A bivariate model of the number of children and the age at first birth

Jacek Osiewalski[1], Beata Osiewalska[2]

Abstract

We formulate a joint statistical model of two important demographic variables: (i) the number of children born by a given woman and (ii) her age at the birth of her first child. The proposed specification is based on the so-called ZIP-CP model of bivariate Poisson-type regression that enables to easily examine dependence between two count variables. In our specification the number of children is a ZIP-type variable (in the hurdle model version), while the conditional distribution of the age at first childbirth given the number of children is a Poisson distribution either left-truncated (when a woman has not had any child) or right-truncated (if a woman gave birth to at least one child). The expected values of the underlying Poisson distributions as well as the relation between both variables are functions of the age of a woman and some socio-economic explanatory variables.

***Keywords and Phrases*:** *count data models, bivariate Poisson regression models, Bayesian inference, fertility, fertility forecasting*

JEL Classification: J13, C35, C51

1 Introduction

Identifying socio-economic factors determining fertility (as well as its forecasting) is a crucial issue of current demographic research. The very low level of fertility observed nowadays in many European countries might be caused by the combination of both the postponement of childbearing to older age (the tempo effect) and the tendency to have smaller families (the quantum effect) (Bongaarts and Feeney, 1998; Sobotka, 2003). Examining which of these two effects plays a major role in the reproductive behaviour of a contemporary woman is necessary to effectively forecast her completed fertility (Lee, 1981; Kohler et al., 2001).

It seems that the completed family size and the age at first birth are usually negatively related, meaning that women who enter motherhood at late ages have fewer children (Trussell and Menken, 1978; Kohler et al., 2001). There are many biological and social reasons for such negative correlation (Leridon and Slama, 2008; Schmidt et al., 2012), nevertheless the dependence may change with contextual factors and socio-economic characteristics of a woman (Neels and De Wachter 2010; Berrington et al., 2015). Thus, it is important to jointly model the two basic variables: age at entry into motherhood and the number of

[1] Corresponding author: Cracow University of Economics, ul. Rakowicka 27, 31-510 Kraków, Poland, e-mail: eeosiewa@cyf-kr.edu.pl.

[2] Cracow University of Economics, ul. Rakowicka 27, 31-510 Kraków, Poland, e-mail: beata.osiewalska@uek.krakow.pl.

children – first, in order to test their dependence and its possible changing character; second, to make statistical inferences more efficient when the tested dependence is present.

In this paper we propose a joint bivariate statistical model of the number of children born by a given woman and her age at the birth of her first child. The first variable can take only non-negative integer values, while the second can be treated either as a continuous variable or as a count variable. Here we assume that age is measured in full years (above some threshold, e.g. 15 years old), so we jointly model two count variables. By adopting such an approach we can take advantage of recent specifications proposed in the statistical literature.

Modelling univariate count data by means of Poisson type regression models is nowadays a routine approach, and several specifications have been proposed for the bivariate case; see e.g. Cameron and Trivedi (1998, 2005). In this paper we modify the so-called ZIP-CP (*zero inflated Poisson – conditional Poisson*) model, analysed by Osiewalski (2012) and Osiewalski and Marzec (2016), which is the generalised version of the P-CP (*Poisson – conditional Poisson*) specification, proposed by Berkhout and Plug (2004). We replace the regular Poisson conditional part by a more appropriate distribution of the age at first childbirth given the number of children. This conditional distribution is a Poisson distribution, either left-truncated (when a woman has not had any child) or right-truncated (if a woman gave birth to at least one child). We obtain the likelihood function for our non-standard bivariate Poisson-type model, formulate important parametric hypotheses and consider forecasting issues. In order to conduct exact small-sample inference, we propose the Bayesian approach equipped with MCMC simulation tools. A preliminary empirical illustration is based on the *Generations and Gender Survey* (GGS) data for Poland.

In the next section we present the probabilistic foundations of our model, i.e. the discrete bivariate distribution used to jointly describe two demographic variables. Section 3 is devoted to our statistical model, the form of the likelihood function and the Bayesian analysis. Section 4 contains a preliminary empirical example.

