The Canadian Journal of Program Evaluation Vol. 21 No. 2 Pages 181–200
ISSN 0834-1516

SIMULATING OR IMPUTING NON-PARTICIPANT INTERVENTION DURATIONS USING A FLEXIBLE SEMI-PARAMETRIC MODEL

Ian Cahill
Human Resources and Social Development Canada
Gatineau, Québec

Paula Folkes
Treasury Board Secretariat
Ottawa, Ontario

Les Szabo
Szabo Consulting
Ompah, Ontario

Abstract: In the evaluation of labour market training programs using matching, evaluators must decide when to start comparing participant outcomes against non-participant outcomes. Measurement relative to an intervention period permits the separation of training opportunity costs from possible benefits, but an equivalent period for the comparison group must be determined. One method imputes the timing of the intervention for comparisons from that of the participant match. However, with Propensity Score Matching, this may produce biased outcome estimates. Instead, the authors develop and apply semi-parametric duration models using Human Resources and Social Development Canada data to simulate the positioning and duration of the intervention for non-participants.

Résumé: Dans le cadre de l'évaluation des programmes de formation liée au marché du travail utilisant l'appariement, les évaluateurs doivent décider à quel moment ils commenceront à comparer les résultats des participants avec ceux des non-participants. La mesure liée à une période d'intervention permet la séparation entre les coûts d'option de la formation et les avantages potentiels, mais une période équivalente pour le groupe de comparaison doit être déterminée. Certaines méthodes suggèrent d'attribuer la durée de l'intervention pour chaque individu du groupe de comparaison selon le participant avec lequel il est apparié. Cependant, avec l'appariement sur la base de scores de propension

Corresponding author: Ian Cahill, Audit and Evaluation, Human Resources and Social Development Canada, 140 Promenade du Portage, Gatineau, QC K1A 0J9; <cahill@magma.ca>

(*Propensity Score Matching*) des résultats biaisés peuvent se produire. Les auteurs élaborent et appliquent plutôt les modèles semi-paramétriques de durée en se servant de données provenant de Ressources humaines et Développement social Canada pour simuler le positionnement et la durée de l'intervention pour les non-participants.

INTRODUCTION

Issues to Consider When Evaluating Active Labour Market Programs

Several countries, including Canada, offer a comprehensive suite of programs for their workers that provide training, employment counselling or other assistance after a job loss. In 1996, the Government of Canada introduced a series of Employment Benefits and Support Measures (EBSMs) to help clients return to work as quickly and efficiently as possible. Employment Benefits such as Skills Development are major interventions that tend to be longer in duration, while Support Measures such as Employment Assistance Services (EAS) are less intensive.

When evaluating EBSMs and other Active Labour Market Programs (ALMP), a number of methodological decisions must be made. Some evaluators have addressed the classic evaluation problem by using quasi-experimental designs with popular methods such as Propensity Score Matching (PSM), to compare the outcomes of program participants and otherwise-similar non-participants over some period of interest. However, when PSM is used, evaluators need to decide when to stop measuring pre-participation outcome variables for inclusion in models used to estimate propensity scores. They also need to decide when to start measuring and comparing participant and comparison outcomes for the purposes of estimating the program effect.

In addition, where measurement relative to a period of intervention is desired, evaluators must determine how to obtain the timing and duration of the intervention for non-participants, who by their definition would not have intervention start or end dates. While developing alternatives for the quasi-experimental evaluation of EBSMs in Ontario, we found that although it may seem natural when PSM is used to simply impute these dates from the participant match, this may lead to biased impact estimates. Instead of using imputation based on PSM, we develop and apply duration models to simulate EBSMs for non-participants. The result of this work, and the discussion around

how the timing and duration of the intervention for non-participants should be determined when PSM is used, constitute the main focus of this article.

This introduction continues with a discussion of issues related to defining the period of measurement relative to an intervention period. The second section provides an overview of PSM and shows how imputation of intervention dates based on the propensity score could lead to bias. The third section presents duration models as an alternative to imputation by PSM, providing a theoretical discussion and empirical evidence to demonstrate the advantages of their use. That section also shows how the decision to simulate or impute using duration models depends on the nature of the available data. The conclusion summarizes our proposal and provides some considerations when deciding on possible approaches.

When to Stop Measuring Pre-participation Variables Used in Matching

With the use of PSM, it is important to determine when to stop measuring any pre-participation outcome variables used in matching. These variables reflect labour market experience prior to program participation. It is important to ensure that they are exogenous, that is, that they do not contain information on selection into participation. Pre-participation outcomes could be measured up until the job separation date, a proxy for which is the start of the EI claim called the Benefit Period Commencement (BPC). Alternatively, provided that endogeneity is not a strong concern, one could measure pre-participation outcomes up until the intervention start date, and match on the information from the BPC to the start date (what we call the "gap"). It may be useful to see these dates illustrated in Figure 1.

When to Start Measuring Post-participation Outcomes

Another important task is to decide when to start measuring post-participation outcomes. One option is to simply measure and compare post-participation outcomes from the BPC. The obvious advantage is that this date exists for both participants and non-participants. However, measurement of post-participation outcomes from the BPC would combine outcomes *before and during* participation as being part of the outcomes *due* to participation. This may not be desirable, particularly if a significant amount of time passes following BPC before participation in an EBSM is even considered.

