Tutorials in Quantitative Methods for Psychology
2012, Vol. 8(2), p. 96-107.

DOI: 10.20982/tqmp.08.2.p096

Statistical simulation and counterfactual analysis in social sciences

François Gélineau, Pierre-Olivier Bédard, Mathieu Ouimet
Université Laval

In this paper, we present statistical simulation techniques of interest in substantial interpretation of regression results. Taking stock of recent literature on causality, we argue that such techniques can operate within a counterfactual framework. To illustrate, we report findings using post-electoral data on voter turnout.

The analysis of quantitative data, and the estimation of regression models in particular, can now be considered commonplace in the social sciences. There are, of course, notable variations in the ways those analyses are generated (research design, estimation methods, etc.). In the same way, there are discrepancies in terms of standards when it comes to the interpretation of the results and their proper communication.

Depending on the nature of the data at hand and the chosen estimation methods, the interpretation phase can be rather equivocal. For instance, displaying the odd ratios, or their natural logarithm, following logistic regressions can be far from intelligible, especially when one is interested in parameters beyond their statistical significance threshold and the direction of their coefficients. Consequently, a greater analytical effort appears to be required to flesh out the proper signification and meaning of parameters, most particularly to express their magnitude. The interpretation of statistical results appears crucial, especially under the lens of knowledge transfer, which involves non-statistical experts (decision-makers, policy analysts, etc.). We contend here that statistical simulation can be put to profit to this end.

We also make the argument that this approach is compatible with a counterfactual conception of causality. Although this is not the place to develop a full-fledged argument about causality, it has been suggested that counterfactual logic is central to the study of causality (see among others, Morgan & Winship, 2007; Pearl, 2000; Antonakis *et al.*, 2011; Woodward, 2003: 191) and to scientific thought more generally. In effect, a counterfactual conception of causality refers more to an overarching logic, to broad features of research designs than to specific analytical techniques. In the case of observational research (which is our focus here), a counterfactual framework would lean towards answering specific research questions about the likely effect of a given phenomenon under specific conditions. Or, put differently, such a design would be driven by an interest in « what-if-things-had-been-different questions » (Woodward, 2003). More precisely, a counterfactual analysis thus implies a comparison between two sets of conditions in the explanation of a given problem: one where the presumed cause is present (in the case of an experiment, the latter would be called a 'treatment'), and one where it is absent (again, in experimental language, the latter would be called 'control'). Consequently, the presumed causal effect would be the difference between the two states. The present paper aims to expose the advantages of specific simulation procedures applied to social science research problems framed in counterfactual language.

Objectives

Recently, multiple methods have been put forth to assist researchers in a more substantial interpretation of their data, notably through the application of statistical *simulations*. The current paper is largely based on existing works, notably from King, Tomz & Wittenberg (2000; see also Tomz,

François Gélineau, Département de science politique, Université Laval, Pavillon Charles-De Koninck, local 4403, Québec, Canada, G1K 7P4. Phone: (418) 656-2131 ext. 3073, Fax: (418) 656-7861. Email: francois.gelineau@pol.ulaval.ca

Wittenberg & King: 2001). Our goal here is not to repeat the authors' argument, but rather to build on their methodological insights and further their discussion of the interest in statistical simulation. Therefore, the objectives of the present paper are to 1) provide a brief overview of some existing simulation methods and their underlying principles (mainly, from a frequentist and Bayesian perspective), 2) discuss the relevance and implications of counterfactual analysis, 3) provide a few empirical illustrations of the simulation methods discussed, and 4) propose possible applications in other branches of statistics.

Simulation in frequentist and Bayesian statistics

Without resorting to a full review of existing statistical approaches and their philosophies, we can nonetheless frame our topic with respect with the two main statistical approaches – that is frequentism and Bayesianism. Both approaches, while often presented as largely opposed, display several features relevant to capture how statistical simulations operate. Essentially, by simulation we mean here two things: 1) the manipulation of the variables to compute quantities of interest and their variations given different values assigned to them, and 2) the generation of these estimates while taking into account the variables' distributional characteristics. This definition, however, comes with a caveat, as it is not a unifying definition. There exists multiple forms of simulation, but we limit our comments and propositions to *empirically based simulations*. The aim of such approach is mainly to explore the distributional properties of parameters and to convert this information into the language of probabilities (i.e. conditional probabilities)

Frequentism and Bayesianism differ on a number of aspects but we explore here three conceptions – of probabilities, inferences and analytical output – all of which have bearing on the topic of simulation. First of all, both operate (or can operate) as functions of maximimum likelihood (Jackman, 2000: 376) but differ on their views of parameters to be estimated, and more generally on the probabilistic notions underlying those estimations. For the frequentists, the parameters of the population are fixed (θ), but unknown, and the properties of the sample ($\hat{\theta}$) are random (meaning that multiple sampling would generate different results). Each sample is a random draw from an unknown population. As in the case of ordinary least squares (OLS), it is posited that the sample is normally distributed, that its distributional properties are similar to that of the population. Therefore, the inference that one is making is about the sample and its likelihood, and not directly about the population. In sum, in this perspective we seek to maximize the likelihood of the model, or rather to minimize its deviation from a perfect model (Pétry & Gélineau, 2009: 185).

