Sdü Sağlık Bilimleri Dergisi / Cilt 11 Sayı 2 / 2020

# Comparison Of Piecewise Regression and Polynomial Regression Analyses In Health and Simulation Data Sets

## Simülasyon ve Sağlık Veri Setlerinde Parçalı Regresyon ile Polinom Regresyon Analizlerinin Karşılaştırılması

**Buğra Varol[1], İmran Kurt Ömürlü[2], Mevlüt Türe[2]**

[1]Aydin Adnan Menderes University, Institute of Health Sciences, Division of Biostatistics, Aydın, Turkey.

[2]Aydin Adnan Menderes University, Medical Faculty, Division of Biostatistics, Aydın, Turkey.

## Abstract

**Objective:** Piecewise regression, which one or more pieces are combined in breakpoints, is widely used as a statistical technique. It was aimed to compare piecewise regression analyses and polynomial regression analysis using both simulated data and real data sets.

**Material-Method:** In the application step of the study, algorithms were created by using R software for simulation practice. Polynomial and piecewise regression analysis methods were compared using data sets with n=100 units and 1000 times running simulation. Additionally, estimation performances of piecewise and polynomial regression were built by using the data sets which contained in the number of tuberculosis cases according to age in 2010 year and the number of measles cases from 1970 to 2015 years in Turkey were compared according to the coefficient of determination ($R^2$), mean square error (MSE), Akaike information criteria (AIC) and Bayes information criteria (BIC).

**Results:** It was found that there was a significant difference between all of the polynomial and piecewise regression models ($p<0.001$). $R^2$ values of piecewise regression models were higher than polynomial regression models; MSE, AIC and BIC values were observed to be lower. According to the result of both simulation and real data set applications, piecewise regression models which were generated according to optimal knots were found to have better estimation performance than polynomial regression models according to $R^2$, MSE, AIC and BIC criteria.

**Conclusions:** This study revealed that data analysis with piecewise regressions having optimal knots provided superiority statistically, although polynomial regression methods are preferred in the field of health studies mostly.

**Keywords:** Piecewise Regression, Simulation, Tuberculosis, Measles, Knot.

## Özet

**Amaç:** Bir veya daha fazla parçanın kırılma noktalarında birleştirildiği parçalı regresyon, istatistiksel bir teknik olarak yaygın bir şekilde kullanılmaktadır. Bu çalışmada hem simülasyon verisi hem de gerçek veri setleri kullanılarak tek değişkenli polinom regresyon analizi ile karesel ve kübik parçalı regresyon analizlerinin karşılaştırılması hedeflendi.

**Materyal-Metot:** Çalışmanın uygulama basamağında R yazılım programı kullanılarak simülasyon uygulaması için algoritmalar yazıldı. Polinom ve sürekli parçalı regresyon analiz yöntemlerinin karşılaştırılması n=100 birimlik veri setleri için 1000 tekrarlı simülasyon ile gerçekleştirildi. Ayrıca Türkiye'de 2010 yılındaki tüberküloz vaka sayılarını içeren tüberküloz veri seti ile Türkiye'deki 1970-2015 yılları arasındaki kızamık vaka sayılarını içeren kızamık veri setleri kullanılarak oluşturulan polinom ve parçalı regresyon modellerinin tahmin performansları; belirtme katsayısı ($R^2$), hata kareler ortalaması (HKO), Akaike bilgi kriteri (ABK) ve Bayes bilgi kriteri (BBK) değerlerine göre karşılaştırıldı.

**Bulgular:** Tüm polinom ve parçalı regresyon modellerinin $R^2$, HKO, ABK ve BBK değerleri bakımından performansları istatistiksel olarak birbirinden farklı bulundu ($p<0,001$). Parçalı regresyon modellerinin $R^2$ değerlerinin polinom regresyon modellerine göre daha yüksek; HKO, ABK ve BBK değerlerinin ise daha düşük olduğu gözlendi. Gerçek veri setleri ile yapılan uygulamalarda en uygun dönüm noktalarına göre oluşturulan tüm parçalı regresyon modellerinin $R^2$ değerlerinin polinom regresyonlardan daha yüksek; HKO, ABK ve BBK değerlerinin ise daha düşük olduğu belirlendi Oluşturulan parçalı regresyon modellerinin veri setlerini polinom regresyonlara göre daha iyi tahmin ettiği belirlendi.

**Sonuç:** Sağlık alanında yapılan çalışmaların çoğunda polinom regresyon yöntemlerinin tercih edilmesine rağmen bu çalışma ile en uygun dönüm noktalı parçalı regresyonlarla veri analizinin istatistiksel açıdan üstünlük sağladığı uygulamalarla ortaya konmuştur.