Alternatively, one could measure post-participation outcomes from the intervention start or end date. Using the start date would also include the in-program effect in the determination of program incremental impacts, which may not always be desirable. The likely negative in-program effect may outweigh the post-program effect, leading to a picture of program net impact that does not accurately reflect the true program impact, particularly if benefits and costs are only weighed over the short term. In contrast, the end date approach measures employment outcomes after the intervention, in the period during which we would expect to see changes such as job finding. Ideally, using both the start-date and the end-date approaches might allow separation of the costs and benefits of a program.

How to Obtain the Measurement Period for Non-participants

An apparent barrier to the measurement of outcomes from the intervention start or end date is that non-participants do not have such dates. Nevertheless, as Smith (2000, p. 15) pointed out, the outcome variables for the participants and non-participants should be measured in the same way. Traditionally, with the use of propensity score matching where measurement is done from the intervention start or end date, the timing and duration of participation for the non-participant is imputed from that of the matched participant. For example, Sianesi (2001), who uses PSM, derives the non-participant intervention start date by imputing it from the matched participant and measures and compares outcomes from the start date.

While it may seem intuitive that, when using a matching technique, intervention dates for each comparison group member could be imputed from the matched participant, we will demonstrate in this article that, in the case of propensity score matching, this may result in biased incremental impact estimates. The proposed alternative is to use duration models to simulate or impute the timing and duration of interventions for non-participants. We will demonstrate that duration models can be used to accurately create a distribution of simulated interventions for the comparison group that is similar to the distribution of participant interventions. However, before we proceed, we will describe the data sets and the unit of analysis used in the study.

The Data and the Unit of Analysis

Human Resources and Social Development Canada (HRSDC) datasets contain data rich enough to answer a variety of empirical questions. Variables related to labour market experience, the Employment Insur-

ance (EI) claim, and EBSM uptake (for participants) are observed for EI clients. Also, tax data from Canada Revenue Agency were available for use in developing the duration models, and provided characteristics such as marital status, self-employment income, total earnings, and income (a variable list is presented in the Appendix).

If participants had more than one intervention, these quality data sets captured it. Thus, because there could be multiple interventions in respect of the same unemployment spell, we needed to determine what would constitute an appropriate unit of analysis. We decided to use a construct called the Action Plan Equivalent (APE), which is defined as a single EBSM or series of EBSMs taken by a client that are no more than six months apart. For the purposes of this article, for APEs containing multiple EBSMs, the APE type is determined by the longest EBSM in the APE, and we use the terms APE and intervention interchangeably.

PROPENSITY SCORE MATCHING (PSM)

An Overview

PSM can reduce the "curse of dimensionality" that occurs when the dimension of the vector of characteristics X increases. Cell-based matching on all of the components of X becomes impractical when, for some values of X among participants, no close matches are found in the comparison group (Smith, 2000). PSM reduces this problem by matching on a single value called the propensity score summarizing relevant information in X (Dehejia & Wahba, 1998).

To implement PSM, each individual i in both the participant group and the eligible non-participant group is assigned an estimate of the propensity score P_i, which is the conditional probability[1] of participation defined by $P_i = e(X_i) = \Pr\{D = 1 \mid X_i\}$. The vector X_i represents observed characteristics of individual i, and D is an indicator of participation, taking the value 1 for participation. According to the key PSM theorem, X and D are stochastically independent conditional on the propensity score, so that the matched participant and non-participant samples will be balanced on the distribution of characteristics X (Rosenbaum & Rubin, 1983). The simplest version of PSM is pairwise nearest neighbour matching on the propensity score. For a discussion of several variants, see Smith and Todd (2004). The propensity score is estimated through logit or probit models, to which Greene (2003) provides a good introduction.

Imputation of Intervention Dates Based on the Propensity Score—Potential for Bias

As noted in the introduction, when the evaluator wishes to measure outcomes throughout periods established relative to an intervention period, it will be necessary to obtain intervention start and/or end dates for non-participants. In the case of PSM using a pairwise approach, it may seem natural to impute the intervention start and end dates of participants onto their matched non-participant "twin." One might conjecture that this must lead to an unbiased estimate of the program impact. We prove that this is not the case by offering a counterexample.

The problem lies in the fact that PSM may match individuals with the same propensity score, but with different observed characteristics X, and these different characteristics may imply different expected intervention start and end dates. It is not sufficient to simply ensure, as imputation by PSM does, that the distribution of the intervention start or end dates through the non-participant population is the same as that for participants. The dates assigned to non-participants must be appropriate at the individual level. This is because there may be variations in the rate at which individuals recover from labour market difficulties without treatment, and this "spontaneous recovery" may depend on the characteristics X that distinguish individuals.

A simple counterexample:

Suppose that $X = [Region, Gender]$, where $Region \in \{A,B\}$, $Gender \in \{M,F\}$. Let $Status \in \{P,C\}$ determine if they were a Participant (P) or a Control (C). The propensity score $e(X) = \text{Prob}(\text{participation} \mid X)$. Let the eligible populations in the four cells defined by X be equal in size. Suppose that $e([A,F]) = e([B,M]) = 0.3$ and $e([A,M]) = e([B,F]) = 0.1$. With PSM, men in Region A are often paired with women in Region B, and women in Region A are often paired with men in Region B. About half of the matches would be to the opposite gender.

