Comparing two Age-Period-Cohort methodologies to model Grade Progression Probabilities*

Raquel Rangel de Meireles Guimarães[1]

Eduardo Luiz Gonçalves Rios-Neto[2]

Abstract

The aim of this article is to realize a comparative methodological exercise of two APC estimators: the conventional restrict estimator obtained by the generalized linear models (REGLM) and the so-called intrinsec estimator (EI). The object of our interest are the contributions of age, period and cohort effects to temporal changes in the progression probabilities to the first grade of Elementary School for the brazilian women. The APC modelling of grade progression probabilities is justified as age, period and cohort effects may signficantly affect school transitions: age effects reflect both the mininal age of school entry as the trade-off between study and work, which becames strong along the educacional carrier; period effects are associated with different economical and political conjuctures, as well as with the state of the educational policies; finally, cohort effects reflect social attributes proper of some group of individuals. Both instruments were contraposed in terms of the efficiency and significance of the parameters. The results reveal the potentiality of the solution to the age-period-cohort model based on intrinsec estimator, which presents excelent statistical properties, namely: small variance and a large number of significative parameters. Therefore, projections of grade progression probabilities based on this estimator may be very promissing.

KEYWORDS: AGE-PERIOD-COHORT MODELS; INTRINSEC ESTIMATOR; GRADE PROGRESSION PROBABILITY

1 Introduction

The application of age-period-cohort models (APC) has been the subject of intense debate in demography since 1970, especially from the work (Mason et al., 1973). Shortly, these models seek to evaluate to what extent a phenomenon of interest was being determined by period, age and cohort variations in the period. Effects of age are generally associated with age differentals in the risk of observing some characteristic. Furthermore, age may reflect the evolution of biological, psychological and social change in the roles of each age group. The effects of period, on the other hand, reflect changes in the phenomenon of interest that affect neutraly all age groups. In general, period variations give us a measure of environmental, economic and social laws. Finally, cohort effects may reflect changes between a group of individuals of the same age. These effects can be represented by genetic or social change (Rodgers, 1982; Halli and Rao 1992; Yang, Fu and Land, 2004).

The main controversy in APC modelling is the choice of the estrategy to deal with the so-called *identification problem*. Since there is a linear dependence between age, period and cohort (period = age + cohort), the design matrix X^TX is singular, i.e., the inverse of X^TX does not exist. Therefore, the estimable solution to this equation is not unique and it is not possible to estimate the three separate effects without imposing identification restrictions. The usual strategy for identification is the imposition of an equality of the parameters of the model (Fienberg and Mason, 1985). However, this solution has received much criticism, which tend to strengthen that the choice of identifying restrictions are *ad hoc* or even atheoretical (Smith, 2004).

Biostatistics researchers have made a meaningful contribution to this literature by estimating functions that would be invariant to identification restrictions on the APC parameters. One of these developments was the so-called intrinsic estimator (IE), described by Yang, Fu and Land (2004). This estimator is based on estimable functions of the singular value decomposition of matrices, and provides unique solutions for age, period and cohort estimators. In addition, the intrinsic estimator requires for the model identification minimal

* Special Acknowledgments: Capes/INEP/SECAD, Ministry of Education, Grant 001/2008, “Demografia da Educação: Fluxo, Proficiência, Seletividade, Desigualdades e Políticas Públicas”

[1] PhD student at Demography Dept/CEDEPLAR/UFMG - raquel@cedeplar.ufmg.br

[2] Full professor at Demography Dept/CEDEPLAR/UFMG - eduardo@cedeplar.ufmg.br

assumptions or prior information. Moreover, the authors showed that the IE has many desirable statistical properties in the APC analysis with fixed periods of time (Yang, Fu and Land, 2004; Yang, 2008).

The aim of this paper is, therefore, a methodological comparison exercise of two APC estimators: the restricted estimator obtained by conventional generalized linear models (REGLM) and the intrinsic estimator (IE). The two instruments were compared in terms of the estimated parameters, the fit to the observed values and model efficiency. Of our interest are the contributions of the age, period and cohort effects on the temporal changes in the progression probability to the first grade of elementary school to women. The APC modeling of grade progression probabilities in the field of Demography of Education has a clear substantive interpretation: effects of age reflect the minimum age for entry into the education system as well as the dilemma between work and study that comes along the educational career, period effects are associated the different economic and political conjucture, as well as with the state of the educational policies and, finally, cohort effects reflect social characteristics peculiar to certain groups of students.

The choice in this study of studying the progression probability to the first grade of elementary school was due to the fact that the temporal changes in the chances of progression in this educational grade were responsible for most of the variation in average years of schooling of the brazilian population between 1981 and 2008, as we demonstrated in our previous work (Guimarães, 2010). Moreover, it was found that exists in Brazil a difference by sex in the behavior over time of the progression probabilities. Thus, we opted for the assessment of the women progression probabilities.

Besides of its great substantive importance, the APC analysis of grade progression probabilities (GPP) has an important applicability to the development of educational projections. Once it is shown that there is a formal relationship between the GPP and the average years of schooling (Rios-Neto, 2004), a consistent modeling of GPP then allows the construction of feasible scenarios of future changes in its age, period and cohort components. From these estimated components it is possible to obtain the average years of schooling of the population in the future. In this sense, the comparison of methodologies acquires a fundamental importance to obtain educational projections that have good quality and accuracy.

This article is organized into four sections, including this introduction. The second section presents a literature review of the APC model, particularly whose concerned with strategies and solutions to overcome the identification problem. The third section deals exclusively with intrinsic estimator described by Yang, Fu and Land (2004). The fourth section describes the data and methodological steps. The fifth section reports the procedures and results comparing the conventional restricted estimator and intrinsic estimator of the age, period and cohort effects of the temporal change in the women progression probability to the first grade of elementary school. The sixth section summarizes the evidence from this study and proposes an agenda for future research.

2 The issue of identification in age-period-cohort models: a brief review

The problem of identification within the APC framework can be described as follows. Consider a general linear model whose dependent variable is a demographic rate T_{ij}, expressed in terms of the ratio between the number of occurrences O_{ij} and the number of individuals at risk E_{ij} for each age group *i* e period *j*:

$$T_{ij} = \frac{O_{ij}}{E_{ij}} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ij} \qquad (1)$$

In this model, i=1,...,a indexes the age group, j=1,...,p indexes the period e k=a+p−1 identifies the cohort. Furthermore, εij is the disturbance term with zero expectance.

The model falls within the Generalized Linear Models class (GLM). According to Rios-Neto and Oliveira (1999), a GLM is constructed from a choice of a appropriate link function to the phenomenon of interest and of a probability distribution for the dependent variable.

