

An Actuarial View of Correlation and Causation— From Interpretation to Practice to Implications

JULY 2022

Definitions

- Correlation is a type of association and measures increasing or decreasing trends quantified using correlation coefficients.*
- Causality is the empirical relation between two events, states, or variables such that a change in one (the cause) brings about a change in the other (the effect).
- The term “unfairly discriminatory” is used in the regulatory pricing sense regarding inadequate rates such that insurance rates are not permitted to be excessive, inadequate, or unfairly discriminatory.

* Altman, N., & Krzywinski, M. (2015). “Points of Significance: Association, Correlation and Causation.” *Nature Methods*, 12(10).



AMERICAN ACADEMY
of ACTUARIES

1850 M Street NW, Suite 300
Washington, DC 20036

202-223-8196 | www.actuary.org

Craig Hanna, Director of Public Policy

© 2022 American Academy of Actuaries. All rights reserved.

Introduction

What is the purpose and who is the audience for this issue brief?

Predictive models have become almost ubiquitous within risk classification and ratemaking practice today. The prevalence of these techniques has brought an increased focus on the significance of causation vs. correlation in risk classification applications and to the limitations of predictive models in differentiating between correlation and causation. This issue brief aims to provide a discussion of some key questions that actuaries may encounter as they work in the risk classification domain. The body of this paper starts by discussing the theoretical and practical differences between causation and correlation. It then turns attention to some of the challenges that actuaries may face as they evaluate the ramifications of correlation, such as the potential for unintended impacts from a model variable and the risk of spurious correlations. It concludes with a discussion of rational explanations as an alternative to proving “causation” in the work of risk selection, approaches for detecting spurious correlations, as well as a set of recommendations for future work and partnership between regulators and actuaries.

Why is this issue important?

The distinction between correlation and causation has grown in importance as insurance risk classification techniques grow in complexity and more and more data—including third-party data (“external data”)—becomes available for risk classification purposes. This data can be used by insurers to develop models that consider hundreds or thousands of potential rating variables, many of which may bear little intuitive relationship to the risk of an insured filing a claim in the upcoming policy period. There is significant public and regulatory concern related to the potential for these complex models to create unfairly discriminatory results for individuals within protected classes. Even if insurance practices

are blind to certain protected class information, unintentional impacts to individuals within protected classes can still occur. It has been argued that credit scores, level of education, and other rating variables are examples of this phenomenon.

The concepts of correlation and causation are highly relevant to the ongoing discussion of what rating variables can be used for risk classification purposes from a regulatory and public policy perspective. Two related questions explored in this paper are:

- When non-insurance data is shown to be predictive of insurance losses, what additional considerations should be included in the decision to rely on this data?
- What considerations should regulators and legislators have in determining laws and regulations for insurers to demonstrate that risk classification systems are not creating unfairly discriminatory results for protected classes?

Establishing Causation

What are the difficulties in establishing causation?

In a scientific context, causation is usually preferred to correlation because it answers “why” two variables interact in a predictable and rational manner. Recognized causal relationships include the link between smoking and lung cancer,¹ speed and the number of highway fatalities,² and hurricanes and roof damage.³ These examples imply an association, which is a necessary, but not a sufficient, condition for establishing causation. Causation means that a change in one variable causes a change in another variable. However, the converse of these relationships is not necessarily true: for example, lung cancer does not cause smoking.

¹ Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B., & Wynder, E.L. (1959). “Smoking and lung cancer: recent evidence and a discussion of some questions.” *Journal of the National Cancer Institute*, 22(1), 173-203.

² *Relationship between Speed and Risk of Fatal Injury: Pedestrians and Car Occupants*; Department for Transport (London); September 2010.

³ “[Understanding the Risks](#)”; Hurricane Retrofit Guide; FloridaDisaster.org.

This issue brief is a work product of the Data Science and Analytics Committee and the Property/Casualty Racial Equity Task Force. Its primary drafters were Dorothy Andrews, MAAA, ASA, CSPA; Steven Armstrong, MAAA, FCAS; Dave Heppen, MAAA, FCAS; and Julia Romero, MAAA, FSA, CERA.

