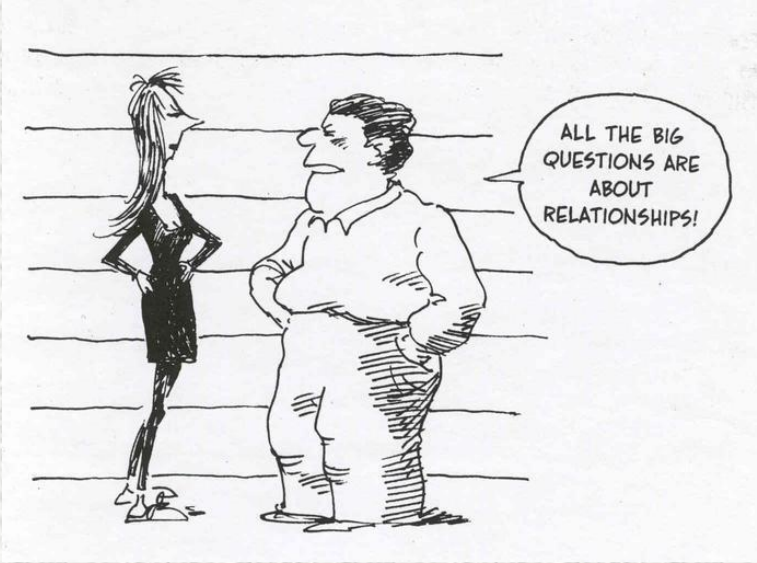
Linear Regression and Correlation

Correlation Coefficient, Least Squares,, Test and Confidence Intervals, Estimation and Prediction

Dr.Mamatha.H.R
Professor
Department of Computer Science and Engineering
PES University
Bangalore

Course material created using various Internet resources and text book

Introduction



So far we have done statistics on one variable at a time. We now interested in relationships between two variables and how to use one variable to predict another variable.

- Does weight depend on height?
- Does blood pressure level predict life expectancy?
- Do SAT scores predict college performance?
- Does taking STATISTICS make you a better person?

Example: Age and Fatness

The following data was collected in a study of age and fatness in humans.

Age	23	23	27	27	39	41	45	49	50
% Fat	9.5	27.9	7.8	17.8	31.4	25.9	27.4	25.2	31.1

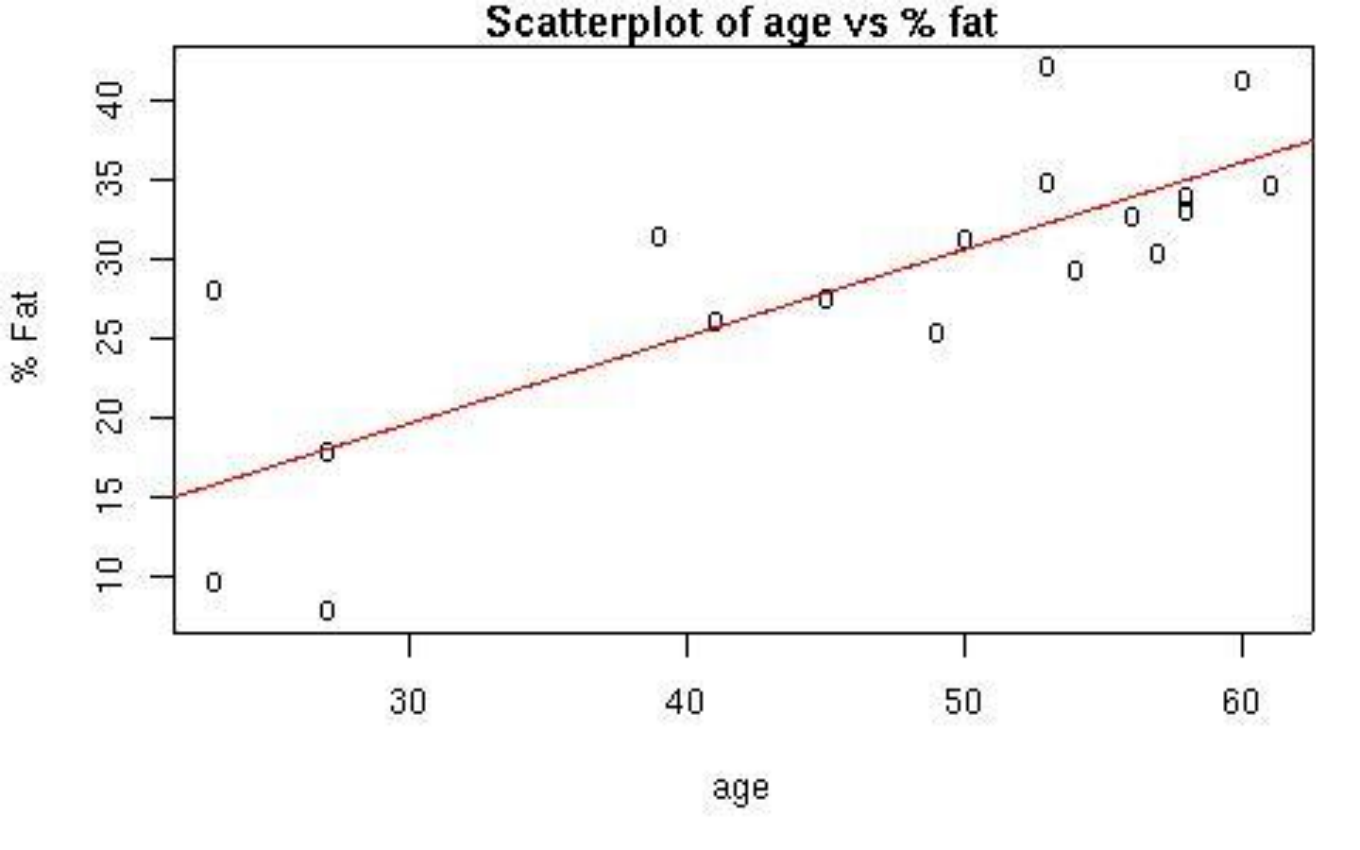
Age	53	53	54	56	57	58	58	60	61
% Fat	34.7	42	29.1	32.5	30.3	33	33.8	41.1	34.5

One of the questions was, "What is the relationship between age and fatness?"

* Mazess, R.B., Peppler, W.W., and Gibbons, M. (1984) Total body composition by dual-photon (^{153}Gd) absorptiometry. *American Journal of Clinical Nutrition*, **40**, *834-839*

Example: Age and Fatness

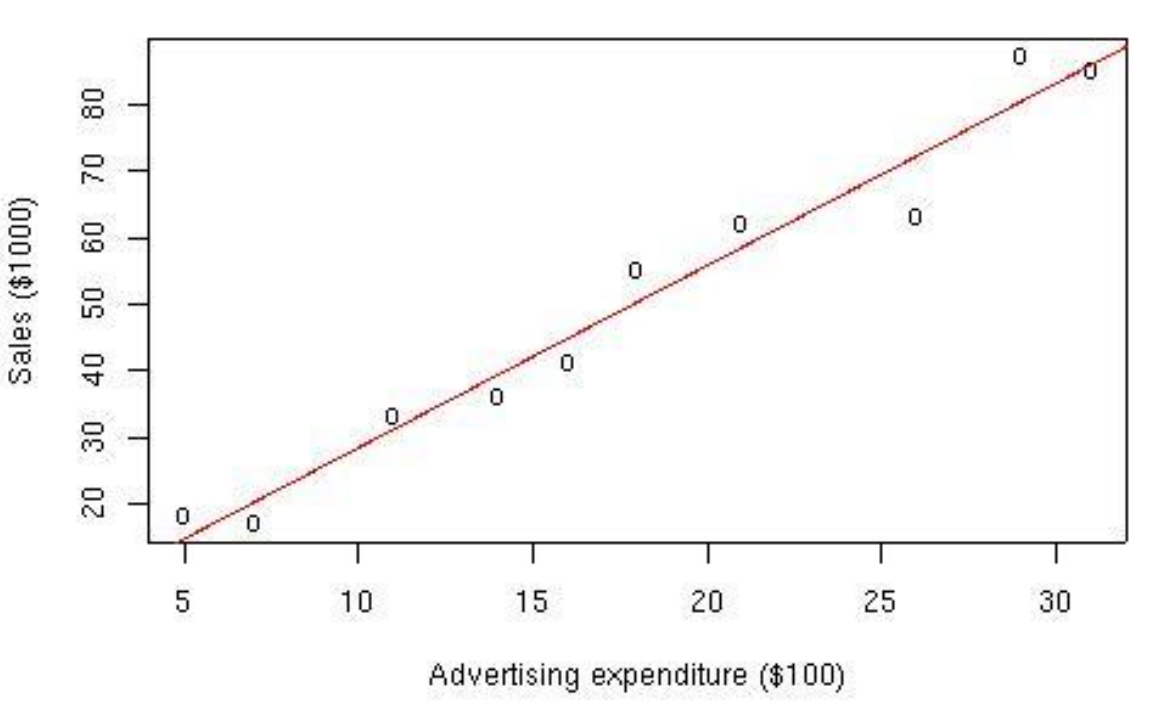
The following scatterplot shows that % fat in general tend to increase with age. The relationship is close, but not exactly, linear.



Example: Advertising and Sale

The following table contains sales (y) and advertising expenditures (x) for 10 branches of a retail store.

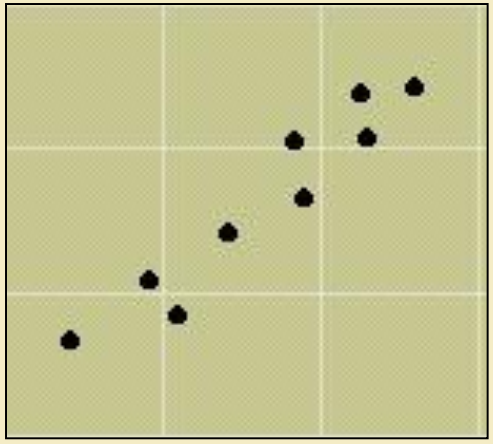
x ($100)	18	7	14	31	21	5	11	16	26	29
y ($1000)	55	17	36	85	62	18	33	41	63	87



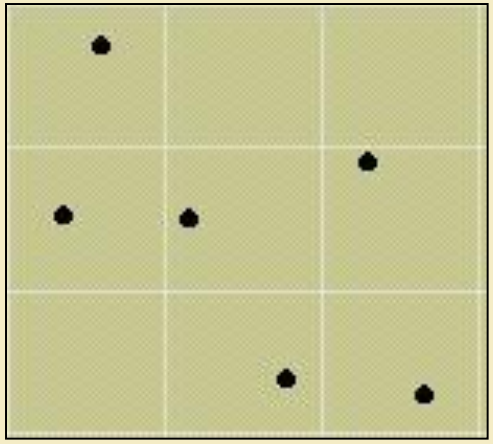
Describing the Scatterplot

- What **pattern** or **form** do you see?
  - Straight line upward or downward
  - Curve or no pattern at all
- How **strong** is the pattern?
  - Strong or weak
- Are there any **unusual observations**?
  - Clusters or outliers

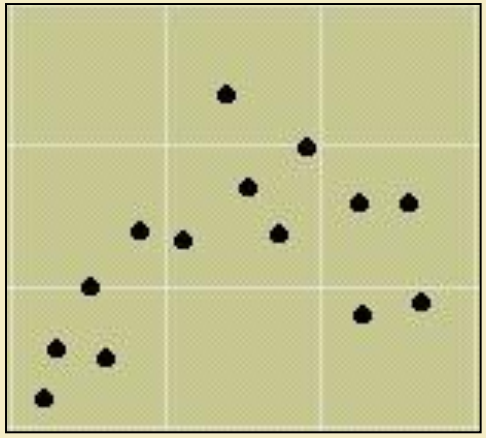
Examples



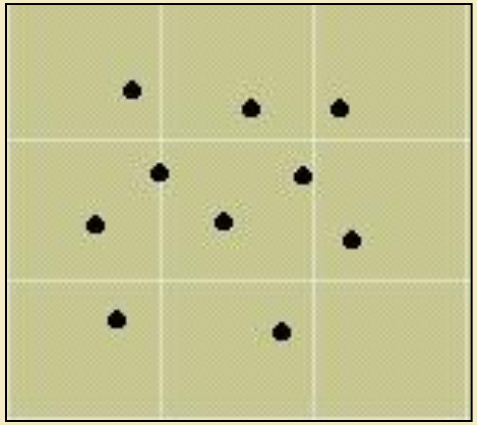
Positive linear - strong



Negative linear -weak



Curvilinear



No relationship

Investigation of Relationship

There are two approaches to investigate linear relationship

- Correlation coefficient: a numerical measure of the **strength** and **direction** of the linear relationship between *x* and *y*.
- Linear regression: a linear equation expresses the relationship between *x* and *y*. It provides a **form** of the relationship.

