Causality: Holland (1986)

3rd Lecture

André Yoshizumi Gomes

IME/USP - Instituto de Matemática e Estatística

October 13th, 2015

About the author: Paul W. Holland

Currently holder of the Frederic M. Lord Chair in Measurement and Statistics at Educational Testing Service.

Held faculty positions at the Graduate School of Education, University of California Berkeley (1993-2000) and the Harvard Department of Statistics (1966-1972).

Elected Member of the International Statistical Institute and past president of the Psychometric society.

He was awarded the (AERA/ACT) E. F. Lindquist Award in 2000.

Received an MA and Ph.D. in Statistics from Stanford University and BA in Mathematics from the University of Michigan.

Introduction

The reaction of many statisticians when confronted with the possibility that their profession might contribute to a discussion of causation is immediately to deny that there is any such possibility.

Barnard (1982): *"That correlation is not causation is perhaps the first thing that must be said."*

A well-designed randomized experiment can be a powerful aid in investigating causal relations to question the need for such a defensive posture by statisticians.

Introduction

The statistical models used to draw causal inferences are distinctly different from those used to draw associational inferences.

The emphasis here will be on measuring the *effects of causes* (like Rubin on his 1974 paper).

An emphasis on the effects of causes rather than on the causes of effects is, in itself, an important consequence of bringing statistical reasoning to bear on the analysis of causation.

Model for associational inference

Population or universe U of "units" (u).

For each u in U there is an associated value $Y(u)$ of a *response variable* Y and another value $A(u)$ of an *attribute variable*.

All probabilities, distributions, and expected values involving variables are computed over U.

For associational inference, the role of *time* is simply to affect the definition of the population of units or to specify the operational meaning of a particular variable.

Model for associational inference

The *joint distribution* of Y and A over U is specified by $P(Y = y, A = a)$, the proportion of u in U for which $Y(u) = y$ and $A(u) = a$.

The associational parameters are determined by this joint distribution (such as the conditional expectation of Y given A).

Associational inference consists of making statistical inferences (estimates, tests, posterior distributions, etc.) about the associational parameters relating Y and A on the basis of data gathered about Y and A from units in U.

Rubin's model of causal inference

Units in the model for causal inference are the objects of study on which *causes* or *treatments* may act.

The effect of a cause is **always** relative to another cause.

For example, the phrase "A causes B" almost always means that A causes B relative to some other cause that includes the condition "not A".

We will deal with two causes (or levels of treatment): t (the treatment) and c (the control).

What can be a cause?

For causal inference, it is critical that each unit be **potentially exposable** to any one of the causes.

- The schooling a student receives can be a cause of the student's performance on a test;
- The student's race or gender cannot be a cause.

Let S be a variable that indicates the cause to which each unit in U is exposed. The critical feature of the notion of cause in this model is that the value of $S(u)$ for each unit *could have been different*.

Response variables

The role of time now becomes important.

Variables divide into two classes: *pre-exposure* – those whose values are determined prior to exposure to the cause; and *post-exposure* – those whose values are determined after exposure to the cause.

The values of post-exposure variables are potentially affected by the particular cause, t or c, to which the unit is exposed.

We need *two* variables, Y_t and Y_c, to represent two potential responses.

The causal effect

- $Y_t(u)$: the value of the response that would be observed if the unit were exposed to t;
- $Y_c(u)$: the value of the response that would be observed *on the same unit* if it were exposed to c.

The effect of the cause t on u as measured by Y and relative to cause c is the difference between $Y_t(u)$ and $Y_c(u)$:

$$Y_t(u) - Y_c(u). \tag{1}$$

Fundamental Problem of Causal Inference

It is impossible to *observe* the value of $Y_t(u)$ and $Y_c(u)$ on the same unit and, therefore, it is impossible to *observe* the effect of t on u.

The implicit threat is that causal inference is impossible. But we should not jump to that conclusion too quickly!

Even though simultaneous observation of $Y_t(u)$ and $Y_c(u)$ is impossible, it doesn't mean that knowledge relevant to these values is completely absent.

