STA3010 Regression Analysis

Feng YIN

The Chinese University of Hong Kong (Shenzhen)

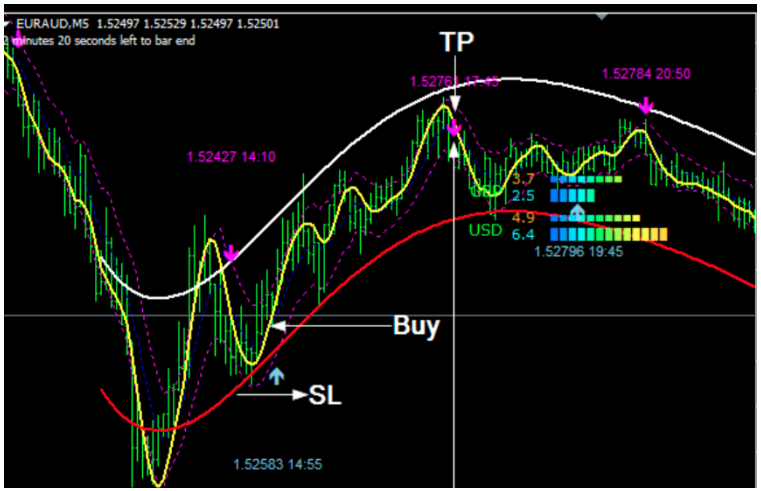
yinfeng@cuhk.edu.cn

February 23, 2020

Overview

Motivating Example

Stock price modeling and prediction



Source: website

Questions:

1. Can we predict the stock price well with a simple linear regression model in one input?
2. How about using the polynomial regression model?

Polynomial Regression Model

- Polynomial regression models can be used for curvilinear response surface of the input(s).
- The gist behind is the Taylor series representation of a function $f(x)$:

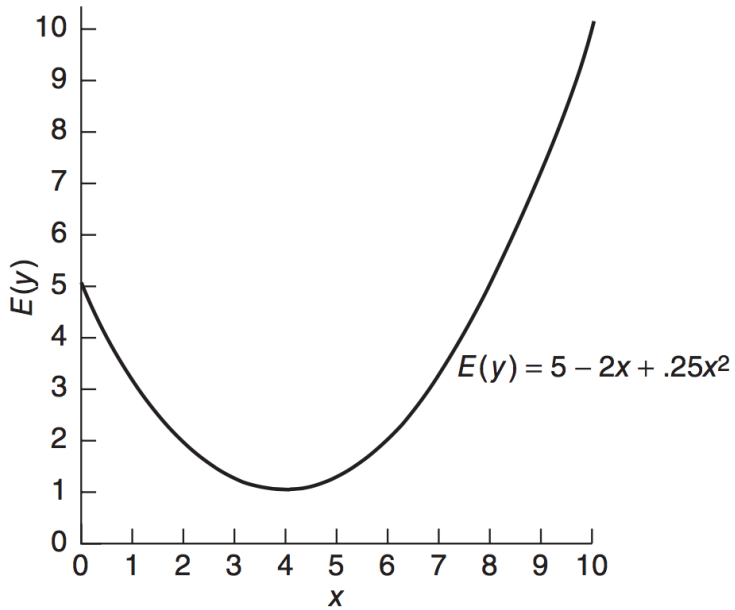
$$f(x) = \sum_{j=0}^{\infty} \frac{f^{(j)}(a)}{j!}(x-a)^j \approx \sum_{j=0}^{M} \frac{f^{(j)}(a)}{j!}(x-a)^j. \tag{1}$$

Polynomial Model in 1 Input: Example

A polynomial regression model in one variable,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon, \tag{2}$$

also known as quadratic model in one variable.



Source: textbook

Polynomial Model in 1 Input: General Expression

A hierarchical k-th order polynomial model in one input is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ...\beta_k x^k + \varepsilon = \sum_{j=0}^{k} \beta_j x^j + \varepsilon. \tag{3}$$

Remark: If we set $x_j = x^j, j = 1, 2, ..., k$ in the above expression, we get a multiple linear regression model in the conventional form with $x_1, x_2, ..., x_k$.

Issues with polynomial models

- Order of the Model (Overfitting)
- Model Building Strategy
- Extrapolation/Interpolation
- Numerical ill-conditioning