A similar notion is correlation, which also implies association between two variables but not necessarily a causal relationship. It is possible to have causation but not correlation, and correlation but not causation. However, both causation and correlation imply association. As an aside, it is less intuitive how causation can exist without correlation, but one need only consider where, for example, a causative relationship of a nonlinear rather than linear nature. Correlation is a common measure of a linear or monotonic⁴ relationship between two variables and might assign a correlation of zero to a nonlinear or nonmonotonic causative relationship. One such example is the relationship between plant growth and watering. While watering aids plant growth, too much watering results in negative growth, producing a nonlinear relationship and the plant eventually dies. It is also possible to demonstrate mathematically that causation can occur without correlation.^{5,6}

Establishing causation is often much more difficult than establishing correlation, because the former typically requires rigorous scientific experimentation, while establishing correlation can be achieved through a straightforward statistical analysis. The randomized control trial (RCT), long considered the gold standard in establishing causation, is an experiment involving random assignments of participants to test and control groups. The test group participants receive a treatment, while the control group participants do not; the experimenters are blind to who received the treatment until the experiment is over and it is time to measure the effectiveness of the treatment.⁷ The blindness feature of RCTs is to prevent researcher bias from creeping into the results of the experiment.

Predictive models do not offer the same causative insights as randomized control trials. At best, algorithmic predictive models can demonstrate an associative relationship between model variables and outcomes. A relevant question to ask is whether correlation alone is an adequate measure for helping consumers and regulators get comfortable with the inclusion of variables in a model that cannot be substantiated as causative of loss. With respect to property and casualty (P&C) insurance, the National Association of Insurance Commissioners (NAIC) has provided guidance to regulators to consider adopting “rational explanations” to explain “why a rating variable is correlated to expected loss or expense, and why that correlation is consistent with the expected direction of the relationship.”⁸

4 Having the property either of never increasing or of never decreasing as the values of the independent variable or the subscripts of the terms increase.

5 “[Can Causation Exist Without Correlation? Yes!](#)”; Awaken Your Superhero; Aug. 10, 2018.

6 “[Causation without Correlation is Possible](#)”; The Incidental Economist; Dec. 16, 2009.

7 Hariton, E., & Locascio, J.J. (2018). Randomised controlled trials—the gold standard for effectiveness research.

BJOG: An International Journal of Obstetrics and Gynaecology, 125(13), 1716.

8 [Regulatory Review of Predictive Models](#) white paper; NAIC Casualty Actuarial and Statistical (C) Task Force; 2020.

The NAIC has deemed “rational explanations” as defensible narratives, in lieu of causation, to explain rating variable selections to consumers, legislators, and the media. As will be discussed in detail later in this issue brief, the definition of “rational explanation” excludes the requirement of causality, recognizing the limitation of predictive models to prove cause-and-effect relationships. A natural question to ask is what makes for a “good” rational explanation? Determining the criteria will require close collaboration between industry and regulators to ensure the interests of the public are served.

Correlation versus Causation

What is the difference between correlation and causation?

It is often assumed that when two events happen together, one caused the other. However, establishing that a statistical relationship exists between two variables is one thing, while establishing causation is quite another. In the context of insurance risk classification, there are well-established statistical techniques for determining the degree to which two variables, such as driver age and accident frequency, are correlated. However, there are no universally accepted approaches to establishing causation for insurance risk classification. Guidance on causality as provided in Actuarial Standard of Practice No. 12, *Risk Classification (for All Practice Areas)* (“ASOP No.12”), adopted by the Actuarial Standards Board in December 2005, states, “While the actuary should select risk characteristics that are related to expected outcomes, it is not necessary for the actuary to establish a cause and effect relationship between the risk characteristic and expected outcome in order to use a specific risk characteristic.”

What are potential issues resulting from the differences between correlation and causation?

Some states have introduced legislation that prohibits the use of certain rating variables unless various conditions can be met. When Colorado Senate Bill 21-169 (“SB 21-169”) was introduced on March 2, 2021,⁹ there was a requirement for insurers to demonstrate that external data, algorithms, and predictive models used for risk selection—to bear a direct causal relationship to insurance losses. This causation requirement was removed in the final bill that became law (Colorado Revised Statutes 10-3-1104.09, signed into law on July 6, 2021). Because causation is very difficult and potentially impossible to demonstrate for a given rating variable—even with a well-accepted variable such as driving experience in auto insurance—this requirement could preclude the use of any external data, algorithms, or predictive models in practice. To the extent causation standards are established for insurers, it will be very important to include explanations of what constitutes causation.