**Anahtar kelimeler:** Parçalı Regresyon, Simülasyon, Tüberküloz, Kızamık, Dönüm noktası.

**DOI:** 10.22312/sdusbed.636203
**Müracaat tarihi / Received date:** 22.10.2019
**Kabul tarihi / Accepted date:** 20.04.2020
**ORCID:** BV 0000-0001-8052-7782, İKÖ 0000-0003-2887-6656, MT 0000-0003-3187-2322

**Yazışma Adresi / Corresponding:** Buğra Varol,
Aydin Adnan Menderes University, Institute of Health Sciences, Division of Biostatistics,
09100 Aydin Turkey.
**Tel:** +90 256 225 31 66
**Fax:** +90 256 212 01 46
**E-posta / E-mail:** bugravarol87@gmail.com

## Introduction

Scientific studies in the field of health, polynomial regression models are used in addition to the linear regression models which are widely used for examining the relationship between the dependent and independent variables (1-4). However, if the point distribution of the relationship between dependent and independent variables shows deviations that cannot be expressed by polynomial regression models, the distribution of these models is reduced. In such cases, piecewise regression models are used (4-6). Examination of point distribution with piecewise regressions provides the researcher with possibilities to deal with low-level polynomials, to estimate for any interval of the independent variable, to obtain more flexible curves and to easily model complex distributions that cannot be explained by the known models according to the optimal knots (7, 8).

Polynomials that form piecewise functions can easily be combined in computer programs. Therefore, the use of piecewise polynomials is suitable for predicting particularly the experimental data or modelled curves (9, 10).

This work aims to compare the performances of quadratic and cubic piecewise regression analyses and univariate polynomial regression analysis using simulation data. Moreover, performances of quadratic, cubic piecewise regression and univariate polynomial regression were compared using 2010 tuberculosis data set and from 1970 to 2015 measles data set in Turkey. The performance of the generated models evaluated according to the coefficient of determination ($R^2$), mean square error (MSE), Akaike information criterion (AIC) and Bayesian information criterion (BIC).

## Material and Methods

### Piecewise Regression

"Piecewise regression" refers to the examination of point distributions of dependent and independent variables divided into pieces at specific points called knot (6, 11-13).

It may not always be appropriate to estimate a large number of (x,y) data points in a data set in the form of $\{(x_i; y_i): =1,\ldots, n\}$ with a single curve. As the number of points increases, deviations from the point distribution will also increase and estimation power of the generated model will be lower since the degree of polynomial representing the relation between x and y will increase too (14, 15). The piecewise regression approach is recommended for such cases. Piecewise regression is based on the principle involving a division of the data sets at specified intervals and an method in each interval with polynomials of an appropriate degree (4, 11, 16). If the knot is determined by the researcher at the beginning of the trial, such knots are called “fixed knots”. If it is not previously known and is determined by examining the point distribution obtained through the research, such knots are called the “variable knots” (17-19). The functions formed between the knots starting from the first knot are polynomials in the d’th degree (20, 21). Determination of location and number of a knot is closely related to the shape of the distribution. The approximate location and number of knots in piecewise regressions can usually be detected by visual inspection (22). Visual inspection of point distribution gives essential clues to the researcher about the division form to be applied, how a function (quadratic or cubic) will be used and how many pieces of point distribution will be examined (4, 23).

It is important to note that the areas with sudden directional changes are the possible knot regions. Furthermore, the function to be used in the division also determine the point at which the knot or points will be formed in the point distribution (7, 19).

### “+” Functions

The “+” functions are commonly used to create piecewise regressions. The regression models can be divided into pieces according to the knots determined by the functions “+”. A function “+” expressed by $(x-t)_+$, t being the knot, and (x-t) being the independent variable of piecewise function, is defined as follows:

$$(x-t)_+ = \begin{cases} x-t, & x > t \\ 0, & x \leq t \end{cases}$$

If x is less than or equal to knot point t, then function equals to 0. Otherwise, the function is equal to (x-t). Thus, an expression given by the function "+" does not affect the part of the piecewise regression model before the corresponding knot (24-26).

### Piecewise Regression Models

It is called to be “continuous piecewise regression” when different regions of point distributions show the distributions in respect of the same function or different functions, the case where functions created before and after any specified knot gives the same “$y_i$” value at this knot (11, 16, 27).

Each part that forms the piecewise regression has a unique fixed-term causing discontinuity. The coefficients that impair continuity can be removed from the model by applying continuity constraints into the piecewise regression model. The general form of piecewise regression without any restrictions is as follows (25, 26, 28, 29).

$$S(x_i) = y_i = \sum_{j=0}^{k} b_{0j} x_i^j + \sum_{l=1}^{m} \sum_{v=0}^{d} b_{lv} (x_i - t_l)_+^v + e_i \qquad i = 1,2,\ldots,n$$

In this equation; x represents the value of the independent variable, t is the knot value, $b_{0j}$ is the regression coefficient, $b_{lv}$ is the regression coefficient of the function “+”, k is the degree of the independent variable, d is the degree of the function “+”, e is the error term, m is the number of knots, and n is the sample unit number (25).