The Correlation Coefficient

The strength and direction of the relationship between x and y are measured using the **correlation coefficient (Pearson product moment coefficient of correlation), *r*.**

$$r = \frac{s_{xy}}{s_x s_y}$$

where

$$s_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n-1}$$

s_x = standard deviation of the x's

s_y = standard deviation of the y's

Example

The table shows the heights and weights of $n = 10$ randomly selected college football players.

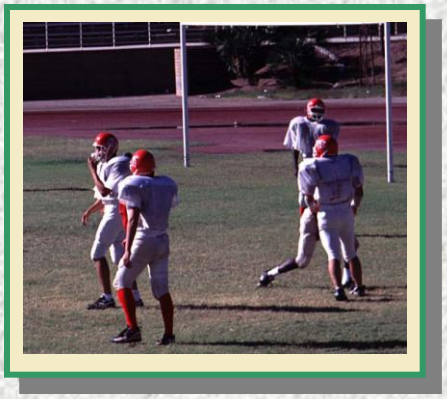
Player	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Height, x	73	71	75	72	72	75	67	69	71	69
Weight, y	185	175	200	210	190	195	150	170	180	175

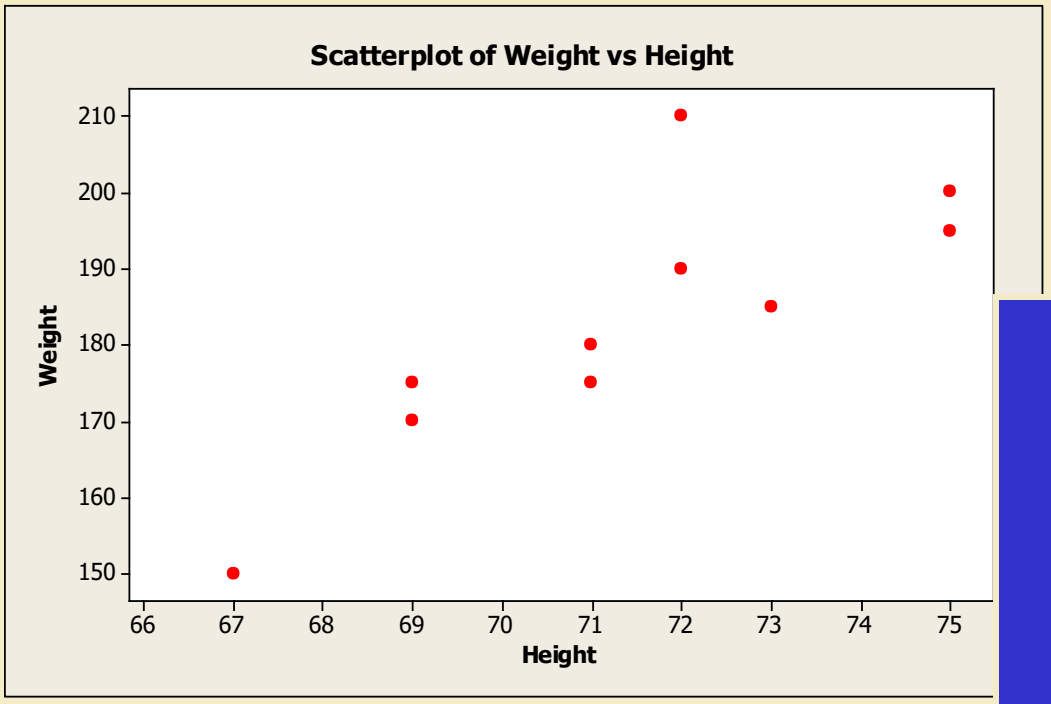
Use your calculator to find the sums and sums of squares.

$$S_{xy} = 328 \quad S_{xx} = 60.4 \quad S_{yy} = 2610$$

$$r = \frac{328}{\sqrt{(60.4)(2610)}} = .8261$$

Football Players





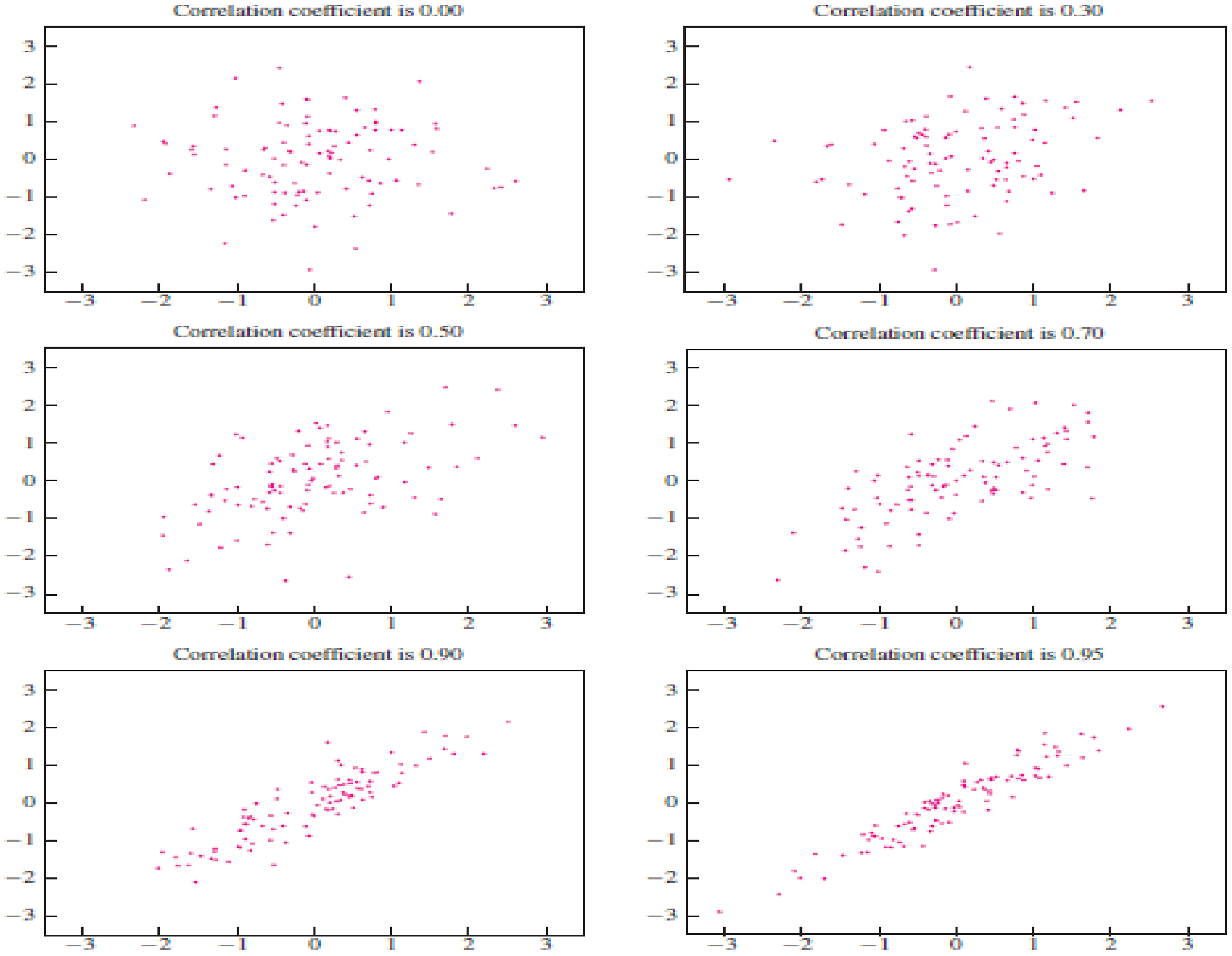
$r = .8261$

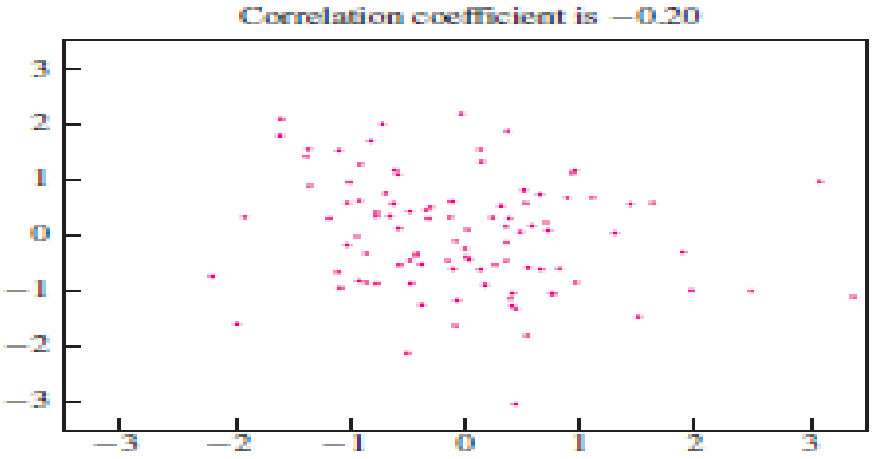
Strong positive correlation

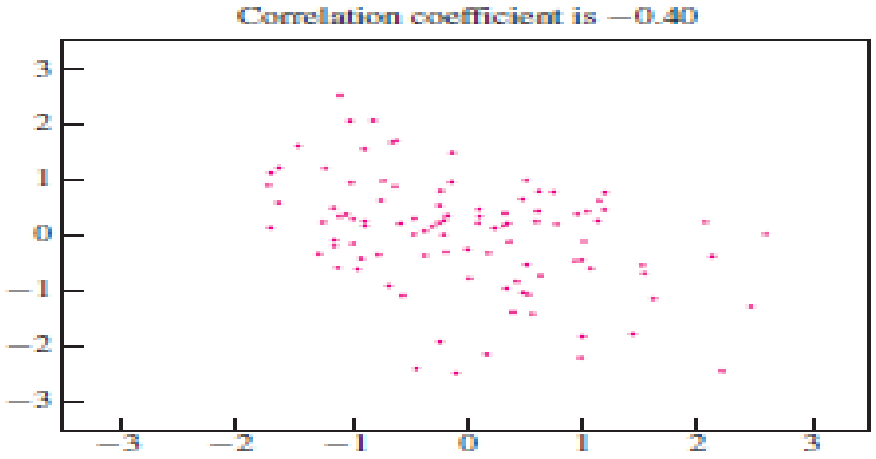
As the player's height increases, so does his weight.

