NOVEMBER 2024

Natural Experiments

A Public Policy Issue Paper

American Academy of Actuaries
ACTUARY.ORG

Data Science and Analytics Committee

Drafting group:

Dorothy Andrews, Ph.D., MAAA, ASA, CSPA—*Chairperson*

Brad Herrman, Ph.D., MAAA, ACAS

Leonard Reback, MAAA, FSA

Julia Romero, MAAA, FSA

The American Academy of Actuaries is a 20,000-member professional association whose mission is to serve the public and the U.S. actuarial profession. For more than 50 years, the Academy has assisted public policymakers on all levels by providing leadership, objective expertise, and actuarial advice on risk and financial security issues. The Academy also sets qualification, practice, and professionalism standards for actuaries in the United States.

American Academy of Actuaries

AMERICAN ACADEMY OF ACTUARIES
1850 M STREET NW, SUITE 300, WASHINGTON, D.C. 20036
202-223-8196 | ACTUARY.ORG

November 2024

Any references to current laws, regulations, or practice guidelines are correct as of the date of publication.

Natural Experiments

Introduction

Establishing causation has been a frequent topic of discussion within the actuarial and U.S. state regulatory community, and for the broader insurance industry. It is generally accepted that replicated randomized controlled trials (RCTs) are the gold standard methodology for establishing causation between two variables.

However, in practice, there are several challenges associated with RCTs that make them less suitable, if not impossible to use, as an approach for research in several areas of actuarial work. Consider, for example, the relationship between behaviors like driving and mortality or slippery roads and auto crashes. To use an RCT to investigate either of the relationships above, the investigator would need to deliberately manipulate a variable that they believe would cause someone to have an auto crash. Such experiments would likely be unacceptable to regulators and society at large, necessitating an alternative to RCTs when evaluating causation when there are practical or ethical challenges.

This paper discusses *natural experiments*, an approach that approximates the construction of an RCT, but has lower ethical, practical, and financial barriers. This paper is intended to assist actuaries in cases where they seek to establish or evaluate the causal structure between predictor variables and outcomes. This paper provides a reference for alternative causal analysis and provides context for natural experiments within the larger field of causal analysis, with a brief overview of RCTs. Definitions and several detailed examples are offered, followed by a discussion of some approaches to performing an analysis using a natural experiment. Strengths and weaknesses of natural experiments are discussed, followed by conclusions and regulatory considerations.

Natural Experiment Literature Review

John Snow is generally credited with performing the first natural experiment, which is also often credited as the birth of epidemiology. This occurred during a cholera outbreak in London in 1854. At the time, the cause of cholera was unknown. Snow suspected that it may be related to drinking contaminated water. Snow conducted his experiment by surveying two categories of households—those that received their drinking water from the Southwark and Vauxhall Company, whose water supply came from the Thames River downstream from where wastewater was discharged, and those households that received water from the Lambeth Company, whose water supply came from a location on the Thames that was upstream from the wastewater discharge. This was a natural experiment because Snow could not assign which households received their water from which company. Snow found that households that received their water from Southwark and Vauxhall were several times more likely to have suffered a cholera death than those that received their water from Lambeth (or other sources, such as wells), thus demonstrating that the wastewater was causing cholera.[1]

Other historical examples of health-related natural experiments

The collapse of the Soviet Union provided a set of opportunities for natural experiments across the former USSR and its dependent economies, including Cuba.[2] Because Cuba had significant economic dependence on the Soviet Union, the fall of the Soviet Union led to a reduction in fuel and food availability in Cuba. This led to reduced calorie intake and increased physical activity, e.g., walking or biking instead of driving. In the four years following the collapse of the Soviet Union, Cubans lost on average 5 kg per person and the incidence of diabetes decreased. When Cuba's economy recovered, Cubans weight increased on average and incidence of obesity and diabetes increased. This offered evidence that increasing energy expenditure while simultaneously decreasing energy intake may have potential health benefits to a population. The implications also served as an example of an interrupted time series study, since there was a time series of Cubans' weights and incidence of obesity and diabetes before the fall of the Soviet Union, the fall of the Soviet Union caused an interruption in that time series, which changed the form of the time series after the fall of the Soviet Union.[3]

1 "Snow's Grand Experiment of 1854"; UCLA Department of Epidemiology; undated. "Remembering Dr. John Snow on the sesquicentennial of his death"; *Canadian Medical Association Journal*; June 17, 2008. "Chapter 7. John Snow and the Natural Experiment"; *Applied Statistics in Healthcare Research*; 2020.
2 "Population-wide weight loss and regain in relation to diabetes burden and cardiovascular mortality in Cuba 1980-2010: repeated cross sectional surveys and ecological comparison of secular trends;" *British Medical Journal*; 2013.
3 "The COVID-19 Pandemic"; *Circulation*; April 23, 2020. "Population-wide weight loss and regain in relation to diabetes burden and cardiovascular mortality in Cuba 1980-2010: repeated cross sectional surveys and ecological comparison of secular trends"; op. cit.

Child-development-related natural experiments

Another natural experiment tested whether the effects of early deprivation on cognition persist into early adolescence.[4] This experiment utilized the fact that in Romania, under its former communist dictator, Ceausescu, children living in state-run orphanages were subject to abuse and often received insufficient care and inadequate nutrition. This treatment resulted in impaired social and cognitive development, as well as high sensitivity to stress. In 2006, Beckett, et al., studied the difference in adolescent cognition between Romanian orphans who had been subject to Ceausescu-era orphanages and subsequently adopted in the U.K. and U.K.-born children who were adopted in the U.K. by the age of 6 months. The study found that the Romanian-born adolescents who were adopted in the U.K. by the age of 6 months had similar IQ scores to the U.K.-born adolescents. However, the Romanian-born adolescents adopted in the U.K. at an older age, having spent more time in the Romanian orphanages, had significantly lower IQ scores. This was a "difference in differences study" because it compared the difference in cognition between two different populations.[5]

In-utero natural experiments

Another natural experiment leveraged the various measures China took to reduce air pollution in Beijing shortly before the 2008 Summer Olympics.[6] This study compared the weight of babies born during the weeks around the Olympics (Aug. 8–Sept. 24, 2008) to the weight of babies born in the same period in 2007 and 2009. The study, which reports on another interrupted time series experiment, found that air pollution impacts baby weight, given that the babies born in 2008 were heavier than those born in 2007 and 2009.[7]

Automobile-related natural experiments

Natural experiments involving topics relevant to auto insurance issues are also possible, such as the impact of speed limits on crashes.[8] In 2016, both Edinburgh and Belfast reduced speed limits on a substantial portion of each city's streets, shifting from 30 or 40 mph to 20 mph. Each city implemented the speed limit reductions differently, with variations in the scope of the changes, pre-implementation education, signage, and enforcement. In Edinburgh, 12 months after the implementation of the speed limit changes, average speeds decreased by 1.34 mph, which is a statistically significant result. Accidents and fatalities decreased by more than 30%, indicating that even a modest decrease in average speeds led to a significant impact on road safety.

4 "Do the Effects of Early Severe Deprivation on Cognition Persist Into Early Adolescence? Findings From the English and Romanian Adoptees Study"; *Child Development*; 2006.
5 "List of 19 Natural Experiments"; *Economics, Psychology, Policy* (blog); June 30, 2015.
6 "Differences in Birth Weight Associated with the 2008 Beijing Olympics Air Pollution Reduction: Results from a Natural Experiment"; *Environmental Health Perspectives*; September 2015.
7 "List of 19 Natural Experiments"; op. cit.
8 "Use of natural experimental studies to evaluate 20mph speed limits in two major UK cities"; *Journal of Transport & Health*; September 2021.