The piecewise regression equation given in the general form above can also be expressed as:

$$\begin{aligned} S(x_i) = {} & b_{00} + b_{01}x_i + \ldots + b_{0k}x_i^k \\ & + \left[b_{10}(x_i - t_1)_+^0 + b_{11}(x_i - t_1)_+^1 + b_{12}(x_i - t_1)_+^2 + \ldots\ b_{1d}(x_i - t_1)_+^d\right] \\ & + \left[b_{20}(x_i - t_2)_+^0 + b_{21}(x_i - t_2)_+^1 + b_{22}(x_i - t_2)_+^2 + \ldots\ b_{2d}(x_i - t_2)_+^d\right] + \ldots \\ & + \left[b_{m0}(x_i - t_m)_+^0 + b_{m1}(x_i - t_m)_+^1 + b_{m2}(x_i - t_m)_+^2 + \ldots\ b_{md}(x_i - t_m)_+^d\right] + e_i \end{aligned}$$

This equation is constrained by applying continuity constraints and $b_{10}, b_{20}, \ldots, b_{m0}$ the constants can be removed from the model. So the model becomes continuous (25, 28, 30).

$$S(x_i) = b_{00} + b_{01}x_i + \ldots + b_{0k}x_i^k + \left[b_{11}(x_i - t_1)_+ + b_{12}(x_i - t_1)_+^2 + \ldots b_{1d}(x_i - t_1)_+^d\right] + \left[b_{21}(x_i - t_2)_+ + b_{22}(x_i - t_2)_+^2 + \ldots b_{2d}(x_i - t_2)_+^d\right] + \ldots + \left[b_{m1}(x_i - t_m)_+ + b_{m2}(x_i - t_m)_+^2 + \ldots b_{md}(x_i - t_m)_+^d\right] + e_i$$

**Simulation Applications**

Two different simulation algorithms were performed in this study. These algorithms have some differences in the data generation phase. The comparison of polynomial and continuous piecewise regression analysis methods was performed according to $R^2$, MSE, AIC, and BIC values that were calculated after the simulation of n=100 with 1000 runs. Descriptive statistics were specified as median (25th-75th percentiles). The data was analyzed using the base and stats packages in R software.

*Second Degree Simulation Application:*

Independent variable, one knot and error term were randomly derived from x~U(1,100), t~U(40,60), e~U(-15,15) distributions, respectively.

Then the independent variable z=x-t and the dependent variable $y=-10+15*x-0.22*x^2+15*z+0.07*z^2-e$ were generated.

The quadratic regression model was estimated by the least square method (LSM) and the $R^2$, MSE, AIC and BIC values of this model were calculated.

The LSM method was used to estimate a piecewise regression model including two partial pieces in quadratic+quadratic structure divided into two according to the knot t, the first piece being formed only with the variable x, and the second piece being formed with the variables x and z. The $R^2$, MSE, AIC and BIC values of the model were calculated.

*Third Degree Simulation Application:*

Independent variable, two different knots and error term were randomly derived from the distributions x~U(1,100), $t_1$~U(30,50), $t_2$~U(51,70), e~U(-4,4), respectively.

Then the independent variables $z_1=x-t_1$, $z_2=x-t_2$ and the dependent variable $y=-10+7*x-0.2*x^2+0.0012*x^3+5.2*z_1-0.02*z_1^2+0.00009*z_1^3-2*z_2+0.01*z_2^2-0.00009*z_2^3-e$ were generated.

The cubic regression model was estimated by the LSM method and the $R^2$, MSE, AIC and BIC values of this model were calculated.

A piecewise regression model with two partial pieces in quadratic+cubic structure divided into two according to the knot $t_1$, the first part is formed only with the variable x and the second part being formed with the variables x and $z_1$, was estimated by the LSM method. The $R^2$, MSE, AIC and BIC values of the model were calculated.

Knot $t_3 = \frac{t_1+t_2}{2}$ and the independent variable $z_3=x-t_3$ were created to estimate a piecewise regression model with two partial pieces in a cubic+cubic structure. $t_3$ was determined as the knot and a piecewise regression model with two partial pieces in a cubic+cubic structure was estimated by the LSM method. The first part is formed only with the variable x; the second part is built with the variables x and $z_3$ in the generated model. The $R^2$, MSE, AIC and BIC values of the model were calculated.

A piecewise regression model with two partial pieces in cubic+cubic structure divided into two pieces according to the knot 3, the first piece being formed only with the variable x and the second part being formed with the variables x and $z_3$, was estimated by the LSM method. The $R^2$, MSE, AIC and BIC values of the model were calculated.

**Real Data Applications**

*Tuberculosis Data Set:*

The 2010 tuberculosis data set used in the study was retrieved from the study called "The Battle of Tuberculosis in Turkey 2012 Report" by the Turkish Public Health Institution (https://hsgm.saglik.gov.tr). The data set consists of 96 units and contains total tuberculosis cases according to age values ranging from 0-99. In order to create regression models, the total number of cases was taken as the dependent variable (y); age variable was taken as an independent variable (x) from this data set.

*Measles Data Set:*

The measles data set used in the study was retrieved from the webpage on "The Statistical Data of the Department of Vaccine-Preventable Diseases" on the website of the Turkish Public Health Institution (https://hsgm.saglik.gov.tr). The data set consists of 46 units and contains measles cases ranging from 1970 to 2015. In order to create regression models, the number of cases variable (y) was taken as the dependent variable and time variable (x) was taken as an independent variable from this data set.