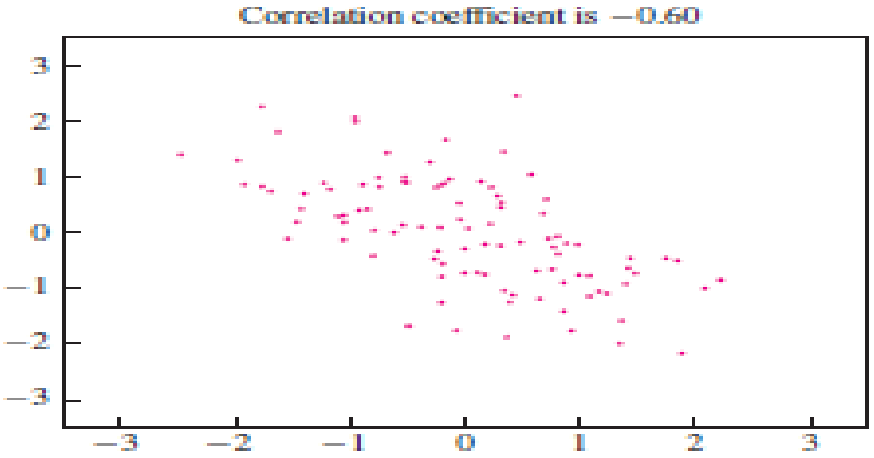
Interpreting *r*

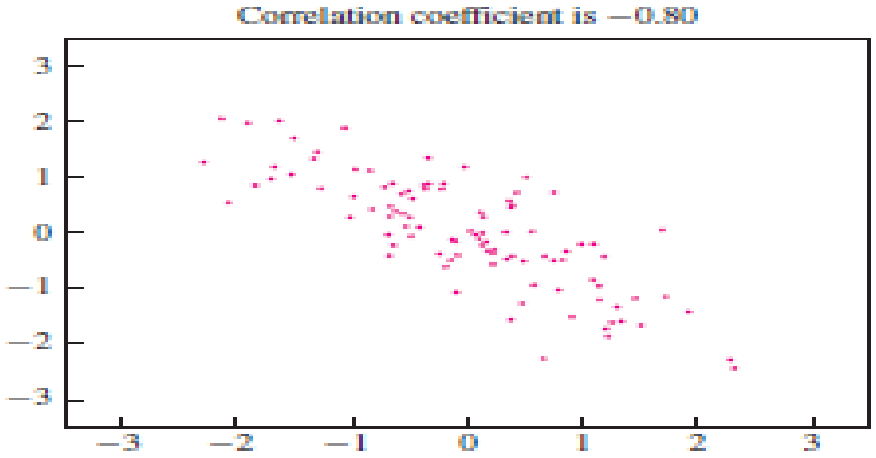
- $-1 \le r \le 1$ Sign of *r* indicates direction of the linear relationship.
- $r \approx 0$ No relationship; random scatter of points
- $r \approx 1$ or -1 Strong relationship; either positive or negative
- $r = 1$ or -1 All points fall exactly on a straight line.

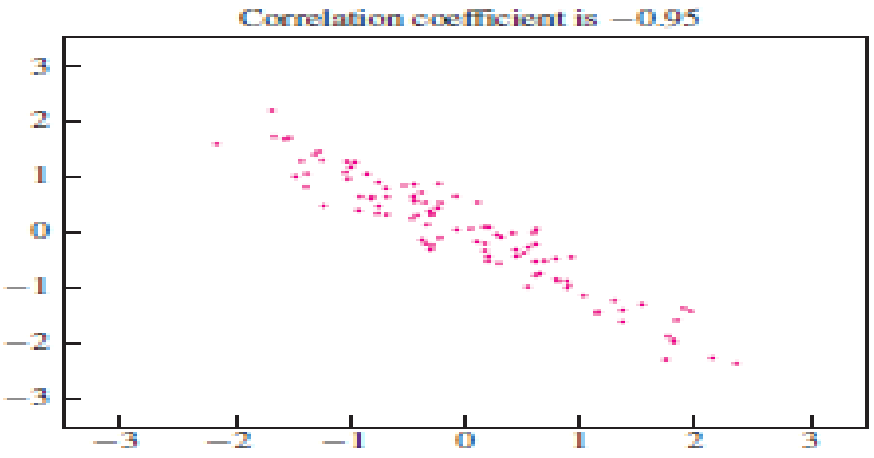

Correlation coefficient is 0.00
Correlation coefficient is 0.30
Correlation coefficient is 0.50
Correlation coefficient is 0.70
Correlation coefficient is 0.90
Correlation coefficient is 0.95

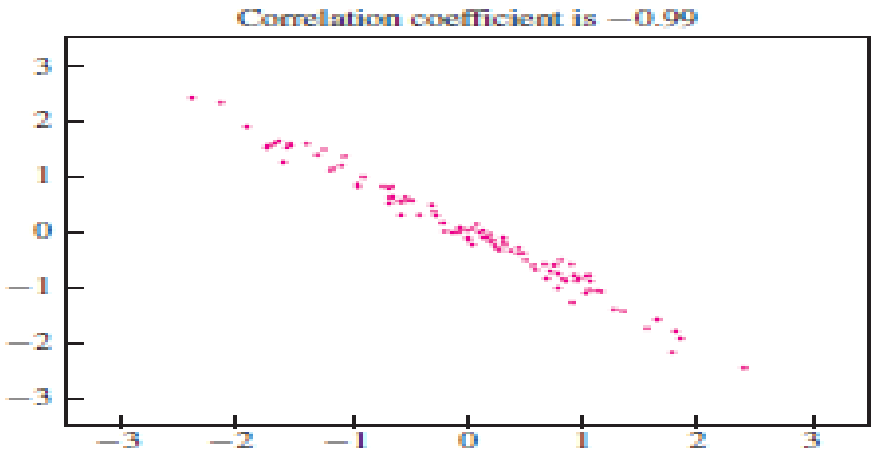

Correlation coefficient is −0.20
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


Correlation coefficient is −0.40
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


Correlation coefficient is −0.60
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


Correlation coefficient is −0.80
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


Correlation coefficient is −0.95
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


Correlation coefficient is −0.99
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

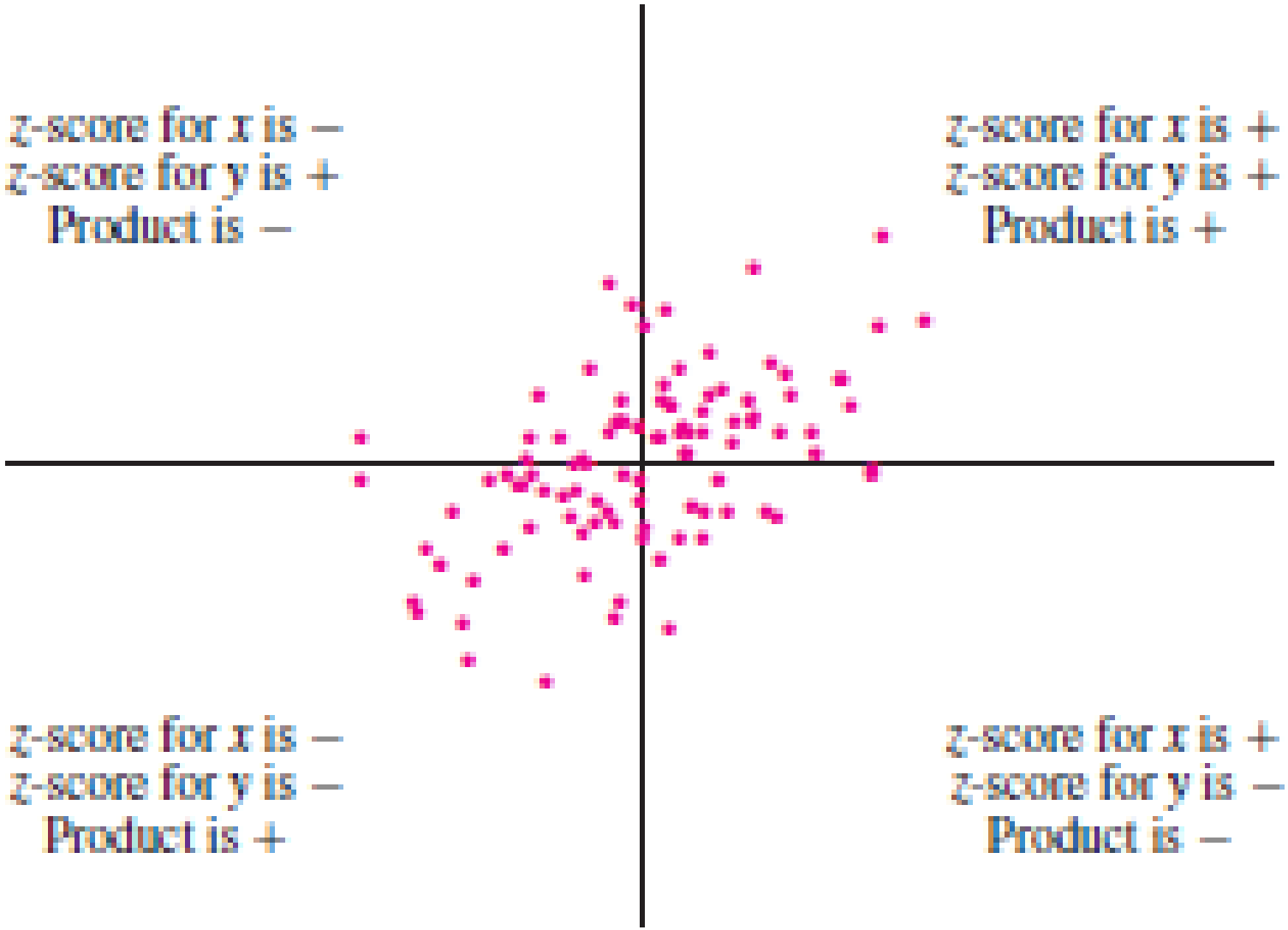


FIGURE 7.5 How the correlation coefficient works.

Summary

The correlation coefficient remains unchanged under each of the following operations:

- Multiplying each value of a variable by a positive constant.
- Adding a constant to each value of a variable.
- Interchanging the values of x and y.

The Correlation Coefficient Is Unit less

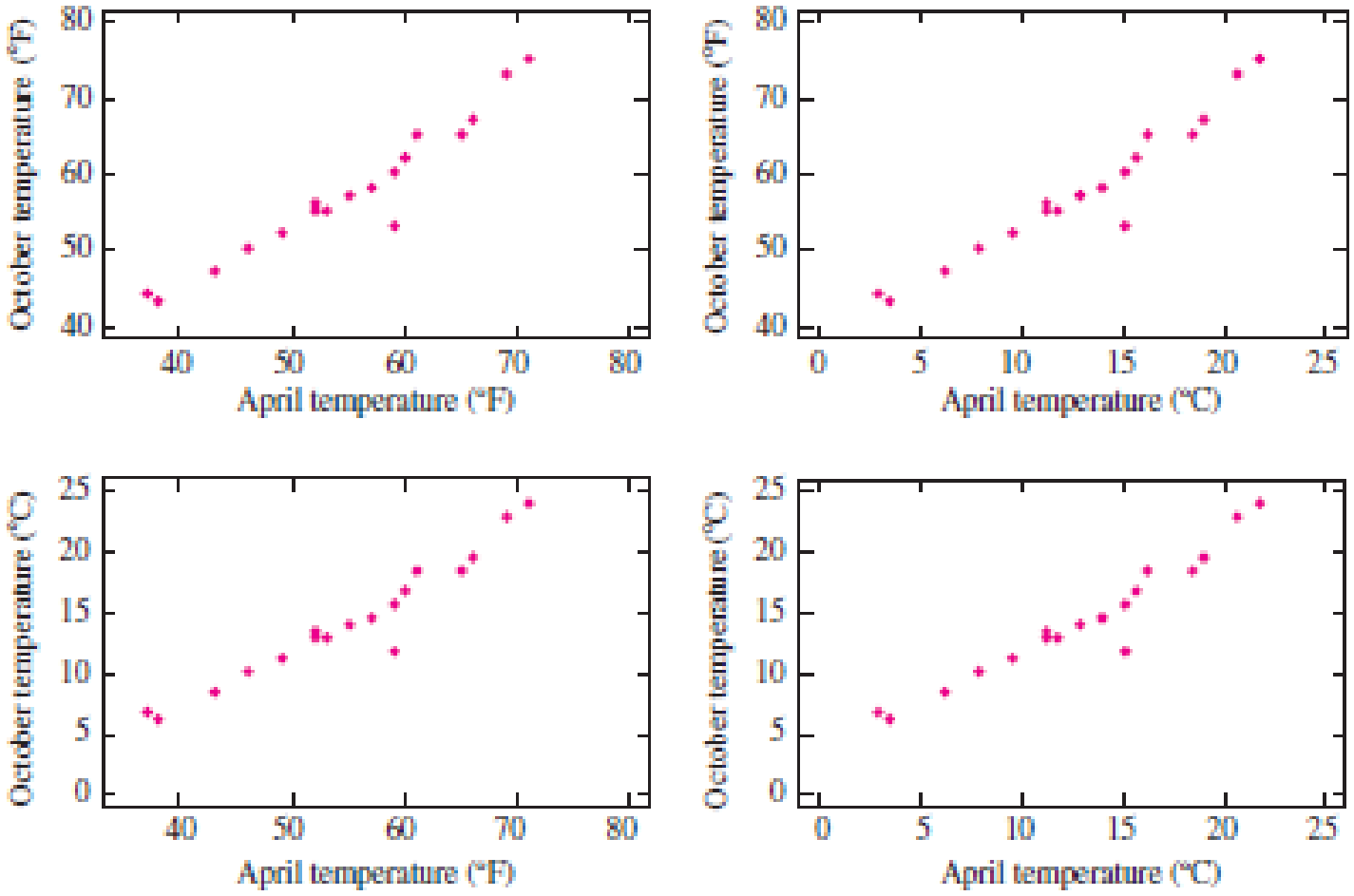


FIGURE 7.6 Mean April and October temperatures for several U.S. cities. The correlation coefficient is 0.96 for each plot; the choice of units does not matter.

