Chapter 2: Introduction to Regression Analysis

- **Regression analysis** is a statistical technique for investigating and modeling the relationship between variables.
- Equation of a straight line (classical)

$$y = mx + b$$

deterministic versus
probabilistic relationship

we usually write this as

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

parameters are fixed
usually unknown quantities

Modeling a Response

- Not all observations will fall exactly on a straight line.

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

where ε represents error

- it is a random variable that accounts for the failure of the model to fit the data *exactly*.
- $\varepsilon \sim N(0, \sigma^2)$

Regression and Model Building

Delivery time example

If we let y represent delivery time and x represent delivery volume, then the equation of a straight line relating these two variables is

$$y = \beta_0 + \beta_1 x \qquad (1.1)$$

true regression line

Regression and Model Building

- Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (1.2)$$

where y – dependent (response) variable

x – independent (regressor/predictor) variable

$\boxed{\text{covariate}}$

β_0 - intercept

β_1 - slope

ε - random error term

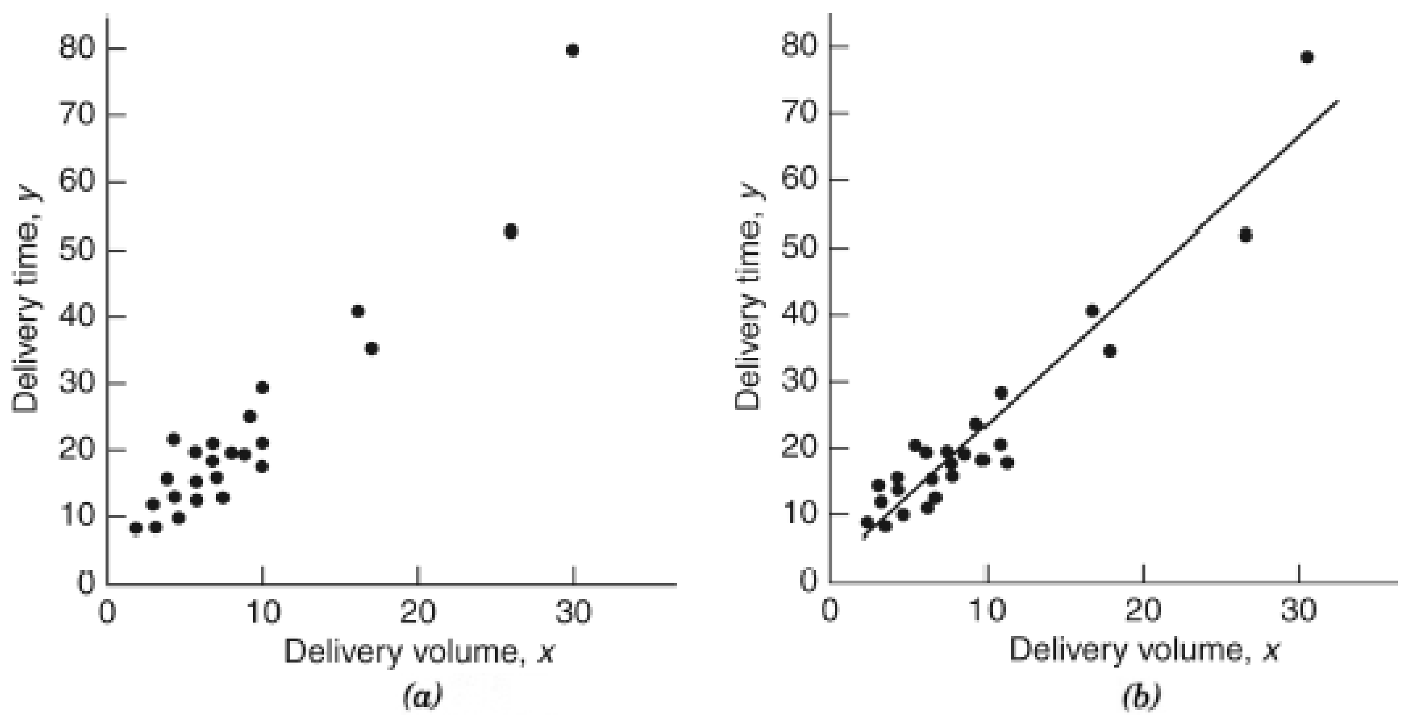

Figure 1.1 (*a*) Scatter diagram for delivery volume. (*b*) Straight-line relationship between delivery time and delivery volume.