Two general solutions

The scientific solution: to exploit various homogeneity or invariance assumptions.

"The value of $Y_c(u)$ measured at an earlier time is equal to the value of $Y_c(u)$ for the current experiment (homogeneity). All we need to do now is to expose u to t and measure $Y_t(u)$."

By careful work he may convince himself and others that this assumption is right, but he can never be absolutely certain.

Two general solutions

The statistical solution: replaces the unobservable causal effect of t on a specific unit with the estimable *average causal effect* of t over a population of units.

The average causal effect, T, of t (relative to c) over U is the expected value of the difference $Y_t(u)$ - $Y_c(u)$ over the u's in U, which can be expressed by

$$T = E(Y_t) - E(Y_c). \tag{2}$$

Information on *different* units that *can be observed* can be used to gain knowledge about T.

Observed response

It is useful to have a notation to express the fact that the causal indicator variable S determines which value, Y_t or Y_c, is observed for a given unit.

- If $S(u) = t$, then $Y_t(u)$ is observed, and
- If $S(u) = c$, then $Y_c(u)$is observed.

The observed response on unit u is $Y_{S(u)}(u)$. Therefore, the observed response variable is Y_S.

Even though the model contains three variables, S, Y_t and Y_c, the process of observation involves only two, S and Y_S.

Temporal stability and causal transience

We will see ahead how specific assumptions added to the model allow causal inferences of particular types.

The *temporal stability* and *causal transience* assumptions tell respectively that:

1. the value of $Y_c(u)$ does not depend on when the sequence "apply c to u then measure Y on u" occurs;
2. the value of $Y_t(u)$ is not affected by the prior exposure of u to the sequence in 1.

This is one way to apply the scientific solution to the Fundamental Problem of Causal Inference.

Unit homogeneity

To assume that $Y_t(u_1) = Y_t(u_2)$ and $Y_c(u_1) = Y_c(u_2)$ for two units u_1 and u_2.

The causal effect of t is taken to be the value of $Y_t(u_1) - Y_c(u_2)$.

Laboratory scientists convince themselves that the units are homogeneous by preparing them carefully so that they "look" identical in all relevant aspects.

Independence

The average causal effect T is the difference between the two expected values $E(Y_t)$ and $E(Y_c)$. The observed data (S, Y_S), however, can only give us information about

$$E(Y_S|S = t) = E(Y_t|S = t) \tag{3}$$

and

$$E(Y_S|S = c) = E(Y_c|S = c). \tag{4}$$

But what happens when units are randomly assigned either to cause t or c?

Independence

If the physical randomization is carried out correctly, then it is plausible that S is *independent* of Y_t and Y_c and all other variables over U.

If this assumption holds, then

$$E(Y_t) = E(Y_t|S = t) \tag{5}$$

and

$$E(Y_c) = E(Y_c|S = c). \tag{6}$$

Independence

Hence under the independence assumption the average causal effect T satisfies the equation

$$T = E\left(Y_S|S = t\right) - E\left(Y_S|S = c\right). \tag{7}$$

What about (7) if independence does not hold?

- *Prima facie causal effect*.

$$T_{PF} = E\left(Y_S|S = t\right) - E\left(Y_S|S = c\right). \tag{8}$$

Constant effect

If the variability in the causal effects $Y_t(u) - Y_c(u)$ is large over U, then T may not represent the causal effect of a specific unit, u_0, very well.

The assumption of *constant effect* (or *additivity*) is that the effect of t on every unit is the same:

$$T = Y_t(u) - Y_c(u), \ \forall u \in U. \tag{9}$$

The unit homogeneity assumption implies constant effects.

Comments on selected philosophers

So much has been written about causality by philosophers, and we will show some of these ideas in the context of Rubin's model for causal inference.

The first one: Aristotle's four "causes" of a thing (*Physics*).

- The material cause (that out of which the thing is made);
- the formal cause (that into which the thing is made);
- the efficient cause (that which makes the thing);
- the final cause (that for which the thing is made).

Comments on selected philosophers

Locke (1690): "*That which produces any simple or complex idea, we denote by the general name 'cause', and that which is produced, 'effect'.*"

Aristotle emphasized the *causes* of a thing rather than the effects of causes. Locke seems a little more even-handed.