The Correlation Coefficient Measures Only *Linear* Association

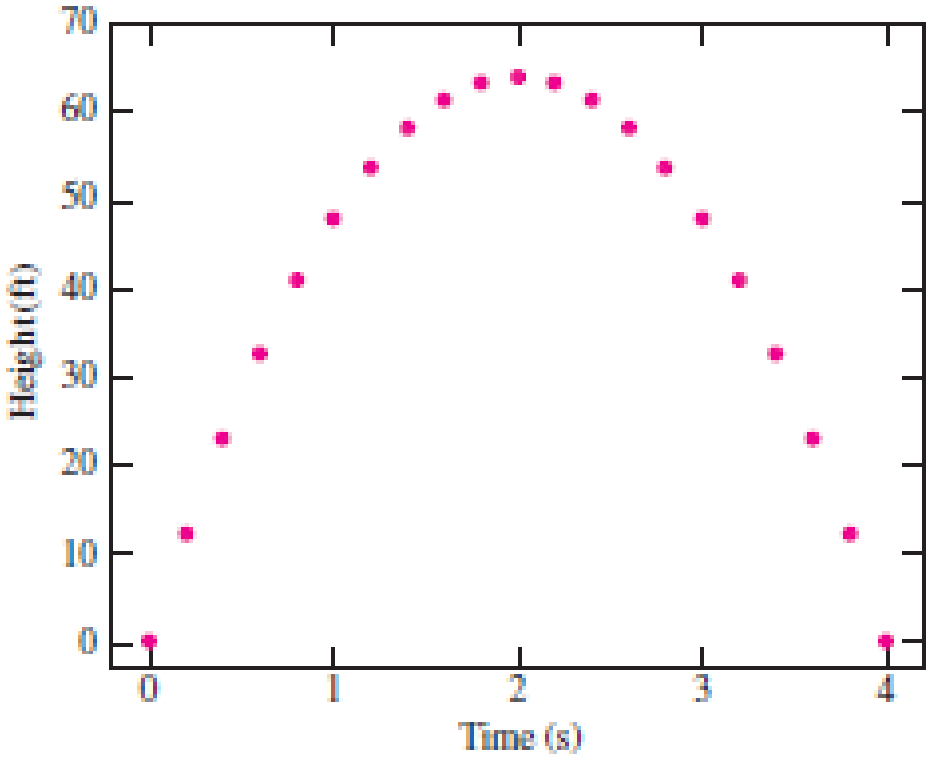


FIGURE 7.7 The relationship between the height of a free-falling object with a positive initial velocity and the time in free fall is quadratic. The correlation is equal to 0.

- the correlation between x and y is equal to 0.
- Is something wrong?
- No. The value of 0 for the correlation indicates that there is no *linear* relationship between x and y, which is true.
- The relationship is purely *quadratic*.
- the correlation coefficient should only be used when the relationship between the x and y is linear.
- Otherwise the results can be misleading.

The Correlation Coefficient can be Misleading when Outliers are Present

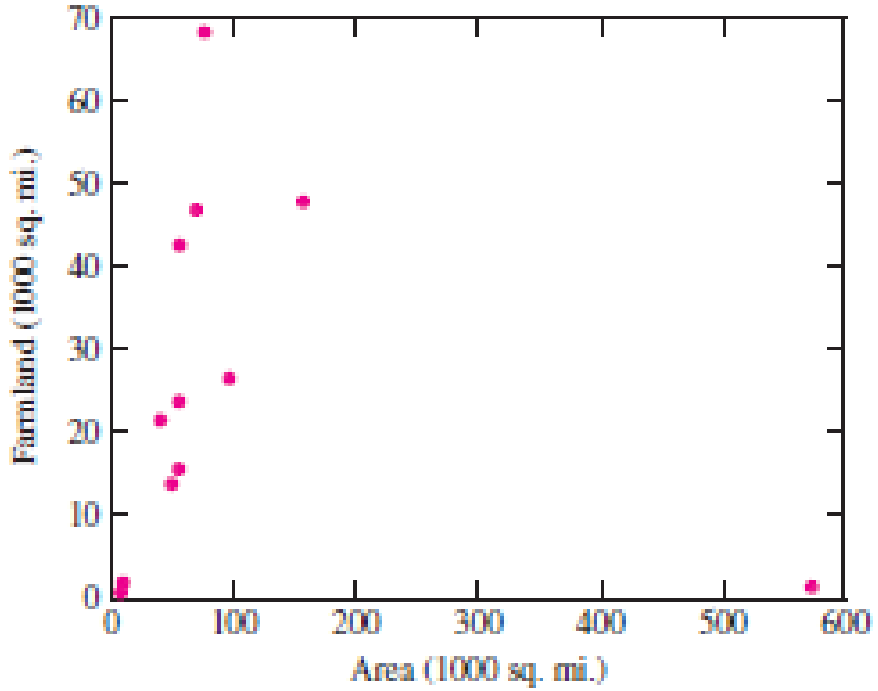


FIGURE 7.8 The correlation is −0.12. Because of the outlier, the correlation coefficient is misleading.

Correlation Is Not Causation

Confounding occurs when there is a third variable that is correlated with both of the variables of interest, resulting in a correlation between them

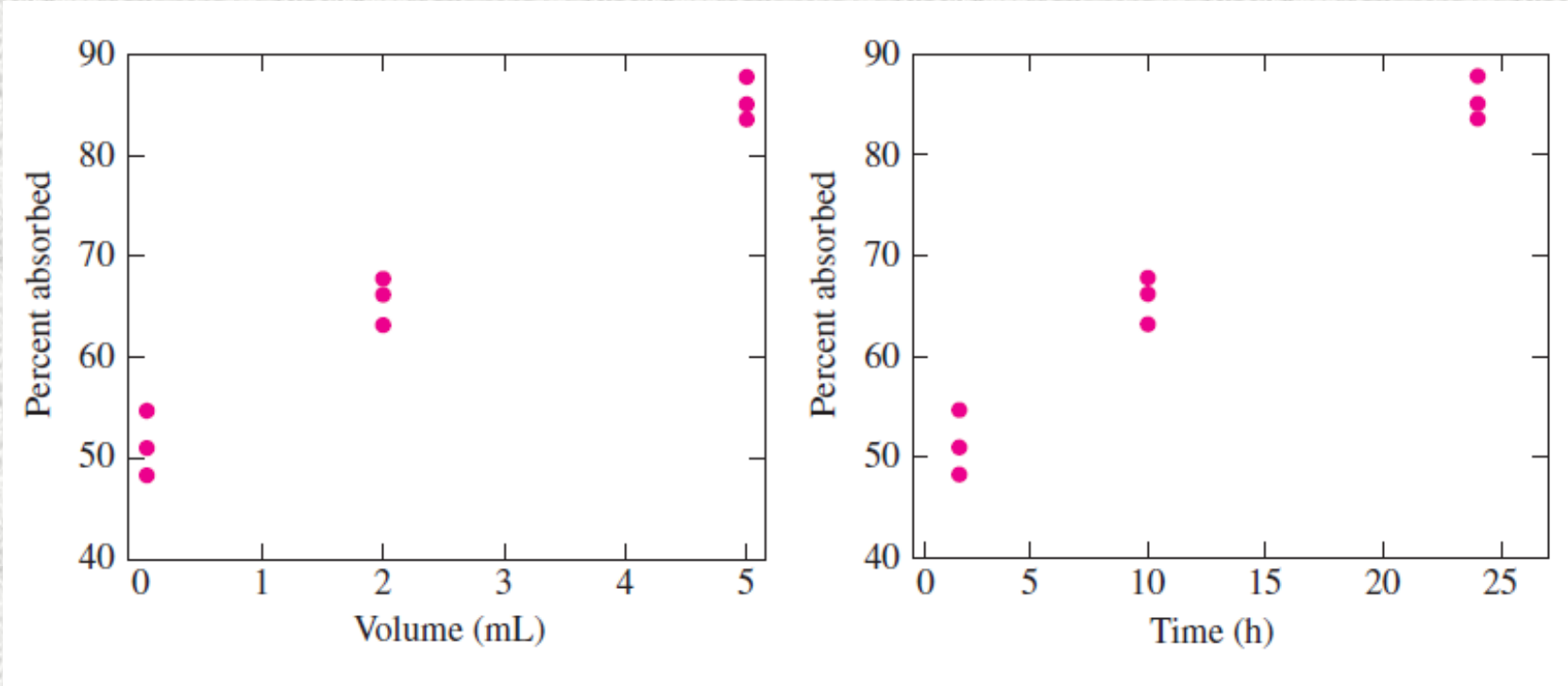
Note:Before we can conclude that two variables have a causal relationship, we must rule out the possibility of confounding.

- Sometimes multiple regression can be used to detect confounding.
- Sometimes experiments can be designed so as to reduce the possibility of confounding.

An environmental scientist is studying the rate of absorption of a certain chemical into skin. She places differing volumes of the chemical on different pieces of skin and allows the skin to remain in contact with the chemical for varying lengths of time. She then measures the volume of chemical absorbed into each piece of skin. She obtains the results shown in the following table.

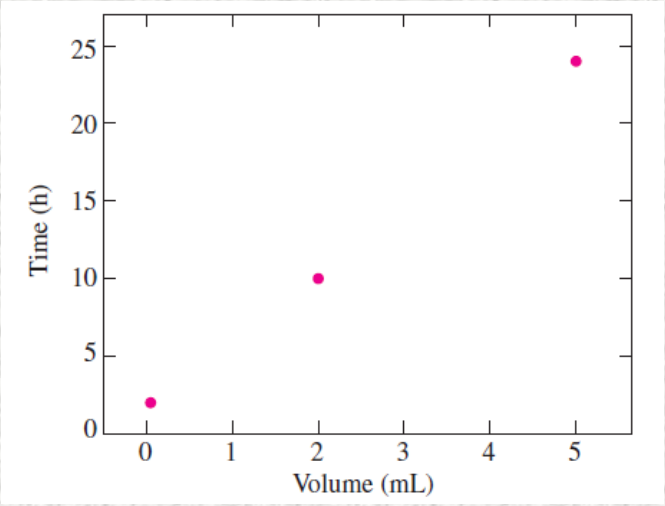
Volume (mL)	Time (h)	Percent Absorbed
0.05	2	48.3
0.05	2	51.0
0.05	2	54.7
2.00	10	63.2
2.00	10	67.8
2.00	10	66.2
5.00	24	83.6
5.00	24	85.1
5.00	24	87.8

The scientist plots the percent absorbed against both volume and time, as shown in the following figure. She calculates the correlation between volume and absorption and obtains $r = 0.988$. She concludes that increasing the volume of the chemical causes the percentage absorbed to increase. She then calculates the correlation between time and absorption, obtaining $r = 0.987$. She concludes that increasing the time that the skin is in contact with the chemical causes the percentage absorbed to increase as well. Are these conclusions justified?