## Results

**Simulation Results**

In the study conducted with the data derived from the second degree, it was observed that none of the $R^2$, MSE, AIC and BIC values were normally distributed according to the quadratic regression model. According to the piecewise regression model in quadratic+quadratic structure, only $R^2$ value was not normally distributed; MSE, AIC and BIC values were found to be normally distributed. For this reason, the Mann-Whitney U test was used to compare differences between formed models in terms of the $R^2$, MSE, AIC and BIC values in simulation with second-degree derived data. The models were found statistically different with respect to these values ($p<0.001$). As shown in Table 1, the piecewise regression model in quadratic+quadratic structure has higher $R^2$ value and lower MSE, AIC and BIC values. In the study conducted with the data derived from the third degree, it was observed that $R^2$ and MSE values were not normally distributed and AIC and BIC values were normally distributed. According to the piecewise regression model in quadratic+cubic structure, none of the $R^2$, MSE, AIC and BIC values were normally distributed. According to the cubic+cubic structure

in piecewise regression model, it was determined that only MSE value was not normally distributed, $R^2$, AIC and BIC values were normally distributed. For this reason, the Kruskal-Wallis test was used to compare differences between formed models in terms of the $R^2$, MSE, AIC and BIC values in simulation with third-degree derived data. The models were found statistically different with respect to these values. ($p<0.001$). In addition, as a result of multiple comparisons, it was concluded that all models are statistically different from each other with respect to $R^2$, MSE, AIC and BIC values ($p<0.001$ for comparison of all regression models). Piecewise regression models were found to have higher values of $R^2$ and lower values of MSE, AIC and BIC. The highest $R^2$ value was found in quadratic+cubic regression model, and the lowest values of MSE, AIC and BIC were also found in quadratic+cubic regression model (Table 2).

**Table 1.** Descriptive statistics and comparison results of $R^2$, MSE, AIC and BIC values of quadratic and piecewise regression models (quadratic+quadratic)

| Criteria | Model | | p |
|---|---|---|---|
| | Quadratic | Quadratic+Quadratic | |
| $R^2$ | 0.82 (0.65-0.93) | 0.98 (0.97 - 0.99) | <0.001 |
| MSE | 876.32 (760.00-1050.96) | 71.30 (67.02 - 76.45) | <0.001 |
| AIC | 969.36 (955.12 - 987.54) | 722.48 (716.30 - 729.45) | <0.001 |
| BIC | 979.78 (965.54-997.96) | 738.11 (731.93-745.08) | <0.001 |

**Table 2.** Descriptive statistics and comparison results of $R^2$, MSE, AIC and BIC values of cubic regression model and piecewise regression models (quadratic+cubic and cubic+cubic)

| Criteria | Model | | | p |
|---|---|---|---|---|
| | Cubic | Quadratic +Cubic | Cubic+Cubic | |
| $R^2$ | 0.77 (0.69 - 0.85) | 0.97 (0.96 - 0.98) | 0.96 (0.95 - 0.97) | <0.001 |
| MSE | 60.14 (52.24 - 70.59) | 7.68 (6.88 - 8.72) | 9.74 (8.53 - 11.4) | <0.001 |
| AIC | 703.46 (689.37 - 719.47) | 501.62 (490.59 - 514.39) | 527.44 (514.18 - 543.15) | <0.001 |
| BIC | 716.47 (702.40 - 732.50) | 519.85 (508.83 - 532.63) | 548.28 (535.02 - 564.00) | <0.001 |

**Tuberculosis Data Set Application Results**

The quadratic and cubic models were examined in polynomial and piecewise regression as the distribution of the number of cases according to age was more appropriate for the cubic structure in the tuberculosis data set.

*Model Estimation By Polynomial Regressions:*

Created quadratic and cubic regression equations to estimate the age-related tuberculosis cases were estimated by the LSM method.

According to the results, quadratic and cubic regression models were found statistically significant (For quadratic model: $F=101$; $df_1=2$, $df_2=93$; $p<0.001$; for cubic model: $F=128$; $df_1=3$, $df_2=92$; $p<0.001$). Additionally, all coefficients of both models were found statistically significant (Table 3).

*Model Estimation By Piecewise Regression:*

The distribution graph of the number of tuberculosis cases varying by age was given in Figure 1. When Figure 1 is examined, it is noticed that the number of cases of tuberculosis increased between the age of 6 and 21, the number of cases started to decrease after the age of 21, there was a steady course between ages 33-57 and a quick decline after age 57. Therefore, it was decided to model with two knots and three partial pieces functions. Ages 21 and 33 were determined as the best knots for applying piecewise regression as a result of experiments in regions where the number of cases of tuberculosis had jumps or deviated direction of distribution.

