San José State University

Math 261A: Regression Theory & Methods

Polynomial Regression Models

Dr. Guangliang Chen

This lecture is based on the following part of the textbook:

- Sections 7.1 – 7.4

Outline of the presentation:

- **Polynomial regression**
  - Important considerations
  - Model fitting
- **Nonparametric methods**
  - Kernel regression
  - Loess regression

Introduction

Previously we talked about **transformations** on the response and/or the predictor(s) for linearizing a nonlinear relationship.

When this fails, we can turn to **polynomial regression** models such as

$$y = \beta_0 + \beta_1 x + \cdots + \beta_k x^k + \epsilon$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

to represent the nonlinear patterns:

- The top model is called a kth-order polynomial model in one variable;
- The bottom model is called a quadratic model in two variables.

Polynomial models are very powerful to handle nonlinearity, because polynomials can approximate continuous functions within any given precision.

However, such fitting problems can still be treated as **linear regression**:

$$y = \beta_0 + \beta_1 \underbrace{x}_{x_1} + \cdots + \beta_k \underbrace{x^k}_{x_k} + \epsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \underbrace{\beta_{11} x_1^2}_{\beta_3 x_3} + \underbrace{\beta_{22} x_2^2}_{\beta_4 x_4} + \underbrace{\beta_{12} x_1 x_2}_{\beta_5 x_5} + \epsilon$$

Thus, we can readily utilize the corresponding theory, tools and techniques for linear regression to carry out polynomial regression.

Important considerations

However, there will be important considerations in polynomial regression:

- **Order of the polynomial model**
- **Model-building strategy**
- **Extrapolation**
- **Ill-conditioning**
- **Hierarchy**

Order of the polynomial model

First, remember that it is always possible to fit a polynomial model of order $n-1$ perfectly to a data set n points (however, this will almost surely be overfitting!!!)

Transformations should be tried first to keep the model first order.

A low-order model in a transformed variable is almost always preferable to a high-order model in the original metric.

One should always maintain a sense of **parsimony**, that is, **use the simplest possible model** that is consistent with the data and knowledge of the problem environment.

Model-building strategy

There are two standard procedures for building a polynomial model:

- **Forward selection**: Successively fit models of increasing order until the t test for the highest order term is nonsignificant.
- **Backward elimination**: Appropriately fit the highest order model and then delete terms one at a time, starting with the highest order, until the highest order remaining term has a significant t statistic.

Interestingly, these two procedures do not necessarily lead to the same model.

Extrapolation

Because polynomial models may turn in unanticipated and inappropriate directions, extrapolation with them can be extremely hazardous.

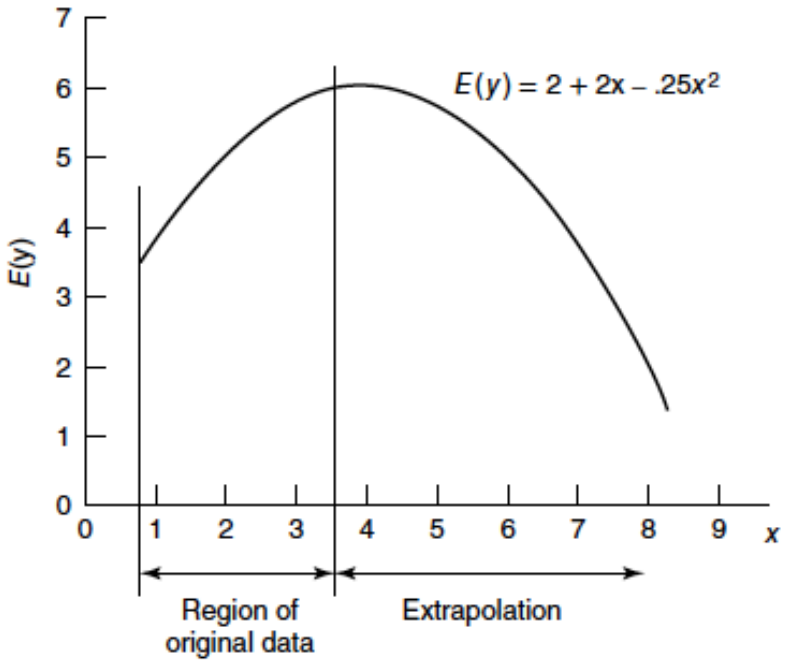


Figure 7.2 Danger of extrapolation.

Ill-conditioning

In the setting of polynomial regression, the design matrix may have lots of columns corresponding to just one predictor or two.

Those columns will have a significant multicolinearity, especially when the values of x are limited to a narrow range.

As the order of the polynomial model increases, $\mathbf{X}'\mathbf{X}$ become more and more ill-conditioned, meaning that matrix inversion calculations are more and more inaccurate.

Centering the data (i.e., letting $\tilde{x}_i = x_i - \bar{x}$) may remove some nonessential ill-conditioning.