⁹ See “[Introduced](#)” bill.

An unintended consequence of implementing causation standards is that it could limit a number of variables associated with incentives for demonstrating low risk behavior. For example, implementing a causation standard might eliminate a discount for completing a driver's education course.

Can predictive models distinguish between correlation and causation?

No, because predictive models cannot address questions about causation, only correlation. Predictive models are typically designed and calibrated based on a review of the correlation between potential rating variables and historical insurance losses. Correlation analysis is purely statistical in nature and can be managed through sheer computing power. Causation analysis, on the other hand, requires more than just computing power. Consider an accident that occurs at night, in the rain, in which a 17-year-old driver loses control of the car and causes an accident with another vehicle. More than one factor may have contributed to the accident—the weather conditions, the time of day, and the driver's inexperience among them. All such factors can be statistically demonstrated to correlate to insurance loss propensity. But can any of them be demonstrated to be causal, within the model, in and of themselves?

Is there a way to balance the desire to demonstrate causation with the capabilities of predictive models?

In scientific studies, causation is typically demonstrated through randomized controlled testing. In a controlled experiment, testers can attempt to isolate individual explanatory variables and hold all other potential explanatory variables constant. If the dependent variable can be explained by the independent variable, then the independent variable is viewed as causal.¹⁰

Predictive models do not follow randomized controlled experimental designs. Randomized control trials are ways of creating data and predictive model are ways of analyzing data.^{11, 12} As a result, predictive models are not designed to demonstrate causation.

¹⁰ A dependent variable depends on other (independent) variables for its values. Independent variables do not derive their value from other variables, but are used to explain the value of dependent variables.

¹¹ "Be Careful When Interpreting Predictive Models in Search of Causal Insights"; *Towards Data Science*; May 17, 2021.

¹² Also, all predictive models implicitly assume that everyone will keep behaving the same way in the future, and therefore correlation patterns will stay constant.

There are no widely accepted minimum thresholds for determining a strong correlation between a proposed rating variable and insurance losses.¹³ Strong correlation between a rating variable and insurance losses indicates that the variable may be useful in predicting future insurance losses, even if it does not definitively demonstrate causation. This approach would have the advantage of being straightforward to demonstrate. A disadvantage is that the selected threshold would be arbitrary. This disadvantage may be outweighed by the relative simplicity of applying this approach in practice, as well as the ability to apply a consistent standard for all insurers in all states.

Empirical studies can often shed light on the strength of a relationship between variables, and recent award-winning research suggests natural experiments can answer questions regarding causation where it is either unethical or not possible to conduct randomized control trial experimentation.¹⁴

In summary, it is understandable that it is hard to distinguish causation from correlation. It is important to draw clear distinctions between these terms if insurers are required by regulators to address the differences between correlation and causation in demonstrating the appropriateness of model rating variables. Implementation of such requirements will likely require clearly defined thresholds, such as the amount of correlation needed for a rating variable to be deemed acceptable.

Rational Explanations Explained

Is establishing variable rationales the next best thing to establishing causation?

ASOP No.12 states, “The actuary should select risk characteristics that are related to expected outcomes. A relationship between a risk characteristic and an expected outcome, such as cost, is demonstrated if it can be shown that the variation in actual or reasonably anticipated experience correlates to the risk characteristic.”

As regulators, consumers, and other interested parties seek more from the actuarial profession on the concepts of causation (beyond correlation), regulators have proposed supporting model variables with “rational explanations” as an alternative and acceptable approach to supporting their inclusion in models.

¹³ Akoglu H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3), 91–93.
<https://doi.org/10.1016/j.tjem.2018.08.001>

¹⁴ Nobel Prize for Economic Sciences: “[Natural experiments help answer important questions.](#)”

The Sept. 9, 2020, version of the NAIC Casualty Actuarial and Statistical Task Force’s *Regulatory Review of Predictive Models* white paper¹⁵ adopted by the full NAIC in April 2021, recommends rate filing reviewers to:

“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. If no rational explanation can be provided, greater scrutiny may be appropriate. ... For example, the regulator should look for unfamiliar predictor variables and, if found, the regulator should seek to understand the connection that variable has to increasing or decreasing the target variable.”