2 Foundations of the new statistical model

We consider the joint distribution of two random variables (Y_1, Y_2) that can take non-negative integer values. In the bivariate P-CP distribution analysed by Berkhout and Plug (2004) the probability distribution of (Y_1, Y_2) is as follows:

$$\Pr\{Y_1 = i, Y_2 = j\} = \Pr\{Y_1 = i\} \Pr\{Y_2 = j \mid Y_1 = i\} = g(i)\, h(j,i), \qquad (1)$$

where $i, j \in N \cup \{0\}$, the marginal distribution of Y_1 is Poisson with mean and variance λ_1, and the conditional distribution of Y_2 given Y_1 is Poisson with mean and variance $\lambda_2\exp(\alpha Y_1)$, i.e.

$$g(i) = \exp(-\lambda_1)(\lambda_1)^i / i!, \quad h(j,i) = \exp[-\lambda_2 \exp(\alpha i)](\lambda_2)^j \exp(\alpha ij) / j!. \tag{2}$$

If $\alpha \neq 0$, then two count variables are stochastically dependent and the variance of Y_2 is greater than its expectation. The dependence between these variables leads to the inflated variance of Y_2, which is often observed in empirical count data. The Poisson distribution of Y_1 does not have this property. Also, the P-CP model puts restrictions on the dependence of two variables, as the sign of covariance between Y_1 and Y_2 depends only on the sign of α, and not on λ_1 or λ_2, which are described by explanatory variables in statistical applications. An appropriate generalisation was proposed by Osiewalski (2012); it allows for the dependence of the sign of covariance on λ_1. This more general class of distributions (called ZIP-CP) is characterized by the same conditional distribution of Y_2 given Y_1, $\Pr\{Y_2 = j \mid Y_1 = i\} = h(j,i)$, and by the ZIP-type distribution of Y_1, with zero treated separately:

$$\Pr^*\{Y_1 = i\} = g^*(i) = \begin{cases} \gamma & for\ i = 0, \\ \dfrac{1-\gamma}{1-g(0)}\, g(i) & for\ i \in N, \end{cases} \tag{3}$$

where γ belongs to the (0, 1) interval, and g and h are the functions as in (2). If $\gamma = g(0)$, then $\Pr^*\{Y_1 = i\} = g^*(i) = g(i) = \Pr\{Y_1 = i\}$; we have the P-CP case. If $\gamma > g(0)$, the distribution of Y_1 is of the ZIP type. However, the specification (3) is more general as it also allows $\gamma < g(0)$; it is known in the literature as the *hurdle model* (Cameron and Trivedi, 2005, p. 680) and is compared to the original ZIP model by Winkelman (2008). The *hurdle model* form of our ZIP type specification for Y_1 leads to a very simple statistical specification, making estimation – as well as testing of the standard Poisson case – relatively easy. The ZIP-CP distribution enables inflating variances of both count variables (although they are not symmetrically treated) and making their dependence more complex than in the P-CP case.

As yet we have not focused on the interpretation of Y_2 (the age at first birth); this variable cannot have the regular Poisson conditional distribution given Y_1. Let a be the actual age of a woman and let b denote the end of the reproductive age. We assume that the conditional distribution of Y_2 given Y_1 is Poisson, but either right-truncated at $\min\{a, b\}$ (if $Y_1 > 0$) or left-truncated at a (if $Y_1 = 0$). In the first case (a woman has already had at least one child) the age at first birth has to be between the beginning of the reproductive age and either woman's current age a or the end of the reproductive age b (whichever is smaller). In the latter case, when a woman has had no children ($Y_1 = 0$), the age at first birth has to be between a and b.

