Multivariate Adaptive Generalized Poisson Regression Spline (MAGPRS) on the number of acute respiratory infection infants

B W Otok\textsuperscript{1*}, S Hidayati, and Purhadi

Departement of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya

\textsuperscript{*}Corresponding author e-mail: dr.otok.bw@gmail.com

Abstract. Acute Respiratory Infection (ARI) is an infectious disease of the respiratory tract that affects the structure of the respiratory tract. The ARI is a health problem that should not be ignored because it causes high infant mortality. Therefore, it is important to know the factors that influence the ARI. Generalized Poisson Regression (GPR) is one of the methods that can handle cases of overdispersion (variance is greater than the mean) or underdispersion (variance is less than the mean). Multivariate Adaptive Regression Spline (MARS) as a statistical method for fitting the relationship between a set of input variables and dependent variables. This research is the development of the MARS method and GPR namely MAGPRS. The application of the MAGPRS model was carried out in the case of the number of ARI infants from Surabaya health department 2017. The results showed that the importance of predictor variables in MAGPRS, the variables affecting the number of ARI patients in infants are the percentage of low birth weight (X\textsubscript{2}), the percentage of unhealthy houses (X\textsubscript{5}), and the percentage given non-exclusive breastfeeding to infants (X\textsubscript{1}).

1. Introduction
Acute Respiratory Infection (ARI) is a disturbing infection respiratory processes caused by viruses or bacteria that attacks nose, respiratory tube, sinuses, and pharynx larynx. Channel infections acute breathing is recorded as the most common disease the community, especially the babies and children who usually experience the symptoms of a cold cough are at least three to six times a year. ARI is classified as an airborne disease, contact with people who are infected and can be contaminated by carrying a virus and bacteria. If ARI is not immediately there will be special treatment severe illness because it enters the lung tissue and will cause it pneumonia, so that it affects death to people who experience it especially the age of a baby child. ARI is a disease caused by viruses, bacteria and rickets and fungi which are transmitted through the air (water borne disease). Causative factor which can increase the incidence of ARI, namely environmental factors, factors babies, behavioral factors, family factors, and health services \cite{1}.

A generalized poisson regression is an alternative model for the data in the form of count where there is an assumption violation in the poisson distribution, the mean and variance have the same value \cite{3}. One of the based model that can overcome overdispersion or underdispersion cases has been developed is MARS. Multivariate Adaptive Regression Spline (MARS) as a statistical method for fitting the relationship between a set of input variables and dependent variables \cite{4}.
Multivariate Adaptive Generalized Poisson Regression Spline (MAGPRS) model is used as the development of the MARS and Generalized Poisson Regression (GPR). The purpose of this study is the first parameter estimation of MAGPRS model using Weighted Least Squares (WLS) method. The next do the test statistics from the MAGPRS model using Maximum Likelihood Ratio Test (MLRT). The application of MAGPRS model was carried out in the case of the number of Acute Respiratory Infection in babies from surabaya health department 2017.

2. Literature review

2.1. Generalized Poisson Regression (GPR)

Poisson regression cannot be used in overdispersion or underdispersion data. Let \( Y_i \sim GP(\mu, \theta) \), \( i = 1, 2, ..., n \) then \( Y_i \) has a probability function

\[
 f(y_i; \mu, \theta) = \left( \frac{\mu}{1 + \theta \mu} \right)^{y_i} \frac{(1 + \theta y_i)^{y_i - 1}}{y_i!} \exp \left\{ -\theta(1 + \theta y_i) \right\}
\]

where mean of \( Y_i \) is given by \( E(Y_i) = \mu \) and the variance of \( Y_i \) is given by \( Var(Y_i) = \mu(1+\theta\mu)^2 \). The parameter \( \theta \) measures the dispersion. When \( \theta = 0 \), the model reduces to the Poisson regression (PR) model. When \( \theta > 0 \), the model represents count data with over-dispersion and when \( \theta < 0 \), the GPR model represents count data with under-dispersion.

