DigitalCommons @WayneState

Journal of Modern Applied Statistical Methods

Volume 14 | Issue 1 Article 16

5-1-2015

Modeling Probability of Causal and Random Impacts

Stan Lipovetsky
GfK, stan.lipovetsky@gfk.com

Igor Mandel
Telmar Group Inc., igor.mandel@gmail.com

Modeling Probability of Causal and Random Impacts

Cover Page Footnote
Acknowledgements: we thank Drs. I. Lipkovich for fruitful discussions, S. Brodsky for help in planning experiments, V. Levin for programming, and L. Sergeeva for editorial assistance.

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 180-195.

ISSN 1538 − 9472

Modeling Probability of Causal and Random Impacts

Stan Lipovetsky
GfK Custom Research North America
Minneapolis, MN

Igor Mandel
Telmar Group Inc.
New York, NY

The method of the estimation of the probability of an event occurring under the influence of the causal and random effects is considered. Epistemological differences from the traditional approaches to causality are discussed, and a new model of the statistical estimation of the parameters of each effect is proposed. The simple and effective algorithms of the model parameters estimation are presented, and numerical simulations are performed. A practical marketing example is analyzed. The results support the validity of the estimation procedure and open the perspective for the application of the method for various decision making problems, where different causes can yield the same outcome.

Keywords: causal and random effects, categorical data, causal modeling

Introduction

Modern decision making actively uses statistical methods, but there is one paradoxical aspect in it. To apply the results of statistical modeling and forecasting in practice, a decision maker, or a manager should be sure that the decision is based on a causal relationship: for instance, a positive correlation between advertising and sales could mean that it makes sense to increase spending on advertising for getting higher revenue. However, most of the statistical methods do not produce “causal models”, they only agree that “correlation is not causation”. For instance, Leo Breiman (2001) emphasized the indifference of the statistical learning to causal problematics (see also Hastie, Tibsharani, & Friedman 2009). So, a positive relationship between advertising and sales may simply indicate that with bigger sales, a company has a higher profit and thus is able to spend more on advertising. More questions related to statistical and causal approaches in sociosystemics and mediaphysica are considered in (Kuznetsov & Mandel, 2007, Mandel, 2011).

Email Mr Lipovetskty at: stan.lipovetsky@gfk.com. Email Mr. Mandel at: igor.mandel@gmail.com.

The past two decades have witnessed a burst of works on various causality problems and methods. Three main approaches intensively used in causality studies are: simultaneous structural equations founded by S. Wright (1921, 1960; more references within Kline, 2010); potential outcomes proposed by J. Splawa-Neyman (1990), and advanced by D. Rubin (1974, 2006), and the concept of do-operators and associated with them acyclic graphs developed by J. Pearl (2000, 2013). There are many other authors and proposals combining and modifying these ideas, although according to J. Pearl, almost all of these approaches in fact talk about the same things, using different terms and stressing different aspects of the problem. One thing is common for most of these works is that they consider a situation when many variables are interlinked, and the main aim of the causal analysis consists in disentangling of these influences and evaluating the pure impact of each cause on the effect. For instance, in the influential J. Pearl's book (2000), all descriptions begin only when graphs have complex structures, with several arrows targeting each node, but it is not clear what to do, if there is only one outcome and many potential causes.

While most applications of causal inference focus on a complex situation with multiple outcomes, the current paper revisits a seemingly simple case of a single binary outcome variable with multiple sources of causal and additional random effects. Randomness is understood here not as a "remaining part" of the unexplained variance, which is typical in statistics, but as the source of the unknown (not associated with any variables) causes, resulting in the same effect. This concept and a general model was proposed in (Mandel, 2013), where one can also find a discussion about the correct definition of causes and effects, the differences between individual and statistical causes, and other methodological issues, partly touched on here. This current paper considers the problems of the parameters estimation in such a model.

