International Journal of Applied Mathematics and Statistics,
Int. J. Appl. Math. Stat.; Vol. 33; Issue No. 3; Year 2013,
ISSN 0973-1377 (Print), ISSN 0973-7545 (Online)

Statistical Modeling for Mortality Data Using Local Generalized Poisson Regression Model

Erni Tri Astuti[1], **I Nyoman Budiantara**[2], **Sony Sunaryo**[3] and **M.Dokhi**[4]

[1,2,3]Department of Statistics, Faculty of Mathematics and Science,
Institut Teknologi Sepuluh Nopember, Surabaya (Indonesia);
Email : erni@mhs.statistika.its.ac.id

[1,4]Department of Applied Statistics,
Sekolah Tinggi Ilmu Statistik, Jakarta (Indonesia)
Email : dokhi@stis.ac.id

ABSTRACT

Mortality data by age are known have a non linier pattern likes a bath-tub shape. There are some parametric regression models developed to reveals the relationship between age and mortality rates. However using such models needs more efforts including a lots of parameter needed in the models and also numerical instabilities. To overcome this difficulties, recently, researchers give much attention to nonparametric regression models. Instead of assuming some restricted regression function, this approach allows for more flexible and robust of smooth function. There are several nonparametric model have been studied intensively, including Kernel and Smoothing Spline Regression. In this paper we developed a new approach by using local polynomial modeling. This model is a likelihood based model assuming Generalized Poisson distribution for the number of deaths at specific age. Using Generalized Poisson distribution instead of Poisson distribution makes this model robust for over or under dispersion problems. We apply this model to Indonesians mortality data based on the result from Population Census 2010, and found that this model performs well

Keywords: Mortality, nonparametric regression, local polynomial, generalized Poisson distribution.

Mathematics Subject Classification: 62G08

1. INTRODUCTION

Modeling mortality rates is a fundamental issues in epidemiology and population studies generally, and for government, the insurance and pensions industry in particular. For the government the projection of mortality rates are useful for development planning, such as building schools, provides more hospitals, health care program, pension provision for employee etc. For the insurance and pensions industry future mortality rates are needed to determine the pricing and reserving of annuities.

Mortality data by ages are well known shows non linear pattern likes a bath-tub shape. There are numbers of approaches to the problem. In the earlier researcher developed some methods based on the forecasting of parameters in some parametric model. For example, Age-Period-Cohort (APC) models in Currie et al. (2004) are a well established method of smoothing mortality rates. Heligman and Pollard (1980) developed 8 parameters non linear logistic regression representing relationship between age and mortality rates. Lee and Carter (1992) introduced a simple bilinear model of mortality in which the time dependent component of mortality is reduced to a single index which is then forecast using time series methods. The model is fitted by ordinary least squares (OLS) with the observed log mortality rates as dependent variable. Brouhns et al. (2002) improved on the OLS approach by modeling the number of deaths directly by a Poisson distribution and using maximum likelihood for parameter estimation. There are others modification and extensions version of the Lee-Carter methods, some of them are Lee and Miller (2001), Booth et al. (2006), De Jong and Tickle (2006) and Hyndman and Ullah (2007). These models have been used in the past for a wide range of mortality data resulting in satisfactory representations of a variety of pattern. But, these approaches make strong assumptions about the functional form of the mortality surface. Using parametric regression for such a non linier relationships needs more efforts including a lots of parameter needed in the models. Estimation of the parameters is problematic due to over parameterization and also numerical instabilities.

