Chapter 14

Linear Regression

Numerical Methods
Fall 2019

Statistics Review: Measure of Location

- ***Arithmetic mean***: the sum of the individual data points (y_i) divided by the number of points n:

$$\bar{y} = \frac{\sum y_i}{n}$$

- ***Median***: the midpoint of a group of data.

- ***Mode***: the value that occurs most frequently in a group of data.

Statistics Review: Measures of Spread

- ***Standard deviation*:**

$$s_y = \sqrt{\frac{S_t}{n-1}}$$

where S_t is the sum of the squares of the data residuals:

$$S_t = \sum (y_i - \bar{y})^2$$

and n–1 is referred to as the *degrees of freedom*.

- ***Variance*:**

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{\sum y_i^2 - \left(\sum y_i\right)^2 / n}{n-1}$$

- ***Coefficient of variation*:**

$$\text{c.v.} = \frac{s_y}{\bar{y}} \times 100\%$$

Example 14.1

Normal Distribution

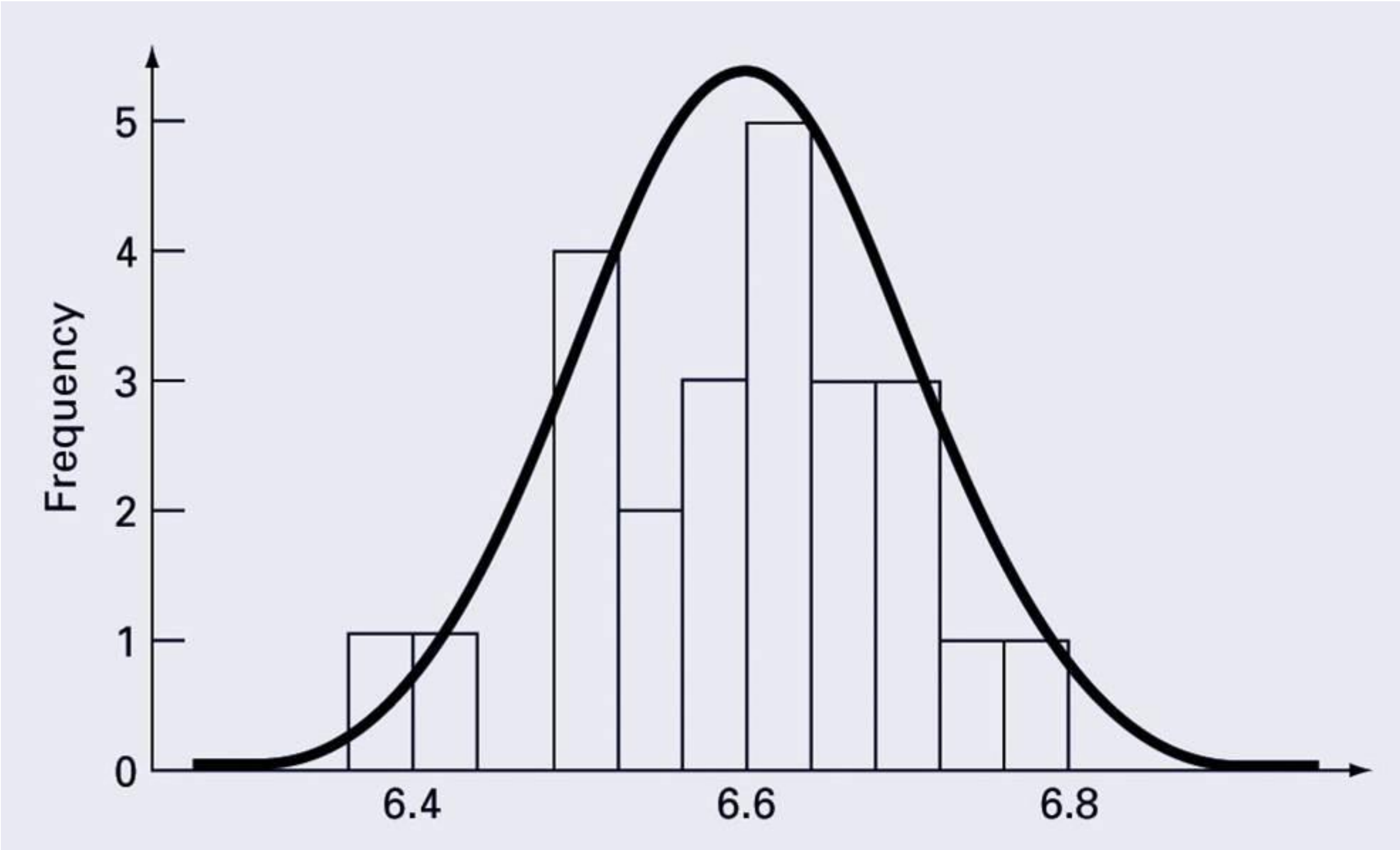


TABLE 14.2: Measurements of the coefficient of thermal expansion of structural steel.