Thus $\Pr\{a \le Y_2 \le b \,|\, Y_1 = 0\}$ is the probability that a woman, which is childless at age *a*, will have a child (and obviously equals to zero if *a*>*b*), while $\Pr\{Y_2 > b \mid Y_1=0\}$ is the probability that she will remain childless (it equals to one if $a \ge b$). For further consideration, assume that age is counted in full years exceeding some threshold, e.g. in years over 15; that is, *b*=34 (as the reproductive age of a woman is [15, 49]) and at this threshold *a*=0 and Y_2=0.

While the interpretation of Y_1 is straightforward (the number of children ever born by a woman), the meaning of Y_2 is more subtle, as only its values up to *b* can represent the age at first birth. In the conditional distribution of Y_2 given Y_1=0, which is not truncated at *b*, the values above *b* serve to describe childlessness after the reproductive age, by attaching a positive probability to such situation. Summing up the assumptions we have already introduced, we propose the following joint distribution of Y_1 and Y_2:

$$\Pr^*\{Y_1 = i, Y_2 = j\} = \begin{cases} \gamma \dfrac{h(j,0)}{1-\sum_{l=0}^{a-1} h(l,0)}, & i = 0 \wedge j \ge a, \\ \dfrac{1-\gamma}{1-g(0)} \dfrac{g(i)\,h(j,i)}{\sum_{l=0}^{\min\{a,b\}} h(l,i)}, & i > 0 \wedge j \le \min\{a,b\}, \\ 0, & i > 0 \wedge j > \min\{a,b\},\ i = 0 \wedge j < a. \end{cases} \tag{4}$$

3 The statistical model, its likelihood function and Bayesian analysis

The statistical model proposed in this paper is designed to cope with cross-section micro-level data for women that may differ in terms of age and other characteristics. Consider *K* bivariate observations (Y_{1k}, Y_{2k}; k = 1,2,...,*K*), where Y_{1k} is the number of children born by the *k*-th woman and Y_{2k} denotes her age at first birth (in full years over 15). The pairs (Y_{1k}, Y_{2k}) are independent and have different distributions with the probability function of the general form (4), so our model amounts to the following parametric class of distributions:

$$\Pr^*\{Y_{1k} = i, Y_{2k} = j;\ \theta\} = \begin{cases} \gamma_k \dfrac{h_k(j,0)}{1-\sum_{l=0}^{a_k-1} h_k(l,0)}, & i = 0 \wedge j \ge a_k, \\ \dfrac{1-\gamma_k}{1-g_k(0)} \dfrac{g_k(i)\,h_k(j,i)}{\sum_{l=0}^{\min\{a_k,b\}} h_k(l,i)}, & i > 0 \wedge j \le \min\{a_k,b\}, \\ 0, & i > 0 \wedge j > \min\{a_k,b\},\ i = 0 \wedge j < a_k, \end{cases} \tag{5}$$

where a_k is the actual age of the k-th woman (in full years over 15),

$$g_k(i) = \exp(-\lambda_{1k})(\lambda_{1k})^i / i!, \quad \lambda_{1k} = \exp(x_k\beta_1), \quad \gamma_k = \exp(-e^{z_k\delta}\lambda_{1k}) = \exp(-\exp(z_k\delta + x_k\beta_1)),$$

$$h_k(j,i) = \exp[-\lambda_{2k}\exp(\alpha_k i)](\lambda_{2k})^j \exp(\alpha_k ij)/\, j!, \quad \lambda_{2k} = \exp(w_k\beta_2), \quad \alpha_k = s_k\beta_3.$$

In the above formulas x_k, z_k and w_k are row vectors of explanatory variables that determine the marginal probabilities of Y_{1k} and conditional probabilities of Y_{2k} given Y_{1k}, respectively, while s_k is a row vector of explanatory variables explaining possible differences in dependence between Y_{2k} and Y_{1k} for different groups of women. The age variable, a_k, seems an obvious explanatory variable, appearing in all four vectors – x_k, z_k, w_k and s_k. Obviously, the role of the explanatory variables depends on the column vectors of parameters β_1, β_2, β_3 and δ, grouped in θ, the vector of all parameters. In particular, stochastic independence between the number of children (Y_{1k}) and mother's age at first birth (Y_{2k}) is equivalent to $\beta_3=0$, while $\delta\neq 0$ means that $\Pr^*\{Y_{1k}=0;\ \theta\}$ deviates from the value corresponding to the Poisson distribution with mean (and variance) λ_{1k}.