See Bunge (1959) for a discussion of the history of many ideas about the essential meaning of causation.

Hume (1740, 1748)

In the analysis of the idea that A causes B, Hume has recognized three basic criteria for causation:

1. spatial/temporal contiguity: A and B are homogeneous in space and time;
2. temporal succession: A precedes B in time, and
3. constant conjunction: A and B always occur (or do not occur) together.

In terms of the model, the latter criterion might not hold. One reason is "measurement error", and another is that the casual effects $Y_t(u) - Y_c(u)$ may vary with u.

Hume (1740, 1748)

The author pointed out some notions that were missing from Hume's analysis:

- *"(...) the effect of cause is always relative to another cause. The notion that a cause could have been different from what it was and that it is this difference that defines the effect."*
- *"(...) causes are not delineated in any way. Anything can be a cause."*

Mill (1843)

John Stuart Mill was positively disposed toward experiments.

- *"Observation, in short, without experimentation (supposing no aid from deduction) can ascertain sequences and co-existences, but cannot prove causation."*
- *"We have not yet proved that antecedent to be the cause until we have reversed the process and produced the effect by means of that antecedent artificially, and if, when we do so, the effect follows, the induction is complete."*

Mill identified four general methods of experimental inquiry.

Mill (1843)

Method of Concomitant Variation

- *"Whatever phenomenon varies in any manner, whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation."*

"Where there is correlational smoke there is likely to be causational fire." (But is it really?)

Mill (1843)

Method of Difference

- *"If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring in the former; the circumstances in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause of the phenomenon."*

Almost an exact statement of what we mean by a causal effect (although its proposed use is to identify causes and effects).

Mill (1843)

- "Phenomenon under investigation" occurs – $Y = 1$.
- "Phenomenon under investigation" does not occur – $Y = 0$.
- "The circumstance in which the instances differ" – when present $= t$, when absent $= c$.

Then $Y_t(u) = 1$ denotes the fact that when the circumstance was present the phenomenon occurs, and $Y_c(u) = 0$ denotes the fact that when the circumstance was absent the phenomenon did not occur.

Thus $Y_t(u) - Y_c(u) = 1$: "the cause or an indispensable part of the cause of the phenomenon." (the causal effect)

Mill (1843)

Method of Residues

- *"Subduct from any phenomenon such part as is known by previous inductions to be the effect of certain antecedents, and the residue of the phenomenon is the effect of the remaining antecedents."*

Let the antecedents (i.e., causes) be denoted by a = those whose effect is known and b = the remaining antecedents.

The causal effect of ab relative to a is simply $Y_{ab}(u) - Y_a(u)$, which is Mill's residue.

Mill (1843)

Method of Agreement

- *"If two or more instances of a phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the cause (or effect) of the given phenomenon."*

All that the method really does is to *rule out* possible causes.

If the phenomenon occurs when the cause t occurs and also when the cause t does not occur, that is, c, we have $Y_t(u) = 1$ and $Y_c(u) = 1$, and so, $Y_t(u) - Y_c(u) = 0$ (t is a cause with *null effect*).

Suppes (1970)

Patrick Suppes's goal was to improve upon Hume's analysis, specifically the constant conjunction criterion.

- *"Roughly speaking, the modification of Hume's analysis I propose is to say that one event is the cause of another if the appearance of the first event is followed with a high probability by the appearance of the second, and there is no third event that we can use to factor out the probability relationship between the first and second events. "*

Then he made three definitions:

Suppes (1970)

1. If $r < s$ denote two time values, the event C_r is a *prima facie cause* of the event E_s if

$$P(E_s|C_r) > P(E_s). \tag{10}$$

2. C_r is a *spurious cause* of E_s if C_r is a prima facie cause of E_s and for some $q < r < s$ there is an event D_q such that

$$P(E_s|C_r, D_q) = P(E_s|D_q) \tag{11}$$

and

$$P(E_s|C_r, D_q) \geq P(E_s|C_r). \tag{12}$$

3. C_r is a *genuine cause* of E_s if C_r is a prima facie cause of E_s. but C_r is not a spurious cause of E_s.

Suppes (1970)

The author states that Suppes's theory does not appear to get to the heart of the notion of causation which Rubin's model does. It does capture some useful ideas though.