Hierarchy

The regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

is said to be **hierarchical** since it contains all terms of order 3 and lower.

In contrast, the models

$$y = \beta_0 + \beta_1 x + \beta_3 x^3 + \epsilon, \qquad y = \beta_0 + \beta_1 x_1 + \beta_{12} x_1 x_2 + \epsilon$$

are not hierarchical.

Polynomial regression

We present the model in the following two cases:

- Polynomial regression in one variable (Hardwood example)

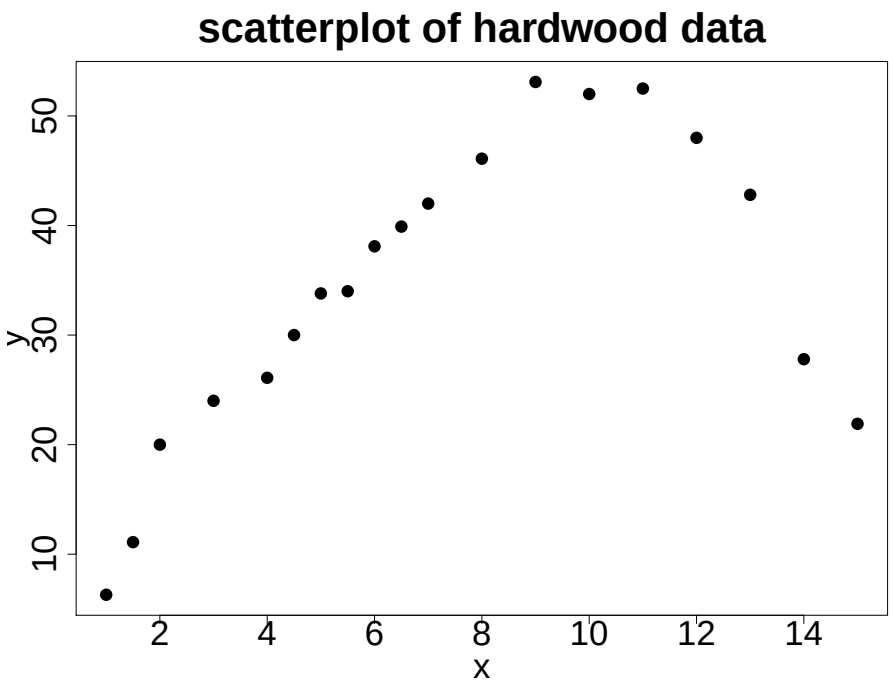
$$y = \beta_0 + \beta_1 x + \cdots + \beta_k x^k + \epsilon$$

- Polynomial regression in two variables (Chemical Process example)

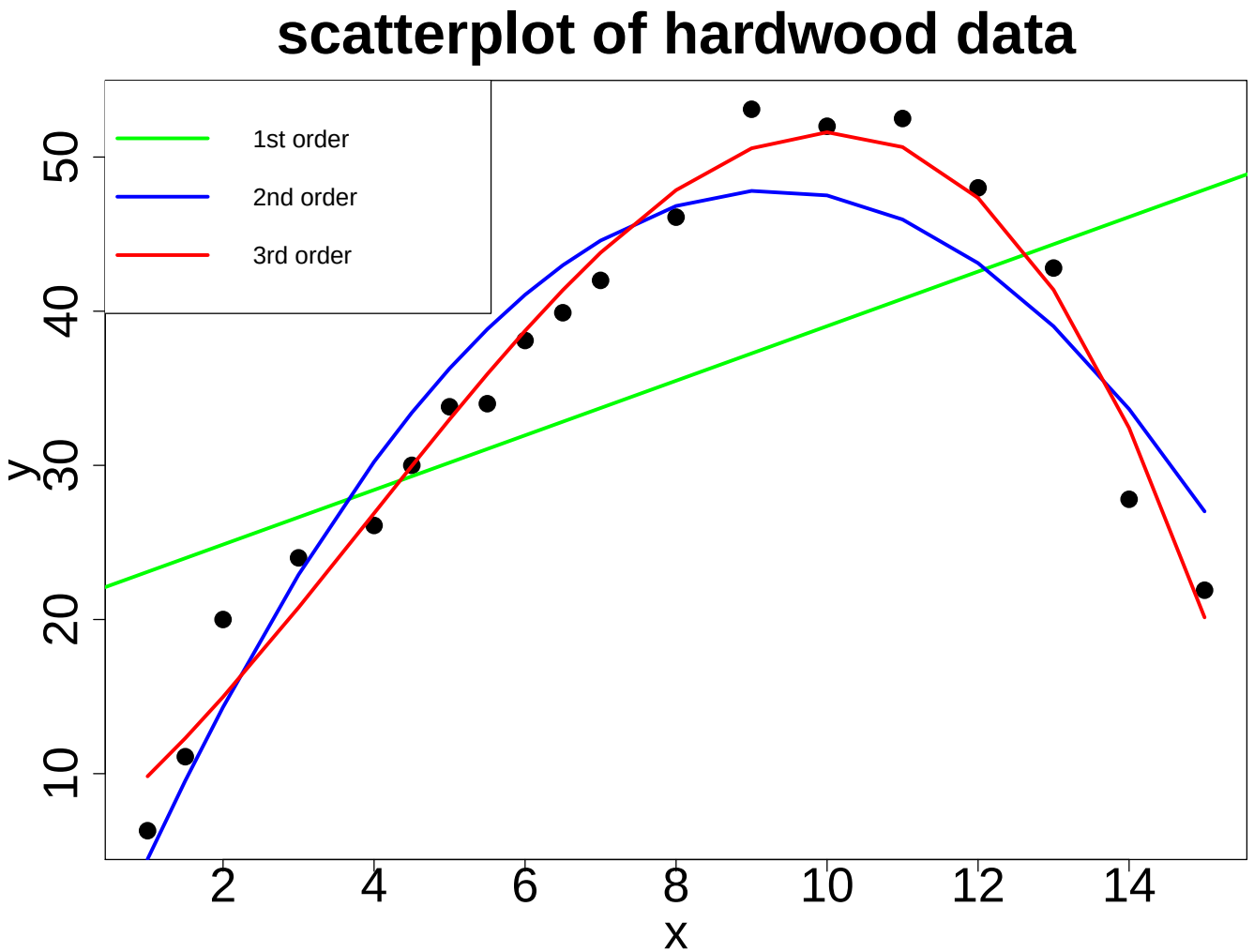
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