Regression and Model Building

- **The** **mean response** at any value, *x*, of the regressor variable is

expectation is a linear operator

$$E(y|x) = \mu_{y|x} = E(\beta_0 + \beta x + \varepsilon) = \beta_0 + \beta_1 x$$

- The **variance** of y at any given *x* is

$$\mathrm{Var}(y|x) = \sigma^2_{y|x} = \mathrm{Var}(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2$$

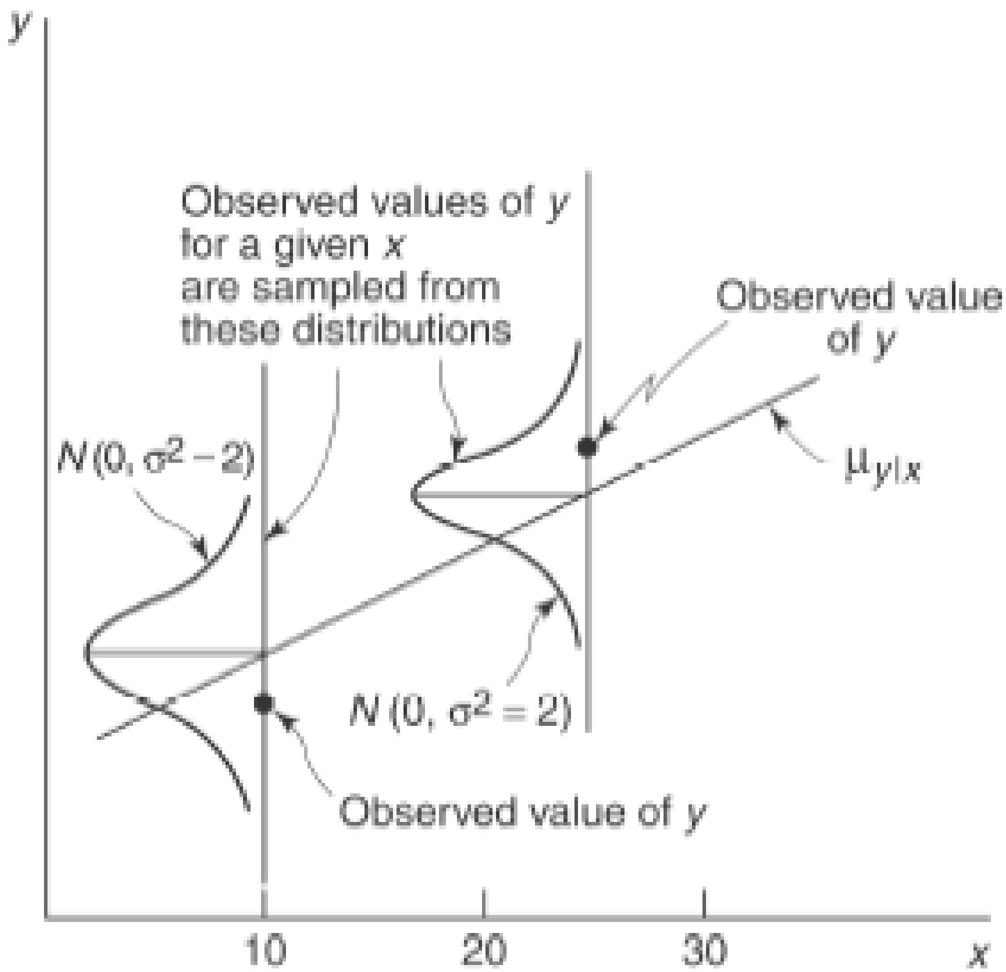

Figure 1.2 How observations are generated in linear regression.