Holland was able to conclude that "*the treatment in a randomized experiment is a spurious cause of the effect if and only if it has a positive average causal effect, but a subpopulation of units can be identified on the basis of pre-exposure variables*

1. *on which the average causal effect is 0 and*
2. *for which the response under t is more likely to occur than it is for all of U.*"

Comments from a few statisticians

Kempthorne (1952): discussion of the analysis of randomized block designs.

- *"We shall denote the yield with treatment k... on plot j... of block i... by y_{ijk}."* (different versions of the response – one for each k)
- *"In fact we do not observe the yield of treatment k on plot j but merely the yield of treatment k on a randomly chosen plot in the block. ... we denote the observed yield of treatment k in block i by y_{ik}.*

Comments from a few statisticians

D. R. Cox (1958): defined *true treatment effects*.

In an experiment with treatments T_1, T_2, it is the difference between *"the observation obtained on any unit when, say, T_1 is applied"*, and *"the observation that would have been observed had, say, T_2 been applied"*.

Cox also made the assumption of constant effect in defining true treatment effects. His reasons appear to be primarily technical rather than conceptual. He did not reject the idea of variable causal effects, however.

Comments from a few statisticians

R. A. Fisher (1926): curiously never dealt directly with the idea of multiple versions of the response.

He could have considered the possibility of the null hypothesis of no treatment effect.

- *"What reason is there to think that, even if no manure had been applied, the acre which actually received it would not still have given the higher yield?"*

Controversy between Fisher and Neyman

The controversy revolves around the choice of null hypothesis in experiments such as randomized block designs.

Fisher was quite clear that the null hypothesis that he proposed is that the causal effect (as we have defined it) is 0 for each unit.

In 1935 Fisher first quoted Neyman, as follows:

- *"... this bias vanishes when... the objects compared are reacting to differences in soil fertility in exactly the same manner. ... This is not always true."*

Controversy between Fisher and Neyman

Then Fisher wrote:

- *"However, it was always true in the case for which it was required, namely, when the hypothesis to be tested was true, that differences of treatment made no difference to the yields."*

Then Neyman, in responding to Fisher's remarks, emphasized his interest in the average causal effect.

Controversy between Fisher and Neyman

- *“ ‘Our purpose in the field experiment consists in comparing numbers such as $X..(k)$, or the average true yields which our objects are able to give when applied to the whole field.’ It is seen that this problem is essentially different from what Professor Fisher suggested. So long as the average yields of any treatments are identical, the question as to whether these treatments affect separate yields on single plots seems to be uninteresting and academic.”*

Controversy between Fisher and Neyman

Fisher's sardonic reply indicates that, at least, he agreed that Neyman stated their differences clearly.

- *"It may be foolish, but that is what the z test was designed for, and the only purpose for which it has been used."*

What can be a cause? (again)

"In this article I take the position that causes are only those things that could, in principle, be treatments in experiments."

"I believe that the notion of cause that operates in an experiment and in an observational study is the same. The difference is in the degree of control an experimenter has over the phenomena under investigation compared with that which an observer has."

(In Rubin's model, total control can make S independent of Y_t and Y_c)

What can be a cause? (again)

Take three examples of statements that involve the word *cause*:

1. She did well on the exam because she is a woman. (an attribute)
2. She did well on the exam because she studied for it. (a voluntary activity)
3. She did well on the exam because she was coached by her teacher. (an imposed activity)

What can be a cause? (again)

In Example 1, an attribute cannot be a cause in an experiment, because the notion of potential exposability does not apply to it.

At the other extreme is Example 3, easily interpreted in terms of the model. Had she not been coached by her teacher she would not have done as well as she did.

Example 2 is just one of many types of examples in which the applicability of the model is not absolutely clear. The problem arises because of the voluntary aspect of the supposed cause – studying for the exam.

