ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII^th SISMEC Congress - Vol. 20 Suppl. 1

A Comparison of Methods to Fit Mortality Curves

Stefano Renzetti[1], *Matteo Charles Malvezzi*[1]

(1) *Department of Medicine and Surgery, Università degli Studi di Parma, Parma, Italy*

CORRESPONDING AUTHOR: *Renzetti Stefano, stefano.renzetti@unipr.it*

INTRODUCTION

In the field of time series, in particular in the study of mortality or other disease event curves, one of the most broadly used methods is the joinpoint (JP) described by Kim et al. [1]. Despite the present methodology was specifically designed to fit mortality or incidence data, it has some practical limitations such as high computational time, its usage is constrained to the Joinpoint statistical software, meaning that there is less flexibility in adapting the model to different types of generalized linear model (GLM) regressions or repeated measure outcomes.

OBJECTIVES

In this study, we aimed to compare the JP method, the segmented model proposed by Muggeo [2], and the Optimal Knots for Linear Spline (OKLS) regression, a new method introduced in this work, highlighting the strengths and weaknesses of each approach across various scenarios.

METHODS

Given the general formula that represents the regression model of the dependent variable y on a regressor x with k knots:

$$y_i = \beta_0 + \beta_1 x_i + \delta_1 (x_i - \tau_1)^+ + \cdots + \delta_k (x_i - \tau_k)^+ + \varepsilon_i^{(k)}$$

where the τ_j's are the unknown joinpoints, δ_k are the difference in slopes between consecutive segments and $(x_i - \tau_j)^+ = x_i - \tau_j$ for $x_i - \tau_j > 0$ and 0 otherwise. In JP regression a grid search over the $\tau_1,\ldots,\tau_k$ is applied where at each step of the grid search the least squares estimates for the other parameters are found by linear model methods, and the set that minimizes the residual sum of squares is chosen. A permutation test is then performed to assess the significance of the new knots included in the model sequentially.

The segmented regression proposes to define the term $(x_i - \tau_j)^+ = (x_i - \tau_j^{(0)})^+ + (\tau_j - \tau_j^{(0)})(-1)I(x_i > \tau_j^{(0)})$ following the first-order Taylor's expansion around the knot $\tau_j^{(0)}$. In the case of a single breakpoint, we can iterate the optimization of the finding of the optimal knot redefining the formula at each step s as follows:

$$y_i = \beta_0 + \beta_1 x_i + \delta (x_i - \tau^{(s)})^+ + \gamma(-1) \cdot I(x_i > \tau^{(s)}) + \varepsilon_i^{(k)}$$

where $\gamma = \delta(\tau - \tau^{(s)})$. After having fitted the model, the breakpoint can be updated as $\tau^{(s+1)} = \frac{\hat{\gamma}}{\hat{\delta}} + \tau^{(s)}$ until convergence.

In OKLS regression we propose to fit a linear spline starting from a high number of knots (we suggest k=$\sqrt{n}$) and then iteratively remove the knots where a significant change in the slope is found after having defined the place of the breakpoints that maximize the likelihood of the model. To be more parsimonious and to avoid overfitting, a Bonferroni adjustment is applied when testing for the change in slope dividing the level of significance α=0.05 by the number of knots. The algorithm stops when all changes in slopes are statistically significant, if present.

Eight different scenarios were simulated to compare the method performances. Four pairs of different numbers of observations and knots were considered and, for each pair, a dependent variable imposing a pseudo R2 of 0.3 and 0.7 was generated: 15 observations and $\overline{0}$ knots (scenarios 1 and 2), 10 observations and 1 knot (scenarios 3 and 4), 25 observations and 3 knots (scenarios 5 and 6) and 50 observations and 5 knots (scenarios 7 and 8).

RESULTS

When the number of knots was not prespecified, the three methods showed similar performances in scenarios 1, 2 and 4 while the OKLS model showed to be more accurate in the other scenarios (Table 1). When fixing the correct number of knots, based on the Root Mean Squared Error (RMSE), JP showed the best performances in estimating the regression parameters and the knots 16.7% and 0% of the times, respectively, segmented 20.8% and 0% while OKLS 62.5% and 100% of the

DOI: 10.54103/2282-0930/29483

times (Table 1). JP was the most efficient method in scenarios 1,2,3 and 4 where the number of observations and knots was smaller showing the lowest computational times. Segmented and OKLS were faster at increasing number of observations and knots (scenarios 5,6,7 and 8).

CONCLUSIONS

This study allowed us to compare the performance of three methodologies for fitting mortality curves. The method we proposed demonstrated strong performance in estimating regression parameters as well as in identifying the number and placement of knots, outperforming both the JP method, considered the gold standard in this field, and the segmented model, a commonly used approach for fitting piecewise regressions. Future research should focus on evaluating the predictive performance of these methodologies.

REFERENCES

1. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. Stat Med. 2000 Feb 15;19(3):335-51. doi: 10.1002/(sici)1097-0258(20000215)19:3<335::aid-sim336>3.0.co;2-z.
2. Muggeo VM. Estimating regression models with unknown break-points. Stat Med. 2003 Oct 15;22(19):3055-71. doi: 10.1002/sim.1545.

DOI: 10.54103/2282-0930/29483