Following the Bayesian approach, the reverse is true. The properties of the sample ($\hat{\theta}$) are fixed and the parameters of the population (θ) are random (in the sense that they appear unknown to the researcher). This last idea refers to the stochastic component of the models estimated in Bayesianism. The Bayesian methods are a way to integrate known information about a problem, and confronting this information to the data at hand, through Bayes' famous theorem, which we can describe as follows:

$$Pr(H|E) = \frac{Pr(E \bigcap H)}{Pr(E)} = \frac{Pr(E|H)Pr(H)}{Pr(E)}$$

where, Pr (H | E) is the probability that H is true, as a function of the data (therefore expressing a justification for our *belief* in H). The left-hand side of the equation is the *posterior density*, obtained through the estimation. The right-hand side can be explained as follows: Pr(E | H) is the probability of obtaining the observed results, reached through maximimum likelihood. Pr(H) is the expected probability (called *prior density function*), expressed and quantified *a priori*, about the research hypothesis. Finally, Pr (E) is the marginal probability. Therefore, the density function is related to the true parameter β (beta), and not the estimated $\hat{\beta}$, as is the case in standard regression: "it is interpreted as reflecting the odds the researcher would give when taking a bets on the true value of β." (Kennedy, 2008: 214) The formula just described is thus a "weighted average of the prior density and the likelihood (the "conditional" density of the data, conditional on the unknown parameters)." (Kennedy, 2008, 214) Or to put it again differently, "the posterior is proportional to the prior times the likelihood." (Jackman, 2004: 485)

Consequently, the frequentist and Bayesian approaches also differ in their analytical output. Without going into too much detail here, let's just recall that frequentist analysis produces *point estimates*, that is, punctual estimations, whereas Bayesian analysis produces as we have seen *posterior density functions*. Also, the Bayesian approach isn't based on sampling distributions. Nonetheless, and this is central to our upcoming discussion, Bayesian approaches tend to privilege statistical simulation as a way to increase statistical power. And this is consequent with what has been exposed earlier: "Anything we want to know about a random variable θ, we can learn by repeated sampling from the probability density function of θ." (Jackman, 2004: 493)

The main reason we considered this very brief incursion into the Bayesian approach, is that is has had a notable influence on several procedures such as Markov Chain Monte Carlo: "MCMC has a distinctly Bayesian heritage and is associated with a resurgence in Bayesian statistics." (Jackman, 2000: 376) Also, the reach of such statistical procedures goes well beyond the Bayesian field of statistics, and reach the frequentists as well: "we have also become more at ease with using Bayesian ideas such as simulation, whether from a "classical" (Herron 2000; King, Tomz, and Wittenberg 2000) or a Bayesian (Jackman 2000) perspective." As we will see later on, some existing methods allowing for the simulation of parameters (for instance, CLARIFY) have a clear Bayesian signature, in that they operate largely through Monte Carlo simulation (Bartels & Sweeney, 2004: 5), while still remaining in the frequentist tradition and without deflecting from central limit theorem.

The relevance of such methods and its applications, which will be described below, tend to illustrate that frequentism and Bayesianism, though quite different on various aspects, are nonetheless complementary (Williamson, 2011). What is more, we can go on to suggest that their analytical differences are mainly a consequence of the distinct, yet complementary, inferential objective each pursues: the frequentist mainly produces inferences about *classes* of events (e.g. the effect of a given variable), whereas Bayesianism appears a useful tool to produce inferences about *particular* cases or events (the probability of occurrence of a phenomena, given what we know).

Counterfactuals, causality and simulations

The recent interest in simulations has been equally fraught with enthusiasm (Reiss, 2011), doubt (Kästner & Arnold, 2011) and skepticism (about a specific form of simulation; Funcke, 2011). To some extent, this development is interestingly paralleled (although non necessarily co-extensive) with a renewed interest in the philosophical and methodological aspects of causality. We can distinguish between at least two types of concerns in this respect, that is, the proper conceptualization of causality (its ontology and its formalization) and the conditions of its materiality and the related methodological issues in the study of presumed causality.

As was suggested above, the counterfactual logic is closely related to the language of experiments ("treatment vs. control"). But this need not be limited to experimental designs, and we argue that observational analyses can emulate, so to speak, such counterfactual logic. What is more, statistical simulations can be put to profit in this respect. While it would probably be excessive to suggest that statistical simulations could act as substitute for experiments (Kästner & Arnold, 2011) – especially in social sciences, simulations can be designed and described using *counterfactual* language, therefore approximating experimental designs, at least *in principle*. That is to say "the epistemology of simulation is essentially an experimental epistemology" (Reiss, 2011: 250). As we shall see below, it is possible to proceed to postestimation analyses in a counterfactual fashion, where the intervention is emulated by a manipulation of the values of given variables.

