Probability & Statistics

for Engineers & Scientists

NINTH EDITION

WALPOLE | MYERS | MYERS | YE

Probability & Statistics for Engineers & Scientists

This page intentionally left blank

Probability & Statistics for Engineers & Scientists

NINTH EDITION

Ronald E. Walpole
Roanoke College

Raymond H. Myers
Virginia Tech

Sharon L. Myers
Radford University

Keying Ye
University of Texas at San Antonio

Prentice Hall

Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editor in Chief: *Deirdre Lynch*
Acquisitions Editor: *Christopher Cummings*
Executive Content Editor: *Christine O'Brien*
Associate Editor: *Christina Lepre*
Senior Managing Editor: *Karen Wernholm*
Senior Production Project Manager: *Tracy Patruno*
Design Manager: *Andrea Nix*
Cover Designer: *Heather Scott*
Digital Assets Manager: *Marianne Groth*
Associate Media Producer: *Vicki Dreyfus*
Marketing Manager: *Alex Gay*
Marketing Assistant: *Kathleen DeChavez*
Senior Author Support/Technology Specialist: *Joe Vetere*
Rights and Permissions Advisor: *Michael Joyce*
Senior Manufacturing Buyer: *Carol Melville*
Production Coordination: *Lifland et al. Bookmakers*
Composition: *Keying Ye*
Cover photo: *Marjory Dressler/Dressler Photo-Graphics*

Library of Congress Cataloging-in-Publication Data

Probability & statistics for engineers & scientists/Ronald E. Walpole ... [et al.] — 9th ed.
p. cm.
ISBN 978-0-321-62911-1
1. Engineering—Statistical methods. 2. Probabilities. I. Walpole, Ronald E.
TA340.P738 2011
519.02'462–dc22

2010004857

1 2 3 4 5 6 7 8 9 10—EB—14 13 12 11 10

Prentice Hall
is an imprint of

PEARSON

www.pearsonhighered.com

ISBN 10: 0-321-62911-6
ISBN 13: 978-0-321-62911-1

Chapter 11

Simple Linear Regression and Correlation

11.1 Introduction to Linear Regression

Often, in practice, one is called upon to solve problems involving sets of variables when it is known that there exists some inherent relationship among the variables. For example, in an industrial situation it may be known that the tar content in the outlet stream in a chemical process is related to the inlet temperature. It may be of interest to develop a method of prediction, that is, a procedure for estimating the tar content for various levels of the inlet temperature from experimental information. Now, of course, it is highly likely that for many example runs in which the inlet temperature is the same, say 130°C, the outlet tar content will not be the same. This is much like what happens when we study several automobiles with the same engine volume. They will not all have the same gas mileage. Houses in the same part of the country that have the same square footage of living space will not all be sold for the same price. Tar content, gas mileage (mpg), and the price of houses (in thousands of dollars) are natural **dependent variables**, or responses, in these three scenarios. Inlet temperature, engine volume (cubic feet), and square feet of living space are, respectively, natural **independent variables**, or **regressors**. A reasonable form of a relationship between the **response** Y and the regressor x is the linear relationship

$$Y = \beta_0 + \beta_1 x,$$

where, of course, β_0 is the **intercept** and β_1 is the **slope**. The relationship is illustrated in Figure 11.1.

If the relationship is exact, then it is a **deterministic** relationship between two scientific variables and there is no random or probabilistic component to it. However, in the examples listed above, as well as in countless other scientific and engineering phenomena, the relationship is not deterministic (i.e., a given x does not always give the same value for Y). As a result, important problems here are probabilistic in nature since the relationship above cannot be viewed as being exact. The concept of **regression analysis** deals with finding the best relationship