In Belfast, the average speed decreased by 0.91 mph 12 months after the implementation of the new speed limits; however, this result was not statistically significant at the 95% level. Accidents and road fatalities decreased much more modestly than in Edinburgh, although the fatality rate per accident decreased by over 40%. The discrepancy between the Edinburgh and Belfast results suggests that different methods of implementation, as well as cultural differences between the cities may have created disparities in the outcomes. Both of these speed limit studies are instances of the interrupted time series structure.[9]

COVID-19-related natural experiments

The COVID-19 lockdowns also offered opportunities for natural experiments. With a reduction in cancer screenings during the lockdowns, it is possible to test whether some conditions are being over-treated or over-diagnosed, as it would be unethical for doctors to have patients intentionally skip screenings as part of a randomized experiment.[10]

The COVID-19 lockdowns permitted testing of the impact of primary pollutants from industrial activity, such as nitrogen oxides, on the development of secondary pollutants, such as ozone.[11] Due to the Covid lockdowns beaches were suddenly devoid of tourists. This led to a natural experiment in which scientists were able to study the impact that people on beaches have on biodiversity.[12] This study incorporated propensity scoring by creating indicators that controlled for the levels of noise, litter, odor, and activities to measure the levels of stressors to biodiversity on each beach.[13]

COVID-19 lockdowns also provided a natural experiment to study the incidence of asthma hospitalizations in children. Asthma hospitalizations decreased during the lockdowns, although the cause was not definitive. Possible causes included lower air pollution levels, better hygiene, and a reduction of non-Covid respiratory infections due to social distancing.[14]

Another Covid-related natural experiment using an interrupted time series focused on the sudden shift to virtual learning by many students during the lockdowns. This natural experiment compared the effectiveness of in-person versus virtual learning in Switzerland. It found that for primary school students, learning slowed down and variability in performance between students increased in the virtual setting. However, the experiment did

9 Ibid.
10 "Pandemic upheaval offers a huge natural experiment"; *Nature*; August 2021.
11 "Global Changes in Secondary Atmospheric Pollutants During the 2020 COVID-19 Pandemic." *Journal of Geophysical Research: Atmospheres*; April 2021.
12 "How does the beach ecosystem change without tourists during COVID-19 lockdown?"; *Biological Conservation*; March 2021.
13 Ibid.
14 "Initial effects of the COVID-19 pandemic on pediatric asthma emergency department utilization"; *Journal of Allergy and Clinical Immunology in Practice*; September 2020.

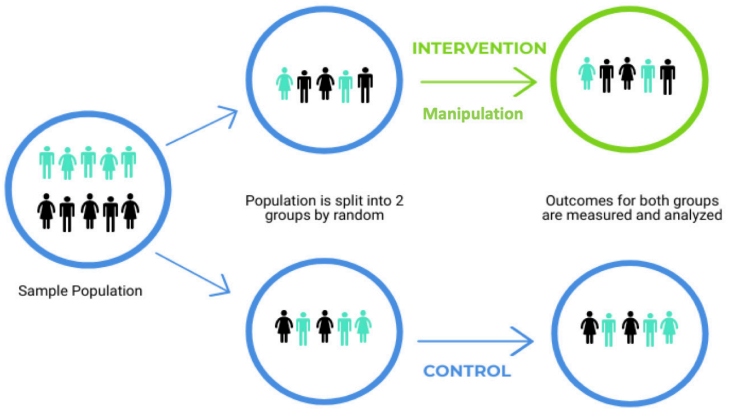
not find a statistically significant difference between in-person and virtual learning among secondary school students.[15]

A study in Germany leveraged the increase of recreational screen time to perform a natural experiment investigating the relationship between recreational screen time and physical activity. The experiment showed that while screen time increased during the lockdowns, in some cases, physical activity increased as well. This interrupted time series showed that recreational screen time and physical activity are not necessarily offsetting and the impact of one on the other is heavily influenced by context.[16]

Randomized Controlled Trials—The Established Gold Standard in Causal Research[17]

A randomized controlled trial (RCT) is an experiment designed to establish a cause-and-effect relationship by isolating the influence of a particular intervention on a certain outcome, while controlling for or holding constant other variables in the experiment. RCTs have long been considered the gold standard for establishing causation.

Figure 1 **RCT schematic showing how random population splitting allows for an intervention to be tested against a control**



Source: Nicholas Nam/World Bank

15 "Educational gains of in-person vs. distance learning in primary and secondary schools: A natural experiment during the COVID-19 pandemic school closures in Switzerland"; *International Journal of Psychology*; August 2021.
16 "Physical activity and screen time of children and adolescents before and during the COVID-19 lockdown in Germany: a natural experiment"; Scientific Reports; December 11, 2020.
17 "Randomised controlled trials—the gold standard for effectiveness research"; *BJOG: An International Journal of Obstetrics and Gynaecology*; June 19, 2018.

Participants in an RCT are randomly assigned to different groups: a control group and a treatment group. The treatment group receives the program or intervention being evaluated, while the control group does not. Statistically, both the control and treatment groups are assumed to be representative of the larger group from which they are drawn, so any finding is reflective of the larger group as well. Control and treatment groups can be segregated at various levels, including an individual level or cluster level. These could reflect households, schools, villages, blocks, etc., according to feasibility and ethical factors. Before the program or intervention is introduced, the two groups are thought to be the same. The premise of RCTs is that any difference that subsequently arises between groups can then be attributed to the program or intervention.

An important aspect of an RCT is that participants are blind to whether they are in the treatment or the control group. Those running the experiment may also be blind to which participants are in which group. Blindness helps prevent experimenters and participants from biasing results toward statistical significance when such significance may not exist. The identity of participants is only revealed when the experiment is over and the results are analyzed.

Consider an RCT for auto insurance. The research question is: Do slippery roads cause auto crashes? Researchers would start with a sample population where all participants share similar attributes so that differences would not confound experimental results. The next step is to randomly assign participants to a treatment group and a control group. The treatment group gets the manipulation—driving on a slippery road—while the control group does not. However, in this example, the principles of an RCT are violated because once the participants begin to drive, they know that they are driving on a slippery road. This approach may lead to the answer being sought, but the scenario introduces three problems. First, it may not be practical to set up experiments for every variable that you may want to consider—in this case, it is not possible to conceal which group the participant is in, treatment vs. control. Second, when drivers are aware they are driving on slippery roads, they may modify their driving behavior, potentially confounding the results. Finally, there are ethical and legal concerns related to experiments that put human lives at risk.

Machine learning algorithms approach the problem differently. These algorithms require participant data as input, while RCTs generate participant data for analysis. The amount of data may be more limited in RCTs due to the cost of collecting data from live human study

subjects and other logistical considerations. There is random sampling of treatment/control groups in RCTs and of training and test datasets in machine learning. There are metrics applied to treatment outcomes in RCTs and to algorithmic outcomes on training and test results to measure model fit to purpose. The most significant difference between machine learning algorithms and RCTs is that there is no *a priori* experimental manipulation of variables in machine learning algorithms. Machine learning algorithms, at best, leverage correlations between the predictor variables and the target (outcome) variable. They cannot prove causation as presupposed by RCTs.

Natural Experiment Methodologies

Natural experiments do not involve a predetermined experimental setup. The actual research happens *post hoc,* making it the result of a "happy accident." For example, there were two separate studies conducted on school districts with mask mandates. One found school districts with mask mandates had a 23% reduction[18] in transmission during the COVID-19 Delta wave of the pandemic, while another showed a 72% reduction[19] in transmission. Both studies measured schools with mandates against those without a mask mandate. Other examples were discussed in the "Natural Experiment Literature Review" section of this paper.

Standard statistical approaches to analyzing results of a study apply to natural experiments. But, because natural experiments do not have an *a priori* experimental design, the data collected can be disjointed, with significant discontinuities. Numerous techniques can help address these issues and allow for objective statistical analysis. The set of standard techniques described in this paper are not exhaustive. They include Interrupted Time Series, Difference in Differences, Propensity Score-Based Methods, Regression Discontinuity Methods, and Instrumental Variables.

