VOL. 11, NO. 23, DECEMBER 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

MODELLING LOCAL POLYNOMIAL FOR LONGITUDINAL DATA A CASE STUDY: INFLATION SECTORS IN INDONESIA

Suparti[1], Alan Prahutama[1] and Rita Rahmawati[1] and Tiani W. Utami[2]
[1]Departement of Statistcs, Diponegoro University, Semarang, Indonesia
[2]Departement of Statistcs, University of Muhammadiyah Semarang, Semarang, Indonesia
E-Mail: suparti702@gmail.com

ABSTRACT

Regression analysis is one of statistical methods for modelling the relation between response variable and predictors variable. Analysis of regression approach was doing three ways such as parametric regression, nonparametric regression and semiparametric regression. One of nonparametric regression methods is Local Polynomial, which is it need kernel function to modelling. Longitudinal data is the data tha combine time series and cross sectional data. Nowdays, we can developed modelling longitudinal data used local polynomial. The first steps to modelling it, we should find optimum bandwidth. One of method to find optimum bandwidth is use Generalized Cross Validation (GCV) method. The optimum bandwidth is smallest GCV's value. In this paper, we modelling seven inflation sector in Indonesia. The sectors are (1) foodstuffs; (2) food, beverages, cigarettes and tobacco; (3) housing, water, electricity, gas and fuel; (4) clothing; (5) health sector; (6) education and sport sector; and (7) transportation, communication and financial services. We used Gaussian kernel function as weighted. The results of this research produce R-square 82.01%.

Keywords: local polynomial, longitudinal data, GCV, inflation sectors.

1. INTRODUCTION

Regression analysis is one of statistical methods for modelling the relation between response variable and predictors variable. Analysis of regression approach was doing three ways such as parametric regression, nonparametric regression and semiparametric regression. Parametric regression could be used if curve's shape of regression has been known. Whereas, nonparametric regression could used done if curve's shape of regression has been unknown. Semiparametric regression could be used if part of curve's shape of regression has been known and others part of curve's shape has been unknown (Eubank, *et. al,* 2004). Curve of nonparametric regression has been assumed smooth that in certain function space, such as sobolev. In nonparametric regression's procedurs, the data will find shape of curve regression itself, without subjectivity influenced of researchers (Wu and Zhang, 2006).

One of nonparametric regression methods that have been developed besides spline and kernel is Local Polynomial. Local polynomial has some excees such as can relieve bias asymptotic and produce good estimate (Xiao, *et al.,* 2003). Estimation of polynomial local used weighted Least Square (WLS). The procedure of WLS was minimizing function to got the parameters. While to determine optimal bandwidth can be used Generalized Cross Validation method (Huang, 2003). Some of research about local polynomial has been developed such as, Bellhouse and Stafford (2003) developed local polynomial for complex survey. Another research were nonparamteric regression with local polynomial in autocorrelation error (Xiao *et al*., 2003) and bandwidth with plug ini for local polynomial regression in correlation errors (Fransisco *et al*., 2004). In addition, Delaigle *et al.,* (2009) developed local polynomial could been developed with adaptive design.

However, in regression analysis there are two kinds of data, such as time series data and cross section data. The time series data is the data from a subject that be repeated observed from time to times. Whereas cross section data is the data from some of subject that be done one repeated observed in one subject, so that one subject to another subjects are independent. The combination of time series data and cross section data formed longitudinal data. Longitudinal data is the data form repeated measured in each subject in different time's interval (Wu and Pourahmadi, 2003). There are some advantages of research in longitudinal data, such as we known alteration of each subject. In longitudinal data, is not needed much subject because one subject repeated measure. In addition the estimation of longitudinal data is efficient because to estimate can be done as collective for all subjects and observations. Some of research about longitudinal data has been developed using kernel methods (Lin *et al*., 2004). Another research used Partially Linear Model (Hu *et al*., 2004) and used smoothing spline (Ibrahim and Suliadi, 2008).