Issues with Polynomial Model

Issues with polynomial models

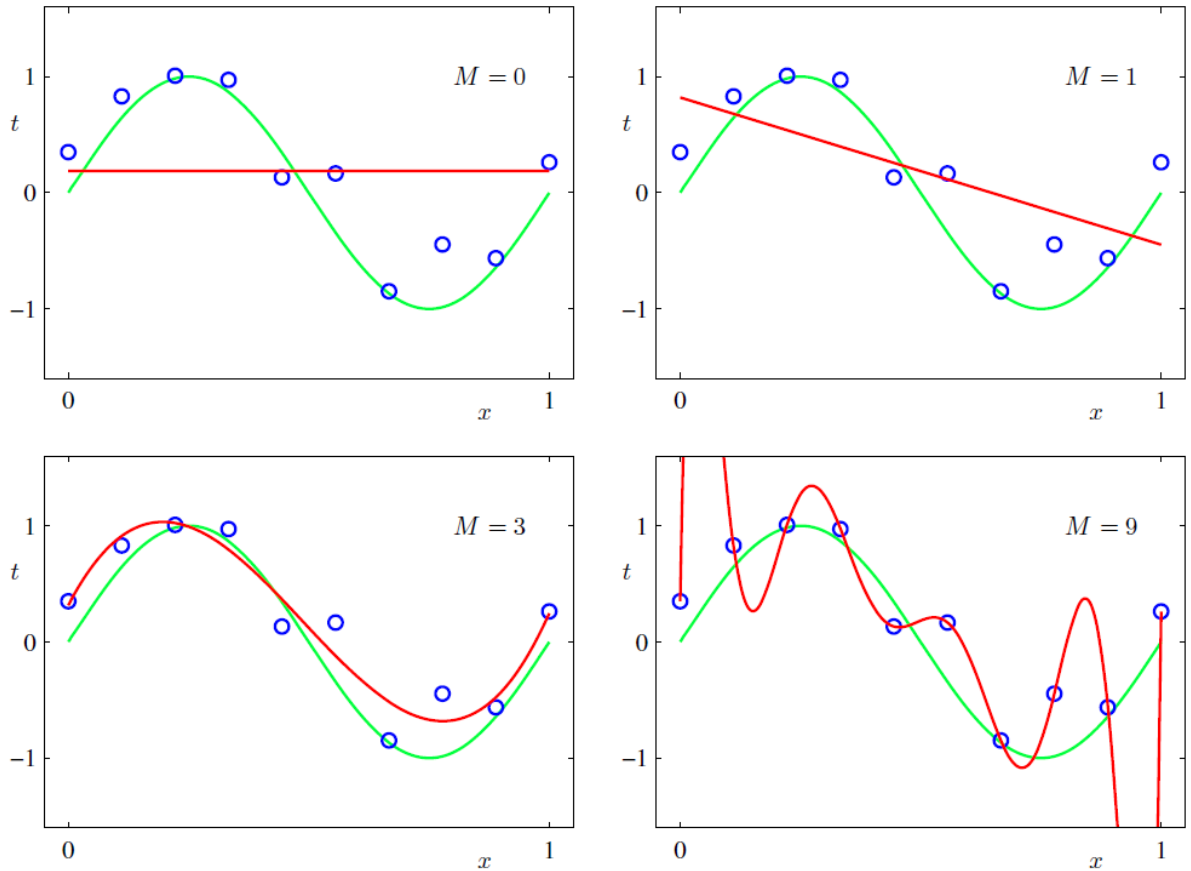
- Order of the Model (Overfitting)
- Model Building Strategy
- Extrapolation/Interpolation
- Numerical ill-conditioning

1. Order of the Model (Overfitting): It is dangerous to adopt high-order ($k > 2$) polynomial regression models for two reasons. First, overfitting may occur; Second, ill-conditioning may occur.

$\rightsquigarrow$ It is recommended to set k as low as possible, otherwise regularization terms have to be taken into account when fitting the model parameters.

Issues with Polynomial Model

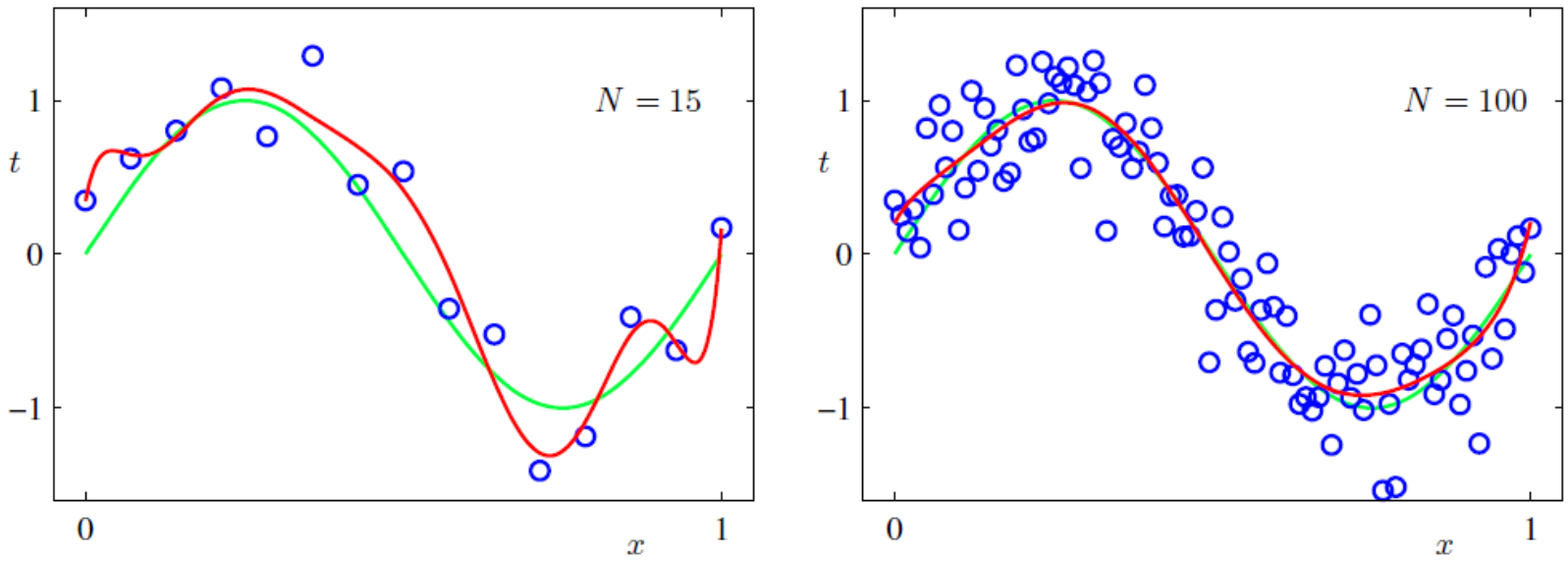
Fitting a sinusoidal function $sin(2\pi x)$ by polynomials with different orders $k = M = 0, 1, 3, 9$. The number of the output data is $n = 10$.