Interrupted Time Series (ITS) study

A time series is a family of random variables indexed by time. In practical terms, a time series is a continuous sequence of observations taken repeatedly, generally at equal intervals over time, though the "equal intervals" condition is not necessary. In an Interrupted Time Series (ITS) study, a time series of a particular outcome of interest is used to establish an underlying trend, which is "interrupted" by an intervention at a known point in time.[20] If no

18 "Masking In K-12 Schools Significantly Reduces Covid-19 Among Staff And Students"; *Forbes*; March 9, 2022.
19 "Mandatory masking in schools reduced COVID-19 cases during Delta surge"; National Institutes of Health press release; March 10, 2022.
20 "Interrupted time series regression for the evaluation of public health interventions: A tutorial"; *International Journal of Epidemiology*; 2016.

intervention were to occur, it is assumed that the underlying trend would continue indefinitely. Therefore, the underlying trend, even after intervention at a known point in time, is extrapolated because it is not directly observable. This extrapolation of the underlying trend after intervention is known as the counterfactual. The values that would be implied by the counterfactual trend serve as a "negative control" even though the counterfactual is not observable. See Figure 2 for an illustration.[21] If there is a significant shift in the actual observed data after the intervention, versus the counterfactual one might interpret this to mean that the intervention caused the change. Hypothetical results of interventions in an ITS are shown in Figure 2. A form of a regression model for an ITS is the following:

$$Y_t = \beta_0 + \beta_1 t + (\beta_2 + \beta_3 t)X_t + (\beta_4 + \beta_5 t)Z_t \qquad (1)$$

where: Y_t = Outcome

$\beta_o = Intercept\ (Baseline)$

$\beta_1 =$ Pre − intervention trend

(change in the outcome per unit change in time prior to the intervention)

$\beta_2 =$ Level change following the intervention

$\beta_3 =$ Post − intervention trend

(change in the outcome per unit change in time after the intervention)

$X_t =$ Binary variable, which is 0 if during the period before the intervention and 1 if it is during the period at or after the intervention

$\beta_4 =$ Level change after some period post intervention

$\beta_5 =$ trend after some period post intervention

(change in the outcome per unit change in time after some period post intervention

$Z_t =$ Binary variable, which is 0 if during the period pre and post intervention and 1 after some period post intervention

t = Time

As in most time series regressions, autocorrelation is a significant limitation as regression models assume uncorrelated error terms. The next series of figures illustrate the meaning of the various coefficients and binary variables in the regression form above.

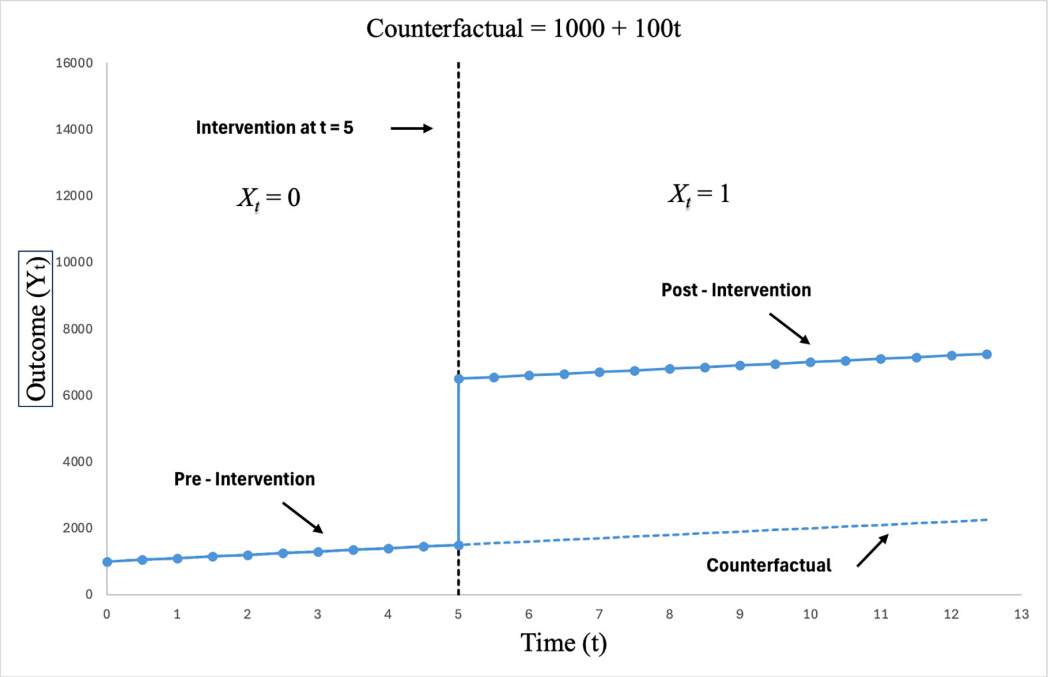
21 Ibid.

Figure 2 Different Cases of Interrupted Time Series (ITS)

$$Y_t = 1000 + 100t + 5000\,X_t$$

$\beta 0 = 1000$
$\beta 1 = 100$
$\beta 2 = 5000$
$\beta 3 = 0$
$\beta 4 = 0$
$\beta 5 = 0$

$$X_t = \begin{cases} 0, t < 5 \\ 1, t \geq 5 \end{cases}$$



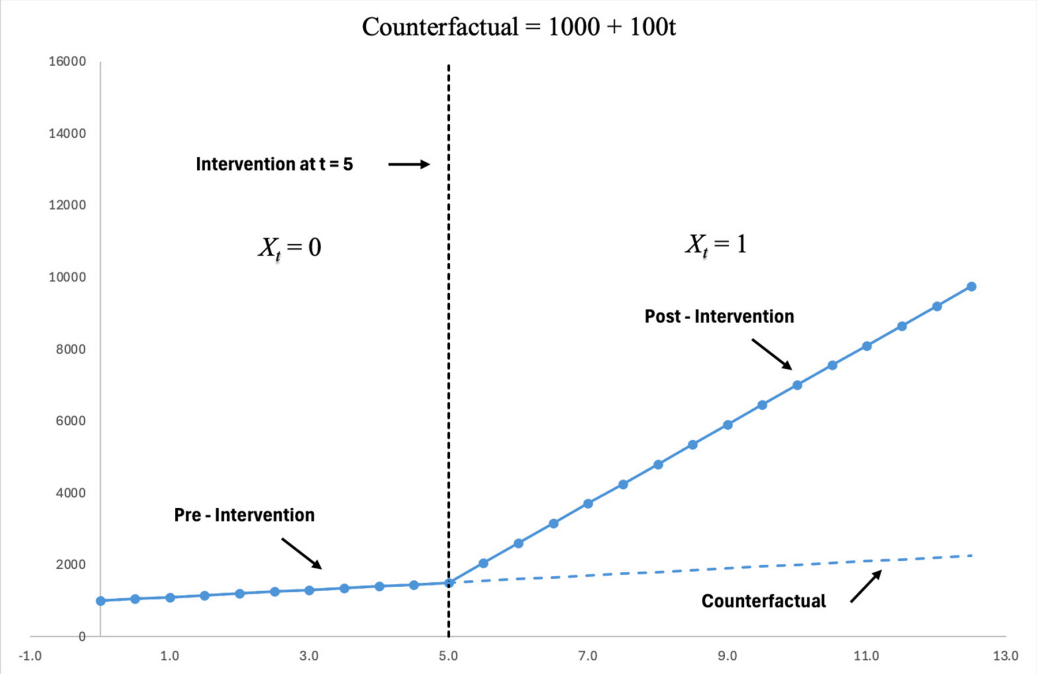
In this case, the intervention at t = 5 causes an upward shift in the outcome but no change in trend (unit change in outcome per unit change in time). The counterfactual is assumed to continue the intercept and trend post-treatment that existed prior to treatment.

Figure 2 (cont.)

$$Y_t = 1000 + 100t + (-5000 + 1000t)X_t$$

$\beta 0 = 1000$
$\beta 1 = 100$
$\beta 2 = -5000$
$\beta 3 = 1000$
$\beta 4 = 0$
$\beta 5 = 0$

$$X_t = \begin{cases} 0, t < 5 \\ 1, t \geq 5 \end{cases}$$



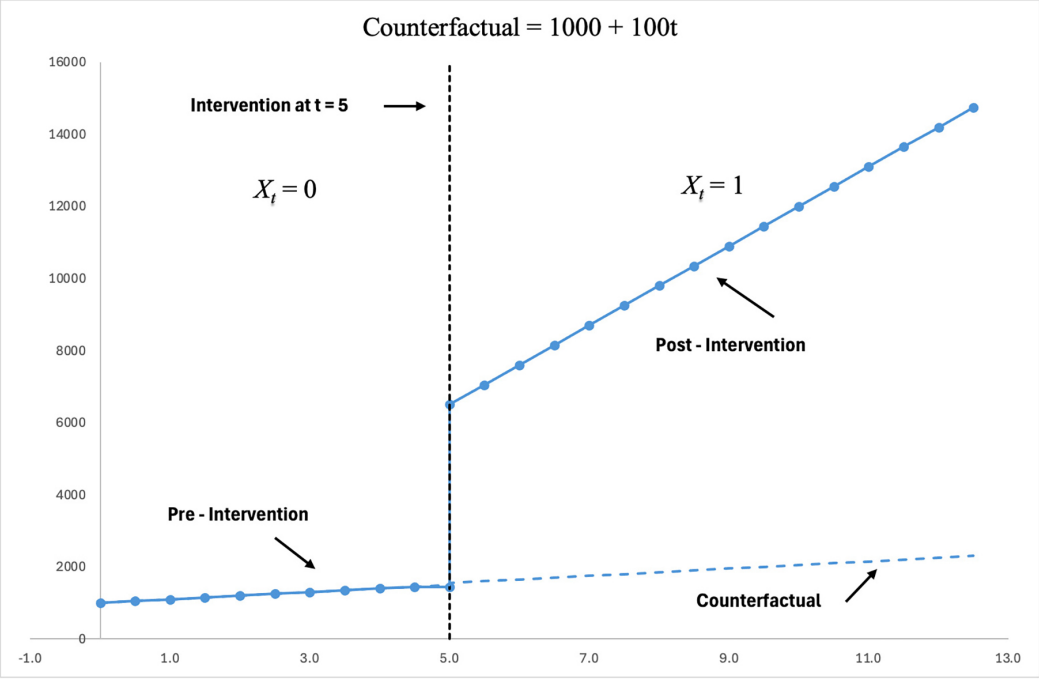
Case (b):
In this case, the intervention at t = 5 leads to no discontinuity in the outcome but causes a change in the trend.

