Linear Regression

Outline

- Introduction
- Linear Regression
- Least Squares Estimators of the Regression Parameters
- Analysis of Residuals: Assessing the Model
- Polynomial Regression

Introduction

- In many situations, there is a single response variable Y, also called the dependent variable, which depends on the value of a set of input, also called independent variables $x_1, \dots, x_r$.
- The simplest type of relationship between the dependent variable Y and the input variables $x_1, \dots, x_r$ is a linear relationship.
- That is, for some constants $\beta_0, \beta_1, \dots, \beta_r$ the relationship can be written as $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r$.

Introduction

- However, in practice, such precision is almost never attainable, and the most that one can expect is that previous equation would be valid subject to some random error. By this we mean that the explicit relationship is

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_r x_r + e$$
$$\equiv \alpha + \beta \mathbf{x} + e.$$

- where e, representing the random error, is assumed to be a random variable having mean 0.

Linear Regression Equation

- $E[Y|\boldsymbol{x}]$ is the expected response given the inputs x,

$$E[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r$$

  where $x = (x_1, \ldots, x_r)$ is the set of independent variables.

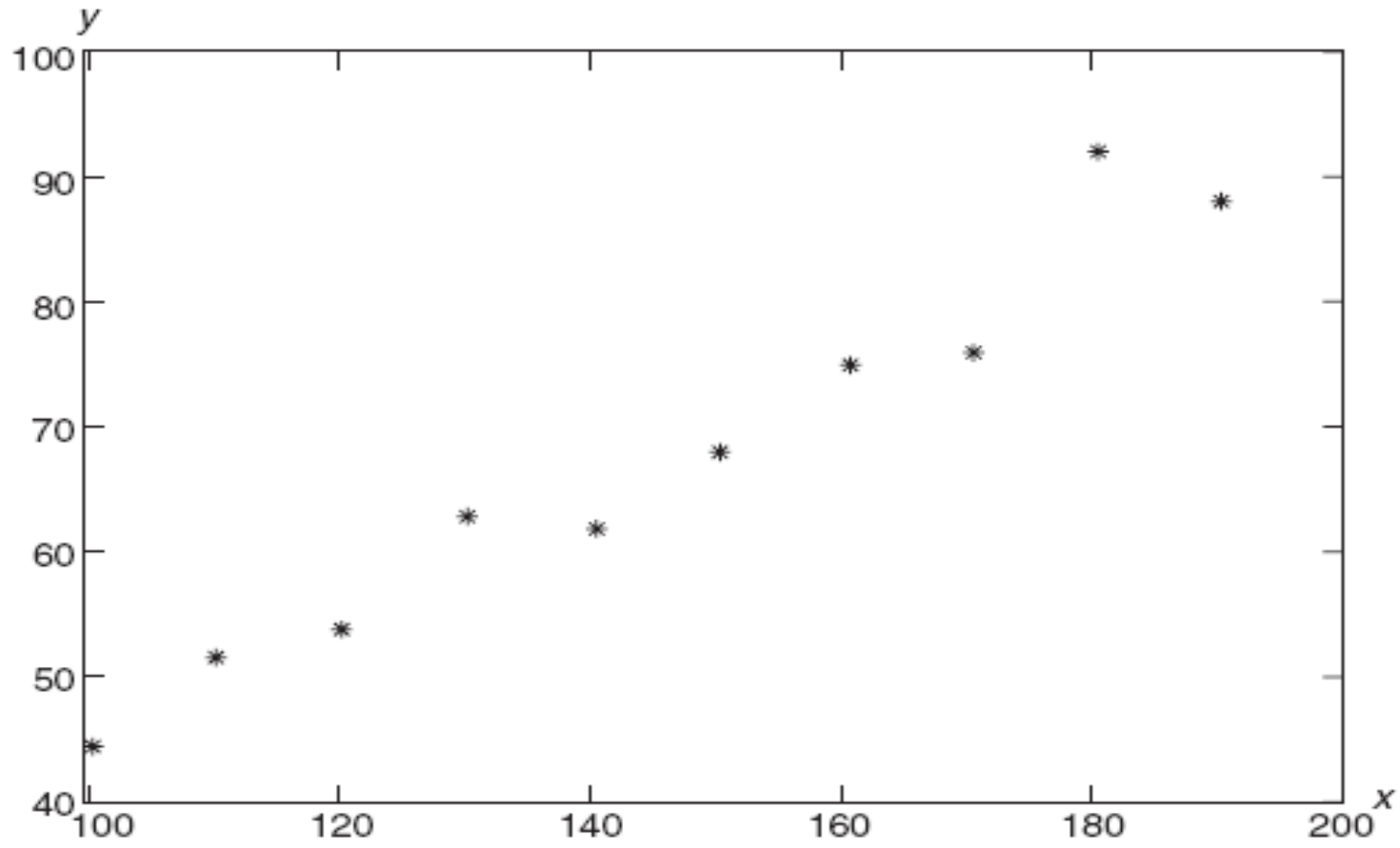
- It describes the regression of Y on the set of independent variables $x_1, \ldots, x_r$.
- $\beta_0, \beta_1, \ldots, \beta_r$ are called the regression coefficients.