“Rational explanation” is defined in the white paper as “a plausible narrative connecting the variable and/or treatment in question with real world circumstances or behaviors that contribute to the risk of insurance loss in a manner that is readily understandable to a consumer or other educated layperson. A ‘rational explanation’ does not require strict proof of causality but should establish a sufficient degree of confidence that the variable and/or treatment selected are not obscure, irrelevant, or arbitrary.”

The NAIC white paper goes on to say, “a ‘rational explanation’ can assist the regulator in explaining an approved rating treatment if challenged by a consumer, legislator, or the media. Furthermore, a ‘rational explanation’ can increase the regulator’s confidence that a statistical correlation identified by the insurer is not spurious, temporary, or limited to the specific datasets analyzed by the insurer.”

The NAIC white paper provides an example, “...if ‘murder’ or ‘theft’ are used to predict the wind peril, the company should provide support and a rational explanation for their use.” This reflects current themes of understandability and transparency in how risk classification works and is used to fairly price insurance, bringing the actuary’s obligation to the public into the forefront as well. It should be noted that rational explanations are subject to subjective interpretation, and to be actionable it will be important to develop clear criteria—whether through regulation, legislation, or otherwise—by which an explanation can be deemed rational. This would especially be the case for P&C insurers.

¹⁵ [Regulatory Review of Predictive Models—Executive \(EX\) Committee and Plenary](#).

Precept 1 of the actuaries' Code of Professional Conduct states, "an Actuary shall act honestly, with integrity and competence, and in a manner to fulfill the profession's responsibility to the public and to uphold the reputation of the actuarial profession." Rational explanations can enhance the actuary's work product.

Hence, while not explicitly required per any ASOP, the prospect of an actuary providing a "rational explanation" as a bridge between solely focusing on correlation and proving causation has merit. Actuaries and regulators—both of whom have shared responsibilities to the public—may want to contemplate "rational explanations" as an ongoing principle that will evolve and be refined over time.

Unintended Consequences of Model Variables

Is the use of some model variables causing unintentionally biased rates for certain protected classes? What can be done about it?

The use of model variables poses the risk of creating unintentional bias in rates or other insurance practices when they are correlated with attributes of membership in a protected class (e.g., race). The variables do not have to necessarily reflect protected class characteristics to create this bias. For example, income, ZIP code, education, and occupation appear neutral, but the use of these variables in insurance ratemaking can result in unintentionally biased results. Federally protected classes include people who can be classified based on race, color, religion, national origin, gender, age, or disability. Insurance companies, because of regulated restrictions, do not typically collect many protected class attributes, thereby making it difficult to assess with certainty how their models are performing based on protected class attributes.

Blinding models to the protected class characteristics is not a guarantee that disparate impacts will be removed. Amazon's automated hiring algorithm differentiated based on gender even though gender was not among the variables included in the model.¹⁶ The algorithm was trained on the resumes of Amazon's then current workforce which was decidedly male. The resumes of the few female employees could be distinguished from male employee resumes because of female nuanced words relating to women's colleges and sports teams and the lack of verbs more prominent in the resumes of men, such as "executed" and "captured." The algorithm found patterns in the resumes that reflected the company's biased hiring. It was eventually retired when it became clear it could not be remediated. According to one women's rights advocate, algorithms such as Amazon's "are not eliminating human bias—they are merely laundering it through software."¹⁷ Algorithms often replicate and amplify biases that are hidden in data and exist in society and can do so in the absence of protected class characteristics.

¹⁶ ["Amazon scraps secret AI recruiting tool that showed bias against women"; Reuters; Oct. 10, 2018.](#)

¹⁷ ["Why Amazon's Automated Hiring Tool Discriminated Against Women"; ACLU; 2018.](#)

Algorithms will find less intuitive proxies for discrimination when they are deprived of directly predictive traits, but by using race, insurers can be required to employ statistical models that isolate only the predictive power of non-suspect variables and remove the effects of race.¹⁸ This raises a question as to whether the inclusion of protected class characteristics as control variables can absorb the discriminatory effects in a model.