Figure 2 (cont.)

$$Y_t = 1000 + 100t + 1000tX_t$$

$\beta 0 = 1000$
$\beta 1 = 100$
$\beta 2 = 0$
$\beta 3 = 1000$
$\beta 4 = 0$
$\beta 5 = 0$

$$X_t = \begin{cases} 0, t < 5 \\ 1, t \geq 5 \end{cases}$$



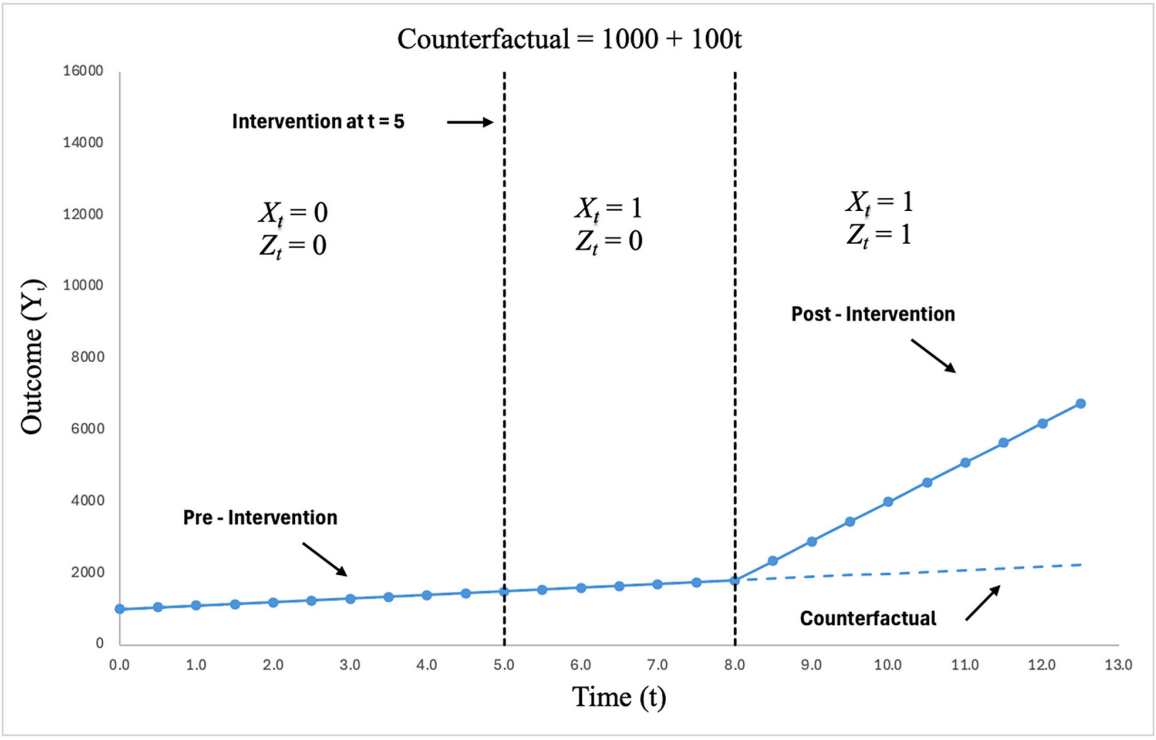
In this case, the intervention at t = 5 causes a discontinuity and a change in trend due to no change in the intercept (outcome at t=0).

Figure 2 (cont.)

$$Y_t = 1000 + 100t + (-8000 + 1000t)\, Z_t$$

$\beta 0 = 1000$
$\beta 1 = 100$
$\beta 2 = 0$
$\beta 3 = 0$
$\beta 4 = -8000$
$\beta 5 = 1000$

$$X_t = \begin{cases} 0, t < 5 \\ 1, t \geq 5 \end{cases} \qquad Z_t = \begin{cases} 0, t < 8 \\ 1, t \geq 8 \end{cases}$$



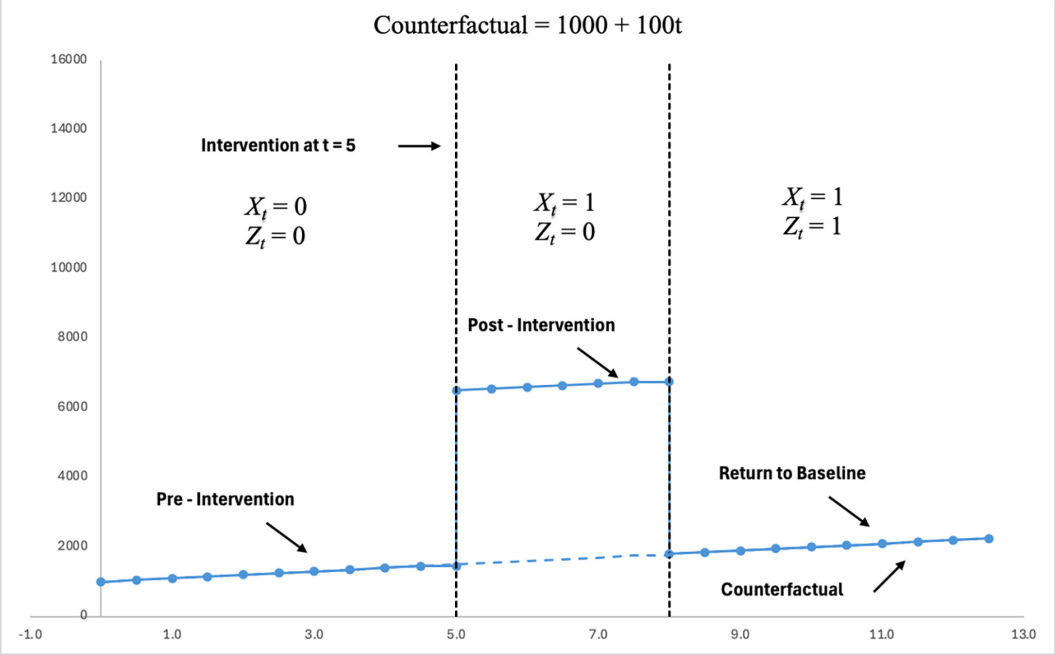
In this case, there is a lag of 3 units after the intervention at t = 5. At t = 8, there is no discontinuity in the outcome but there is a change in the trend.