Of course, this does not exempt the researcher to be methodologically sensitive to the usual disclaimer about the validity of causal inferences in non-experimental and the concerns about *endogeneity*. We can define this generalized problem as such: "If the relation between x and y is due, in part, to other reasons, then x is endogenous, and the coefficient of x cannot be interpreted, not even as a simple correlation (i.e., the magnitude of the effect could be wrong as could be the sign)" (Antonakis *et al.*, 2010: 1080). As it is understood here, the simulation procedures we explore have empirical contents (in that it is not merely intended to explore mathematical properties), and could be more or less conceived as a way to model reality. Therefore, the methods we discuss are valid inasmuch as the data, specification and estimations are valid. We believe that statistical modeling, simulation and postestimation procedures offer an epistemic access to causal inference (Russo, 2009: 55; Khander, Koolwal & Samad, 2010) and that this should be accompanied by a focus on the conditions of *validity* of those causal inferences (Antonakis *et al.*, 2010: 1090; Shadish, Cook, & Campbell, 2002; Pearl, 2004).

Simply put, the idea behind the counterfactual approach allows us to mimic the ex ante treatments and controls that are usually introduced in laboratory experiment, but with an ex post statistical strategy. In a laboratory setting, we would select the participants on the basis of some specific characteristics (e.g., age) in order to neutralize its effect on the dependent variable (e.g., voter turnout). This would allow us to measure the impact of another variable (e.g., years of schooling) on voter turnout through simple bivariate correlation. In the postestimation approach we use random participant selection, and assign a fixed value to every participant on one variable (age) while letting the other (years of schooling) vary when computing the point estimates. In doing so, we can isolate the effect of schooling on voter turnout while controlling for age.

Approaches to the computation of conditional quantities of interest

Predicted quantities of interest

It is quite common in regression analysis to report the coefficient, its direction, and its statistical significance at a given threshold. Although this can be instructive to some extent, it can also be limited in maximum likelihood models as the scale of the coefficient is not always intuitively and directly interpretable. This phenomenon holds true for different quantities of interest such as probabilities or other values. In this paper, predicted probabilities are used as an illustrative example of the simulation of a specific quantity of interest. The first step to obtain a more substantial interpretation of the data is to compute the predicted probabilities for each (or only one) categories of the dependent variable. It is possible to compute the predicted probability for each observation contained in the data set and obtain a new variable expressing the mean of those probabilities for each category of the dependent variable. In logistic regression, the predicted probability for each observation is obtained by the following equation:

$$P(Y=1) = \frac{1}{1+e(Exp-L)}$$

where L is the predicted log-odds ratio obtained by resolving the logit equation with individual data points. Averaging the values obtained with this formula using the observed data points in the dataset informs us of the mean probability of obtaining a positive value in the case of a dichotomous variable. This provides some information, but in theory, it shouldn't be very different from the mean value of this variable. That is to say that restraining yourself to this information would be of little value. To the extent that we are interested in obtaining a refined analysis of the marginal effect of one (or more) explanatory variable(s) when they take on different values it is possible to apply the counterfactual scheme described above by comparing the predicted probabilities for two given scenarios.

By scenario, we mean a situation where we assign a specific set of values to the independent variables in the model in order to obtain a predicted probability. The interest of such procedure is that it allows us to measure the relative impact of a single variable on the predicted probability. To do so, we repeat the simulation by maintaining the variables at their same values, except for the variable of interest which would be free to vary within the range of interest (say, an increase of one unit in the case of a continuous variable, and from the minimum to maximum (0 to 1) in the case of a dichotomous variable). Although extremely simple, this method can be informative of the effect of certain variables and can be useful to depict realistic scenarios of interest.

A notable caveat of this simple method is that it doesn't take into account the notion of uncertainty. The computation of the probabilities is based on the observations in the data set at hand and therefore the probabilities are expressed as point estimates, and not as distributions of probabilities. If there were no uncertainty surrounding our estimation, we wouldn't obtain a distribution but a specific value. This absence of uncertainty is posited implicitly if we report on predicted probabilities without confidence intervals.

This would be paradoxical if we return to the frequentist postulates regarding the status of estimated coefficients. Uncertainty, whether it is caused by measurement errors or by pure randomness, is inherent to regression analysis and should therefore be reported in analyses built on those estimations. As signaled by Herron (1999: 85):

> [...] if β is a random variable after estimation, then functions of β are random as well. In particular, randomness in β implies that the values of such functions cannot be known with complete certainty even after probit estimation. Therefore, when researchers estimate probit models and report functions of estimated β vectors, it is incumbent on them to identify residual uncertainty by also reporting standard errors and/or confidence intervals for the estimated function values.

One of the ways we can compute a confidence interval for predicted probabilities, also suggested by Herron, is to draw random vectors from the normal distribution of the variables and to compute the predicted probabilities in a repeated fashion (*ibid*: 87). This is precisely what CLARIFY (King *et al.*, 2000) allows one to do, through a simplified sequence. We turn next to the sequence used to randomly draw parameters from the normal distribution, as implemented in CLARIFY, to then turn to possible expansion and applications of this methods and its principles.