Example 1: The Hardwood Data

We have 19 observations concerning the **strength of kraft paper** (y) and the **percentage of hardwood** (x) in the batch of pulp from which the paper was produced.



Three polynomial models along with the linear model were fitted to the data.


scatterplot of hardwood data
1st order
2nd order
3rd order
y
x
10
20
30
40
50
2
4
6
8
10
12
14

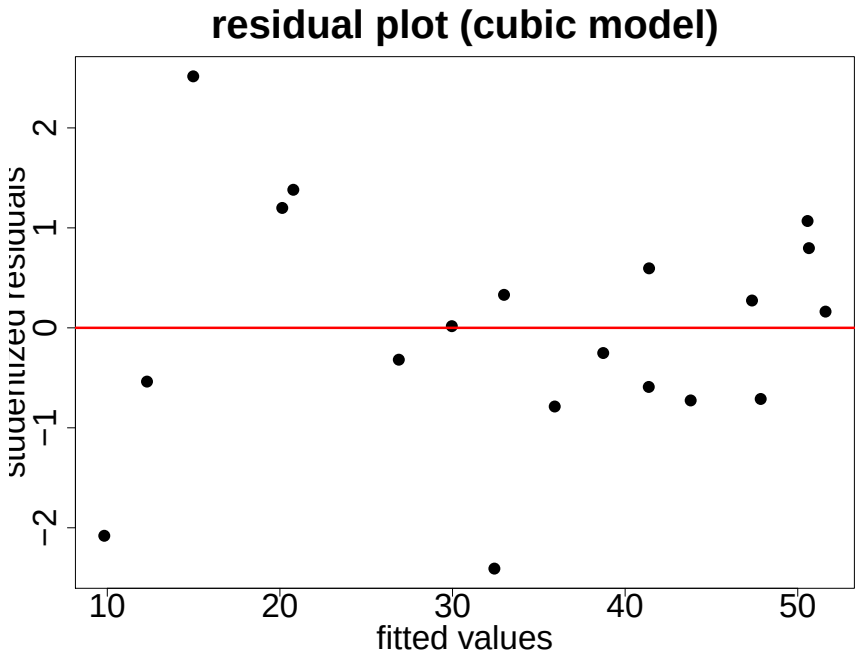
The summary statistics of the four models are reported below:

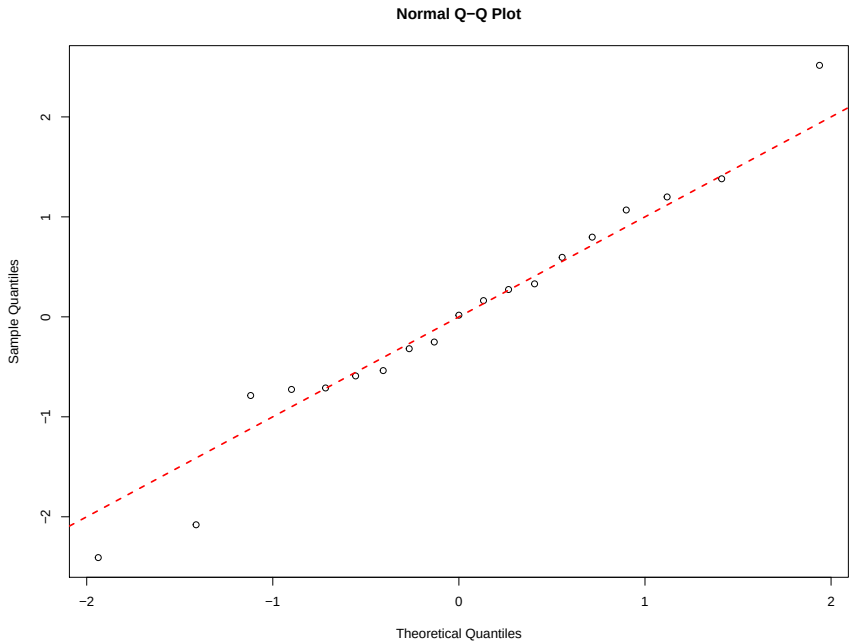
Order of poly. model	$\mathbf{R^2}$	$\mathbf{R^2_{adj}}$	$\hat{\sigma}^2$	**p-value***
1 (linear model)	0.3054	0.2645	11.82^2	0.01414
2 (quadratic model	0.9085	0.8971	4.42^2	1.89e-08
3 (cubic model)	0.9707	0.9648	2.585^2	4.72e-05
4	0.9736	0.9661	2.539^2	0.233

