WIKIPEDIA

Polynomial regression

In statistics, **polynomial regression** is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted $E(y|x)$. Although *polynomial regression* fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function $E(y | x)$ is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression.

The explanatory (independent) variables resulting from the polynomial expansion of the "baseline" variables are known as higher-degree terms. Such variables are also used in classification settings.[1]

Contents

History

Polynomial regression models are usually fit using the method of least squares. The least-squares method minimizes the variance of the unbiased estimators of the coefficients, under the conditions of the Gauss–Markov theorem. The least-squares method was published in 1805 by Legendre and in 1809 by Gauss. The first design of an experiment for polynomial regression appeared in an 1815 paper of Gergonne.[2][3] In the twentieth century, polynomial regression played an important role in the development of regression analysis, with a greater emphasis on issues of design and inference.[4] More recently, the use of polynomial models has been complemented by other methods, with non-polynomial models having advantages for some classes of problems.

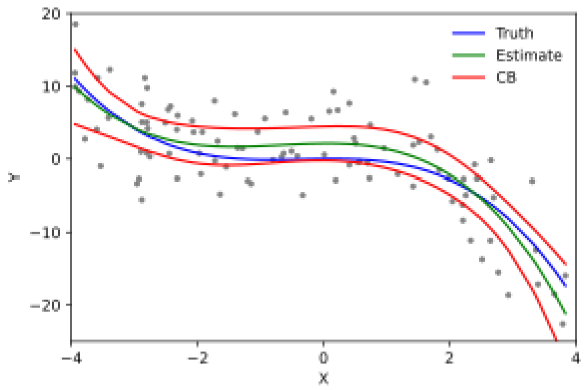
Definition and example

The goal of regression analysis is to model the expected value of a dependent variable y in terms of the value of an independent variable (or vector of independent variables) x. In simple linear regression, the model

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

is used, where ε is an unobserved random error with mean zero conditioned on a scalar variable *x*. In this model, for each unit increase in the value of *x*, the conditional expectation of *y* increases by β_1 units.

In many settings, such a linear relationship may not hold. For example, if we are modeling the yield of a chemical synthesis in terms of the temperature at which the synthesis takes place, we may find that the yield improves by increasing amounts for each unit increase in temperature. In this case, we might propose a quadratic model of the form



A cubic polynomial regression fit to a simulated data set. The confidence band is a 95% simultaneous confidence band constructed using the Scheffé approach.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.$$

In this model, when the temperature is increased from *x* to *x* + 1 units, the expected yield changes by $\beta_1 + \beta_2(2x + 1)$. (This can be seen by replacing *x* in this equation with *x*+1 and subtracting the equation in *x* from the equation in *x*+1.) For infinitesimal changes in *x*, the effect on *y* is given by the total derivative with respect to *x*: $\beta_1 + 2\beta_2 x$. The fact that the change in yield depends on *x* is what makes the relationship between *x* and *y* nonlinear even though the model is linear in the parameters to be estimated.

In general, we can model the expected value of *y* as an *n*th degree polynomial, yielding the general polynomial regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon.$$

Conveniently, these models are all linear from the point of view of estimation, since the regression function is linear in terms of the unknown parameters β_0, β_1, ... Therefore, for least squares analysis, the computational and inferential problems of polynomial regression can be completely addressed using the techniques of multiple regression. This is done by treating *x*, x^2, ... as being distinct independent variables in a multiple regression model.

Matrix form and calculation of estimates

The polynomial regression model

$$y_i \;=\; \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \varepsilon_i \ (i = 1, 2, \dots, n)$$

can be expressed in matrix form in terms of a design matrix $\mathbf{X}$, a response vector $\vec{y}$, a parameter vector $\vec{\beta}$, and a vector $\vec{\varepsilon}$ of random errors. The *i*-th row of $\mathbf{X}$ and $\vec{y}$ will contain the *x* and *y* value for the *i*-th data sample. Then the model can be written as a system of linear equations:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ 1 & x_3 & x_3^2 & \dots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

which when using pure matrix notation is written as

$$\vec{y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}.$$

The vector of estimated polynomial regression coefficients (using ordinary least squares estimation) is

$$\widehat{\vec{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\ \mathbf{X}^\mathsf{T}\vec{y},$$

assuming $m < n$ which is required for the matrix to be invertible; then since $\mathbf{X}$ is a Vandermonde matrix, the invertibility condition is guaranteed to hold if all the x_i values are distinct. This is the unique least-squares solution.

Interpretation

Although polynomial regression is technically a special case of multiple linear regression, the interpretation of a fitted polynomial regression model requires a somewhat different perspective. It is often difficult to interpret the individual coefficients in a polynomial regression fit, since the underlying monomials can be highly correlated. For example, x and x^2 have correlation around 0.97 when x is uniformly distributed on the interval (0, 1). Although the correlation can be reduced by using orthogonal polynomials, it is generally more informative to consider the fitted regression function as a whole. Point-wise or simultaneous confidence bands can then be used to provide a sense of the uncertainty in the estimate of the regression function.