Inflation is the tendency of trend or the general price level rising movements which take place continuously from one period to the next periods. Controlled inflation can support the preservation of purchasing power in market. Meanwhile, unstable inflation will complicate business planning of business activities, both in production and investment activities as well as in the pricing of goods and services produced. According to Bank Indonesia (2013), a grouping of inflation in Indonesia as measured by the Consumer Price Index (CPI) can be divided into seven sectors of expenditure, namely (1) foodstuffs; (2) food, beverages, cigarettes and tobacco; (3) housing, water, electricity, gas and fuel; (4) clothing; (5) health sector; (6) education and sport sector; and (7) transportation, communication and

financial services. Seventh sectors such spending greatly affect inflation in Indonesia.

According to Statistics Department of Indonesia (2013), the increase in inflation in March 2013 contributed by the increase in foodstuff sector amounted 2.04%; food sector, beverages, tobacco and cigarettes by 12.04%; housing, water, electricity, gas and fuel reaches 0.21%; Health sector reached 0, 24%; education sector reached 0.12%; the transport and communication sectors reaches 0.19%, while inflation in August 2013 due to the increase of horticultural crops and beef. Inflation in March 2014 reached 7.32% with the largest contribution caused of inflation is the increase in price of chili and rice. The increase in the price of fuel oil (BBM) and rice accounted for inflation during March 2015.

Therefore, frequent changes to inflation in every commodity and every sector of expenditure than modelling of inflation sector are needed. It aims to determine the movements from time to time change of value inflation in each sector. Seventh sectors inflation in Indonesia were conducted as longitudinal data. For that reason, in this paper discuss about modelling local polynomial for longitudinal data in Inflation sector in Indonesia.

2. MATERIAL AND METHODS

2.1 Kernel function

One of estimate methods in local polynomial is using Weighted Least Square (WLS) so that needs weighted. The weighted that can be used to get estimate is kernel function. The definition of Kernel function is

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right) \quad (1)$$

for $-\infty < x < \infty$ and $h > 0$.

One of kernel function is Gaussian kernel function. It can be defined as (Wu and Zhang, 2006)

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2}\left(-x^2\right)\right) \quad (2)$$

2.2 Local polynomial for longitudinal data

Given longitudinal data in j-th subject and k-th observed as:

Given longitudinal data in j-th subject and k-th observed as (t_{ij}, y_{ij}) with $i = 1,2,...,m$ and $j = 1,2,...,n$. The regression equation is obtained as follow as:

$$y_{ij} = \eta(t_{ij}) + \varepsilon_{ij} \quad (3)$$

$\eta\left(t_{ij}\right)$ is a curve that unknown of the shape, so that we can estimate by nonparametric regression. One of approach to estimate $\eta\left(t_{ij}\right)$ is local polynomial method. Local polynomial estimator obtained by Taylor series that includes a polynomial of degree p. If $\eta\left(t_{ij}\right)$ could be conducted according to Taylor series with polynomial degree p obtained as (Welsh and Yee, 2005)

$$\eta\left(t_{ij}\right) \approx \eta(t) + \left(t_{ij} - t\right)\eta^{(1)}(t) + \cdots + + \left(t_{ij} - t\right)^p \eta^{(p)}(t) \Big/ p!$$

where $t_{ij} \in \left[t-h, t+h\right]$.

For example $\beta_r(t) = \dfrac{\eta^{(r)}(t)}{r!}$ with $r = 0,1,2,..., p$ then the equation (4) can be written by

$$\eta\left(t_{ij}\right) \approx \beta_0(t) + \left(t_{ij} - t\right)\beta_1(t) + \cdots + + \left(t_{ij} - t\right)^p \beta_p(t) \quad (5)$$

To get estimator $\hat{\boldsymbol{\beta}}$ could be minimizing use *Weighted Least Square* (WLS) such as:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} (y_{ij} - x_{ij}^T \beta)^2 K_h(t_{ij} - t) \quad (6)$$

where K_h is kernel function [9].