*of the t test for the highest-order term in each model.

Which model should we select?

The cubic model is the best!





Example 2: Chemical Process data

The following table presents data 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).

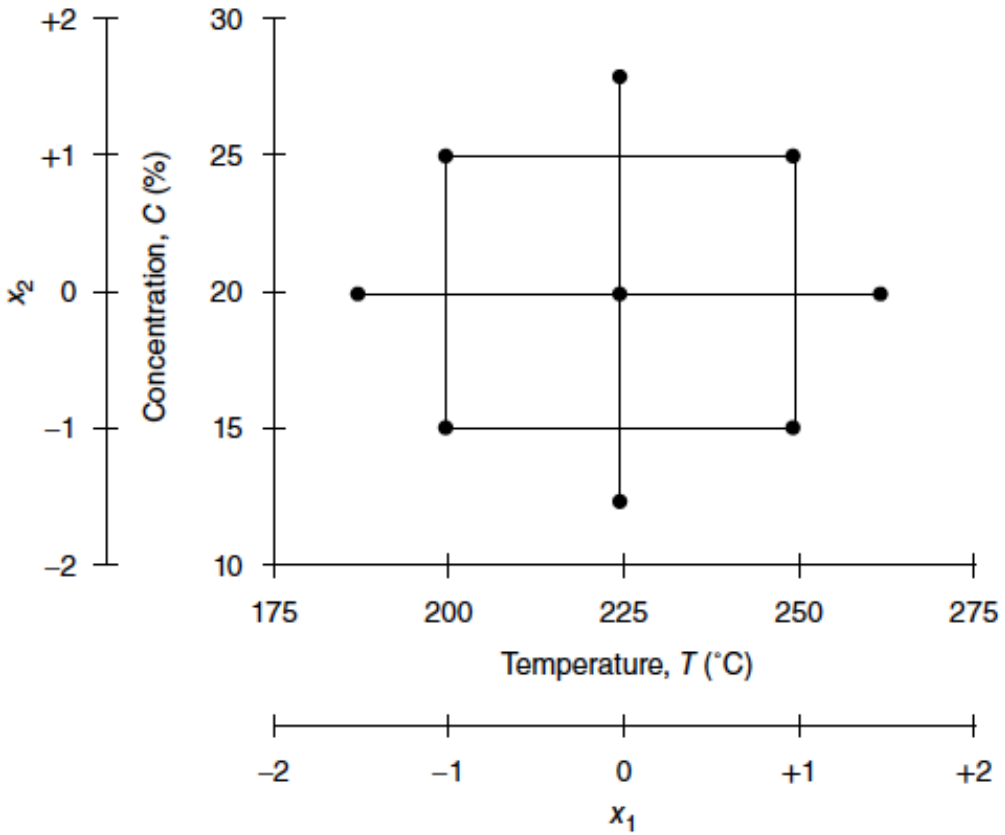
Panel A of the table shows the levels used for T and C in the natural units of measurements, and panel B shows the levels in terms of coded variables x_1 and x_2:

$$x_1 = \frac{T - 225}{25}, \quad x_2 = \frac{C - 20}{5}$$

TABLE 7.7 Central Composite Design for Chemical Process Example

		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

The process engineers adopted a **central composite design** in order to fit a second-order model:



We fit a full quadratic model in the coded variables x_1, x_2:

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

The design matrix and the response vector for this model are

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

By direct calculation, we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (79.75, 9.83, 4.22, -8.88, -5.13, -7.75)'$$