Descriptive Statistics in MATLAB

- MATLAB has several built-in commands to compute and display descriptive statistics. Assuming some column vector *s*:
  - `mean(s), median(s), mode(s)`
    - Calculate the mean, median, and mode of *s*. `mode` is a part of the statistics toolbox.
  - `min(s), max(s)`
    - Calculate the minimum and maximum value in *s*.
  - `var(s), std(s)`
    - Calculate the variance and standard deviation of *s*
- Note - if a matrix is given, the statistics will be returned for each column.

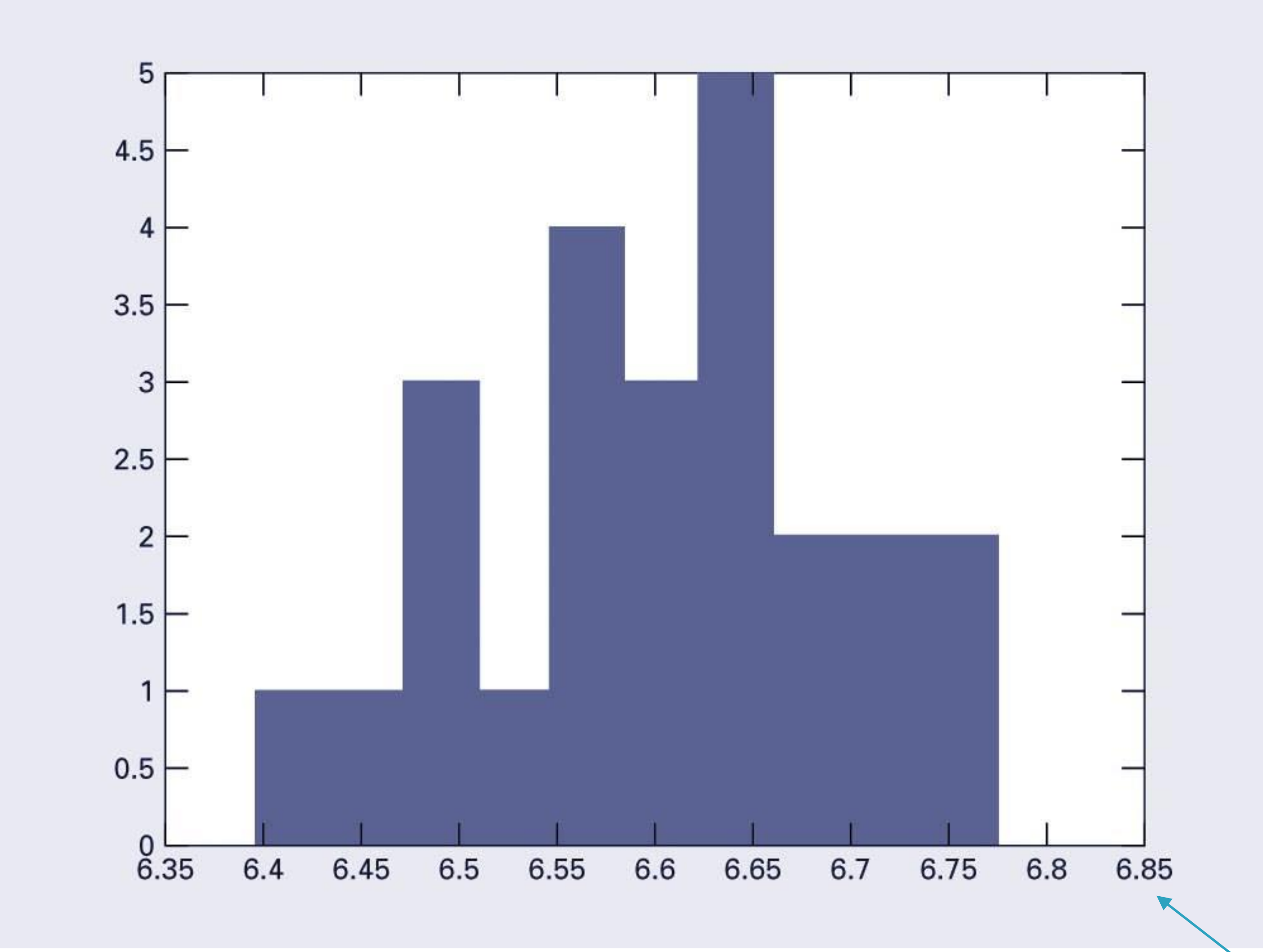
Histograms in MATLAB

- `[n, x] = hist(s, x)`
  - Determine the number of elements in each bin of data in *s*. *x* is a vector containing the center values of the bins.