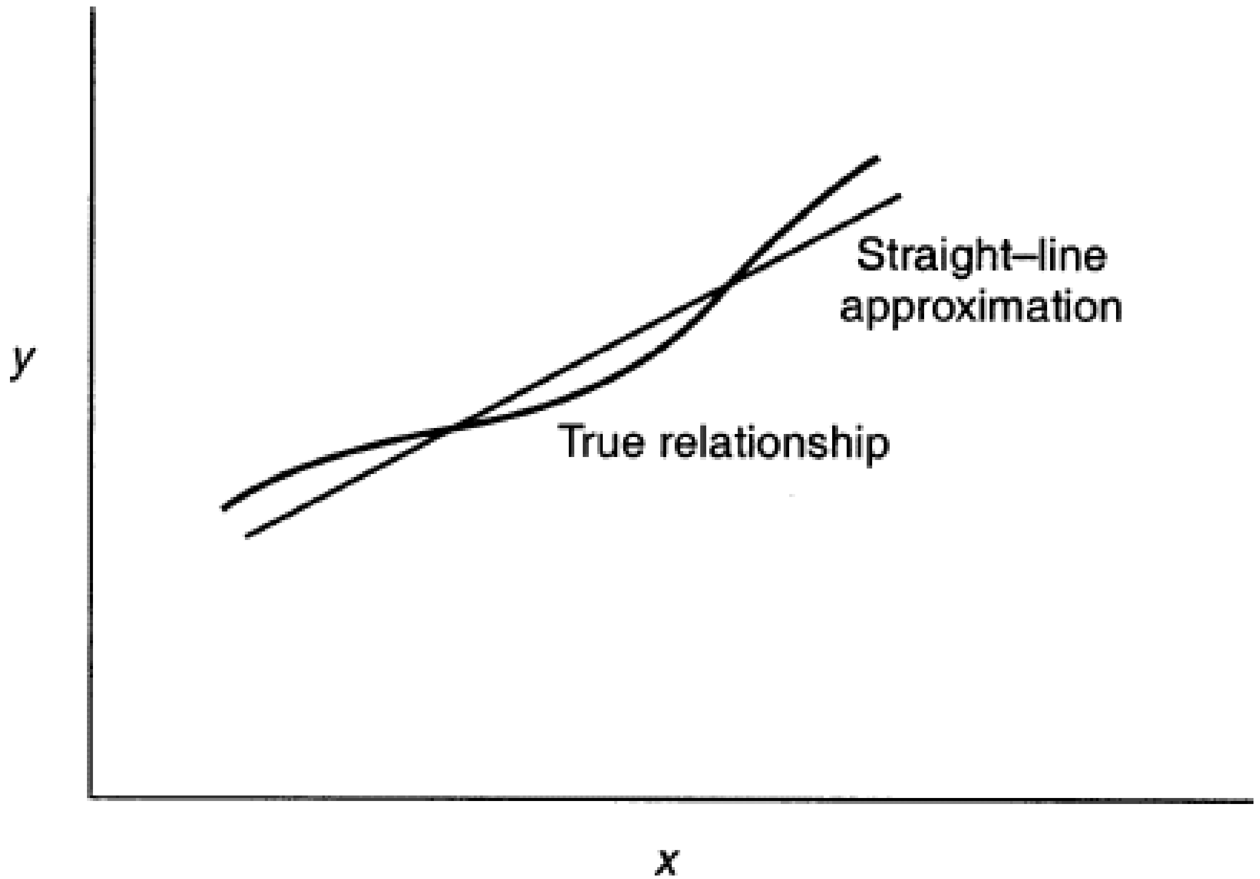

Figure 1.3 Linear regression approximation of a complex relationship.

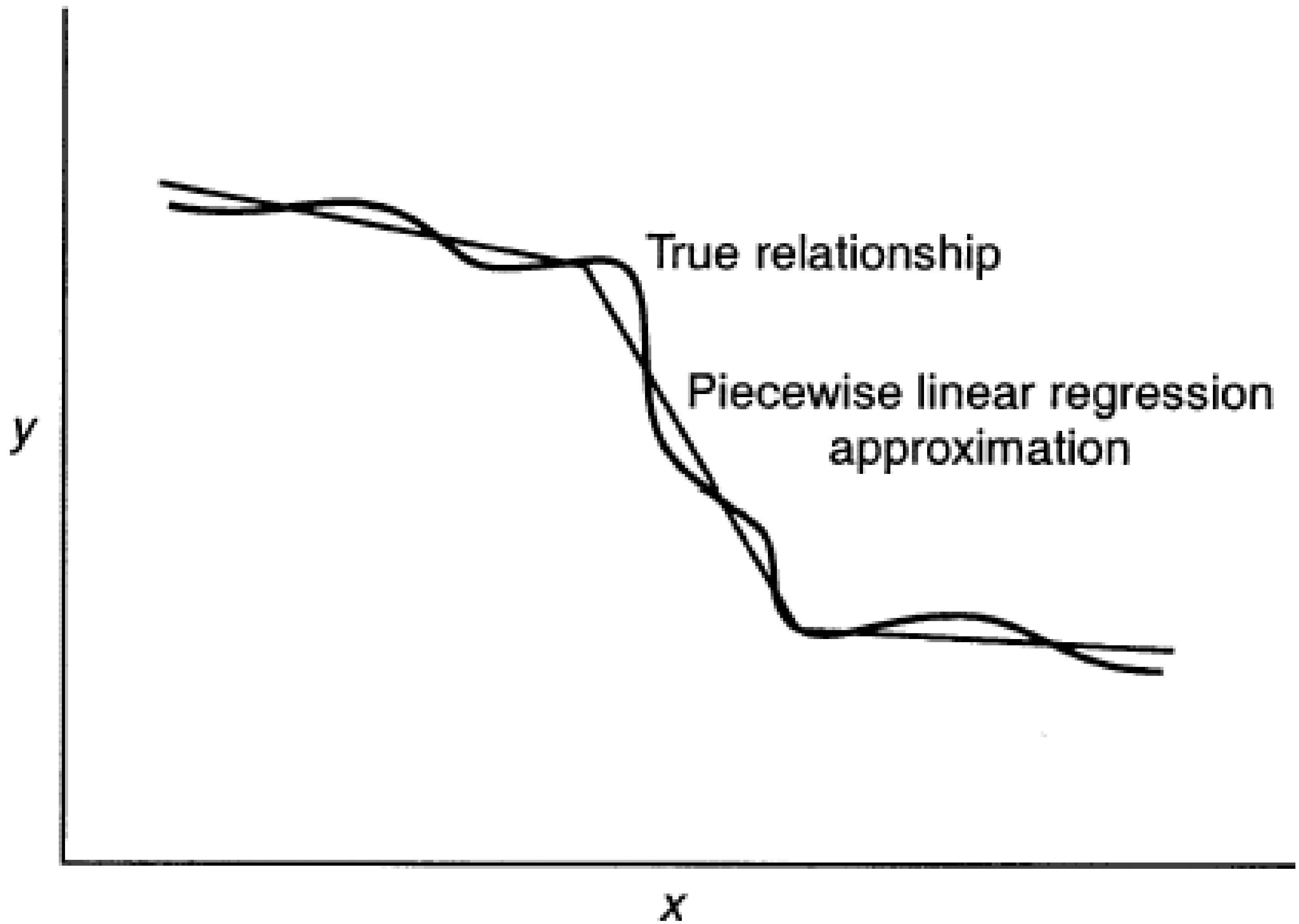

Figure 1.4 Piecewise linear approximation of a complex relationship.