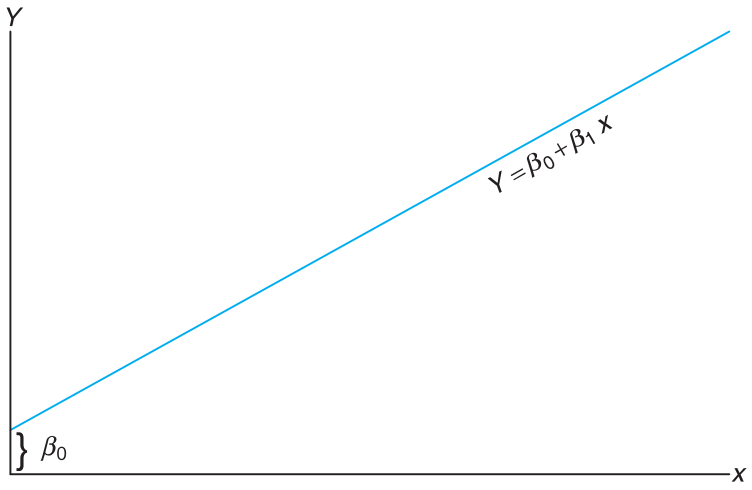

Figure 11.1: A linear relationship; β_0: intercept; β_1: slope.

between Y and x, quantifying the strength of that relationship, and using methods that allow for prediction of the response values given values of the regressor x.

In many applications, there will be more than one regressor (i.e., more than one independent variable **that helps to explain Y**). For example, in the case where the response is the price of a house, one would expect the age of the house to contribute to the explanation of the price, so in this case the multiple regression structure might be written

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where Y is price, x_1 is square footage, and x_2 is age in years. In the next chapter, we will consider problems with multiple regressors. The resulting analysis is termed **multiple regression**, while the analysis of the single regressor case is called **simple regression**. As a second illustration of multiple regression, a chemical engineer may be concerned with the amount of hydrogen lost from samples of a particular metal when the material is placed in storage. In this case, there may be two inputs, storage time x_1 in hours and storage temperature x_2 in degrees centigrade. The response would then be hydrogen loss Y in parts per million.

In this chapter, we deal with the topic of **simple linear regression**, treating only the case of a single regressor variable in which the relationship between y and x is linear. For the case of more than one regressor variable, the reader is referred to Chapter 12. Denote a random sample of size n by the set $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$. If additional samples were taken using exactly the same values of x, we should expect the y values to vary. Hence, the value y_i in the ordered pair (x_i, y_i) is a value of some random variable Y_i.

11.2 The Simple Linear Regression (SLR) Model

We have already confined the terminology *regression analysis* to situations in which relationships among variables are not deterministic (i.e., not exact). In other words, there must be a **random component** to the equation that relates the variables.

This random component takes into account considerations that are not being measured or, in fact, are not understood by the scientists or engineers. Indeed, in most applications of regression, the linear equation, say $Y = \beta_0 + \beta_1 x$, is an approximation that is a simplification of something unknown and much more complicated. For example, in our illustration involving the response $Y =$ tar content and $x =$ inlet temperature, $Y = \beta_0 + \beta_1 x$ is likely a reasonable approximation that may be operative within a confined range on x. More often than not, the models that are simplifications of more complicated and unknown structures are linear in nature (i.e., linear in the **parameters** β_0 and β_1 or, in the case of the model involving the price, size, and age of the house, linear in the **parameters** β_0, β_1, and β_2). These linear structures are simple and empirical in nature and are thus called **empirical models**.

An analysis of the relationship between Y and x requires the statement of a **statistical model**. A model is often used by a statistician as a representation of an **ideal** that essentially defines how we perceive that the data were generated by the system in question. The model must include the set $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$ of data involving n pairs of (x, y) values. One must bear in mind that the value y_i depends on x_i via a linear structure that also has the random component involved. The basis for the use of a statistical model relates to how the random variable Y moves with x and the random component. The model also includes what is assumed about the statistical properties of the random component. The statistical model for simple linear regression is given below. The response Y is related to the independent variable x through the equation

Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

In the above, β_0 and β_1 are unknown intercept and slope parameters, respectively, and ϵ is a random variable that is assumed to be distributed with $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$. The quantity σ^2 is often called the error variance or residual variance.