Source: C. Bishop, Machine Learning and Pattern Recognition

Issues with Polynomial Model

Fitting a sinusoidal function $sin(2\pi x)$ by polynomials with order $k = 9$. Increasing the number of the output data to $n = 15$ and $n = 100$.



Source: C. Bishop, Machine Learning and Pattern Recognition

Issues with Polynomial Model

Issues with polynomial models

- Order of the Model (Overfitting)
- Model Building Strategy
- Extrapolation/Interpolation
- Numerical ill-conditioning

2. Model Building Strategy: A preferred strategy is forward selection. The idea is to start with a low order polynomial model and successively fit the model parameters of increasing order until the t test for a higher order (say k is now the highest) term is non-significant, i.e.,

$$t_0 = \frac{\hat{\beta}_k}{\sqrt{MS_{Res} C_{kk}}} = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)} \sim t_{n-p} < t_{c/2,n-p}. \tag{4}$$

Issues with Polynomial Model

Issues with polynomial models

- Order of the Model (Overfitting)
- Model Building Strategy
- Extrapolation/Interpolation
- Numerical ill-conditioning

2. Model Building Strategy: The forward selection strategy is fairly conservative. Recent trend, however, is to use as many inputs/features as possible and let the algorithm determine which features are more important.

Issues with Polynomial Model

Issues with polynomial models

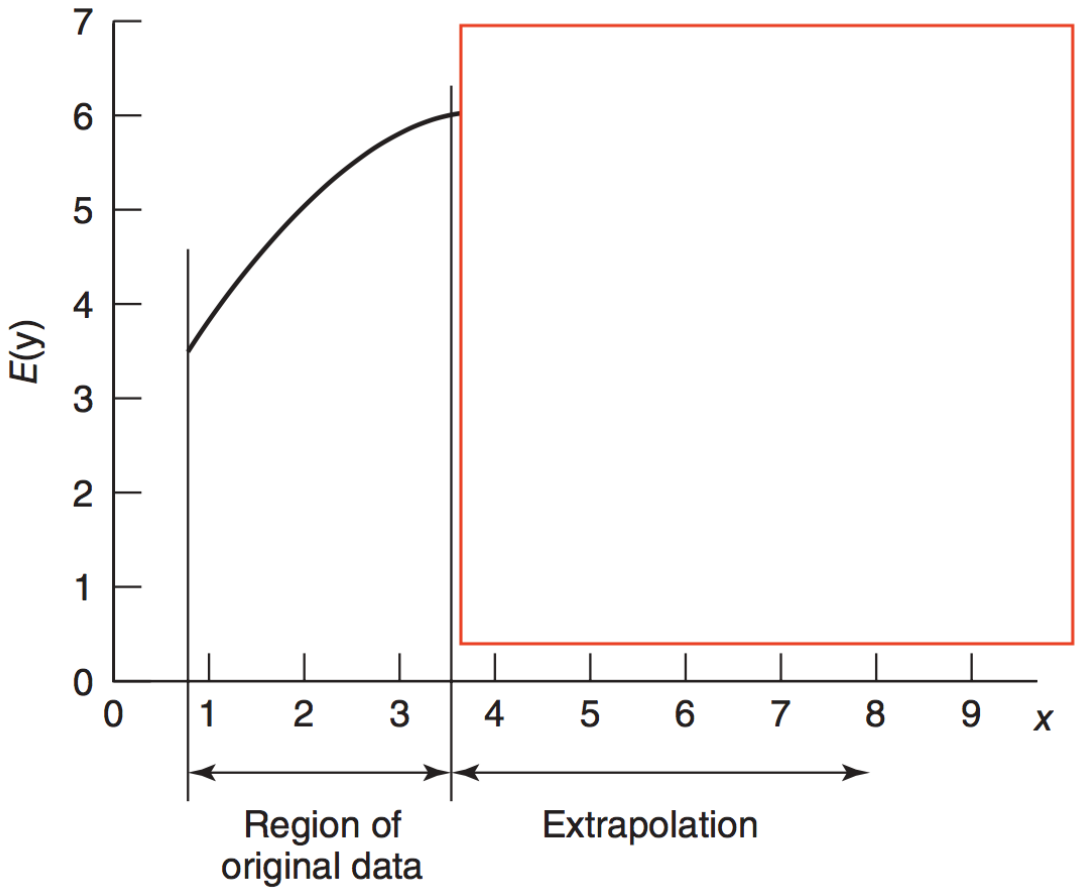
- Order of the Model (Overfitting)
- Model Building Strategy
- Extrapolation/Interpolation
- Numerical ill-conditioning

3. Extrapolation/Interpolation: By interpolation/extrapolation, we mean conducting prediction in/beyond the range of the observed input values x.

Special Note: Extrapolation and interpolation using polynomial models with high order can be extremely hazardous.