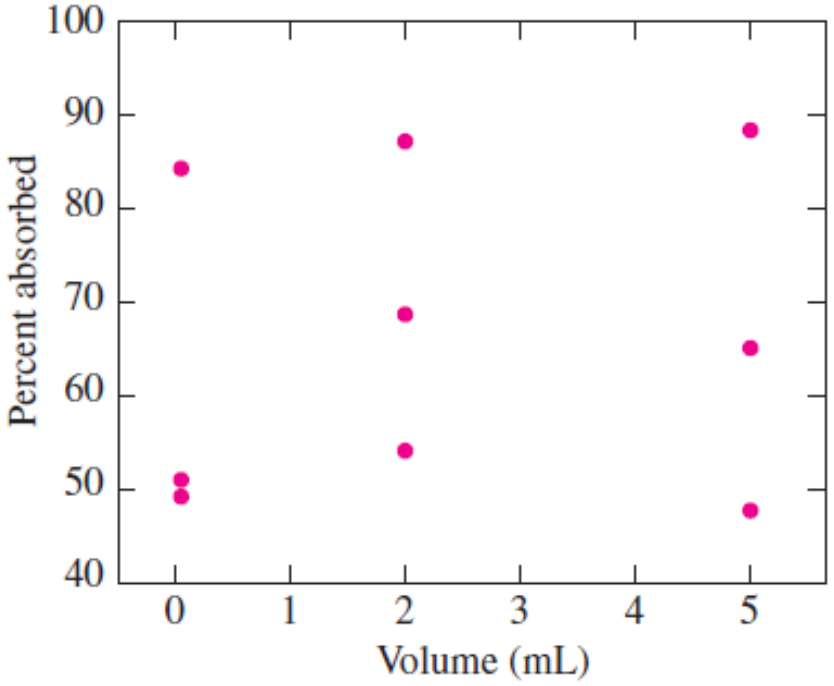
Solution

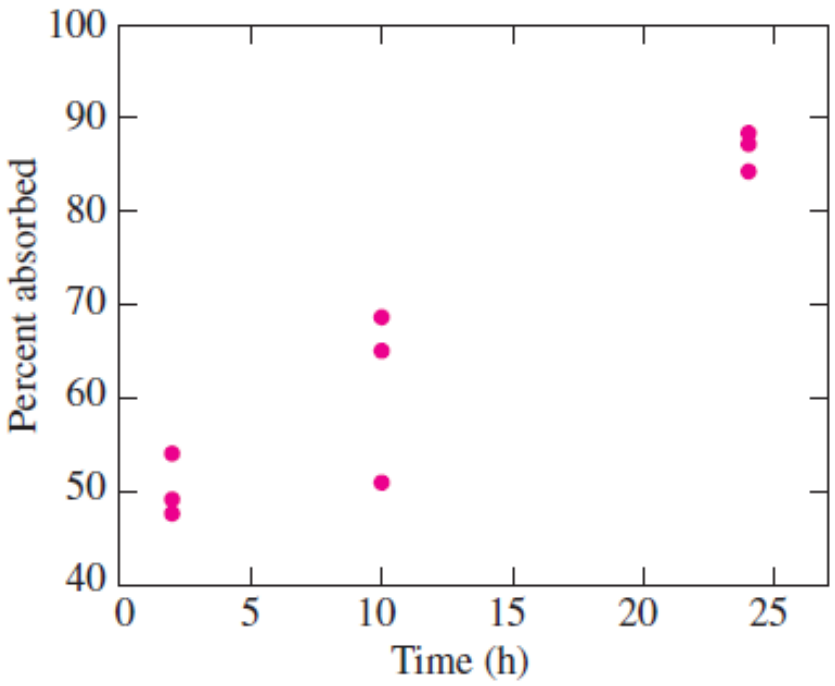
- No. The scientist should look at the plot of time versus volume,
- The correlation between time and volume is $r = 0.999$, so these two variables are almost completely confounded.
- If *either* time or volume affects the percentage absorbed, *both* will appear to do so, because they are highly correlated with each other.
- For this reason, it is impossible to determine whether it is the time or the volume that is having an effect.
- This relationship between time and volume resulted from the design of the experiment and should have been avoided.



The scientist has repeated the experiment, this time with a new design.
The results are presented in the following table.

Volume (mL)	Time (h)	Percent Absorbed
0.05	2	49.2
0.05	10	51.0
0.05	24	84.3
2.00	2	54.1
2.00	10	68.7
2.00	24	87.2
5.00	2	47.7
5.00	10	65.1
5.00	24	88.4

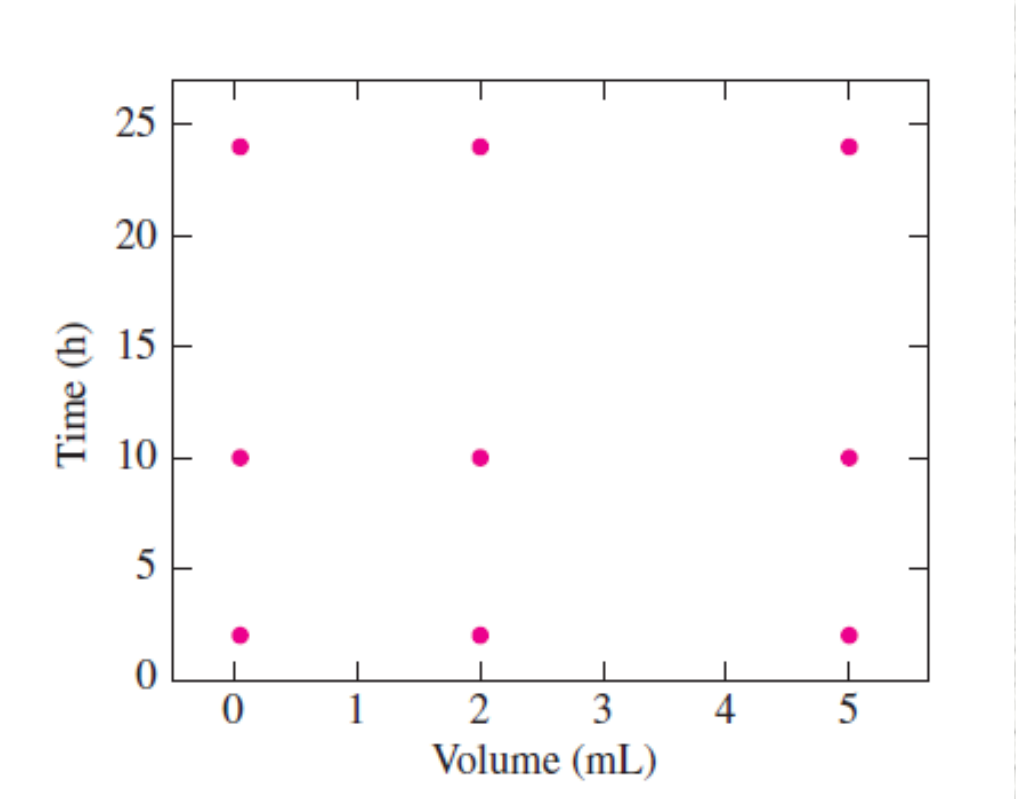




The scientist plots the percent absorbed against both volume and time, She then calculates the correlation between volume and absorption and obtains r =0.121. She concludes that increasing the volume of the chemical has little or no effect on the percentage absorbed. She then calculates the correlation between time and absorption and obtains r = 0.952. She concludes that increasing the time that the skin is in contact with the chemical will cause the percentage absorbed to increase.Are these conclusions justified?

This experiment has been designed so that time and volume are uncorrelated.

It now appears that the time, but not the volume, has an effect on the percentage absorbed. Before making a final conclusion that increasing the time actually causes the percentage absorbed to increase, the scientist must make sure that there are no other potential confounders around



Controlled Experiments Reduce the Risk of Confounding

- In controlled experiments, confounding can often be avoided by choosing values for factors in a way so that the factors are uncorrelated.
- confounding is difficult to avoid in observational studies and they must generally be repeated a number of times, under a variety of conditions, before reliable conclusions can be drawn

The Least-Squares Line

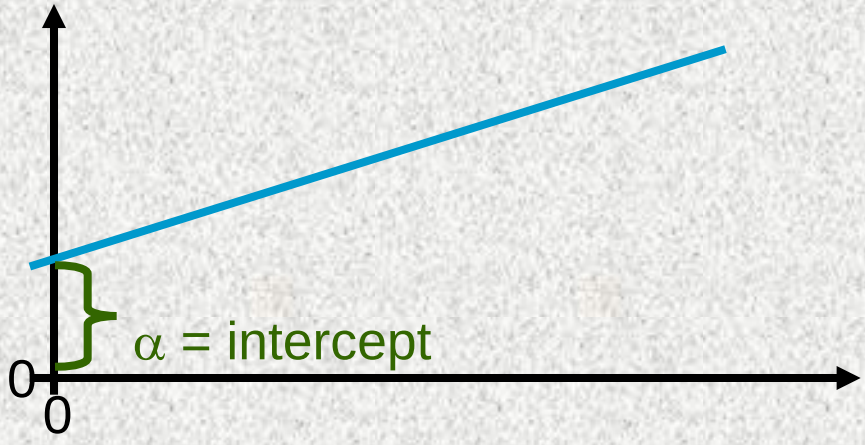
When two variables have a linear relationship, the scatterplot tends to be clustered around a line known as the least-squares line

Linear Deterministic Model

- Denote x as **independent variables** and y as **dependent variable**. For example, x=age and y=%fat in the first example; x=advertising expenditure, y=sales in the second example

- We want to find how y depends on x, or how to predict y using x

- One of the simplest deterministic mathematical relationship between two variables x and y is a linear relationship $y = \alpha + \beta x$



α: intercept

β: slope

Reality

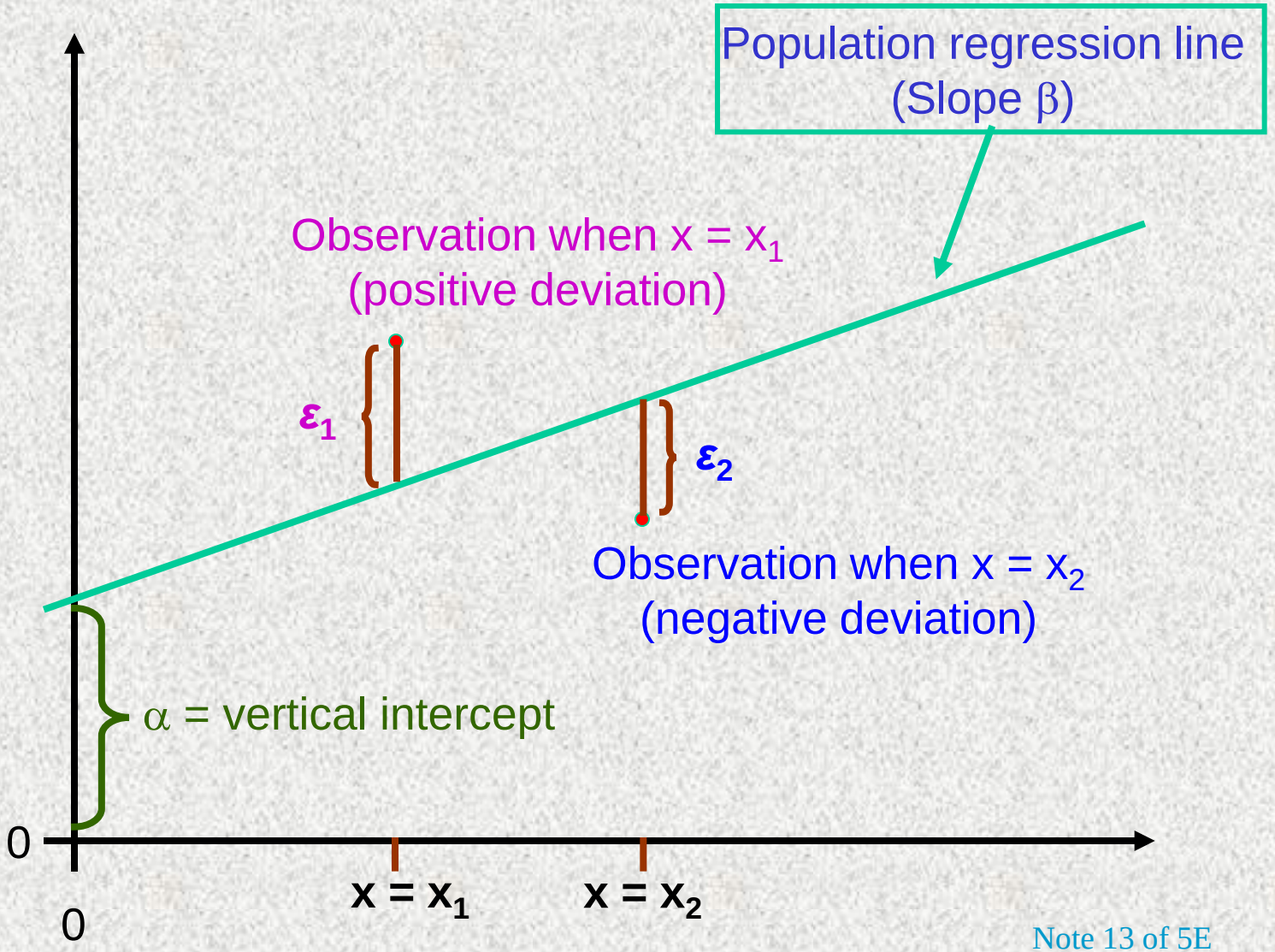
- In the real world, things are never so clean!
  - Age influences fatness. But it is not the sole influence. There are other factors such as gender, body type and random variation (e.g. measurement error)
  - Other factors such as time of year, state of economy and size of inventory, besides the advertising expenditure, can influence the sale
- Observations of (x, y) do not fall on a straight line