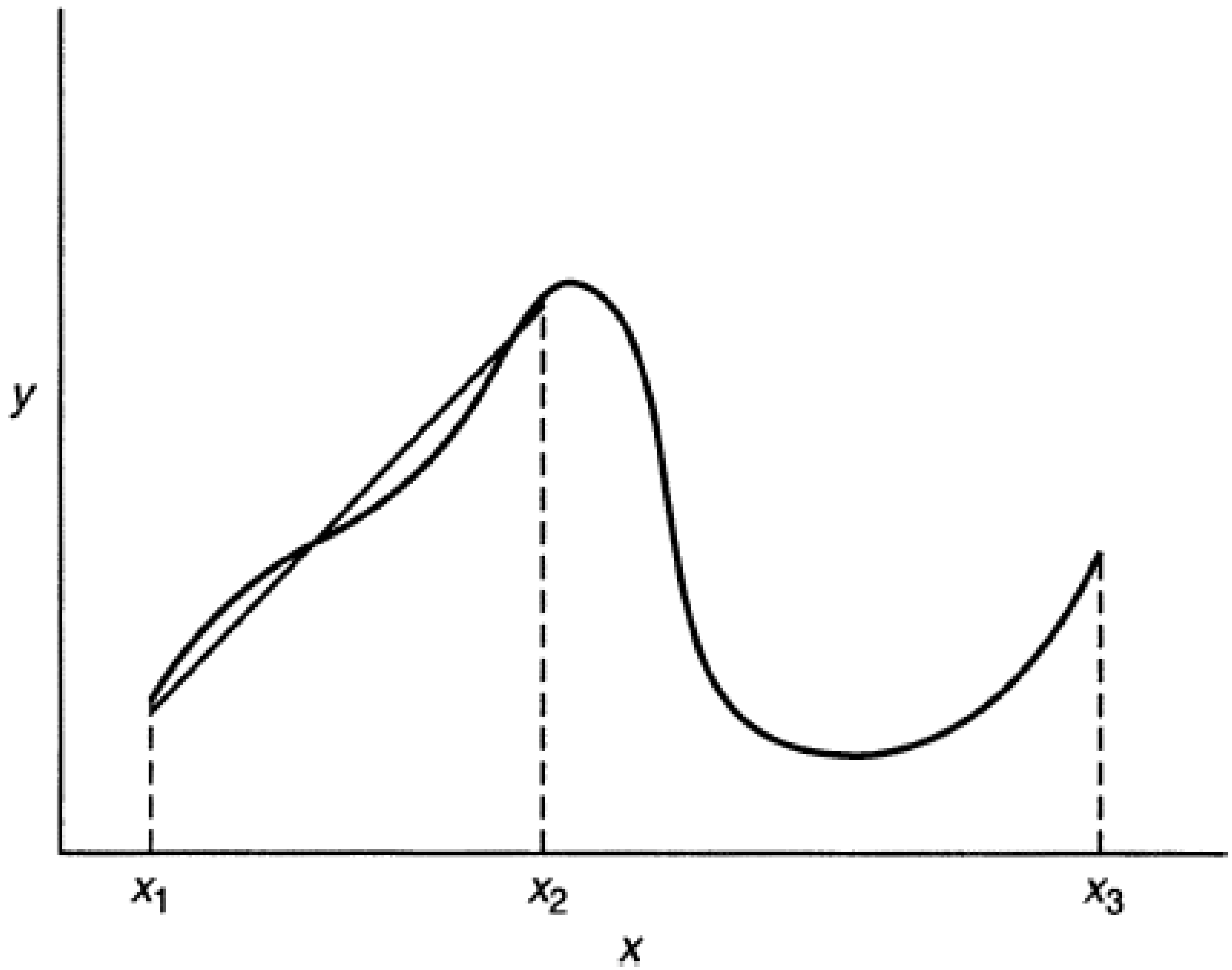

Figure 1.5 The danger of extrapolation in regression.

Regression and Model Building

"Essentially, all models are wrong, but some are useful" Since model building is the essence of science, this quote has a bit of a bite to it. It is from George E. P. Box (1919 – 2013), who was not only an eminent statistician but also an eminently quotable one.

- Multiple Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (1.3)$$

Regression analysis is perhaps the most widely used statistical technique, and probably the most widely misused.

Cause and Effect Relationships

- Caution: just because you *can* fit a linear model to a set of data, does not mean you should.
- It is relatively easy to build “nonsense” relationships between variables
- Regression does not necessarily imply causality

Chapter 3: Simple Linear Regression

Simple Linear Regression Model

- Single regressor, *x*; response, *y*

Population regression model

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

- β_0 – intercept: if $x = 0$ is in the range, then β_0 is the mean of the distribution of the response *y*, when $x = 0$; if $x = 0$ is not in the range, then β_0 has no practical interpretation

If x=0 is in the scope of the model then Bo has practical interpretation.