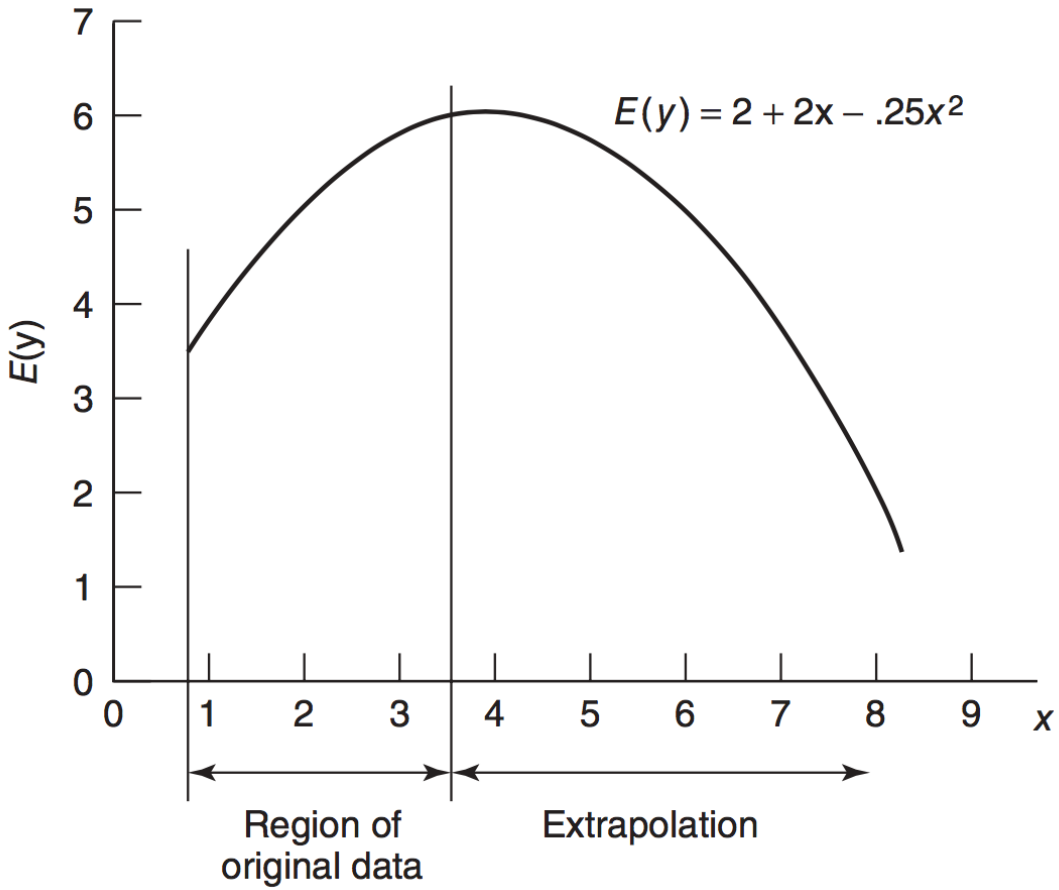
Issues with Polynomial Model

3. Extrapolation/Interpolation: Extrapolation and interpolation with high-order polynomial models can be extremely hazardous, e.g.,



Source: textbook

Actual answer:



Source: textbook

Issues with Polynomial Model

Issues with polynomial models

- Order of the Model (Overfitting)
- Model Building Strategy
- Extrapolation/Interpolation
- Numerical ill-conditioning

4. Numerical ill-conditioning: Mainly due to the fact that the design matrix $\mathbf{X}$ is not of full column rank, incurred by multicolinearity.

- Centering the inputs can help reduce the possibility of having ill-conditioned design matrix.
- Deficient column rank is often the case when we use high-order model or the values of x are limited to a narrow range, say 1 to 2, then x and x^2 are highly correlated.

Issues with Polynomial Model

When there are near-linear dependencies among the inputs/regressors, the problem of multicollinearity is said to exist.

We can formally define multicollinearity in terms of the linear dependence of the columns of $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p]$, i.e., there is a non-zero α such that $\sum_{i=1}^{p} \alpha_i \mathbf{x}_i = \mathbf{0}$.

Two ways of detecting multicollinearity:

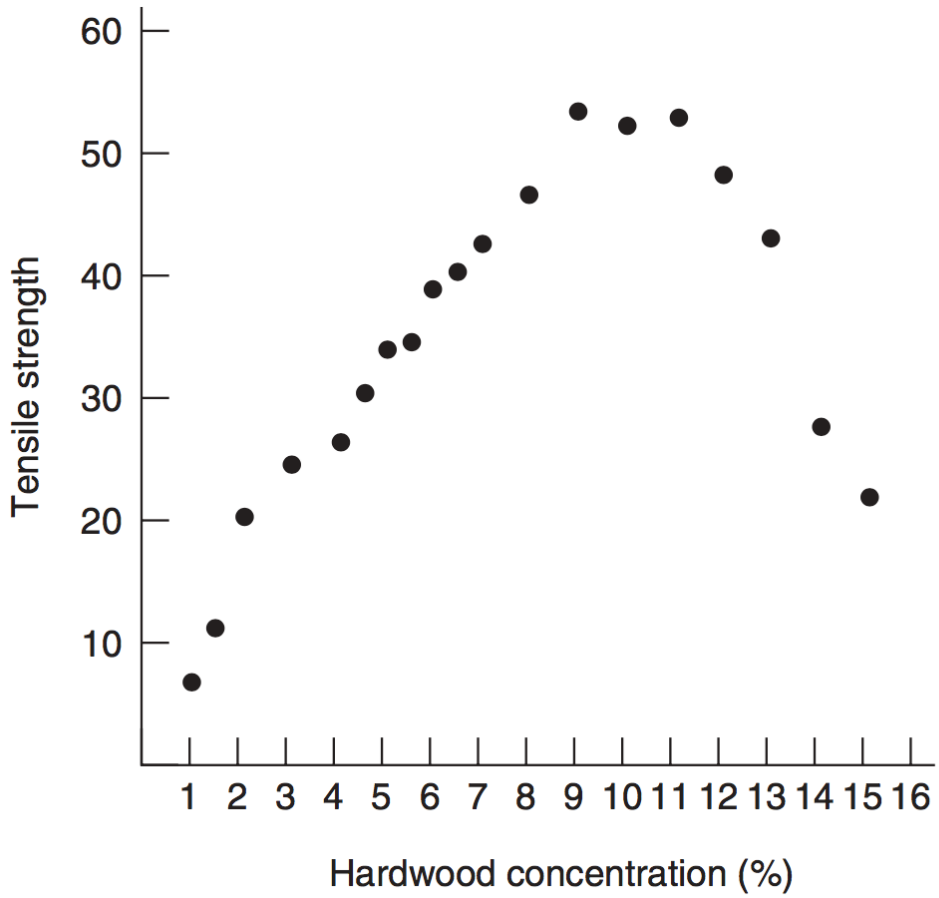
1. A very simple measure of multicollinearity is inspection of the off-diagonal elements r_{ij} in $\mathbf{X}^T\mathbf{X}$.
2. Use the eigenvalues or condition indices of $\mathbf{X}^T\mathbf{X}$ to measure the extent of multicollinearity in the data.