The Concept of the Causal Intrinsic Probabilities

Consider a model of the direct impact of multiple causes onto the binary outcome Y with $Y = 1$ and $Y = 0$ meaning that the effect of the interest has occurred or has not, respectively. Consider a case of K attributes $A_1, A_2, ..., A_K$ (where $A_k = 1$ and $A_k = 0$ denote the presence and the absence of a k-th attribute, with $k = 1, 2, \ldots, K$). The attributes are represented by the categorical variables which may be binary, ordinal, or nominal variables. A vector of the realized values of such attributes can be denoted as $a = (a_1, a_2, .., a_K)$, and this may represent levels of the same and/or different categorical variables, e.g., $A_1 = 1$ means male, $A_2 = 1$ means female, $A_3 = 1$ means kids, $A_4 = 1$ means teenagers, $A_5 = 1$ means adults, etc. Let us assume that the attribute A_k creates the causal effect $Y = 1$ with probability p_k. In the simplest case $k = 2$, the probability that $Y = 1$ would follow the rule of the union of the independent events: $S = p_1 + p_2 - p_1 * p_2$. In essence, it just reflects the fact that the coincidence of two causes does not produce anything more than one effect. Respectively, the probability of not having the causal effect would be presented as $1 – S = (1 - p_1)(1 - p_2)$.

For any K, the causal effect of outcome $Y = 1$ is defined as an intrinsic (latent) probability $S_{\text{causal}}(a)$, where a is a vector of the realized set of attributes, so that the probability of not-occurring of the event is:

$$1 - S_{causal}(a) = \prod_{\{k:\, a_k = 1\}} (1 - p_k) \tag{1}$$

where p_1, p_2,.., p_K are parameters which represent the causal strength associated with the presence of each attribute A_k. Note that the absence of an attribute may imply the presence of the opposite attribute (e.g., the absence of the “male” attribute A_1 contributes to the presence of the “female” attribute, A_2). In other situations it could vary: for instance, a road accident may happen due to fog (A_1), reckless driving (A_2), ice conditions (A_3), and other non-mutually exclusive causes.

There is also an irreducible latent probabilistic "random cause" that represents other factors that are not explicitly accounted for by the set of attributes. It is assumed that this random effect is: a) independent of other attributes; b) its outcome (denoted as r in the sequel) is constant across all configuration of attributes that may be present for a particular individual. These assumptions yield the expected probability at the population level as the union of the causal and random sources, $S(a) = S_{causal}(a) + S_{random} – S_{causal}(a) \cdot S_{random}$, or in the explicit form:

$$\begin{aligned} S(a) &= S_{causal}(a) + r - r \cdot S_{causal}(a) \\ &= 1 - (1 - r)\left(1 - S_{causal}(a)\right) \\ &= 1 - (1 - r) \prod_{\{k:\, a_k = 1\}} (1 - p_k) \end{aligned} \tag{2}$$

The aim of the proposed causal model is the estimation of $K + 1$ parameters, p_1,..,p_K, and r, on the basis of the sample of the realized outcomes $Y_i(a) = \{1, 0\}$ and the associated attribute vectors.

Concerning the motivation for the model, we can see the following arguments. Our setup acknowledges the *asymmetric* nature of causality, and the model (1)-(2) for intrinsic (causal) probability assumes that a *single* cause is sufficient for an event to happen ("fire"), whereas for an event not to occur, all potential causes should be ineffectual. It can be seen in a diagram with parallel pathways, where at least one of them would fire the event. It contrasts with a common binary logistic regression, where all the attributes contribute additively to the probability of the event occurring, or not occurring. Also, the model assumes that a random cause is *irreducible* and is presented within the sample probabilities $S(a)$. Finally, in the considered model, the main role is played by the *presence of attributes*, rather than by the changing levels of the factors in classical methods based on the concept of regression, potential outcomes, and other models.