The Brookings Institution has made a number of public policy recommendations to mitigate algorithmic bias; among them was the development of a regulatory sandbox to conduct anti-bias experimentation.¹⁹ It is important to understand the extent to which the inclusion of protected class characteristics in a model can aid in removing bias. To date, there has been limited experimentation measuring the effect of protected class characteristics on insurance risk classification and pricing models, but any such experimentation could be conducted in a regulatory sandbox to ensure the integrity of the results.

Ultimately, it may be impossible to know whether a model is inappropriately affecting a protected class without including an identifier for the protected class in the model. For example, the clearest way that one could know the impact of race on a model is to include an indicator for race as a model variable. Some have called for this approach, including the Brookings Institution. It is also an approach advocated for by fair machine learning approaches.²⁰

What P&C regulatory and legislative activity is going on with respect to model rating variables?

Variables such as gender and credit history are commonly used in insurance risk classification and pricing models but are prohibited from use in certain states.²¹ California, Hawaii, Michigan, and Massachusetts are among the most restrictive states, banning one or more of the following in auto insurance pricing: gender, age, credit history, education, occupation, employment status, years of driving experience, and residential status.²²

U.S. Senator Cory Booker introduced a bill in 2020 to prohibit the use of non-driving-related factors in the setting of insurance rates.²³ Specifically, the bill proposes to ban the use of education, occupation, employment status, homeownership status, credit scores, gender, ZIP code, census tract data, marital status, previous insurer, and prior purchase of insurance to determine rates. Senator Booker has said, “The use of factors unrelated to an individual’s driving record to determine auto insurance rates and eligibility is unfair and hurts working families.”

¹⁸ “[Proxy discrimination in the age of artificial intelligence and big data](#)”; *Iowa Law Review*; 2020.

¹⁹ “[Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms](#)”; Brookings Institution; May 22, 2019.

²⁰ Dwork, C., & Ilvento, C. (2018). Fairness under composition. arXiv preprint arXiv:1806.06122.

²¹ “[Car Insurance Rating Factors by State](#)”; The Zebra; Jan. 10, 2022.

²² Ibid.

²³ “[Booker Introduces Bill to Prevent Automotive Insurance Discrimination](#)”; Cory Booker press release; Sep. 24, 2020.

Congresswomen Bonnie Watson Coleman and Rashida Tlaib introduced a companion bill in the U.S. House of Representatives. Watson Coleman and Tlaib have found that for some people, auto insurance costs can make up as much as 18% of their income.²⁴ These bills are indicative of the legislative attention that is being focused on insurance models and model rating variables.

More recently, Colorado Gov. Jared Polis signed Senate Bill 21-169, *Restrict Insurers' Use Of External Consumer Data*, into law, which prohibits insurers from “unfairly discriminating based on an individual’s race, color, national or ethnic origin, religion, sex, sexual orientation, disability, gender identity, or gender expression in any insurance practice; or using any external consumer data and information source, algorithm, or predictive model (external data source) with regard to any insurance practice that unfairly discriminates against an individual based on an individual’s race, color, national or ethnic origin, religion, sex, sexual orientation, disability, gender identity, or gender expression.” Oklahoma (House Bill 3186) and Rhode Island (House Bill 7230) introduced similar bills that are nearly identical to Colorado’s bill. Delaware is introducing a bill to address gender disparities in auto insurance pricing (Senate Bill 231).

What life regulatory and legislative activity is going on with respect to model rating variables?

As life risk classification practices have incorporated a broader variety of data types—such as public records, wearables, credit attributes, and professional licenses—regulators and legislative bodies are revisiting standards to ensure that these new variables are used in a intuitive manner and do not introduce unfair bias into risk classification processes. Much of the regulatory attention on these new data sources has been focused on their use in accelerated underwriting. However, it is important to bear in mind that accelerated underwriting is just one aspect of risk classification. Existing regulations from New York and Colorado place particular focus on the use of “external data sources” and “predictive models.” Specifically, these regulations focus on ensuring that customers have confidence and transparency into the ways in which these new data sources impact their underwriting outcomes.