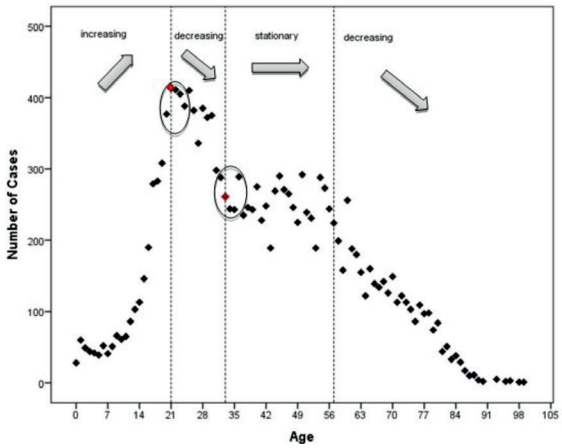


**Figure 1.** Distribution of tuberculosis cases according to age and candidate knot regions

**Table 3.** Quadratic and cubic regression parameter estimates

| Variable | Model | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Quadratic | | | | Cubic | | | |
| | b | $s_b$ | t | p | b | $s_b$ | t | p |
| Constant | 60.42 | 20.6 | 2.93 | 0.004 | -44.39 | 21.2 | -2.09 | 0.04 |
| x | 10.53 | 1 | 10.51 | <0.001 | 24.13 | 1.94 | 12.42 | <0.001 |
| $x^2$ | -0.13 | 0.01 | -12.58 | <0.001 | -0.49 | 0.05 | -10.24 | <0.001 |
| $x^3$ | | | | | 0.003 | 0.0003 | 7.66 | <0.001 |

The following piecewise function structure was created according to the determined knots.

$$S(x) = \begin{cases} b_{00} + b_{01}x + b_{02}x^2 & ,0 \le x \le 21 \\ b_{00} + b_{01}x + b_{02}x^2 + b_{11}(x-21) + b_{12}(x-21)^2 & ,21 < x \le 33 \\ b_{00} + b_{01}x + b_{02}x^2 + b_{11}(x-21) + b_{12}(x-21)^2 + b_{21}(x-33) + b_{22}(x-33)^2 & ,33 < x \le 99 \end{cases}$$

According to the function above, it is the first piece between 0-21 years old, the second piece between 21-33 years old and the third piece between 33-99 years old. A piecewise regression model with restricted fixed coefficients was formed by the "+" functions based on the specified knot as follows:

$$y_i = b_{00} + b_{01}x_i + b_{02}x_i^2 + b_{11}(x_i - 21)_+ + b_{12}(x_i - 21)_+^2 + b_{21}(x_i - 33)_+ + b_{22}(x_i - 33)_+^2 + e_i \qquad i = 1,2,\dots,96$$

The regression equation was estimated by the LSM method. Piecewise regression model was found statistically significant according to the obtained results (F=340; $df_1$=6, $df_2$=89; p<0.001). In addition, all coefficients of the model were found statistically significant (Table 4).

Graphical representation of formed models was given in Figure 2. Marked points are knots of the piecewise regression.

**Table 4.** Parameter estimates in piecewise regression with quadratic+quadratic+quadratic structure

| Variable | b | $s_b$ | t | p |
|---|---|---|---|---|
| Constant | 69.19 | 14.81 | 4.67 | <0.001 |
| x | -15.80 | 3.18 | -4.97 | <0.001 |
| $x^2$ | 1.52 | 0.14 | 10.71 | <0.001 |
| $(x-21)_+$ | -45.35 | 7.37 | -6.15 | <0.001 |
| $(x-21)_+^2$ | -2.65 | 0.37 | -7.27 | <0.001 |
| $(x-33)_+$ | 23.29 | 4.95 | 4.71 | <0.001 |
| $(x-33)_+^2$ | 1.08 | 0.39 | 2.75 | 0.007 |

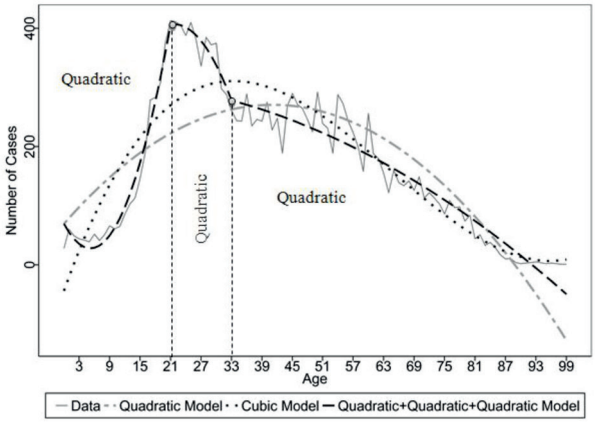


**Figure 2.** Representation of estimated and observed values by regression models of the number of age-related tuberculosis cases

**Table 5.** The $R^2$, MSE, AIC and BIC values of models

| Model | $R^2$ | MSE | AIC | BIC |
|---|---|---|---|---|
| Quadratic | 0.68 | 4569.95 | 1089.45 | 1099.71 |
| Cubic | 0.81 | 2791.50 | 1044.13 | 1056.95 |
| Quadratic+ Quadratic+Quadratic | 0.96 | 604.59 | 903.37 | 923.79 |

*Comparison Of Models For Tuberculosis Data Set:*

The comparison results according to the calculated model selection criteria were given in Table 5.