Thus, each cause works as an independent entity and is associated not with the whole variable (like a binary “gender”), but with the separated levels (grades) of the variable (like two variables of “males” and “females”). It is different from the traditional statistical way of making models: one should look at these “grades related yields” rather than at the coefficients of general association (or regression), linking the whole “gender” to the outcome. Each level of the potentially causal variable produces an outcome with its own intrinsic probability. And if there are some causes, which cannot be associated with any measured variables, but still produce the outcome, then we relate them to the random cause. A typical example of such random causes is as follows: customers can buy a product regardless of advertising or promotions (a “baseline” which is hard to estimate). The purpose of the causal analysis is to evaluate the intrinsic probabilities, or the parameters of the outcome $Y = 1$ generated by different causes, including the random ones, with the observed data.

Causal Analysis and Parameters Estimation

The causes and the effect are associated with the usual statistical variables. Consider a data set containing variables X – the attribute causes of the outcome variable Y. With categorical causal variables, each grade of a causal X variable has its probability of generating the occurrence of the event, or the value $Y = 1$ in the outcome. A categorical variable with n grades can be presented as a set of n binary variables x_1, x_2, …, x_n, or the so-called Gifi-system (Gifi, 1990; Lipovetsky, 2012), where each j-th of these binary variables has 1s in the positions of j-th grade, and 0s in other positions. It allows the estimation of the causal effect only for values 1 for each variable, and the random cause can also have the impact inducting the appearance of the event $Y = 1$. So, the outcome $Y = 1$ occurs as a union of the independent events coming from two different sources – those associated with the measured variables and random noise (2).

As an explicit example, consider data with three x, so in total there are eight cells of all combinations of their values, and in each cell we find the proportions S_i of the outcome variable $S(a)$, so the proportion of $Y = 1$ in the base size of each cell. The cells and corresponding proportions S_i are presented in Table 1. Of course, in a particular real data set, some cells can be empty. The variables in Table 1 are orthogonal (see in Appendix A), so they are statistically independent.

Table 1. Example of data set with three binary variables.

i	x_1	x_2	x_3	S_i
1	0	0	0	0.09141
2	1	0	0	0.73409
3	0	1	0	0.25630
4	1	1	0	0.80300
5	0	0	1	0.57608
6	1	0	1	0.86570
7	0	1	1	0.63409
8	1	1	1	0.89563

In a general case of many variables, each presented via the Gifi-system of binaries with their total number of K variables, model (2) can be presented in a generalized form:

$$S_i = 1 - (1-r)\prod_{k=1}^{K}(1-p_k)^{x_{ik}} \tag{3}$$

where $k = 1, 2, \ldots, K$ is a number of variable x_{ik} identifying the k-th parameter of the probability in the i-th cell ($i = 1, 2, \ldots, N$). The values of x_{ik} are 1 or 0 when the variable is presented or not, respectively, as in Table 1. So, for the i-th value S_i, with values $x_{ik} = 1$, the term 1-p_k enters the product in (3), and for values $x_{ik} = 0$ the term 1- p_k is absent in the i-th row of the data. The cells as the new units are denoted by the current index $i = 1, 2, \ldots, N$. The relations (3) show that if any probability p_k or r is close to 1, the total probability of event occurrence S_i is close to 1 as well. This system corresponds to the feature of the independence of the variables' levels when the value of union S_i is defined by the criterion of "at least" one variable impacting on the event appearance. It is important to note that due to the definition (3), any additional cause with the term (1-p_k) can only increase S_i, as can be expected.

Consider how to estimate parameters of the model (3) by data like those given it Table 1. Regrouping and taking logarithm of equation (3), and using notations

$$y_i = \ln(1-S_i), \quad b_0 = \ln(1-r), \quad b_k = \ln(1-p_k) \tag{4}$$

we represent (3) in the linearized form:

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_K x_{iK} \tag{5}$$

So, the problem of estimation of the parameters b_k is reduced to the ordinary least squares (OLS) linear regression, with the known solution

$$b = (X'X)^{-1} X'y \tag{6}$$

where y (4) is a vector of N-th order, X is the design matrix of x_{ik} values (completed by the additional column of all 1s, which corresponds to the intercept b_0 in the model), b is the vector of all $K+1$ parameters in (5). If there are not enough observations, the matrix of the second moments $X'X$ in (6) could be close to singular, and its inversion is impossible, or it yields too inflated coefficients. In such a case, we can use a regularization imposed onto the parameters which produces the so-called ridge-regression:

$$b = (X'X + qI)^{-1} X'y \tag{7}$$

When the profile parameter of the ridge regression q is close to zero, the solution (7) reduces to OLS (6). More complicated ridge-regressions with a high quality of fit see in (Lipovetsky, 2010).