2.2. Multivariate Adaptive Regression Spline (MARS)

MARS use a set of special function to capture the nonlinearity of data, which is approximated with different regression slope in the corresponding interval of each predictor. The intervals are closed and non-overlapping except for the boundaries, which are called knots. MARS model which is a two-stage process; forward and backward. In the first stage, MARS constructs a model with an extra large number of Basis Function (BF), which deliberately overfits the data. Then, some of the BFs that contribute least to the overall performance are removed. Therefore, the forward construction may initially include insignificant model terms. In the backward pruning step, these terms are excluded. Thus, the backward step reduces the complexity of the model without degrading the fit to the data.

The MARS model is formulated as follows:

\[
 f(x_i) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} \left[ s_{km}(x_{(k,m)i}) - t_{km} \right] + \varepsilon_i = \alpha_0 + \sum_{m=1}^{M} \alpha_m B_{mi}(x_i) + \varepsilon_i
\]

where,

- \( \alpha_0 \) : intercept of the MARS model
- \( \alpha_m \) : regression coefficient on the \( m \)th basis function
- \( B_{mi}(x_i) \) : the \( m \)th basis function in MARS
- \( t_{km} \) : knot
- \( s_{km} \) : the basic function sign is if +1 knot is located on the right or -1 if the knot is on the left
- \( \varepsilon_i \) : error in observation to i
- \( x_{(k,m)i} \) : independent variable

At the end of the forward stage of the MARS algorithm, the largest model that overfits the data is obtained. In the backward stage, in each step, the least contributing BF that causes the smallest increase in the residual sum of squares is deleted from the model set. In order to estimate the optimal value of the model size, MARS again uses the GCV, defined as follows:
\[
GCV = \frac{MSE}{n} \left( 1 - \frac{C(M)}{n} \right) = \frac{n^{-1} \left[ (y - \hat{f}(x))^T (y - \hat{f}(x)) \right]}{1 - \frac{C(M)}{n}}
\]

where,

- \( C(M) \): penalty function of model complexity
- \( \hat{f}(x) \): estimated function in MARS
- \( GCV \): generalized cross validation statistic

2.3. Multivariate Adaptive Generalized Poisson Regression Spline (MAGPRS)

The Multivariate Adaptive Generalized Poisson Regression Spline model is a development from MARS and GPR. The response variable is generalized poisson distribution \( Y \sim GP(\mu, \theta) \) with \( E(Y) = \mu \), and \( \text{Var}(Y) = \mu(1 + \theta \mu)^2 \) \([10]\). The assumptions can be written as,

\[
y = \exp(B\varphi) + \varepsilon
\]

where,

\[
y = (y_1, y_2, \ldots, y_n)^T, \varphi = (\alpha_0, \alpha_1, \ldots, \alpha_M)^T, \varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T
\]

Probability density function is said to be an exponential family if the function can be transformed into an exponential family function. Recall that the single-parameter exponential family is expressed as:

\[
P(x|\eta) = h(x) \exp \left\{ \eta^T T(x) - A(\eta) \right\}
\]

where \( \eta \) is the natural parameter, \( T(x) \) is the sufficient statistics, \( A(\eta) \): log partition function and \( \mu \) is the mean parameter. We can write out the Generalized Poisson distribution in the exponential family form by applying the \( \exp(\log(.) ) \) function:

\[
P(x|\eta) = \exp \left\{ \log \left( \frac{\mu}{1 + \theta \mu} \right)^x \frac{1}{x!} \exp \left( -\mu(1 + \theta \mu) \right) \right\}
\]

\[
= \exp \left\{ \log \left( \frac{\mu}{1 + \theta \mu} \right)^x + \log \left( \frac{(1 + \theta \mu)}{1 + \theta \mu} \right) + \log \left( \exp \left( -\mu(1 + \theta \mu) \right) \right) \right\}
\]

\[
= \frac{1}{x!} (1 + \theta \mu)^{-x} \exp \left\{ x \log \mu - \mu(1 + \theta \mu) \right\}
\]

where \( \mu = \exp\{\eta\} \) and \( \eta = \log \mu, \eta = B\varphi \)