Figure 2 (cont.)

$$Y_t = 1000 + 100t + 5000X_t - 5000t\,Z_t$$

$\beta 0 = 1000$
$\beta 1 = 100$
$\beta 2 = 5000$
$\beta 3 = 0$
$\beta 4 = -5000$
$\beta 5 = 0$

$$X_t = \begin{cases} 0, t < 5 \\ 1, t \geq 5 \end{cases} \qquad Z_t = \begin{cases} 0, t < 8 \\ 1, t \geq 8 \end{cases}$$



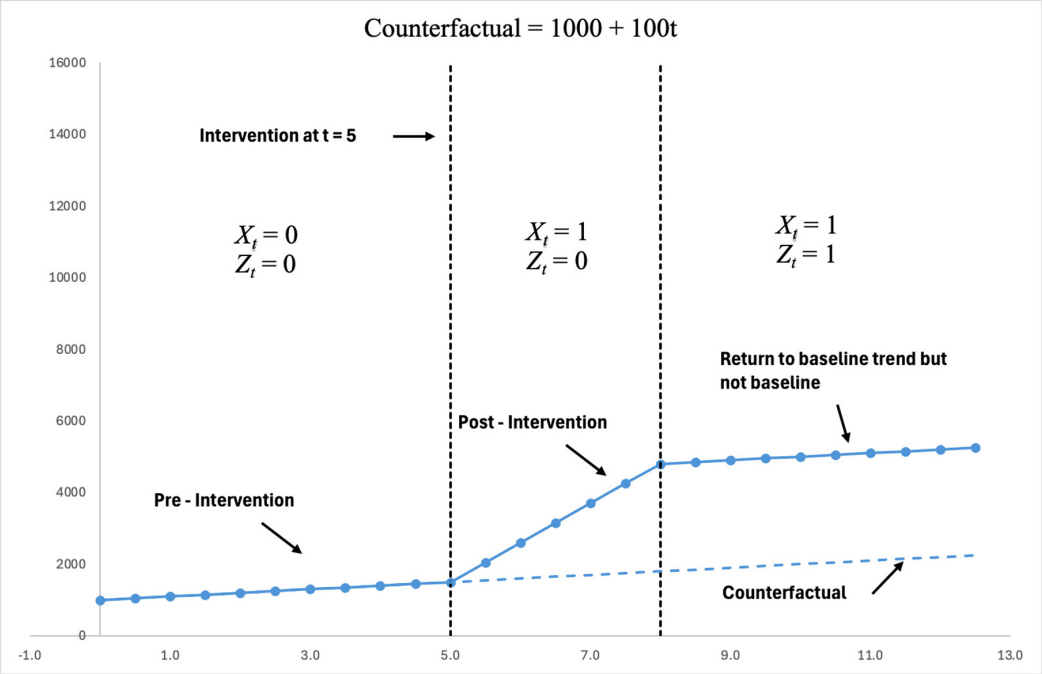
In this case, the intervention as t = 5 causes an upward shift in the outcome and then after 3 units of time, there is a return to baseline intercept and trend (the counterfactual).

Figure 2 (cont.)

$$Y_t = 1000 + 100t + (-5000 + 1000t)X_t + (8000 - 1000t)Z_t$$

$\beta 0 = 1000$
$\beta 1 = 100$
$\beta 2 = -5000$
$\beta 3 = 1000$
$\beta 4 = 8000$
$\beta 5 = 0$

$$X_t = \begin{cases} 0, t < 5 \\ 1, t \geq 5 \end{cases} \qquad Z_t = \begin{cases} 0, t < 8 \\ 1, t \geq 8 \end{cases}$$

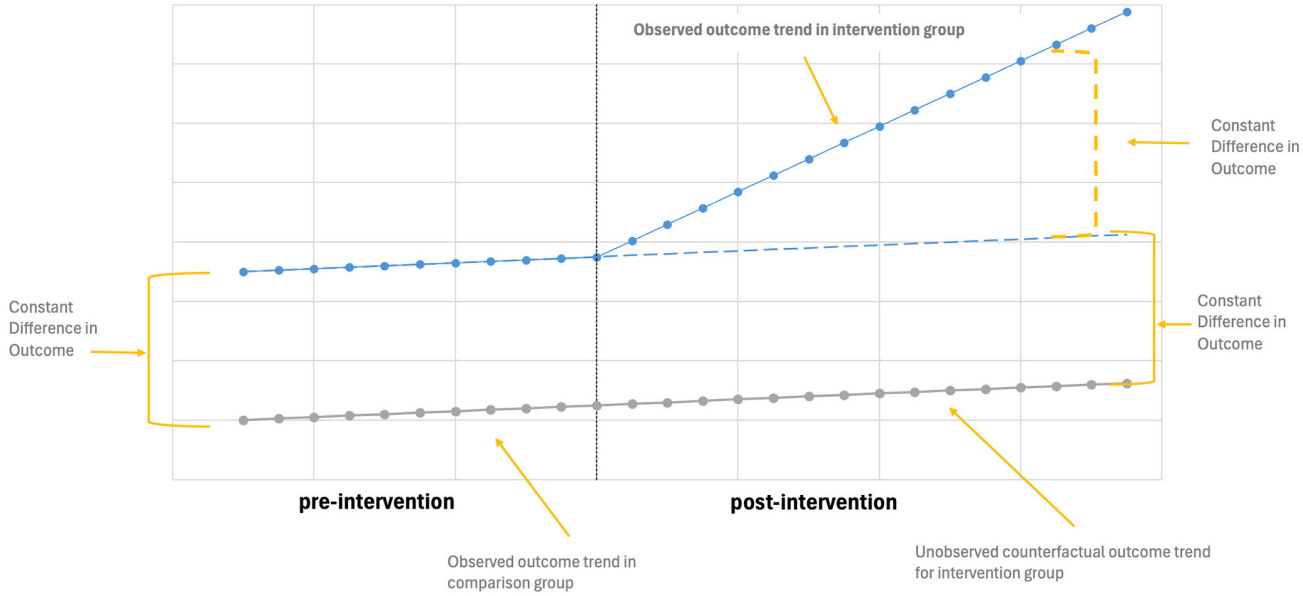


In this case, the intervention as t = 5 leads to no discontinuity in the outcome but causes a changes in the trend. After 3 units of time, there is no discontinuity but there is a return to baseline trend.

Difference in Differences (DID) design

Difference in Differences (DID) is a quasi-experimental design that makes use of longitudinal data from treatment and control groups to obtain an appropriate counterfactual to estimate a causal effect. Longitudinal data are observations of experimental groups made over a specified period of time. A counterfactual in the context of Difference in Differences is the extrapolation of the trendline in the treated group(s) that would maintain the same difference in outcome metric over time between the treated and non-treated group(s). DID is typically used to estimate the effect of a specific intervention or treatment, such as the passage of a law, enactment of a policy, or a large-scale program implementation. DID compares the changes in outcomes over time between a population that is impacted by a program (the intervention group) and a population that is not (the control group). This is a useful technique when there are attributable differences between individuals in the control and treatment groups. DID assumes that without intervention, observable differences between samples of a population are constant and after intervention, if the intervention had an effect, those observable differences should change, hence the term "Difference in Differences." A schematic is shown in Figure 3.

Figure 3 Difference in Differences (DID) Design Schematic



Propensity scoring

Propensity scoring is a statistical technique that creates a composite score for all the individuals based on selected characteristics. This technique is important for generating matched pairs of treated and untreated subjects when the experimental design cannot support the matching. When there is not random assignment, the sample may not have the same distribution as the population and, therefore, the average treatment effect for the entire population will not necessarily be the same as the average treatment effect for the treated subjects.

Consider the following:

$$Y_i(Z_i) = Z_i Y_i(1) + (1 - Z_i) Y_i(0) \quad (2)$$

$$Y_i = outcome\ variable\ for\ subject\ i$$

$$Z_i = binary\ variable\ which = \begin{cases} 1, if\ treatment\ occurred \\ 0, if\ treatment\ did\ not\ occur \end{cases}$$

The average treatment effect for the population, known as ATE, is given by:

$$ATE = E\big(Y_i(1) - Y_i(0)\big) \quad (3)$$

The average treatment effect for the treated group, known as ATT, is given by:

$$ATT = E\left(\big(Y_i(1) - Y_i(0)\big) | Z_i = 1\right) \quad (4)$$

In a non-randomized trial, there could be fundamental differences between the group(s) that received the treatment and the group(s) that did not, which would cause ATE≠ATT.

In an RCT, ATE and ATT are approximately equal, because—due to the randomization—there should be no difference in characteristics between the group of people treated and the group of people not treated.

The propensity score is the probability of the individual being selected into the treatment group based on observed covariates. Mathematically, the propensity score is expressed as:

$$e(x) = P(Z_i = 1 | X_i = x_i)$$

where Z_i is the treatment indicator for a subject, i, and X_i represents background variables, so the propensity score is the probability of treatment, given background variables (also known as covariates).