To overcome the problems, recently researchers pay much attention to nonparametric regression. Instead of assuming some restricted regression function, this approach allows for more flexible and robust of smooth function. Currie et al. (2004) used a penalized generalized linear model (PGLM) with Poisson errors and show how to construct regression and penalty matrices appropriate for two dimensional modeling. They illustrated the methods with two data sets provided by the Continuous Mortality Investigation Bureau (CMIB), a central body for the collection and processing of UK insurance, and found that the model performs nicely. Shyamalkumar (2006) is also using smoothing spline Poisson regression model with penalized likelihood estimation method for the analysis of mortality data. The model was applied to estimate the mortality rates from the Female, English and Welsh Mortality data (1988-1992). In the other hands Peristera and Kostaki (2005) proposed using Kernel Regression for modeling mortality rates by age. They provided a critical presentation and an evaluation of the various kernel regression estimators such as Nadaraya-Watson estimate, Gasser-Muller estimate and kernel weighted local linear estimates for graduating age specific mortality data. For that, they applied the alternative kernel estimators to empirical mortality data sets of different national populations and different time periods. These data are provided by Human Mortality Database (HMD).

The methods presented above, can only estimate the smooth regression function itself without the capability to infer the estimator. In this paper we developed a new approach for analyzing mortality data. We used local polynomial modeling by Fan and Gijbels (1997), which extends nonparametric regression to maximum likelihood-based regression models. In this research we present an estimator for regression function namely a local maximum likelihood estimator based on generalized Poisson distribution. Using generalized Poisson locally instead of Poisson distribution is for avoiding over or under dispersion problems that often occurred in count data modeling. This model has the advantage that we can construct a confidence interval for the regression function which is hard to do by other nonparametric methods.

2. MATERIAL AND METHODS

In this chapter we discuss about the theoretical model for Generalized Poisson Regression model and the data used for implementation the model.

2.1. Generalized Poisson Regression Model

Generalized Poisson regression is a generalization of Poisson regression model. The last model is a well known statistical tools that can reveals the relationship between count response with some covariates. However, the Poisson regression model assumes that the mean and variance of the response variable is equal, whereas in practice, the data may display over dispersion or extra-Poisson variation, i.e., a situation where the variance exceeds the mean, or the opposite situation called under dispersion. Inappropriate imposition of the Poisson may underestimate the standard errors and overstate the significance of the regression parameters, and consequently, giving misleading inference about the regression parameters.

In order to overcome this problem there is generalized Poisson regression models (Famoye et al, 2004) that take into account over and under dispersion which is often appear in count data. Suppose we have pairly observational data $(Y_i, \mathbf{x_i}),\ i = 1,2,\cdots,n$ which is distributed independently with $\mathbf{x}$ is a vector of covariates. The count response $\mathrm{Y_i}$ is assuming follows the generalized Poisson distribution, with the probability density function given by:

$$f(y;\mu,\varphi) = \left(\frac{\mu}{1+\varphi\mu}\right)^{y}\frac{(1+\varphi y)^{y-1}}{y!}exp\left[-\frac{\mu(1+\varphi y)}{1+\varphi\mu}\right], y = 0,1,\cdots$$

with $E(Y) = \mu$ and $V(Y) = \mu(1+\varphi\mu)^2$. The parameter φ plays as dispersion parameter. When $\varphi = 0$, it will reduce to Poisson probability density. When $\varphi < 0$ this model is underdispersed, and when $\varphi > 0$ it will overdisperse relative to Poisson distribution respectively (Famoye,2000). In generalized Poisson regression, the mean of the response variable $E(Y|\mathbf{x}) = \mu(\mathbf{x})$ are assumed depends on some covariates through a link function or regression function

$$\mu_i(\mathbf{x}) = \exp(\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta})$$

where $\boldsymbol{\beta}$ is a parameter vector. The maximum likelihood estimates of $\boldsymbol{\beta}$ and φ is obtained by maximized the logarithmic of likelihood function:

$$l(\mathbf{x}^T\boldsymbol{\beta},\varphi,y_i) = \sum_{i=1}^{n}\left[y_i\ln\left(\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{1+\upphi\mathbf{x}_i^T\boldsymbol{\beta}}\right) + (y_i - 1)\ln(1+\upphi y_i) - \frac{\mathbf{x}_i^T\boldsymbol{\beta}(1+\upphi \mathrm{y_i})}{1+\upphi\mathbf{x}_i^T\boldsymbol{\beta}} - \ln(\mathrm{y_i}!)\right]$$