Therefore, the fitted model (in the coded variables) is

$$\hat{y} = 79.75 + 9.83x_1 + 4.22x_2 - 8.88x_1^2 - 5.13x_2^2 - 7.75x_1x_2$$

In terms of the original variables, the model is

$$\hat{y} = -1105.56 + 8.024T + 22.994C - 0.0142T^2 - 0.205C^2 - 0.062TC$$

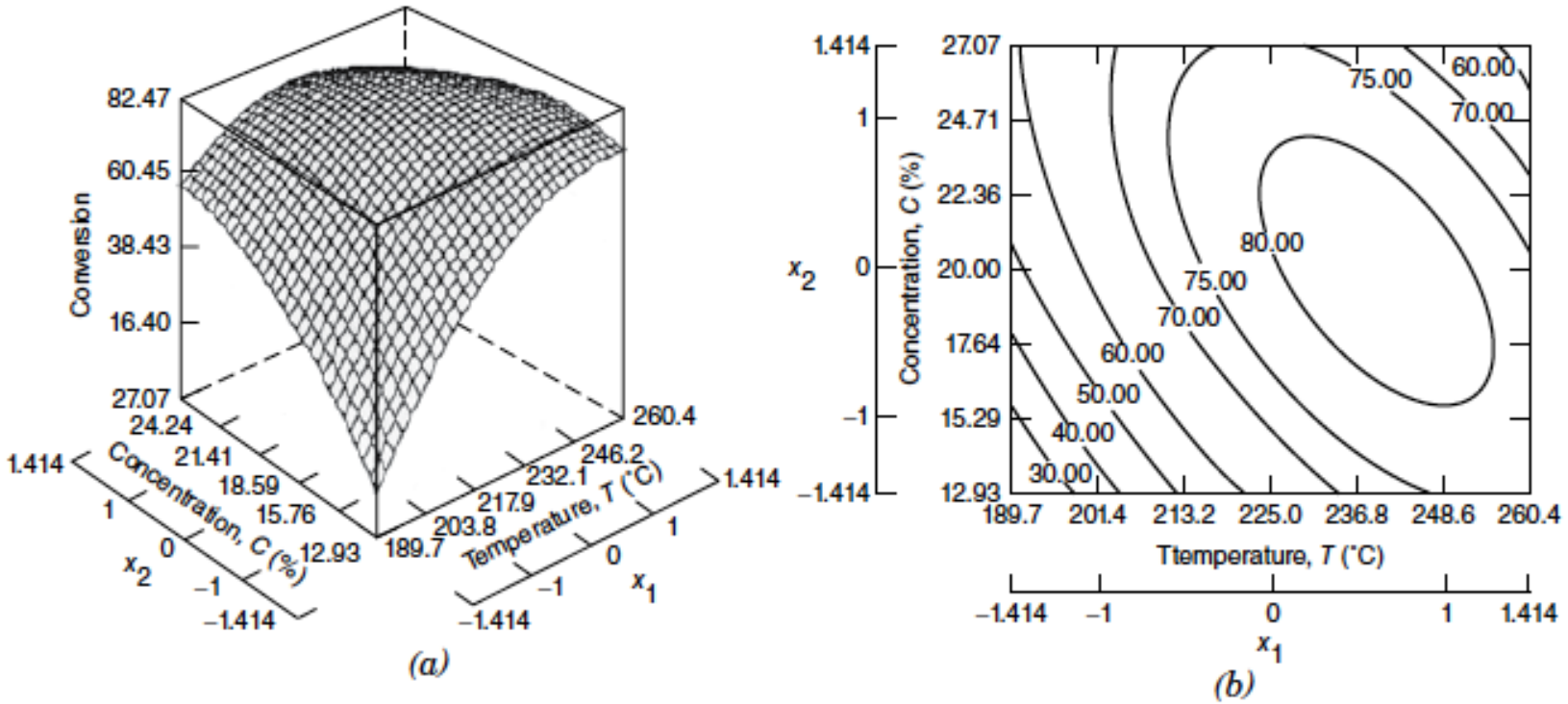


Figure 7.17 (*a*) Response surface of predicted conversion. (*b*) Contour plot of predicted conversion.

Such a **response surface methodology (RSM)** is widely applied in industry for optimizing the response.

Nonparametric regression models

So far we have discussed linear and polynomial regression models which all specify a functional relationship betwee the predictors and the response:

$$y = f(\underbrace{\mathbf{x}}_{\text{predictors}}, \underbrace{\boldsymbol{\beta}}_{\text{parameters}}) + \epsilon$$

They are examples of **parametrized regression** where we

- first choose a class of the function f (linear, quadratic, etc.) and
- then use the data to estimate the parameters $\boldsymbol{\beta}$.

Nonparametric regression methods do not need to specify the form of the function f (thus there is no parameter)

$$y = f(\underbrace{\mathbf{x}}_{\text{predictors}}) + \epsilon$$

but use the data to directly make predictions in some way (In some sense, the goal is to estimate the function f itself).