In this paper, we treat grade progression probability as a dichotomous variable that has a binomial distribution. That is, for each combination of age-period in this study we have the absolute frequency of those who had progressed in a particular school transition and those who had not. From these quantities we derive the chances of progressing in relation to not progressing. The canonical connection is then performed by the logistic function, resulting in a logit model as follows:

$$e_{ij} = log(\frac{p_{ij}}{1-p_{ij}}) = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ij} \quad (2)$$

Where e_{ij} expresses the logarithm of the odds ratio of school progression in a transition to a certain age group and period and p_{ij} is the progression probability in cell i, j. According to Yang, Fu and Land (2004), model 2 can only be operationalized through the centralization of the parameters or by imposing an identification of one of the categories of the age, period and cohort covariates as the reference category.

Once is chosen the operationalizations for the model, we can rewrite the second model in terms of a general linear equation, where Y represents the logarithm of the chance of progression in each cell, X is the matrix consisting of indicator variables with dimension m = 1 + (a-1) + (p-1) + (a + p-2) and ε is the vector of random error:

$$Y = X\beta + \varepsilon \quad (3)$$

The numerical solution to obtain the sample parameters for β by maximum likelihood in the model 3 can be described by:

$$b = \left(X^T X\right)^{-1} \left(X^T X\right) \quad (4)$$

The key issue is that there is not only one possible vector of estimated coefficients for model 3. This is because X^TX is not invertible (singular matrix), due to a perfect linear relationship between the effects of age, period and cohort. This impasse is called in the literature of as identification problem. Therefore, it becomes impossible to separate the estimated effects of cohort, age and period without imposing additional restrictions on the model coefficients, ie, beyond the centralization or the adoption of reference categories.

A intense debate, then, emerged in the social sciences and epidemiology concering the best identification restriction that should be adopted in the APC model. Mason et al. (1973) and Fienberg and Mason (1985) were the first researches that proposed a solution to this problem. According to the authors, a strategy to make the matrix X^TX invertible and thus obtain a unique solution for the parameters, would be imposing an equality constraint on the coefficient vector β. Thus, it would be sufficient to admit that the coefficients of the first and second period, or first and second cohort, or the first and second age group, for example, were equal.

Recent literature that addresses the identification problem has received significant contributions from biostatistics and epidemiology. In one of the lines of this branch of study that employs estimable functions, we find the so-called intrinsic estimator. He was introduced by Fu, Knight and Fu, Fu, Hall, and Rohan (apud YANG, FU; LAND, 2004) and is based on the singular value decomposition of matrices. Also, this method provides robust estimators of the effects of age, period and cohort. Since this method is of central interest in this article, the formalization of this method will be presented in the next section.

3 The Intrinsec Estimator (IE)

In this section we formalize the construction of the intrinsic estimator described by Yang, Fu and Land (2004), Yang (2008), Yang et al. (2008), as well as their statistical properties. Consider the APC general linear model (Equation 3). There the linear dependence between the effects of age, period and cohort can be represented in a matrix form as follows, from a non-null vector B_0:

$$XB_0 = 0 \quad (5)$$

Equation 5 results of the fact that the matrix X is singular, ie, there is some linear combination of columns of the design matrix X that results in a zero vector. In terms of Linear Algebra, it is said that the matrix X does not have a full rank.

Kupper et al. (apud YANG, FU; LAND, 2004) showed that, if a matrix has a less-than- a- full rank, its parametric space can be decomposed into the direct sum of two linear subspaces that are mutually perpendicular:

$$P = N \oplus T \qquad (6)$$

Where $\oplus$ represents the direct sum of two linear subspaces N and T, which are perpendicular to each other. N is the null space with a dimension X measured by the vector sB_0 with a real number s and τ is the complementary orthogonal subspace N. Due to this orthogonal decomposition of parametric space, each of the infinite solutions of the unrestricted APC model can be written as:

$$\hat{b} = B + SB_0 \qquad (7)$$

Where S is a scalar corresponding to a specific solution to the problem of identifying and B_0 is an eigenvector of norm 1 or Euclidean length 1. Yang, Fu and Land (2004) argue that this eigenvector $B_0$is independent of observed rates Y, and hence is completely determined by the number of age groups and time periods, ie, B_0 has a specific form that is solely a function of the design matrix X.

Kupper et al. (Apud YANG, FU; LAND, 2004) showed that B_0 has the following form: $B_0 = \frac{\tilde{B}_0}{\|\tilde{B}_0\|}$ (8)

The direct implication of equation 8 is that B_0 is the normalized vector of $\tilde{B}_0$ which corresponds to:

$$\tilde{B}_0 = (0, I, P, C)^T \qquad (9)$$

Where:

$$I = (1 - \frac{i+1}{2}, \cdots, [a-1] - \frac{i+1}{2}),\; P = (\frac{p+1}{2} - 1, \cdots, \frac{p+1}{2} - [p-1]),$$

$$C = (1 - \frac{i+p}{2}, \cdots, [a+p-2] - \frac{i+p}{2})$$

and *i*, *p* e *c* denote, respectively, age groups, periods and cohorts. Yang, Fu e Land (2004) underscore the main issue of equation 9 is that the vector B_0 is fixed, ie, it is independent of the response variable Y and, therefore, has no role in determining the model coefficients. However, when any restriction is imposed on the coefficients vector, as it is proposed by Fienberg and Mason (1985), then this principle is violated in the sense that *s* in the equation 6 assumes a nonzero value.

Therefore, it can be shown that any conventional APC model can be described by a constraint on the identification of design matrix X, according to equation 6, and B is called the intrinsic estimator. This estimator is orthogonal to the null space and is determined by the generalized inverse of Moore-Penrose.

Fu, Hall e Rohan (2004) and Yang, Fu and Land (2004) described mathematically some statistical properties of the intrinsic estimator. Following the authors, the first statistical advantage of IE is that it satisfies the condition for estimating linear functions of the parameter vector *b*. This is a positive result of the approaches based on estimable functions: they are invariant with respect to which the solution to normal equations are obtained. Moreover, these functions should be desirable as statistical estimators since they are linear functions of the unidentified parameter vector, and the latter that can be estimated without bias - in other words, the IE provides unbiased estimators of age, period and cohort effects of. It is worth to mention that this condition, properly formalized by Kupper et al. (apud YANG, FU; LAND, 2004) implies that any restricted estimator, ie, that which is obtained by imposing equality constraints on the vector of parameters, will always produce biased estimates of age, period and cohort effects. In summary, the first statistical property of IE is that it produces non-biased estimates of coefficients in the APC framework for the analysis of population rates considering fixed finite time periods *p*. The asymptotic property of the intrinsic estimator suggests that, as the number of periods increases, the arbitrariness of the many possible estimators for REGLM is removed and these estimators converge to the intrinsic estimator B (Yang, Fu and Land 2004).