The maximum likelihood estimator for vector parameter is found to be a solution of maximum likelihood equation:

$$\frac{\partial l}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n}\frac{\left(y_i - \exp[\mathbf{x}_\mathrm{i}^\mathrm{T}\boldsymbol{\beta}]\right)}{(1+\varphi\exp[\mathbf{x}_\mathrm{i}^\mathrm{T}\boldsymbol{\beta}])^2} = \mathbf{0}$$

$$\frac{\partial l}{\partial \varphi}=\sum_{i=1}^{n}\left\{-\frac{y_i \exp[\mathbf{x}_\mathrm{i}^\mathrm{T}\boldsymbol{\beta}]}{1+\varphi\exp[\mathbf{x}_\mathrm{i}^\mathrm{T}\boldsymbol{\beta}]}+\frac{y_i(y_i-1)}{1+\varphi y_i}-\frac{\exp[\mathbf{x}_\mathrm{i}^\mathrm{T}\boldsymbol{\beta}]\left(y_i-\exp[\mathbf{x}_\mathrm{i}^\mathrm{T}\boldsymbol{\beta}]\right)}{(1+\varphi\exp[\mathbf{x}_\mathrm{i}^\mathrm{T}\boldsymbol{\beta}])^2}\right\}=0$$

Since the equations are nonlinear in parameters, the solutions can be obtained by using some iterative procedure such as Newton Raphson method..

2.2. Local Linear Estimator for Generalized Poisson Regression Model

The parametric regression approach like generalized Poisson regression model are assumed that the regression function has some prespecified functional form, As an alternative one could try to estimate regression function nonparametrically without refer to a specified form, which called nonparametric regression. The objective of using nonparametric regression model is to minimize the assumption of regression function and let the data found the form of the function itself (Hardle, 1990). In nonparametric regression context, the simplest way to obtain the estimate of regression function is scatter plot smoothing, and one of them is by applying polynomial regression locally which called local polynomial technique (Fan and Gijbels,1997).

Local polynomial technique can be illustrated simple as: for a given value of x, we applied polynomial regression locally in the neighborhood of x resulting a separate line in a window around each x value. The value of the estimated line at x is the estimate of the regression function at x . The size of the data in the local neighborhood is called bandwidth which plays as smoothing parameter in the model. The local polynomial technique has the advantages comparing to other smoothing techniques, not only that it is mathematically intuitive and simple but also the capability of making inference for the estimator of regression function by constructing a confidence interval.

Tibshirani and Hastie (1987) extend the idea of local fitting in nonparametric regression analysis to the class of generalized linear model, also known as maximum likelihood-based regression models. They called this approach as local likelihood technique. Fan et al. (1998) present a framework for assessing the bias and the variance of the local likelihood estimator as well as a bandwidth selection procedure and also applying this approach to logistic regression model and Cox Regression Model. Santos and Neves (2008) adopted this method to perform local Poisson regression model.

In the local generalized Poisson regression model, instead of considering some specified regression function, the dependence of mean response with a covariate is describe by a smooth nonparametric regression function s:

$$E(Y_i|x)=\mu_i(x)=\exp[s(x_i)]$$

Assume that the function *s* has a $(p+1)^{th}$ continuous derivative at the point x_0. For data points x_i in a neighborhood of x_0 or $x_i\in(x_0-h,x_0+h)$, with h is a bandwidth, we approximate $s(x_i)$ via a Taylor expansion by a polynomial of degree *p*:

$$s(x_i)\approx s(x_0)+s^{'}(x_0)(x_i-x_0)+\cdots+\frac{s^{(p)}(x_0)}{p!}(x_i-x_0)^p=\mathbf{x}_\mathrm{i}^\mathrm{T}\boldsymbol{\beta}$$

where $\mathbf{x}_i=(1,(x_i-x_0),\cdots,(x_i-x_0)^p)^T$, $\boldsymbol{\beta}=(\beta_0,\cdots,\beta_p)^T$ with $\boldsymbol{\beta_v}=\boldsymbol{s}^{(v)}(\boldsymbol{x_0})/\boldsymbol{v}!$ and $v=0,1,\cdots,p$.