2.3 Optimal bandwidth selection with GCV method

Bandwidth is a controller between function with data, so that generated function become smooth. Selection optimal bandwidth will produce optimal estimator. Therefore, the selection optimal bandwidth is very important in nonparametric regression. One of ways to determine optimal bandwidth can be use Generalized Cross Validation method (Wu and Zhang, 2006). Suppose an estimation curve generating function, where every bandwidth h are *A(h)*.

$$\hat{\mathbf{y}} = \mathbf{A}(h)\mathbf{y}$$

so

$$GCV(h) = \frac{MSE(h)}{\left(\frac{1}{2mn} tr\left[\mathbf{I} - \mathbf{A}(h)\right]\right)^2} \quad (7)$$

With

$$MSE(h) = \frac{1}{2mn}\sum_{j=1}^{m}\sum_{k=1}^{n} (y_{jk} - \hat{y}_{jk})^2 \quad . \quad (8)$$

The smallest value of GCV will provide optimal bandwidth value.

2.4 Methods

This research is development from nonparametric regression method with local polynomial approach for longitudinal data in case sectors inflation in Indonesia. Response variables (as subject variable) in this research is

inflation value year on year in every sectors, while predictor variable (as observation of time) is time, from January 2009 until Maret 2016. The subjects consist of inflation in the foodstuffs sector (Y_1); Food sector, beverages, cigarettes and tobacco (Y_2); Housing, Water, electricity, gas and fuels (Y_3); Clothes (Y_4); Health (Y_5); Education, recreation and sport (Y_6); transportation, communication and financial services (Y_7). The steps to modelling local polynomial for longitudinal data, as follows as:

a) Determine degree of local polynomial *p*
b) Determine weighted kernel function
c) Find optimal bandwidth with GCV method
d) Find model estimate based on degree of local polynomial

3. RESULTS

The results of this research are getting estimator local polynomial in longitudinal data and application this model to inflation sector in Indonesia.

3.1. Estimator local polynomial for longitudinal data

Given set longitudinal data (y_{ij}, t_{ij}) for $i = 1,2,\ldots,m$; $j = 1,2,\ldots,n$, with m shows number of subjects while n shows number of observation. Predictor variable consist of t while y as response variable. Nonparametric regression model for longitudinal data can be written:

$$y_{ij} = f(t_{ij}) + \varepsilon_{ij}$$

Error of observation in each subject $(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{in})$ correlating each other dan *error* between one subject to others $(\varepsilon_{1j}, \varepsilon_{2j}, \ldots, \varepsilon_{mj})$ are independent. So that, if model of longitudinal data, curve of $f(t_{ij})$ approach by local polynomial with degree of p then the model can be written as:

$$f_{ij}(t_{ij}) = f_{ij}(t) + (t_{ij} - t) f_{ij}^{(1)}(t) + (t_{ij} - t)^2 \frac{f_{ij}^{(2)}(t)}{2!} + \ldots + (t_{ij} - t)^p \frac{f_{ij}^{(p)}(t)}{p!} \quad (8)$$

If $\alpha_{ijr} = \dfrac{f_{ij}^{(r)}(t)}{r!}$ as $r = 0,1,2,\ldots,p$ then the equation of (11) and (12) can be written:

$$f_{ij}(t_{ij}) = \alpha_{i0} + (t_{ij} - t)\alpha_{i1} + (t_{ij} - t)^2 \alpha_{i2} + \ldots + (t_{in} - t)^p \alpha_{ip} \quad (9)$$

If the model can be written as matrix as:

$$\mathbf{y} = \mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \quad (10)$$

as $\mathbf{D}$ and $\boldsymbol{\alpha}$ definited by:

$$\mathbf{D} = \left[diag(\mathbf{D_1}, \mathbf{D_2}, \ldots, \mathbf{D_m}) \right]_{nm \times m(p+1)}$$

$$\mathbf{D_1} = \begin{bmatrix} 1 & (t_{11} - t) & (t_{11} - t)^2 & \cdots & (t_{11} - t)^p \\ 1 & (t_{12} - t) & (t_{12} - t)^2 & \cdots & (t_{12} - t)^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & (t_{1n} - t) & (t_{1n} - t)^2 & \cdots & (t_{1n} - t)^p \end{bmatrix}_{n \times (p+1)} \text{ and}$$