The MAGPRS model as follows,

\[
\mu = \exp\{\eta\}
\]

\[
\mu = \exp(B\varphi)
\]

\[
\mu = \exp\left( \alpha_0 + \sum_{m=1}^M \alpha_m B_m(x) \right)
\]

the parameter estimation Multivariate Adaptive Generalized Poisson Regression Spline use Weighted Least Squares method. The parameter estimation of the MAGPRS model as follows \([10] \),
\[
\hat{f}(x) = \exp(B\hat{\alpha}) = \exp\left(B(B^{-1}\ln\left( (B^T \exp(\hat{\alpha}^T B^T) W)^{-1} B^T \exp(\hat{\alpha}^T B^T) W_2^T ) \right) \right)
\]

\[
\hat{f}(x) = (B^T \exp(\hat{\alpha}^T B^T) W)^{-1} B^T \exp(\hat{\alpha}^T B^T) W_2^T
\]

2.4. Statistics Test of Multivariate Adaptive Generalized Poisson Regression Spline (MAGPRS)

The parameter estimation in Multivariate Adaptive Generalized Poisson Regression Spline (MAGPRS), carried out testing simultaneously use the Maximum Likelihood Ratio Test (MLRT) method. The hypothesis of simultaneously testing is \( H_0: \alpha_1 = ... = \alpha_n = 0 \) and \( H_1: \) at least have one \( \alpha_j \neq 0, j = 1,...,m \). The set of parameters that under the population \( \Omega = \{\theta, \alpha_1, \ldots, \alpha_m\} \), and set of parameters that under \( H_0 \) is \( \omega = \{\theta\} \). Then, the test statistics with equation below \( G^2 = -2 \ln \Lambda = 2 \left( \ln L(\hat{\Omega}) - \ln L(\hat{\omega}) \right) \) reject \( H_0 \) if \( G^2 > \chi^2_{\alpha,n} \) when \( v = n(\Omega) - n(\omega) \) [10].

3. Research methodology

The application of MAGPRS was carried out in the case of the number of Acute Respiratory Infection in babies in Surabaya, East Java Province 2017. Data sourced from surabaya health department. The dependent variable of this study is the number of ARI infants (Y). The seven independent variables which influence the response variables are the percentage of babies given non-exclusive breastfeeding (X_1), the percentage of birth weight is low (X_2), the percentage of basic immunization is incomplete (X_3), the percentage of unclean and unhealthy life behavior (X_4), the percentage of unhealthy houses (X_5), the percentage of babies not getting vitamin A (X_6), the percentage of not getting coverage of infant health services (X_7). Step of analyses in this study are as follows:

- Describing the characteristics of the number of ARI infants in babies as research object
- Carrying out overdispersion and underdispersion testing
- Estimating the parameters of the MAGPRS model with the Weighted Least Squares (WLS)
- Conducting hypothesis testing simultaneously and partially with Maximum Likelihood Ratio Test (MLRT)
- Drawing conclusions from the analysis of the MAGPRS model

4. Analysis and discussion

The first step of analysis is to detect overdispersion or underdispersion of the number of ARI in infants. In the Poisson regression, the assumption of equidispersion between a mean and a variance of the response variable must be satisfied. However, this assumption is often violated in real cases. The variance of the response is often greater than the mean resulting an overdispersion problem. To detect the overdispersion, we use a value of deviance/df or pearson/df. If the deviance/df or pearson/df is greater than 1, then the overdispersion occurs.

| Variable                  | Criteria       | Value/df |
|---------------------------|----------------|----------|
| The number of ARI         | Deviance       | 14.44    |
|                           | Pearson chi-square | 16.53    |

\[
H_0 : \frac{\text{Var}(Y)}{E(Y)} = 1
\]

\[
H_1 : \frac{\text{Var}(Y)}{E(Y)} > 1
\]

\[
D = 2 \sum_{i=1}^{n} \left( Y_i \ln \frac{Y_i}{\hat{\mu}_i} - (Y_i - \hat{\mu}_i) \right)
\]
Table 1 shows that the Deviance/df value of 14.44 is greater than 1. Thus, the null hypothesis is rejected and concluded that the number of ARI in infants has an overdispersion.