Detecting Spurious Correlations

Telling the difference between spurious and “real” correlations

Spurious correlation is where there is no explanation, direct or indirect, for the relationship between the correlated variables. In practice, actuaries and data scientists work to ensure that the relationships they capture in their predictive models are not spurious. Some accidental correlations can be easily dismissed as statistical noise (e.g., the correlation between divorce rates in Maine and per capita consumption of margarine).²⁵

²⁴ “Sen. Booker’s PAID Act Looks To Eliminate Discriminatory Non-Driving Factors In Auto Insurance Pricing”; *Forbes*; Oct. 5, 2020.

²⁵ Additional illustrative examples can be found at <https://tylervigen.com/spurious-correlations>.

The practitioner is often faced with cases where the relationship between variables is at least somewhat plausible but there is not support in the research literature for the relationship and the practitioner has to decide whether to include the variables in a model. Spurious correlations are generally “artifacts of method and arise from factors such as sample selection bias; use of an inappropriate correlation coefficient; large sample size; or errors of sampling, measurement, and computation.”²⁶

A non-spurious correlation reflects a relationship between two variables that is “amenable to a proper causal interpretation” whether through direct causation or more indirect means such as “common or intervening causes.”²⁷ Modelers may encounter instances of non-spurious correlation where there is a well-established direct causal link between the two variables: for example, the correlation between incidence of mesothelioma and historical exposure to asbestos.²⁸ However, in other cases the practitioner may have to search more deeply to understand the relationship reflected in the association because the relationship is not directly causal. An example of a non-spurious indirect correlation may be the correlation between mortality and a familial history of breast cancer disease.²⁹ In this case, the medical literature tells us that there is a genetic component to many breast cancers, and that the presence of certain genes can be even more predictive than family history.³⁰ Thus, while a family history of breast cancer certainly cannot be said to “cause” future breast cancers, the correlation between the two is non-spurious as it reflects a relationship between two factors that are linked.

Are there tests for detecting spurious results?

There are several ways of detecting spurious correlations to prevent faulty conclusions in data analysis.

Tests of Causality

1. Check for the directionality of the two correlated variables. Does variable A precede variable B or does variable B precede variable A? If a relationship fails to meet either one of these conditions, then the relationship may be spurious. Let’s say data show that total fire damage amounts increase as the number of firefighters at the scene increases.³¹ If the antecedent is the number of firefighters, is it logical that the dollars of fire damage is the consequent, or vice versa?³²

²⁶ *Encyclopedia of measurement and statistics, Vol 3.*

²⁷ Ibid.

²⁸ “Impact of asbestos on public health: a retrospective study on a series of subjects with occupational and non-occupational exposure to asbestos during the activity of Fibronit plant (Broni, Italy);” *Journal of public health research.*

²⁹ “Breast Cancer Facts & Figures 2019-2020”; American Cancer Society; 2019.

³⁰ “A Population-Based Study of Genes Previously Implicated in Breast Cancer”; *New England Journal of Medicine*; 2021.

³¹ “APA Dictionary of Psychology: Spurious Correlation.”

³² The antecedent is the if-clause of a cause-and-effect hypothetical proposition, while the consequent is the then-clause.

2. Identify a third mediating variable that explains the relationship between the two variables and that would reduce the statistical significance between the first two variables. For example, in the fire example, the mediating variable might be the initial size of the fire, which affects both the damage amount and the number of firefighters called to the scene.
3. Examine the residuals (i.e., prediction errors). If the residuals exhibit autocorrelation (i.e., correlated with a lagged set of the same residual stream), then some key variable is likely missing from the analysis that is likely mediating the relationship between the two variables. Autocorrelation of residuals can be checked using the Durbin-Watson test, which measures the extent to which residuals are successively correlated.³³ For example, define a simple linear regression as fire damage amount as a function of the number of firefighters. If the residuals fail the Durbin-Watson test, then the correlation between fire damage amount and number of firefighters is likely spurious.