Secondly, it was shown that IE is more efficient than the concentional restricted estimator - ie, has lower variance. other words, for any finite number p of time periods, the intrinsic estimator B has a smaller variance than any other restricted estimator MLGR.Thus, it is possible to derive that $var\left(\hat{b}\right)-var\left(B\right)$ is a defined and positive function for any non-trivial constraint identification (Yang, Fu and Land 2004). Another important property of the intrinsic estimator is that it is asymptotically consistent, ie, when $p \rightarrow \infty$ it converges to the true parameters that generated the sequence of APC rates.

One limitation, however, still persists in the IE solution for the APC models. The intrinsic estimator, although resulting in different coefficients of GLM framework - ie, free of any bias - will present the same set of quality measuresof the latter, such as log-likelihood and deviance. Therefore, these measures should not be used to select the best APC model (Yang et al., 2008).

4 Data and Methods

For this exercise we use brazilian microdata from PNAD-IBGE in the period between 1981 and 2008. Although the PNAD does not have a panel survey design, ie, which tracks individuals over time, their analysis in the context of the APC model is plausible. This is true as we should follow each cohort over repeated surveys, looking at the cohort members who were randomly selected in each year of the survey (Oliveira, 2002). Thus, for example, individuals who were 7 years old in 1981 would have eight years in 1982, and so on.

The individuals of our interest were that aged 7 to 29 years, since it is required for the study of school transition that the individuals were exposed to the risk of completing the educational grade. Thus, we have 20 unit intervals of age × 28 periods. As the PNAD series has three discontinuities (1991, 1994 and 2000), we used a linear interpolation of the number of individuals promoted and at risk and in adjacent years in order to complete the series.

For estimation of the model, we employed a logistic function as the canonical link. The choice of this functional form of the model aimed to ensure that the predited odds of grade progression were restricted to the interval (0.1). The estimation method used was maximum likelihood.

With respect to the model specification, we estimate the age-period-cohort model in its complete form, in order to check the differences between IE and REGLM methodologies and evaluate the substantive importance of the effects of age, period and cohort for the grade progresion probability. We kept in mind, however, that the ideal procedure would be to test the importance of each variable, starting from a null model, and then adding one by one of the variables of age, period and cohort and assess their significance, using the deviance statistic and R-squared. Another important procedure in the APC analysis that was not implemented in this paper was the test for interactive effects between age, period and cohort, or even the inclusion of quadratic terms quadratic terms. However, because this is an comparative-exercise article, we were interested not primarily in reaching a perfect fit to the data, but rather to verify the potential of each framework for the estimation of a APC model in its complete form.

For estimation of the intrinsic estimator model we employed the algorithm provided in STATA by Schulhofer-Wohl and Yang (2006). These authors point out that in the algorithm to compute the intrinsic estimator uses the constraint the constraint that the sum of the coefficients is zero. For computational purposes, the algorithm creates indicator variables for age, period and cohort variables in the design matrix X, but one of the categories of each of them is omitted. After the principal components regression estimation, however, the restriction that the parameters should be zero-sum allows the researcher to get estimates for the omitted categories. Note that this does not occur in the REGLM framework.

In order to estimate the generalized linear model, we employed the algorithm *glm* also available in STATA (StataCorp 2007). Our identification strategy was the imposition of the two oldest cohorts should have the same coefficients (ie, cohorts of 1952 and 1953 would have equal APC effects.) It is considered that this alternative is plausible since it can be assumed that the two oldest cohorts have not gone through a significant process of social change. We believe that the APC effects fot the two most recent periods could not be assumed as equal as several educational policies were recently adopted in Brazil. We had not had na intention here to restrict the age parameters, because we assume that the behavior of age-progression is unique and of substantive interest, because the change in odds of progression by age in a given school transition may be reflectinf the chronic pattern of age-grade-distortion in brazilian education.

5 Results

We report in this section the results of the age-period-cohort estimates for the progression probability to the first grade of elementary school of brazilian women and compare the performance of the intrinsic estimator and the restricted estimator (with the assumption that the two older cohorts were equal). The values of estimated coefficients and statistics for goodness of fit of the two models are reported in Table A.1 (Appendix). It is noteworthy that, as explained in section 3, the two models have the same measures of goodness of fit (deviance, AIC, BIC and log-likelihood), and therefore employment of these criteria in selecting the best model is infeasible. The main evidences from this table is that the Intrinsec Estimator coefficients were more efficient than the Restricted Estimator ones. 94 coefficients were estimated by IE, and 79 of them were statistically significant at a 5% level. On the other hand, of a total of 90 coefficients estimated by RE, only 13 of them were significant at this level.

Figure 1 reports the estimated coefficients of the IE and REMLG age-period-cohort models. By the graphical analysis we try to verify to what extent the parameter estimates derived from each method are discrepant. It is worth mentioning that in the estimation of the model, the IE uses the constraint that the sum of the coefficients of age, period and cohort is equal to zero. In turn, the REGLM uses the restriction to omit a reference category, where the first category of age, period and cohort. Therefore, to maintain comparability between the coefficients of the two models, the REGLM model parameters were centered around the average of the coefficients of age, period and cohort. This procedure is known as effect coding, and from there, the intercept is equal to the global average and the intercept for each variable expresses the difference between the group and overall average (Hosmer and Lemeshow, 2000).

Comparing the magnitude of the estimated coefficients for the probability progression to the first grade of elementary school of Brazilian women, we realized that the age effects showed little difference between the methods (Figure 1a), while the effects of period (Figure 1b), and cohort (Figure 1c) showed quite different magnitudes. A substantive analysis of the behavior of the coefficients of age period and cohort reveals interesting sources of variations on the observed progression probabilities. We note, first, that the probability progression to the first grade of elementary school of Brazilian women is very low for women aged 7, and then rises at an accelerated rate until the women reaches 14 years old, and since then it stabilizes (Figure 1a). This pattern by age of the grade progression probability is consistent with the high standard of holding back students in Brazil, due to the fact that individuals tend not to complete the first grade on the appropriate age (Rios-Neto et al. 2010). The period effects obtained by the intrinsic estimator indicate a increase in the probability progression to the first grade of elementary school of Brazilian women, and this behavior is consistent with the expansion of primary education that occurred in Brazil (Figure 1b). Finally, the cohort effects behave increasingly and unstable until 1994, when it reaches its greatest magnitude (Figure 1c).

We now examine the efficiency of both methods through the comparison of the magnitude of the confidence intervals. We saw in section 3 that has been shown mathematically that the IE has less variance than either estimator REGLM, ie, any estimator REGLM obtained by restriction of identification. To investigate this assertion in the study of the probability progression to the first grade of elementary school of Brazilian women, we build graphs showing the behavior of the estimated coefficients for IE and the REGLM with their respective confidence intervals at a 95% level.

It is possible to conclude by Figures 2, 3 and 4 that, in fact, the intrinsic estimator is more efficient - ie, has a smaller variance - than REGLM, both for the age, period and cohort effects. When analyzing the graphs for the effects of age, we found that the intrinsic estimator has an excellent efficiency in relation to REGLM. The variance in EI, however, increases when the coefficients for period or cohort are not significant, yet in fact this estimator has a smaller variance than the REGLM.