For simplicity, we assume that the duration T_{GAP} from BPC to the start of the APE (real or potential) is fixed for both participants and controls, and we are concerned only with end dates. Almost all of the variation in APE duration is explained by one component of X: *Gender*. As an outcome measure, consider average $ weekly earnings Y in the 52 week period beginning at a point T weeks from the start of an APE (real or potential). For evaluation purposes we are

interested in the case where $T = T_{APE}$, the APE duration. Men tend to have short APEs of about 4 weeks duration, while women tend to have long APEs of about 52 weeks duration. All controls experience unaided income recovery, but recovery is much faster for women. Explicitly we assume

$$E(Y \mid T = 4, Status = C, Gender = M) = \$100$$

$$E(Y \mid T = 52, Status = C, Gender = M) = \$500$$

$$E(Y \mid T = 4, Status = C, Gender = F) = \$500$$

$$E(Y \mid T = 52, Status = C, Gender = F) = \$500.$$

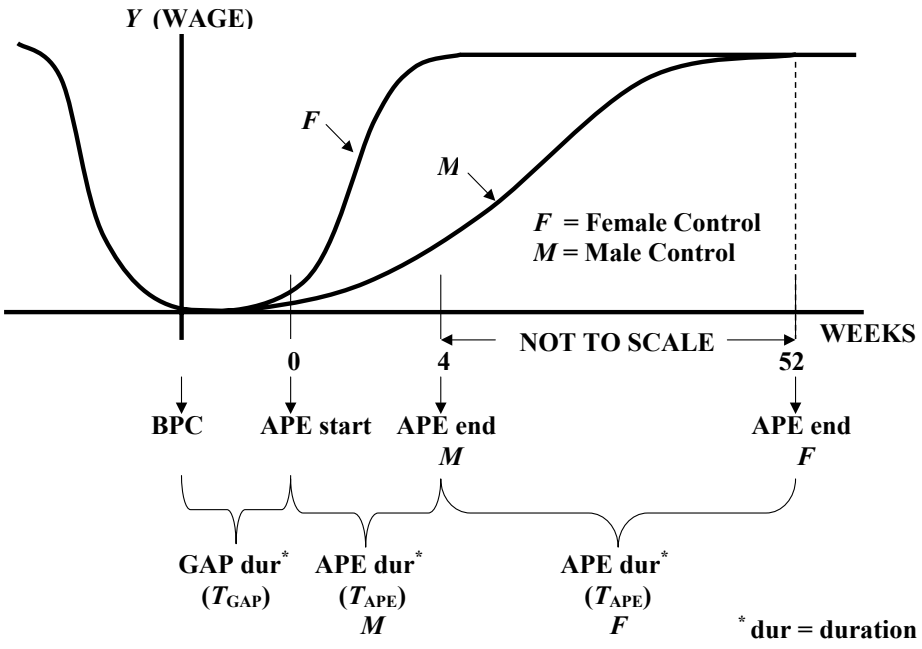
Now we can see how imputation can lead to biased estimates. When a male control is matched with a female participant, the duration of the control APE will be imputed to be about 52 weeks. For evaluation purposes, the expected outcome estimate is $E(Y \mid T = 52, Status = C, Gender = M) = \$500$, which is wrong. We assumed that he will have recovered to equilibrium level already by time $T = 52$, so his measured outcome is higher than it should be for a male control. The correct measure is $E(Y \mid T = 4, Status = C, Gender = M) = \$100$. For a female control matched with a male participant, the estimated control outcome is correct. Even though we impute her APE duration to be 4 weeks (too short), we assumed that she will nevertheless have already recovered to a stable level when measurement begins, so the outcome measure is correct. Naturally the estimate is also correct when participant and control have the same gender. The overall average estimated outcome for controls is biased due to overestimates for male controls matched with female participants. Hence we will underestimate program impact. Figure 1 is intended to illustrate the argument.

In applied work it is common to force a match on gender (we agree with this practice), but in the counterexample gender appears merely as a formal representative characteristic. We are forced all the way to cell-based matching if we wish to eliminate potential bias of this sort.

In the example, T represents the end date. Since we have assumed that everyone has a start date that is the same distance from BPC, a simple change converts the argument to start dates. The time T must now refer to APE start, measured as distance from BPC, instead of

referring to APE end, measured as distance from APE start. Change the phrase "men tend to have short APEs of about 4 weeks duration" to "men tend to begin their APE about 4 weeks from BPC," and change the phrase "women tend to have APEs of about 52 weeks duration" to "women tend to begin their APEs about 52 weeks from BPC."

Figure 1
A Simple Counterexample



DURATION MODELS AS AN ALTERNATIVE

An Overview

Duration models are a type of econometric model specialized to deal with phenomena involving the length of time spent continuously in a particular state. Equivalent terms are survival models, from medical applications, and failure time models, from engineering applications. These models deal with the particular characteristics of duration data, such as the restriction to positive values, the wide variety of potential distributional forms, and the possibility of durations being interrupted at the time of data collection (censoring). The effect on duration of

a vector of explanatory variables or covariates X is often represented. See the Appendix for more details and some references.