When specifying the likelihood function that corresponds to (5) we have to remember that the age at first birth (Y_{2k}) is not observed when the woman has not born a child ($Y_{1k}=0$). Thus, the likelihood function is the product of $K=K_0+K_1$ factors, where K_0 factors (of the form γ_k) correspond to the probability of zero in the marginal distribution $\Pr^*\{Y_{1k}=i;\ \theta\}$ and K_1 factors correspond to the joint probability (5) for $i > 0$ and $j \le \min\{a_k, b\}$. Denote the observed values of Y_{1k} and Y_{2k} as y_{1k} and y_{2k}, respectively; then the likelihood function takes the form

$$L(\theta; y) = \left[\prod_{k:\, y_{1k}=0} \gamma_k\right] \left[\prod_{k:\, y_{1k}>0} \frac{1-\gamma_k}{1-g_k(0)} \frac{g_k(y_{1k})h_k(y_{2k}, y_{1k})}{\sum_{l=0}^{\min\{a_k, b\}} h_k(l, y_{1k})}\right], \tag{6}$$

where y groups all the values y_{1k} and y_{2k}. The likelihood function (6) enables us to test many specific parametric hypotheses, but the most important from the theoretical point of view is the one stating that $\beta_3=0$. If $\beta_3=0$, Y_{1k} and Y_{2k} are independent random variables that lead to two separate models and likelihood functions: one built for $K=K_0+K_1$ values y_{1k} and involving β_1 and δ, the other built for K_1 values y_{2k} and involving β_2. Only for $\beta_3\neq 0$ our joint bivariate model can lead to inferential gains and makes joint forecasting of both demographic variables more efficient than treating them separately.

Our inference on the parameters and unobserved values of both demographic variables will follow the Bayesian statistical approach, where a probability measure (prior distribution) on the parameter space is defined. We assume prior independence among all parameters in θ and the

standard normal prior $N(0, 1)$ for each individual parameter. Zero prior expectations mean that the simplest model (with no ZIP effect, no dependence and no explanatory variables) gets the highest prior chance, but unitary standard deviations ensure significant prior chances for specifications being far from the simplest one. It seems that such simple joint prior distribution brings little initial information and guarantees easy Monte Carlo simulations from the posterior distribution. Obviously, the sensitivity of inferences with respect to the form of the prior distribution is an empirical question, to be answered with the data at hand. Following Bayesian statistical paradigm makes our inference not only exact (small-sample), despite a non-standard form of the likelihood (6), but also coherent and intuitive.

The statistical analysis based on our model (5) can serve different purposes. First, using the posterior density $p(\theta \mid y) \propto p(\theta)L(\theta; y)$, where $p(\theta)$ denotes the prior density, we can test basic hypotheses and point at the explanatory variables that are most important in determining the number of children and the age at first birth. The simplest way to test hypotheses is the so-called Lindley-type approach, see e.g. Osiewalski and Marzec (2016).

Second, our model can serve different forecasting purposes. For a woman outside the dataset, but with given characteristics represented by the row vectors x_f, z_f, w_f and s_f, the predictive probability that $Y_{1f} = i$ and $Y_{2f} = j$ is obtained by averaging (5), interpreted as the conditional probability given θ, with $p(\theta \mid y)$ as the weight function:

$$\Pr^*\{Y_{1f} = i, Y_{2f} = j \mid y\} = \int_\Theta \Pr^*\{Y_{1f} = i, Y_{2f} = j \mid \theta\} p(\theta \mid y)\, d\theta\,. \tag{7}$$

In order to examine the explanatory power of our model and to infer on fertility of women with certain characteristics, we compute the predictive probabilities: marginal for $Y_{1f} = i$ and conditional for $Y_{2f} = j$ given $Y_{1f} > 0$:

$$\Pr^*\{Y_{1f} = i \mid y\} = \int_\Theta \Pr^*\{Y_{1f} = i \mid \theta\} p(\theta \mid y)\, d\theta\,, \tag{8}$$

$$\Pr^*\{Y_{2f} = j \mid y, Y_{1f} > 0\} = \sum_{i>0} \Pr^*\{Y_{1f} = i, Y_{2f} = j \mid y\} \Big/ \sum_{i>0} \Pr^*\{Y_{1f} = i \mid y\}\,; \tag{9}$$

they can be compared to observed frequencies.