It is noteworthy that the use of parameters with small variance in the APC model is very important and desirable for educational projections based on GPP, so that any extrapolation of the estimates standard error based on the intrinsic estimator will tend to be quite efficient.

6 Final remarks

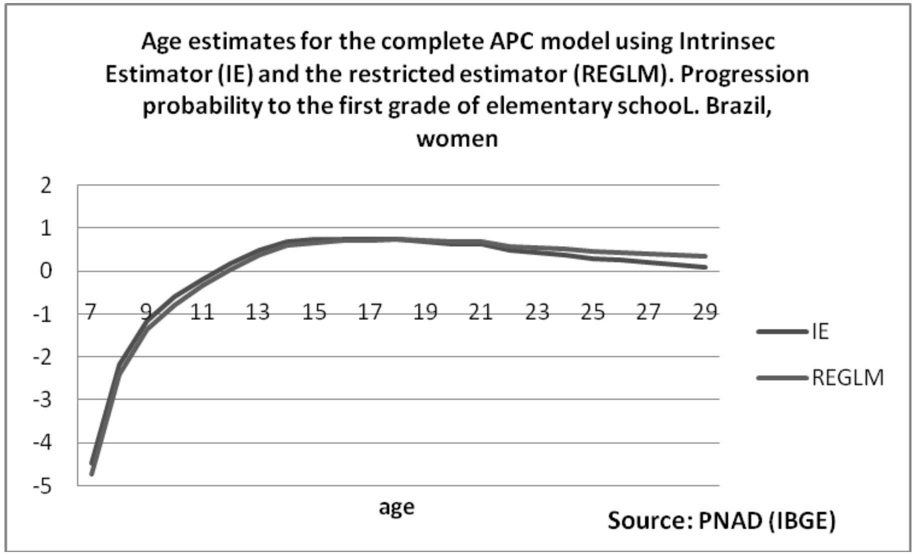
Age-period-cohort model have a strong tradition in Demography. In this article, we sought to analyze the APC research on the different sources of variation in the probability progression to the first grade of elementary school of Brazilian women.

This article has, however, a strong methodological character. We sought to compare two methods, the first being the usual demographic literature procedure, based on the work of Fienberg and Mason (1985), which we named restricted estimator by generalized linear models (REGLM), and a methodology that has emerged from recent advances in epidemiology, called the intrinsic estimator (IE). It was argued in the literature that the solution based on EI addresses a solution to a great impasse regarding the identification problem in the APC framework. The major innovation of this estimator is that, through the decomposition of parametric space of the APC unrestricted model (ie, the non identified model), this particular estimator B can be derived by both the projection method as by the principal components regression method. Moreover, another uniqueness of this estimator is that the only restriction necessary is based on the orientation of the estimator in the parametric space, which depends crucially on the design matrix X fixed, ie, the number of periods and age groups.

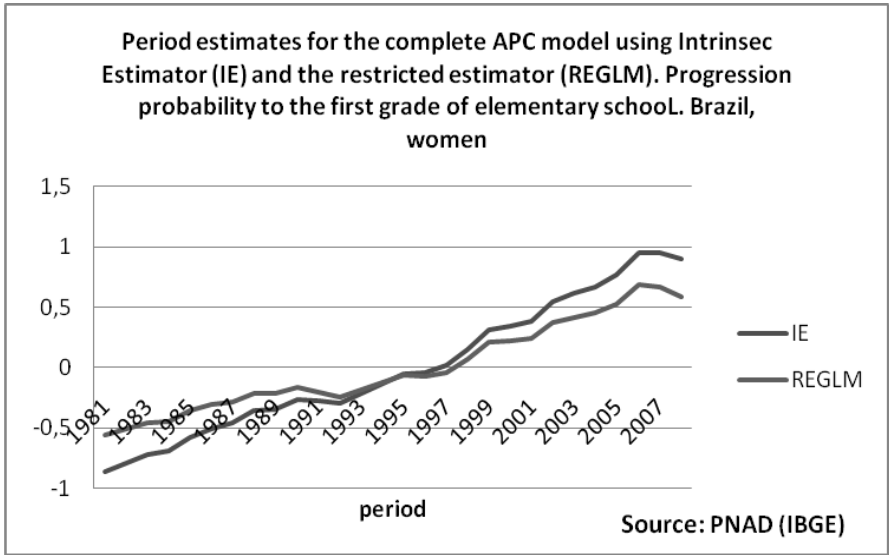
We have also seen that the intrinsic estimator presents, according to the literature, excellent statistical properties. To a large extent, our empirical evidence corroborate this evidence. In addition to presenting the estimated parameters consistent with the historical evolution of educational policies in Brazil and converge to the true values of parameters in large samples, this estimator is more efficient than the restricted estimators based on generalized linear models (REGLM). In turn, the majority of the estimators for REGLM with the assumption that the parameters of the two oldest cohorts were equal were not statistically significant and had a greater variance than the intrinsec estimators.

Given all this, we argue that the intrinsic estimator presents itself indeed as a powerful tool in the APC framework. Therefore, the construction of probabilistic projections of GPP based in those estimators proves promising. This is the next step we intend to implement in future articles. Therefore, it is necessary to proceed with a step-by-step construction of the best APC model to our educational data, testing for inclusion of each indicator variables of age, period and cohort, comparing the fit of the models and the testing for the inclusion of variables interactive or quadratic terms.

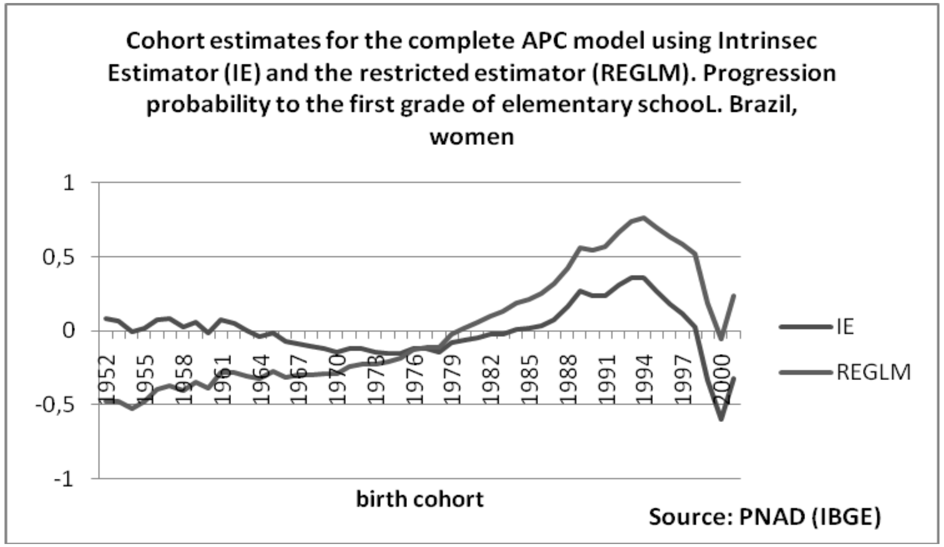
Figure 1: Comparison between the estimated values of age, period and cohort effects by the Intrinsec Estimator and Restricted Estimator. Progression probability to the first grade of elementary school. Brazilian women.