Detecting Spurious Correlation

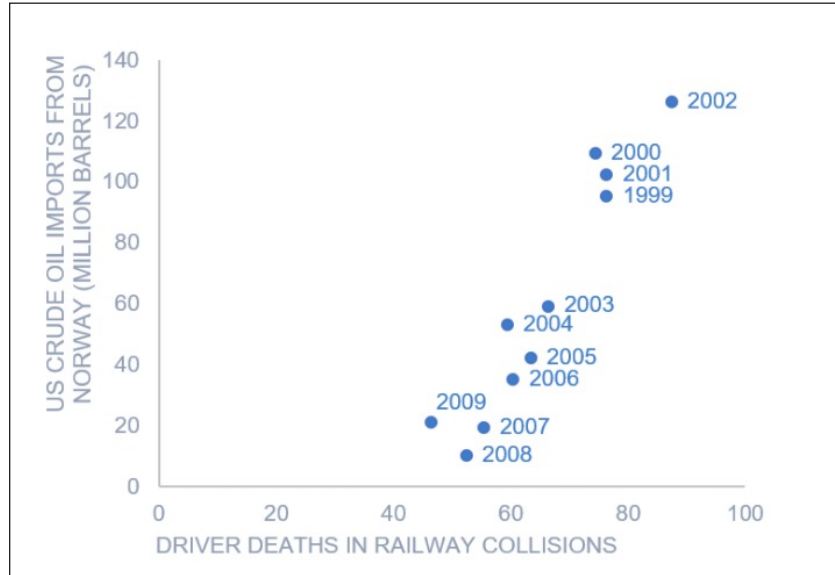
1. In linear regression, if a variable is statistically significant in the absence of other variables but becomes statistically insignificant in the presence of other variables, then it is likely the variable has a spurious relationship with the dependent variable. In classic regression modeling, the relationship between the number of breeding stork pairs in Europe (independent variable) and country-specific human birth rates (dependent variable) was found to have a correlation of 0.62, which was statistically significant with a p-value < 0.05 .³⁴ But when country size entered the equation, the stork variable became insignificant. Larger countries have more babies and more storks.³⁵
2. Examine the quality of the theory behind the correlated variables. Is there good reason to believe, as validated by research, the variables would occur together? If such validation does not exist, then the relationship may be spurious. For example, is there any validation to the relationship between the number of driver deaths in railway collisions by year (the horizontal axis), and the annual imports of Norwegian crude oil by the U.S., as depicted below?³⁶ This is an example of a spurious correlation. It is not clear what a rational explanation would be for this relationship.

³³ “14.3—Testing and Remedial Measures for Autocorrelation”; STAT 501; Penn State Eberly College of Science; 2022.

³⁴ “Storks deliver babies (p= 0.008).” *Teaching Statistics*; 2000.

³⁵ Ibid.

³⁶ <https://www.displayr.com/learn-what-is-spurious-correlation/>



3. Check for dependence between the two variables. Dependence means the occurrence of one variable is accompanied by a change in the other. Also, if one variable does not occur, then the other variable also does not occur. On the other hand, independence between two variables means one variable can occur without the other variable occurring and this may be an indication of a spurious relationship. For example, crude oil imports are not dependent on driver deaths in railway collisions, or vice versa.
4. Testing the model on out-of-sample holdout data for predictive models developed for risk classification purposes. If the relationships hold on the holdout dataset, that may provide some evidence that the relationships among the variables are non-spurious. When working with time series data, it is often a good practice to create an out-of-time holdout set that falls at the very end of the observation period, to ensure that the model is not being overfitted to the specific historical period of the training set.

How to avoid spurious correlations?

In practice, it is inevitable that one will encounter spurious correlations. However, being vigilant and applying common sense modeling habits while constructing and validating models can help reduce the risk that a spurious correlation makes it “out of the lab.” Starting in the data-gathering phase, the modeler should take care to review the relevant domain literature to ensure that only relevant variables, where plausible relationships may exist, are included among potential predictor variables. Moreover, as spurious correlations are often the result of datasets that are too small or where the features are improperly specified, it is critical that the modeler ensure that there are enough data to support all the features included in the model and that the features are well formed for the purpose and structure of the model. There is also a risk of spurious correlations when dealing with

large datasets if there are too many candidate variables. The modeling team should also take care, both during data set construction and model development, to directly inspect and interrogate the correlations both between predictors and outcomes as well as among the predictor variables.