Random simulation of coefficients with CLARIFY

As was suggested above; "We can learn about the distributional properties of a random variable, *y*, by sampling many (*m*) times from the probability distribution that generated *y*." (Bartels & Sweeney, 2004: 4) In this respect, the programmed sequence CLARIFY (King, Tomz et Wittenberg, 2000; Tomz, Wittenberg & King, 2001) implemented for STATA software, allows for the generation, through random simulation, of parameters distributions. These distributions can then be used to estimate predicted probabilities involving a confidence interval. This can be done in a counterfactual fashion, as already suggested, and in this sense, can yield more substantive interpretation of data.

The procedure employed in CLARIFY is essentially a three-step operation, corresponding to three implemented commands in STATA. The first step is to estimate the regression model, just as would be done without resorting to the CLARIFY program. At that stage, the command generates as many coefficients as there are variables in the model, plus the constant (i.e. the mean value of Y when X variables are fixed at 0). However, for each coefficient β, one thousand observations are simulated, using the known properties of the variable (mean, standard deviation). In doing so, CLARIFY randomly generates a series of parameters with the same distributional characteristics. Those simulated parameters can thus be considered as multiple observations on the initial coefficients.

We can easily see at this stage that the procedure is not unfaithful to the central limit theorem postulates, in that the simulations are built around a normal distribution. On the other hand, the simulations include a stochastic component kindred to Bayesian techniques (e.g. MCMC). Consequently, as a result of this random component, the estimation generated by this method a likely to differ lightly when repeated.

There is also a difference to signal between the approach just exposed and other common techniques, such as *bootstrapping*. While the latter also simulates data, it is distinct in that it considers the sample as a *pseudo-population* from which other samples are drawn. In contrast, CLARIFY doesn't generate a subsample, but rather new distributions from the known properties of the parameters.

The second step in the process, once the model is estimated and the parameters are simulated, is to set the explanatory variables at given values. These can be the average value of the explanatory variables, or any other value that illustrates a given scenario. This is where the counterfactual logic comes into play.

Once the values are set, we can compute the quantities of interest (predicted or expected probabilities or first differences). This step is essentially the resolution of the equation for each combination of simulated coefficients (the default in CLARFY is 1000). In the case of predicted probabilities, the output is to be captured as the mean probability (average of the 1000 predicted probabilities) of obtaining, say, a positive value in a dichotomous dependent variable. This method is all the more interesting because it allows the result to be reported with a confidence interval. This interval stands as the average of the upper and lower bounds of each distribution, given any given level of confidence. It should be noted that in the case of first differences (the reported difference in probabilities when a variable of interest take successively two different values), the estimation of the probability is sensitive to the values to which the variables are set. Therefore, as the distribution is curvilinear (as it is a log function), once the values of all the explanatory variables allow the probability to go beyond the curve's inflexion threshold (where the curve flattens), the marginal effect of the variable of interest will be minimal, as it approaches a probability of 1.

We can clearly see how CLARIFY provides the advantage of reporting predicted probabilities contained in a confidence interval. The ensuing estimation is more reflective of the uncertainty and margin of error surrounding the estimation of β.

Simulation of parameters through sequences (or loops) operating in CLARIFY

The interest of such a method resides precisely in the fact that the parameters are simulated on the basis of the observation values themselves. Concretely, as we suggested before, the first three steps described above remains essentially the same. Yet we introduce a small change in the second step (after the estimation of the model). Instead of imposing a single set of values to the explanatory variables, we set, in as many iterations as there are observations in the dataset, the variables to the real values of each observation. The fist iteration thus uses the values of individual *i*. In doing so, we obtain a distribution of parameters based on the observed values of the first individual in the dataset. At that stage, we obtain an averaged predicted probability, with its confidence interval (lower and upper bound), for individual *i*. The sequence is then rerun based on the observed values in individual *ii*. The sequence is then repeated for every individual. The results of this procedure are stored in a new variable. In order to obtain an overall predicted value, we simply take the mean value of the newly created variable. The main advantage of this procedure is that it makes no explicit assumption about the distribution of the independent variables.

Empirical illustrations

To illustrate possible applications of the procedures described above, we develop a straightforward voter turnout model. The extant literature provides many plausible explanations to determine why some individuals vote while others abstain. [1] In the Canadian context, authors generally identify two series of explanations: sociodemographics and attitudinal factors (Blais 2000; Blais et al. 2000; Blais et al. 2002; Nevitte et al. 1999; O'Neil 2003; Pammett and LeDuc 2003; Rubenson *et al*. 2003). For instance, on the one hand, men, French-speakers, older people, wealthier, and more educated individuals are all expected to have a greater propensity to vote. On the other hand, people who immigrated recently as well as more cynical and less politically interested individuals are expected to vote less.