- β_1 – slope: change in the mean of the distribution of the response produced by a unit change in *x*
- ε - random error

Simple Linear Regression Model

- The response, y, is a random variable
- There is a probability distribution for y at each value of x
  - Mean:

$$E(y \mid x) = \beta_0 + \beta_1 x$$

  - Variance:

$$Var(y \mid x) = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2$$

Least-Squares Estimation of the Parameters

- β_0 and β_1 are unknown and must be estimated

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \ldots, n$$

Sample regression model

Graphical depiction?

- Least squares estimation seeks to minimize the *sum of squares* of the differences between the observed response, y_i, and the straight line.

$$S(\beta_0, \beta_1) = \sum_i \varepsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

What is the LS estimator of mu in univariate statistics?
Answer: Xbar

Least-Squares Estimation of the Parameters

- Let $\hat{\beta}_0, \hat{\beta}_1$ represent the least squares estimators of β_0 and β_1, respectively.
- These estimators must satisfy:

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Least-Squares Estimation of the Parameters

- Simplifying yields the least squares **normal equations**:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Least-Squares Estimation of the Parameters

- Solving the normal equations yields the ordinary least squares estimators:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

Under the assumption of normal error terms, these are the LS estimators and also the Maximum Likelihood estimators.

Computation formulas - can be easily programmed in Excel, for example

Least-Squares Estimation of the Parameters

- The fitted simple linear regression model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Sum of Squares Notation:

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

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = \sum_{i=1}^{n} y_i \left(x_i - \overline{x}\right)$$

Least-Squares Estimation of the Parameters

- Then

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

Least-Squares Estimation of the Parameters

- Residuals: $e_i = y_i - \hat{y}_i$

observed - fitted
data - fit

- Residuals will be used to determine the **adequacy** of the model

Example 2.1– The Rocket Propellant Data

TABLE 2.1 Data for Example 2.1

Observation i	Shear Strength (psi) y_i	Age of Propellant (weeks) x_i
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.50
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2256.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.50

Description: igniter propellent and a sustainer propellant are bonded to manufacture a rocket motor. Shear strength of the bond of the two types of propellant is an important quality characteristic.

