

**Combining Registration-System and Survey Data
to Estimate Birth Probabilities~**

Mark S. Handcock*, Sami M. Huovilainen, and Michael S. Rendall*****

25 January 2000

~ Financial support from U.K. Leverhulme Trust grant F/353/G, from NICHD grant R-01-HD34484-01A1, and from NICHD Center Grant P-30-HD-29263 is gratefully acknowledged. The British Household Panel Survey data used in this article were made available through the Data Archive, and were originally collected by the ESRC Research Centre on Micro-Social Change at the University of Essex. Neither the original collectors of the data nor the Data Archive bear any responsibility for the analyses or interpretations presented here. An earlier version of this article was presented at the 1999 Annual Meeting of the Population Association of America, and benefitted from comments received there and from two anonymous reviewers.

* Department of Statistics and Center for Statistics and the Social Sciences, University of Washington, WA 16802-2111.

** Department of Statistics and Population Research Institute, Pennsylvania State University, PA 16802-2111.

*** Department of Sociology and Population Research Institute, Pennsylvania State University, PA 16802-6211.

Abstract

With the widespread availability of event history data, demographers have increasingly eschewed registration-system data in favour of survey data. We propose instead using survey and registration-system data in combination, via a constrained maximum likelihood framework for demographic hazard modeling. As an application, we combine panel survey data and birth registration data to estimate annual birth probabilities by parity. The General Fertility Rate obtained from registration-system data serves to constrain the weighted sum of parity-specific birth probabilities. The variances about the parity-specific birth probabilities are halved when registration-system data are used to constrain the estimates. Other demographic applications are discussed.

Frequently, demographers have access to complete or near-complete annual counts of births, deaths, marriages, and divorces from registration systems. Yet with widespread availability of event-history data, demographers have increasingly eschewed registration-system data in favour of survey data. Survey data contain more detail with respect both to the demographic events themselves, and to the attributes of the individuals who do, and do not, experience the events. In the case of demographic events relating to family status, event detail is becoming increasingly important as family types become more varied. Registration systems record only marriage and divorce, and not the formation and dissolution of non-marital unions; and record births often by formal marital status of the mother. Yet registration-system data have an important advantage over survey data: They are collected for all events in a population. This reduces or eliminates sampling error (depending on whether a super-population or finite-population approach is taken), and eliminates sample-selection biases. Registration-system data collection may also reduce reporting biases, especially in comparison to retrospectively collected data.

Instead of discarding the less detailed registration system data, demographic analyses might exploit them with methods that formally combine registration-system and sample-survey data. Imbens and Lancaster (1994) use this approach with economic data, and report large gains in efficiency by incorporating marginal moments from census data with sample-survey joint distributions (see also Qin and Lawless 1994). The availability of population counts of demographic events in registration-system data, combined with near-complete exposed-population counts in population censuses, may allow for even greater scope for efficiency gains to be realized in demographic applications.

Another concern with the reliance on survey data alone is bias. Schoen and Weinick (1993) contrast registration-system divorce estimates with survey-based estimates incorporating complicated corrections for divorce under-reporting in surveys. Studies of men's fertility histories are also subject to under-reporting (see Rendall, Clarke, Ranjit, Peters, and Verropoulou 1999, and references therein). Hellerstein and Imbens (1999) demonstrate the bias-reduction possibilities of including aggregate data in the context of estimating wages from survey data.

We use a similar technique to those developed by Imbens and colleagues, adapted for demographic applications involving estimation of probabilities of transition between

discrete demographic statuses. These transitions may be any for which registration-system data are also available, i.e., deaths, births, marriages, divorces, changes of residence, etc. We show how to estimate such transition probabilities by adapting maximum likelihood estimation of discrete or continuous hazard models to a constrained maximum likelihood framework, in which the constraint(s) come from registration-system data. We further show that widely available software can be used for this estimation.