- `[n, x] = hist(s, m)`
  - Determine the number of elements in each bin of data in *s* using *m* bins. *x* will contain the centers of the bins. The default case is $m = 10$

- `hist(s, x)` or `hist(s, m)` or `hist(s)`
  - With no output arguments, `hist` will actually produce a histogram.

Histogram Example



Histogram generated with the MATLAB `hist` function.

10 bins

Linear Least-Squares Regression

- Linear least-squares regression is a method to determine the “best” coefficients in a linear model for given data set.
- “Best” for least-squares regression means minimizing the sum of the squares of the *estimate* residuals. For a straight line model, this gives:

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \qquad (14.12)$$

- This method will yield a unique line for a given set of data.

Least-Squares Fit of a Straight Line

- Using the model:

$$y = a_0 + a_1 x$$

- To determine values for a_0 and a_1, Eq. (14.12) is differentiated with respect to each unknown coefficient:

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i)$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum [(y_i - a_0 - a_1 x_i) x_i]$$

- The slope and intercept producing the best fit can be found using:

$$a_1 = \frac{n \sum x_i \, yi - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

Least-Squares Fit of a Straight Line

- EXAMPLE 14.4
  - A free-falling object such as a bungee jumper is subject to the upward force of air resistance.

$$F_u = C_d v^2$$

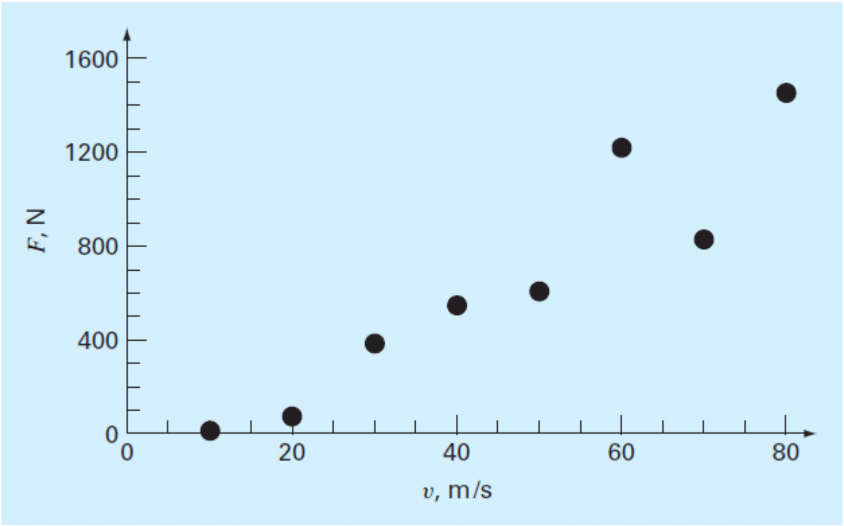


TABLE 14.1 Experimental data for force (N) and velocity (m/s) from a wind tunnel experiment.

v, m/s	10	20	30	40	50	60	70	80
F, N	25	70	380	550	610	1220	830	1450

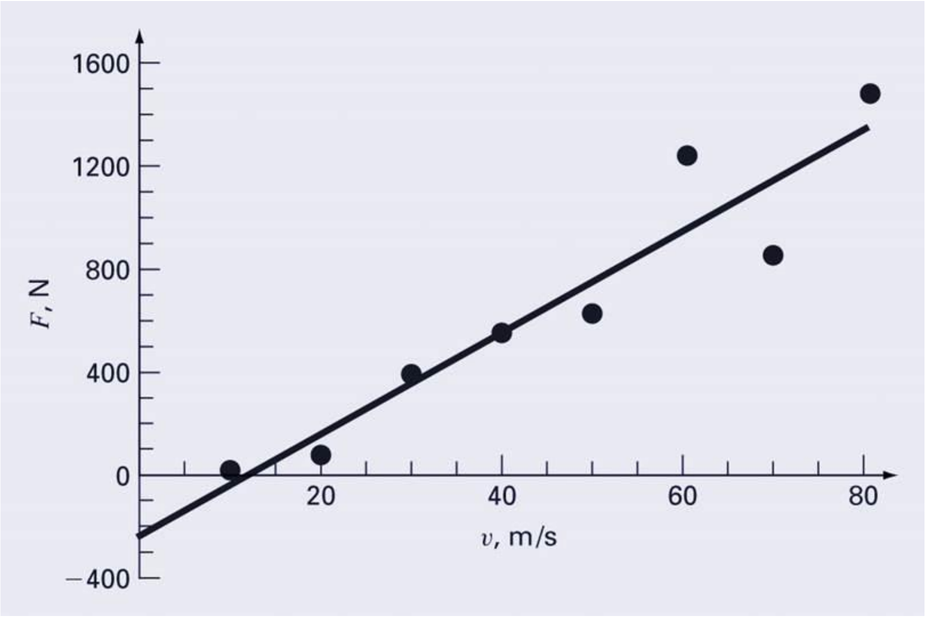
Least-Squares Fit of a Straight Line Example