As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality. **Albert Einstein**

Probabilistic Model

- Probabilistic model:
  y = deterministic model + random error
- Random error represents random fluctuation from the deterministic model
- The probabilistic model is assumed for the population
- **Simple linear regression model**:
  $y = \alpha + \beta x + \varepsilon$
- Without the random deviation ε, all observed points (x, y) points would fall exactly on the deterministic line. The inclusion of ε in the model equation allows points to deviate from the line by random amounts.

Simple Linear Regression Model



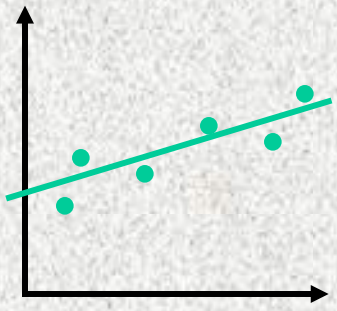
Basic Assumptions of the Simple Linear Regression Model

1. The distribution of ε at any particular x value has mean value 0.
2. The standard deviation of ε is the same for any particular value of x. This standard deviation is denoted by σ.
3. The distribution of ε at any particular x value is normal.
4. The random errors are independent of one another.

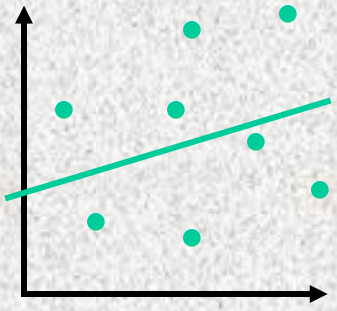
Interpretation of Terms

1. The line $\alpha + \beta x$ describes **average** value of y for any fixed value of x.
2. The **slope** β of the population regression line is the **average** change in y associated with a 1-unit increase in x.
3. The **intercept** α is the height of the population line when $x = 0$.
4. The size of σ determines the extent to which (x, y) observations deviate from the population line.

Small σ

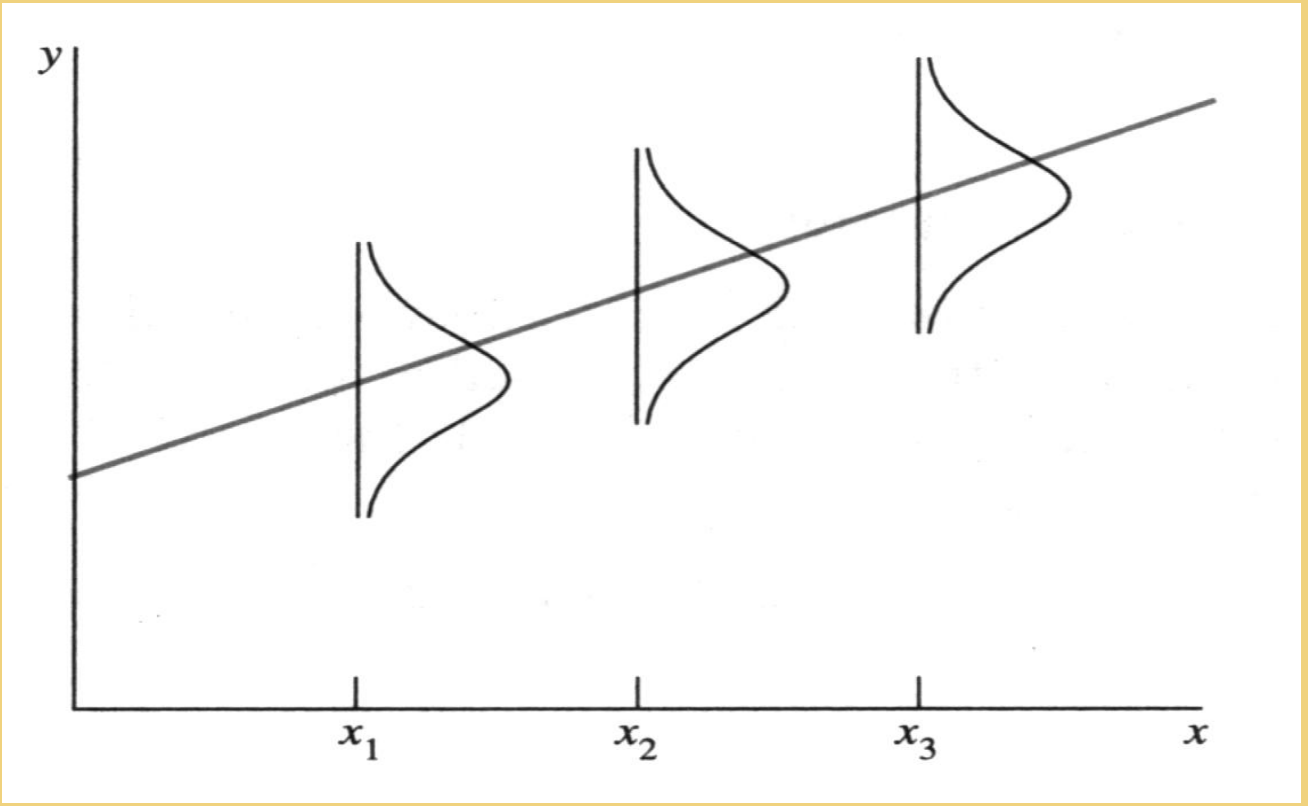


Large σ



The Distribution of y

For any fixed x, y has normal distribution with mean $\alpha + \beta x$ and standard deviation σ.



Data

1. So far we have described the population probabilistic model.
2. Usually three population parameters, α, β and σ, are unknown. We need to estimate them from data.
3. Data: n pairs of observations of independent and dependent variables

$$(x_1,y_1),\ (x_2,y_2),\ \ldots,\ (x_n,y_n)$$

4. Probabilistic model

$$y_i = \alpha + \beta\, x_i + \varepsilon_i, \qquad i=1,\ldots,n$$

ε_i are independent normal with mean 0 and standard deviation σ.

Steps in Regression Analysis

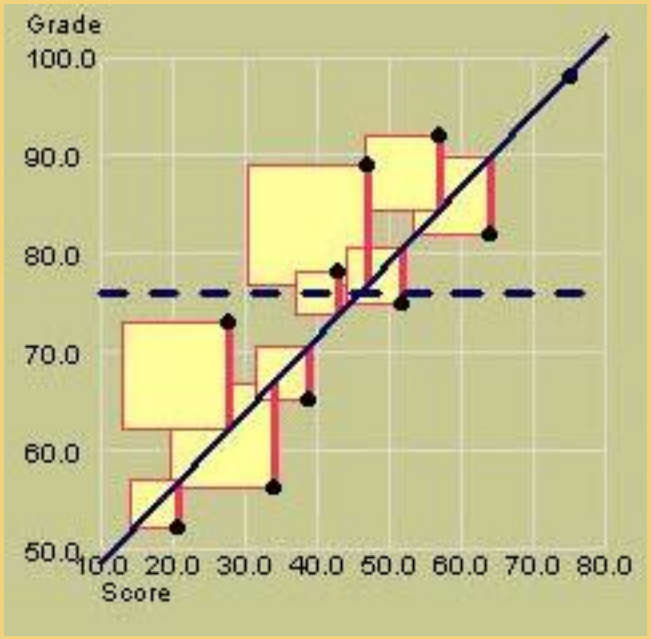
When you perform simple regression analysis, use a step-by step approach:

1. Fit the model to data – estimate parameters.
2. Use the analysis of variance F test (or t test) and r^2 to determine how well the model fits the data.
3. Use diagnostic plots to check for violation of the regression assumptions.
4. Proceed to estimate or predict the quantity of interest

We now discuss statistical methods for each step and use the fatness example to illustrate each step

The Method of Least Squares

- The equation of the best-fitting line is calculated using n pairs of data (x_i, y_i).
- We choose our estimates $\hat{\alpha}$ and $\hat{\beta}$ to estimate α and β so that the vertical distances of the points from the line, are minimized.



Best fitting line : $\hat{y} = \hat{\alpha} + \hat{\beta}x$

Choose $\{\hat{\alpha}$ and $\{\hat{\beta}$

to minimize SSE $= \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$

Least Squares Estimators

Compute $\bar{x} = \dfrac{\sum x_i}{n}, \quad \bar{y} = \dfrac{\sum y_i}{n},$

$S_{xx} = \sum x_i^{\,2} - \dfrac{(\sum x_i)^2}{n}, \quad S_{yy} = \sum y_i^{\,2} - \dfrac{(\sum y_i)^2}{n},$

$S_{xy} = \sum x_i\, y_i - \dfrac{(\sum x_i)(\sum y_i)}{n}.$ Then

$\hat{\beta}$ = point estimator of $\beta = \dfrac{S_{xy}}{S_{xx}}$

$\hat{\alpha}$ = point estimator of $\alpha = \bar{y} - \hat{\beta}\,\bar{x}$

Example: Age and Fatness

The following data was collected in a study of age and fatness in humans.

Age	23	23	27	27	39	41	45	49	50
% Fat	9.5	27.9	7.8	17.8	31.4	25.9	27.4	25.2	31.1

Age	53	53	54	56	57	58	58	60	61
% Fat	34.7	42	29.1	32.5	30.3	33	33.8	41.1	34.5

One of the questions was, "What is the relationship between age and fatness?"

* Mazess, R.B., Peppler, W.W., and Gibbons, M. (1984) Total body composition by dual-photon (^{153}Gd) absorptiometry. *American Journal of Clinical Nutrition*, **40**, *834-839*

Example: Age and Fatness

Age (x)	% Fat y	x^2	xy
23	9.5	529	218.5
23	27.9	529	641.7
27	7.8	729	210.6
27	17.8	729	480.6
39	31.4	1521	1224.6
41	25.9	1681	1061.9
45	27.4	2025	1233
49	25.2	2401	1234.8
50	31.1	2500	1555
53	34.7	2809	1839.1
53	42	2809	2226
54	29.1	2916	1571.4
56	32.5	3136	1820
57	30.3	3249	1727.1
58	33	3364	1914
58	33.8	3364	1960.4
60	41.1	3600	2466
61	34.5	3721	2104.5
834	515	41612	25489.2

$$n = 18$$

$$\sum X = 834$$

$$\sum y = 515$$

$$\sum X^2 = 41612$$

$$\sum XY = 25489.2$$

Example: Age and Fatness

$$n = 18,\ \sum x = 834,\ \sum y = 515$$

$$\sum x^2 = 41612,\ \sum xy = 25489.2$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$= 41612 - \frac{834^2}{18} = 2970$$

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}$$

$$= 25489.2 - \frac{(834)(515)}{18} = 1627.53$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

$$= \frac{1627.53}{2970}$$

$$= .55$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

$$= \frac{515}{18} - .55\,\frac{834}{18}$$

$$= 3.22$$

$$\hat{y} = 3.22 + .55\,x$$

The Analysis of Variance

- The total variation in the experiment is measured by the **total sum of squares**:

$$\text{Total SS} = S_{yy} = \sum (y - \bar{y})^2$$

- The Total SS is divided into two parts:

✓SSR (**sum of squares for regression**): measures the variation explained by including the independent variable *x* in the model.

