# THE CROSS-VALIDATION METHOD IN THE POLYNOMIAL REGRESSION

**by**
**Nicoleta Breaz**

**Abstract.** One of the methods used for the degree selection in the polynomial regression is the cross-validation method(CV). In this paper, we implement a CV-based algorithm, in Matlab 6.5 medium and we apply it on some test functions. Then we analyze the results by performing an ordinary regression analysis. Finally, we propose a new algorithm that combines the CV method with the classical degree selection.

## 1.Introduction

We consider the regression model,

$$Y = f(X) + \varepsilon,$$

with *X*, *Y*, two random variables and $\varepsilon$, the error term. After a sampling, we obtain the observational model,

$$y_i = f(x_i) + \varepsilon_i, \quad i = \overline{1,n}$$

and we suppose that $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)' \sim N(0, \sigma^2 I)$.

One of the most used regression models is the polynomial regression, that is

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + ... + \alpha_q x_i^q + \varepsilon_i .$$

It is well known that such a model can be estimate with least squares method, after a reduction to a multiple linear model, with the explicative variables, $X, X^2, ..., X^q$. But, before make this estimation, it is necessary to establish the form of the regression function, or more precisely in this case, the polynom's degree.

An ordinary solution to this problem is to estimate the model, for different values of $q$ and then, to compare these models, by performing a regression analysis.

As an alternative, there exist some data-based selection methods that give the appropriate value for *q*. One of such method is the cross-validation method(CV).

## 2.Degree selection based on the regression analysis

The regression analysis for a fitted model can be made, either by graphical comparison or quantitative methods.

For the graphical comparison of two or more models, obtained for different values of *q*, it is necessary to plot the fitted curves versus the data and also, the residuals, $e_i = y_i - (a_0 + a_1 x_i + ... + a_q x_i^q), \quad i = \overline{1,n}$, with $a_i, i = \overline{1,n}$, the least squares estimators.

Obviously, we choose the model that is more closely to data and which has the residuals curve as an white noise. For a good model, the residuals need to be randomly scattered around zero.

Another graphical method is to plot the fitted curves together with the prediction bounds. If the prediction interval is too wide, we need to have caution about using such a model for prediction.

Also, as a quantitative comparison method, we can analyze the accuracy of the estimation, if we look at the confidence bounds for estimated coefficients. We cannot trust in a model, for that the coefficients have wide confidence intervals.

Another quantitative method consists in the comparison of some usual regression statistics as squared-R, $\mathrm{R}^2$, adjusted squared-R, $\overline{R}^2$, root mean square error, *s* and sum of squared residuals, $S_R^2$. A good model will have small values for $S_R^2$ and *s*, respectively, values closed to one, for $R^2$ and $\overline{R}^2$. Anyway, in the polynomial regression, is preferred $\overline{R}^2$ instead of $R^2$, because $R^2$ depends on the number of explicative variables that occur linearly in the model.

**3.Degree selection based on the CV method**

A natural way to select the polynom's degree, *q*, based on data information, is to minimize the expected prediction error,

$$PSE(q) = E\left(y' - f_q(x')\right)^2,$$

where $x', y'$ are new data and $f_q$ is the fitted polynom of $q$ degree.

Since additional data are not usually available, we can use just an estimator of $PSE(q)$. One of such estimator is the (leaving-out-one) cross-validation function, given by

$$CV(q) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f_q^{(-i)}(x_i)\right)^2,$$

where $f_q^{(-i)}$ is the regression polynom, fitted from all data, less the *i*-th data.

A leaving-out-one resampling method is used here. We obtain the fitted models, $f_q^{(-i)}, i = \overline{1,n}$, from $n$ learning samples(each one, with $n-1$ data), then we validate these models by other $n$ test samples, formed with one-leaving-out data.