For data points (X_i, Y_i) in a neighborhood of x_0, the contribution to the log likelihood function is weighted by some kernel function $K_h(\cdot) = K(\cdot/h)/h$. By assuming generalized Poisson distribution for response variable Y_i , these considerations yield the conditional local kernel weighted log-likelihood:

$$L_p(\boldsymbol{\beta}, \varphi, h, x_0) = \sum_{i=1}^{n} \left\{ y_i ln \left(\frac{\mu_i(x)}{1 + \varphi \mu_i(x)} \right) + (y_i - 1) \ln(1 + \varphi y_i) - (1 + \varphi \mathrm{y_i}) \frac{\mu_i(x)}{1 + \varphi \mu_i(x)} \right\} K_h(x_i - x_0)$$

where $\mu_i(x) = \exp\left[\mathbf{x}_{\mathrm{i}}^{\mathrm{T}} \boldsymbol{\beta}\right]$ and $K_h(\cdot) = K(\cdot/h)/h$ is a Kernel weight. The choice of the kernel function is not a crucial issues, because the result is almost similar for any kind of kernel function including Epachnecnikov, Gaussian or Boxcar Kernels.

The kernel-weighted maximum likelihood estimator for regression function, is the solution of (*p*+2) equation :

$$\frac{\partial \mathrm{L}}{\partial \boldsymbol{\beta}} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{\left(\mathrm{y_i} \text{- } \exp[\mathbf{x}_{\mathrm{i}}^{\mathrm{T}} \boldsymbol{\beta}]\right)}{(1 + \varphi \exp[\mathbf{x}_{\mathrm{i}}^{\mathrm{T}} \boldsymbol{\beta}])^2} \mathrm{K_h}(\mathrm{x_i} \text{-} \mathrm{x_0}) \mathbf{x}_{\mathrm{i}} = \mathbf{0}$$

$$\frac{\partial L}{\partial \varphi} = \sum_{i=1}^{n} \left\{ - \frac{y_i \exp\left[\mathbf{x}_{\mathrm{i}}^{\mathrm{T}} \boldsymbol{\beta}\right]}{1 + \varphi \exp[\mathbf{x}_{\mathrm{i}}^{\mathrm{T}} \boldsymbol{\beta}]} + \frac{y_i(y_i - 1)}{1 + \varphi y_i} - \frac{\exp\left[\mathbf{x}_{\mathrm{i}}^{\mathrm{T}} \boldsymbol{\beta}\right] \left(y_i - \exp\left[\mathbf{x}_{\mathrm{i}}^{\mathrm{T}} \boldsymbol{\beta}\right]\right)}{(1 + \varphi \exp[\mathbf{x}_{\mathrm{i}}^{\mathrm{T}} \boldsymbol{\beta}])^2} \right\} K_{\mathrm{h}}(x_i - x_0) = 0$$

The solution of the system which is called local polynomial (maximum likelihood) estimator can be solved by iterative procedure such as Newton Raphson Methods. The log-likelihood function above depends on two quantities, the smoothing parameter (*h*) and the order of polynomial (*p*). The model complexity is effectively controlled by the bandwidth *h*. As *h* increases from 0 to +∞, the model runs from the most complex model (interpolation) to the simplest model (polynomial model). Fan and Gijbels (1997) stated that a too large bandwidth under parameterizes the regression function causing a large modeling bias, while too small bandwidth over parameterizes the unknown faction and result in noisy estimates. Ideal or optimal model is lying between the two models, which can be obtained by different criteria. One of the criteria that are easy to interpret is Akaike's Information Criterion (AIC), so we seek for a model that minimizes the AIC. Since the modeling bias is controlled by the bandwidth, the choice of order of polynomial is less crucial. Unlike traditional polynomial parametric regression, local polynomial model just requires a small degree of polynomial for a complex pattern of data. In the case when p=0, the technique is called local constant and when p=1, it is known as local linear model. In this paper we will use local linear estimator for the application to mortality data.