Example

- Consider the following 10 data pairs $(x_i, yi), i = 1, \ldots, 10$, relating y, the percent yield of a laboratory experiment, to x, the temperature at which the experiment was run

i	x_i	y_i	i	x_i	y_i
1	100	45	6	150	68
2	110	52	7	160	75
3	120	54	8	170	76
4	130	63	9	180	92
5	140	62	10	190	88

A plot of y_i versus x_i — called a *scatter diagram* — is given in Figure 9.1. As this scatter diagram appears to reflect, subject to random error, a linear relation between y and x, it seems that a simple linear regression model would be appropriate.



Scatter plot.

Least Squares Estimators of the Regression Parameters

PROPOSITION 9.2.1 The least squares estimators of β and α corresponding to the data set $x_i, Y_i, i = 1, \ldots, n$ are, respectively,

$$B = \frac{\sum_{i=1}^{n} x_i Y_i - \bar{x} \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$A = \bar{Y} - B\bar{x}$$

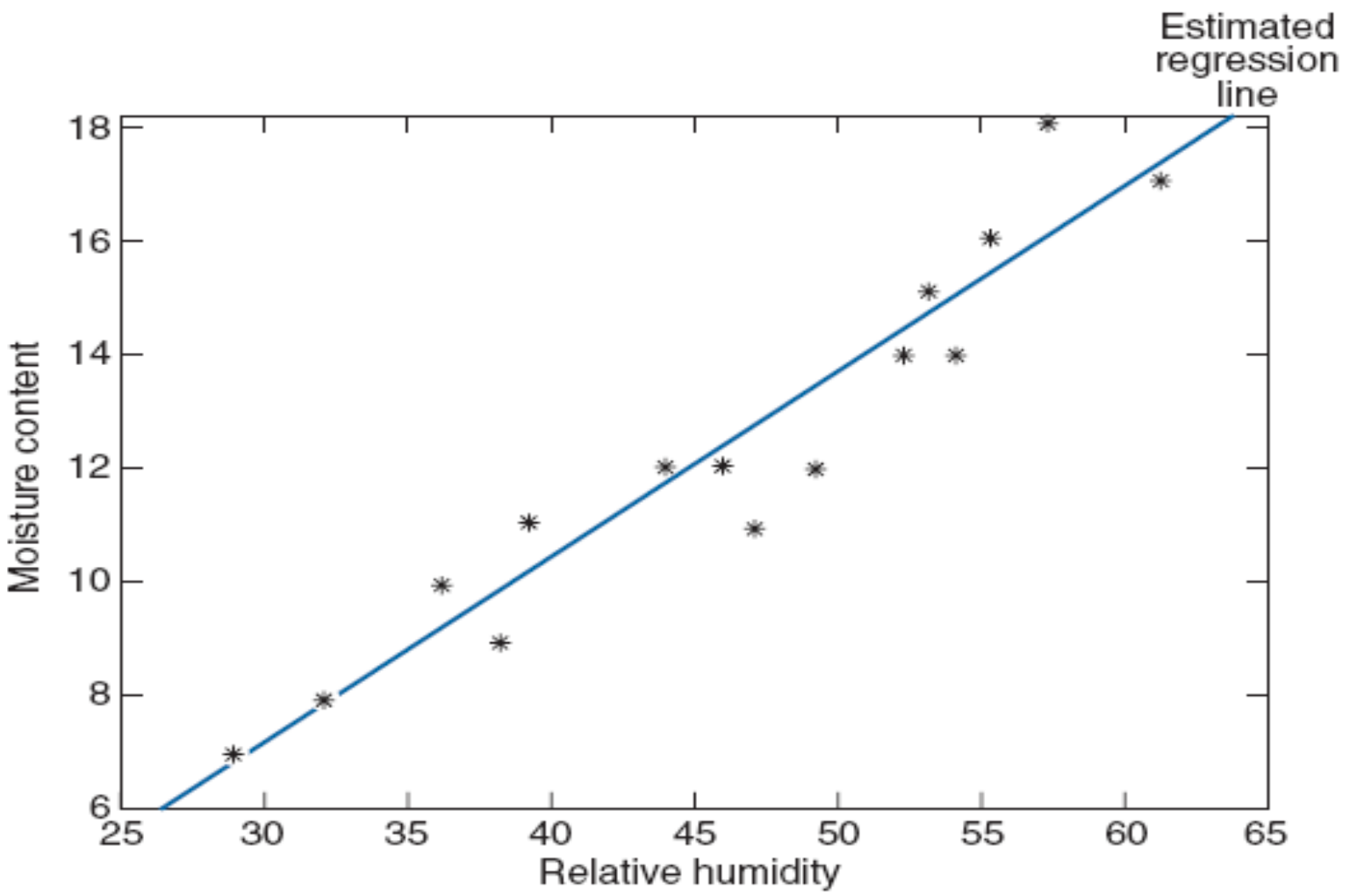
The straight line $A + Bx$ is called the estimated regression line.