Third, interesting forecasts can be made for women from the dataset. In this case we can consider one of the complimentary predictive probabilities: that a woman, which is childless at age a, will have a child or that she will remain childless. However, this is left for future research.

In order to simulate samples from the joint posterior distribution of θ, the vector of model parameters, and to approximate the integrals above, we will use the Metropolis-Hastings sequential chain, one of the Markov Chain Monte Carlo (MCMC) simulation techniques.

4 A preliminary empirical example

Our empirical example is based on the first wave GGS data for Poland. The survey was conducted in 2011 and includes respondents between 18 and 79 years old. The only purpose of our preliminary analysis is to check whether the proposed model can describe small data sufficiently well. Thus we have selected only women at the age of 33, who have completed tertiary education, are married and live in urban areas. Our final sample consists of only 52 women, of which 7 are childless (13,5%). The half of the women have one child (48,1%), one-third have two children (34,6%) and the rest (3,8%) have already three children. The most common age at first birth is 28 (22,2%). Almost half of the women gave birth after that age (46,7%), and one-third have the first child before the age of 28 (31,1%). At this stage we do not include any explanatory variables (only intercepts), thus the final model includes four scalar parameters β_1, β_2, β_3 and δ.

The basic characteristics of marginal posterior distributions (means, standard deviations, 0.05 and 0.95 quantiles) are presented in Table 1. All the distributions, besides the marginal posterior distribution of β_1, are separated from zero. In particular, it is *a posteriori* almost certain that β_3 is negative. This confirms the negative dependence between the number of children and the age at first birth; it also proves the necessity to jointly model these two variables. In addition, the parameter δ is positive (with very high posterior probability), thus the ZIP effect is present and childlessness seems more frequent than the standard Poisson distribution may capture.

θ_i	$E(\theta_i\|y)$	$D(\theta_i\|y)$	$q_{0.05}(\theta_i\|y)$	$q_{0.95}(\theta_i\|y)$
β_1	-0.144	0.212	-0.499	0.198
β_2	3.054	0.166	2.786	3.327
β_3	-0.256	0.099	-0.419	-0.095
δ	0.814	0.285	0.355	1.290

Table 1. Characteristics of marginal posterior distributions.

The marginal predictive distribution of the number of children, given by (8), is compared to the frequencies observed in the data and presented in Figure 1 (left-hand side). The model efficiently represents the data and properly predicts the number of children of a given woman with chosen characteristics. Although at this stage it underestimates the probability of having two children, we believe that the accuracy will be improved through enlarging the sample size

(by considering women of different age and other characteristics) and, therefore, including explanatory variables. The conditional predictive distribution (9) of the age at first birth and the frequencies observed in the data (for 45 women with at least one child) are shown in Figure 1 (right-hand side). As for the very limited sample, the model performs exceptionally well and accurately represents the data.