Polynomial Regression Example: Data

Data ($n = 19$ samples) concerning the strength of kraft paper and the percentage of hardwood in the batch of pulp from which the paper was produced.

Hardwood Concentration, x_i (%)	Tensile Strength, (psi) y, (psi)
1	6.3
1.5	11.1
2	20.0
3	24.0
4	26.1
4.5	30.0
5	33.8
5.5	34.0
6	38.1
6.5	39.9
7	42.0
8	46.1
9	53.1
10	52.0
11	52.5
12	48.0
13	42.8
14	27.8
15	21.9

Polynomial Regression Example: Scatter Plot



Source: textbook

Polynomial Regression Example: Model Fitting

1. We fit a low-order polynomial model with centered inputs

$$y = \beta_0 + \beta_1(x - \bar{x}) + \beta_2(x - \bar{x})^2 + \varepsilon \tag{5}$$

2. We perform least-squares fit and get $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
3. We get the fitted model

$$y = 45.295 + 2.546(x - 7.263) + 0.635(x - 7.263)^2 \tag{6}$$

4. Lastly, we test for significance of the regression model (see next page)

Polynomial Regression Example: Test for Significance

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_{Res}
Residual	SS_{Res}	$n-k-1$	MS_{Res}	
Total	SS_T	$n-1$		

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	3104.247	2	1552.123	79.434
Residual	312.638	16	19.540	
Total	3416.885	18		

1. $R^2 = \frac{SS_R}{SS_T} \approx 0.9085$, $R^2_{adj} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)} \approx 0.8971$.
2. In order to test the significance of individual model parameter, for instance $H_0 : \beta_2 = 0, H_1 : \beta_2 \neq 0$, we require the estimated standard error $se(\hat{\beta}_2) = 0.062$.

Piecewise Polynomial Fitting (Splines)

Problem: Underlying function behaves differently in different ranges of x.
Solution: Divide the range of x into a few segments and fit them separately. Use spline functions to perform piecewise polynomial fitting.

Splines in General and Cubic Splines

Splines are piecewise polynomials of order k in general. The joint points of the pieces are usually called knots. A cubic spline ($k = 3$) is normally adequate for practical problems. A cubic spline with h knots, $t_1 < t_2 < ... < t_h$, and continuous first- and second order derivatives can be written as:

$$S(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{h} \beta_i (x - t_i)_+^3. \tag{7}$$

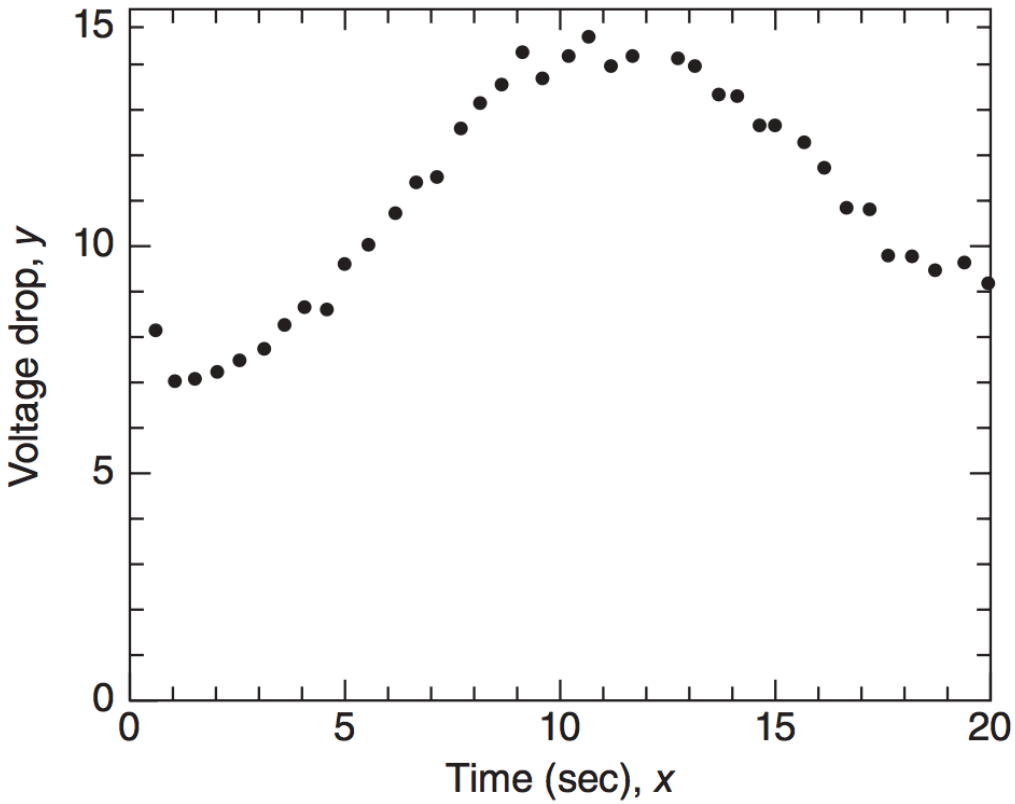
Note: If the knot positions are known, ordinary LS fitting works; otherwise, the joint estimation problem is nonlinear and difficult.