Use of Duration Models in Program Evaluation

Several recent evaluation studies have used duration models to evaluate the effectiveness of active labour market programs similar to EBSMs in Canada. Program effectiveness has been evaluated by using duration models to examine the transition out of unemployment to employment after participation in an ALMP, and sometimes to examine the transition out of subsequent spells of employment. This is the "event history" or "timing of events" approach introduced into the evaluation literature by Ridder (1986). Studies using this approach include Bonnal, Fougère, and Sérandon (1997) and Lalive, van Ours, and Zweimüller (2002). Analysis using duration models has also been combined with the use of a comparison group, as in Dolton and O'Neil (1996) and Brännäs (2000).

This article appears to be the first published study using duration models to assign analogues of intervention start and end dates to non-participants. Recall that the dates are positioned relative to BPC according to T_{GAP} and T_{APE} (see Figure 1), which means that assigning dates is equivalent to assigning durations. We consider three possible methods of assigning durations using an estimated model. The first method, prediction, is shown to be inappropriate because it leads to biased program impact estimates. Simulation may be suitable for large datasets, such as for the administrative data that we are using. Imputation on predicted duration may provide more accuracy in the case of small datasets.

Predict:

Once a duration model is estimated, it can be used to predict the duration T based on the values of X. Since the stochastic model has been completely specified, we could estimate a conditional mean $E(T \mid X)$ and use this as a prediction of T. Ignoring the stochastic element of T could be a problem, however. To simplify, suppose that T_{GAP} and X are constant, and the outcome Y for the control group is simply a non-linear function of $T = T_{APE}$. In order to compare mean outcomes with those of the participant group, we wish to estimate $E(Y(T))$ for the control group, but we have only an estimator of $Y(E(T))$, and according to Jensen's inequality $E(Y(T)) \leq Y(E(T))$ if Y is concave, and the reverse if it is convex (see Parzen [1960, p. 434]).

Simulate:

Deterministic algorithms can be used to generate series of numbers that appear to be random, and these form the basis of simulation. With many duration models one can use these pseudo-random number generation algorithms to simulate values of the duration T that will incorporate all of the information contained in the estimated model (see our discussion in the Appendix). As we demonstrate through graphs (Figures 3–6), simulation using a duration model can reproduce the distribution of durations among participants very accurately. In the case of small samples, however, it may be that the randomness in the sample distribution introduced by simulation would be significant enough to suggest consideration of an alternative method, such as imputation.

Impute:

It is possible to use the randomness incorporated in the dates of participants to add a stochastic element to the dates of non-participants, in addition to the dependence of duration T on the characteristics X. After carrying out PSM we have a matched set of non-participants that we will now refer to as the set of "controls." We propose matching each control with the participant who is the nearest neighbour on predicted gap duration, which will be a function of $X\hat{\beta}_{GAP}$, where $\hat{\beta}_{GAP}$ is the vector of estimated coefficients from the gap duration model. Controls and participants can then be divided into strata on imputed gap duration for controls and on observed actual gap duration for participants. Within each of these strata, each control is matched with the participant who is the nearest neighbour on the predicted APE duration, a function of $Z\hat{\beta}_{APE}$ from the estimated APE duration model (as explained in the next section, the characteristics Z should include the gap duration T_{GAP} as well as the vector X). Each control will then have an imputed APE duration that incorporates correlation with the gap duration.

The Advantages of Duration Models in Determining Non-participant Dates

If we use PSM for imputation of T_{GAP}, the duration of the gap from BPC to start date, and T_{APE}, the duration of the APE, then we are using a circuitous path. First each vector of characteristics X is assigned a propensity score. Then imputation from matched participants maps each propensity score to a pair of durations. Yet we have no reason to suppose that individuals with the same propensity score but different vectors of characteristics X will have similar duration.

We propose estimation of the parameters of a duration model that maps each vector of characteristics directly to durations. Most duration models are univariate—that is, they show the dependence of a single duration variable on a vector of covariates. Nevertheless, this type of model can be used in our situation if we accept a model where the determination of T_{GAP} and T_{APE} is sequential rather than simultaneous. We assume that T_{GAP} is determined first, and may be used as a predetermined variable to be included, along with the covariates in X, as part of a new set of covariates Z used in the duration model for T_{APE}.

If the duration model is correctly specified, both the simulation and imputation methods described in the previous section will assign to each non-participant a duration with a distribution, conditional on characteristics X, that is accurate up to unbiased estimates of the model parameters. This will prevent the sort of bias that we have shown could result from imputation using PSM. In the following sections we introduce the duration model that we believe to be sufficiently flexible to provide an adequate specification of duration distributions that are likely to arise in practice. We then provide evidence of the effectiveness of this approach, compared to imputation using PSM.

The Cox Model

The Cox proportional hazard model that we use allows for great flexibility in the form of the duration distribution. A key feature is the ability to estimate the dependence of duration on the characteristics X, while leaving the "baseline" distribution (where all characteristics take fixed values such as sample means) free to take any shape. With little theory to guide us, this is a desirable feature. Very accessible discussions of the Cox model are provided by Allison (1984, 2000).

We estimate the Cox model using a maximum likelihood technique and then use a pseudo random number generator to simulate non-participant durations according to the estimated model. Details are presented in the Appendix.