We mention two nonparametric regression methods:

- **Kernel regression**
- **Locally weighted regression (Loess)**

Kernel regression

Recall that in ordinary linear regression (with the least squares criterion), the fitted values are collectively given by

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}.$$

Individually, we have for each point i,

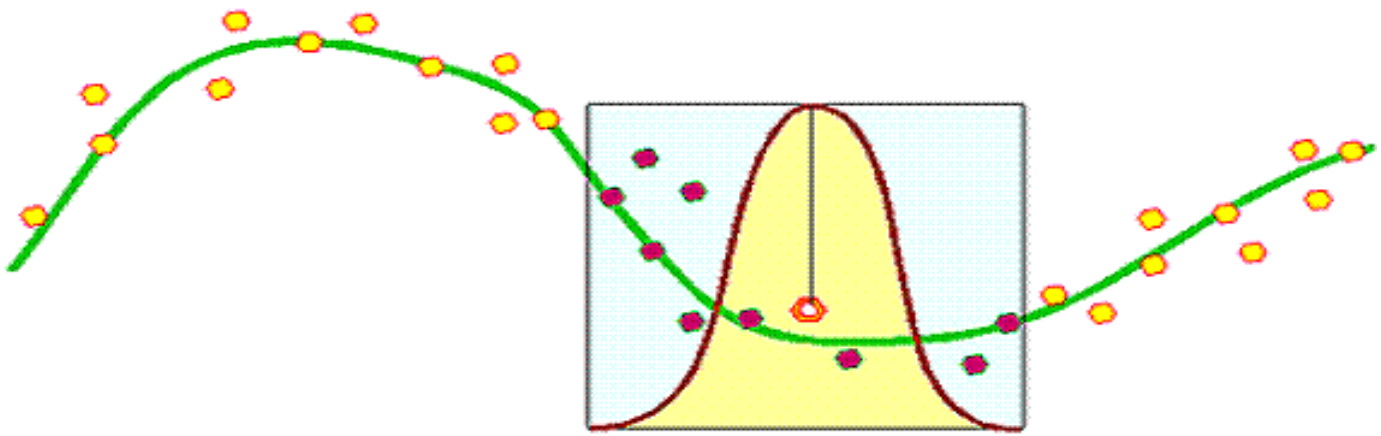
$$\hat{y}_i = \sum_j h_{ij} y_j.$$

This shows that each fitted value $\hat{y}_i$ is a linear combination of the observations y_i but with different weights.

Kernel regression maintains the form of estimation but extends the weights using a **kernel function** $K(\cdot)$ to predict y at any specific location x_0:

$$\tilde{y} = \sum_j w_j y_j, \qquad \text{where } w_j = \frac{K(x_j - x_0)}{\sum_k K(x_k - x_0)} \text{ for each } j.$$

Note that $\sum w_j = 1$. This operation is called **kernel smoothing**.



Typically, the kernel functions have the following properties:

- $K(t) \geq 0$ for all t
- $\int_{-\infty}^{\infty} K(t)\,\mathrm{d}t = 1$
- $K(-t) = K(t)$ for all t

Note that these are properties of symmetric probability density functions.

Additionally, K is often required to **peak at zero** and **become (nearly) zero outside a neighborhood of 0**, so that only points in the neighborhood are used for prediction at x_0.

Common kernel functions

- Gaussian kernel function

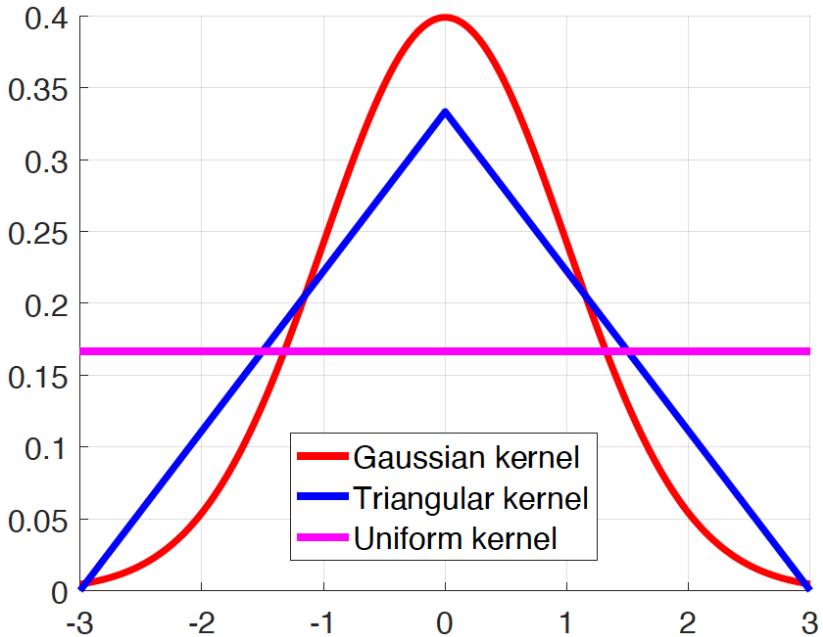
$$K(t;b) = \frac{1}{\sqrt{2\pi(b/3)^2}} e^{-\frac{t^2}{2(b/3)^2}}$$