As an example, we follow up on Bongaarts and Feeney's (1998) complaint that registration systems in much of Western Europe collect birth data by parity only within marriage. We show that the registration-system data of one such offending country (England and Wales) may nevertheless be used to improve the efficiency of estimation from survey data of annual birth probabilities of all women by parity. The function used to derive the birth probability is a binomial logit with a single regressor for whether the woman is childless. The General Fertility Rate (GFR) is used to constrain the weighted sum of the birth probabilities for women alternately with and without children. To keep our example simple, we choose not to use additional registration-system information, such as period- and age-specific fertility rates, that might further improve our survey-based estimates. In the course of the paper, we describe how such additional information can be incorporated within the framework provided here.

To evaluate the gains obtained by our approach, we estimate also an alternate, unconstrained version of our model. Because the GFR is an overall mean, we expect and find gains in efficiency of estimating the intercept parameter of our logistic regression. Imbens and Lancaster (1994) give a simple example of how efficiency gains in estimating the intercept parameter may be easily obtained by using aggregate data for the overall mean. But they note that the intercept parameter is seldom of interest to economists, who are typically interested in marginal responses to changes in covariates. In contrast, the intercept parameter typically does matter in demographic applications because it figures importantly in the calculation of the levels of all transition probabilities. Levels are critical whenever the transition probabilities are to serve as inputs to projections of spells, life courses, etc. In our example, since the birth probabilities for women either with or without previous children depend on the precision

of estimation of the intercept parameter, we expect and find gains in the efficiency of estimating both birth probabilities.

The General Estimation Method

We apply nonlinear constrained optimization techniques to the special case of maximum likelihood (ML) estimation. Maximum likelihood estimators are a mainstay of demographic hazard modeling due to their good statistical properties and because likelihood-based methods provide a general conceptual and inferential framework. Imbens and colleagues (Imbens and Lancaster 1994; Hellerstein and Imbens 1999) instead use a Generalized Method of Moments estimator (GMM, Manski 1988) to incorporate the information from aggregate data as population moments that constrain the estimation. They show elsewhere (Lancaster and Imbens 1996) the asymptotic equivalence between the GMM and ML estimators, noting that the GMM approach avoids computational problems of constrained maximization likelihood. We overcome these computational problems by implementing our constrained ML estimator using commercially available nonlinear programming (NLP) software (SAS Institute 1997). This provides a flexible environment for imposing a wide range of objective functions and nonlinear or linear constraints. The NLP procedure also calculates standard errors. The use of this software should greatly facilitate the incorporation of the methods we describe here into common demographic practice.

We maximize the likelihood function subject to constraints that arise from the registration data. Our objective function is a nonlinear log-likelihood function, $\log(L(\theta; \mathbf{y}, \mathbf{x}))$, where θ is a vector of parameters representing population characteristics of interest, y is an outcome variable, and \mathbf{x} is a vector of predictor variables. In the case that the registration information is not used, we solve the unconstrained optimization problem

$$\max_{\theta} \log(L(\theta; \mathbf{y}, \mathbf{x})) \tag{1}$$

to find the maximum likelihood estimate of θ . Suppose the registration system provides J aggregate values representing population characteristics that are known functions of θ . Denote the j^{th} such function by $C_j(\theta)$, and the aggregate value of that function by r_j . We then solve the problem of maximizing equation (1) subject to a set of constraints:

$$C_j(\theta) = r_j, \quad j = 1, \dots, J \tag{2}$$

The nature and number of constraints depends on the model we wish to estimate and on the form of registration-system data we have available. In some applications the relationship between population parameters θ and registration value r_j cannot be expressed as an equality, but can be expressed as a bound. That is, the constraint on the parameters may be expressed as an inequality such as $C_j(\theta) < r_j$. Combinations of equality and one- or two-sided inequality constraints may also be used in this framework.

Typically the likelihood and constraint functions are smooth, i.e., the second order partial derivatives of the functions exist. Thus the optimization problem specified by equations (1) and (2) can be efficiently solved using standard, quasi-Newton methods. This is the case in our example below. Were the functions not smooth, the Nelder-Mead simplex method could be used (see Gill, Murray, and Wright 1981 for exposition of these and other numerical optimization methods).