From the model above, several things become apparent. The quantity Y is a random variable since ϵ is random. The value x of the regressor variable is not random and, in fact, is measured with negligible error. The quantity ϵ, often called a **random error** or **random disturbance**, has constant variance. This portion of the assumptions is often called the **homogeneous variance assumption**. The presence of this random error, ϵ, keeps the model from becoming simply a deterministic equation. Now, the fact that $E(\epsilon) = 0$ implies that at a specific x the y-values are distributed around the **true**, or population, **regression line** $y = \beta_0 + \beta_1 x$. If the model is well chosen (i.e., there are no additional important regressors and the linear approximation is good within the ranges of the data), then positive and negative errors around the true regression are reasonable. We must keep in mind that in practice β_0 and β_1 are not known and must be estimated from data. In addition, the model described above is conceptual in nature. As a result, we never observe the actual ϵ values in practice and thus we can never draw the true regression line (but we assume it is there). We can only draw an estimated line. Figure 11.2 depicts the nature of hypothetical (x, y) data scattered around a true regression line for a case in which only $n = 5$ observations are available. Let us emphasize that what we see in Figure 11.2 is not the line that is used by the

scientist or engineer. Rather, the picture merely describes what the assumptions mean! The regression that the user has at his or her disposal will now be described.

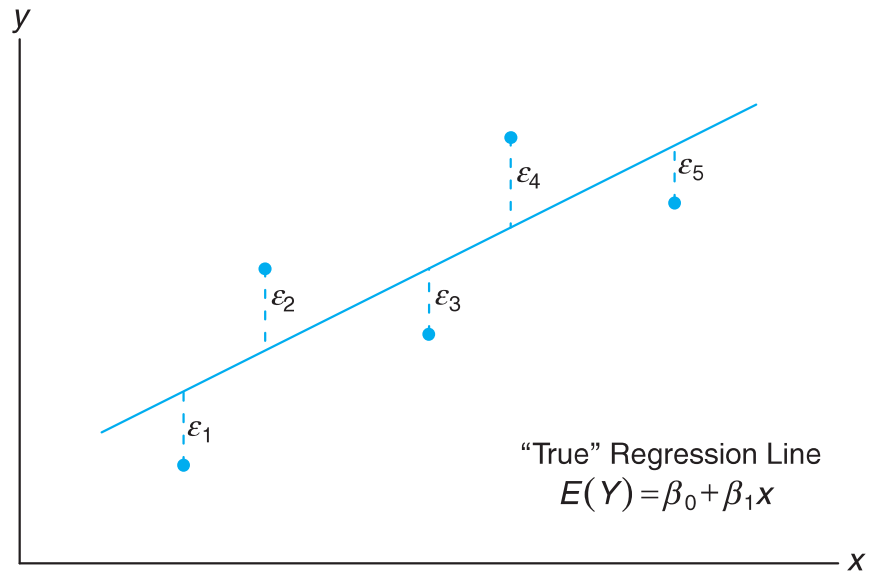

Figure 11.2: Hypothetical (x, y) data scattered around the true regression line for $n = 5$.

The Fitted Regression Line

An important aspect of regression analysis is, very simply, to estimate the parameters β_0 and β_1 (i.e., estimate the so-called **regression coefficients**). The method of estimation will be discussed in the next section. Suppose we denote the estimates b_0 for β_0 and b_1 for β_1. Then the estimated or **fitted regression** line is given by

$$\hat{y} = b_0 + b_1 x,$$

where $\hat{y}$ is the predicted or fitted value. Obviously, the fitted line is an estimate of the true regression line. We expect that the fitted line should be closer to the true regression line when a large amount of data are available. In the following example, we illustrate the fitted line for a real-life pollution study.

One of the more challenging problems confronting the water pollution control field is presented by the tanning industry. Tannery wastes are chemically complex. They are characterized by high values of chemical oxygen demand, volatile solids, and other pollution measures. Consider the experimental data in Table 11.1, which were obtained from 33 samples of chemically treated waste in a study conducted at Virginia Tech. Readings on x, the percent reduction in total solids, and y, the percent reduction in chemical oxygen demand, were recorded.

The data of Table 11.1 are plotted in a **scatter diagram** in Figure 11.3. From an inspection of this scatter diagram, it can be seen that the points closely follow a straight line, indicating that the assumption of linearity between the two variables appears to be reasonable.