The number of predictors in this study is 7 so the appropriate number of BF includes 14, 21, and 28. Maximum Interaction (MI) is the number of interactions that can occur in the model. The maximum interaction used in this study is 1, 2, and 3 because according to Friedman (1991) if the maximum interaction used is more than 3, the value of GCV will increase and the model used will be increasingly complex. Minimum Observation (MO) is the minimum number of observations between knots. In this study, the MO used is 0, 1, 2, 3, 4, and 5 because the GCV value above will increase. The number of possible models based on the combination are as many as 54 models. Based on the empirical using Multivariate Adaptive Generalized Poisson Regression Spline (MAGPRS) of trial and error combination BF, MI, and MO, the combination of which produces minimum GCV value is combination of 28, 3, 1 with a value of GCV = 6.08E-05 with $R^2 = 0.878131$. Based on the results of this combination, it is known MAGPRS models produced are as follows:

$$
\hat{\mu} = \exp \left( \hat{a}_0 + \sum_{m=1}^{M} \hat{a}_m \sum_{k=1}^{5} [s_{km}(x_{i(k,m)} - f_{km})] \right)
$$

$$
\hat{\mu} = \begin{bmatrix}
2.05874 + 1.27478 h(x_2 - 0.12) + 1.18897 h(96.64 - x_5) + 3.19983 h(x_5 - 96.64) + 0.34207 h(x_2 - 12) \\
\times h(x_5 - 96.47) + 0.89811 h(x_2 - 0.12) \times h(96.47 - x_5) + 0.99921 h(x_1 - 85.17) \times h(x_2 - 0.12) \\
\times h(96.47 - x_5) + 1.00466 h(85.71 - x_1) \times h(x_2 - 0.12) \times h(96.47 - x_5) + 1.00356 h(x_1 - 72.3) \\
\times h(96.64 - x_5) + 0.99399 h(72.3 - x_1) \times h(96.64 - x_5) + 1.00450 h(x_1 - 72.3) \times h(x_2 - 1.38) \times h(96.64 - x_5) \\
+ 0.97645 h(x_1 - 72.3) \times h(1.38 - x_2) \times h(96.64 - x_5) + 1.0043 h(x_4 - 91.31) \times h(96.64 - x_5) \\
+ 1.00018 h(x_1 - 72.3) \times h(91.31 - x_4) \times h(96.64 - x_5) + 0.99099 h(72.3 - x_4) \times h(96.64 - x_5) \times h(x_7 - 99.88) \\
+ 0.99908 h(72.3 - x_1) \times h(96.64 - x_5) \times h(99.88 - x_7) + 0.99999 h(x_1 - 72.3) \times h(96.64 - x_5) \\
\times h(x_5 - 83.15) + 0.99559 h(x_1 - 72.3) \times h(96.64 - x_5) \times h(83.15 - x_5) + 0.99402 h(x_1 - 92.07) \\
\times h(96.64 - x_5) + 1.00260 h(92.07 - x_3) \times h(96.64 - x_5) + 1.0557 h(x_1 - 78.36) \times h(x_2 - 0.12) \\
\times h(x_5 - 96.47) + 1.0619 h(78.36 - x_1) \times h(x_2 - 0.12) \times h(x_5 - 96.47) + 1.00114 h(x_3 - 92.07) \times h(x_4 - 70.2) \\
\times h(96.64 - x_5) + 1.00047 h(x_3 - 92.07) \times h(70.2 - x_4) \times h(96.64 - x_5) + 0.99972 h(72.3 - x_1) \\
\times h(x_1 - 59.56) \times h(96.64 - x_5) + 0.99991 h(72.3 - x_1) \times h(59.56 - x_4) \times h(96.64 - x_5)
\end{bmatrix}
$$

Simultaneous hypothesis testing of parameters was carried out to find out predictor variables together to influence the response variable. The testing hypothesis of the MAGPRS parameters simultaneously as follows,

$$H_0 : \alpha_1 = \ldots = \alpha_M = 0$$

$$H_1 : \text{there is at least one } \alpha_j \neq 0, j = 1, \ldots, M$$

The test statistics of $G^2$ is 419.45 which is greater than $\chi^2_{(0.05;25)}$ of 37.65. So that the decision to reject $H_0$ is obtained which means that at least one basis function selected in the MAGPRS model has an influence on the response variable. Therefore it is necessary to conduct partial test as presented in Table 2.