Example: Binomial Logit for Annual Birth Probabilities by Parity

Our universe consists of all women of childbearing age in a given population. The response variable Y has two levels: 0 denotes no birth, and 1 denotes a birth, during the year $[t-1, t)$. The only predictor in this model is a dichotomous “parity” variable X for whether the woman is childless ($X=0$) or has had at least one child ($X=1$) at time $t-1$. The binomial logit model for the birth probability $P(Y=1|X=x)$ is:

$$P(x) = \exp(\beta_0 + \beta_1 x) / [1 + \exp(\beta_0 + \beta_1 x)]. \quad (3)$$

Thus if we knew parameters β_0 and β_1 , we could calculate the probabilities of a birth for a childless woman, $P(0) = \exp(\beta_0) / [1 + \exp(\beta_0)]$, and for a woman with children, $P(1) = \exp(\beta_0 + \beta_1) / [1 + \exp(\beta_0 + \beta_1)]$. If, in addition, we knew the proportion of women in each parity category, π and $1 - \pi$, we could then use these expressions for $P(0)$ and $P(1)$ to calculate the annual probability of childbearing among all women as:

$$P = P(0)\pi + P(1)(1 - \pi) \quad (4)$$

The form of this equation is very general in demographic applications. It expresses an overall or “crude” rate P as a weighted sum of covariate-dependent or “specific” rates $P(0)$ and $P(1)$. The weights are given by the population distribution of the covariate $\{ \pi, 1 - \pi \}$. This type of equation is used in demographic standardization and decomposition. As a typical example, a Crude Death Rate may be expressed as a weighted sum of age-

specific death rates, where the age distribution of the population provides the weights (Smith 1992:63). A crucial difference is that our conditional probabilities are additionally nonlinear functions of regression parameters β_0 and β_1 . We further show that the parameter π for the covariate distribution $\{\pi, 1-\pi\}$ in the constraint equation may be estimated jointly with the regression parameters.

The birth probability among all women, P , may be approximated by the General Fertility Rate (GFR) estimated from registration data. The mid-year population of 15 to 44 year-old women approximates the group who were at risk of a birth at the beginning of a given year, assuming negligible mortality of women in the course of a year. The number of births in a year approximates the number of women of this group who actually give birth in the year, assuming negligible occurrences of multiple births to a woman in a single calendar year.

We assume that our survey data are a simple random sample of n person-years, drawn from the population of women of childbearing age. Therefore an unbiased estimate of the covariate distribution parameter π may be obtained from the survey data. Conveniently, this can be achieved by specifying π as an additive component in the log-likelihood function. Additivity in the log-likelihood holds for any covariate vector \mathbf{X} , following from the standard assumption that the conditional distribution of Y given \mathbf{X} is independent of the distribution of \mathbf{X} .

Let $\{y_i, x_i\}$ be the data for the i^{th} person-year, and let $\theta = (\beta_0, \beta_1, \pi)$. The constrained optimization estimation of the binary logit model of equation (3) is then achieved by maximizing the log-likelihood:

$$\begin{aligned} \log[L(\theta; \mathbf{y}, \mathbf{x})] = & \sum_i^n y_i \cdot (\beta_0 + \beta_1 x_i) - \sum_i^n \log[1 + \exp(\beta_0 + \beta_1 x_i)] \\ & + \sum_i^n (1 - x_i) \cdot \log(\pi) \end{aligned} \quad (5)$$

subject to:

$$\text{GFR} = \left\{ \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \right\} \cdot \pi + \left\{ \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} \right\} \cdot (1 - \pi) \quad (6)$$

In the unconstrained version of the estimation, the optimization problem is simply to maximize the log-likelihood:

$$\log[L(\theta; \mathbf{y}, \mathbf{x})] = \sum_i^n y_i (\beta_0 + \beta_1 x_i) - \sum_i^n \log[1 + \exp(\beta_0 + \beta_1 x_i)] \quad (7)$$

The constrained maximum likelihood problem given by equations (5) and (6) is just a special case of maximizing equation (1) with constraint equations (2). Similarly, the

unconstrained version of our model is just a special case of maximizing equation (1). Note that the approach does not require the birth-probability equation to be specified as a binary logit. For example, a probit or proportional-hazard specification could be used. Similarly, there is nothing restrictive about the particular constraint of equation (6). Aggregate data could also be used to impose a constraint on, for example, one or more of a set of covariate-dependent probabilities. Our data are pooled over several years. The pooled GFR could be expressed as a weighted sum of year-specific GFRs as well as of parity, with the year-specific GFRs taken from the registration-system data. In this case, efficiency gains would also be expected in the estimation of covariate parameters (for period trends). Census data could also be used to constrain the population distribution of components of covariate vector \mathbf{x} . More generally, inequality constraints could be added.