(a) Age

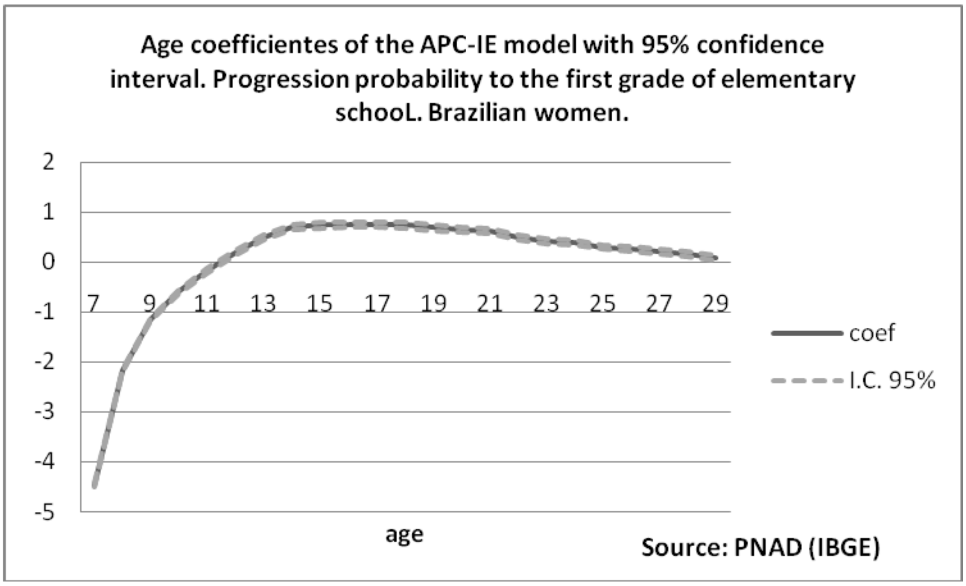


(b) Period

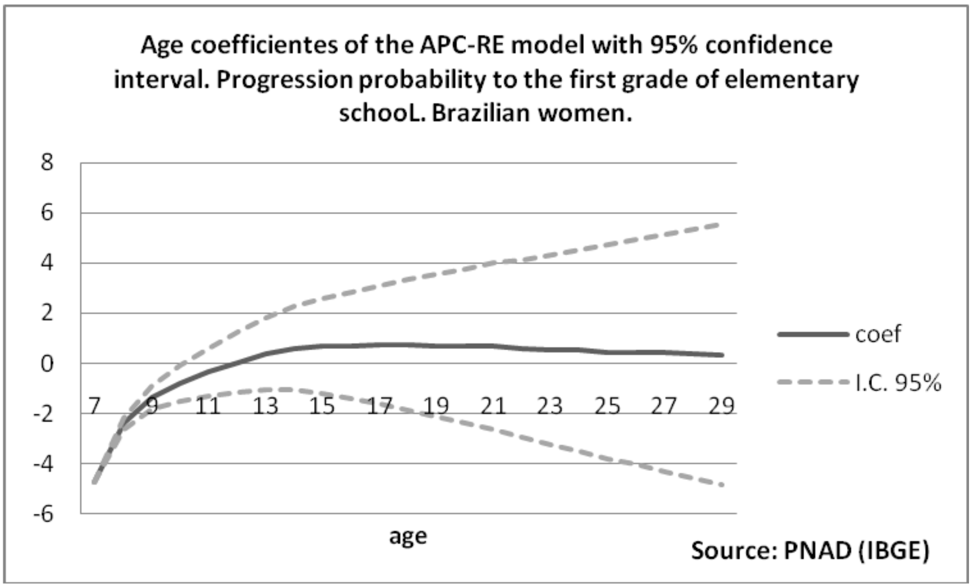


(c) Cohort

Figure 2: Comparison between the 95% confidence intervals for the age effects. Intrinsec versus Restricted estimator. Progression probability to the first grade of elementary school. Brazilian women.

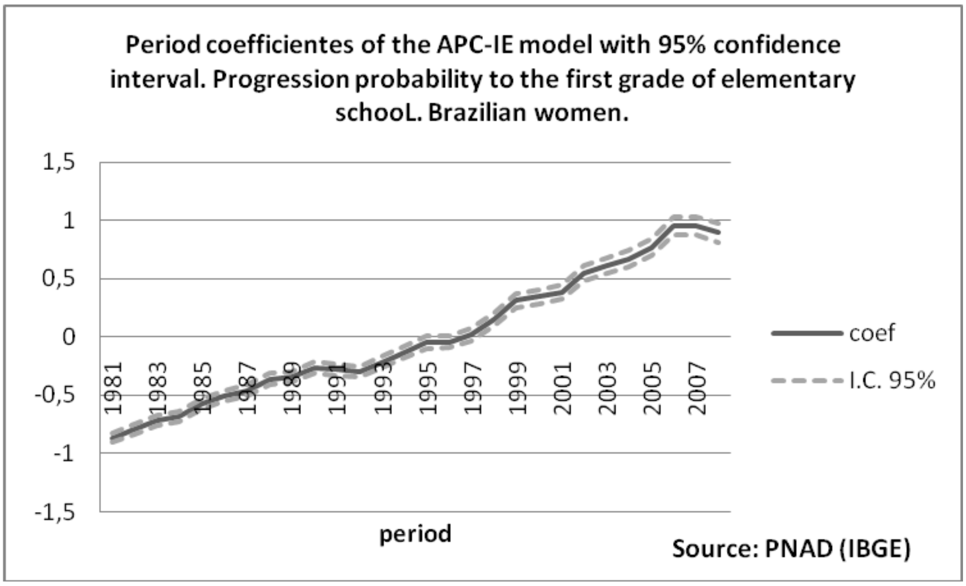


(a) Intrinsec Estimator (IE)