2.3. The Data

There are commonly 3 methods of mortality data collection: population census, civil registration and households survey. Among these 3 methods, civil registration is the preferred ones, because the number and cause of deaths can be achieved directly. But, in the developing countries like Indonesia, the quality of data from civil registration is very poor and also incomplete. There is no willingness from the peoples, to report events like deaths or births to the civil registration office. United Nation (UN) reported that more than half the world population lives in countries in which civil registration provides incomplete coverage of deaths and births, including Indonesia. In the absence of civil registration data, population censuses and households surveys are used to collect mortality data. The population census is a potentially rich source of mortality data. Like civil registration, it will provide complete data on all geographic areas, but it will also provide data for any persons live in the households. In the population census there is commonly a question about recent deaths in the households. Questions on recent household deaths have several important advantages. In the absence of complete and accurate death registration data, they are the only possible source of information on the age pattern of mortality.

The mortality data used in this paper are from Indonesians Population Census 2010 organized by BPS Statistics Indonesia. This census is the latest and largest population census held in Indonesia, covering the entire Indonesian's territory of 33 provinces, 497 districts, and 7.000 sub district and about 75.000 villages. The 2010 Population Census has been designed to meet various data needs, including valuable input in monitoring the progress for achieving the Millennium Development Goals. One of the goals is to decrease the mortality rates and because of that reason for the first time this census covering the mortality data resulting direct mortality data. For each households was asked whether there are "event of deaths" in the households in the past 12 months before enumeration time (around May 2010). If there are member of the household that died in the range of time, the age and sex of the deaths also recorded. This is also known as current mortality data. The aggregation of number of deaths for the whole country, resulting number of death by age will be used in this paper as the observational data. The number of deaths at specific age is a response variable, and the age itself is a covariate. Because number of deaths is come from the aggregation by hierarchical procedure from households, villages, sub district, districts and provinces, it must be some intersubject variability in counting number of deaths. As a consequences, we suspicious that the data exhibit over dispersion problem the one hand, we considered the classical method of least squares without variables selection and on the other hand, the *stepwise* selection method of variables is used. These methods were adopted, because they are among the most used methods, and are available in almost all statistical software.

3. RESULTS

The scatter plot of total number of deaths (in a logarithmic scale) by age is given in figure 1. There are obviously two things that seen very attractive. First, is the pattern itself that is far from linier and also there are different slope in growths between some interval value of age. There are very high number of deaths at the young ages (between 0 and 1) also known as infant mortality, and decreased sharply until age of 12 or 13. Then it increased quite fast in the intervals of 13 to 20, but the growth is become more slowly increased after that. At the end, it will reach a high level in the interval above age of 70. For such kind of different slope of growth, we will need some smooth and flexible function for describing this pattern rather than some restricted function in parametric regression model. The second thing that also found attractive, is the systematic sharp jump in the ages ending with 0 or 5, especially for elderly age . This is a well known data problem among demographers and applied statistician, called “age heaping”. In case of mortality data, age heaping occurs when people do not remember exactly the time when the deceased born and or death. For that case, they are unable to calculate the deceased’s age at death exactly and tend to round it up or down into the nearest age ending with 0 or 5.