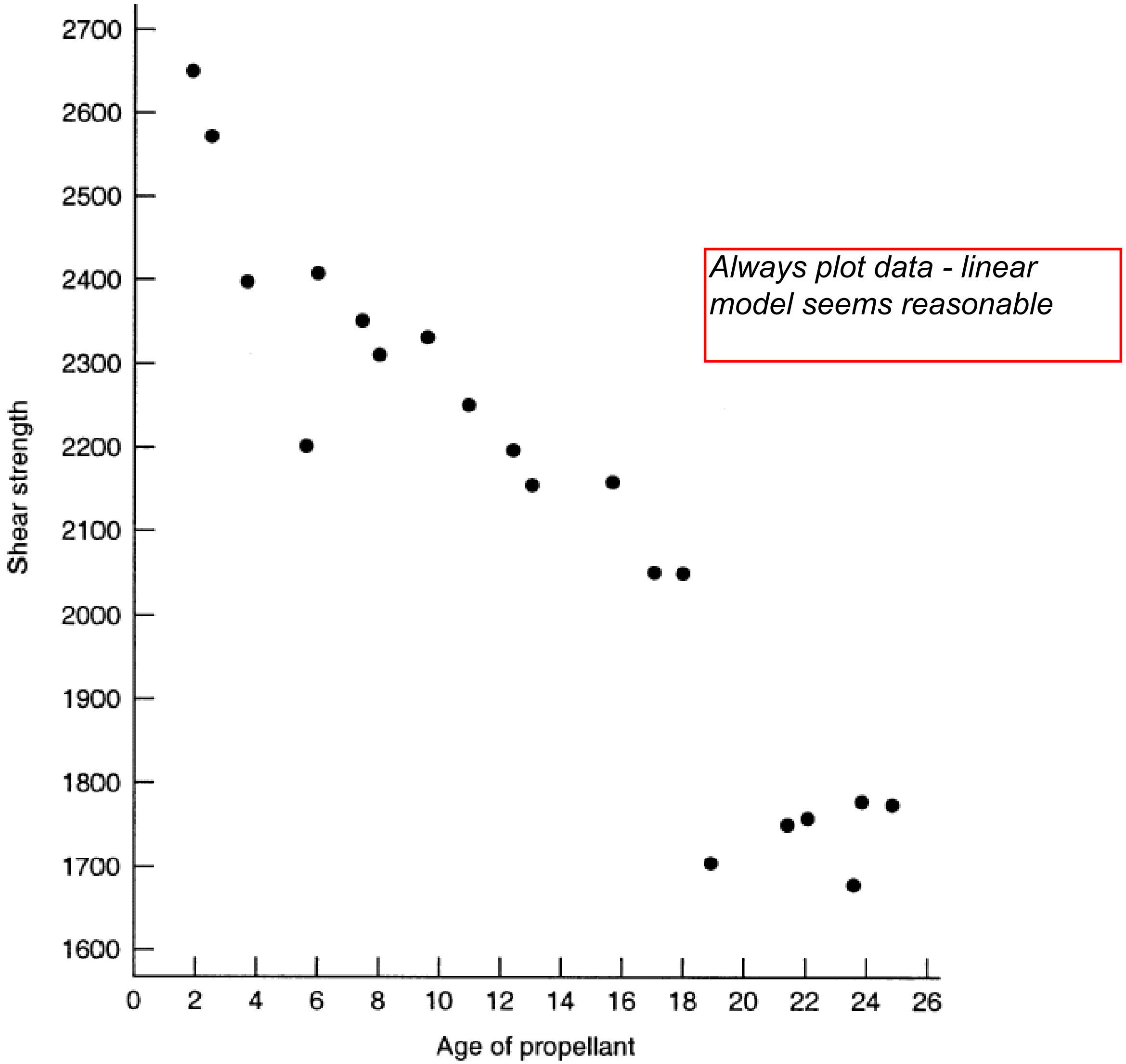

Figure 2.1 Scatter diagram of shear strength versus propellant age. Example 2.1.

Example 2.1- Rocket Propellant Data

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 4677.69 - \frac{71{,}422.56}{20} = 1106.56$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} = 528{,}492.64 - \frac{(267.25)(42{,}627.15)}{20}$$

$$= -41{,}112.65$$

Example 2.1- Rocket Propellant Data

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-41,112.65}{1106.56} = -37.15$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2131.3575 - (-37.15)13.3625 = 2627.82$$

- The least squares regression line is

$$\hat{y} = 2627.82 - 37.15x$$

SAS program and SAS output shown later

TABLE 2.2 Data, Fitted Values, and Residuals for Example 2.1

Observed Value, y_i	Fitted Value, $\hat{y}_i$	Residual, e_i
2158.70	2051.94	106.76
1678.15	1745.42	−67.27
2316.00	2330.59	−14.59
2061.30	1996.21	65.09
2207.50	2423.48	−215.98
1708.30	1921.90	−213.60
1784.70	1736.14	48.56
2575.00	2534.94	40.06
2357.90	2349.17	8.73
2256.70	2219.13	37.57
2165.20	2144.83	20.37
2399.55	2488.50	−88.95
1799.80	1698.98	80.82
2336.75	2265.58	71.17
1765.30	1810.44	−45.14
2053.50	1959.06	94.44
2414.40	2404.90	9.50
2200.50	2163.40	37.10
2654.20	2553.52	100.68
1753.70	1829.02	−75.32
$\Sigma y_i = 42627.15$	$\Sigma \hat{y}_i = 42627.15$	$\boxed{\Sigma e_i = 0.00}$

Least-Squares Estimation of the Parameters

- Just because we can fit a linear model doesn't mean that we should
  - How well does this equation fit the data?
  - Is the model likely to be useful as a predictor?
  - Are any of the basic assumptions (such as constant variance and uncorrelated errors) violated, if so, how serious is this?

Least-Squares Estimation of the Parameters

Computer Output (Minitab)

We will learn how to populate the ANOVA table below

TABLE 2.3 MINITAB Regression Output for Example 2.1

Regression Analysis

When p is low, H0 must go

The regression equation is
Strength = 2628 - 37.2 Age

Predictor	Coef	StDev	T	P
Constant	2627.82	44.18	59.47	0.000
Age	-37.154	2.889	-12.86	0.000

S = 96.11 R-Sq = 90.2% R-Sq(adj) = 89.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1527483	1527483	165.38	0.000
Error	18	166255	9236		
Total	19	1693738			

Properties of the Least-Squares Estimators and the Fitted Regression Model

- The ordinary least-squares (OLS) estimator of the slope is a linear combinations of the observations, y_i:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i$$

Useful in showing expected value and variance properties

where

$$c_i = (x_i - \overline{x}) / S_{xx}, \sum_{i=1}^{n} c_i = 0, \sum_{i=1}^{n} c_i^2 = 1$$

Properties of the Least-Squares Estimators and the Fitted Regression Model

- The least-squares estimators are **unbiased estimators** of their respective parameter:

$$E\left(\hat{\beta}_1\right) = \beta_1 \qquad E\left(\hat{\beta}_0\right) = \beta_0$$

- The variances are

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}} \quad Var\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)$$

- The OLS estimators are **Best Linear Unbiased Estimators** (BLUE)

Gauss-Markov theorem

Properties of the Least-Squares Estimators and the Fitted Regression Model

- Useful properties of the least-squares fit

1. $\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$

2. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$

3. The least-squares regression line always passes through the centroid $(\bar{y}, \bar{x})$ of the data.

4. $\sum_{i=1}^{n} x_i e_i = 0$ 5. $\sum_{i=1}^{n} \hat{y}_i e_i = 0$

Should be (xbar,ybar)

An estimate of sigma**2 is needed for making inferences

- Residual (error) sum of squares

$$SS_{\mathrm{Res}} = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} e_i^2$$

$$= \underbrace{\sum_{i=1}^{n} y_i^2 - n\bar{y}}_{SS_T = \sum\left(y_i - \bar{y}\right)^2} - \hat{\beta}_1 S_{xy}$$

$$= SS_T - \hat{\beta}_1 S_{xy}$$

- **<u>Unbiased estimator</u>** of σ^2

$$\hat{\sigma}^2 = \frac{SS_{\mathrm{Res}}}{n-2} = MS_{\mathrm{Res}}$$

- The quantity $n-2$ is the number of degrees of freedom for the residual sum of squares.