The effects of the particular constraint we have imposed (the pooled GFR) will be limited to the intercept parameter β_0 , as the constraint applies to the overall level of the birth probabilities, and not to the differential by parity (which is captured by covariate parameter β_1). Gains with respect to both efficiency and bias of the estimator of β_0 are possible. These gains will, in turn, apply to the efficiency and bias of both $P(0)$ and $P(1)$, as both are affected by the intercept parameter.

Data and Estimation

The registration-system data are for all births in 1992 to 1996 to women in England and Wales (Office for National Statistics 1998). Parity information is provided only for married women. Thus we do not know with certainty the order of the birth for any woman. We sum the 1992 to 1996 births and the 1992 to 1996 mid-year populations of 15 to 44 year-old women to calculate a 1992 to 1996 General Fertility Rate (GFR): 0.06179. The mid-year population estimates are based on the 1991 Census of England and Wales (found also in Office for National Statistics 1998). Thus we assume the GFR calculated from these data has zero sampling variance.

Our survey data are from the British Household Panel Survey (BHPS, Taylor et al 1995). We exclude women living in Scotland to maintain consistency with the birth-registration system's geographical boundaries. The survey design consists of annual

panel observation, beginning in 1991. We use the years 1991 to 1996. Births are not directly recorded in these data. Instead we code births in the 1992 to 1996 period for women aged between 15 to 44 years old when the woman's co-resident family unit experiences an increase in the number of dependent-aged children from one year ($t-1$) to the next (t). This assumes that the mother and child live together at the survey interview immediately following the birth, and that infant mortality between birth and survey is zero. We code whether a woman is childless or not at $t-1$ from whether one or more children are in her co-resident family unit. Any general upward or downward bias on the birth probabilities is corrected for by our constrained optimization method, whereby the weighted sum of the parity-specific birth probabilities is constrained to equal the unbiased overall birth probability represented by the GFR.

[Table 1 about here]

The BHPS uses an approximately equal-probability sample design. We take advantage of this to conduct our estimation without the added complication for variance computation of using sample weights. Thus we ignore initial sample design effects and biases introduced by attrition. We select our sample to consist only of those women who responded in all six years, thus excluding all attriters. This results in a sample of 2,316 women who were ever aged between 15 and 44 in the 1992 to 1996 period. Because we include their years of observation only when aged between 15 and 44, our potential person year-pairs sample is slightly less than five times 2,316. We ignore variance-estimation complications due to the repeated observation of individuals in the panel. Because we use the same data, without weights, for both the constrained and unconstrained estimates, introducing these complications should not change our main results.

[Figures 1 and 2 about here]

We use these survey data to conduct a Monte Carlo simulation study. We draw 2,000 samples of 700 women from our total of 2,316 to estimate the constrained and

unconstrained models. Table 1 illustrates a typical sample. The total number of year-pairs is 2,966, of which 179 include a birth. The results of estimating the constrained and unconstrained models 2,000 times each are given in box-plots in Figures 1 and 2, and in summary form as ratios of the constrained and unconstrained model variances in Table 2. The distributions of the regression parameter estimates and of their standard errors are given in box plots. The box portion is defined by the 1st and 3rd quartiles. The line in middle of the box is the median. The more symmetric the distribution, the more centered is the median in the box. The location of b_0 is seen to be little changed after imposing the constraint. This indicates that the survey sample is not biased relative to the registration-system data.