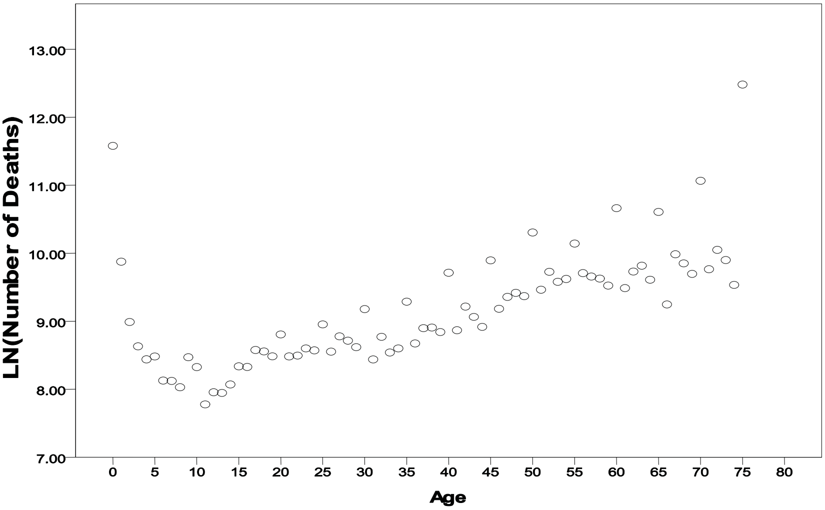

Figure 1: Plot of number of deaths by age from Indonesian Population Census 2010

Age heaping phenomenon is a kind of misreporting data, and using this data for further analysis carelessly, can leads to biased conclusion. Camarda et al. (2008), using composite link models combined with smoothing spline technique and applying this model to Portuguese mortality data. They redistributed the number of deaths at preference ages to the neighboring ages using composition matrix. The concept of local regression in this paper can work for correction of age heaping in a similar way. Because the regression is running locally, in each region where age heaping is occurred, the kernel function will redistributed this value to the neighborhood with a symmetric weight automatically.

We apply the local linier generalized Poisson model to the data, and the result with different value of bandwidth is given in Figure 2. As we can see, when the bandwidth parameter is small (h=1), the model is simply interpolate the data and as the bandwidth is increased (h=5 and h=10) we have a more smooth estimates. But when we set the bandwidth at high value (h=50), we got over smooth estimates and it is almost similar with running one linier regression model. So we have to look for best model, according to their AIC value. The AIC for different values of bandwidth are given by Figure 3, and we found the optimum ones at *h*=5 with AIC equal to -756.2546

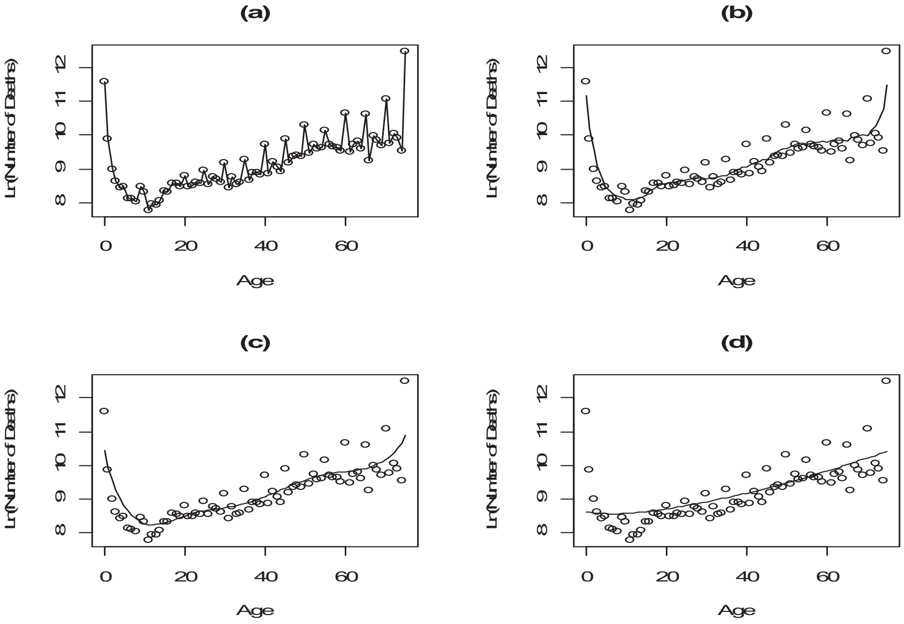

Figure 2: Local Linear Generalized Poisson Estimates for Indonesian Mortality data with different value of bandwidth (a) h=1, (b) h=5, (c) h=10 and (d)h=50