According to (1), the cross validation function is equal to $n \cdot PRESS$, where $PRESS$ is a prediction power measure for the model. Small values for $PRESS$ give models with large prediction power. So, by minimizing the CV function in respect with *q*, we obtain the appropriate degree for the polynomial model.

**4.Numerical experiments**

For computational aspects, we implement in Matlab medium, the next algorithm, based on the CV method:

**Algorithm 1**

*Step1.* Read the sample data $(x_i, y_i), i = \overline{1,n}$ and if is necessary, order and weight the data, in respect with data sites, $x_i$.

*Step 2.* For each $i, i = \overline{1,n}$, determine the fitted polynom of *q* degree, $f_q^{(-i)}$, based on the leaving-out-one resampling.

*Step 3.* Calculate the value of $CV(q)$ function.

*STOP.*

In order to obtain *q*, for which $CV(q)$ is minimum, the following adequate step must be added:

*Step 4.* Calculate $CV(q)$ for integer and strictly positives different values of *q*.

The appropriate value of *q* is $q_{CV}$, with

$$CV(q_{CV}) = \min_q CV(q).$$

In this paper, we search *q* from 1 to 7.

In order to see how it works the *CV* selection, we first consider a fourth degree polynom as a test function. The goal of the *CV* method will be to reconstitute the degree of the polynom from noisy data, obtained by the test function and the random number generator.

Let be the test function

$$f_1(x) = 6x^4 + 3x^3 + 7x^2 - 2x + 5$$

and for the beginning, the sample of exact data, $(x_i, y_i)$, where $x_i = \frac{i}{100}$ and $y_i = f_1(x_i),\ i = \overline{1,100}$.

The implementation of the algorithm 1, in Matlab medium, gives us the following values for $CV(q)$:

| $q$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $CV(q)$ | 2,92357 | 0,09029 | 0,00093 | 0 | 0 | 0 | 0 |

We mention that the zeros are in fact, some values of $1 \cdot 10^{-29}$, magnitude order, so are insignificantly different from zero.

Now we use noisy data, namely $(x_i, y_i)$, with $x_i = \frac{i}{100}$ and $y_i = f_1(x_i) + \varepsilon_i$, $i = \overline{1,100}$, where $\varepsilon_i$, $i = \overline{1,100}$, come from a random number generator simulating independently and identically distributed, $N(0;0,1)$, random variables.

If we set to 0 the seed of the random number generator, after applying the algorithm 1, we obtain the value $q_{CV} = 4$.

Otherwise, if we repeat the simulation for 100 replicates, with distinct seeds, we obtain an average of $q_{CV}$ equal to 4,37. But, tacking into account the possible values for *q*, the average is not very representative. So, we look at the distribution of the values of $q_{CV}$, in order to retain the most frequently case.

For the same 100 replicates of average 4,37, we obtain the distribution

$$q_{CV} : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 0 & 8 & 63 & 18 & 6 & 5 \end{pmatrix}.$$

Now, we can conclude that, $q_{CV} = 4$ is the optimal value since it occurs in 63% of the cases.

Consequently, the *CV* method recognizes the degree of the test function.

However, if the data are very noisy, the comparison of $q_{CV}$ with the real *q* from the test function will not be relevant, anymore. So, for the validation of the results obtained by the *CV* method, it is necessary to perform an ordinary regression analysis.

After we perform this analysis in the case of the mentioned data, we obtain that the case $q = 4$ has a small advantage, from statistics comparison point of view and $q = 3$ is recommended by the accuracy of confidence bounds. Consequently, the *CV* method can be viewed as a selection method between two appropriate values indicated by the regression analysis. Anyway, the *CV* method selects one of the most recommended cases by the regression analysis and does this, in a more simple manner, with less time.

Next, we will consider another test function that is not a polynomial one so we validate the *CV* method just in the regression analysis and not by comparison with the real degree.