[Table 2 about here]

As expected, the variance of b_0 is clearly smaller with the imposition of the GFR constraint, while the variance about b_1 is unchanged. We used the delta method (Agresti 1990) to obtain asymptotic standard errors for the birth probabilities $P(0)$ and $P(1)$, respectively for childless women and women of parity one. From Figure 2, it is clear that both probabilities have a much tighter distribution. Table 2 summarizes the efficiency gains in terms of the ratio of the variance of the estimates from the constrained to the unconstrained model. The variance of the intercept parameter estimate b_0 in the constrained model is only 48.2% as large as that in the unconstrained model. The variance ratio for the slope parameter b_1 meanwhile is 97.5%, implying no variance reduction. Since the birth probabilities are functions of both parameters, the reduction in variance in the constrained model is similarly large for both birth probabilities: 53.4% and 40.7% respectively of the variance for the unconstrained-model estimates of $P(0)$ and $P(1)$. That is, parity-specific birth probabilities estimated from survey data are re-estimated with a halving of their variance by introducing a GFR constraint.

Discussion

We have shown with a simple empirical example that parity-specific birth probabilities estimated from panel survey data may be re-estimated with a large reduction of their

variance simply by introducing a GFR constraint. The model we estimated is simpler than one that would be specified in substantive applications. As covariates are added, more constraints from registration-system data could be incorporated. We described, for example, how period covariates could be added and constrained. The joint distribution of the covariates would then need to be estimated. We have shown that this can be done easily by incorporating the covariate distribution as a set of additive components of the log-likelihood of equation (5). The general form given by equations (1) and (2) also allows for continuous covariates. Imbens and Lancaster (1994) prove explicit formula for the variance reductions relevant to these general models.

Other demographic hazard processes that could be estimated in this constrained-optimization framework for combining survey and registration-system data are death, marriage and divorce, and migration. For example, marriage and divorce statistics in the United States are now available nationally only with respect to total numbers of marriage and divorces. These numbers, however, can be combined with census-based estimates of the total numbers at risk to provide unbiased estimates of the national marriage and divorce rates (National Center for Health Statistics 1998). Divorce is known to be under-reported in survey retrospective histories, and correlated with respondent attrition in panel studies. Therefore divorce probabilities will be downwardly biased when derived from a hazard model that is estimated from survey data. A constrained binary logit model of the annual probability of divorce similar to that of the model of the birth-probability example given earlier would eliminate this bias, while simultaneously reducing sampling variance. A more general treatment of bias reduction by combining aggregate and survey data is given in Hellerstein and Imbens (1999).

Finally, while the methods here are adapted from the econometrics (and statistics) literature, we note also a connection to an established literature in demographic techniques. Region-specific “model life tables” are appropriately used to improve the efficiency of a particular country’s life table estimates when “...life tables are produced from sample data in which the number of events reported is small” (Smith 1992:191). In that case, bias is introduced by using as the larger, constraining data set, a population that does not correspond exactly to the sampled population. This bias is traded off against a reduction in sampling variance. Similarly, with indirect standardization, the demographic

practitioner is counseled for reasons of statistical efficiency (Smith 1992:58) to use the more populous country's data to provide the pattern of conditional probabilities of death by age. Schoen and Weinick (1993) similarly apply indirect standardization techniques to impute national age-conditional probabilities of marriage and divorce by combining the age-conditional probabilities from the Marriage and Divorce Registration Area subsets of states with national total marriage and divorce rates and national age distributions. In the context of these established demographic techniques, the constrained optimization model of the present paper may be seen as a more general method for combining demographic data collections within a regression framework.

References

- Agresti, A. (1985) Categorical Data Analysis New York: Wiley.
- Bongaarts, J., and G. Feeney (1998). On the quantum and tempo of fertility Population and Development Review 24(2):271-291.
- Gill, P., M.H. Wright, and W. Murray (1981) Practical Optimization London: Academic Press.
- Hellerstein, J., and G.W. Imbens (1999) Imposing moment restrictions from auxiliary data by weighting Review of Economics and Statistics 81(1):1-14.
- Imbens, G.W. and T. Lancaster (1994) Combining micro and macro data in microeconomic models. Review of Economic Studies 61: 655-680.
- Lancaster, T., and G.W. Imbens (1996) Case control studies with contaminated controls Journal of Econometrics 71:145-160.
- Manski, C. (1988) Analog Estimation Methods in Econometrics New York: Chapman Hall.
- Office for National Statistics (1998) 1997 Birth statistics London: HMSO.
- National Center for Health Statistics (1998) Births, Marriages, Divorces, and Deaths for 1997 Monthly Vital Statistics Report 46(12). Hyattsville, MD: National Center for Health Statistics.
- Qin, J., and J. Lawless (1994) Empirical likelihood and general estimating equations Annals of Statistics 22(1):300-325.
- Rendall, M.S., L. Clarke, H. E. Peters, N. Ranjit, and G. Verropoulou (1999) Incomplete Reporting of Male Fertility in the United States and Britain: A Research Note. Demography 36(1):135-144.