[1] See Blais 2000, chapter 2, for a nice review of the socioeconomic determinants of turnout in comparative perspective.

Data set

The model we estimate uses individual level data from a post-electoral survey administered immediately after the December 8, 2008 Quebec general election. The questionnaire was administered by *Jolicoeur et Associés*, a Montreal-based survey firm. Interviews were completed between December 9, 2008 and January 24, 2009. Respondents were selected from a list of randomly sampled phone numbers. The overall response rate was established at 38.4%.

A sampling strategy was developed to obtain a greater number of non-voters than we generally obtain through phone surveys. This strategy assigned a higher selection probability to non-voters in households that contain both voters and non-voters (mixed households). The resulting sample of 9992 respondents included 742 voters and 257 non-voters. Even with such a strategy, we were not able to match the real ratio of non-voters to voters. Our non-weighted estimate of turnout is 74.27%. The real turnout was almost 17% lower, at 57.43%.

The models

The model we use is fairly straightforward and follows the strategy used by Blais et al. (2004). We estimate a probit regression to assess the direct effect of a series of sociodemographic and attitudinal variables on the individual propensity to vote3. "Age" is inserted as the number of years4. "Education" is a categorical variable that distinguish respondents (1) with no schooling to secondary incomplete from (2) those with a secondary diploma completed and (3) those with a university diploma completed. The variable "francophones" is a dichotomous variable that identify respondents who mostly speak French at home. The variable labeled "immigrants" is a dichotomous variable that identifies respondents who immigrated to Quebec during the last 10 years. Finally two attitudinal variables are added to the model. These were generated from a series of 21 survey questions through a factor analysis. These variables measure political interest and cynicism. These two variables are zero-centered and range from about -3 to 3.

The simulation phases

In order to draw a comparison of the methods surveyed above, we proceed to compute the predicted probabilities, following the estimated model. We hereby intend to show that the different procedures, although similar in features, can yield variable results, even when ran with the same data. Before presenting the results, let us give an overview of how the procedures discussed above were applied in our demonstration.

We first estimated the predicted probability by setting the values of all explanatory variables to their mean, as in the simple CLARIFY strategy presented above. We resolved the equation to generate the average probability of occurrence of a positive value in *y* (i.e. propensity to vote). This analysis allows us to answer the question concerning the probabilities of voting for an average voter (as defined by its sociodemographic characteristics). Of course, by using the mean values, this strategy assumes that the independent variables follow a normal distribution.

Secondly, we proceeded to the estimation of the same quantity of interest, but instead of resolving the equation using the sample means, we worked by iterations, using the actual observed values for each individual in the dataset. The first iteration used the values of individual *i*, the second used the values of individual *ii*, the third used those of individual *iii*, and so forth. In the end, we completed as many iterations as there are individuals in the dataset. The final estimations were obtained by taking the mean predicted probabilities of the *n* iterations. The interest of this method is that it is reflective of real observed values, and that it does not assume the normal distribution of the independent variables.

Third, we estimated the differences in predicted probabilities (*first difference*) by setting the explanatory variables at their mean. As we were interested in describing the marginal effect of specific variables (mainly, the attitudinal disposition – cynicism), we allowed the latter to vary. Starting from the initial position to its mean, the variable of interest was then downgraded by one standard deviation below the mean, therefore capturing the effect of lower cynicism in voting behavior. This was done by following the steps implemented in the simplest CLARIFY strategy – as described above in the first step.

Although the first differences are useful to flesh out the magnitude of a given variable, the results of such estimations (using the simple CLARIFY strategy) remain bounded by the values set to the explanatory variables (that

[2] Because of missing values, the final sample contains 793 individuals, of which 587 (or 75%) declared having voted.

[3] The reader is referred to the Apendix to consult the coding used to generate the basic probit model (see, "BASIC PROBIT MODEL AND ESTIMATION OF Y USING THE MEAN VALUES OF THE IND VARS") as well as the coding used for all subsequent analyses.

[4] Note that in this section, the life cycle effect is not modeled as a curvilinear effect. Doing so would have made the path analysis too complex to interpret.

Listing 1. A probit model of voting propensity

```
Iteration 0:   log likelihood = -442.54388
Iteration 1:   log likelihood = -392.82019
Iteration 2:   log likelihood = -392.43731
Iteration 3:   log likelihood = -392.43719
Iteration 4:   log likelihood = -392.43719

Probit regression                                 Number of obs   =        786
                                                  LR chi2(7)      =     100.21
                                                  Prob > chi2     =     0.0000
Log likelihood = -392.43719                       Pseudo R2       =     0.1132

------------------------------------------------------------------------------
       statu |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     interet |   .3188387   .0620054     5.14   0.000     .1973104     .440367
     cynisme |  -.1999815   .0635844    -3.15   0.002    -.3246047   -.0753583
        age1 |   .0202896   .0036323     5.59   0.000     .0131705    .0274087
        educ |   .0632003   .0863653     0.73   0.464    -.1060726    .2324731
       femme |   .1798704     .10392     1.73   0.083     -.023809    .3835498
       anglo |  -.4082725   .2198382    -1.86   0.063    -.8391475    .0226025
     nouvimm |  -.9579363    .344262    -2.78   0.005    -1.632678   -.2831952
       _cons |  -.3960619   .2726144    -1.45   0.146    -.9303763    .1382526
------------------------------------------------------------------------------
```