Let be $f_2(x) = 5^x + 3e^{-2x}$ and the data $(x_i, y_i)$, with $x_i = \frac{i}{100}$ and $y_i = f_2(x_i) + \varepsilon_i$, $\varepsilon_i \sim N(0;0,1)$, $i = \overline{1,100}$.

If we set the seed of random number generator at 0, we obtain by applying algorithm 1, the value $q_{CV} = 2$. Else, if we repeat the algorithm for 100 replicates, with different seeds, we obtain the distribution

$$q_{CV} : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 74 & 9 & 7 & 3 & 3 & 4 \end{pmatrix},$$

with average 2,64.

So, after a quick view on the distribution, we retain the value $q_{CV} = 2$, as the appropriate fitting polynomial degree.

On the other side, we make the comparative analysis for polynomial fittings, with $q = \overline{1,7}$.

For simplicity, we plot in the following figure just the first degree fitting polynomial and the second degree fitting polynomial, versus the data. But we need to mention that in a complete plot, with all seven fitting curves, the curves for $q = \overline{3,7}$ are not too different from the curve, $q = 2$.

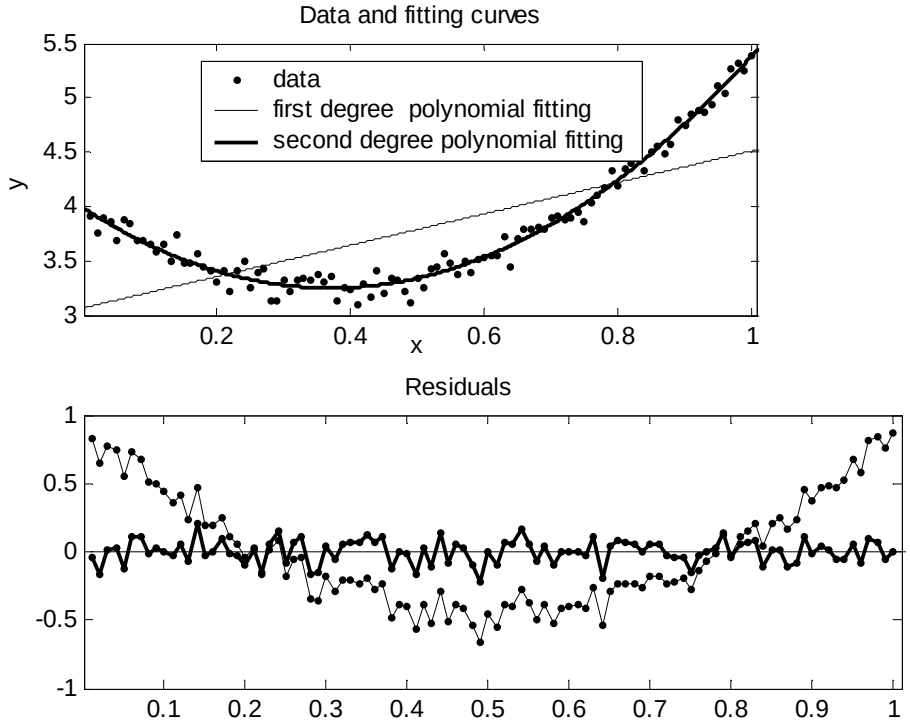


**Fig. 1**

We can see that we must eliminate the case $q=1$, since it is not close enough to data and its residuals present some trend, consequently, this case doesn't offer a satisfactory fitting.

For more deep analysis, in the following plot, we look at the 95%-prediction bounds of the fitting curves. Again, for simplicity, we plot just the second degree fitting polynomial and the sixth degree fitting polynomial.

In this plot, the solid curve, together with the dashed curves, correspond to the case $q=2$ and the remaining curves are for the case $q=6$. The cases $q=3$ and $q=4$ are likewise to $q=2$ and the cases $q=5$ and $q=7$ are likewise to $q=6$.