Comparison of Methods

We worked with administrative data and large samples. Thus our choice is between imputation by PSM and simulation by duration

models. We have given theoretical arguments against imputation by PSM, and now we provide empirical evidence.

Comparing the methods empirically is challenging. We do not have data on the potential start and end dates for non-participants, nor do we have data on the true impact of the programs. Nevertheless it is possible to use the participant group to test how well each of the two methods assigns start and end dates. We do this by first assigning propensity scores with a logit model using the entire participant and non-participant samples. Then we split the participant sample randomly into two parts. One part serves as a sample of participants, while the other part serves as a sample of pseudo non-participants. PSM is then applied to the two groups, matching each participant with the pseudo non-participant who is the nearest neighbour on propensity score. We can then apply both methods of assigning durations, comparing results with actual durations.

In order to compare results of the two methods we calculate the average squared difference between assigned and actual for both methods. For the case of APE duration we also take a look at the difference between the mean assigned and actual values across categories for certain explanatory variables. Results for the comparison of methods are in Table 1 and Figure 2.

Table 1
Compare Simulation vs Imputation from PS Match

Duration	*Var #*	*Variable*	*Mean Sqr (Days²)*	*Std Error*	*t Value*	*Pr > |t|*
APE	1	|Imputed - Actual|²	35760.2	778.8	45.9	<.0001
APE	2	|Simulated - Actual|²	16884.6	329.5	51.2	<.0001
Gap	3	|Imputed - Actual|²	17281.8	281.0	61.5	<.0001
Gap	4	|Simulated - Actual|²	13054.2	230.4	56.7	<.0001
APE	5	Var1 - Var2	18875.6	764.7	24.7	<.0001
GAP	6	Var3 - Var4	4227.6	309.1	13.7	<.0001

The results in Table 1 for APE duration show a distinct superiority of the simulation method. The mean squared differences for imputation by PSM are more than twice those for simulation. The results for the gap duration are less striking, but they also favour the simulation method. Examining the Cox model output in Table 2 presented below, we see that the Cox model fit statistic (-2 LogLike) and the three tests of the Global Null Hypothesis (all coefficients are zero) all indicate that the Cox model for the APE duration fits better than the model for the Gap, so we would expect that simulation would perform relatively better for APE duration.

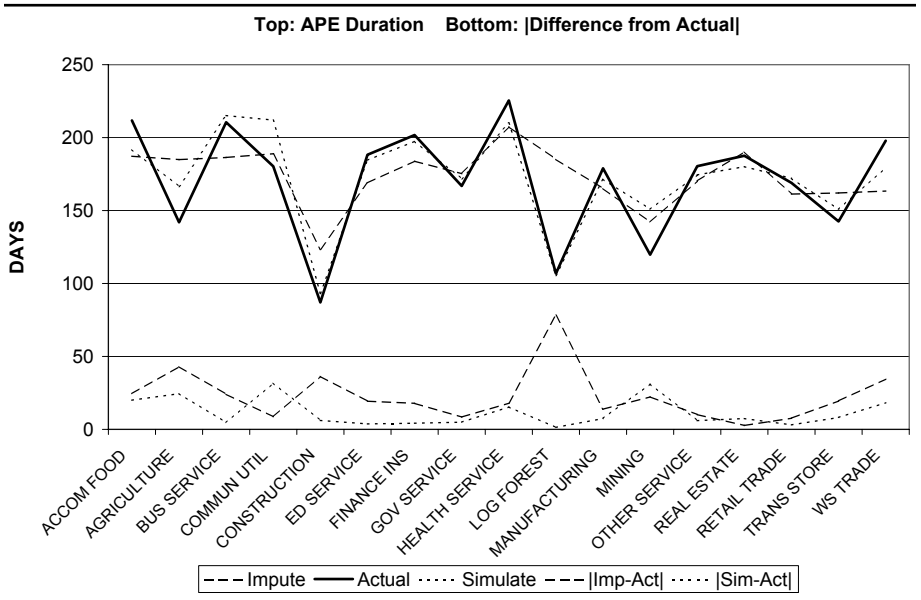
We include the t-statistic in order to test whether the mean squared differences for imputation and simulation are distinct from each other. It can be seen from the p values in the last two rows of the last column that, for both APE duration and gap duration, the null hypothesis of equal means is easily rejected.

Figure 2 shows that the simulation method tracks the mean APE duration better than the imputation method in 14 out of the 17 industries. We expect this sort of result from the arguments presented earlier concerning the advantages of duration models. The duration model incorporates industry categories as part of the vector of characteristics X.

Results of Duration Model Simulations Compared to Actual Durations

Observations were randomly assigned to one of two datasets. One dataset was used to develop and estimate the models. The other dataset was used to apply the models for simulation, producing the results presented in this section. Successful model development and an appropriate application of the model would show similar distributions of intervention gaps and durations for both participants and non-participants.

Figure 2
By Industry: Compare Simulation vs. Imputation from PS Match



The duration models worked very well to simulate distributions of gaps and APE durations for non-participants that are similar to participant durations. We exhibit results for Skills Development (SD) APEs, which are the largest category of major interventions. From Table 2 one can see that three tests (Likelihood ratio, Score, and Wald) easily reject the null hypothesis that the vector of characteristics X has no effect on duration.