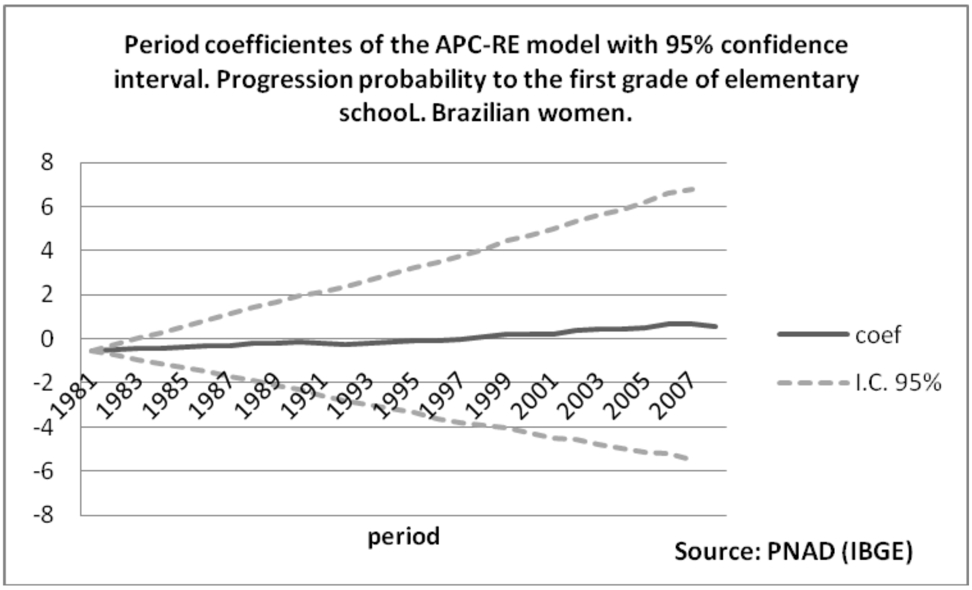


(b) Restricted estimator (REGLM)

Figure 3: Comparison between the 95% confidence intervals for the period effects. Intrinsec versus Restricted estimator. Progression probability to the first grade of elementary school. Brazilian women.

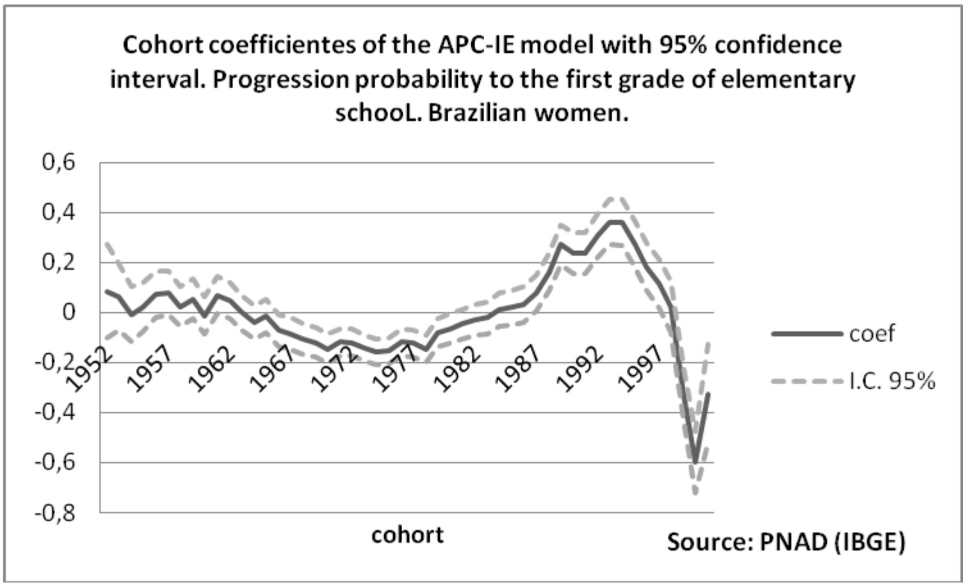


(a) Intrinsec Estimator (IE)

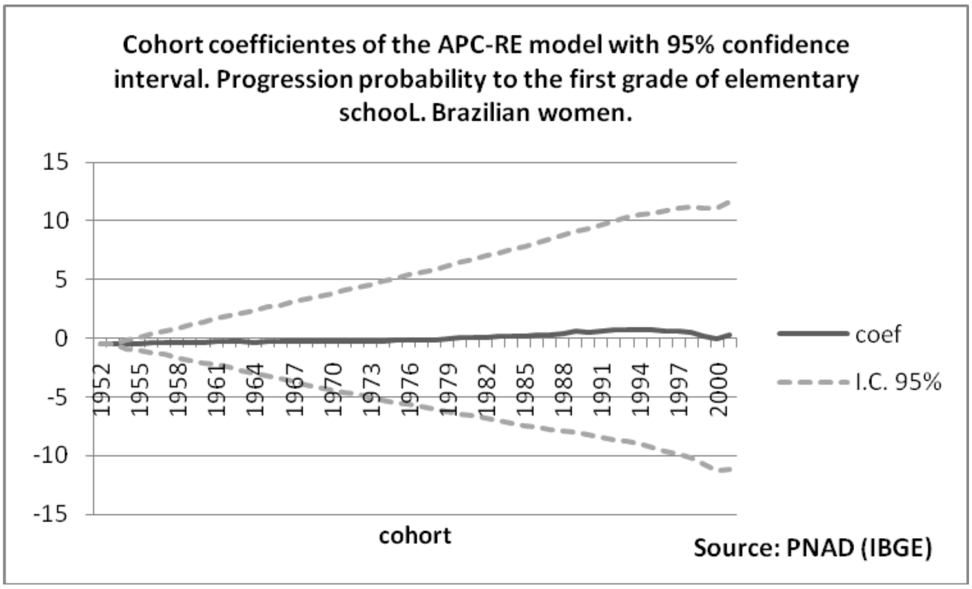


(b) Restricted estimator (REGLM)

Figure 4: Comparison between the 95% confidence intervals for the cohort effects. Intrinsec versus Restricted estimator. Progression probability to the first grade of elementary school. Brazilian women.



(a) Intrinsec Estimator (IE)



(b) Restricted estimator (REGLM)

References

FIENBERG, S. E.; MASON, W. M. Specification and implementation of age, period and cohort models. In: MASON, W. M.; FIENBERG, S. E. (Ed.). *Cohort analysis in social research*. (S.l.): Springer Verlag, 1985.

FU, W. J. Ridge estimator in singular design with application to age-period-cohort analysis of disease rates. *Communications in Statistics: Theory and Method*, v. 29, p. 263–278, 2000.

FU, W. J.; HALL, P.; ROHAN, T. E. Age-period-cohort analysis: Structure of estimators, estimability, sensitivity, and asymptotics. *Journal of the American Statistical Association*, 2004.

GUIMARÃES, R. R. de M. *Probabilidade de Progressão por Série no Brasil: evolução, seletividade e aplicação de modelos de idade-período-coorte*. Dissertação (Mestrado) — Centro de Desenvolvimento e Planejamento Regional, 2010.

HALLI, S. S.; RAO, K. V. *Advanced techniques of population analysis*. (S.l.): Plenum Press, 1992.

HECKMAN, J.; ROBB, R. Using longitudinal data to estimate age, period and cohort effects in earnings equation. In: MASON, W. M.; FIENBERG, S. E. (Ed.). *Cohort analysis in social research*. (S.l.): Springer Verlag, 1985.