Propensity scores can be used to match participants from different groups.[22] For example, suppose the only covariates are age and education, with age broken down into two categories, old and young, and education into educated and uneducated. The "treatment" is implementation of highways with no speed limit. Suppose that there are three cities with the covariates shown below:

Figure 4: Propensity scoring—theoretical characteristics of different groups (cities)

City A	City B	City C
• 45% young and educated	• 45% young and educated	• 60% young and educated
• 15% young and uneducated	• 15% young and uneducated	• 10% young and uneducated
• 20% old and educated	• 20% old and educated	• 15% old and educated
• 20% old and uneducated	• 20% old and uneducated	• 15% old and uneducated

To compute the ATE, City A's outcomes would be compared to City B's outcomes but would not be compared to City C's outcomes, because City C's drivers have a different distribution of covariates.

22 "An Introduction to Propensity Scores: What, When, and How"; *The Journal of Early Adolescence*; 2014.

Propensity scores can also be used as a weight when working with data[23] and for stratification, where categories are created based on ranges of propensity scores and analyses are performed separately on the different strata.[24] Different models can then be applied to different strata. Propensity scores can also be used in regression equations. In the context of a regression equation, this method investigates whether the treatment variable matters when holding constant the likeliness of receiving treatment.[25] The regression becomes a weighted least squares regression. A form of this least squares regression is given by the following:[26]

$$Y = a + bZ + \sum_{j=1}^{n} c_j X_j + \varepsilon$$

where: Y = outcome,

$$Z = binary\ variable\ which = \begin{cases} 1, if\ treatment\ occurred \\ 0, if\ treatment\ did\ not\ occur \end{cases}$$

$$X_j = covariate\ j\ (can\ be\ something\ like\ an\ age\ group, smoker\ or\ nonsmoker, level\ of\ education, etc)$$

$$\varepsilon = error\ term\ which\ is\ independent\ of\ Z\ and\ X_j$$

The following regression weights are applied where *e(x)* as defined above is the propensity score:

$$\frac{1}{e(x)}\ for\ the\ treated\ subjects\ (Z = 1)\ and\ \frac{1}{1 - e(x)}\ for\ the\ untreated\ subjects\ (Z = 0)$$

Note the sum of the weights do not need to be one. For example, in many cases, weighted least squares sets weights as the inverse of the variance-covariance matrix. The inverse of the variance-covariance matrix is known as the information matrix.

Regression Discontinuity Design (RDD)

Regression discontinuity methods are non-randomized study designs that permit strong causal inference with relatively weak assumptions. The Regression Discontinuity Design (RDD) is implemented whenever a treatment is assigned based on some threshold value. This could include: antiretrovirals given to HIV+ patients when the CD4[27] count drops below a certain threshold, a cholesterol lowering medication is given when the low-density lipoprotein (LDL) to high-density lipoprotein (HDL) ratio exceeds a certain amount, and consumption of alcoholic beverages in people above the minimum drinking age. By

23 "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study"; *Statistics in Medicine*; Aug. 24, 2004.
24 "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score"; *Journal of the American Statistical Association*; Feb. 1, 1983.
25 "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation"; *Health Services & Outcomes Research Methodology*; 2001.
26 "Weighting regressions by propensity scores"; *Evaluation Review*; August 2008.
27 "CD" means "Cluster of Differentiation." CD4 cells are helper T-cells. T-cells are a type of lymphocyte, which is an immune cell. The CD4 helps to stimulate the B-cell to produce antibodies against the invading virus or other antigen, helps stimulate macrophages, and helps stimulate CD8 (cytotoxic, or killer T-cells) cells.

comparing observations lying closely on either side of the threshold, it is possible to estimate the average treatment effect in environments in which randomization is not feasible.[28] Consider the cholesterol- lowering medication example. The medication may be given to patients with an LDL-to-HDL ratio above 5.0. RDD would then consider patients with a ratio between 4.8 and 5.2. It can be reasonably assumed that most patients with ratio in this range are quite similar in terms of important covariates, but may face different outcomes for receiving the cholesterol-lowering medication. RDD requires that all other potentially relevant variables outside the treatment variable and outcome variable be continuous at the point where the treatment and outcome discontinuities occur. If the treatment assignment is "as good as random" at the threshold for treatment, then it guarantees that those who just barely received treatment are comparable to those who just barely did not receive treatment, as treatment status is effectively random. This is a sufficient condition and, in effect, simulates an RCT.[29, 30]

A nonparametric representation of a Regression Discontinuity Method can be expressed as:

$$Y = a + \tau D + \beta_1(X - c) + \beta_2 D(X - c) + \varepsilon$$

Where:

$$\tau = size\ of\ the \text{ jump due to the treatment effect}$$

$$D = \begin{cases} 1, if\ above\ the\ threshold \\ 0, otherwise \end{cases}$$

$$c = treatment\ threshold$$

$$\epsilon = error\ term$$

Note that there is an interaction term, D(X-c), to account for a shift in slope following treatment, similar to what occurs in interrupted time series analysis.

An example of a parametric representation of a Regression Discontinuity Method is shown as:

$$Y = \alpha + \beta_1 X + \beta_2 c + \beta_3 c^2 + \beta_4 c^3 + \varepsilon$$

Where:

$$X = \begin{cases} 1, if\ c > c_T \\ 0, otherwise \end{cases}$$

$$c_T = treatment\ threshold$$

28 "Regression-discontinuity analysis: An alternative to the ex post facto experiment"; *Journal of Educational Psychology*; 1960.
29 "Misunderstandings about the Regression Discontinuity Design in the Study of Close Elections" *Annual Review of Political Science*; 2016.
30 "Randomized experiments from non-random selection in U.S. House elections" *Journal of Econometrics*; February 2008.

Instrumental variables

Instrumental variables are variables that are correlated to the independent variable but uncorrelated to the dependent variable. They are used to estimate causal relationships when controlled experiments are not feasible. When explanatory variables are correlated with the error term in a regression equation, biased results can occur. The instrumental variable design aims to identify an exogenous variable that is correlated to the independent variable but is uncorrelated to the dependent variable. This variable is referred to as the "instrument." Subjects are assigned not to the key independent variable of interest, but rather to the instrumental variable.[31] This is important because in many of these "real-world studies," there are omitted variables that affect both the dependent and explanatory variables or there are situations in which changes in the dependent variable can change the value of at least one of the independent variables, also known as reverse causation. An instrumental variable allows a relationship to be established between the dependent and independent variables without bias that results from correlation between independent variables and error terms.

An application of instrumental variables is illustrated in the investigation of the impact of smoking on health. Health and smoking are impacted by a myriad of other shared factors, such as psychiatric illness, age, education, and demographics. Furthermore, smoking can be impacted by health. In this example, given it is presumed that smoking would be the independent variable and health the outcome variable, this would be considered reverse causation. A proposed instrumental variable that would reasonably be assumed not to be correlated to health by itself would be the tax rate on cigarettes. Significantly, the tax rate on cigarettes would be assumed to be highly correlated to smoking. Thus, if a health outcome, such as blood pressure, is considered a dependent variable and the tax rate on cigarettes an independent variable, if this relationship is found to hold in the data then by extension we can assert that health is impacted by smoking. That is, health and the tax rate on cigarettes,

31 *Natural Experiments in the Social Sciences: A Design-Based Approach*; Dunning, Thad; 2012.

by construct, are correlated through the effect of the cigarette tax rate on smoking. The tax rate on cigarettes is then considered an instrumental variable. To be a good instrumental variable, the potential instrumental variable should be highly correlated to the independent variable and uncorrelated with the error terms in the regression. Mathematically, instrumental variables can be implemented as a two-stage least squares regression.[32]

Suppose we have:

$$Y_i = X_i\beta + \varepsilon_i$$

where:

$$Y_i = dependent\ variable$$

$$X_i = independent\ variable$$

$$\varepsilon_i = regression\ error$$

$$\beta = regression\ coefficient$$

Then suppose we have:

$$X_i = Z_i\gamma + \omega_i$$

Where:

$$Z_i = instrumental\ variable$$

$$\gamma = regression\ coefficient$$

$$\omega_i = regression\ error$$

The matrix solution is:

$$\gamma = (Z^TZ)^{-1}Z^TX$$

$$X_{predicted} = Z(Z^TZ)^{-1}Z^TX = P_ZX$$

Where:

$$P_Z = Z(Z^TZ)^{-1}Z^T$$

Substituting $X_{predicted}$ for X yields the following:

$$Y_i = P_ZX_i\beta + \varepsilon_i$$

The matrix solution to this equation is:

$$\beta = (X^TP_ZX)^{-1}X^TP_ZY$$

32 *Two-stage predictor substitution for time-to-event data*; Master's thesis, University of Oslo Department of Mathematics, Simon Lergenmuller; Spring 2017.