LSE: Example

EXAMPLE 9.2a The raw material used in the production of a certain synthetic fiber is stored in a location without a humidity control. Measurements of the relative humidity in the storage location and the moisture content of a sample of the raw material were taken over 15 days with the following data (in percentages) resulting.

Relative humidity	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Moisture content	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

Example: The Scatterplot

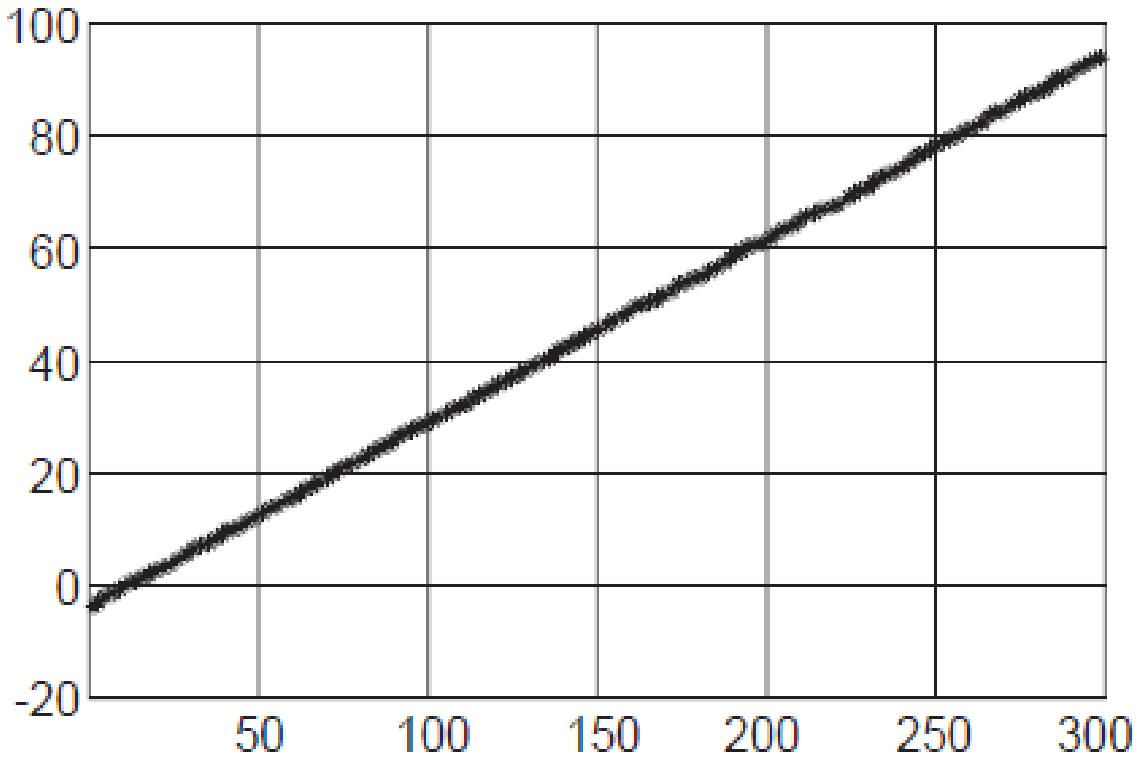


The least squares estimators are as follows:

a = –2.51 Average x value = 46.13

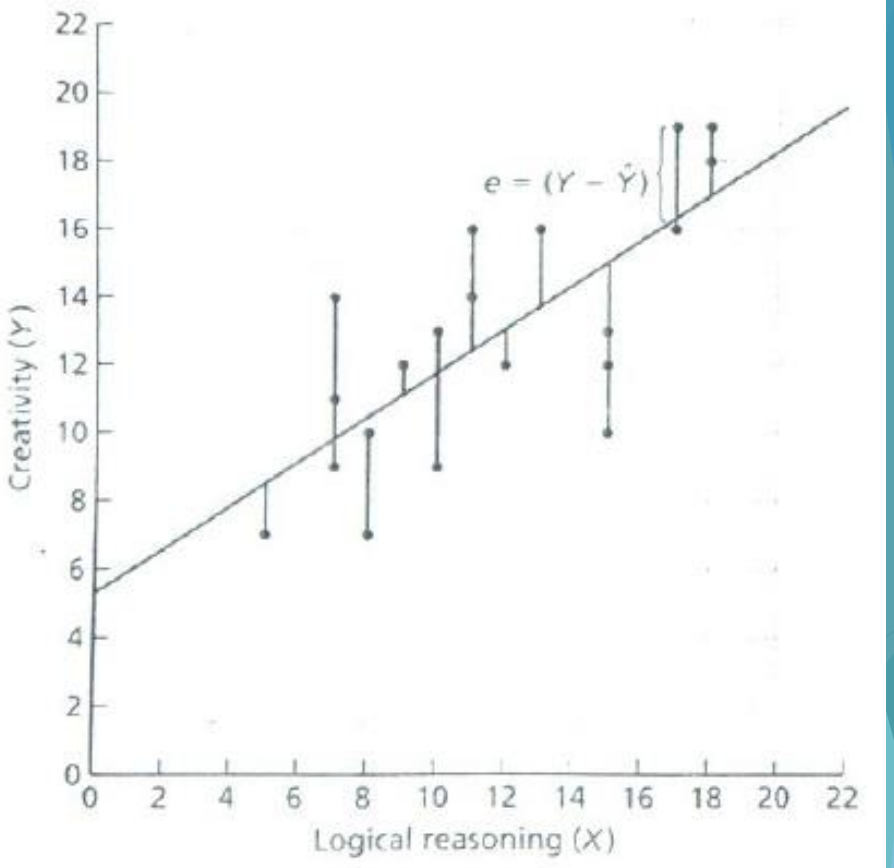
b = 0.32 Sum of squares of the x values = 33212.0