SAS Institute (1997) SAS/OR Technical Report: The NLP Procedure Cary, NC:

SAS Institute Inc.

Schoen, R., and R. M. Weinick (1993) The Slowing Metabolism of Marriage:

Figures from 1988 Marital Status Lifetables Demography 30(4):737-746.

Smith, D.P. (1992) Formal Demography New York: Plenum.

<u>Presence of Children (t-1)</u>	Frequency	Y=0 for no births.		Total
	Percent	Y=1 for a birth.		
	Row Pct	0	1	
	Col Pct			
No children at time (t-1)		1419	71	1490
		47.84	2.39	50.24
		95.23	4.77	
		50.91	39.66	
At Least one child at time (t-1)		1368	108	1476
		46.12	3.64	49.76
		92.68	7.32	
		49.09	60.34	
Total		2787	179	2966
		93.96	6.04	100.00

Table 1. Frequency table for births between times (t-1) and (t) by presence of children at time (t-1), taken from one Monte Carlo sample of 700 women.

Source: British Household Panel Survey, 1991 to 1996 waves.

Binomial Logit Models

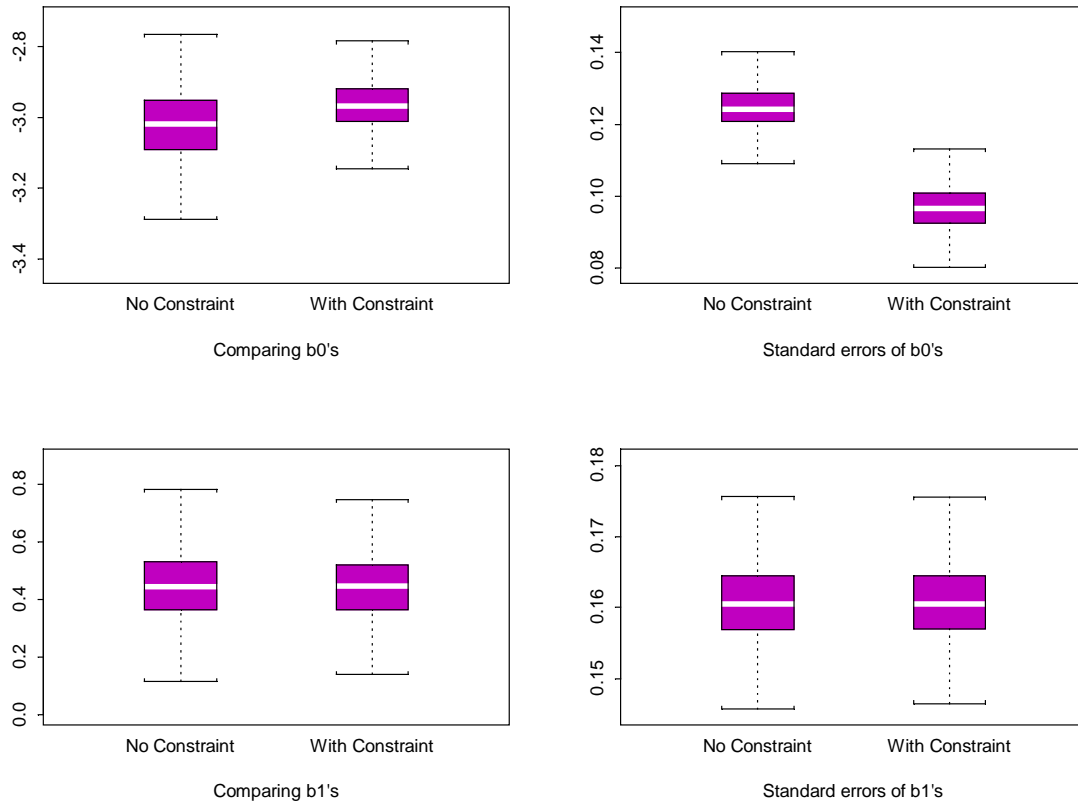


Figure 1. Box-plots for the parameter estimates and standard errors of the single predictor binomial logit model. The predictor variable is the presence of children at time (t-1). The response variable is a dichotomous variable that describes whether or not a birth takes place between times (t-1) and (t).