$$H_0 : \alpha_j = 0$$

$$H_1 : \alpha_j \neq 0, j = 1, \ldots, M$$
| Parameter | Estimation value | Standard Error (SE) | t value | P-Value |
|-----------|-----------------|---------------------|---------|---------|
| (Intercept) | 7.22E-01 | 3.09E-16 | 2.341E+15 | < 2E-16 |
| h(X2-0.12) | 2.43E-01 | 1.14E-16 | 2.12E+15 | < 2E-16 |
| h(X2-0.12)*h(Xs-96.47) | -1.07E+00 | 4.75E-16 | -2.26E+15 | < 2E-16 |
| h(X2-0.12)*h(96.47-Xs) | -1.08E-01 | 2.93E-17 | -3.67E+15 | < 2E-16 |
| h(Xs-96.64) | 1.16E+00 | 3.37E-16 | 3.45E+15 | < 2E-16 |
| h(96.64-Xs) | 1.73E-01 | 4.44E-17 | 3.90E+15 | < 2E-16 |
| h(X2-85.71)*h(X2-0.12)*h(96.47-Xs) | -7.90E-04 | 7.07E-18 | -1.12E+14 | < 2E-16 |
| h(85.71-X1)*h(X2-0.12)*h(96.47-Xs) | 4.46E-03 | 1.52E-18 | 2.94E+15 | < 2E-16 |
| h(X2-72.3)*h(96.64-Xs) | -3.6E-03 | 4.55E-18 | -7.82E+14 | < 2E-16 |
| h(72.3-X1)*h(96.64-Xs) | -6.02E-03 | 2.82E-18 | -2.13E+15 | < 2E-16 |
| h(X2-72.3)*h(X2-1.38)*h(96.64-Xs) | 4.49E-03 | 4.01E-18 | 1.12E+15 | < 2E-16 |
| h(X2-72.3)*h(1.38-X1)*h(96.64-Xs) | -2.38E-02 | 8.66E-18 | -2.75E+15 | < 2E-16 |
| h(X2-72.3)*h(Xs-91.31)*h(96.64-Xs) | 4.31E-04 | 1.41E-18 | 3.05E+14 | < 2E-16 |
| h(X2-72.3)*h(91.31-X1)*h(96.64-Xs) | 1.82E-04 | 2.22E-19 | 8.21E+14 | < 2E-16 |
| h(72.3-X1)*h(96.64-Xs)*h(Xs-99.88) | -9.05E-03 | 2.94E-18 | -3.07E+15 | < 2E-16 |
| h(72.3-X1)*h(96.64-Xs)*h(99.88-Xs) | -9.18E-04 | 1.82E-19 | -5.03E+15 | < 2E-16 |
| h(X2-72.3)*h(96.64-Xs)*h(Xs-83.15) | -8.31E-07 | 4.57E-19 | -1.82E+12 | < 2E-16 |
| h(X2-72.3)*h(96.64-Xs)*h(83.15-X1) | -4.41E-03 | 4.52E-18 | -9.77E+14 | < 2E-16 |
| h(Xs-92.07)*h(96.64-Xs) | -5.59E-03 | 5.82E-18 | -1.03E+15 | < 2E-16 |
| h(92.07-X1)*h(96.64-Xs) | 2.60E-06 | 9.83E-19 | 2.65E+15 | < 2E-16 |
| h(X1-78.36)*h(Xs-0.12)*h(Xs-96.47) | 5.37E-02 | 3.44E-17 | 1.56E+15 | < 2E-16 |
| h(78.36-X1)*h(Xs-0.12)*h(Xs-96.47) | 6.01E-02 | 3.62E-17 | 1.66E+15 | < 2E-16 |
| h(Xs-92.07)*h(Xs-70.2)*h(96.64-Xs) | 1.14E-03 | 5.05E-19 | 2.27E+15 | < 2E-16 |
| h(X2-92.07)*h(70.2-X1)*h(96.64-Xs) | 4.79E-04 | 4.33E-19 | 1.11E+15 | < 2E-16 |
| h(72.3-X1)*h(Xs-59.56)*h(96.64-Xs) | -2.76E-04 | 1.54E-19 | -1.80E+15 | < 2E-16 |
| h(72.3-X1)*h(59.56-X1)*h(96.64-Xs) | -8.61E-05 | 4.07E-19 | -2.11E+14 | < 2E-16 |

\[ \theta = -0.080326 \quad 0.005364267 \quad -1495687 \quad 0.000 \]