$$\mathbf{D_m} = \begin{bmatrix} 1 & (t_{m1} - t) & (t_{m1} - t)^2 & \cdots & (t_{m1} - t)^p \\ 1 & (t_{m2} - t) & (t_{m2} - t)^2 & \cdots & (t_{m2} - t)^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & (t_{mn} - t) & (t_{mn} - t)^2 & \cdots & (t_{mn} - t)^p \end{bmatrix}_{n \times (p+1)}.$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha_1} \\ \boldsymbol{\alpha_2} \\ \boldsymbol{\alpha_3} \\ \vdots \\ \boldsymbol{\alpha_m} \end{bmatrix}_{m(p+1) \times 1} ; \boldsymbol{\alpha_1} = \begin{bmatrix} \alpha_{10} \\ \alpha_{11} \\ \alpha_{12} \\ \vdots \\ \alpha_{1p} \end{bmatrix}$$

Given $\mathbf{K_h}$ as weighted kernel, can be definition as:

$\mathbf{K_h} = diag(K_h(t_{11} - t), K_h(t_{12} - t), \ldots, K_h(t_{mn} - t))_{N \times N}$;

$\mathbf{y}$ as a vector with size $mn \times 1$

To get estimator $\boldsymbol{\alpha}$ can use optimizing by *Weighted Least Square* (WLS) (Takezawa, 2006). If $\psi(\boldsymbol{\alpha})$ as a function of WLS then

$$\psi(\boldsymbol{\alpha}) = (\mathbf{y} - \hat{\mathbf{y}})^{\mathbf{T}} \mathbf{K_h} (\mathbf{y} - \hat{\mathbf{y}}) \quad (11)$$

If $\mathbf{V}\boldsymbol{\alpha} + \mathbf{H}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\theta}$ then

$$\psi(\boldsymbol{\alpha}) = \mathbf{y}^{\mathbf{T}}\mathbf{K_h}\mathbf{y} - 2\boldsymbol{\alpha}^{\mathbf{T}}\mathbf{D}^{\mathbf{T}}\mathbf{K_h}\mathbf{y} + \boldsymbol{\alpha}^{\mathbf{T}}\mathbf{D}^{\mathbf{T}}\mathbf{K_h}\mathbf{D}\boldsymbol{\alpha}$$

To get estimator $\boldsymbol{\alpha}$ can use optimizing by *Weighted Least Square* (WLS) so that the estimator is:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{D}^{\mathbf{T}}\mathbf{K_h}\mathbf{X})^{-1}\mathbf{D}^{\mathbf{T}}\mathbf{K_h}\mathbf{y} \quad (12)$$

If we definnition of $\hat{\mathbf{y}} = \mathbf{A}(h)\mathbf{y}$ then we got the value of A*(h)* as:

$$\mathbf{A}(h) = \mathbf{D}(\mathbf{D}^{\mathbf{T}}\mathbf{W}^{-1}\mathbf{K_h}\mathbf{D})^{-1}\mathbf{D}^{\mathbf{T}}\mathbf{W}^{-1}\mathbf{K_h} \quad (13)$$

After we got estimator of local polynomial for longitudinal data, we must find optimal bandwidth. The optimal bandwidth can be find with GCV method. The GCV method was calculated by MSE value. The MSE can be definition by

$MSE(h) = (mn)^{-1}(\mathbf{y} - \mathbf{D}\hat{\boldsymbol{\alpha}})^{\mathbf{T}}\mathbf{K_h}(\mathbf{y} - \mathbf{D}\hat{\boldsymbol{\alpha}})$.

While GCV value can be written by:

$$GCV(h) = \frac{MSE(h)}{\left[(mn)^{-1} trace(\mathbf{I} - \mathbf{A})\right]^2}.$$

The smallest GCV value is the optimal bandwidth.

3.2. Modelling local polynomial for longitudinal data in inflation sector in Indonesia

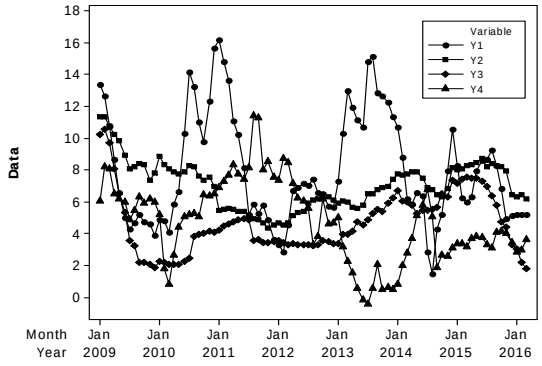

Figure-1. Time series plot subject Y1-Y4.