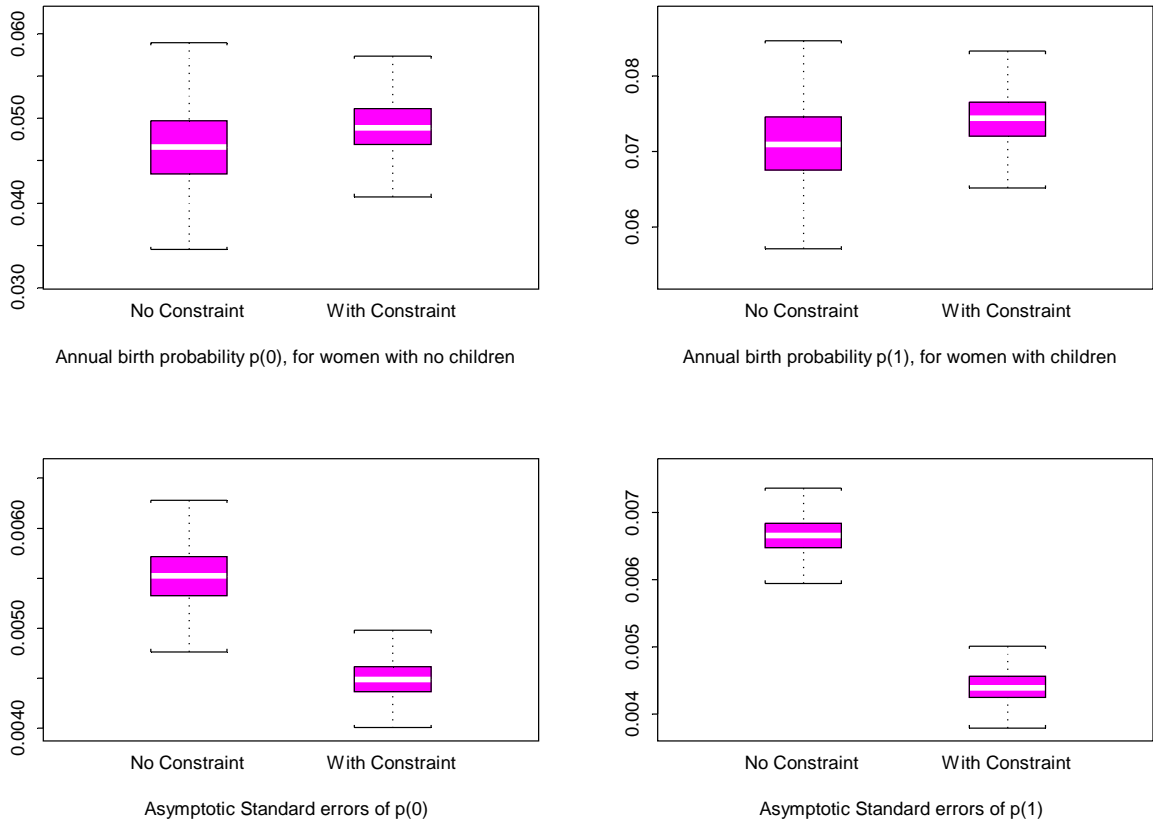


Figure 2. Box-plots for the birth probability estimates and for the asymptotic standard errors of the parameters. $P(0)$ and $P(1)$ are the probabilities of birth between $t-1$ and t , respectively for a woman with no children and a woman with children.

Parameter	Variance Ratio*100
b_0	48.2
b_1	97.5
P(0)	53.4
P(1)	40.7

Table 2. Variance ratios of constrained to unconstrained models' estimates.