Piecewise Polynomial Regression Example: Data

Data (size $n = 41$ samples) concerning battery voltage drop in a guided missile motor observed over the time of missile flight.

Observation, i	Time, x_i (seconds)	Voltage Drop, y_i	Observation, i	Time, x_i (seconds)	Voltage Drop, y_i
1	0.0	8.33	21	10.0	14.48
2	05	823	22	105	14.92
ODEL INADEQUACIES	1.0	7.17	23	11.0	14.37
4	1.5	7.14	24	11.5	14.63
5	2.0	7.31	25	12.0	15.18
6	2.5	7.60	26	12.5	14.51
7	3.0	7.94	27	13.0	14.34
8	3.5	8.30	28	13.5	13.81
9	4.0	8.76	29	14.0	13.79
10	4.5	8.71	30	14.5	13.05
11	5.0	9.71	31	15.0	13.04
12	5.5	10.26	32	15.5	12.60
13	6.0	10.91	33	16.0	12.05
14	6.5	11.67	34	16.5	11.15
15	7.0	11.76	35	17.0	11.15
16	7.5	12.81	36	17.5	10.14
17	8.0	13.30	37	18.0	10.08
18	8.5	13.88	38	18.5	9.78
19	9.0	14.59	39	19.0	9.80
20	95	14.05	40	19.5	9.95
			41	20.0	9.51

Piecewise Polynomial Regression Example: Scatter Plot



Two knots locate at $t_1 = 6.5$ and $t_2 = 13$ seconds after the missile launch determined from the expert knowledge.

Piecewise Polynomial Regression Example: Fitted Model and ANOVA

The cubic spline model is

$$y = \beta_{00} + \beta_{01}x + \beta_{02}x^2 + \beta_{03}x^3 + \beta_1(x - 6.5)_+^3 + \beta_2(x - 13)_+^3 + \varepsilon \quad (8)$$

Performing LS fit and ANOVA test of significance, yields:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	260.1784	5	52.0357	725.52
Residual	2.5102	35	0.0717	
Total	262.6886	40		

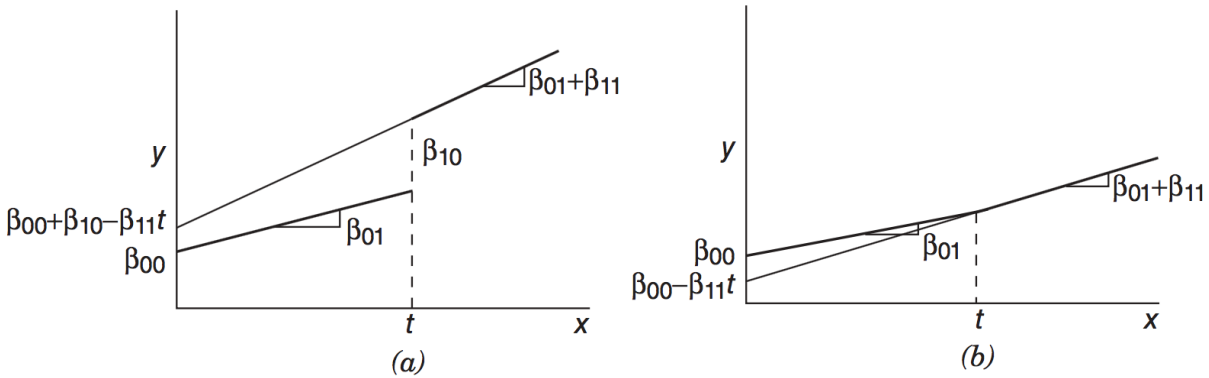
Piecewise Polynomial Fitting (Splines)

Linear Splines

Linear splines can be used to fit piecewise simple linear regression models. suppose there is a single knot at t and that there could be both a slope change and a discontinuity at the knot.

$$\text{discontinuous}: \quad S(x) = \beta_{00} + \beta_{01}x + \beta_{10}(x-t)_+^0 + \beta_{11}(x-t)_+^1 \quad \text{or} \quad (9)$$

$$\text{continuous}: \quad S(x) = \beta_{00} + \beta_{01}x + \beta_{11}(x-t)_+^1 \quad (10)$$