Table 11.1: Measures of Reduction in Solids and Oxygen Demand

Solids Reduction, x (%)	**Oxygen Demand Reduction, y (%)**	**Solids Reduction, x (%)**	**Oxygen Demand Reduction, y (%)**
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

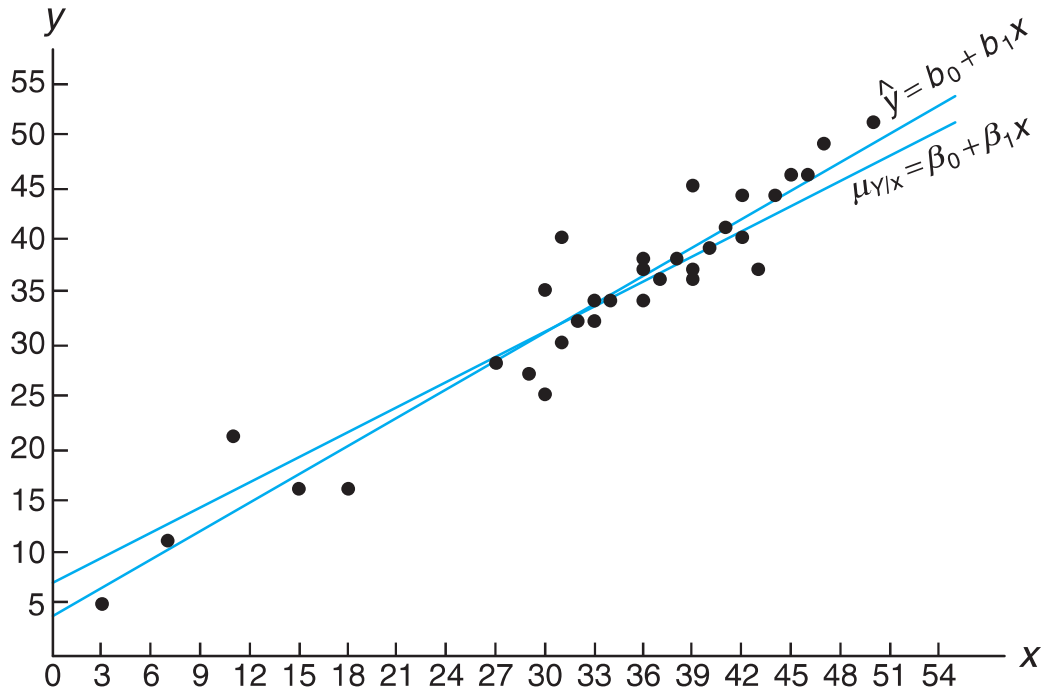

Figure 11.3: Scatter diagram with regression lines.

The fitted regression line and a *hypothetical true regression line* are shown on the scatter diagram of Figure 11.3. This example will be revisited as we move on to the method of estimation, discussed in Section 11.3.

Another Look at the Model Assumptions

It may be instructive to revisit the simple linear regression model presented previously and discuss in a graphical sense how it relates to the so-called true regression. Let us expand on Figure 11.2 by illustrating not merely where the ϵ_i fall on a graph but also what the implication is of the normality assumption on the ϵ_i.

Suppose we have a simple linear regression with $n = 6$ evenly spaced values of x and a single y-value at each x. Consider the graph in Figure 11.4. This illustration should give the reader a clear representation of the model and the assumptions involved. The line in the graph is the true regression line. The points plotted are actual (y, x) points which are scattered about the line. Each point is on its own normal distribution with the center of the distribution (i.e., the mean of y) falling on the line. This is certainly expected since $E(Y) = \beta_0 + \beta_1 x$. As a result, the true regression line **goes through the means of the response**, and the actual observations are on the distribution around the means. Note also that all distributions have the same variance, which we referred to as σ^2. Of course, the deviation between an individual y and the point on the line will be its individual ϵ value. This is clear since

$$y_i - E(Y_i) = y_i - (\beta_0 + \beta_1 x_i) = \epsilon_i.$$

Thus, at a given x, Y and the corresponding ϵ both have variance σ^2.