Alternative approaches

Polynomial regression is one example of regression analysis using basis functions to model a functional relationship between two quantities. More specifically, it replaces $x \in \mathbb{R}^{d_x}$ in linear regression with polynomial basis $\varphi(x) \in \mathbb{R}^{d_\varphi}$, e.g. $[1, x] \xrightarrow{\varphi} [1, x, x^2, \ldots, x^d]$. A drawback of polynomial bases is that the basis functions are "non-local", meaning that the fitted value of y at a given value $x = x_0$ depends strongly on data values with x far from x_0.[5] In modern statistics, polynomial basis-functions are used along with new basis functions, such as splines, radial basis functions, and wavelets. These families of basis functions offer a more parsimonious fit for many types of data.

The goal of polynomial regression is to model a non-linear relationship between the independent and dependent variables (technically, between the independent variable and the conditional mean of the dependent variable). This is similar to the goal of nonparametric regression, which aims to capture non-linear regression relationships. Therefore, non-parametric regression approaches such as smoothing can be useful alternatives to polynomial regression. Some of these methods make use of a localized form of classical polynomial regression.[6] An advantage of traditional polynomial regression is that the inferential framework of multiple regression can be used (this also holds when using other families of basis functions such as splines).

A final alternative is to use kernelized models such as support vector regression with a polynomial kernel.

If residuals have unequal variance, a weighted least squares estimator may be used to account for that.[7]

See also

- Curve fitting
- Line regression
- Local polynomial regression
- Polynomial and rational function modeling
- Polynomial interpolation
- Response surface methodology
- Smoothing spline

Notes

- Microsoft Excel makes use of polynomial regression when fitting a trendline to data points on an X Y scatter plot.[8]

References

1. Yin-Wen Chang; Cho-Jui Hsieh; Kai-Wei Chang; Michael Ringgaard; Chih-Jen Lin (2010). "Training and testing low-degree polynomial data mappings via linear SVM" (http://jmlr.csail.mit.edu/papers/v11/chang10a.html). *Journal of Machine Learning Research*. **11**: 1471–1490.
2. Gergonne, J. D. (November 1974) [1815]. "The application of the method of least squares to the interpolation of sequences" (https://doi.org/10.1016%2F0315-0860%2874%2990034-2). *Historia Mathematica* (Translated by Ralph St. John and S. M. Stigler from the 1815 French ed.). **1** (4): 439–447. doi:10.1016/0315-0860(74)90034-2 (https://doi.org/10.1016%2F0315-0860%2874%2990034-2).
3. Stigler, Stephen M. (November 1974). "Gergonne's 1815 paper on the design and analysis of polynomial regression experiments" (https://doi.org/10.1016%2F0315-0860%2874%2990033-0). *Historia Mathematica*. **1** (4): 431–439. doi:10.1016/0315-0860(74)90033-0 (https://doi.org/10.1016%2F0315-0860%2874%2990033-0).
4. Smith, Kirstine (http://www.webdoe.cc/publications/kirstine.php) (1918). "On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial Function and its Constants and the Guidance They Give Towards a Proper Choice of the Distribution of the Observations" (https://zenodo.org/record/1431591). *Biometrika*. **12** (1/2): 1–85. doi:10.2307/2331929 (https://doi.org/10.2307%2F2331929). JSTOR 2331929 (https://www.jstor.org/stable/2331929).
5. Such "non-local" behavior is a property of analytic functions that are not constant (everywhere). Such "non-local" behavior has been widely discussed in statistics:
    - Magee, Lonnie (1998). "Nonlocal Behavior in Polynomial Regressions". *The American Statistician*. **52** (1): 20–22. doi:10.2307/2685560 (https://doi.org/10.2307%2F2685560). JSTOR 2685560 (https://www.jstor.org/stable/2685560).
6. Fan, Jianqing (1996). *Local Polynomial Modelling and Its Applications: From linear regression to nonlinear regression*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC. ISBN 978-0-412-98321-4.

7. Conte, S.D.; De Boor, C. (2018). *Elementary Numerical Analysis: An Algorithmic Approach* (https://books.google.com/books?id=oYZXDwAAQBAJ&pg=PA259). Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104). p. 259. ISBN 978-1-61197-520-8. Retrieved 2020-08-28.
8. Stevenson, Christopher. "Tutorial: Polynomial Regression in Excel" (https://facultystaff.richmond.edu/~cstevens/301/Excel4.html). *facultystaff.richmond.edu*. Retrieved 22 January 2017.

External links

- Curve Fitting (https://phet.colorado.edu/en/simulation/curve-fitting), PhET Interactive simulations, University of Colorado at Boulder