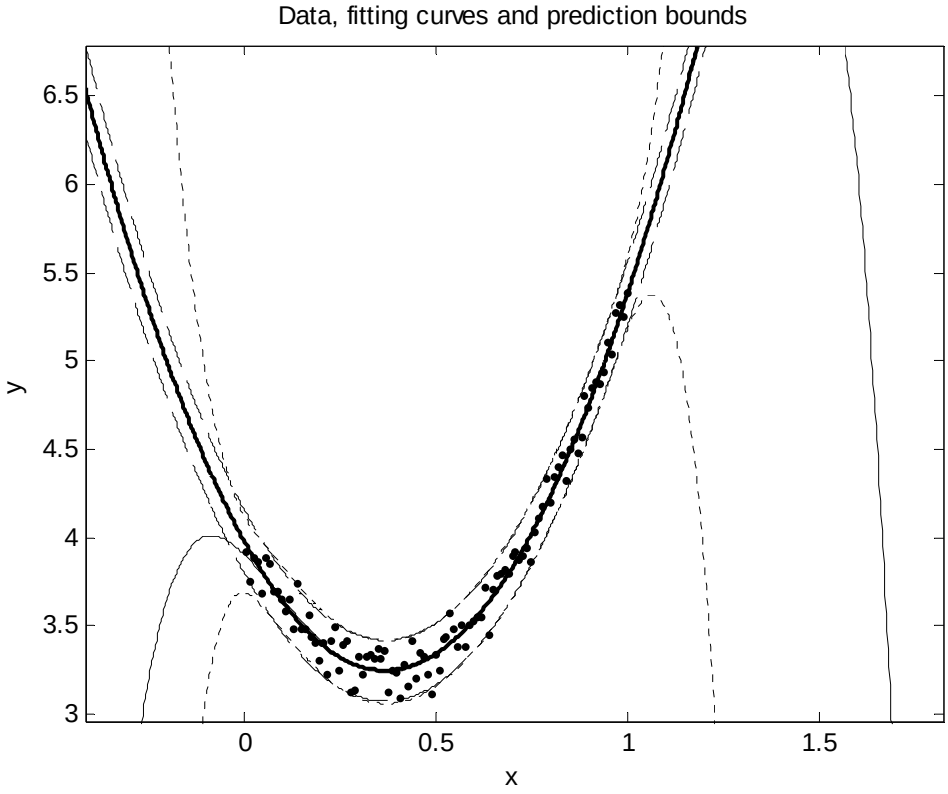


**Fig. 2**

We observe that, out of the data range, the prediction interval for $q=6$ is too wide, so we cannot trust in the prediction on the sixth degree polynomial fitting and we obtain the same conclusion, for $q=5$ and $q=7$.

Consequently, just the cases $q=2$, $q=3$ and $q=4$ remain in the competition. For these cases, we extend the plot interval, in order to compare the width of the prediction intervals.

The following plot contains the cases $q=2$ and $q=4$ and we observe that $q=4$ has more wide prediction interval, than $q=2$. Again, for $q=2$, we have the solid curve, together with the dashed curves.