The estimated regression line is Y = –2.51 + 0.32x



Residual

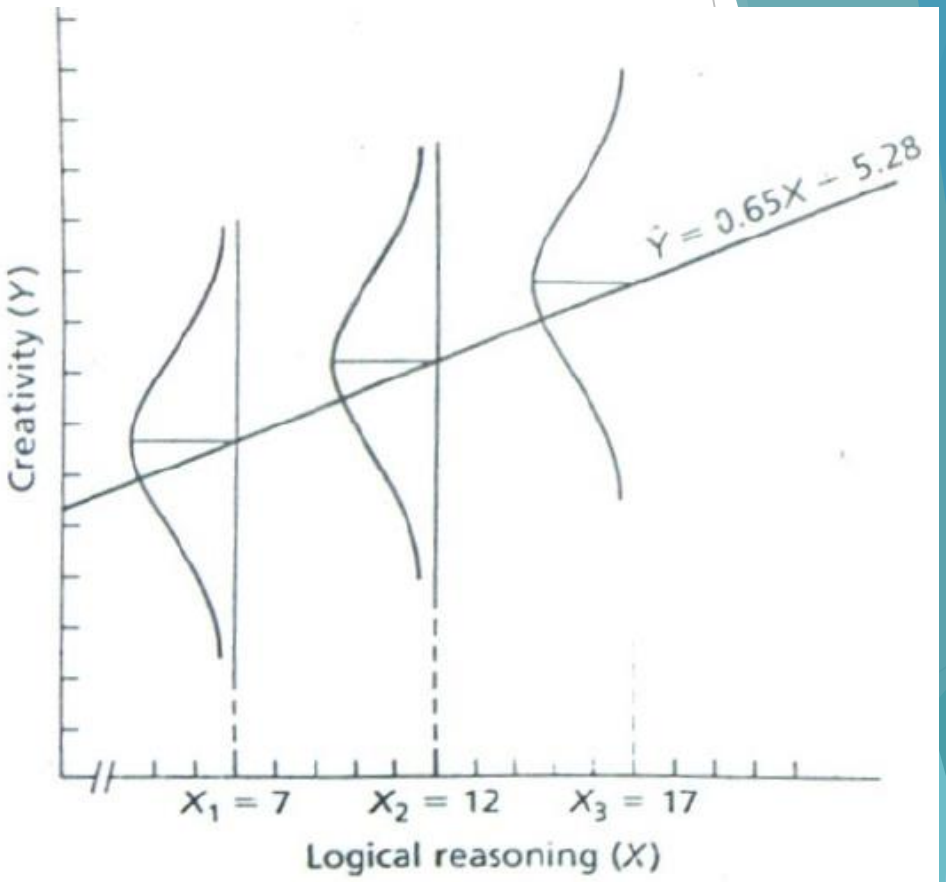
- The quantities $Y_i - A - Bx_i$, $i = 1, \ldots, n$, which represent the differences between the actual responses (that is, the Y_i) and their least squares estimators (that is, $A + Bxi$) are called the residuals.



Standard Error of Estimate

$$SS_R = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

***SSR* can be utilized to estimate the unknown variance σ^2 .**



Standard Error of Estimate (2)

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

That is, SS_R/σ^2 has a chi-square distribution with $n-2$ degrees of freedom, which implies that

$$E\left[\frac{SS_R}{\sigma^2}\right] = n - 2$$

or

$$E\left[\frac{SS_R}{n-2}\right] = \sigma^2$$

Thus $SS_R/(n-2)$ is an unbiased estimator of σ^2. In addition, it can be shown that SS_R is independent of the pair A and B.

Notation

If we let

$$S_{xY} = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = \sum_{i=1}^{n} x_i Y_i - n\overline{x}\,\overline{Y}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

$$S_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$$

$$B = \frac{S_{xY}}{S_{xx}}, \qquad A = \overline{Y} - B\,\overline{x} \qquad SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

Analysis of Residuals: Assessing the Model

The initial step for ascertaining whether or not the simple linear regression model

$$Y = \alpha + \beta x + e, \qquad e \sim \mathcal{N}(0, \sigma^2)$$

is appropriate in a given situation is to investigate the scatter diagram.

Analysis of Residuals: Assessing the Model (2)