According to the results, the piecewise regression model with quadratic+quadratic+quadratic structure formed by three partial pieces is more successful than quadratic and cubic regression models in estimating the number of age-related tuberculosis cases (Table 5 and Figure 2).

**Measles Data Set Application Results**

Because the distribution of the number time-varying measles cases in the measles data set is more appropriate for the cubic structure, cubic models were examined in polynomial regressions, and quadratic models were examined in piecewise regressions.

*Model Estimation By Polynomial Regression:*

Created regression equation for estimating the distribution of time-varying measles cases was estimated by the LSM method. A quadratic regression model was statistically significant according to the obtained results (F=13.81; $df_1$=3, $df_2$=42; p=0.002). In addition, all coefficients of the model were statistically significant. (Table 6).

**Table 6.** Parameter estimation for cubic regression

| Variable | b | $s_b$ | t | p |
|---|---|---|---|---|
| Constant | 16 880 000 000 | 6 048 000 000 | 2.79 | 0.008 |
| x | -25 420 000 | 9 107 000 | -2.79 | 0.008 |
| $x^2$ | 12 760 | 4571 | 2.79 | 0.008 |
| $x^3$ | -2.14 | 0.77 | -2.79 | 0.008 |

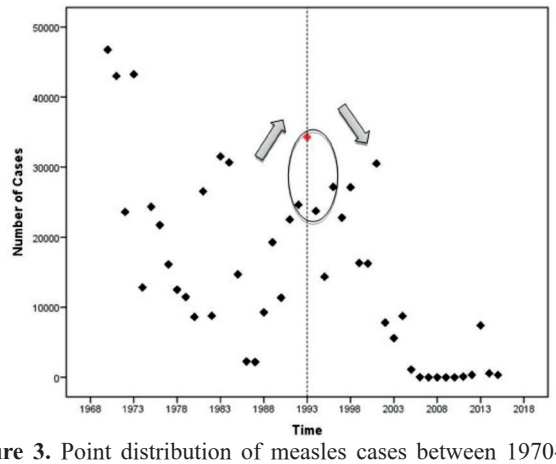


**Figure 3.** Point distribution of measles cases between 1970-2015 years

*Model Estimation By Piecewise Regressions:*

Distribution of measles cases from 1970 to 2015 was given in Figure 3. The knot was determined according to jump or sudden change points. As shown in Figure 3, the number of cases of measles is fluctuating in some regions. So each of the points in these regions is a candidate knot. For this reason, it was decided to model with one knot and piecewise function

with two pieces and year 1993 was determined to be the best knot for applying piecewise regression as a result of the trials.

According to the determined knot, the following piecewise function structure was formed:

$$S(x) = \begin{cases} b_{00} + b_{01}x + b_{02}x^2 & ,1970 \leq x \leq 1993 \\ b_{00} + b_{01}x + b_{02}x^2 + b_{11}(x-1993)_+ + b_{12}(x-1993)_+^2 & ,1993 < x \leq 2015 \end{cases}$$

According to the function above, it is the first piece between the 1970-1993 years; the second piece is between the 1993-2015 years. A piecewise regression model with restricted fixed coefficients were created by "+" functions based on the specified knot as follows:

$$y_i = b_{00} + b_{01}x_i + b_{02}x_i^2 + b_{11}(x_i - 1993)_+ + b_{12}(x_i - 1993)_+^2 + e_i \qquad i = 1,2,\dots,46$$

The regression equation was estimated by the LSM method. Piecewise regression model was found statistically significant according to the obtained results ($F=18.63$; $df_1=4$, $df_2=41$; $p<0.001$). In addition, all coefficients of the model were found as statistically significant (Table 7). Graphical representation of formed models was given in Figure 4. Marked points are knots of the piecewise regression.

**Table 7.** Piecewise regression with quadratic+quadratic structure parameter estimates

| Variable | b | $s_b$ | t | p |
|---|---|---|---|---|
| Constant | 680 600 000 | 135 600 000 | 5.02 | <0.001 |
| x | -686 300 | 136 800 | -5.02 | <0.001 |
| $x^2$ | 173 | 34.52 | 5.01 | <0.001 |
| $(x-1993)_+$ | -5757 | 1431 | -4.02 | <0.001 |
| $(x-1993)_+^2$ | -125 | 40.56 | -3.08 | 0.004 |

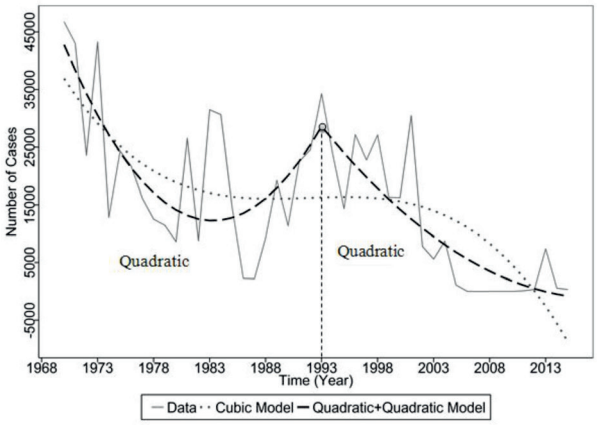


**Figure 4.** Estimated and observed values of measles cases between 1970-2015 by the regression models

### *Comparison Of Regression Models For Measles Data Set:*

The comparison results according to the calculated model selection criteria were given in Table 8.