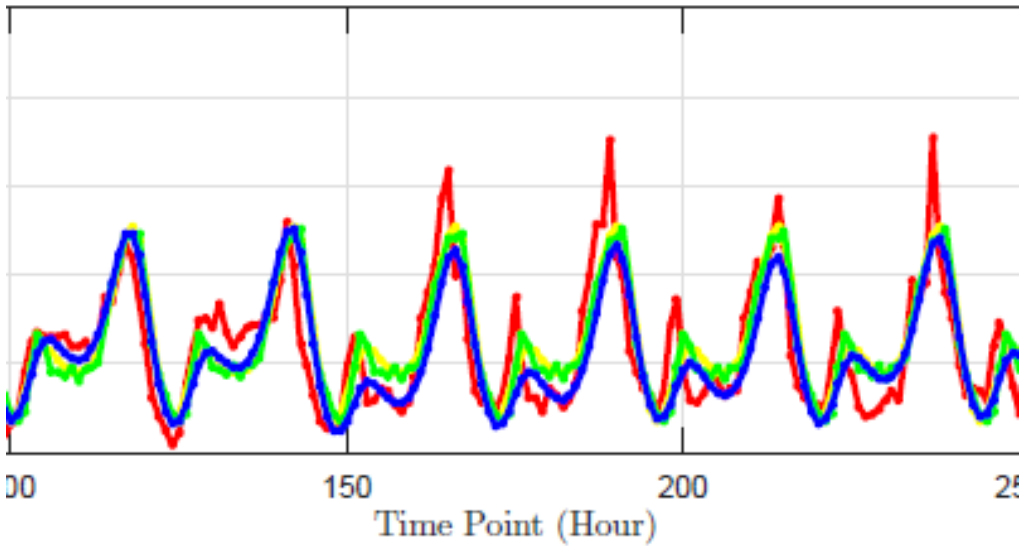
(a) discontinuity at the knot; (b) continuous at the knot.

Polynomial and Trigonometric Terms

A better model combines polynomial- and trigonometric terms, in particular when the data profile shows periodicity pattern.

The modified model is:

$$y = \beta_0 + \sum_{i=1}^{d} \beta_i x^i + \sum_{j=1}^{r} (\delta_j \sin(jx) + \gamma_j \cos(jx)) + \varepsilon \tag{11}$$



wireless traffic versus time.

Polynomial Regression Model in ≥ 2 Inputs

Fitting a polynomial regression model in two or more input/regressor variables is a straightforward extension. For instance, a second-order polynomial model in two inputs:

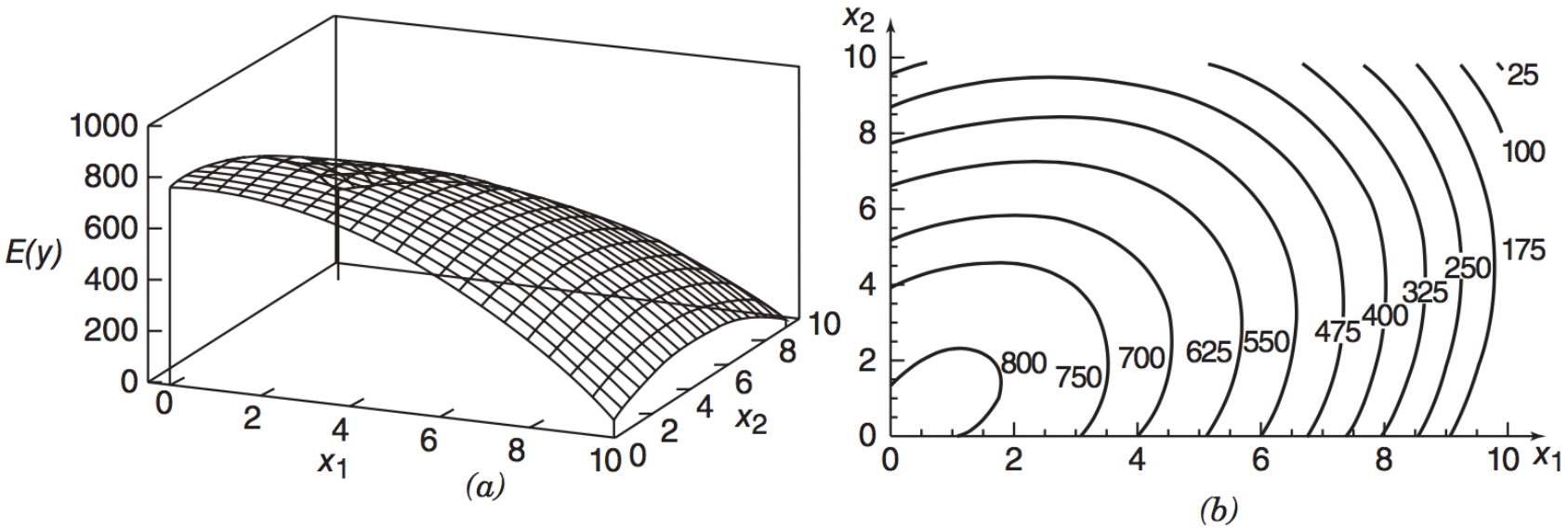
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon, \tag{12}$$

where

- β_1 and β_2 are two linear effect parameters;
- β_{11} and β_{22} are two quadratic effect parameters;
- β_{12} is an interaction effect parameter.