Other Approaches to Causal Analysis

Koch's Postulates

Koch's Postulates were applied to the study of infectious diseases to determine the causative agent of pathogenic disease. The postulates are a set of action items to be completed to be sure that a microbe is the causative agent of disease. Specifically, Koch's Postulates are:[33]

1. The microorganism must be found in the diseased animal, and not found in healthy animals.
2. The microorganism must be extracted and isolated from the diseased animal and subsequently grown in culture.
3. The microorganism must cause disease when introduced to a healthy experimental animal.
4. The microorganism must be extracted from the diseased experimental animal and demonstrated to be the same microorganism that was originally isolated from the first diseased animal.

These postulates are used as criteria for determining whether the microbe is the etiology—that is, the cause—of a disease.[34] With more advanced medical knowledge and a greater understanding of the study of infectious disease, numbers 1 and 3 above have been shown to not always be true.[35] A very current example of this is the SARS-CoV2 virus, also known as COVID-19, which does not always cause symptomatic disease.

The set of action items in Koch's Postulates can be conceptually applied to other areas of scientific research to demonstrate a causative agent for an outcome.[36] For example, in medicine, a lesion is a generic clinical term for the presence of something that can cause a medical problem, where that something can be:

1. A microbe
2. A Prion, which is typically a misfolded protein that causes cellular and intercellular dysfunction,
3. An immune response (like a "tubercle" in tuberculosis),
4. A neoplasm/malignancy (cancerous growth), general inflammatory response,
5. A foreign object
6. Area of organ damage, etc.

Instead of "microorganism" in Koch's Postulates above, replace it with "lesion."

33 *Science*; Vol 351, Issue 6270, pp. 224-226; Jan. 15, 2016.
34 *Journal of Investigative Dermatology*; Volume 133, Issue 9, pp. 2141-2142; Sept. 1, 2013.
35 *Taxonomic Guide to Infection Diseases (Second Ed.)*; "Chapter 8—Changing how we think about infectious diseases"; 2019.
36 Dr. Mitchell Schaffler, personal communication.

When studying how bone loss occurs in osteoporosis, Koch's Postulates could be used to determine whether osteocyte (tissue resident bone cells) apoptosis (programmed or regulated cell death) is a causative agent of osteoclastic bone resorption. Osteoclastic bone resorption is when osteoclasts (specialized cells) resorb bone by effectively taking "Pacman bites" out of packets of bone.[37, 38, 39] A modified set of Koch's Postulates would apply when:

1. Osteocyte apoptosis is prevalent in osteoporotic bone in postmenopausal women and older men and in bone with overuse damage.[40]
2. Osteocyte-like cells in culture, when undergoing apoptosis upregulates osteoclastic bone resorption on bone like material.[41]
3. Activating osteocyte apoptosis by introducing microdamage leads to a spatial and temporal increase in bone resorption.[42]
4. Inducing microdamage in the experimental animal leads to increased osteocyte apoptosis, which is the same lesion that was found in animals with increased bone resorption.[43]

Inhibiting osteocyte apoptosis shuts off bone resorption in experimental animals.[44] The conclusion is that osteocyte apoptosis is a causal agent of osteoclastic bone resorption in bone remodeling and osteoporotic bone loss.

Can the action items or criteria for causation be applied in insurance losses? Consider this natural thought experiment under the framework of Koch's Postulates—the addition of side (curtain) airbags reduces medical costs and death in automobile accidents.[45] At the time they were introduced, many older cars did not have side airbags. Hence, we had a natural experiment of automobiles with and without side airbags. To place in the framework of Koch's Postulates:

1. No side airbags were found in automobiles where more fatalities occurred.[46]
2. Crash test dummies sustained more injury when there were no side airbags.[47]

37 "Loss of Osteocyte Integrity in Association with Microdamage and Bone Remodeling After Fatigue In Vivo"; *Journal of bone and mineral research: the official journal of the American Society for Bone and Mineral Research*; 2000.
38 "Spatial Distribution of Bax and Bcl-2 in Osteocytes After Bone Fatigue: Complementary Roles in Bone Remodeling Regulation?" *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*; 2002.
39 "Osteocyte apoptosis controls activation of intracortical resorption in response to bone fatigue"; *Journal of Bone and Mineral Research*; April 2009.
40 "The Implications of Osteocyte Biology for Bone Disease"; *Osteoporosis* (Third Ed.); 2008.
41 "Osteocyte Signals for Bone Resorption"; *Osteoporosis* (Third Ed.); 2008.
42 "Loss of Osteocyte Integrity in Association with Microdamage and Bone Remodeling After Fatigue In Vivo"; op. cit.
43 "Spatial Distribution of Bax and Bcl-2 in Osteocytes After Bone Fatigue: Complementary Roles in Bone Remodeling Regulation?"; op. cit.
44 "Osteocyte apoptosis controls activation of intracortical resorption in response to bone fatigue"; op. cit.
45 "Updated estimates of fatality reduction by curtain and side air bags in side impacts and preliminary analyses of rollover curtains (Report No. DOT HS 811 882)"; National Highway Traffic Safety Administration; January 2014.
46 Ibid.
47 "Vehicles that earn good side-impact ratings have lower driver death risk"; IIHS.org; Jan. 19, 2011.

3. Removal of side airbags from a brand of automobile that has them would result in more fatalities in that brand of automobile. This item is strictly a thought experiment. Most "real-world" natural experiments will not necessarily satisfy all 4 of Koch's Postulates due to practicality.
4. Lack of side airbags in the brand of automobile that normally has side airbags leads to more fatalities, as observed in step 1, in vehicles with no side air bags.[48] The results after evaluating outcomes in vehicle accidents clearly have shown efficacy of curtain side airbags.

Mendelian randomization

Mendelian randomization[49] is an increasingly popular methodological alternative to RCTs in establishing causation, particularly in medical literature. Mendelian randomization utilizes the essentially random nature of genetic inheritance and the independence of inheritance across genes to construct post hoc "treatment" and "control" groups.[50] Put another way, the expression of a gene or not provides researchers with an instrumental variable whereby, randomly but observably, some members of the population are, in some way, differentially exposed to the risk factor.[51]

In studies that utilize Mendelian randomization, researchers must first develop a theory where a particular genetic expression can act as a strong instrument for exposure to the particular treatment factor. Then, they must construct an observational dataset of a population consisting of both the outcome observation and the requisite genetic data for the members of the population. With the data gathered, the researcher can then use established instrumental variable modeling techniques to evaluate the strength of the causal relationship between the treatment and the outcome via the instrumental variable, genetics.[52]

Working through an example where Mendelian randomization provided a superior alternative to RCTs in establishing causation due to ethical and practical limitations, consider the relationship between moderate alcohol consumption and coronary heart disease (CHD).[53] In this case, researchers utilized genotypes that relate to alcohol metabolism, specifically genetic variations where alcohol is metabolized quickly, moderately, and slowly.[54] In the experimental setup, the researchers would expect that "if there is a biologically protective effect of alcohol on CHD risk then the slow oxidizers may be

48 "Crash Test Dummies Show The Difference Between Cars In Mexico And U.S."; *NPR*; Nov. 20, 2016.
49 "Mendelian randomization"; *Nature Reviews Methods Primers*; 2, 6; 2022.
50 "Mendelian Randomization: Using Genetics to Study Behaviors and Environments that Cause Disease"; Genomics & Precision Health; Centers for Disease Control and Prevention; June 29, 2022.
51 "Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians"; *British Medical Journal*; 2018.
52 "Mendelian Randomization: A Precision Public Health Tool for the COVID-19 Response"; Genomics & Precision Health; Centers for Disease Control and Prevention; July 20, 2021.
53 "'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?"; *International Journal of Epidemiology*, Volume 32, Issue 1, pp. 1-22; Feb. 1 2003.
54 Ibid.

expected to have a lower risk of disease, since any alcohol they drink may be less rapidly cleared from the system."[55] The researchers subsequently found this to be the case, although the relationship was relatively weak.[56] Thus, by looking at the way that genetics control relative exposure to alcohol due to metabolism speed, researchers can effectively construct an essentially randomized study as to the effect of moderate alcohol consumption on CHD.