Listing 2. Predicted probability of voting, given that all variables are set to their means

```
      Quantity of Interest |      Mean       Std. Err.    [95% Conf. Interval]
---------------------------+--------------------------------------------------
          Pr(statu=votants) |   .7749115      .0162244      .7419243    .8065885
```

is, in the preceding example, to their mean). To give the estimations a more realistic tone, we worked by iterations, just like in the second strategy outline above. The procedure[5], akin to that applied by Duch & Stevenson (2008, 2005), consists in downgrading the explanatory variable of interest (in this case, cynicism) by a unit, starting from the actual observed value for each individual. That is to say that individual *i*, who scored 0 will be downgraded to 0 minus one standard deviation, individual *ii*, who scored 3, will be set to 3 minus one standard deviation, and so forth. The first difference is thus estimated for every individual in as many iterations as there are individuals in the dataset. The individual effects are then averaged (so are the confidence interval lower and upper bounds). We contend that this strategy allows the predicted probabilities estimated to be closer to the actual data, and hence be more convincing and realistic.

Results

The basic probit model used for our demonstration was estimated on 786 observations and generated a pseudo R^2 of 0.1132, as reported in Listing 1. As it is the main focus of our demonstration, let us now focus mainly on the results for our variable of interest, that is cynicism. As shown, it is statistically significant at the 0.001 level, and the negative coefficient suggests that a greater level of cynicism induces a lower propensity to vote, when all things remain constant.

We are mainly interested in the substantive effect of the variable and therefore report the predicted probabilities, marginal effects and the like in a counterfactual framework. As described above, our first postestimation strategy yielded a predicted probability for the likelihood of voting when all other explanatory variables contained in the model are set to their mean. As shown in Listing 2, the average predicted probability of voting is 0.7749 (with a confidence interval located between 0.7419 and 0.8066).

This result is interesting in itself, but it comes with a caveat. When one is dealing with categorical and dichotomous variables, the mean, while statistically correct, just isn't realistic. As this was our case ("education" is categorical while "immigration status" and "French speaking" are dichotomous), it is worth turning to an approach that is more representative of the data at hand. As described above, our second strategy was to sequentially impose the scores of each individual, and then take the

[5] As described above, is also close to the second step described.

Listing 3. Predicted probability of voting – given that all observed values are sequentially imposed

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        yhat |       786    .7476261    .1523263   .1739246   .9849181
     yhat_lo |       786    .6691551    .1809765   .0435246    .964664
     yhat_hi |       786    .8183593    .1221224   .3397878   .9957854
```

Listing 4. Marginal effect of cynicism – given that all explanatory variables are set to their means and that cynicism is downgraded by a standard deviation below the mean

```
      Quantity of Interest |     Mean       Std. Err.    [95% Conf. Interval]
---------------------------+--------------------------------------------------
            dPr(statu = 1) |   .0481557      .0144332      .0197009    .0753839
```

Listing 5. Marginal effect of cynicism – given that all explanatory variables are set to their means and that cynicism is downgraded by a standard deviation below the actual observed value.

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      d_yhat |       786    .0462472    .0169467   .0047315   .0702915
   d_yhat_lo |       786    .0179051    .0062936   .0015096   .0312669
   d_yhat_hi |       786    .0753074    .0275097    .008661   .1188982
```

average predicted probability. As reported in Listing 3, this strategy yields an estimated probability of voting of 0.7476 (with in a confidence interval ranging from 0.6691 (*yhat_lo*) to 0.8183 (*yhat_hi*)).

While the probability is still quite high, it is interesting to note that our second strategy yielded a different predicted probability. We could certainly argue that the results are not strikingly different, but it still goes to show that these kinds of postestimation procedures are sensitive to the values we use to obtain the predicted probabilities, a point to which we will return in our closing comments.

We now turn to our results of differences in the predicted probabilities (*first differences*) by applying the two different approaches described above. At this stage, we proceeded to estimate the marginal effect of cynicism by setting all the explanatory variables to their mean, and by letting the variable of interest (cynicism) be downgraded by one standard deviation from the mean. The marginal effect, in this case, is simply the reported difference between the two predicted probabilities for the two scenarios. As reported in Listing 4, the marginal effect of cynicism, using this strategy, is estimated to be 0.0481 (with a confidence interval ranging from 0.0200 to 0.0753).

Finally, we estimated the marginal effect of our variable of interest by iteratively downgrading it by one standard deviation from the observed value for each individual, as described above. As reported in Listing 5, the marginal effect of cynicism on turnout is estimated to be 0.0462 (with a confidence interval ranging from 0.0179 (*d_yhat_lo*) to 0.0753 (*d_yhat_hi*).