Table 2
Duration Model Fit Statistics and Tests

	BPC-APE GAP			*APE DURATION*		
Test	*Chi-Sq*	*DF*	*Pr > ChiSq*	*Chi-Sq*	*DF*	*Pr > ChiSq*
Likelihood Ratio	3757.92	107	<.0001	8015.01	107	<.0001
Score	3972.73	107	<.0001	7742.12	107	<.0001
Wald	3744.47	107	<.0001	6808.05	107	<.0001
	Num. Obs = 10030			Num. Obs = 10030		
	-2 LogLike = 161012.65			-2 Log Likelihood = 156755.56		

Figures 3–6 show the real and simulated gap and APE duration for Skills Development APEs. Both the gap and APE duration simulations are extremely successful. The Cox model is flexible enough to deal with two duration distributions having radically different shapes.

Figure 3
Histogram: Real Gap (BPC–APE Start)

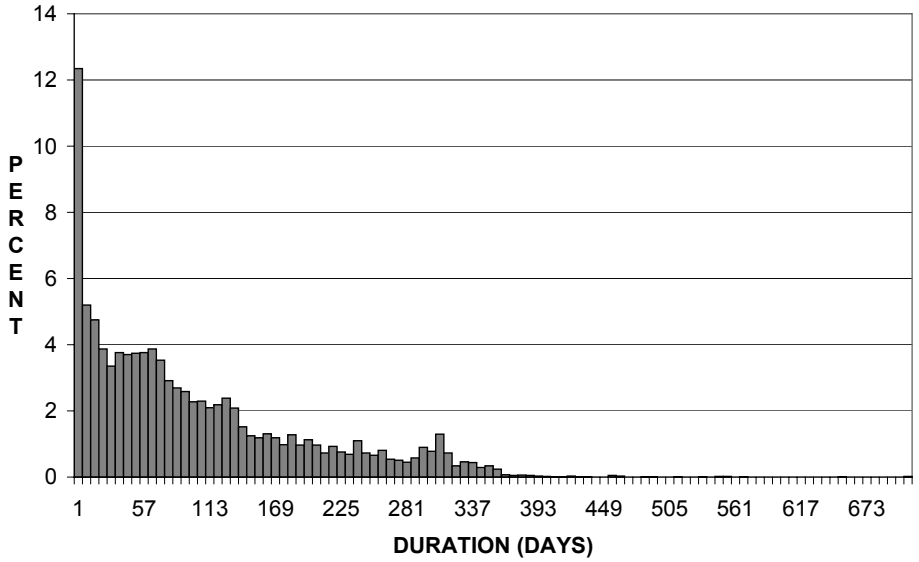


Figure 4
Histogram: Simulated Gap (BPC–APE Start)