In Figure-1 we look that inflation sector in foodstuffs Pada Gambar 1 terlihat bahwa inflasi sektor bahan makanan (Y1) tend higher than Food sector, beverages, cigarettes and tobacco (Y_2); Housing, Water, electricity, gas and fuels (Y_3); and Clothes (Y_4).

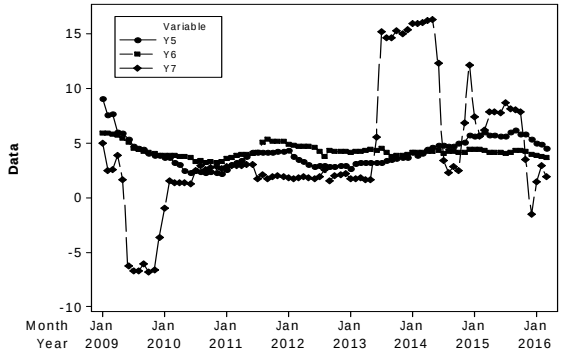

Figure-2. Time Series plot subject Y3-Y7.

In Figure-2. we look that inflation sector in transportation, communication and financial services (Y_7) higher in 2013 than Health sector (Y_5); and Education, recreation and sport (Y_6). While in 2009 until 2010, it lower than others.

3.2.1. Determine optimum bandwidth use GCV method

In this paper, we just use two degrees polynomial such as, linear and quadratic. To determine GCV value, we must trial and error the interval number of bandwidth. The bandwidth greater than zero.

Table-1. The value of minimum GCV for linear local polynomial (*p=1*).

y_i	h	t	GCV
y_1	0.8	22	0.08377421
y_2	1.1	13	0.05068069
y_3	0.9	10	0.01166412
y_4	0.5	40	0.07001320
y_5	1.2	35	0.04361142
y_6	0.4	28	0.02467148
y_7	0.7	63	0.05250300

Table-2. The value of minimum GCV for quadratic local polynomial (*p=2*).

y_i	h	t	GCV
y_1	0.6	30	0.06265444
y_2	1.5	10	0.04368968
y_3	1.2	25	0.12466715
y_4	0.4	22	0.08263722
y_5	1.0	60	0.06150042
y_6	0.6	43	0.01362249
y_7	0.8	55	0.04360300

3.2.2. Estimate local polynomial for longitudinal data in inflation case

Before we estimate model, we should combine optimal bandwidth of each subject. The optimal bandwidth based on Table-1 and Table-2 with minimum GCV. Table-3 shows that combination of degree polynomial for each subject.

Table-3. Combination of optimal bandwidth for each subject.

y_i	h	t	p
y_1	0.6	30	2
y_2	1.5	10	2
y_3	0.9	10	1
y_4	0.5	40	1
y_5	1.2	35	1
y_6	0.6	43	2
y_7	0.8	55	2

Table-4. Estimation of parameter local polynomial longitudinal data in each sector.

y_i	parameter	Estimasi parameter
y_1	α_{11}	0.2842
	α_{12}	0.9112
y_2	α_{21}	3.201
	α_{22}	0.1302
y_3	α_{31}	1.0212
y_4	α_{41}	0.0231
y_5	α_{51}	2.7200
y_6	α_{61}	0.0014
	α_{62}	3.1102
y_7	α_{71}	2.302
	α_{72}	1.1022

4. DISCUSSIONS

Local polynomial estimation of longitudinal data in this paper yields an estimate as weighted linear regression estimation (weighted Regression). However the weighted that used in local polynomial is kernel density function. Meanwhile, to get the optimum bandwidth using GCV method. GCV method is basically to minimize residual value of estimate model. Residual minimum is the optimum bandwidth. Forms of GCV value as a quadratic curve, where the value of the minimum GCV an optimum inflection point on the curve. Longitudinal data modeling is an efficient modeling, because the estimate several models at once to produce the R-square value of the single.