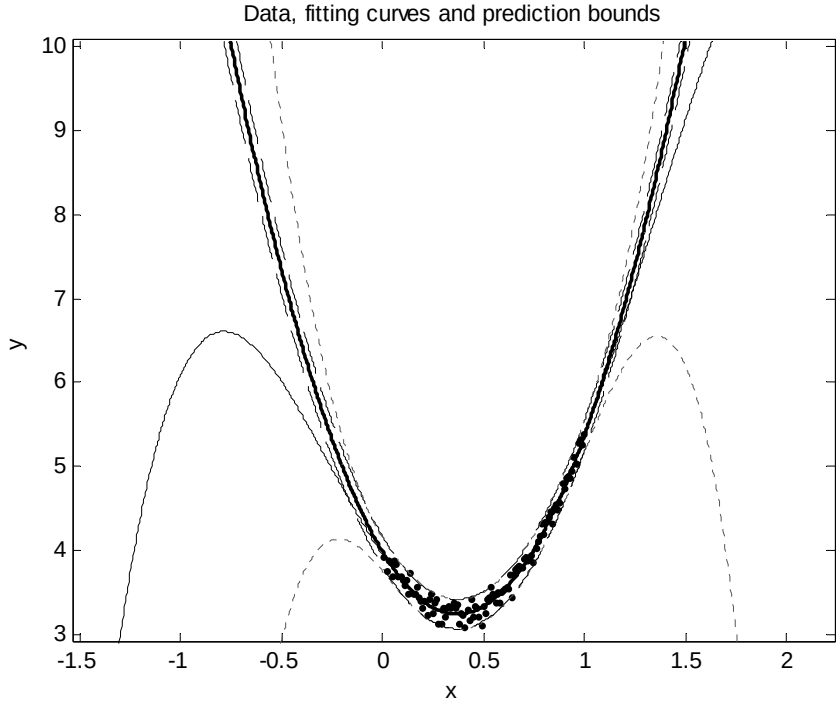


**Fig. 3**

After a comparison between $q=2$ and $q=3$, we obtain the same conclusion: from this point of view, the case $q=2$ is more appropriate.

Anyway, for these last three cases, we make also a quantitative comparison and we obtain the following values for the regression statistics:

| Statistics / Degree | $S_R^2$ | $\overline{R}^2$ | $s$ |
|---|---|---|---|
| 2 | 0,7318 | 0,9781 | 0,0869 |
| 3 | 0,7308 | 0,9779 | 0,0872 |
| 4 | 0,7261 | 0,9778 | 0,0874 |

Since the case $q = 2$ is recommended by $\overline{R}^2$ and *s*, respectively the case $q = 4$ is recommended by $S_R^2$, once again the balance is favorably for $q = 2$.

Also, the following 95%-confidence bounds for the coefficients indicate more accuracy in estimation, for the case $q = 2$.

Second-degree polynomial coefficients and confidence bounds:

$a_2 = 5{,}38 \ (5{,}148; 5{,}611)$,
$a_1 = -3{,}99 \ (-4{,}231; -3{,}749)$,
$a_0 = 3{,}989 \ (3{,}936; 4{,}042)$.

Third-degree polynomial coefficients and confidence bounds:

$a_3 = 0{,}1701 \ (-0{,}7469; 1{,}087)$,
$a_2 = 5{,}122 \ (3{,}713; 6{,}53)$,
$a_1 = -3{,}885 \ (-4{,}499; -3{,}271)$,
$a_0 = 3{,}98 \ (3{,}908; 4{,}052)$.

Fourth-degree polynomial coefficients and confidence bounds:

$a_4 = -1{,}446 \ (-5{,}096; 2{,}204)$,
$a_3 = 3{,}091 \ (-4{,}34; 10{,}52)$,
$a_2 = 3{,}219 \ (-1{,}788; 8{,}226)$,
$a_1 = -3{,}453 \ (-4{,}706; -2{,}201)$,
$a_0 = 3{,}957 \ (3{,}865; 4{,}049)$.

After all these analyses, we conclude that the most appropriate polynomial fitting, for the data $(x_i, y_i)$, coincides with the *CV*-case, $q = 2$.

The following plot contains the test function, the data and the second degree fitting polynomial.