Next Steps

The questions raised in this issue brief provide a context and framework for understanding the limitations of predictive models to explain causation and why correlation is a viable mechanism for supporting the inclusion of predictive variables in insurance risk classification models. However, when selecting model variables, it may be useful to consider the rational basis for their inclusion. A plausible relationship between model variables and the risk being insured may aid in promoting transparency and understandability to consumers, regulators, and other laypersons.

The above discussion suggests several possible actions requiring collaboration between insurers and regulators. They are:

Establish criteria for judging an acceptable rational explanation. The guidelines need to be clear, concrete, and actionable to allow for an objective evaluation of model variables against measures of the fairness and non-discriminatory treatment of consumers in insurance risk classification plans.

Establish regulatory minimum acceptable thresholds for correlations between independent variables, as well as between independent variables and dependent variables. Generally, correlations between predictor and target variables should be high while correlations among predictor variables should be low, depending on the machine learning approach used. Generalized linear models perform better when there are low correlations among predictor variables, which mitigates parameter instability.

Investigate how natural experiments might be used to establish cause-and-effect relationships between predictor and target model variables. Research is emerging that points to the potential use of natural experiments in insurance modeling.^{37, 38}

Build models in a regulatory sandbox to test whether the inclusion of protected class characteristics can identify bias in the outputs of rating models and safeguard protected class data from misuse.

³⁷ [“The impact of health insurance on health”](#); *Annual Review of Public Health*; April 2008.

³⁸ [“Do traffic tickets reduce motor vehicle accidents? Evidence from a natural experiment”](#); *Journal of Policy Analysis and Management*; Winter 2015.

While these are ambitious action items, they are achievable. Regulatory collaboration is essential to formulating solutions that mutually satisfy the objectives of insurers and regulators, and obligations to the public.

The work of the American Academy of Actuaries Property and Casualty Racial Equity Task Force, the Casualty Practice Council, the Life Practice Council, the Health Practice Council, and the Data Science and Analytics Committee will continue to investigate and analyze issues of how rating classifications may impact consumers.

References

“Beware Spurious Correlations.” Harvard Business Review, June 4, 2015.

<https://hbr.org/2015/06/beware-spurious-correlations>

Crossman, Ashley. (2021, January 14). What It Means When a Variable Is Spurious.

Retrieved from <https://www.thoughtco.com/spuriousness-3026602>

Goodman, R. (2018). Why Amazon’s Automated Hiring Tool Discriminated Against Women.

ACLU. Retrieved from <https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against>

Kenton, Will. “Spurious Correlation Definition.” Investopedia. Investopedia, October 14,

2021. https://www.investopedia.com/terms/s/spurious_correlation.asp

Lee, Ines. “4 Reasons Why Correlation Does Not Imply Causation.” Medium. Towards

Data Science, April 18, 2021. <https://towardsdatascience.com/4-reasons-why-correlation-does-not-imply-causation-f202f69fe979>

Haig, Brian. “Spurious Correlation.” In Encyclopedia of Measurement and Statistics, edited by Neil J. Salkind and Kristin Rasmussen. Sage, 2007.

https://www.researchgate.net/publication/315643991_Spurious_correlation

Vigen, Tyler. “15 Insane Things That Correlate with Each Other.” Spurious Correlations,

2015. <http://www.tylervigen.com/spurious-correlations>

Visonà, S. D., Villani, S., Manzoni, F., et al. (2018). Impact of asbestos on public health:

a retrospective study on a series of subjects with occupational and non-occupational exposure to asbestos during the activity of Fibronit plant (Broni, Italy). Journal of public health research, 7(3), 1519. <https://doi.org/10.4081/jphr.2018.1519>

(online access: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6321947/#>)

American Cancer Society. Breast Cancer Facts & Figures 2019-2020. Atlanta: American Cancer Society, Inc. 2019. (Accessed Online 11/20/21; <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf>)

Hu, Chunling, Steven N. Hart, Rohan Gnanaolivu, Hongyan Huang, Kun Y. Lee, Jie Na, Chi Gao, et al. “A Population-Based Study of Genes Previously Implicated in Breast Cancer.” *New England Journal of Medicine* 384, no. 5 (2021): 440–51. <https://doi.org/10.1056/nejmoa2005936>

The American Academy of Actuaries is a 19,500-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.