Causation and Medicine

About the establishment of specific bacteria as the cause of specific infectious diseases, Yerushalmy and Palmer (1959) described three postulates formulated by the great bacteriologist, Robert Koch.

1. *"The organism must be found in all cases of the disease in question."*
2. *"It must be isolated from patients and grown in pure culture."*
3. *"When the pure culture is inoculated into susceptible animals or man, it must reproduce the disease."*

Granger Causation in Economics

In Granger's theory a variable causes another variable; that is, the values of one variable improve one's ability to predict the future values of another variable.

If X, Y, and Z denote three (possibly vector-valued) variables defined on a population, then X and Y are *conditionally independent* given Z if

$$P(Y = y|X = x, Z = z) \geq P(Y = y|Z = z). \qquad (13)$$

And the author says that X *is not a Granger cause of* Y (relative to Z) if X and Y are conditionally independent given Z.

Causal models in Social Science

What if we have a third variable R of a second set of causes, t' and c' which affects the units prior to S?

Suppose that we wish to measure the effect of studying certain material on the performance on a particular test.

We might be able to *encourage* or *not encourage* students to study the material – the R causes, t' and c'.

We might then be able to ascertain whether the students *did* or *did not* study the material – the S causes, t and c.

More on "encouragement designs": Powers and Swinton (1984).

A recent example

Brzeski, Taddy & Draper (September 2015), *Causal Inference in Repeated Observational Studies: A Case Study of eBay Product Releases*.

Question: How to properly relate *User Actions* (tasks completed, photos stored, songs downloaded, links clicked per visit) with the software new version releases?

Problem: *Early-adopter effect*. The majority of users who adopt the treatment at all do so in a relatively short time period after its release, and those who are the earliest adopters exhibit a higher average response compared to those who wait longer to try the treatment.

A recent example

Identifying all relevant PCFs may be difficult (or impossible) given a single observational study (a single release).

The authors argue that if one can observe a sequence of similar treatments (11 new version releases) over the course of a lengthy time period (a year), one can identify patterns of behavior in the experimental subjects (early-adopter effect) that are correlated with the response of interest and control for those patterns directly.

A recent example

The early-adopter effect really exists!

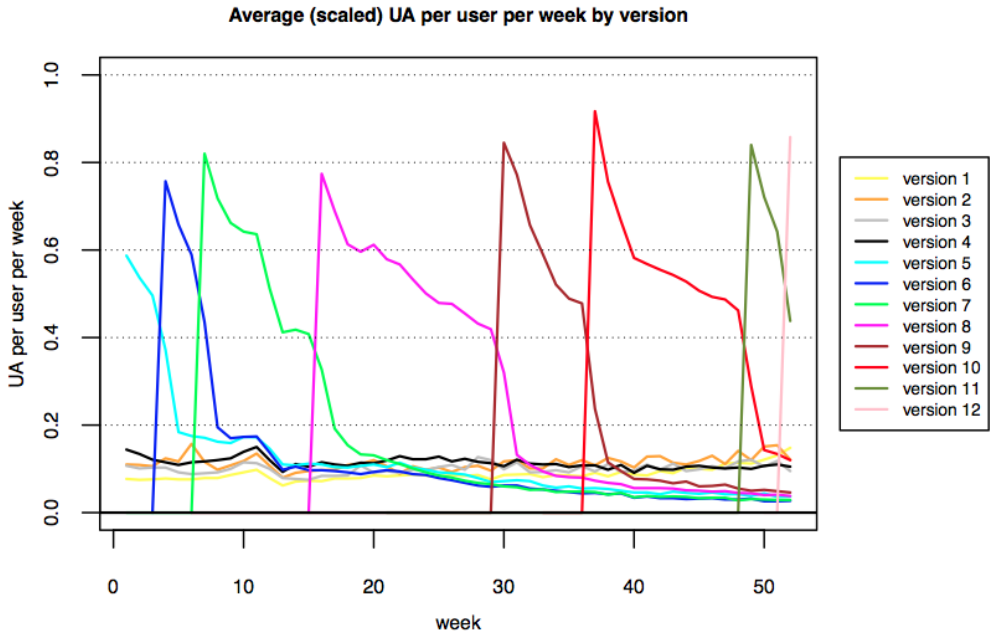

Figure 2: *Average (scaled) true UA per user per week for each individual version. This graph shows that the early adopters of a new release have the highest UA average, and that the early-adopter effect exists and is quite regular from version to version.*

A recent example

As their measure of causal effect, they use the *Conditional Average Treatment (Effect) on the Treated* (CATT – see Imbens, 2004).