- Triangular kernel function

$$K(t;b) = \begin{cases} \frac{1}{b}(1 - \frac{|t|}{b}), & |t| < b \\ 0, & |t| > b \end{cases}$$

- Uniform kernel function

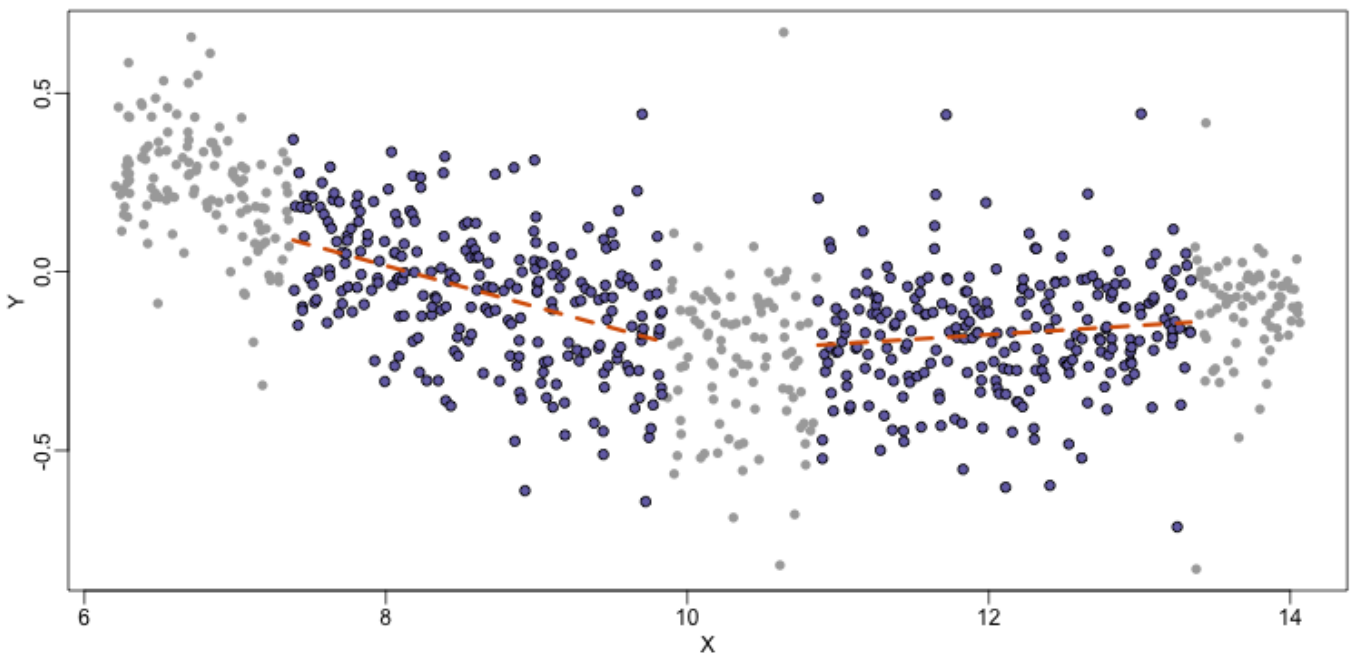
$$K(t;b) = \begin{cases} \frac{1}{2b}, & |t| < b \\ 0, & |t| > b \end{cases}$$



The radius b of the neighborhood is called the **bandwidth** of the kernel.

Locally weighted regression (Loess)

Like kernel regression, loess also uses the data from a neighborhood around the specific location x_0, defined by the **span** (fraction of the total points).



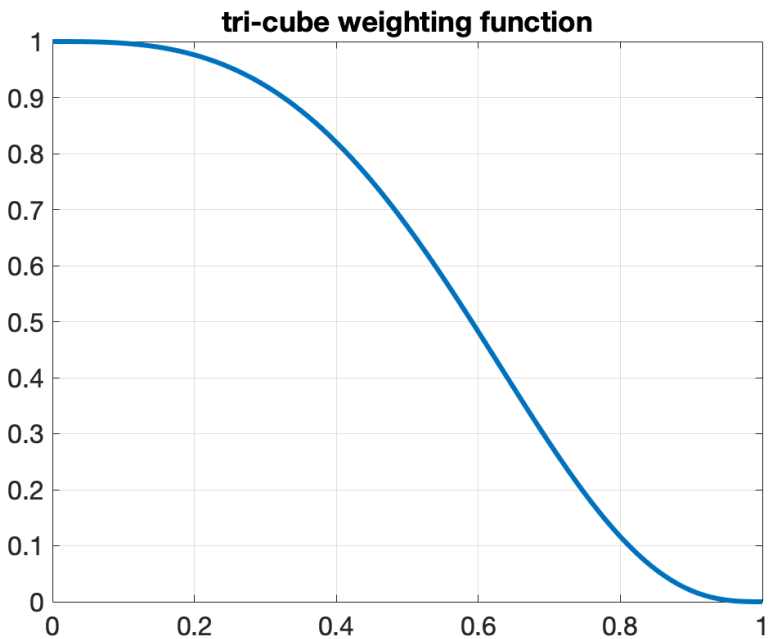
The loess procedure uses the points in the neighborhood to generate a **weighted least-squares** estimate of the specific response y at x_0 (usually through simple linear regression or a quadratic regression model).