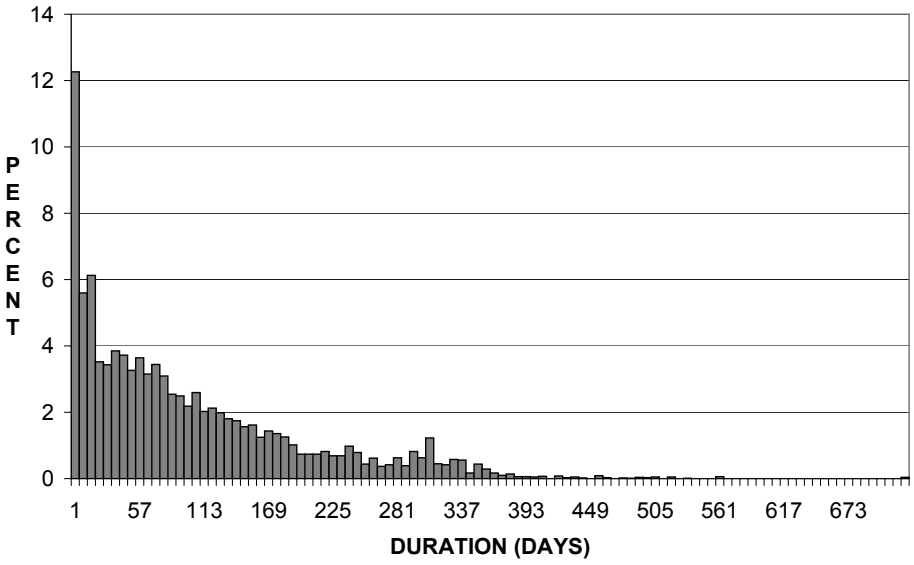


Figure 5
Histogram: Real APE Duration

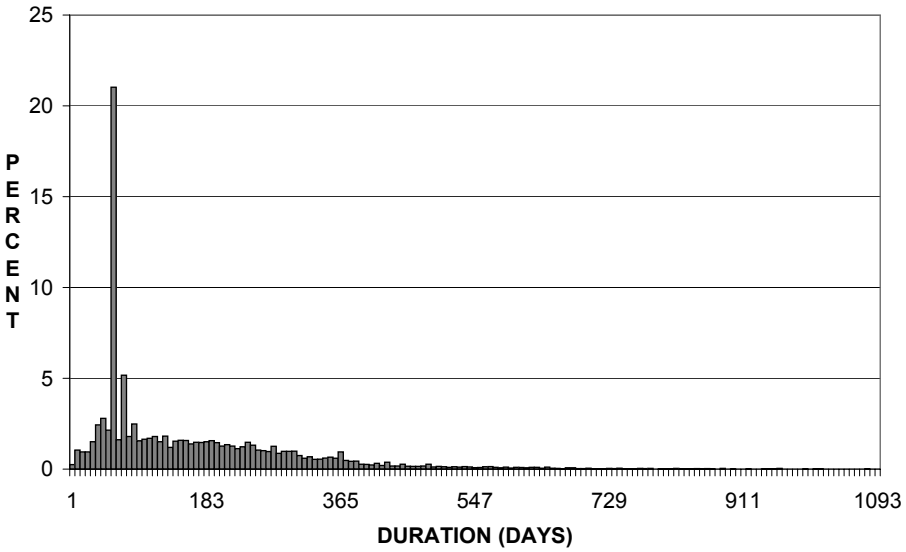
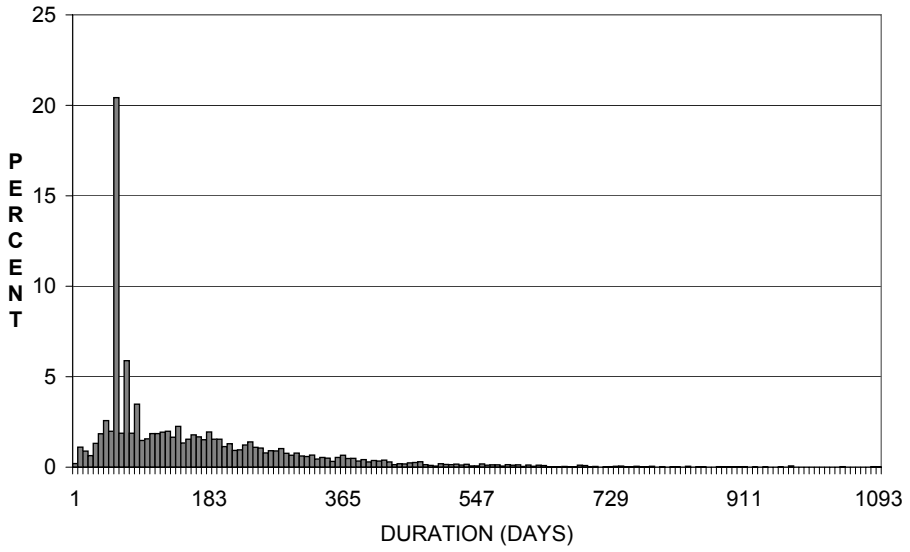


Figure 6
Histogram: Simulated APE Duration



CONCLUSION

In conclusion, we have demonstrated that when evaluating a program where outcomes will be measured and compared relative to a period of participation, duration models can be used to accurately create a distribution of simulated interventions for the comparison group that is similar to the distribution of interventions observed for participants.

Propensity score matching is a useful technique because it balances across characteristics overall, and can match units that have very different characteristics X but that have a similar probability of participating. Once matching is implemented, evaluators often turn their attention to determining intervention dates for comparisons that are analogous to program start and end dates for participants. While it may seem intuitive to simply impute the duration and timing of participation for the comparison from the matched participant, PSM imputation ignores the fact that despite a similar probability of participating, the matches may have little in common with respect to their individual characteristics; it is on these individual charac-

teristics X that the different rates of recovery from labour market difficulties may depend. Imputation by PSM will not avoid this potential source of bias.

Duration models do not rely on the propensity score but rather use the individual characteristics to estimate the timing and duration of participation. Whether one simulates or imputes using the duration model depends on the nature of the data that are available. For larger datasets, we have demonstrated that simulation using duration models may be suitable, while imputation on predicted duration may provide more accuracy in the case of small datasets. In both instances, we recommend the use of the Cox semi-parametric duration model because of its flexibility.

Evaluators strive to obtain estimates of incremental program effect that are as free from bias as possible. Certainly, with the use of quasi-experimental methods, it is arguable that theoretically some bias is likely unavoidable. However, potential bias introduced by imputing the timing and duration for the comparison from its participant match using PSM is indeed avoidable, and we argue it should be avoided wherever possible. In order to ensure robust, bias-free incremental impact estimates, evaluators must scrutinize the methods proposed for use in their evaluations. In this article, a viable alternative has been developed, and evaluators may now discriminate between the methods available for use. In our view, there is no longer a choice with respect to the use of imputation by PSM. The choice now is, with the use of duration models, whether one imputes or simulates the timing and duration of the comparison intervention.

ACKNOWLEDGEMENTS

We thank our colleagues in Program Evaluation, Human Resources Development Canada (HRDC), for fruitful discussions, and also Paul Decker and Peter Schochet of Mathematica Policy Research Inc. for useful comments on related evaluations. This work was performed when the authors worked in Program Evaluation, HRDC (now Human Resources and Social Development Canada). Opinions expressed are solely those of the authors.

NOTE

1. For a classic introduction to conditional probability and the related notion of stochastic independence, see Feller (1968, Chapter V).