Concluding remarks

We opened up by suggesting the importance of applying simulation technique to generate more substantive interpretations of statistical data, to "simulate for substance" (Bartels & Sweeney, 2004). We have shown in this paper that relatively simple methods can be applied, following statistical regressions, to yield estimates reporting quantities of interest expressible in the language of probabilities. We contend that taking greater advantage of such techniques can only be beneficial to flesh out the implications of inferential claims about presumed exogenous variables.

Nonetheless, the simulation procedures discussed above have some limits that should be made explicit. First of all, as the methods presented here intervene mainly after the model has been specified and estimated, it cannot be seen as an easy way out of the difficult problems that hinder statistical analyses more generally (that is, correct model specification, endogeneity, measurement error, multicollinearity, etc.). We can see that the results are *model-dependent*, in that they are an extension of the distributional properties of the variables contained in the model. For the sake of our demonstration, these things were taken for granted, even though in the course of current analysis, these aspects are fundamental.

The second identifiable limit to the simulation procedures discussed here is that they are sensitive to the type of variables mobilized in the estimation, especially when those are not normal-centered. This was the case for some of the variables used in our demonstrations, and the results suggest that the quantities of interest estimated aren't necessarily constant, when shifting the values imposed on the explanatory variables. Although the discrepancy certainly was not a major one, it nonetheless suggests a form of caution when proceeding to this kind of analysis and its interpretation. One should keep in mind that these estimations are partly contingent, in that they are conditional on the values one feeds in the explanatory variables. As see above, the iterative sequences are a possible solution to circumvent this kind of problem, as it remains committed to the estimation of predicted probabilities based on observed values.

In a general sense, our demonstration has hopefully shown the relevance of statistical simulation and counterfactuals but, in our view, it also acts as a reminder of the prime importance to think about the way we define our counterfactuals (see, among others, King & Zeng, 2007; 2006). This can be considered an argument that reinstates the importance of guiding our analyses by theory, or by being pragmatically guided and allowing minimal realism in our statistical operations.

References

Antonakis, J., S. Bendahan, *et al*. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21(6): 1086-1120.

Bartels, B. & K. Sweeney (2004), Simulation and Substantive Interpretation in Statistical Modeling, *Lab Notes*, 3(3).

Blais, A. (2000). *To Vote or Not to Vote : The Merits and Limits of Rational Choice Theory*. University of Pittsburg Press.

Blais, A., E. Gidengil, R. Nadeau, and N. Nevitte (2002). *Anatomy of a Liberal Victory: Making Sense of the Vote in the 2000 Canadian Election*, Peterborough: Broadview Press.

Blais, A., E. Gidengil, N. Nevitte, and R. Nadeau (2004). Where does turnout decline come from? *European Journal of Political Research*, 43: 221-236.

Duch, R.M. & R.T. Stevenson (2005). Context and the Economic Vote: A Multilevel Analysis. *Political Analysis* 13(4): 387-409.

Duch, R.M. & R.T. Stevenson (2008). *The Economic Vote. How Political and Economic Institutions Condition Election Results*. Cambridge University Press.

Funcke, A. (2011). A Skeptic Embrace of Simulation. Complex Adaptive Systems: Energy, Information and Intelligence, Association for the Advancement of Artificial Intelligence.

Jackman, S. (2000). Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo. *American Journal of Political Science*. 44(2): 375-404.

Jackman, S. (2004). Bayesian Analysis For Political Research. *Annual Review of Political Science.* 7(1):483-505.

Herron, M.C. (1999). Postestimation Uncertainty in Limited Dependent Variable Models. *Political Analysis*, 8(1).

Kennedy, P.(2008). *A Guide to Econometrics. Sixth Edition*. Blackwell Publishing

Kästner, J. and A. Eckhart (2011). When can a Computer Simulation act as Substitute for an Experiment? Preprint Series, Stuttgart Research Centre for Simulation Technology.

Khander, S. R., Koolwal, G. B., & Samad, H. A. (2010). *Handbook on Impact Evaluation: Quantitative Methods and Practices*. Washington, DC: The World Bank.

King, G., M. Tomz & Jason Wittenberg (2000). Making the Most of statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science.* 44:341-355.

King, G. & L. Zeng (2007). When Can History Be Our Guide? The Pitfalls of Counterfactual Inference. *International Studies Quarterly,* 51(1): 183-210.

King, G. & L. Zeng (2006). The Dangers of Extreme Counterfactuals. *Political Analysis,* 14(1): 131–159

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Camdridge University Press.

Nevitte, N.(1999). *Unsteady State: The 1997 Federal Election*, Oxford University Press.

O'Neil, B.(2003). Examen du déclin de la participation électorale chez les jeunes du Canada. *Perspectives électorales*, July.

Pammet, J. H. & L. LeDuc. (2003a). La problématique du déclin de la participation électorale chez les jeunes," *Perspectives électorales*, July.