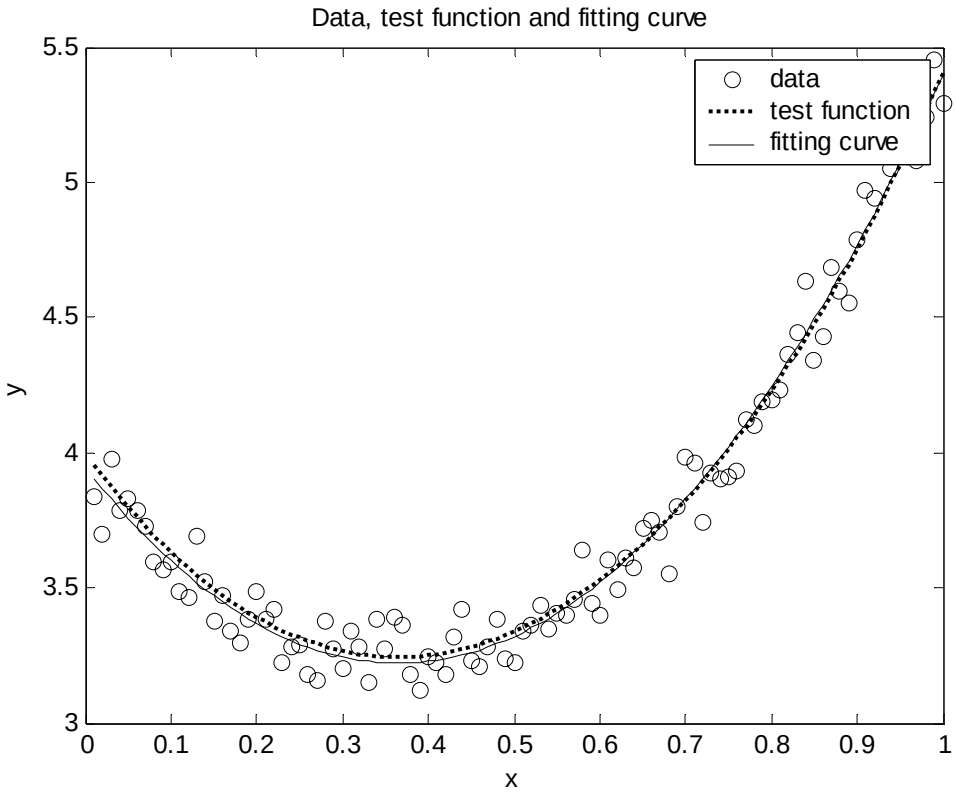


**Fig. 4**

After these numerical experiments, we can state that the *CV*-method works enough well, in the degree selection. Anyway, for an optimal fitting, it is necessary to use, not just a single method, but more and then, the most appropriate fitting will be that, with more recommendations.

With these considerations in mind, we propose a composed algorithm for degree selection, that is cheaper than a full regression analysis and in the same time, is more precisely than algorithm 1.

**Algorithm 2**

*Step 1.* Find a sample with *p* replicates values of $q_{CV}$ and the related distribution.

*Step 2.* Retain the mode of the distribution, $q_{CV}^1$ and also, the mode of the remaining values, $q_{CV}^2$.

*Step 3.* If the fitting with polynomial of order $q_{CV}^1$ is validated by both, graphical and quantitative regression tests, *STOP*.

Else, follow the next step.

*Step 4.* Make the comparative analysis for the cases $q_{CV}^1$, $q_{CV}^2$, $q_{CV}^1 - 1$, $q_{CV}^1 + 1$ and establish the optimal fitting.
*STOP*.

Obviously, for nonsimulated data, the distribution of replicates isn't exist, so at the step 4, we compare just the case $q_{CV}$, with $q_{CV} - 1$ and $q_{CV} + 1$.

**References**

1.Eubank R. L. - Nonparametric Regression and Spline Smoothing-Second Edition, Marcel Dekker, Inc., New York , Basel, 1999
2.Saporta G.- Probabilites, Analyse des Donnes et Statistique. Editions Techniq, Paris, 1990
3.Stapleton J.H.- Linear Statistical Models. A Willey-Interscience Publications, Series in Probability and Statistics, New York, 1995
4.Tassi Ph.- Methodes statistiques, 2$^e$ edition, Economica, Paris, 1989

**Author:**
Nicoleta Breaz, „1 Decembrie 1918" University of Alba Iulia, Romania,
nbreaz@uab.ro