = 1 | H_{it}) < 1$. This is not necessarily guaranteed, under the counterfactual argument.

- **Sequential Exchangeability.** Define

$$H_{it} = \{C_i; A_{i,1}, ..., A_{i,t-1}; Y_{i1}, ...Y_{i,t-1}; ...L_{i1}...L_{i,t}$$

  as being the *history* for a person i at time t, where C_i represents baseline covariates, $A_{i,1:t-1}$ represents past treatments/doses, $Y_{i,1:t-1}$ represents past outcomes/symptoms, and L_{is} represents time-varying covariates measured directly before the t'th decision. Then, we require that $A_{it} \perp\!\!\!\perp (Y_{i,t}^{(0)}, Y_{i,t}^{(1)}) \mid H_{it}$. This is the hardest assumption to justify.

Estimation: The standard BSTS model, as shown in (Brodersen et al. 2015), is given by a system of equations:

$$Y_t = Z_t^\top \alpha_t + \varepsilon_t \qquad \text{[Outcome Equation]},$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t \qquad \text{[State Equation]},$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ and $n_t \sim \mathcal{N}(0, Q_t)$ are independent of all other unknowns. The first equation is the outcome equation and represents what we are actually controlling for.

The variables are represented by:

- Y_t is a scalar, which represents the PHQ-9 scores. It can also include other measures. If running a joint model, this opens up an entire other can of worms.

- Z_t is essentially a control matrix, which includes a dummy variable if treated. For instance, $Z_t = [1, D_t, S_{1,t}, X_{2,t}]'$ represents the case in which a set of seasonal components are included, as well as a set of covariates.

- α_t represents the coefficients associated with each portion of the control matrix. For instance, in the previous example, the corollary $\alpha_t = [\mu_t, \tau, \beta_1, \beta_2]$, where μ_t represents an individual's underlying mental health trajectory absent of treatment and τ represents the inferred causal impact of the treatment itself. Obviously, the question is: *how do we deal with μ_t.* This is where the state equation comes into play.

- α_{t+1} governs how the latent states evolve over time. That is, how does an untreated individual evolve over time. This is essential in modeling the counterfactual.

- T_t is a transition matrix, which tells you how the state evolves, deterministically, over time. The degree to which we model trends, in general, is reflected inside this. *It must match the seasonal trends reflected in α_t.*

- R_t determines which states receive noise and how much noise they receive.

- Q_t represents the state disturbance covariance matrix, and controls how much the latent components are allowed to evolve over time.

To model this more specifically, suppose the outcome equation is given as:

$$Y_t = \mu_t + \tau \cdot D_t + X_{t-1}\beta + \gamma_t + \varepsilon_t,$$

where $X_{t-1} = [\text{Sleep}_{t-1}, \text{Exercise}_{t-1}, \text{Average Heart Rate}_{t-1}, \text{Environmental Stressors}_{t-1},$ $\gamma_t = \gamma_t^{(\text{Week})} + \gamma_t^{(\text{Circadian})} + \gamma_t^{(\text{Seasons})}$ represents latent seasonal effects, and $\mu_t = \mu_{t-1} + \eta_t$ represents how a person's depression would naturally evolve over time, absent of medication or seasonal cycles. In fact, the distribution of η_t can align with aggregate population data with certain priors. Note that each $\gamma_t^{(\cdot)}$ has its own transition matrix, such that $\gamma_t(T) = -\sum_{j=1}^{T-1} \gamma_t^{(j)}$. Essentially, these work as dummy variables that are included in the outcome equation, but which are estimated with some degree of uncertainty; they influence both the outcome equation and the state equation. Q_t is either *learned* through posterior sampling, or estimated from the data, typically using MLE.

g-methods: In the process of learning.

Utility: Rather than evaluating each estimated measure, in isolation, a conglomerate measure of utility should be estimated, based on an individual's ranking of each respective metric. As an example,

$$U_i = W_A \cdot \text{Sleep} + W_B \cdot \text{PHQ-9} + W_C \cdot \text{Fulfillment},$$

where W_A represents arbitrary weights such that the $\sum_{j=1}^{J} W_J = 1$ and j represents the number of categories. Likely, using machine learning techniques is the best way to approach this.

Economic Importance: The goal of causal inference is to evaluate the average effect on an arbitrary member of the population. However, one question is whether this metric is always sufficient, in application. Sometimes, the goal may not be generalizability but individualization. This idea – of integrating individual-level heterogeneity into traditional aggregate models – is already influencing modern macroeconomic models. It is also at the heart of the traditional field experiment vs. lab experiment argument. There is no reason to believe that large-scale "individualized econometrics" is not coming next.

Additionally, there may be large welfare gains – and efficiency gains – from targeted dosing vs. a one-size-fits-all approach, or from a one-size-fits-all approach to corollary subgroups.

Finally – and perhaps more importantly – it may be possible to aggregate these results together to uncover the population effect (obviously, under a different control/placebo experimental design). That is, we may be able to recover an ATE *with individual-level heterogeneity*, which could have intriguing applications.

Data: The only way to conduct this study is to run some manipulation of an experiment which requires lots of funding, planning, and development.

Sources: In Zotero.