By the estimated coefficients b (6)-(7), each original parameter of probability can be obtained from the relations (4) by transformation:

$$r = 1 - \exp(b_0), \quad p_k = 1 - \exp(b_k) \tag{8}$$

The relations (8) show that the parameters b should be negative which can be achieved by their special parameterization (for instance, each b is substituted by another unknown parameter c in the relation $b = -c^2$, and a nonlinear estimation is performed for the free parameters c). But usually the solutions (6)-(7) are feasible for the meaningful values (8).

To illustrate this approach, return to Table 1, take $y_i = \ln(1 - S_i)$ as the dependent variable (4), and construct the model (6). Its coefficients are presented in the first column of Table 2. These coefficients are transformed by (8) to the probabilities r of the random impact and p_i of the causes, which are given in the second numerical column in Table 2. In the next column, Table 2 also presents the original values of cause probabilities used in this simulated data. Comparison of the estimated and the original values shows a pretty good quality of the estimation with the relative errors of several percent or less shown in the last column of Table 2. The coefficient of multiple determination in this model (6) equals 0.998, and its value adjusted by degrees of freedom equals 0.995, so the quality of the model is indeed very high.

Table 2. Regression coefficients and probability estimates.

Coefficients of regression		Estimates of cause probability		Original values used in simulation	Relative error, % to original values
b_0	-0.102722	r	0.09762	0.10	2.38
b_1	-1.240261	p_1	0.71069	0.70	1.53
b_2	-0.224870	p_2	0.20138	0.20	0.69
b_3	-0.697456	p_3	0.50215	0.50	0.43

Note that a design matrix like in Table 1 is orthogonal, so the x-variables have zero correlations. In such situation, coefficients of multiple linear regression equal the coefficients in the pair regression of y on each x separately, which makes calculations even simpler, as shown in Appendix A. If a cell of certain variables' combination is empty, the number of rows in the design table can be reduced. But even in such a case, it is possible to hold the whole design matrix substituting zero by a small proportion value, say, $S = 0.005$.

In application, the interest may be in estimating an additive share of influence of each cause in the effect. In order to achieve this, use the formula:

$$S_{ik} = S_i \frac{\ln\left(1 - p_k\right)^{x_{ik}}}{\ln\left(1 - S_i\right)} \tag{9}$$

where the total of the causes (including the random one corresponding to the index $k = 0$) in each cell equals the predicted proportion:

$$S_i = \sum_{k=0}^{K} S_{ik} \tag{10}$$

The derivation and other properties of the relations (9)-(10) are presented in Appendix B.

Methodology

Numerical simulations

To test validity of the proposed estimation procedure, a series of experiments on the generated data were performed. The varied parameters are described in Table 3. Not all combinations of these parameters (there are about 1700 scenarios) were estimated, some of them are simply impossible. For each combination of factors, several random runs (from one to forty) were performed. In each case, the assignment of value 1 to Y was done, if any of X variables was equal 1. For correlations in Table 3, both signs were used; correlation -1 means that two variables represent in fact one binary variable.

Table 3. Different parameters of simulation and estimation

	Parameters	Value 1	Value 2	Value 3	Value 4
1	Number of observations in a data set	100	500	10,000	
2	Number of causal variables	1	2	3	8
3	Correlations between certain X variables	Low (0-0.3)	Middle (0.3-0.7)	High (> 0.7)	-1
4	Random causal coefficients	0.1	0.5	0.8	Any
5	Causal coefficients for X variables	Equal	Different		

After the modeling, the estimated in (6)-(8) parameters of the causal yields were compared with the original values used in data generation. The estimated and the original parameters for models with one, two, or three causal variables on ten datasets, together with the relative errors, are presented in Table 4.