This partial test uses a \[ |t_{value}| \] compared to \[ t_{table} \] by using a significance level 5% obtained \[ t_{0.05/2;35} = 2.030 \]. Table 2 show that \[ |t_{value}| \] from each predictor variables greater than 2.030 so the decision reject \[ H_0 \] so that all basis functions selected in the MAGPRS model have an influence on the number of ARI infants in Surabaya 2017.

The value of the importance of predictor variables in the grouping function with increasing GCV values in Table 3. It can be seen that the percentage of low birth weight (X2) and the percentage of unhealthy houses (X3) are the most important variables in the MAGPRS model with 100 importance. Then followed by the percentage given non-exclusive breastfeeding to infants (X1) with an importance of 81.4. In the fourth order, the variable the percentage of basic immunization is incomplete (X3) and the percentage of unclean and unhealthy behavior (X2) with a level of importance of 63.6. The last two variables are the the percentage of not getting coverage of infant health services (X1) with a value of importance of 50.6 and the percentage of babies not getting vitamin A (X2) has the smallest value for the variable importance level of 16.
Table 3. The importance of the predictor variable

| variable | nsubsets | gcv | rss |
|----------|----------|-----|-----|
| X₂       | 25       | 100 | 100 |
| X₅       | 25       | 100 | 100 |
| X₁       | 21       | 81.4| 81.4|
| X₃       | 18       | 63.6| 63.6|
| X₄       | 18       | 63.6| 63.6|
| X₇       | 16       | 50.6| 50.6|
| X₆       | 4        | 16  | 16  |

Based on the value importance of predictor variables in MAGPRS, the variables that affect the number of ARI patients in infants are the percentage of low birth weight (X₂), the percentage of unhealthy houses (X₅), and the percentage given non-exclusive breastfeeding to infants (X₁).

5. Conclusion
The results showed that the importance of predictor variables in Multivariate Adaptive Generalized Poisson Regression Splines (MAGPRS), the variables affecting the number of ARI patients in infants are the percentage of low birth weight (X₂), the percentage of unhealthy houses (X₃), and the percentage given non-exclusive breastfeeding to infants (X₁).

References
[1] Notoatmodjo S 2010 Health Research Methodology Jakarta Rineka Cipta
[2] Consul P and Jain G 1973 A Generalization of the Poisson Distribution Technometrics 15 791-799
[3] Wang W And Famoye F 1997 Modeling Household Fertility Decisions with Generalized Poisson Regression Journal of Population Economics 10 273-283
[4] Friedman J H 1991 Multivariate Adaptive Regression Splines (with discussion) Annual Statistics 191-141
[5] Famoye F and Wang W 2004 Censored Generalized Poisson Regression Model Computational Statistics and Data Analysis 46 547-560
[6] Friedman J H 1990a Estimating Functions of Mixed Ordinal and Categorical Variables Using Multivariate Adaptive Regression Splines Technical Report LCS 107 Statistics Department Stanford University
[7] Friedman J H 1990b Some Comments on Multivariate Adaptive Regression Splines Seattle WA 98195
[8] Friedman J H 1991 Multivariate Adaptive Regression Splines (with discussion) Annual Statistics 191-141
[9] Cameron A C and Trivedi P K 1998 Regression Analysis of Count Data Cambridge University Press Cambridge
[10] Hidayati S 2019 Multivariate Adaptive Generalized Poisson Regression Splines (MAGPRS) Thesis Department of Statistics FMKSD ITS Surabaya
[11] Ampulembang A P Otok B W Rumiati A T Budiasih 2015 Biresponses Nonparametric Regression Model Using MARS and its Properties Applied Mathematical Sciences 9 1417-1427