REFERENCES

Allison, P. (1984). *Event history analysis: Regression for longitudinal event data* (Quantitative Applications in the Social Sciences #46). Newbury Park, CA: Sage.

Allison, P. (2000). *Survival analysis using the SAS system: A practical guide*. Cary, NC: SAS Institute.

Bonnal, L., Fougère, D., & Sérandon, A. (1997). Evaluating the impact of French employment policies on individual labour market histories. *Review of Economic Studies, 64*, 683–713.

Brännäs, K. (2000). *Estimation in a duration model for evaluating education programs* (Discussion Paper 103). Bonn, Germany: IZA—Institute for the Study of Labor.

Cox, D.R., & Oakes, D. (1984). *Analysis of survival data*. London: Chapman and Hall.

Dehejia, R., & Wahba, S. (1998). *Propensity score matching methods for non-experimental causal studies* (NBER Working Paper 6829). Cambridge: NBER.

Dolton, P., & O'Neil, D. (1996). Unemployment duration and the restart effect: Some experimental evidence. *Economic Journal, 106*, 386–400.

Feller, W. (1968). *An introduction to probability theory and its applications*. New York: Wiley.

Greene, W.H. (2003). *Econometric analysis (*5th ed.). Upper Saddle River, NJ: Prentice Hall.

Kalbfleisch, J.D., & Prentice, R.L. (1980). *The statistical analysis of failure time data*. New York: Wiley.

Lalive, R., van Ours, J.C., & Zweimüller, J. (2002). *The impact of active labor market programs on the duration of unemployment* (Working Paper 41). Zurich: Institute for Empirical Research in Economics, University of Zurich.

Lancaster, T. (1990). *The econometric analysis of transition data*. Cambridge: Cambridge University Press.

Parzen, E. (1960). *Modern probability theory and its applications*. New York: Wiley.

Ridder, G. (1986). An event history approach to the evaluation of training, recruitment, and employment programs. *Journal of Applied Econometrics, 1*, 109–126.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Sianesi, B. (2001). *An evaluation of the active labour market programmes in Sweden* (Working Paper 2001:5). Uppsala, Sweden: IFAU Office of Labour Market Policy Evaluation.

Smith, J. (2000). A critical survey of empirical methods for evaluating active labor market policies. *Swiss Journal of Economics and Statistics, 136*(3), 1–22.

Smith, J., & Todd, P. (2004). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*(1–2), 305–353.

Stuart, A., & Ord, J.K. (1987). *Kendall's advanced theory of statistics* (originally by Sir Maurice Kendall) (5th ed., Vol. 1: *Distribution Theory*). New York: Oxford University Press.

APPENDIX: DETAILS OF THE ECONOMETRIC METHODOLOGY

In duration models it is convenient to define the survival function $S(t) = 1 - F(t)$ where F is the cumulative distribution function (cdf) of duration, and also the hazard rate $h(t)$, which is the instantaneous termination rate at time t, conditional on having attained the duration t. The Cox proportional hazard model specifies the hazard function according to

$$h(t) = \lambda_0(t)\exp(X'\beta) \tag{1}$$

where X is a vector of covariates (no constant term), β is a vector of coefficients, and $\lambda_0(t)$ is a baseline hazard function with time-dependence left unspecified, so that it provides great flexibility. Good references on duration models are Kalbfleisch and Prentice (1980), Cox and Oakes (1984), and Lancaster (1990). There is also a good introduction in Greene (2003).

For our application we make use of an estimate of the survivor function at mean values of the covariates. The value of the survivor function for any value of the covariates X is then given by

$$S(t,X) = \exp\{\exp[(X-\bar{X})'\beta] \cdot \ln[S(t,\bar{X})]\} \quad (2)$$

Equation (2) is critical for simulation, because it provides us with an estimate of the survivor function, and therefore the cumulative distribution function (cdf), for any individual in the dataset. We use uniformly distributed pseudo random numbers in the inverted Cox cdf, obtained by a table-lookup. See Stuart and Ord (1987, p. 306) for a description of this approach to simulating general distributions.

Table 3 displays the vector of characteristics X, which comprises an extensive list of demographic and socio-economic variables, as well as some detailed information about the use of EI.

Table 3
Variable List for the Vector of Characteristics *X*

Demographic	*Social*	*Economic*	*EI related*
Gender	Aboriginal?	SA receipt?	Insured hours
Age category (5)	Disabled?	Occupation (27)	BPC yr & quarter
Region (19)	Visible minority?	Industry (18)	Insured earnings
Marital Status	Student?	Total Income	Prior EI claims
		Self-employed?	Prior weeks of EI
		Number of jobs	Regular Claim?

Ian Cahill, M.Sc. (mathematics), M.A. (economics), is a Senior Evaluation Officer at Human Resources and Social Development Canada. He has over 20 years of experience with the Canadian government in policy research and evaluation.

Paula Folkes is a Senior Analyst in the Results-Based Management Directorate at Treasury Board Secretariat. She holds a Master's Degree in Public Administration from the University of Toledo (Ohio). Her research interests include innovations in evaluation methodology and she is currently working on improving reporting to Parliament.

Les Szabo, B.A., M.Sc., is currently a consultant in the area of program evaluation. He has 35 years of experience with the Canadian government in policy research, analysis and evaluation.