Pammet, J. H. & L. LeDuc. (2003b). *Explaining the Turnout Decline in Canadian Federal Elections: A New Survey of Non-voters*. Elections Canada (March).

Pearl, J. (2000). *Causality* (2nd edition). New York : Cambridge University Press.

Pearl, J. (2004). *Robustness of causal claims.* Paper presented at the Proceedings of the 20th conference on Uncertainty in artificial intelligence, Banff, Canada.

Pétry, F. & F. Gélineau (2009). *Guide pratique d'introduction à la régression en sciences sociales. Deuxième édition revue et augmentée*. Ste-Foy, Les Presses de l'Université Laval.

Reiss, J. (2010). A Plea for (Good) Simulations: Nudging Economics Toward an Experimental Science. *Simulation & Gaming* 42(2): 243-264.

Rubenson, D., A. Blais, P. Fournier, E. Gidengil and N. Nevitte (2004). Accounting for the Age Gap in Turnout. *Acta Politica*, 39: 407-421

Russo, F. (2009). *Causality and Causal Modelling in the Social Sciences* (Vol. 14): Springer.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inferences*. Boston, MA: Houghton Mifflin Company.

Tomz, M., J. Wittenberg & G. King (2001). CLARIFY: Software for Interpreting and Presenting Statistical Results. Version 2.1. Cambridge, MA: Harvard University.

Williamson, J. (2011). Why Frequentists and Bayesians Need Each Other. *Erkenntnis*, 1(26).

Woodward, J. (2003). *Making Things Happen. A Theory of Causal Explanation*. Oxford, UK: Oxford University Press.

Manuscript received 3 March 2012
Manuscript accepted 7 May 2012.

Appendix Follows

Apendix. Programming used in STATA 11

```
* BASIC PROBIT MODEL AND ESTIMATION OF Y USING THE MEAN VALUES OF THE IND VARS
* [Results are shown in Listing 1 & 2]
use analyse_v1, clear
tab statu
estsimp probit statu interet cynisme age1 educ femme anglo nouvimm
setx mean
simqi, prval(1) genpr(yhat)
summ yhat


*ESTIMATION OF Y USING THE LOOP STRATEGY [Results are shown in Listing 3]
use analyse_v1, clear
probit statu interet cynisme age1 educ femme anglo nouvimm
keep if e(sample)
gen yhat=.
gen yhat_lo=.
gen yhat_hi=.
gen id=_n
summ id
local e = r(max)

forvalues i = 1/`e' {

      quietly: estsimp probit statu interet cynisme age1 educ femme anglo nouvimm

      quietly:  setx  interet  interet[`i']  cynisme  cynisme[`i']  age1  age1[`i']  educ
                  educ[`i'] femme femme[`i'] anglo anglo[`i'] nouvim nouvim[`i']

      quietly: simqi, prval(1) genpr(prob)

      quietly: summ prob

      quietly: replace yhat=r(mean) if id==`i'

      quietly: _pctile prob, p(2.5, 97.5)

      quietly: replace yhat_lo=r(r1) if id==`i'

      quietly: replace yhat_hi=r(r2) if id==`i'

      drop b1 b2 b3 b4 b5 b6 b7 b8 prob
}

summ yhat yhat_lo yhat_hi

*ESTIMATION OF D.Y USING SIMPLE CLARIFY STRATEGY [Results are shown in Listing 4]
use analyse_v1, clear
tab statu
```

```
estsimp probit statu interet cynisme age1 educ femme anglo nouvimm
setx mean
simqi, pr fd(prval(1)) changex(cynisme mean -.88)


*ESTIMATION OF D.Y USING THE LOOP STRATEGY [Results are shown in Listing 5]
use analyse_v1, clear
probit statu interet cynisme age1 educ femme anglo nouvimm
keep if e(sample)
gen d_yhat=.
gen d_yhat_lo=.
gen d_yhat_hi=.
drop cynisme1
gen cynisme1=cynisme-.88
gen id=_n
summ id
local e = r(max)

forvalues i = 1/`e' {

        quietly: estsimp probit statu interet cynisme age1 educ femme anglo nouvimm

        quietly:  setx  interet  interet[`i']  cynisme  cynisme[`i']  age1  age1[`i']  educ
                     educ[`i'] femme femme[`i'] anglo anglo[`i'] nouvim nouvim[`i']

        quietly: simqi, prval(1) genpr(prob1)
        quietly: setx cynisme cynisme1[`i']
        quietly: simqi, prval(1) genpr(prob2)
        quietly: gen d_prob=prob2-prob1

        quietly: summ d_prob

        quietly: replace d_yhat=r(mean) if id==`i'

        quietly: _pctile d_prob, p(2.5, 97.5)

        quietly: replace d_yhat_lo=r(r1) if id==`i'

        quietly: replace d_yhat_hi=r(r2) if id==`i'

        drop b1 b2 b3 b4 b5 b6 b7 b8 prob1 prob2 d_prob
}

summ d_yhat d_yhat_lo d_yhat_hi
```