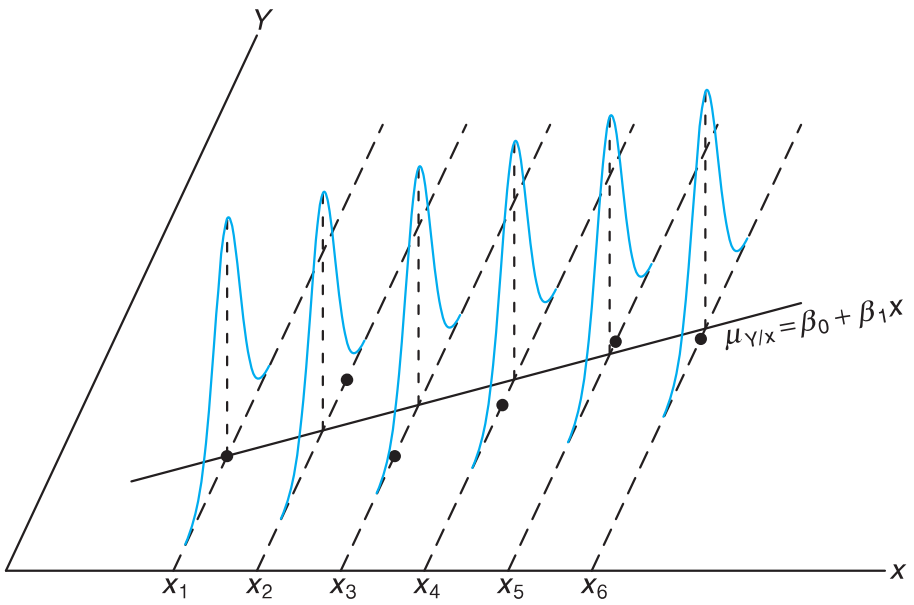

Figure 11.4: Individual observations around true regression line.

Note also that we have written the true regression line here as $\mu_{Y|x} = \beta_0 + \beta_1 x$ in order to reaffirm that the line goes through the mean of the Y random variable.

11.3 Least Squares and the Fitted Model

In this section, we discuss the method of fitting an estimated regression line to the data. This is tantamount to the determination of estimates b_0 for β_0 and b_1

for β_1. This of course allows for the computation of predicted values from the fitted line $\hat{y} = b_0 + b_1 x$ and other types of analyses and diagnostic information that will ascertain the strength of the relationship and the adequacy of the fitted model. Before we discuss the method of least squares estimation, it is important to introduce the concept of a **residual**. A residual is essentially an error in the fit of the model $\hat{y} = b_0 + b_1 x$.

Residual: Error in Fit

Given a set of regression data $\{(x_i, y_i); i = 1, 2, \ldots, n\}$ and a fitted model, $\hat{y}_i = b_0 + b_1 x_i$, the ith residual e_i is given by

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n.$$

Obviously, if a set of n residuals is large, then the fit of the model is not good. Small residuals are a sign of a good fit. Another interesting relationship which is useful at times is the following:

$$y_i = b_0 + b_1 x_i + e_i.$$

The use of the above equation should result in clarification of the distinction between the residuals, e_i, and the conceptual model errors, ϵ_i. One must bear in mind that whereas the ϵ_i are not observed, the e_i not only are observed but also play an important role in the total analysis.

Figure 11.5 depicts the line fit to this set of data, namely $\hat{y} = b_0 + b_1 x$, and the line reflecting the model $\mu_{Y|x} = \beta_0 + \beta_1 x$. Now, of course, β_0 and β_1 are unknown parameters. The fitted line is an estimate of the line produced by the statistical model. Keep in mind that the line $\mu_{Y|x} = \beta_0 + \beta_1 x$ is not known.

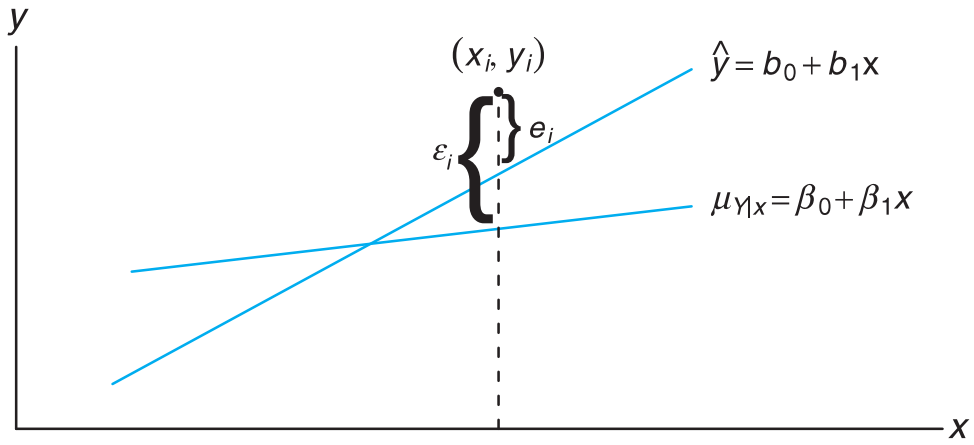

Figure 11.5: Comparing ϵ_i with the residual, e_i.