Example ($\geq$ 2 Inputs)

Example:



$$E(y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2$$

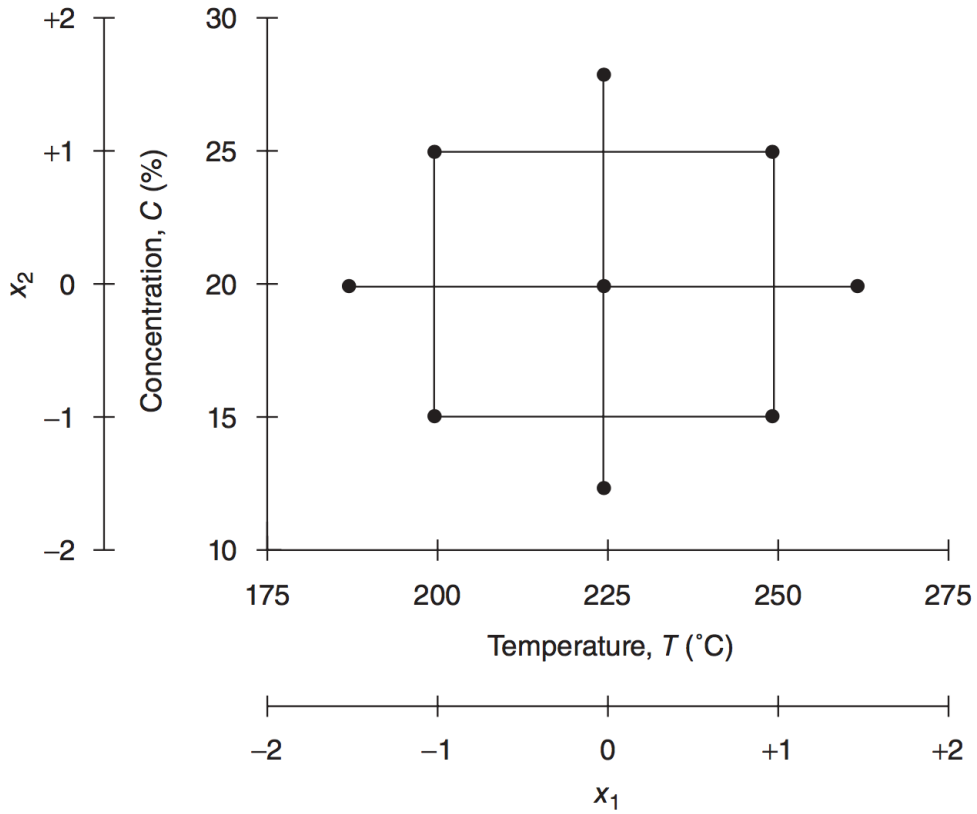
Polynomial Regression Example (≥ 2 Inputs): Data

Data (size $n = 12$ samples) from an experiment that was performed to study the effect of two variables, reaction temperature (T) and reactant concentration (C), on the percent conversion of a chemical process y.

		A		B		
Observation	Run Order	Temperature (°C) T	Cone. (%) C	x_1	x_2	y
1	4	200	15	−1	−1	43
2	12	250	15	1	−1	78
3	11	200	25	−1	1	69
4	5	250	25	1	1	73
5	6	189.65	20	−1.414	0	48
6	7	260.35	20	1.414	0	76
7	1	225	12.93	0	−1.414	65
8	3	225	27.07	0	1.414	74
9	8	225	20	0	0	76
10	10	225	20	0	0	79
11	9	225	20	0	0	83
12	2	225	20	0	0	81

Polynomial Regression Example (≥ 2 Inputs): Data

Central composite design [1], from natural units to levels:



[1] More details about central composite design can be found here: http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/doe/response-surface-designs/what-is-a-central-composite-design/

Polynomial Regression Example (≥ 2 Inputs): Model Fitting

We fit the second-order model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon, \tag{13}$$

Given a dataset,

$$\mathbf{X} = \begin{array}{c} \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_1^2 & x_2^2 & x_1x_2 \end{array} \\ \begin{bmatrix} -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 \\ -1.414 & 0 & 2 & 0 & 0 \\ 1.414 & 0 & 2 & 0 & 0 \\ 0 & -1.414 & 0 & 2 & 0 \\ 0 & 1.414 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}, \quad \mathbf{y} = \begin{bmatrix} 43 \\ 78 \\ 69 \\ 73 \\ 48 \\ 76 \\ 65 \\ 74 \\ 76 \\ 79 \\ 83 \\ 81 \end{bmatrix}$$

$$\mathbf{X'X} = \begin{bmatrix} 12 & 0 & 0 & 8 & 8 & 0 \\ 0 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 0 & 0 & 0 \\ 8 & 0 & 0 & 12 & 4 & 0 \\ 8 & 0 & 0 & 4 & 12 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}, \quad \mathbf{X'y} = \begin{bmatrix} 845.000 \\ 78.592 \\ 33.726 \\ 511.000 \\ 541.000 \\ -31.000 \end{bmatrix}$$

The fitted model is $\hat{y} = 79.8 + 9.8x_1 + 4.2x_2 - 8.9x_1^2 - 5.1x_2^2 - 7.8x_1x_2$.

Summary

To summarize with some keywords:

- Polynomial regression
- Taylor series representation of a function
- Spline model
- Overfitting
- Extrapolation/Interpolation may be hazardous
- Numerical ill-condition due to multicollinearity

Appendix

Sample correlation coefficient or sample Pearson correlation coefficient of two vectors $\mathbf{x} = [x_1, x_2, ..., x_n]^T$ and $\mathbf{y} = [y_1, y_2, ..., y_n]^T$ is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{14}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the centroid of $\mathbf{x}$ and $\mathbf{y}$, respectively.