✓SSE (**sum of squares for error**): measures the leftover variation not explained by *x*.

$$\text{SSR} = \frac{(S_{xy})^2}{S_{xx}} \qquad \text{SSE} = \text{Total SS} - \text{SSR}$$

Coefficient of Determination

The coefficient of determination is defined as

$$r^2 = \frac{SSR}{\text{Total SS}} = 1 - \frac{SSE}{\text{Total SS}}$$

- r^2 is the square of correlation coefficient
- r^2 is a number between zero and one and a value close to zero suggests a poor model.
- It gives the proportion of variation in y that can be attributed to an approximate linear relationship between x and y.
- A very high value of r^2 can arise even though the relationship between the two variables is non-linear. The fit of a model should never simply be judged from the r^2 value alone.

Estimate of σ

An estimator of the variance σ^2 is

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = MSE$$

Thus, an estimator of the standard deviation σ is

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$$

Example: Age and Fatness

$$\text{Total SS} = \sum y^2 - \frac{(\sum y)^2}{n} = 16156.3 - \frac{515^2}{18} = 1421.58$$

$$\text{SSR} = \frac{(S_{xy})^2}{S_{xx}} = \frac{1627.53^2}{2970} = 891.27$$

$$\text{SSE} = \text{Total SS} - \text{SSR} = 1421.58 - 891.27 = 529.71$$

$$r^2 = 1 - \frac{\text{SSE}}{\text{Total SS}} = 1 - \frac{529.71}{1421.58} = .627$$

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n - 2} = \frac{529.71}{18 - 2} = 33.11$$

$$\hat{\sigma} = \sqrt{33.11} = 5.75$$

Example: Age and Fatness

- With r^2=0.627 or 62.7%, we can say that 62.7% of the observed variation in %Fat can be attributed to the probabilistic linear relationship with human age.
- The magnitude of a typical sample deviation from the least squares line is about 5.75(%) which is reasonably large compared to the y values themselves.
- This would suggest that the model is only useful in the sense of provide gross ballpark estimates for %Fat for humans based on age.

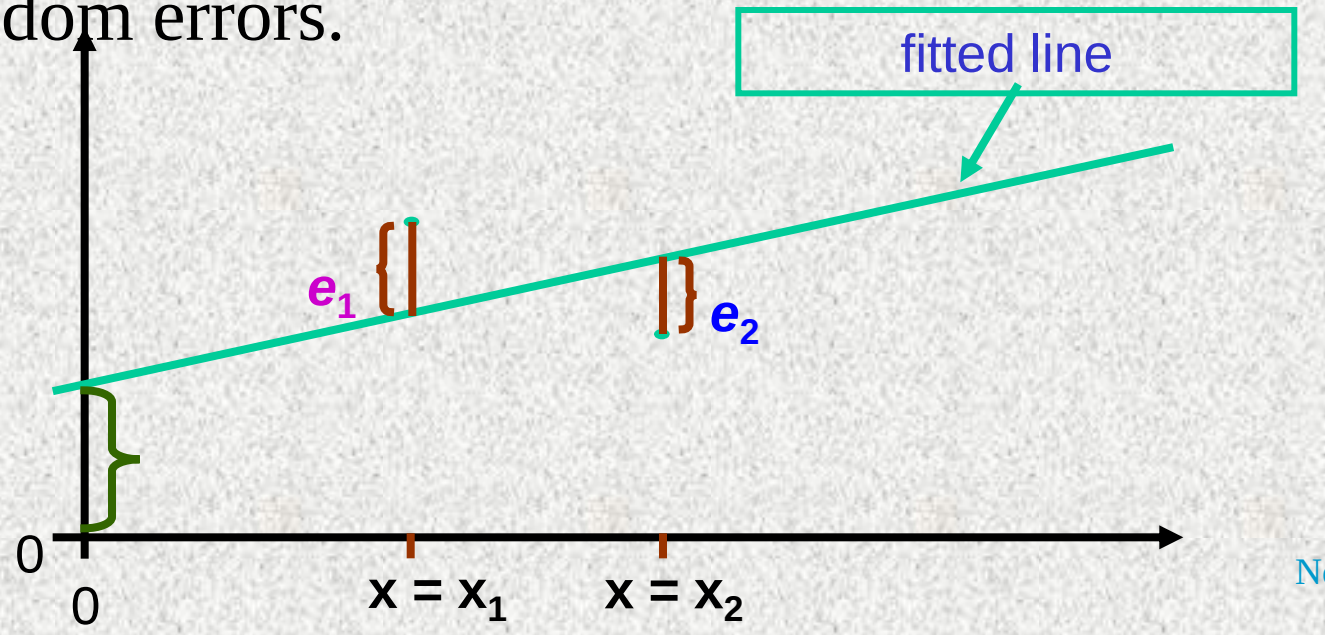
Checking the Regression Assumptions

Remember that the results of a regression analysis are only valid when the necessary assumptions have been satisfied.

1. The relationship between x and y is linear, given by $y = \alpha + \beta x + \varepsilon$.
2. The random error terms ε are independent and, for any value of x, have a normal distribution with mean 0 and variance σ^2.

Residuals

- The residual corresponding to (x_i, y_i) is $e_i = y_i - \hat{\alpha} - \hat{\beta}x_i$
- The **residual** is the "leftover" variation in each data point after the variation explained by the regression model has been removed.
- Residuals reflect random errors, thus we can use them to check violations in the assumptions about random errors.

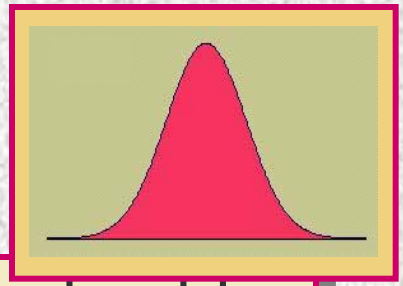


Diagnostic Tools – Residual Plots

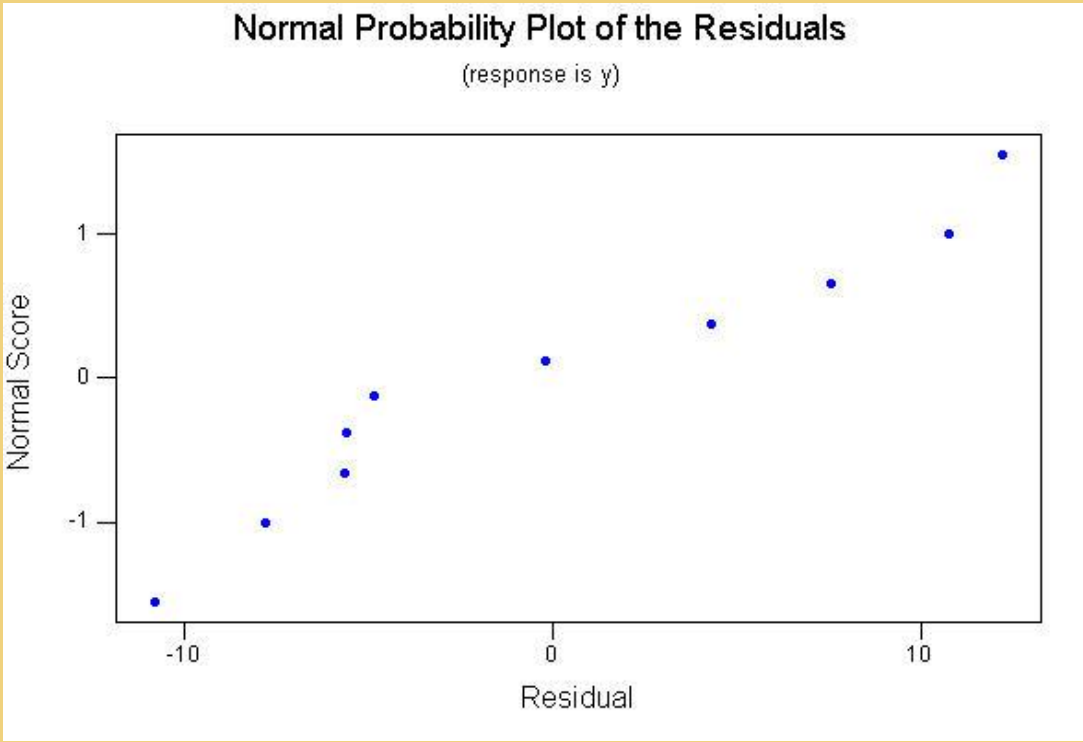
We can check the normality and equal variance assumptions using

1. **Normal probability plot** of residuals
2. Plot of **residuals versus fit** or **residuals versus variables**

Normal Probability Plot

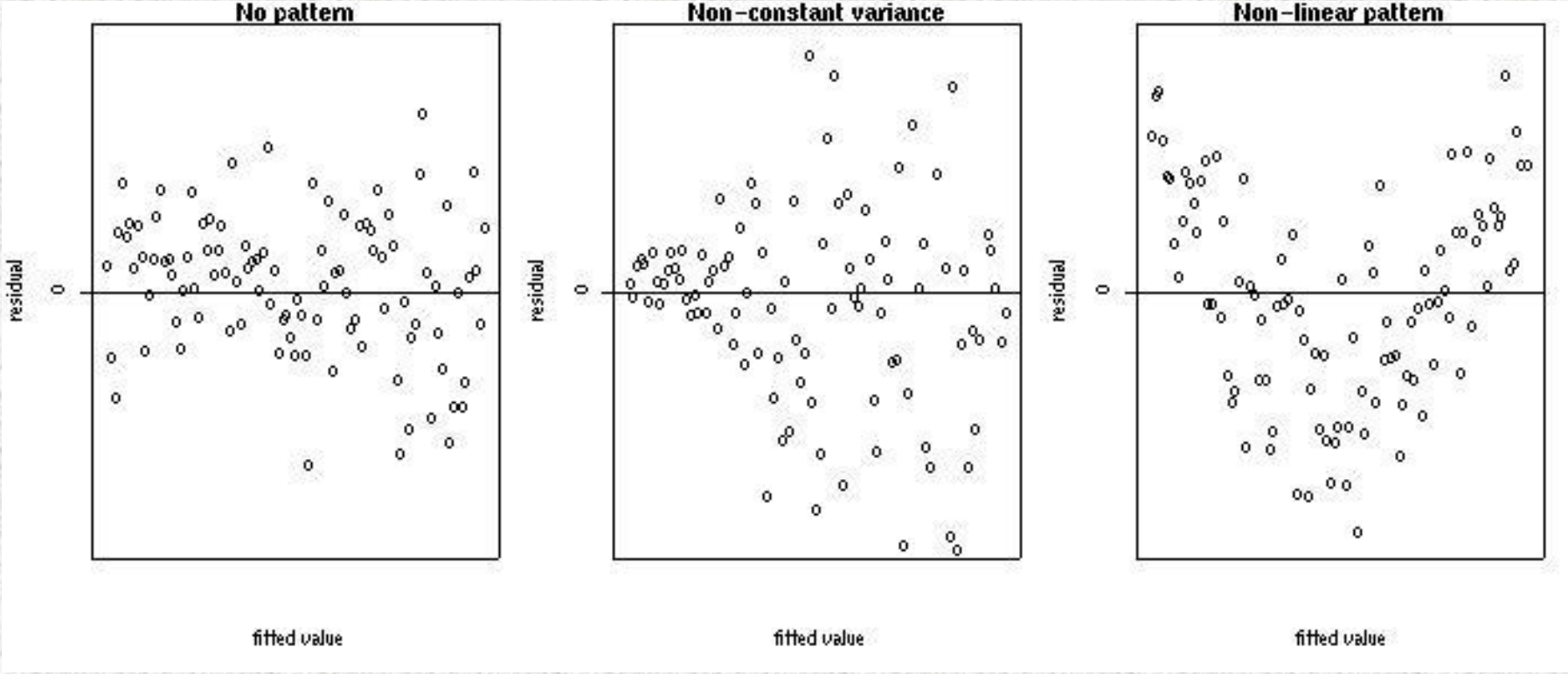


- If the normality assumption is valid, the plot should resemble a straight line, sloping upward to the right.
- If not, you will often see the pattern fail in the tails of the graph.