**Table 8.** $R^2$, MSE, AIC and BIC values of generated regression models

| Model | $R^2$ | MSE | AIC | BIC |
|---|---|---|---|---|
| Cubic | 0.50 | 82 550 036 | 979.07 | 988.22 |
| Quadratic+Quadratic | 0.65 | 58 190 618 | 964.99 | 975.96 |

According to the results, the piecewise regression model with quadratic+quadratic structure that formed by two partial pieces is more successful than cubic regression model about estimating the number of time-varying measles cases (Table 8 and Figure 4).

## Discussion

Although piecewise regressions are used in many areas, their use in the field of health is not common yet. Polynomial regressions were not as popular as piecewise regressions because they provided more straightforward analysis than piecewise regressions and were included more often in statistical package programs. This led researchers to use polynomial regressions. However, the polynomials that form piecewise regressions can easily be combined in computer programs and provide the researcher with the required ease for analysis.

In reviewing the literature, some studies using piecewise regressions are remarkable. Hurley et al. conducted a simulation work with 2000 runs with five data sets of different structures and formed in order to compare the performances of the piecewise regressions and simple regressions using one dependent variable and one independent variable derived from the distribution of x~U(1,100) consisting of 201 units (31). Three of the data sets had a quadratic structure, and the remaining two data sets had a cubic structure. They constructed six regression models for each data structure: linear regression, polynomial regression (cubic and quadratic), and piecewise regression (linear, quadratic and cubic). For each piecewise regression model, they defined the points x=32 and x=68 points to be fixed knots, and formed linear, polynomial and piecewise regression models according to the LSM method conforming to the data structure in order to estimate the data sets they derived. They reported that piecewise regression models had higher $R^2$ values and lower MSE values. Mulla reported that the use of piecewise regression modeling technique would provide significant benefit to the researcher in clinical trials, especially in studies on the dose and response of the drug given to the patient (32). For the said study, it was used records from 117 patients who were given serum albumin ranging from 1.1 to 5.1 g/100 mL. it was used the cubic piecewise regression model and a classical model, both created with the LSM method, and identified a knot determined by visual inspection for the cubic piecewise regression model. According to the results, it was reported that the cubic piecewise regression model predicted the whole of the blood concentration values of 60 patients with real-like accuracy; whereas the classical model only predicted the values of 25 patients with real-like accuracy, and the results of the remaining 35 patients contained considerably more significant differences than the real concentration values which gathered around the knot determined for piecewise regression. In our study, polynomial and piecewise regression models were created by using simulation data and real data sets, and the performances of these models were compared. In the simulation application, fixed knots determined according to the breakpoints created in the data production phase were

used. The number of knots for all piecewise regression models were set to one. Changeable knots were used in applications incluiding real data sets. Knot numbers were determined to be two in the study with the tuberculosis data set and one in the study with the measles data set. In determining the changeable knots, first the number of knots to be used was decided after visual inspection, and the candidate knot areas were identified. After that, the points that provided the highest performance after trials were chosen as knots. If none of the points tested as knots in an area designated as the candidate area had a significant contribution to the model's estimation strength, then no selection of knots was made from that area. According to the results of the applications performed by using both the simulation datasets and real datasets, it was seen that the performances of the piecewise regressions are better than polynomial regressions with higher $R^2$ value and lower MSE, AIC, BIC values.

Determination of appropriate knots is crucial for the estimations with high performance. Parkhurst et al. used regression models with both fixed and variable knots (33). They found that all the variable knot models have higher prediction power than the fixed knot models in the same structure. Seber and Wild and Eubank found that the use of unnecessarily high-degree polynomials for point intervals that are formed according to the most appropriate knots didn't lead to a significant increase in $R^2$, but also caused an excess of parameters and loss of degree of freedom (4, 18). Wold pointed out some details about the determination of the number and location of knots and reported that each interval forming a piecewise regression must contain at least 4-5 observation points and thus the number of knots should be chosen as few as possible (19). Although the point distribution of the measles data set used in our study is more appropriate for the cubic structure, it is predicted by excellent performance through the use of the piecewise regression model with quadratic+quadratic structure formed according to optimal knot. Working with low-degree polynomials is desirable in terms of providing the process if it doesn't lead to the need for an increase in the number of knots. In our study, it was observed that the position of the knot is closely related to the shape of the distribution and optimal degree polynomials were used for sub-intervals formed by using as few knots as possible. Also, it has been found that the contribution of the model to the performance is close to each other if any point in the candidate knot determined by visual inspection is selected as a knot. Firstly the number of knots should be determined, and then the candidate areas to select the knots should be decided. Afterwards, the position of the knots should be identified and the knots providing the best prediction strength should be selected by trials from candidate regions.