Table 4. Probability estimates for datasets with 1, 2, or 3 variables, by 10,000 observations.

		Estimated parameters				Original parameters				Relative error, %			
Data set	Model	r	$p1$	$p2$	$p3$	r	$p1$	$p2$	$p3$	r	$p1$	$p2$	$p3$
1	OLS	0.571	0.773			0.5	0.77			14.3	0.5		
2	OLS	0.107	0.269			0.1	0.31			7.1	14.0		
3	OLS	0.857	0.514			0.8	0.71			7.1	27.4		
4a	OLS	0.893	-0.426			0.8	0			11.6			
4b	Ridge	0.847	0.001			0.8	0			5.9			
5	OLS	0.104	0.026	0.716		0.1	0.03	0.72		4.4	8.2	0.4	
6	OLS	0.803	0.527	0.041		0.8	0.48	0.15		0.4	10.1	72.1	
7	OLS	0.095	0.837	0.911		0.1	0.87	0.87		4.8	4.2	4.3	
8	OLS	0.527	0.883	0.887		0.5	0.87	0.87		5.5	1.1	1.6	
9	OLS	0.489	0.677	0.393	0.677	0.5	0.53	0.53	0.97	2.2	28.8	25.2	30.5
10	OLS	0.099	0.498	0.559	0.613	0.1	0.53	0.53	0.65	0.7	5.2	6.2	5.5

In most cases, the OLS regression (6) works well, producing probabilities close to the original values used for the data simulation. In one dataset #4, the OLS yields the negative probability value (row 4a), so we run the ridge regression (7), which yields all positive probabilities (row 4b). It is interesting to note that the original p_1 in this case equals zero. The relative errors of the estimated probabilities to their original values show a reasonable precision mostly of several percent, but sometimes more (it often corresponds to close to zero or one original values).

What is especially important here, when the causes are dominantly random (like in rows 3, 4, and 6, when $r = 0.8$), the estimation procedure still yields very good results, separating causal related events with low intensity from this very high level (80%) of "noise". In fact, even the biggest deviation (72%) in row 6 for the estimate 0.04 vs. 0.15 doesn't seem bad with this high random influence. The other important feature: the procedure works even when coefficients are equal to each other, like in rows 7 and 8, with different level of randomness. It is remarkable because in traditional statistics, if two values (i.e., males and females) produce the same marginal frequency, the gender is considered having no causal interpretation. Actually, we can say that each cause works with the same intensity, and they both differ from the random cause.

In another experiment with eight variables, the original coefficients might take any values (not controlled). This situation matches the typical data sets in many applied research. The results of 40 simulations are shown in Table 5, where the average correlation of original *Y* with *X*s is 0.05, and the maximum correlation equals 0.15.

Table 5. Quality of the parameters estimation.

	*p*1	*p*2	*p*3	*p*4	*p*5	*p*6	*p*7	*p*8	*r*
Correlations between original and estimated values among 40 runs	0.69	0.87	0.78	0.81	0.83	0.86	0.74	0.80	0.64
Median error, % to original value	35	20	23	32	27	33	21	21	40

The first row in Table 5 shows that correlations between original and estimated values are quite big, so the procedure definitely captures the main features of the data. It is especially important because the original datasets (10,000 observations) have practically no correlations among *Y* and *X* variables, so in this situation the traditional statistical methods fail. The second row in Table 5 shows that median error is about 20-30% of the original values, similar to those in row 9 in Table 4. Of course, it is not an ideal but a good enough result in a situation where original data are uncorrelated. Other experiments showed that the estimations only slightly depend on the level of the mutual correlations between *X* variables, so the problem of multicollinearity is not so troubling in this approach as in common regression modeling.