HOSMER, D.; LEMESHOW, S. *Applied logistic regression*. (S.l.): Wiley, 2000.

KNIGHT, K.; FU, W. J. Asymptotics for lasso-type estimations. *Annals of Statistics*, v. 28, p. 1356–78, 2000.

KUPPER, L. L. et al. Statistical age-period-cohort analysis: A review and critique. *Journal of Chronic Diseases*, v. 38, p. 811–830, 1985.

MASON, K. O. et al. Some methodological issues in cohort analysis of archive data. *American Sociological Review*, v. 38, p. 242–258, 1973.

MASON, W. M.; SMITH, H. L. Age-period-cohort analysis and the study of deaths from pulmonary turbeculosis. In: MASON, W. M.; FIENBERG, S. E. (Ed.). *Cohort analysis in social research*. (S.l.): Springer Verlag, 1985.

OLIVEIRA, A. M. H. C. de. *Acumulando Informações e Estudando Mudanças ao Longo do Tempo: Análises Longitudinais do Mercado de Trabalho Brasileiro*. Tese (Doutorado) — Centro de Desenvolvimento e Planejamento Regional, 2002.

OSMOND, C.; GARDNER, M. J. Age, period and cohort models applied to cancer mortality. *Statistical Medicine*, v. 1, p. 245–259, 1982.

PEARL, J. *Causality: Models, Reasoning and Inference*. (S.l.): Cambridge University Press, 2000.

RIOS-NETO, E. L. G. O método de probabilidade de progressão por série. In: RIOS-NETO, E. L. G.; RIANI, J. de L. R. (Ed.). *Introdução à Demografia da Educação*. Campinas: Associação Brasileira de Estudos Populacionais, 2004. cap. 1.

RIOS-NETO, E. L. G.; OLIVEIRA, A. M. H. C. de. Aplicação de um modelo de idade-período-coorte para a atividade econômica no brasil metropolitano. *Pesquisa e Planejamento Econômico*, v. 29, n. 2, p. 243–272, ago 1999.

RIOS-NETO, E. L. G.; GUIMARÃES, R. R. de M.; PIMENTA, P. S. F.; MORAIS; T. A. *Análise de Indicadores Educacionais no Brasil, 1981 a 2008*. Junho de 2010. Centro de Desenvolvimento e Planejamento Regional. Texto para Discussão n. 386.

ROBERTSON, C.; BOYLE, P. Age, period and cohort models: The use of individual records. *Statistical Medicine*, v. 5, p. 527–538, 1986.

ROBERTSON, C.; GANDINI, S.; BOYLE, P. Age-period-cohort models: A comparative study of available methodologies. *Journal of Clinical Epidemiology*, v. 52, n. 6, p. 569–583, 1999.

RODGERS, W. L. Estimable functions of age, period, and cohort effects. *American Sociological Review*, v. 47, n. 6, p. 774–787, dez 1982.

SCHULHOFER-WOHL, S.; YANG, Y. *APC: Stata module for estimating age-period-cohort effects*. ago. 2006. Statistical Software Components, Boston College Department of Economics. Disponível em: <http://ideas.repec.org/c/boc/bocode/s456754.html>.

SMITH, H. Response: cohort analysis redux. *Sociological Methodology,* v. 34, p. 111–119, 2004.

STATACORP. *Stata Statistical Software: Release 10*. 2007.

WINSHIP, C.; HARDING, D. J. A mechanism-based approach to the identification of age-period-cohort models. *Sociological Methods Research*, v. 36, p. 362–401, 2008.

YANG, Y. Trends in u.s. adult chronic disease mortality, 1960-1999: age, period and cohort variations. *Demography*, v. 45, n. 2, p. 387–416, Maio 2008.

YANG, Y.; FU, W. J.; LAND, K. C. A methodological comparison of age-period-cohort models: the intrinsic estimator and conventional generalized linear models. *Sociological Methodology*, v. 35, p. 75–110, 2004.

YANG, Y. et al. The Intrinsic Estimator for age-period-cohort analysis: What it is an how to use it. *American Journal of Sociology*, v. 113, n. 6, p. 1697–1736, maio 2008.

Appendix

Table A.1: Age-period-cohort estimates using Intrinsec Estimator, the Restricted Estimator by Generalized Linear Models and Ajusted Restricted estimator for the effect coding. Progression probability to the first grade of elementary school. Brazilian women.

	IE	REGLM	Adjusted REGLM*
age_7	-	-	0,000
			-
age_8	-	-	0,000
			-
age_9	-	-	0,000
			-
age_10	-	-	0,000
			-
age_11	-3,188	-	0,000
	[0,029]		-
age_12	-1,217	2,006	0,000
	[0,021]	[0,150]	[0,150]
age_13	-0,546	2,713	0,000
	[0,023]	[0,293]	[0,293]
age_14	-0,179	3,115	0,000
	[0,025]	[0,438]	[0,438]
age_15	0,045	3,375	0,000
	[0,027]	[0,583]	[0,583]
age_16	0,270	3,635	0,000
	[0,028]	[0,728]	[0,728]
age_17	0,392	3,792	0,000
	[0,029]	[0,873]	[0,873]
age_18	0,454	3,890	0,000
	[0,030]	[1,018]	[1,018]
age_19	0,508	3,980	0,000
	[0,030]	[1,163]	[1,163]
age_20	0,473	3,980	0,000
	[0,030]	[1,308]	[1,308]
age_21	0,470	4,013	0,000
	[0,029]	[1,453]	[1,453]
age_22	0,428	4,006	0,000
	[0,029]	[1,598]	[1,598]
age_23	0,395	4,008	0,000
	[0,028]	[1,744]	[1,744]
age_24	0,364	4,013	0,000
	[0,028]	[1,889]	[1,889]

age_25	0,323	4,007	0,000
	[0,027]	[2,034]	[2,034]
age_26	0,251	3,971	0,000
	[0,027]	[2,179]	[2,179]
age_27	0,224	3,979	0,000
	[0,027]	[2,325]	[2,325]
age_28	0,289	4,081	0,000
	[0,027]	[2,472]	[2,472]
age_29	0,245	4,072	0,000
	[0,027]	[2,612]	[2,612]
period_1981	-0,656	-	0,000
	[0,028]		-
period_1982	-0,620	0,000	0,000
	[0,028]	[0,145]	[0,145]
period_1983	-0,589	-0,005	0,000
	[0,028]	[0,290]	[0,290]
period_1984	-0,573	-0,023	0,000
	[0,028]	[0,435]	[0,435]
period_1985	-0,533	-0,019	0,000
	[0,029]	[0,580]	[0,580]
period_1986	-0,502	-0,023	0,000
	[0,029]	[0,725]	[0,725]
period_1987	-0,442	0,001	0,000
	[0,030]	[0,870]	[0,870]
period_1988	-0,416	-0,009	0,000
	[0,030]	[1,015]	[1,015]
period_1989	-0,378	-0,006	0,000
	[0,030]	[1,160]	[1,160]
period_1990	-0,343	-0,007	0,000
	[0,031]	[1,305]	[1,305]
period_1991	-0,306	-0,005	0,000
	[0,031]	[1,450]	[1,450]
period_1992	-0,271	-0,006	0,000
	[0,031]	[1,595]	[1,595]
period_1993	-0,250	-0,020	0,000
	[0,032]	[1,741]	[1,741]
period_1994	-0,212	-0,018	0,000
	[0,032]	[1,886]	[1,886]
period_1995	-0,176	-0,017	0,000
	[0,033]	[2,031]	[2,031]
period_1996	-0,078	0,045	0,000
	[0,034]	[2,176]	[2,176]
period_1997	-0,070	0,018	0,000