The Method of Least Squares

We shall find b_0 and b_1, the estimates of β_0 and β_1, so that the sum of the squares of the residuals is a minimum. The residual sum of squares is often called the sum of squares of the errors about the regression line and is denoted by SSE. This

minimization procedure for estimating the parameters is called the **method of least squares**. Hence, we shall find a and b so as to minimize

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

Differentiating SSE with respect to b_0 and b_1, we have

$$\frac{\partial(SSE)}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i), \quad \frac{\partial(SSE)}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i.$$

Setting the partial derivatives equal to zero and rearranging the terms, we obtain the equations (called the **normal equations**)

$$nb_0 + b_1\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \quad b_0\sum_{i=1}^{n} x_i + b_1\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

which may be solved simultaneously to yield computing formulas for b_0 and b_1.

Estimating the Regression Coefficients

Given the sample $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$, the least squares estimates b_0 and b_1 of the regression coefficients β_0 and β_1 are computed from the formulas

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ and}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1\bar{x}.$$

The calculations of b_0 and b_1, using the data of Table 11.1, are illustrated by the following example.

Example 11.1: Estimate the regression line for the pollution data of Table 11.1.

***Solution*:**

$$\sum_{i=1}^{33} x_i = 1104, \quad \sum_{i=1}^{33} y_i = 1124, \quad \sum_{i=1}^{33} x_i y_i = 41{,}355, \quad \sum_{i=1}^{33} x_i^2 = 41{,}086$$

Therefore,

$$b_1 = \frac{(33)(41{,}355) - (1104)(1124)}{(33)(41{,}086) - (1104)^2} = 0.903643 \text{ and}$$

$$b_0 = \frac{1124 - (0.903643)(1104)}{33} = 3.829633.$$

Thus, the estimated regression line is given by

$$\hat{y} = 3.8296 + 0.9036x.$$

Using the regression line of Example 11.1, we would predict a 31% reduction in the chemical oxygen demand when the reduction in the total solids is 30%. The

31% reduction in the chemical oxygen demand may be interpreted as an estimate of the population mean $\mu_{Y|30}$ or as an estimate of a new observation when the reduction in total solids is 30%. Such estimates, however, are subject to error. Even if the experiment were controlled so that the reduction in total solids was 30%, it is unlikely that we would measure a reduction in the chemical oxygen demand exactly equal to 31%. In fact, the original data recorded in Table 11.1 show that measurements of 25% and 35% were recorded for the reduction in oxygen demand when the reduction in total solids was kept at 30%.

What Is Good about Least Squares?

It should be noted that the least squares criterion is designed to provide a fitted line that results in a "closeness" between the line and the plotted points. There are many ways of measuring closeness. For example, one may wish to determine b_0 and b_1 for which $\sum_{i=1}^{n} |y_i - \hat{y}_i|$ is minimized or for which $\sum_{i=1}^{n} |y_i - \hat{y}_i|^{1.5}$ is minimized. These are both viable and reasonable methods. Note that both of these, as well as the least squares procedure, result in forcing residuals to be "small" in some sense. One should remember that the residuals are the empirical counterpart to the ϵ values. Figure 11.6 illustrates a set of residuals. One should note that the fitted line has predicted values as points on the line and hence the residuals are vertical deviations from points to the line. As a result, the least squares procedure produces a line that **minimizes the sum of squares of vertical deviations** from the points to the line.