- $\hat{\sigma}^2$ depends on the residual sum of squares. Then:
  - Any violation of the assumptions on the model errors could damage the usefulness of this estimate
  - A misspecification of the model can damage the usefulness of this estimate
  - This estimate is **model dependent**

Go to SAS Example 2.1

Hypothesis Testing on the Slope and Intercept

- Three assumptions needed to apply procedures such as **hypothesis testing** and **confidence intervals**. Model errors, ε_i,
  - are normally distributed
  - are independently distributed
  - have constant variance

i.e. $\varepsilon_i \sim NID(0, \sigma^2)$

Use of t-tests

Slope

H_0: $\beta_1 = \beta_{10}$ H_1: $\beta_1 \neq \beta_{10}$

- Standard error of the slope: $se\left(\hat{\beta}_1\right) = \sqrt{\dfrac{MS_{\mathrm{Res}}}{S_{xx}}}$
- Test statistic: $t_0 = \dfrac{\hat{\beta}_1 - \beta_{10}}{se\left(\hat{\beta}_1\right)}$
- Reject H_0 if $|t_0| > t_{\alpha/2,n-2}$
- Can also use the *P*-value approach

Use of t-tests

Intercept

H_0: $\beta_0 = \beta_{00}$ H_1: $\beta_0 \neq \beta_{00}$

- Standard error of the intercept: $se\left(\hat{\beta}_0\right) = \sqrt{MS_{\mathrm{Res}}\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}$
- Test statistic: $t_0 = \dfrac{\hat{\beta}_0 - \beta_{00}}{se\left(\hat{\beta}_0\right)}$
- Reject H_0 if $|t_0| > t_{\alpha/2, n-2}$
- Can also use the *P*-value approach

Testing Significance of Regression

Does X help to predict Y?

$H_0: \beta_1 = 0 \quad H_1: \beta_1 \neq 0$

- This tests the **significance of regression**; that is, is there a linear relationship between the response and the regressor.
- *Failing to reject* $\beta_1 = 0$, implies that there is no linear relationship between y and x

Example 2.3 The Rocket Propellant Data

We test for significance of regression in the rocket propellant regression model of Example 2.1. The estimate of the slope is $\hat{\beta}_1 = -37.15$, and in Example 2.2, we computed the estimate of σ^2 to be $MS_{\text{Res}} = \hat{\sigma}^2 = 9244.59$. The standard error of the slope is

$$\text{se}\left(\hat{\beta}_1\right) = \sqrt{\frac{MS_{\text{Res}}}{S_{xx}}} = \sqrt{\frac{9244.59}{1106.56}} = 2.89$$

Therefore, the test statistic is

$$t_0 = \frac{\hat{\beta}_1}{\text{se}\left(\hat{\beta}_1\right)} = \frac{-37.15}{2.89} = -12.85$$

If we choose $\alpha = 0.05$, the critical value of t is $t_{0.025,18} = 2.101$. Thus, we would reject H_0: $\beta_1 = 0$ and conclude that there is a linear relationship between shear strength and the age of the propellant. ■

Testing significance of regression

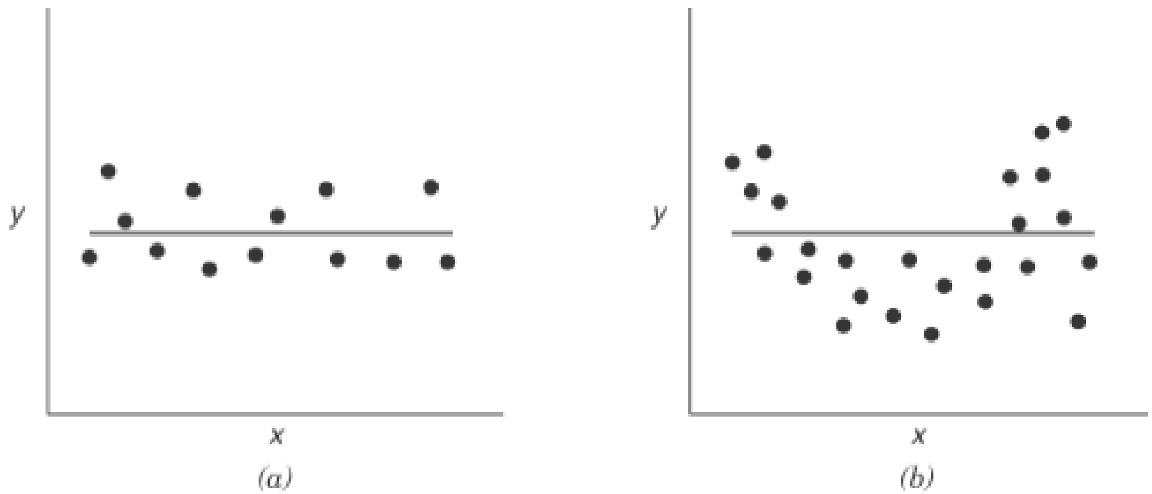

Figure 2.2 Situations where the hypothesis H_0: $\beta_1 = 0$ is not rejected.