Example of estimation of advertising efficiency

A typical phase in media planning is the analysis of mutual frequency distribution of the media vehicles (TV shows, magazines, websites, etc.) and of the particular brand consumption. The high brand frequency for some vehicle is considered as a good indicator, and this vehicle is included in the list of the candidates for making advertising via it. Table 6 in its left part presents cross-tabulation of five products and four media vehicles (all data are real and represent popular magazines and different important products; the number of the respondents is measured in tens of thousands). For example, in a cell Product 1 - Vehicle 2, or *P1*-*V2* (Table 6, left half), 13.8% means that this fraction of the readers of *V2* magazine have bought *P1*, so *V2* is the most promising vehicle (not accounting for circulation, frequency of advertising, and other factors).

Five causal models were constructed using each product as a target – the resulting parameters are presented in Table 6, the right part, with estimates of the random causes in the last column. Comparing the two parts of Table 6 shows a rather dramatic difference. The most promising cell *P1-V2* suggests that just about 3.1% of buyers (instead of 13.8%) might have bought the product due to this magazine's advertising, while the other customers could buy regardless of it. A similar diminishing we see in any cell, for instance, the Vehicle 1 is even not important at all, so all buyers have no relation to this magazine, they would buy the product anyway, without this advertising. This type of analysis shows completely different picture of the media performance, and the decisions about advertising distribution could be changed accordingly.

Table 6. Modeling of the advertising efficiency

	Fraction of vehicle audience consuming particular product, %					Estimated causal coefficients, %				
Product *P*/ Vehicles *V*	*V1*	*V2*	*V3*	*V4*	Total	*V1*	*V2*	*V3*	*V4*	Random causes
P1	7.2	13.8	9.2	11.8	8.5	0	3.1	1.9	6.5	8.3
P2	2.9	5.1	3.9	5	3.2	0	4.8	2.6	0	3.1
P3	0.3	1.2	1.8	0.5	0.5	0	1.1	1.5	0	0.5
P4	6.8	12.6	10.3	10.6	7.5	0	4.3	0	0	7.1
P5	2.2	3.4	5.9	3.8	2.9	0	0	0.7	1.9	3

For each product, the total number of positive outcomes was decomposed by different magazines, according to (9), (10) and (23) from Appendix B. As expected, the found shares reflect the importance of the magazines, as shown in Table 6. For example, for the product *P2* the vehicles *V2*, *V3* and random effect contribute as much as 17%, 7% and 76%, respectively; the random causes dominate (up to 95%) for all the considered products.

Conclusion

A new approach to causal modeling was considered based on the direct accounting for the internal relationship between the causal impacts and the outcome effect. The proposed model is a significant departure from the regular regression, or statistical learning models, as well as from the traditional models of causal analysis. In the suggested model, each causal variable effects the outcome individually, not cumulatively with others, which contrasts with the traditional statistics, where the outcome cumulates the combined effect of all the variables of influence, and adding variables improves the goodness of fit. Also, unlike in the traditional methods, the random cause is not considered as something to be "minimized", but rather as a reflection of all causes which were not captured by the introduced variables. The proposed approach to the analysis and estimation of causal relations demonstrates several important features:

- it offers a way to estimate the causal relationships, when many possible causes generate one effect – a situation very typical for numerous applications;
- it allows to estimate the intensity of the causal relationships in the data, even if there is no correlation between Y and X variables, when causal variables are highly correlated among themselves, when coefficients of variables are equal to each other, when random component in the data is very high; all these features make it very different from the traditional statistical and causal approaches;
- it works just with frequency tables (providing they exist for all or many combinations of the predictors), so there is no need for the original observational data sets, that may be very useful in many practical situations;
- parameter estimation is simple and could be performed with any available software.
- it works with data of high dimensionality, since the orthogonal design matrix allows to reduce estimation to paired regressions.

Future generalization of the main problems of the parameter evaluation for causal and random impacts can be seen in using numerical Y and X variables, and in the framework of complex causal relationships (as in structural equations, or acyclic graphs with do-operators).

References

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3) 199-231 (with comments and rejoinder). doi:10.1214/ss/1009213726

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Chichester, England: Wiley.