$$a_1 = \frac{n\sum x_i\, yi - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{8(312850) - (360)(5135)}{8(20400) - (360)^2} = 19.47024$$

$$a_0 = \bar{y} - a_1\bar{x} = 641.875 - 19.47024(45) = -234.2857$$

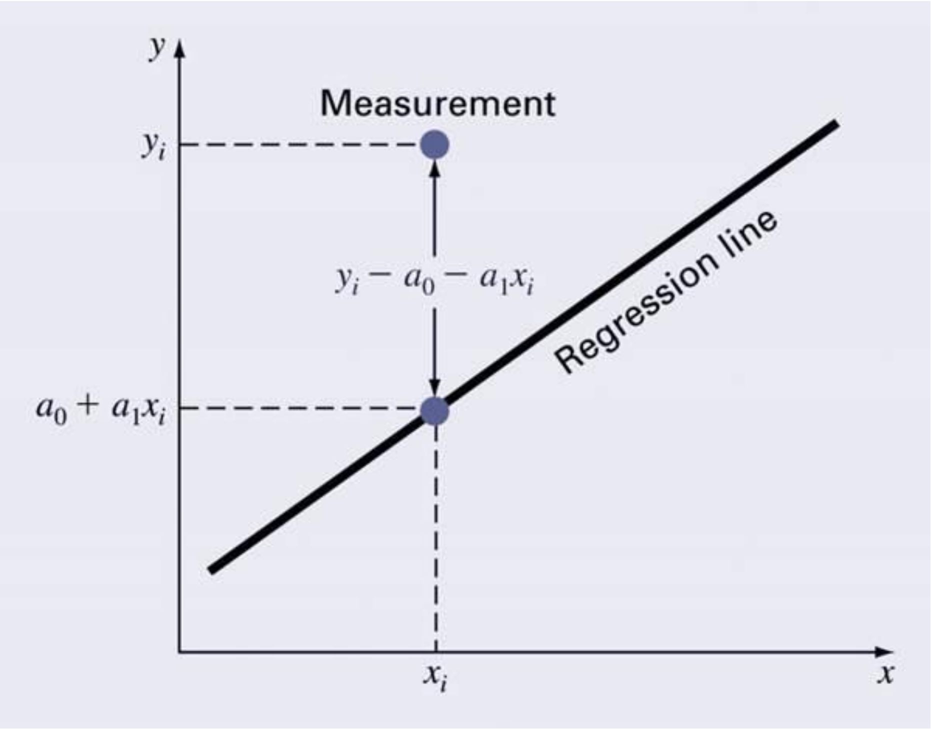
	V (m/s)	*F* (N)		
i	x_i	y_i	$(x_i)^2$	$x_i y_i$
1	10	25	100	250
2	20	70	400	1400
3	30	380	900	11400
4	40	550	1600	22000
5	50	610	2500	30500
6	60	1220	3600	73200
7	70	830	4900	58100
8	80	1450	6400	116000
Σ	360	5135	20400	312850

$$F_{est} = -234.2857 + 19.47024v$$



Quantification of Error

- Recall for a straight line, the sum of the squares of the estimate residuals:



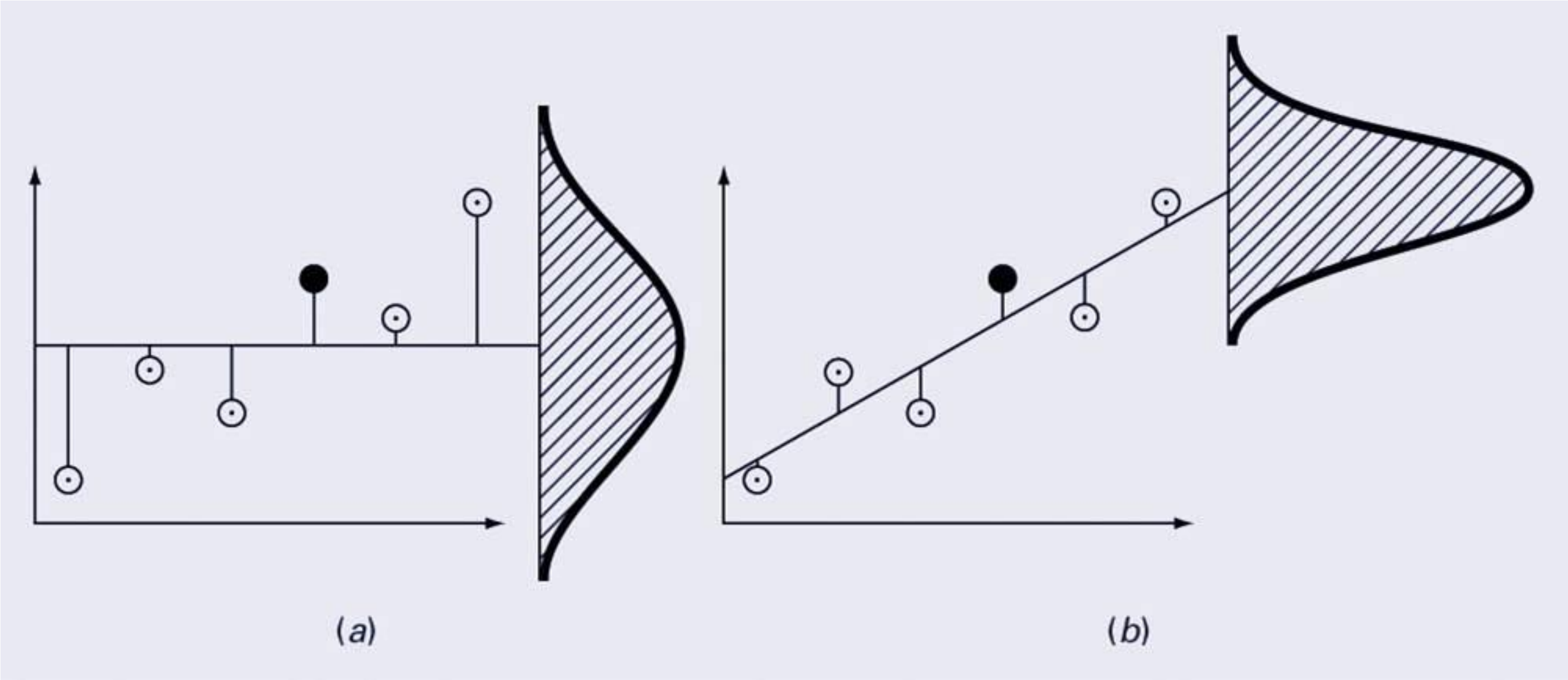
$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

- *Standard error of the estimate:*

$$s_{y/x} = \sqrt{\frac{S_r}{n-2}}$$

Standard Error of the Estimate

- Regression data showing (a) the spread of data around the mean of the dependent data and (b) the spread of the data around the best fit line:



- The reduction in spread represents the improvement due to linear regression.

Coefficient of Determination

- The *coefficient of determination r^2* is the difference between the sum of the squares of the data residuals and the sum of the squares of the estimate residuals, normalized by the sum of the squares of the data residuals:

$$r^2 = \frac{S_t - Sr}{S_t}$$

- r^2 represents the percentage of the original uncertainty explained by the model.
  - For a perfect fit, $S_r = 0$ and $r^2 = 1$.
  - If $r^2 = 0$, there is no improvement over simply picking the mean.
  - If $r^2 < 0$, the model is *worse* than simply picking the mean!

EXAMPLE 14.5

Coefficient of Determination Example

	V (m/s)	F (N)			
i	x_i	y_i	$a_0+a_1x_i$	$(y_i-\bar{y})^2$	$(y_i-a_0-a_1x_i)^2$
1	10	25	−39.58	380535	4171
2	20	70	155.12	327041	7245
3	30	380	349.82	68579	911
4	40	550	544.52	8441	30
5	50	610	739.23	1016	16699
6	60	1220	933.93	334229	81837
7	70	830	1128.63	35391	89180
8	80	1450	1323.33	653066	16044
Σ	360	5135		1808297	216118

$$F_{est} = -234.2857 + 19.47024v$$

$$S_t = \sum (y_i - \bar{y})^2 = 1808297$$

$$S_r = \sum (y_i - a_0 - a_1 x_i)^2 = 216118$$

$$s_y = \sqrt{\frac{1808297}{8-1}} = 508.26$$

$$s_{y/x} = \sqrt{\frac{216118}{8-2}} = 189.79$$

$$r^2 = \frac{1808297 - 216118}{1808297} = 0.8805$$

- 88.05% of the original uncertainty has been explained by the linear model

Coefficient of Determination

- r^2 is close to 1 does not mean that the fit is necessarily "good".
- As in Fig. 14.12, he came up with four data sets consisting of 11 data points each. Although their graphs are very different, all have the same best-fit equation, $y = 3 + 0.5x$, and the same coefficient of determination, $r^2 = 0.67$