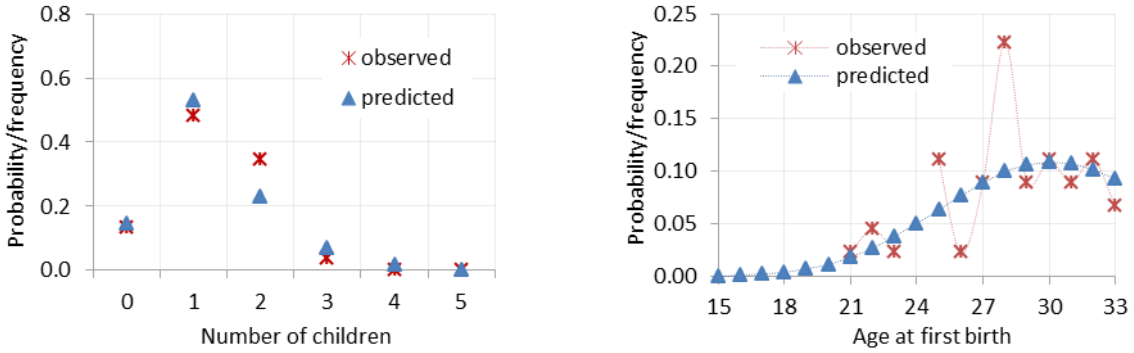

Fig. 1. The marginal predictive distribution of the number of children and the conditional predictive distribution of the age at first birth versus frequencies observed in the data.

To conclude, using our modified ZIP-CP model to jointly analyse the number of children and the age at first birth seems to be well justified by both the dependence between the two variables and the overrepresentation of zero (childlessness). The model also provides with reasonable predictions and thus serves a promising tool to infer about the fertility of a woman (on the basis of much larger and more informative samples).

Acknowledgement

The authors acknowledge support from research funds granted to the Faculty of Management at Cracow University of Economics, within the framework of the subsidy for the maintenance of research potential.

References

Berkhout, P., & Plug, E. (2004). A bivariate Poisson count data model using conditional probabilities. *Statistica Neerlandica*, 58, 349–364.

Berrington, A., Stone, J., & Beaujouan, E. (2015). Educational differences in timing and quantum of childbearing in Britain: a study of cohorts born 1940-1969. *Demographic Research*, 33(26), 733.

Bongaarts, J., & Feeney, G. (1998). On the quantum and tempo of fertility. *Population and development review*, 24(2), 271-291.

Cameron, A.C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge University Press, New York.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and application*. Cambridge University Press, New York.

Kohler, H. P., Skytthe, A., & Christensen, K. (2001). *The age at first birth and completed fertility reconsidered: findings from a sample of identical twins*. MPIDR Working Paper 2001-006. Max Planck Institute for Demographic Research, Rostock.

Lee, R. D. (1981). A model for forecasting fertility from birth-expectations data. In: Hendershot, G. E., & Placek, P. J. (eds.), *Predicting Fertility: Demographic Studies of Birth Expectations*, Lexington, MA: Heath, 75-99.

Leridon, H., & Slama, R. (2008). The impact of a decline in fecundity and of pregnancy postponement on final number of children and demand for assisted reproduction technology. *Human Reproduction*, 23(6), 1312-1319.

Neels, K., & De Wachter, D. (2010). Postponement and recuperation of Belgian fertility: how are they related to rising female educational attainment?. *Vienna Yearbook of Population Research*, 8, 77-106.

Osiewalski, J. (2012). *Dwuwymiarowy rozkład ZIP-CP i jego momenty w analizie zależności między zmiennymi licznikowymi*. In: *Spotkania z królową nauk (Księga jubileuszowa dedykowana Profesorowi Edwardowi Smadze)*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków, 147–154.

Osiewalski, J. & Marzec, J. (2016). Joint modelling of two count variables when one of them can be degenerate. Technical report, Cracow University of Economics.

Schmidt, L., Sobotka, T., Bentzen, J. G., & Andersen, A. N. (2012). Demographic and medical consequences of the postponement of parenthood. *Human reproduction update*, 18(1), 29-43.

Sobotka, T. (2003). Tempo-quantum and period-cohort interplay in fertility changes in Europe: Evidence from the Czech Republic, Italy, the Netherlands and Sweden. *Demographic Research*, 8(6), 151-214.

Trussell, J., & Menken, J. (1978). Early childbearing and subsequent fertility. *Family Planning Perspectives*, 10(4), 209-218.

Winkelman, R. (2008). *Econometric analysis of count data*. Springer-Verlag, Berlin Heidelberg.