Hastie, T., Tibshirani, R., & Friedman, J. (2006). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Kline, R. B. (2010). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.

Kuznetsov D. & Mandel, I. (2007), Statistical physics of media processes: Mediaphysics. *Physica A*, *377*(1), 253-268. doi:10.1016/j.physa.2006.10.098

Lipovetsky, S. (2010). Enhanced ridge regressions. *Mathematical and Computer Modeling*, *51*(5/6), 338-348. doi:10.1016/j.mcm.2009.12.028

Lipovetsky, S. (2012), Regression split by levels of the dependent variable. *Journal of Modern Applied Statistical Methods*, *11*(2), 319-324. Available at: http://digitalcommons.wayne.edu/jmasm/vol11/iss2/4

Mandel, I. (2011), Sociosystemics, statistics, decisions. *Model Assisted Statistics and Applications*, *6*(3), 163–217. doi:10.3233/MAS-2011-0203

Mandel, I. (2013), Fusion and causal analysis in big marketing data sets. *Proceedings of JSM - Section on Statistics in Marketing*. Montreal, Canada: American Statistical Association.

Pearl, J. (2000), *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Pearl, J. (2013). Linear models: A useful “microscope” for causal analysis. *Journal of Causal Inference*. *1*(1), 155-170. doi:10.1515/jci-2013-0003

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688-701. doi:10.1037/h0037350

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge, MA: Cambridge University Press.

Splawa-Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (D. M. Dabrowska & T. P. Speed, Trans. and Ed.). *Statistical Science, 5*(4), 465 472. (Reprinted from 1923, *Annals of Agricultural Sciences, 10*, 1-51.)

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*(7), 557–585.

Wright, S. (1960). Path coefficients and path regressions: Alternative or complimentary concepts? *Biometrics*, *16*(2), 189–202. doi:10.2307/2527551

Appendix A: Closed-form OLS solution for the orthogonal design

The OLS solution (6) for a multiple linear regression can be presented in analytical closed-form, if we have a total set of all possible cells, that is: for K variables there are $N = 2^K$ cells of all possible combinations of 0 and 1 values by each variable. For instance, if $K = 3$, there are eight cells as those presented in Table 1. For the orthogonal design matrix, each coefficient of regression can be estimated by the paired relation. For that we need covariance of y with each x_j:

$$\operatorname{cov}\left(y, x_j\right) = N^{-1}\sum_{i=1}^{N}\left(y_i - \overline{y}\right)\left(x_{ij} - \overline{x}_j\right) = N^{-1}\left(\sum_{i=1}^{N} y_i x_{ij} - N\overline{y}\overline{x}_j\right) \tag{A1}$$

And the variance of each x is:

$$\operatorname{var}\left(x_j\right) = N^{-1}\sum_{i=1}^{N}\left(x_{ij} - \overline{x}_j\right)^2 = N^{-1}\left(\sum_{i=1}^{N} x_{ij}^2 - N\overline{x}_j^{\,2}\right) \tag{A2}$$

Covariance of two binary predictors is:

$$\begin{aligned}\operatorname{cov}\left(x_j, x_k\right) &= N^{-1}\sum_{i=1}^{N}\left(x_{ij} - \overline{x}_j\right)\left(x_{ik} - \overline{x}_k\right) \\ &= N^{-1}\left(\sum_{i=1}^{N} x_{ij}x_{ik} - N\overline{x}_j\overline{x}_k\right) \\ &= \frac{1}{2^K}\left(2^{K-2} - \frac{2^K}{4}\right) = 0\end{aligned} \tag{A3}$$