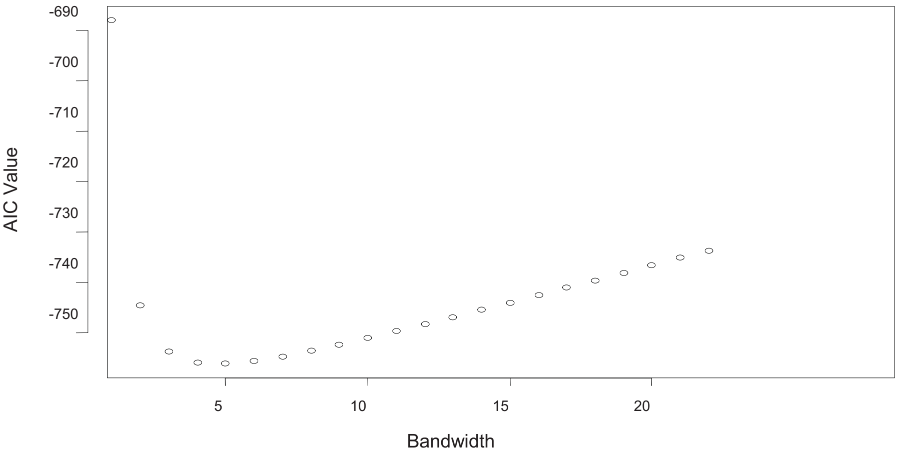

Figure 3: Plot of AIC value with different values of bandwidth

The plot of the model with optimum bandwidth is given in Figure 4. As we can see in the final model, the age heaping is vanished and redistributed to the neighborhood ages.

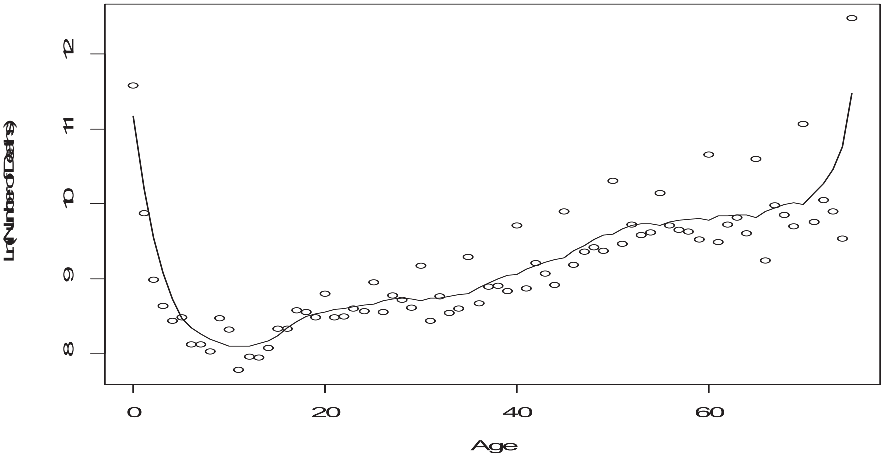

Figure 4: Local Linier Generalized Poisson estimates for Indonesians mortality data

4. DISCUSSION AND CONCLUSION

In this study, a new nonparametric regression method for count response was proposed by using local polynomial modeling. In this research we present an estimator for regression function namely a local maximum likelihood estimator based on generalized Poisson distribution. Using generalized Poisson locally instead of Poisson distribution is for avoiding over or under dispersion problems that often occurred in count data modeling. The proposed model was applied to Indonesians mortality data from Population Census 2010. The pattern of mortality by age was obviously non linear and exhibits some age heaping problem. We found that the model works nicely for smoothing pattern of number of deaths by age and also can handle the age heaping problems. As mention in the previous introduction, one of the advantage of local polynomial modeling is the capability of the construction of a confidence interval. So it is our interest in further study to derive the confidence interval for this local generalized Poisson model.