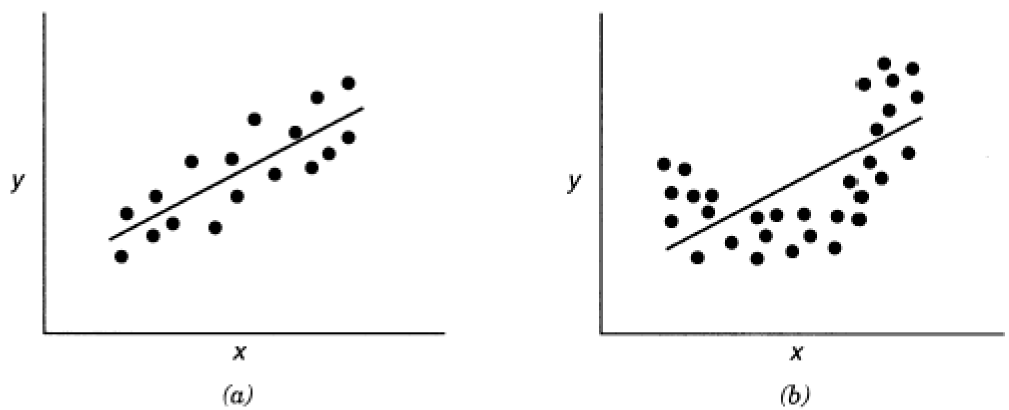

Figure 2.3 Situations where the hypothesis H_0: $\beta_1 = 0$ is rejected.

Always plot data!!

The Analysis of Variance Approach

- Partitioning of total variability

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Add and subtract yihat

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 + 2\sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

see properties on page 31 of these notes to show this cross product term goes to 0.

or

$$\underbrace{\sum (y_i - \bar{y})^2}_{SS_T} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SS_R} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{SS_{\mathrm{Res}}}$$

- It can be shown that $SS_R = \hat{\beta}_1 S_{xy}$

The Analysis of Variance

- Degrees of Freedom

$$\underbrace{\sum\left(y_i-\bar{y}\right)^2}_{SS_T}=\underbrace{\sum\left(\hat{y}_i-\bar{y}\right)^2}_{SS_R}+\underbrace{\sum\left(y_i-\hat{y}_i\right)^2}_{SS_{\mathrm{Res}}}$$

$$n-1 \quad = \quad 1 \quad + \quad (n-2)$$

- Mean Squares

$$MS_R=\frac{SS_R}{1} \qquad MS_{\mathrm{Res}}=\frac{SS_{\mathrm{Res}}}{n-2}$$

The Analysis of Variance

Rely on Cochran's theorem for underlying distribution of MSR/MSRES

- ANOVA procedure for testing H_0: $\beta_1 = 0$

Source of Variation	Sum of Squares	DF	MS	F_0
Regression	SS_R	1	MS_R	MS_R/MS_{Res}
Residual	SS_{Res}	n-2	MS_{Res}	
Total	SS_T	n-1		

- A large value of F_0 indicates that regression is significant; specifically, reject if $F_0 > F_{\alpha,1,n-2}$
- Can also use the *P*-value approach

*Compute the probability under the null hypothesis that the random variable F is greater than the computed F=F**

When p is low H0 must go.

TABLE 2.5 Analysis-of-Variance Table for the Rocket Propellant Regression Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P value
Regression	1,527,334.95	1	1,527,334.95	165.21	1.66×10^{-10}
Residual	166,402.65	18	9,244.59		
Total	1,693,737.60	19			

The Analysis of Variance

Relationship between t^0 and F^0:

- For H_0: $\beta_1 = 0$, it can be shown that:

$$t_0^2 = F_0$$

So for testing significance of regression, the t-test and the ANOVA procedure are equivalent (only true in simple linear regression)

Coefficient of Determination

- R^2 - coefficient of determination
- R^2 - SS(Regression)/SS(Total)
- Proportion of variation explained by the regressor, x
- For the rocket propellant data

$$R^2 = \frac{SS_R}{SS_T} = \frac{1{,}527{,}334.95}{1{,}693{,}737.60} = 0.9018$$

Coefficient of Determination

- R^2 can be misleading!
  - Simply adding more terms to the model will increase R^2
  - As the range of the regressor variable increases (decreases), R^2 generally increases (decreases).
  - R^2 does not indicate the appropriateness of a linear model

Go to Example2_1Rocket.sas and output in Rocket.pdf

Considerations in the Use of Regression

- Extrapolating
- Extreme points will often influence the slope.
- Outliers can disturb the least-squares fit
- Linear relationship does not imply cause-effect relationship

- Sometimes, the value of the regressor variable is unknown and itself be estimated or predicted.