Then each coefficient of regression (6) equals:

$$\begin{aligned}b_j &= \frac{\sum_{i=1}^{N} y_i x_{ij} - N\overline{y}\overline{x}_j}{\sum_{i=1}^{N} x_{ij}^2 - N\overline{x}_j^{\,2}} = \frac{\sum_{i=1}^{N} y_i x_{ij} - 2^{K-1}\overline{y}}{2^{K-1} - 2^{K-2}} \\ &= \frac{1}{2^{K-2}}\sum_{i=1}^{N} y_i x_{ij} - 2\overline{y} = \frac{1}{2^{K-2}}(y'x_j) - 2\overline{y}\end{aligned} \tag{A4}$$

where $y'x_j$ is the scalar product of the vector and vector x_j. With all mean values of *xs* equal 0.5, the intercept in the model (5) equals:

$$b_0 = \bar{y} - b_1\bar{x}_1 - \ldots - b_K\bar{x}_K = \bar{y} - 0.5\left(b_1 + \ldots + b_K\right) \tag{A5}$$

Using (14) in (15) yields:

$$\begin{aligned} b_0 &= \bar{y} - 0.5\sum_{j=1}^{K}\left(\frac{1}{2^{K-2}}\left(y'x_j\right) - 2\bar{y}\right) \\ &= \bar{y} - \frac{1}{2^{K-1}}\sum_{j=1}^{K} y'x_j + K\bar{y} \\ &= \left(K+1\right)\bar{y} - \frac{1}{2^{K-1}}\sum_{j=1}^{K} y'x_j \end{aligned} \tag{A6}$$

The parameters of probability (8) are also related. Indeed, rewriting r using (A5) yields:

$$\begin{aligned} r &= 1 - \exp\left(b_0\right) \\ &= 1 - \exp\left(\bar{y} - 0.5\left(b_1 + \ldots + b_K\right)\right) \\ &= 1 - e^{\bar{y}}\left(e^{b_1}\right)^{-0.5}\ldots\left(e^{b_K}\right)^{-0.5} \\ &= 1 - \frac{e^{\bar{y}}}{\sqrt{\left(1-p_1\right)\ldots\left(1-p_K\right)}} \end{aligned} \tag{A7}$$

So, the relations (A5), or (A7) between the coefficients should be taken into account in simulations of the model parameters.

Appendix B. Decomposition to the additive shares of influence of each cause in the effect.

Consider (3) in a generalized form, where we denote $r = p_0$ and $x_{i0} = 1$ identically:

$$1 - S_i = \left(1-r\right)\prod_{k=1}^{K}\left(1-p_k\right)^{x_{ik}} = \prod_{k=0}^{K}\left(1-p_k\right)^{x_{ik}} \tag{B1}$$

with the aim to present S_i as a total of the items, each related to one of the causes:

$$S_i = \sum_{k=0}^{K} S_{ik} \tag{B2}$$

For the additive decomposition of S_i we take shares proportional to the ratio of logarithms:

$$S_{ik} = S_i \frac{\ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} \tag{B3}$$

The total of (B3), due to (B1), coincides with S_i itself:

$$\begin{aligned} \sum_{k=0}^{K} S_{ik} &= \sum_{k=0}^{K} S_i \frac{\ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} \\ &= S_i \frac{\sum_{k=0}^{K} \ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} = S_i \frac{\ln\left(\prod_{k=0}^{K}\left(1-p_k\right)^{x_{ik}}\right)}{\ln\left(1-S_i\right)} = S_i \end{aligned} \tag{B4}$$

If S_i was defined as the quotient of the counts $S_i = n_i / N_i$, where n_i is the counts of $Y = 1$ in the base size N_i of each cell. Then by using it in (B3), we obtain the estimation for the counts n_{ik} related to each k-th cause:

$$n_{ik} = N_i S_{ik} = N_i S_i \frac{\ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} = n_i \frac{\ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} \tag{B5}$$

Total of (B5) by k, similarly to (B4) yields:

$$\begin{aligned} \sum_{k=0}^{K} n_{ik} &= \sum_{k=0}^{K} n_i \frac{\ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} \\ &= n_i \frac{\sum_{k=0}^{K} \ln\left(1-p_k\right)^{x_{ik}}}{\ln\left(1-S_i\right)} = n_i \frac{\ln\left(\prod_{k=0}^{K}\left(1-p_k\right)^{x_{ik}}\right)}{\ln\left(1-S_i\right)} = n_i \end{aligned} \tag{B6}$$