5. ACKNOWLEDGEMENT

The first author would like to thanks BPS-Statistics Indonesia for the Ph.D. grant through Human Resources Development Project in cooperation with Institut Teknologi Sepuluh Nopember (ITS), and for providing the data used in this research.

6. REFERENCES

Booth, H., Hyndman, R.J., Tickle, L. and De Jong, P., 2006, Lee-Carter Mortality Forecasting: A Multi-Country Comparison of variants and Extensions, *Demographic Research*, 15, 289-310.

Brouhns, N., Denuit, M., Vermunt, J.K., 2002 A Poisson log-bilinear regression approach to the construction of projected life tables. *Insurance: Mathematics & Economics*, 31, 373-93.

Camarda, C.G., Eilers, P.H.C. and Gampe, J., 2008, Modelling general patterns of digit preference. *Statistical Modeling*, 8(4), 385-401.

Currie, I.D., Durban, M. and Eilers,P.H.C, 2004, Smoothing and Forecasting Mortality Rates, *Statistical Modeling*, 4, 279-298.

De Jong P. and Tickle, L., 2006. Extending Lee-Carter mortality forecasting, *Mathematical Population Studies*, 13 (1), 1-18.

Famoye, F., 2000, Restricted Generalized Poisson Regression, *Communication in Statistics-Theory and Methods*, 33, 1135-1154.

Famoye, F., Wulu Jr, J.T. and Singh, K.P., 2004, On the Generalized Poisson Regression Model with an Application to Accident Data, *Journal of Data Science*, 2, 287-295.

Fan, J. and Gijbels, I., 1997, *Local Polynomial Modeling and Its Application*, Chapman and Hall, London.

Fan, J. , Farmen, M. and Gijbels, I, 1998, Local Maximum Likelihood estimation and Inference, *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 60(3), 591-608 .

Gyorfi, L., Kohler, M., Krzyzak, A. and Walk, H., 2002, *A Distribution-Free Theory of Nonparametric Regression*, Springer, New York.

Heligman, M. and Pollard, J.H., 1980, The age pattern of mortality, *Journal of the Institute of Actuaries*, 107, 49-80.

Hyndman, R.J. and Ullah, M.S., 2007, Robust Forecasting of Mortallity and Fertility Rates: A Functional Data Approach, *Computational Statistics and Data Analysis*, 51, 4942-4956.

Lee, R.D. and Carter, L.R., 1992, Modeling and forecasting U.S. mortality, *Journal of the American Statistical Association*, 87, 659-75.

Lee, R.D and Miller, T., 2001, Evaluating the performance of the Lee-Carter method for forecasting mortality, *Demography*, 38 (4), 537–549.

Peristera, P. and Kostaky, A., 2005, An Evaluation of the Performance of Kernel Estimator for Graduating Mortality Data, *Journal of Population Research*, 22, 185-197.

Santos, J.A. and Neves, M.M., 2008, A Local Maximum Likelihood Estimator for Poisson Regression, *Metrika*, 68, 257-270.

Shyamalkumar, N.D, 2006, Analysis of Mortality Data using Smoothing Spline Poisson Regression, *Working Paper*, Dept. of Stat. and Actuarial Science, The University of Iowa. *http://www.soa.org/library/research/actuarial-research-cleasing-house/2006/january/arch06v4n1-ix.pdf*. Accessed on January, 27 2011.

Tibshirani, R. and Hastie, T., 1987, Local Likelihood Estimation, *Journal of the American Statistical Association*, 82, 559-567.