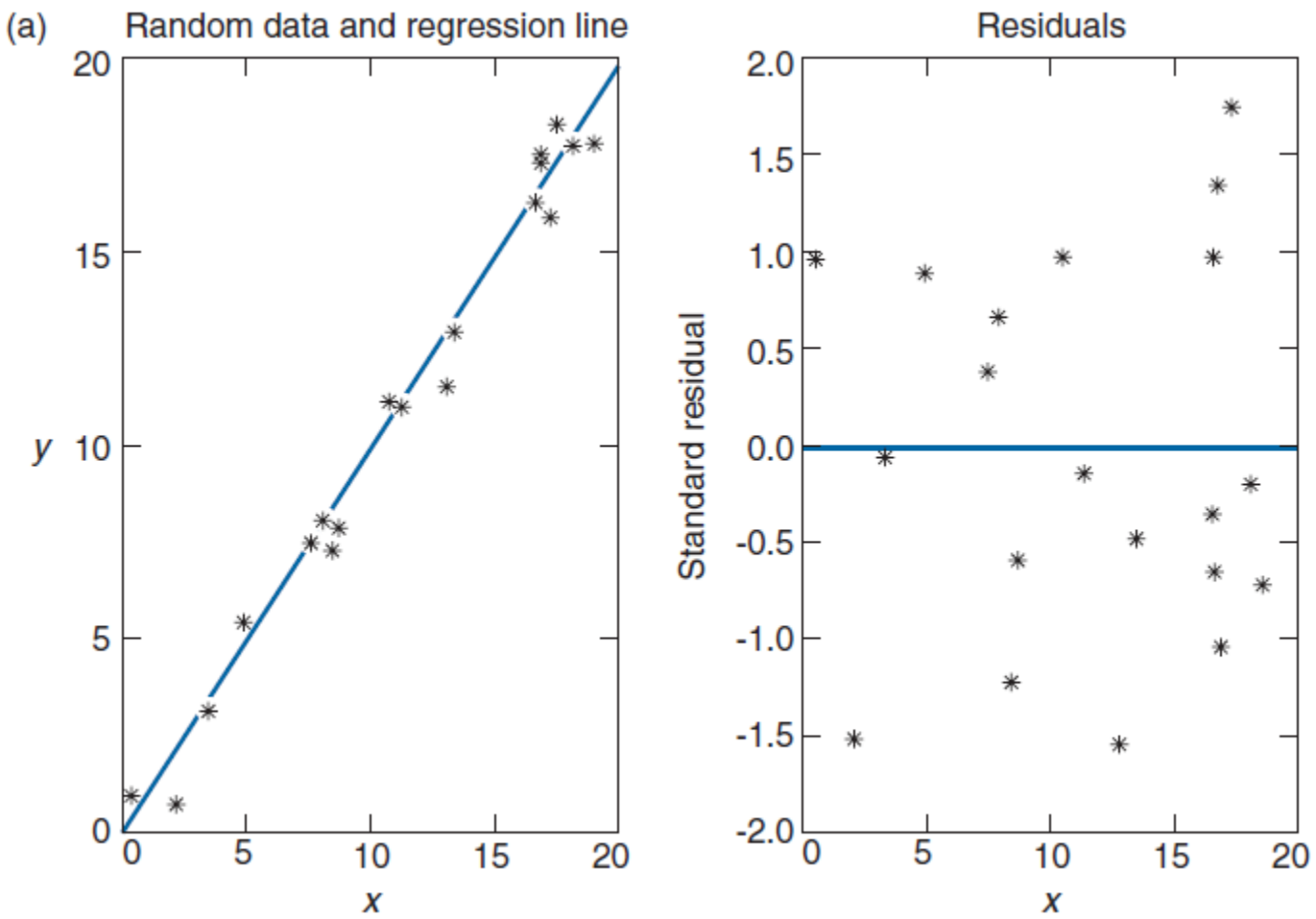
- The analysis begins by normalizing, or standardizing, the residuals by dividing them by $\sqrt{SS_R/(n-2)}$, the estimate of the standard deviation of the Y_i .
- The resulting quantities