Residuals versus Fits

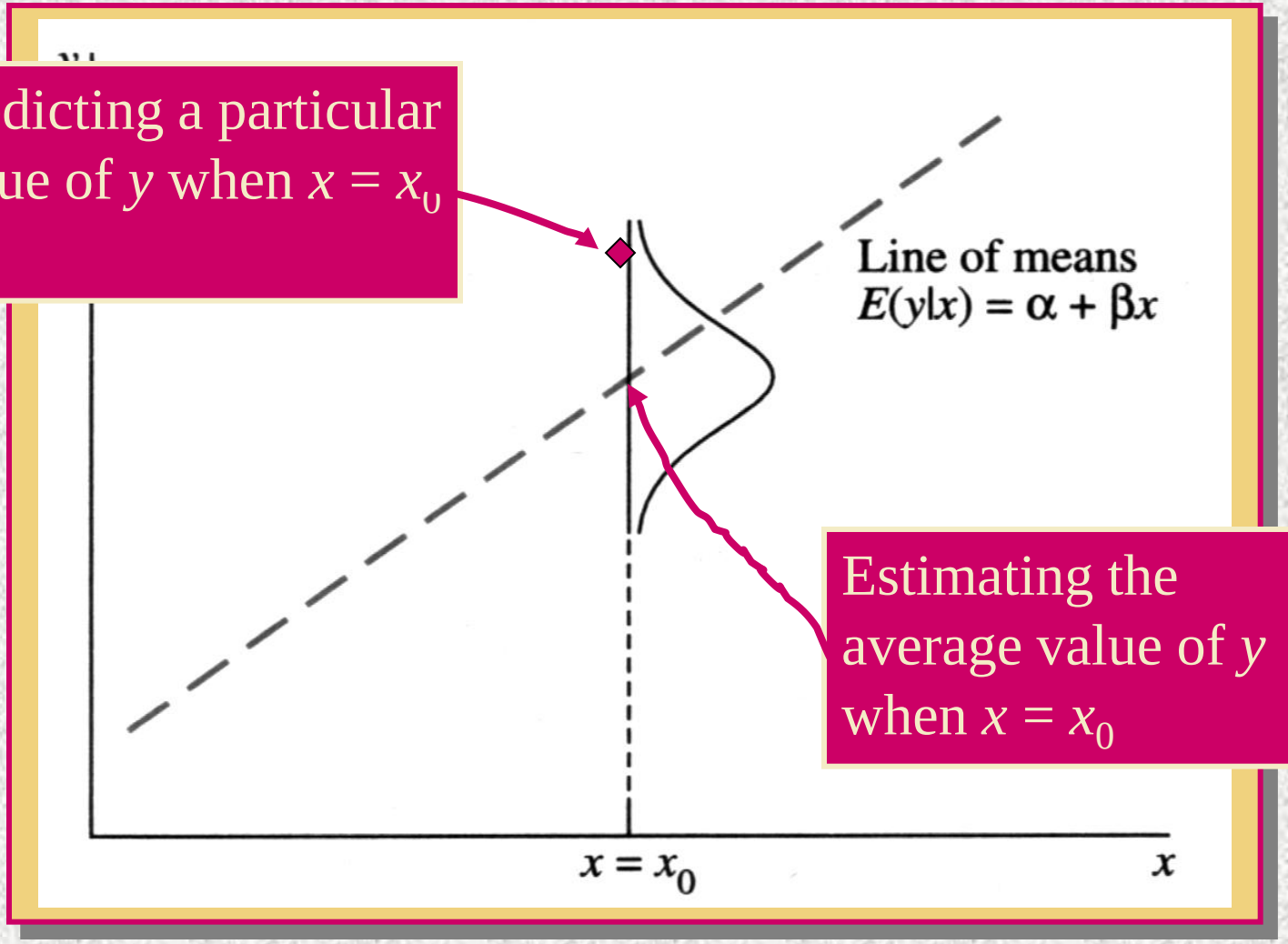
- ✓ If the equal variance assumption is valid, the plot should appear as a random scatter around the zero center line.
- ✓ If not, you will see a pattern in the residuals.



Estimation and Prediction

- Once we have
  - ✓ determined that the regression line is useful
  - ✓ used the diagnostic plots to check for violation of the regression assumptions.
- We are ready to use the regression line to
  - ✓ Estimate the average value of y for a given value of x
  - ✓ Predict a particular value of y for a given value of x.

Estimation and Prediction

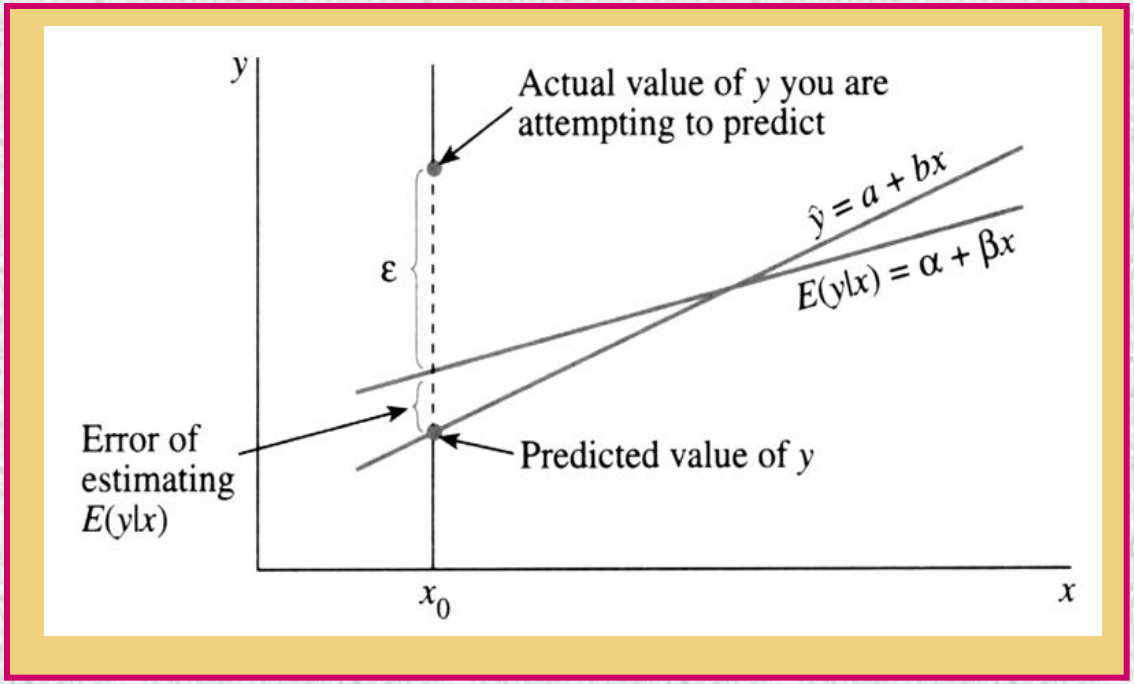


Estimation and Prediction

- The best estimate of the average value of *y* and prediction of y for a given value $x = x_0$ is

$$\hat{y} = \hat{\alpha} + \hat{\beta}x_0$$

- The prediction of *y* is more difficult, requiring a wider range of values in the prediction interval.



Estimation and Prediction

To estimate the average value of y when $x = x_0$:

$$\hat{y} \pm t_{\alpha/2}\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

To predict a particular value of y when $x = x_0$:

$$\hat{y} \pm t_{\alpha/2}\hat{\sigma}\sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Example: Age and Fatness

The fitted line is

$$\hat{y} = 3.22 + 0.548x$$

If x_0=45 is put into the equation for x, we have both an estimated average %Fat for 45 year old humans and a predicted %Fat for a 45 year old human

$$3.22+0.548(45)=27.9\%$$

The two interpretations are quite different.

Example: Age and Fatness

95% confidence interval for the average of y at $x_0 = 45$ is

$$\hat{y} \pm 2.12 \times 5.75 \sqrt{\left(\frac{1}{18} + \frac{(45 - 834/18)^2}{2970} \right)} = 27.9 \pm 2.88$$

or $(25.2, 30.78)$.

95% confidence interval for the prediction of y at $x_0 = 45$ is

$$\hat{y} \pm 2.12 \times 5.75 \sqrt{\left(1 + \frac{1}{18} + \frac{(45 - 834/18)^2}{2970} \right)} = 27.9 \pm 12.53$$

or $(15.37, 40.43)$. The prediction interval is much wider.

Steps in Regression Analysis

When you perform simple regression analysis, use a step-by step approach:

1. Fit the model to data – estimate parameters.
2. Use the analysis of variance F test (or t test) and r^2 to determine how well the model fits the data.
3. Use diagnostic plots to check for violation of the regression assumptions.
4. Proceed to estimate or predict the quantity of interest

Minitab Output

Least squares regression line

Regression Analysis: y versus x

The regression equation is y = 40.8 + 0.766 x

Predictor	Coef	SE Coef	T	P
Constant	40.784	8.507	4.79	0.001
x	0.7656	0.1750	4.38	0.002

S = 8.704 R-Sq = 70.5% R-Sq(adj) = 66.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1450.0	1450.0	19.14	0.002
Residual Error	8	606.0	75.8		
Total	9	2056.0			

$\sqrt{MSE}$

Regression coefficients, a and b

$t^2 = F$

Key Concepts

I. **Scatter plot**

Pattern and unusual observations

II. **Correlation coefficient** a numerical measure of the **strength** and **direction.**

Interpretation of r

Key Concepts

III. A Linear Probabilistic Model

1. When the data exhibit a linear relationship, the appropriate model is $y = \alpha + \beta x + \varepsilon$.
2. The random error ε has a normal distribution with mean 0 and variance σ^2.

IV. Method of Least Squares

1. Estimates $\hat{\alpha}$ and $\hat{\beta}$, for α and β, are chosen to minimize SSE, the sum of the squared deviations about the regression line, $\hat{y}=\hat{\alpha}+\hat{\beta}x$.
2. The least squares estimates are $\hat{\alpha}=S_{xy}/S_{xx}$ and $\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}$

Key Concepts

V. Analysis of Variance

1. Total SS = SSR + SSE, where Total SS = S_{yy} and SSR = $(S_{xy})^2 / S_{xx}$.
2. The best estimate of σ^2 is MSE $=$ SSE / $(n - 2)$.

VI. Testing, Estimation, and Prediction

1. A test for the significance of the linear regression—$H_0 : \beta = 0$ can be implemented using statistic

$$t = \frac{\hat{\beta}}{\hat{\sigma}\sqrt{S_{xx}}}$$

Key Concepts

2. The strength of the relationship between x and y can be measured using

$$r^2 = 1 - \frac{\text{SSE}}{\text{Total SS}}$$

which gets closer to 1 as the relationship gets stronger.

3. Use **residual plots** to check for nonnormality, inequality of variances, and an incorrectly fit model.
4. **Confidence intervals** can be constructed to estimate the slope β of the regression line and to estimate the average value of y, for a given value of x.
5. **Prediction intervals** can be constructed to predict a particular observation, y, for a given value of x. For a given x, prediction intervals are always wider than confidence intervals.