Coefficient of Determination

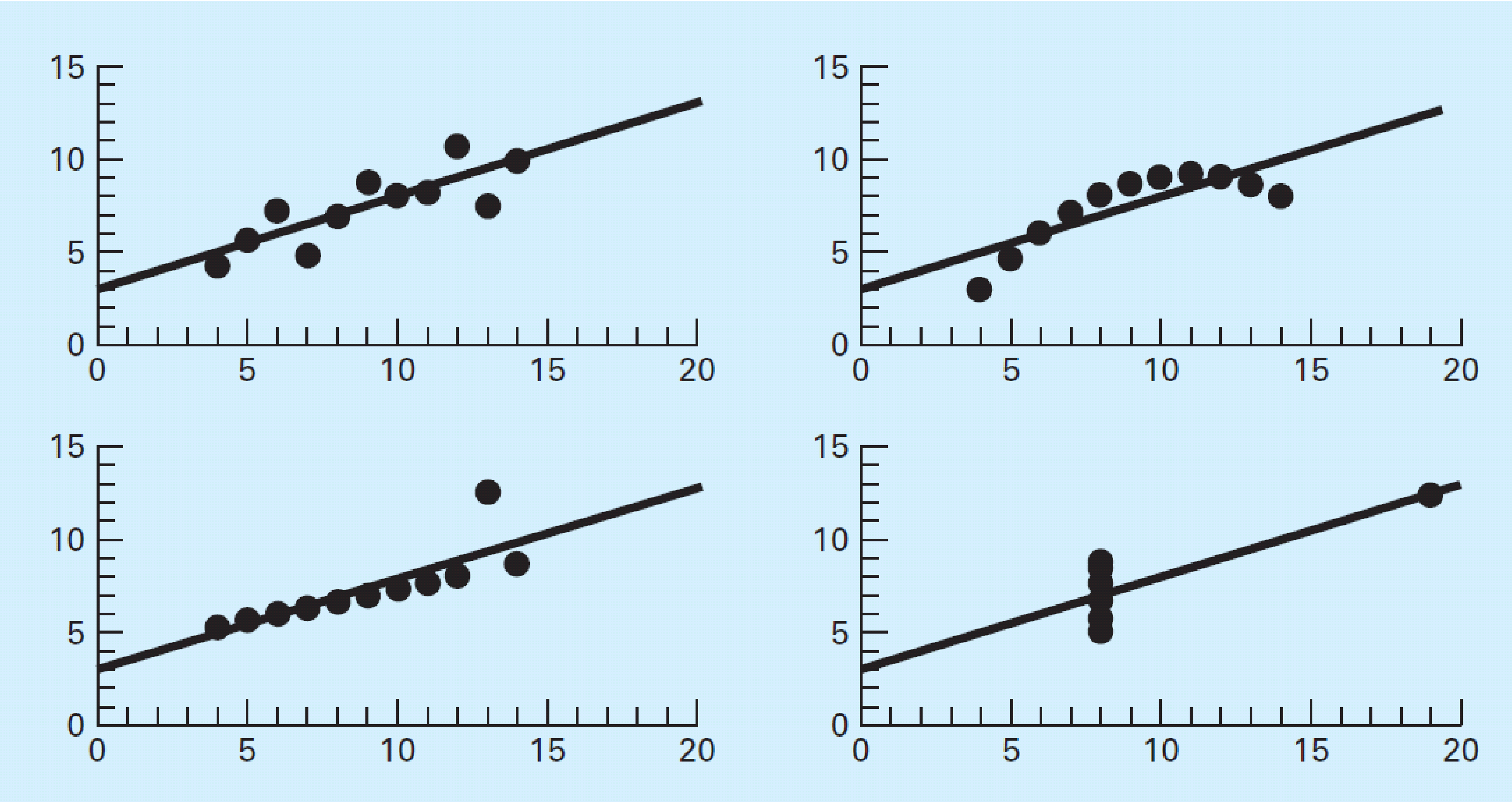


FIGURE 14.12
Anscombe's four data sets along with the best-fit line, $y = 3 + 0.5x$.