Modelling local polynomial for longitudinal data in Inflation sector in Indonesia as well as:

$$\hat{y}_1 = 0.2842(t_{1j} - 30) + 0.9112(t_{1j} - 30)^2$$
$$\hat{y}_2 = 3.201(t_{2j} - 10) + 0.1302(t_{2j} - 10)^2$$
$$\hat{y}_3 = 1.0212(t_{3j} - 10)$$
$$\hat{y}_4 = 0.0231(t_{4j} - 40)$$
$$\hat{y}_5 = 2.72(t_{5j} - 35)$$
$$\hat{y}_6 = 0.0014(t_{6j} - 43) + 3.1102(t_{6j} - 43)^2$$
$$\hat{y}_7 = 2.302(t_{7j} - 55) + 1.1022(t_{7j} - 55)^2$$

In this models produce R-square value is 82.01%.

5. CONCLUSSIONS

Nonparametric regression model more complicated procedure than parametric regression model. In this model, we do not need assumtions of model, it's more practice. To determine optimum bandwidth we still do trial and error, so that it's not guarantee that the bandwidth that we got was optimal ones. We couldn't guarantee that, the model local polynomial in this paper is the best one models.

ACKNOWLEDMENT

This research founded by Ministery of Reseach, Technology and Higher Education as Hibah Fundamental. Thanks to Statistics Laboratory of Statistics Department, Diponegoro University, Semarang, Indonesia.

REFERENCES

Bellhouse D.R., dan Stafford J.E. 2001. Local Polynomial regression in Complex Surveys. Statistics Canada, Catalouge: Survey Methodology. 27(2): hal. 197-203.

Delaigle A., Fan J., dan Carroll R.J. 2009. A Design-Adaptive Local Polynomial Estimator for the errors in Variable Problem, Journal of the Amercan Association. 104(485): hal. 348-359.

Eubank R.L., Huang C., Maldonado Y.M., Wang N., Wang S. and Buchanan R.J. 2004. Smoothing Spline Estimation in Varying-Coefficient Models, J. R. Statist. Soc. 66, Part 3, pp. 653-667.

Fransisco M.F., Opsomer F.J., dan Fernandez M.V. 2004. Plug in Bandwidth Selector for Local Polynomial Regression Estimator With Correlated Error. Nonparametric Statistics. 16(1-2): hal. 127-151.

Hu Z., Wang N., dan Carroll R.J. 2004. Profile Kernel versus Backfitting in the Partially Linier Models for Longitudinal or Clustered Data, Biometrika. 91(2): hal. 251-262.

Huang J.Z. 2003. Asymptotics for Polynomial Spline Regression under Weak Conditions, Statistics and Probability Letters. 65: 207-216.

Ibrahim N.A. dan Suliadi. 2008. Analyzing Longitudinal Data Using Gee –Smoothing Spline, Proceedings of the 8th WSEAS International Conference on Applied Computer and Applied Computational science, hal. 26-33.

Nonparametric Regression Model with Unequal Correlation of Errors. Journal of Mathematics and Statistics. 6(3): hal. 327-332.

Lin X., Wang N., Wels A.H., Carroll R.J. 2004. Equivalent Kernels of Smoothing Splines in Nonparametrics Regression for Clustered/Longitudinal Data. Biometrika. 91(1): hal. 177-193.

Takezawa K. 2006. Introduction to Nonparametric Regression. New Jersey: John Wiley and Sons, Inc.

Welsh A.H dan Yee, T.Y. 2005. Local Regression for Vector Responses. Journal of Statistical Planning and Inference. 136: hal. 3007-3031.

Wu H. dan Zhang J.T. (2006). Nonparametric Regression Methods for Longitudinal Data Analysis. A John-Wiley and Sons Inc. Publication, New Jersey, USA.

Wu W.B dan Pourahmadi M. 2003. Nonparemetric Estimation of Large Covariance Matrices of Longitudinal Data. Journal of Biometrika. 90(4): hal. 831-844.

Xiao Z., Linton O.B., Carroll R.J., Mammen E. 2003. More Efficient Local Polynomial Estimation in Nonparametric Regression with Auto correlated Errors. Jornal of the American Statistical Association. 99(464): hal. pp. 980-992.