They also define the *CATT Causal Ratio* (CCR) as a ratio of the estimated counterfactual and the observed answer for the treatment group.

CCR values less than 1 suggest that the effect caused by the Product release was to increase UA (user satisfaction) on average (but company experts were highly skeptical of CCR values far from 1).

A recent example

They fit their data using variations of a Bayesian hierarchical (mixed effects) model with a Gaussian error distribution.

$$\mathbf{y}_i = \mathbf{f}_i \boldsymbol{\beta}_i + \mathbf{W}_i \boldsymbol{\gamma} + \boldsymbol{\varepsilon}_i$$

$$(\boldsymbol{\beta}_i \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$(\boldsymbol{\varepsilon}_i \,|\, \nu) \sim \mathbf{N}(\mathbf{0}, \nu \mathbf{I}_{T-p})$$

$$\boldsymbol{\mu} \sim \mathbf{N}(\mathbf{0}, \kappa_\mu \mathbf{I}_d)$$

$$\boldsymbol{\gamma} \sim \mathbf{N}(\mathbf{0}, \kappa_\gamma \mathbf{I}_{T-p})$$

$$\nu \sim \text{Inv-Gamma}\left(\frac{\epsilon}{2}, \frac{\epsilon}{2}\right)$$

$$\boldsymbol{\Sigma} \sim \text{Inv-Wishart}_{d+1}(\mathbf{I})$$

A recent example

Modeling a single version in isolation resulted in somewhat suspect mean CCR estimates (0.906, 0.878, 1.188).

In order to capture the early-adopter effect, the proposed method (which models all versions jointly) yielded a more "realistic" estimate (0.998).

Also, an out-of-sample comparison between the proposed model and a simpler non-hierarchical model was performed.

A recent example

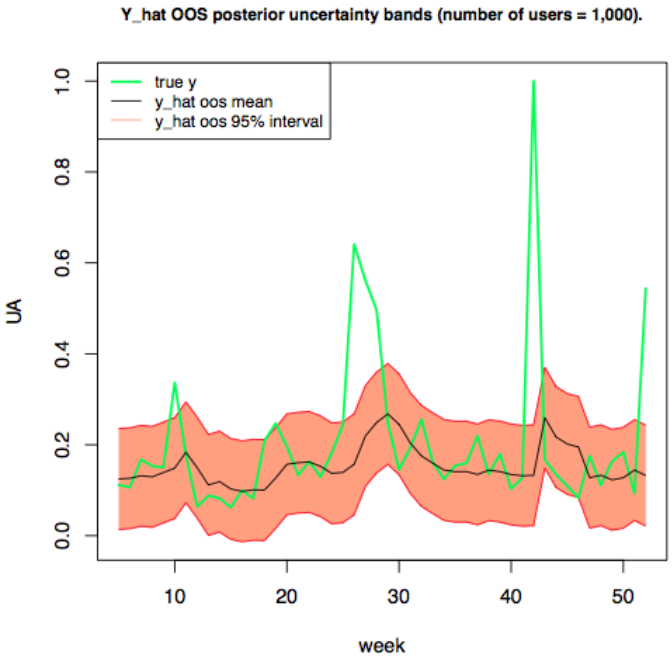

Figure 10: *Model fit results on 1,000 OOS users using the flat model. This model, using the same covariates as its hierarchical counterpart in Figure 11, cannot capture the nuances (e.g., non-normality) of the aggregate response for a small sample of users.*