Nonlinear Relationships

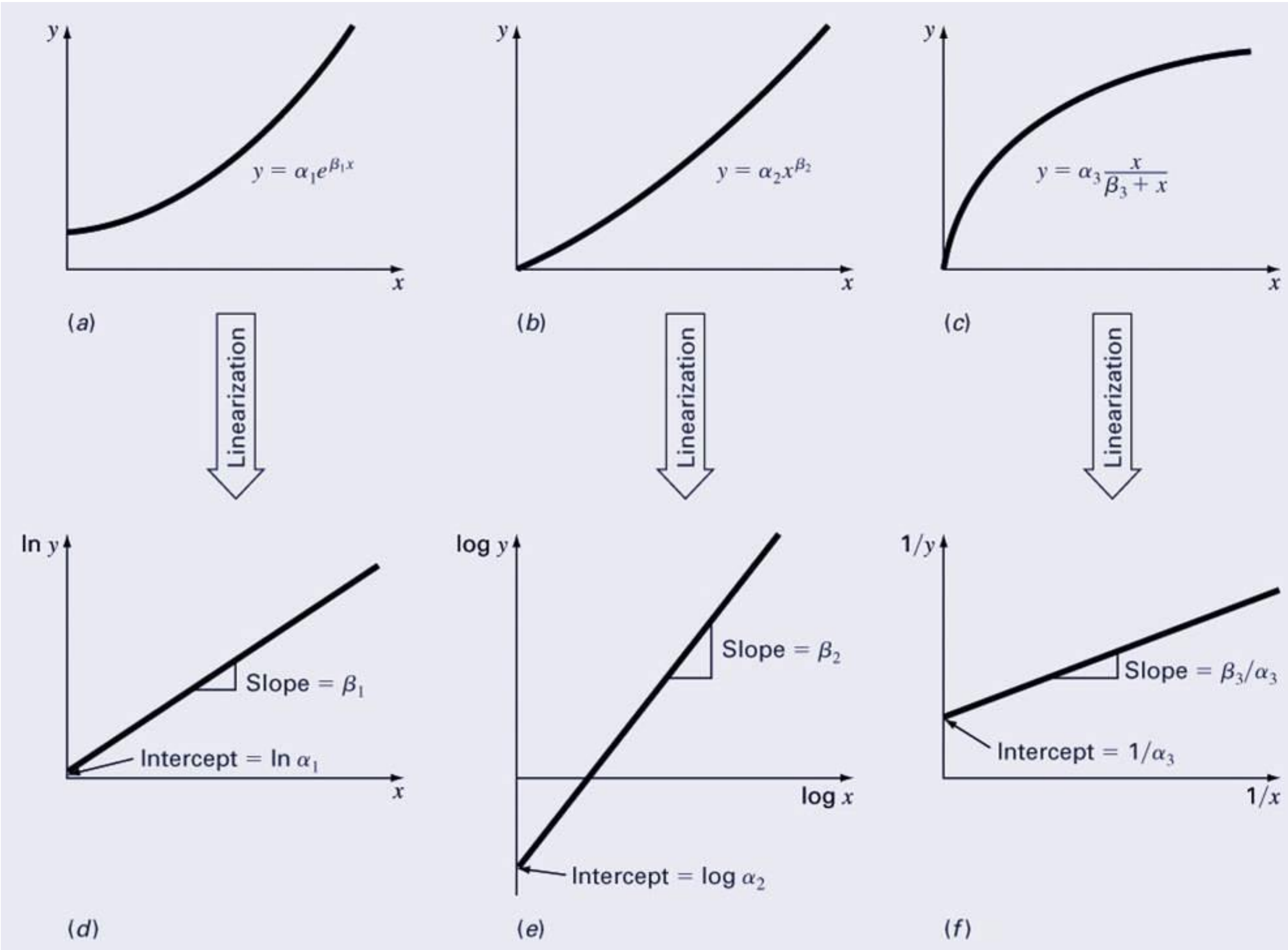
- Linear regression is predicated on the fact that the relationship between the dependent and independent variables is linear - this is not always the case.
- Three common examples are:
  - exponential : $y = \alpha_1 e^{\beta_1 x}$
  - power : $y = \alpha_2 x^{\beta_2}$
  - saturation−growth−rate : $y = \alpha_3 \frac{x}{\beta_3 + x}$

Linearization of Nonlinear Relationships

- One option for finding the coefficients for a nonlinear fit is to linearize it. For the three common models, this may involve taking logarithms or inversion:

Model	Nonlinear	Linearized	
exponential :	$y = \alpha_1 e^{\beta_1 x}$	$\ln y = \ln \alpha_1 + \beta_1 x$	(14.22)
power :	$y = \alpha_2 x^{\beta_2}$	$\log y = \log \alpha_2 + \beta_2 \log x$	(14.23)
saturation-growth-rate :	$y = \alpha_3 \dfrac{x}{\beta_3 + x}$	$\dfrac{1}{y} = \dfrac{1}{\alpha_3} + \dfrac{\beta_3}{\alpha_3} \dfrac{1}{x}$	(14.24)

Transformation Examples



Linearization of Nonlinear Relationships

- EXAMPLE 14.6
  - Fitting Data with the Power Equation.
  - Fit Eq. (14.23) to the data in Table 14.1 using a logarithmic transformation.