	[0,034]	[2,322]	[2,322]
period_1998	0,049	0,101	0,000
	[0,036]	[2,467]	[2,467]
period_1999	0,106	0,123	0,000
	[0,036]	[2,612]	[2,612]
period_2000	0,206	0,187	0,000
	[0,038]	[2,757]	[2,757]
period_2001	0,317	0,263	0,000
	[0,039]	[2,903]	[2,903]
period_2002	0,407	0,317	0,000
	[0,040]	[3,048]	[3,048]
period_2003	0,563	0,438	0,000
	[0,042]	[3,193]	[3,193]
period_2004	0,665	0,504	0,000
	[0,043]	[3,338]	[3,338]
period_2005	0,795	0,599	0,000
	[0,045]	[3,484]	[3,484]
period_2006	0,930	0,698	0,000
	[0,047]	[3,629]	[3,629]
period_2007	1,169	0,902	0,000
	[0,051]	[3,774]	[3,774]
period_2008	1,209	0,906	0,000
	[0,056]	[3,920]	[3,920]
cohort_1952	-0,175	-	0,000
	[0,114]		-
cohort_1953	-0,210	-	0,000
	[0,081]		-
cohort_1954	-0,298	-0,052	0,000
	[0,066]	[0,215]	[0,215]
cohort_1955	-0,350	-0,069	0,000
	[0,057]	[0,349]	[0,349]
cohort_1956	-0,340	-0,023	0,000
	[0,051]	[0,490]	[0,490]
cohort_1957	-0,317	0,035	0,000
	[0,048]	[0,633]	[0,633]
cohort_1958	-0,262	0,126	0,000
	[0,045]	[0,777]	[0,777]
cohort_1959	-0,210	0,213	0,000
	[0,043]	[0,922]	[0,922]
cohort_1960	-0,151	0,308	0,000
	[0,041]	[1,067]	[1,067]
cohort_1961	-0,103	0,392	0,000
	[0,040]	[1,211]	[1,211]

cohort_1962	-0,073	0,457	0,000
	[0,039]	[1,356]	[1,356]
cohort_1963	-0,046	0,519	0,000
	[0,038]	[1,501]	[1,501]
cohort_1964	0,026	0,626	0,000
	[0,038]	[1,646]	[1,646]
cohort_1965	0,032	0,668	0,000
	[0,036]	[1,791]	[1,791]
cohort_1966	0,035	0,706	0,000
	[0,035]	[1,937]	[1,937]
cohort_1967	0,026	0,733	0,000
	[0,034]	[2,082]	[2,082]
cohort_1968	0,048	0,791	0,000
	[0,033]	[2,227]	[2,227]
cohort_1969	0,045	0,823	0,000
	[0,032]	[2,372]	[2,372]
cohort_1970	0,063	0,877	0,000
	[0,032]	[2,517]	[2,517]
cohort_1971	0,082	0,931	0,000
	[0,033]	[2,662]	[2,662]
cohort_1972	0,089	0,974	0,000
	[0,034]	[2,807]	[2,807]
cohort_1973	0,078	0,998	0,000
	[0,034]	[2,953]	[2,953]
cohort_1974	0,086	1,042	0,000
	[0,035]	[3,098]	[3,098]
cohort_1975	0,068	1,059	0,000
	[0,035]	[3,243]	[3,243]
cohort_1976	0,102	1,129	0,000
	[0,036]	[3,388]	[3,388]
cohort_1977	0,132	1,194	0,000
	[0,037]	[3,534]	[3,534]
cohort_1978	0,137	1,235	0,000
	[0,038]	[3,679]	[3,679]
cohort_1979	0,173	1,306	0,000
	[0,039]	[3,824]	[3,824]
cohort_1980	0,229	1,398	0,000
	[0,040]	[3,969]	[3,969]
cohort_1981	0,299	1,503	0,000
	[0,042]	[4,115]	[4,115]
cohort_1982	0,307	1,547	0,000
	[0,044]	[4,260]	[4,260]
cohort_1983	0,336	1,611	0,000

	[0,046]	[4,405]	[4,405]
cohort_1984	0,367	1,678	0,000
	[0,048]	[4,550]	[4,550]
cohort_1985	0,353	1,699	0,000
	[0,050]	[4,696]	[4,696]
cohort_1986	0,329	1,711	0,000
	[0,052]	[4,841]	[4,841]
cohort_1987	0,320	1,737	0,000
	[0,053]	[4,986]	[4,986]
cohort_1988	0,335	1,788	0,000
	[0,056]	[5,131]	[5,131]
cohort_1989	0,274	1,763	0,000
	[0,057]	[5,277]	[5,277]
cohort_1990	0,170	1,693	0,000
	[0,059]	[5,422]	[5,422]
cohort_1991	0,129	1,688	0,000
	[0,061]	[5,567]	[5,567]
cohort_1992	-0,001	1,594	0,000
	[0,063]	[5,712]	[5,712]
cohort_1993	-0,030	1,600	0,000
	[0,066]	[5,858]	[5,858]
cohort_1994	-0,229	1,437	0,000
	[0,070]	[6,003]	[6,003]
cohort_1995	-0,477	1,224	0,000
	[0,077]	[6,148]	[6,148]
cohort_1996	-0,711	1,026	0,000
	[0,091]	[6,294]	[6,294]
cohort_1997	-0,687	1,086	0,000
	[0,143]	[6,440]	[6,440]
Constant	1,383	-3,274	-
	[0,009]	[2,520]	-
Number of observations	532	532	532
L	-162,66	-162,66	-162,66
AIC	0,95	0,95	0,95
BIC	-2773,04	-2773,04	-2773,04
Deviance	1,24	1,24	1,24

Source: Microdata from PNAD 1981-2008

Obs.: Standard erros between brackets.

*Coefficents adjusted for effect coding. Global mean of the REGLM coefficients: age = 4.7198; period = 0.55515; cohort = 0.47896