It is reasonable to point out that the challenges of collecting genetic data on the population within a large observation study should make Mendelian randomization a useful alternative to RCTs. Nevertheless, there is utility to the approach, particularly in cases where ethical considerations of a positive treatment would prove challenging. An additional limitation stems from the need to have a well-established relationship between a genotype and risk factor exposure. Thus, the applicability of Mendelian randomization is limited both by our understanding of the genome available on and the extent to which exposure to a risk factor is controlled by genetics in a material way. Nevertheless, Mendelian randomization has been a critical tool in the evaluation of causal structures throughout the SARS-CoV-2 pandemic.[57]

Regulatory Implications and Perspectives on Natural Experiments

A September 2020 white paper from the National Association of Insurance Commissioners (NAIC), *Regulatory Review of Predictive Models*,[58] recommends that rate filing reviewers

> obtain a rational explanation for why an increase in each predictor variable should increase or decrease frequency, severity, loss costs, expenses, or any element or characteristic being predicted. ... The explanation should go beyond demonstrating correlation. Considering possible causation may be relevant, but proving causation is neither practical nor expected.

This recommendation recognizes the difficulty of demonstrating causation with predictive models but may not have contemplated the potential for natural experiments to provide causal support for modeling insurance relationships to risk.

55 Ibid.
56 Ibid.
57 "Mendelian Randomization: A Precision Public Health Tool for the COVID-19 Response"; op. cit.
58 *Regulatory Review of Predictive Models* white paper; NAIC Casualty Actuarial and Statistical (C) Task Force; 2020.

In lieu of demonstrating causation, the NAIC white paper endorses companies providing rational explanations for including each variable in a model. The white paper defines rational explanations as plausible narratives that connect model variables to the risk being modeled, in a way easy to understand by a consumer or other educated layperson. The explanation is intended to "establish a sufficient degree of confidence that the variable and/or treatment selected are not obscure, irrelevant, or arbitrary." While the intent of the definition of a rational explanation is to ensure arbitrary and spurious variables are not influencing rating decisions, there may be a lack of consensus among insurers and regulators as to what constitutes a rational explanation. What may seem rational to some may lack sufficient explanatory power to others. There are no guardrails for defining and assessing rational explanations. Additional clarification that might help operationalize the definition of rational explanation could include:

1. empirical research support,
2. alternative explanations, or
3. resolving conflicts among competing explanations.

A natural experiment approach could be a sound alternative to rational explanations or add further support. However, there may not be natural experiments to support every relationship between dependent and independent model variables. Where natural experiments exist, regulators may prefer their inclusion when compared to solely utilizing rational explanations.

It should be noted that the NAIC white paper was developed for property and casualty applications of predictive modeling. The property and casualty field has been applying predictive models to insurance problems for longer than the life, health, and retirement practice areas. As a result, the property and casualty field has tackled and resolved issues that other practice areas are only now facing. The NAIC white paper reflects learnings that may benefit the other practice areas, including the discussion of rational explanations.

Limitations and Risks of Natural Experiments

Natural experiments are an extremely useful methodology for investigating causal relationships when RCTs are not viable. However, natural experiments may not exist for all situations. They have their own limitations, shortcomings, and challenges that the practitioner must consider when developing a research approach.

First, the validity of a natural experiment is predicated on the extent to which the circumstances divide the population into a treatment and control group as though they were randomly assigned.[59] The practitioner must be careful in determining whether the circumstances of the natural experiment create a population split that falls close enough to this "as if" it were randomly divided standard.[60]

A second and related limitation comes in the treatment phase of the experiment, where due to the uncontrolled environment, the application of the intervention to the treatment group may be irregular or inconsistent, creating potential for additional confounding.[61] Put another way, does the lack of control over the timing or magnitude of the intervention to the treatment group affect the validity of the experiment? Or similarly, does the potential exposure of the control group to the treatment disqualify the findings?

Finally, a third area where natural experiments can pose a particular challenge to researchers is the limitation around data collection and data quality, as measurement is generally not considered in advance. This could mean that researchers find themselves in a position where a critical data point for analysis is omitted, partially censored, or otherwise of poor quality or reliability.

All of the issues above can fundamentally undermine the validity of an analysis or the integrity of a study and must be taken seriously.

Natural experiments have no formal design phase. They are essentially "happy accidents" that occur due to quirks in public policy, socioeconomic forces, and environmental events—all of which are outside the control of a researcher. Thus, natural experiments are subject to higher risks of confounding, as participants are not necessarily randomly assigned between the treatment and control. There is some other "natural pattern" at work that divides the population between the two courses.[62] Indeed, even identifying a valid control

59 "Improving Causal Inference: Strengths and Limitations of Natural Experiments"; *Political Research Quarterly*; June 2008.
60 Ibid.
61 "Introduction: why natural experiments?"; *Natural Experiments in the Social Sciences: A Design-Based Approach*; op. cit.
62 "The effect of changing the built environment on physical activity: A quantitative review of the risk of bias in natural experiments"; *International Journal of Behavioral Nutrition and Physical Activity*; Oct. 7, 2016.

group "becomes more difficult when evaluating policy interventions, as these are generally implemented acutely, and at the population level."[63] Ultimately, researchers must understand that any apparent natural experiment exists along "a spectrum, in which the assertion of 'as if' random assignment ranges from less to more plausible and valid."[64]

In the execution phase, the most glaring challenge to natural experiments is that the nature of the intervention is out of the researcher's control. In an RCT, the researcher can exert a high degree of control over the environment in which a treatment is delivered, the timing of the delivery, and the magnitude of the dose delivered. However, in a natural experiment, the researcher cannot control any of those variables. This lack of agency can lead to the irregular application of the treatment even within the treatment group.[65] Researchers must be cognizant of this potential for control group contamination, as it can invalidate the entire study.

A further challenge emerges during the analysis phase. Because natural experiments are not set up in advance, the data available to researchers for analysis tends to be limited and of variable quality. For example, consider having to rely on body mass index (BMI) instead of height and weight in a study of malnourishment and cognitive outcomes. The misalignment between available data and data desired in the natural experiment analysis can create logistical challenges in terms of data cleansing and staging, as well as more fundamental challenges when it comes to the strength and interpretability of the results. For example, in the previously mentioned case of John Snow's analysis of cholera outbreaks in London, one data element that was not captured in the initial analysis was that a large number of the households who did not experience a death had fled the city. Had this omission been apparent at the time, and setting aside the preponderance of supplemental evidence, this could have severely eroded the credibility of the conclusions Snow presented.[66]

The challenges presented above mean that natural experiments provide weaker evidence concerning the validity of a given causal hypothesis when compared to what would be possible in a more well-controlled setup, like an RCT. Nevertheless, the importance of understanding causality necessitates looking to tools that can provide some insight in cases where RCTs are impossible, impractical, or potentially unethical.

63 "Using natural experiments to improve public health evidence: A review of context and utility for obesity prevention"; *Health Research Policy and Systems*; May 18, 2020.
64 "Improving Causal Inference: Strengths and Limitations of Natural Experiments"; op. cit.
65 "Using natural experiments to improve public health evidence: A review of context and utility for obesity prevention"; op. cit.
66 *Visual Explanations: Images and Quantities, Evidence and Narrative* (pp. 27-37); Tufte, Edward; 1997.

Conclusion

As actuaries increasingly use broader datasets in their work, adopting new tools and methods may aid in evaluating the nature of the relationship between predictor variables and risk outcomes. The use of new data elements provided by third parties and more advanced techniques has also increased interest in understanding the causal structure between predictor and outcome variables of interest. The gold standard approach to evaluating causation, the RCT, is not always possible due to ethical, practical, or logistical considerations. Robust alternative methods for evaluating causation, such as natural experiments, may be useful.

The natural experiment enjoys a storied history in the domains of health, social, public, and economic policy—providing a critical path to causal analysis when a traditional experiment is not a viable tool for investigation. Moreover, there are a variety of statistical and methodological frameworks available, enabling a variety of different event patterns to serve as natural experiments.

Finally, it is important to note that, as of this writing, there is no regulatory requirement to establish causation. However, the NAIC has published a white paper to encourage actuaries to provide a "rational explanation" for relationships used in risk classification. Natural experiments can provide support for "rational explanations" of the causal relationship between variables.

The domain of natural experiments will be increasingly important. There are many opportunities for actuaries to become involved in frontline research to support innovation and advances.[67]

67 "Mendelian randomization"; op. cit.

1850 M STREET NW, SUITE 300, WASHINGTON, D.C. 20036
202-223-8196 | **ACTUARY.ORG**