Most software packages use the **tri-cube weighting function**

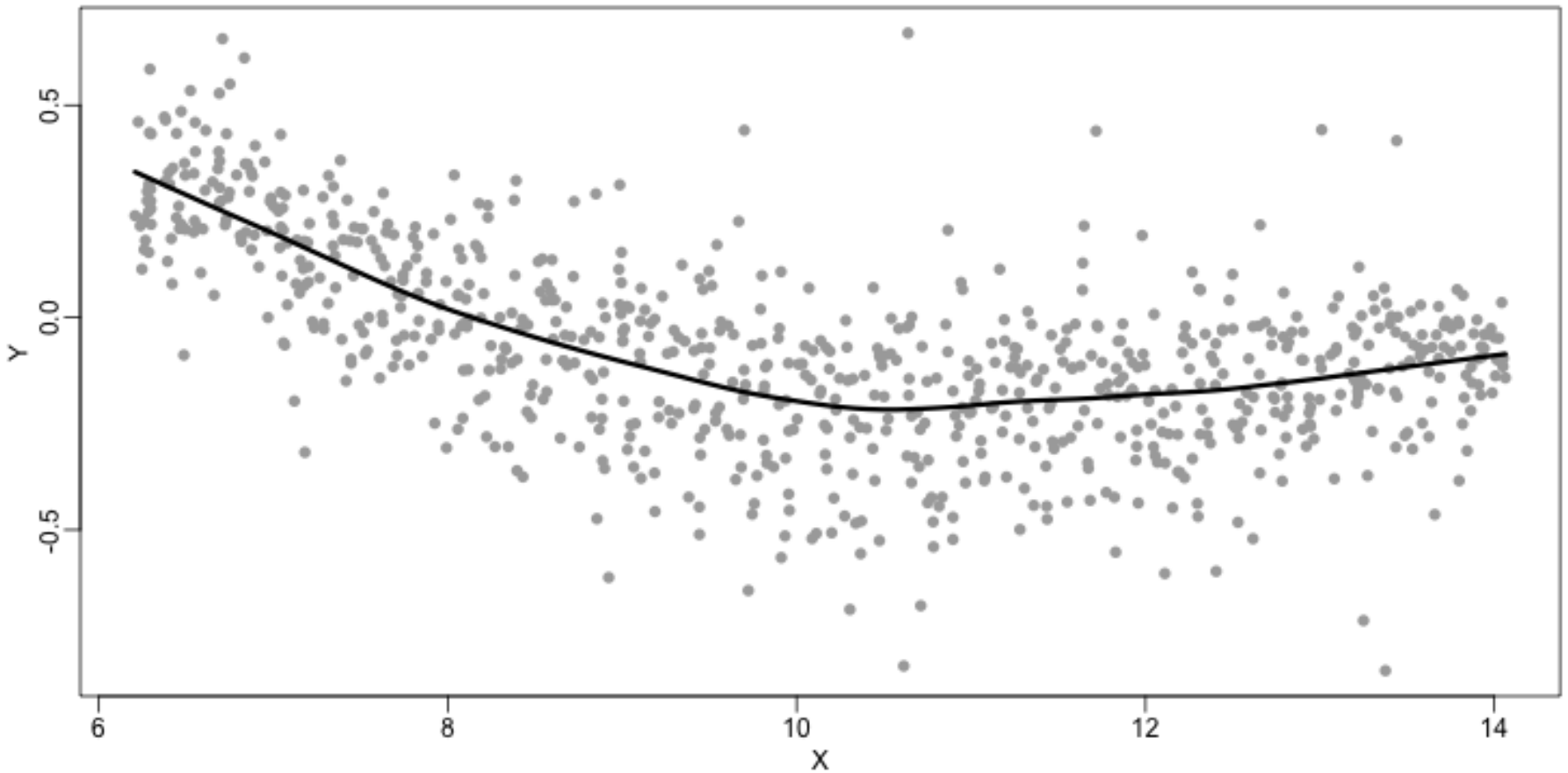
$$W(t) = \begin{cases} (1-t^3)^3, & 0 \le t \le 1 \\ 0, & t > 1 \end{cases}$$

to assign weights for each point x_j in the neighborhood of x_0

$$w_j = W\left(\frac{|x_j - x_0|}{\Delta(x_0)}\right)$$



Polynomial Regression



Summary

We have talked about the following regression methods:

- **Polynomial regression**
- **Nonparametric methods***
  - Kernel regression
  - Loess regression

*These methods are flexible but computationally very intensive.

Further learning

7.2.2 Piecewise Polynomial Fitting (Splines)

7.5 Orthogonal Polynomials