## Conclusion

Although researches in the field of health mostly prefer polynomial regression methods, this study showed through applications that data analysis by piecewise regressions with optimal knot provides statistical superiority. The future studies should consider the piecewise regression method as a powerful alternative for all data sets where the relationship between the dependent and independent variables would be examined. Furthermore, the use of piecewise regression should be extended for estimations with higher performance in health-related researches.

Presented as an oral presantation at the "4th International Researchers, Statisticians and Young Statisticians Congress (IRSYSC)" on April 28-30, 2018.

## References

1. Freedman DA. Statistical models: theory and practice. Cambridge University Press. New York, 2009; 41-60.

2. Freund RJ, Wilson WJ, Sa P. Regression analysis. Academic Press. 2nd ed. New York, 2006; 270-95.

3. Hartley HO, Booker A. Nonlinear least-squares estimation. The Annals of mathematical statistics 1965; 36(2): 638-50.

4. Seber G, Wild C. Nonlinear regression. Hoboken: John Wiley & Sons Google Scholar, 2003, 325-65.

5. Park SH. Experimental designs for fitting segmented polynomial regression models. Technometrics 1978; 20(2): 151-4.

6. Wainer H. Piecewise regression: A simplified procedure. British Journal of Mathematical and Statistical Psychology 1971; 24(1): 83-92.

7. Eubank R. Approximate regression models and splines. Communications in Statistics-Theory and Methods 1984; 13(4): 433-84.

8. Gallant AR, Fuller WA. Fitting segmented polynomial regression models whose join points have to be estimated. Journal of the American Statistical Association 1973; 68(341): 144-7.

9. Berberoglu B, Berberoglu CN. Modeling the Structural Shifts in Real Exchange Rate with Cubic Spline Regression (CSR). Turkey 1987-2008. International Journal of Business and Social Science 2011; 2(17).

10. De Boor C, Rice JR. Least squares cubic spline approximation, II-variable knots. West Lafayette, Purdue University, 1968; 4-13.

11. Poirier DJ. Piecewise regression using cubic splines. Journal of the American Statistical Association 1973; 68(343): 515-24.

12. Porth RW. Application of least square cubic splines to the analysis of edges [The Master of Science Degree Thesis]. New York , Rochester Institute of Technology, 1984; 23-51.

13. Schwetlick H, Schütze T. Least squares approximation by splines with free knots. BIT Numerical mathematics 1995; 35(3): 361-84.

14. Draper NR, Smith H. Applied regression analysis. John Wiley & Sons. 3rd ed. Vol 326. New York, 2014; 158-75.

15. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Statistics in medicine 1984; 3(2): 143-52.

16. Chan S-h. Polynomial spline regression with unknown knots and AR (1) errors [Doctoral Thesis]. Columbus, The Ohio State University, 1989; 22-38.

17. De Boor C, Rice JR. Least squares cubic spline approximation I-Fixed knots. West Lafayette, Purdue University, 1968; 2-19.

18. Eubank RL. Nonparametric regression and spline smoothing. CRC press. 2nd ed. Vol 157. New York, 1999; 227-308.

19. Wold S. Spline functions in data analysis. Technometrics 1974; 16(1): 1-11.

20. Hawkins DM. On the choice of segments in piecewise approximation. IMA Journal of Applied Mathematics 1972; 9(2): 250-6.

21. Ruppert D. Selecting the number of knots for penalized splines. Journal of computational and graphical statistics 2002; 11(4): 735-57.

22. Agarwal GG, Studden W. An algorithm for selection of design and knots in the response curve estimation by spline functions. West Lafayette, Purdue University Department of Statistics, 1978; 78-85.

23. Marsh LC, Cormier DR. Spline regression models. Sage. London, 2001; 7-58.

24. Powell M. The local dependence of least squares cubic splines. SIAM Journal on Numerical Analysis 1969; 6(3): 398-413.

25. Smith PL. Splines as a useful and convenient statistical tool. The American Statistician 1979; 33(2): 57-62.

26. Wegman EJ, Wright IW. Splines in statistics. Journal of the American Statistical Association 1983; 78(382): 351-65.

27. Genç A, Oktay E, Alkan Ö. İhracatın İthalatı Karşılama Oranlarının Parçalı Regresyonlarla Modellenmesi. Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi 2012; 16(1): 497-511.

28. Markov D. Information content in stock market technical patterns: A spline regression approach [Doctoral Thesis]. South Bend, University of Notre Dame, 2003; 58-72.

29. Marsh LC. Estimating the number and location of knots in spline regressions. Journal of Applied Business Research 1986; 2(2): 60-70.

30. Studden WJ, VanArman D. Admissible designs for polynomial spline regression. The Annals of Mathematical Statistics 1969; 40(5): 1557-69.

31. Hurley D, Hussey J, McKeown R, Addy C, editors. An evaluation of splines in linear regression. The 132nd Annual Meeting; 2004 Nov 6-10; Washington.

32. Mulla Z. Spline regression in clinical research. West indian medical journal 2007; 56(1): 77-9.

33. Parkhurst A, Spiers D, Hahn G. Spline models for estimating heat stress thresholds in cattle. Conference on Applied Statistics in Agriculture: 14th Annual Conference Proceedings. 2002, 137-48; New York.