$$\frac{Y_i - (A + Bx_i)}{\sqrt{SS_R/(n-2)}}, \qquad i = 1, \ldots, n$$

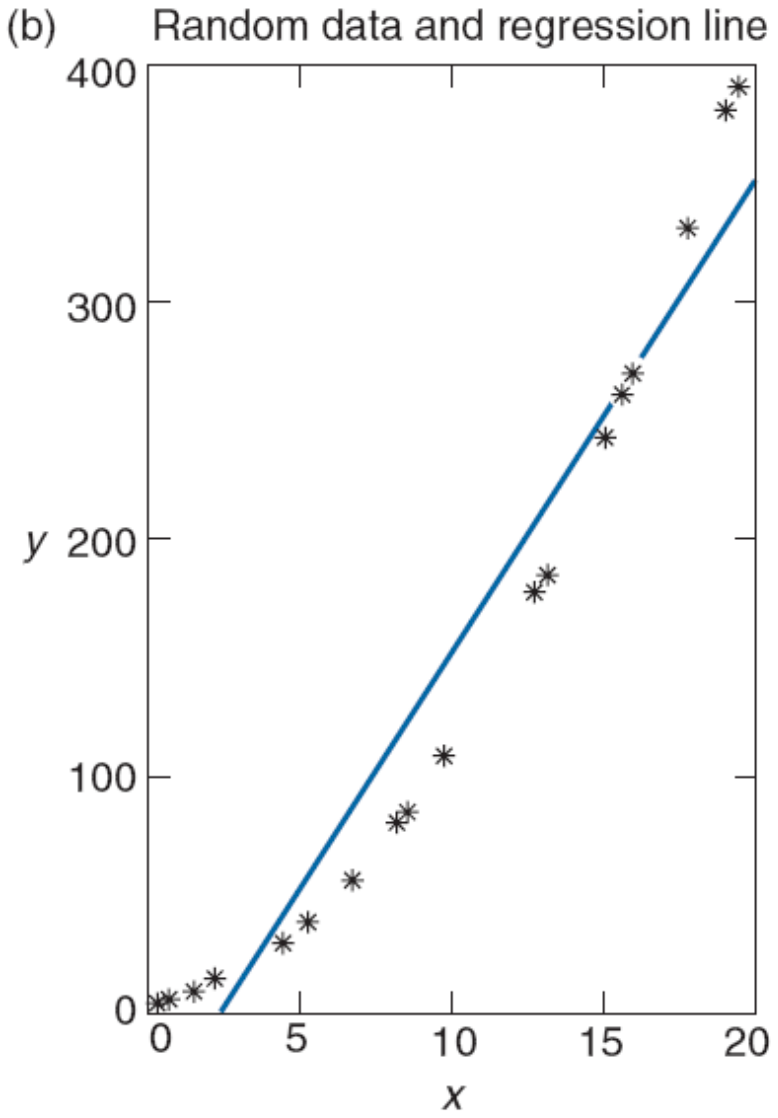
are called the *standardized residuals.*

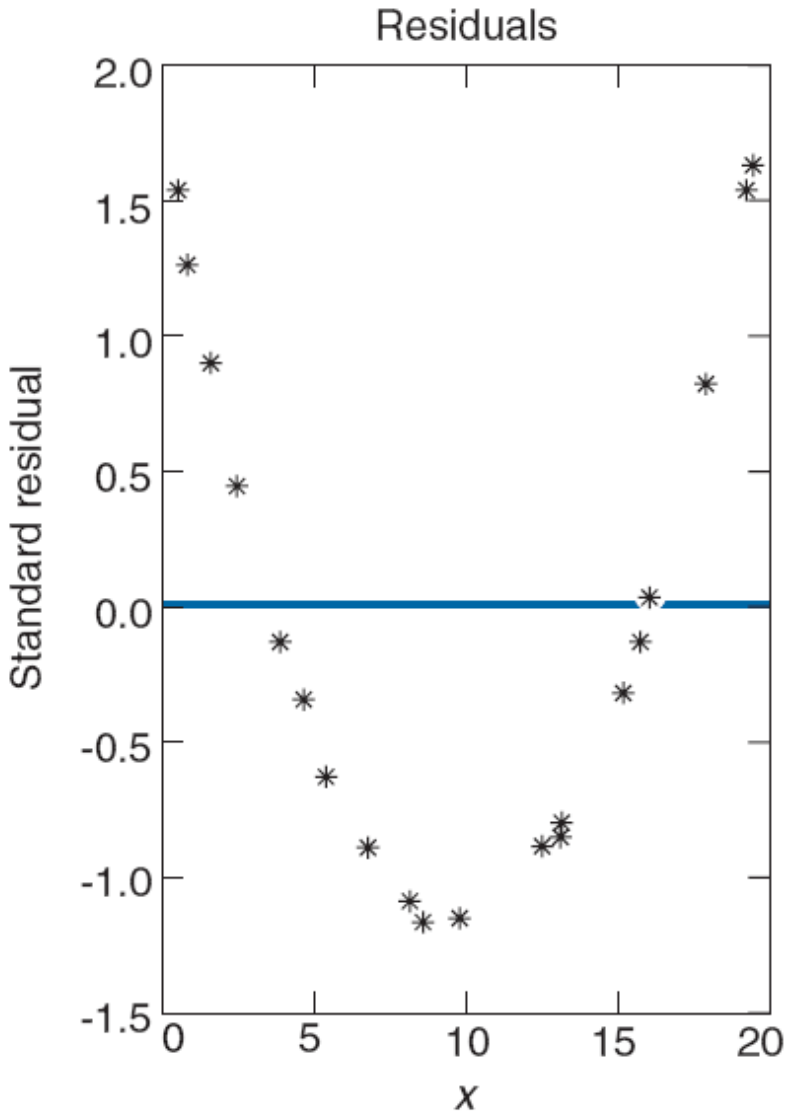
Analysis of Residuals: Assessing the Model (3)