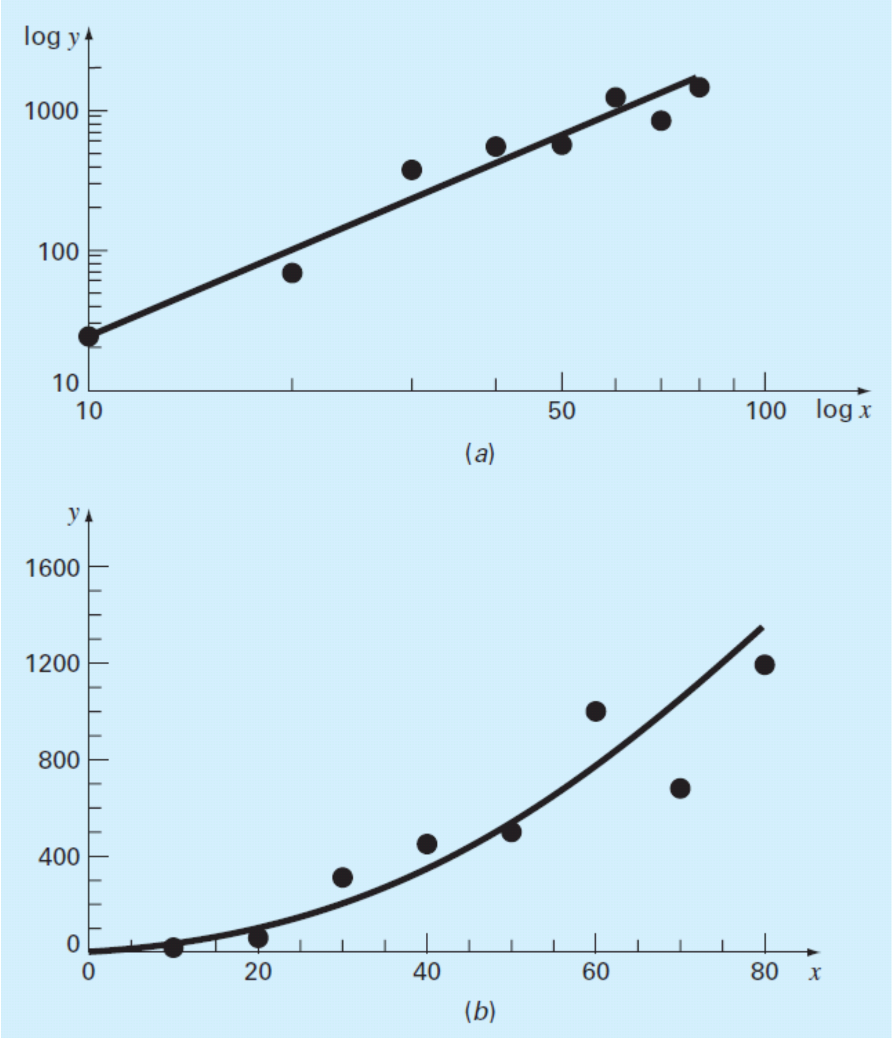


TABLE 14.1 Experimental data for force (N) and velocity (m/s) experiment.

v, m/s	10	20	30	40	50	60	70	80
F, N	25	70	380	550	610	1220	830	1450

MATLAB Functions

- MATLAB has a built-in function `polyfit` that fits a least-squares nth order polynomial to data:
  - `p = polyfit(x, y, n)`
    - x: independent data
    - y: dependent data
    - n: order of polynomial to fit
    - p: coefficients of polynomial

$$f(x)=p_1x^n+p_2x^{n-1}+\ldots+p_nx+p_{n+1}$$

```
>> x = [10 20 30 40 50 60 70 80];
>> y = [25 70 380 550 610 1220 830 1450];
>> a = polyfit(x,y,1)

a =
   19.4702 -234.2857
```

HW#5

- 연습문제

- 13.5
- 14.4, 14.5, 14.6

  (using MATLAB or Equation (14.15), (14.16))