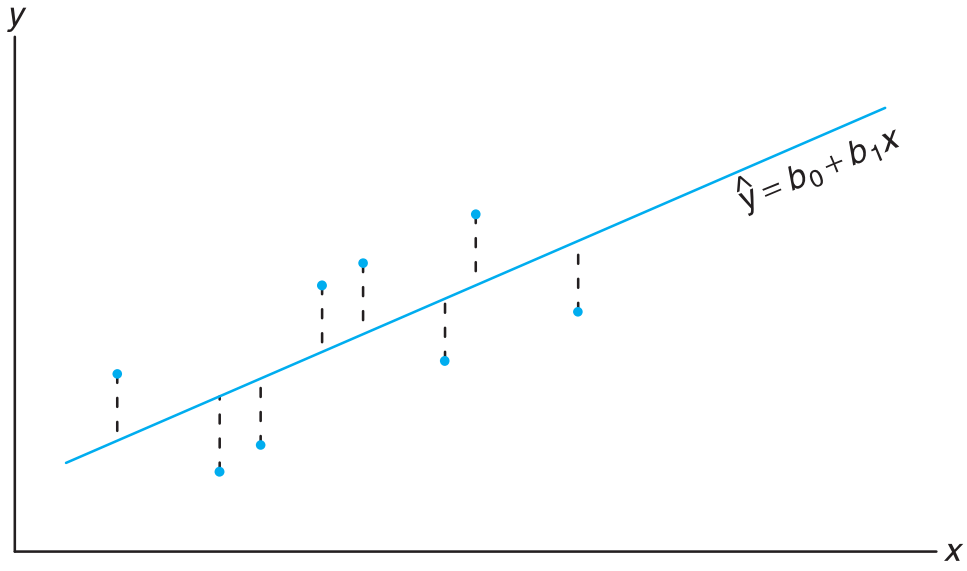

Figure 11.6: Residuals as vertical deviations.

Exercises

11.1 A study was conducted at Virginia Tech to determine if certain static arm-strength measures have an influence on the "dynamic lift" characteristics of an individual. Twenty-five individuals were subjected to strength tests and then were asked to perform a weight-lifting test in which weight was dynamically lifted overhead. The data are given here.

Individual	Arm Strength, x	Dynamic Lift, y
1	17.3	71.7
2	19.3	48.3
3	19.5	88.3
4	19.7	75.0
5	22.9	91.7
6	23.1	100.0
7	26.4	73.3
8	26.8	65.0
9	27.6	75.0
10	28.1	88.3
11	28.2	68.3
12	28.7	96.7
13	29.0	76.7
14	29.6	78.3
15	29.9	60.0
16	29.9	71.7
17	30.3	85.0
18	31.3	85.0
19	36.0	88.3
20	39.5	100.0
21	40.4	100.0
22	44.3	100.0
23	44.6	91.7
24	50.4	100.0
25	55.9	71.7

(a) Estimate β_0 and β_1 for the linear regression curve $\mu_{Y|x} = \beta_0 + \beta_1 x$.

(b) Find a point estimate of $\mu_{Y|30}$.

(c) Plot the residuals versus the x's (arm strength). Comment.

11.2 The grades of a class of 9 students on a midterm report (x) and on the final examination (y) are as follows:

x	77	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	68

(a) Estimate the linear regression line.

(b) Estimate the final examination grade of a student who received a grade of 85 on the midterm report.

11.3 The amounts of a chemical compound y that dissolved in 100 grams of water at various temperatures x were recorded as follows:

x (°C)	y (grams)		
0	8	6	8
15	12	10	14
30	25	21	24
45	31	33	28
60	44	39	42
75	48	51	44

(a) Find the equation of the regression line.

(b) Graph the line on a scatter diagram.

(c) Estimate the amount of chemical that will dissolve in 100 grams of water at 50°C.

11.4 The following data were collected to determine the relationship between pressure and the corresponding scale reading for the purpose of calibration.

Pressure, x (lb/sq in.)	Scale Reading, y
10	13
10	18
10	16
10	15
10	20
50	86
50	90
50	88
50	88
50	92

(a) Find the equation of the regression line.

(b) The purpose of calibration in this application is to estimate pressure from an observed scale reading. Estimate the pressure for a scale reading of 54 using $\hat{x} = (54 - b_0)/b_1$.

11.5 A study was made on the amount of converted sugar in a certain process at various temperatures. The data were coded and recorded as follows:

Temperature, x	Converted Sugar, y
1.0	8.1
1.1	7.8
1.2	8.5
1.3	9.8
1.4	9.5
1.5	8.9
1.6	8.6
1.7	10.2
1.8	9.3
1.9	9.2
2.0	10.5

(a) Estimate the linear regression line.

(b) Estimate the mean amount of converted sugar produced when the coded temperature is 1.75.

(c) Plot the residuals versus temperature. Comment.