- When the simple linear regression model is correct, the standardized residuals are approximately independent standard normal random variables, and thus should be randomly distributed about 0 with about 95 percent of their values being between –2 and +2 (since $P\{-1.96 < Z < 1.96\} = .95$).
- In addition, a plot of the standardized residuals should not indicate any distinct pattern.
- Indeed, any indication of a distinct pattern should make one suspicious about the validity of the assumed simple linear regression model.

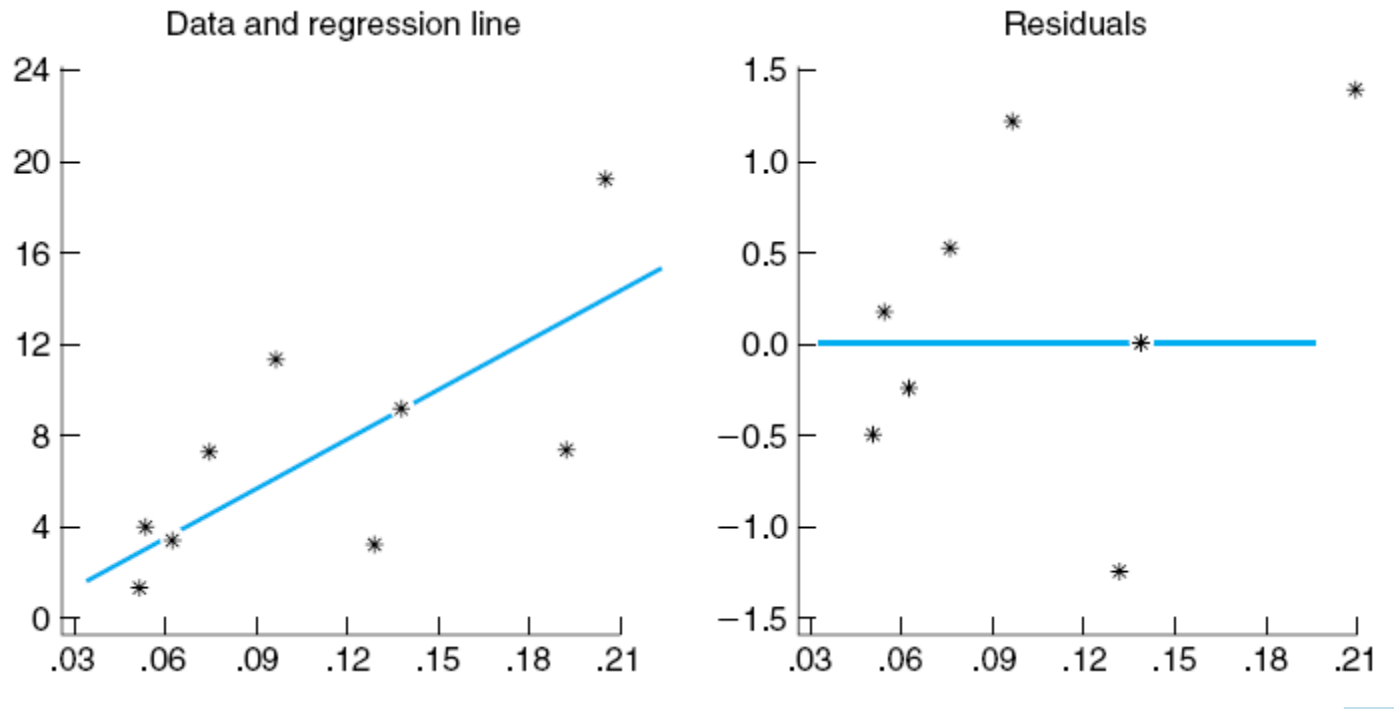


indicated both by its scatter diagram and the random nature of its standardized residuals, appears to fit the straight-line model quite well





This often means that higher-order (than just linear) terms are needed



Increasing as the input level increases; variance is not constant.