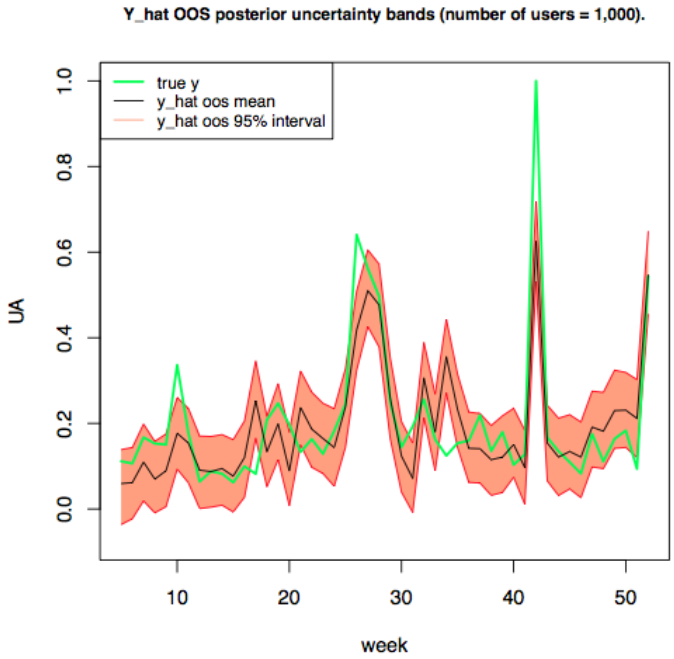

Figure 11: *Model fit results on 1,000 OOS users using our main hierarchical model. Note how much better this model is able to capture the aggregate user response compared to the flat model in Figure 10.*

Summary

1. The analysis of causation should begin with studying the effects of causes rather than the traditional approach of trying to define what the cause of a given effect is.
2. Effects of causes are always relative to other causes (i.e., it takes two causes to define an effect).
3. Not everything can be a cause; in particular, attributes of units are never causes.

NO CAUSATION WITHOUT MANIPULATION!

Summary

And two immediate consequences of Rubin's model are worth emphasizing:

1. The difference between the model (S, Y_t, Y_c) and the process of observation (S, Y_s).
2. The Fundamental Problem of Causal Inference – only Y_t or Y_c but not both can be observed on any unit u.

References

Barnard, G. A. (1982), "Causation", in Encyclopedia of Statistical Sciences (Vol. 1), eds. S. Kotz, N. Johnson, and C. Read, New York: John Wiley, pp. 387-389.

Brzeski, V., Taddy, M., Draper, D. (2015), "Causal Inference in Repeated Observational Studies: A Case Study of eBay Product Releases".

Bunge, M. (1959), Causality and Modern Science (3rd ed.), New York: Dover Publications.

Cox, D. R. (1958), The Planning of Experiments, New York: John Wiley.

References

Fisher, R. A. (1926), "The Arrangement of Field Experiments", Journal of Ministry of Agriculture, 33, 503-513.

Hume, D. (1740), A Treatise on Human Nature.

Hume, D. (1748), An Inquiry Concerning Human Understanding.

Imbens, G. (2004). "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review", Review of Economics and Statistics, 86(1), 4–29.

Kempthorne, O. (1952), The Design and Analysis of Experiments, New York: John Wiley.

References

Locke, J. (1690), An Essay Concerning Human Understanding, Book II, Chapter XXVI.

Mill, J. S. (1843), A System of Logic.

Neyman, J. (with Iwaszkiewicz, K., and Kolodziejczyk, S.) (1935), "Statistical Problems in Agricultural Experimentation (with discussion)", Supplement of Journal of the Royal Statistical Society, 2, 107-180.

Powers, D. E., and Swinton, S. S. (1984), "Effects of Self-Study for Coachable Test Item Types", Journal of Educational Measurement, 76, 266-278.

References

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", Journal of Educational Psy- chology, 66, 688-701.

Suppes, P. C. (1970), A Probabilistic Theory of Causality, Amsterdam: North-Holland.

Yerushalmy, J., and Palmer, C. E. (1959), "On the Methodology of Investigations of Etiologic Factors in Chronic Diseases", Journal of Chronic Diseases. 10, 27-40.