Biostatistics for Health Care Researchers: A Short Course

Correlation and Simple Linear Regression

Chandan Saha, PhD

Division of Biostatistics

Indiana University School of Medicine

Objectives

- ✓ Learn how to interpret and use correlation coefficient
- ✓ Understand principle of fitting linear regression model
- ✓ Learn how to interpret estimates of the model parameters and use for prediction

Outline

- Correlation analysis
  - ♦ Pearson correlation and Spearman rank correlation
  - ♦ Properties and interpretation of correlation coefficient
  - ♦ Test of hypothesis, H_0: $\rho = 0$

- Simple linear regression
  - ♦ Describe a linear regression model
  - ♦ Fit the model to observed data
  - ♦ Interpret the fitted model
  - ♦ Verify model assumptions.

What is Correlation

♦ Pearson’s correlation coefficient

-- Parametric, i.e., two variables are assumed to have a bivariate normal distribution

-- measures a **linear** association between two variables

♦ Spearman’s rank correlation coefficient

- non-parametric, i.e., no distributional assumption is made for two variables

-- measures a monotone association

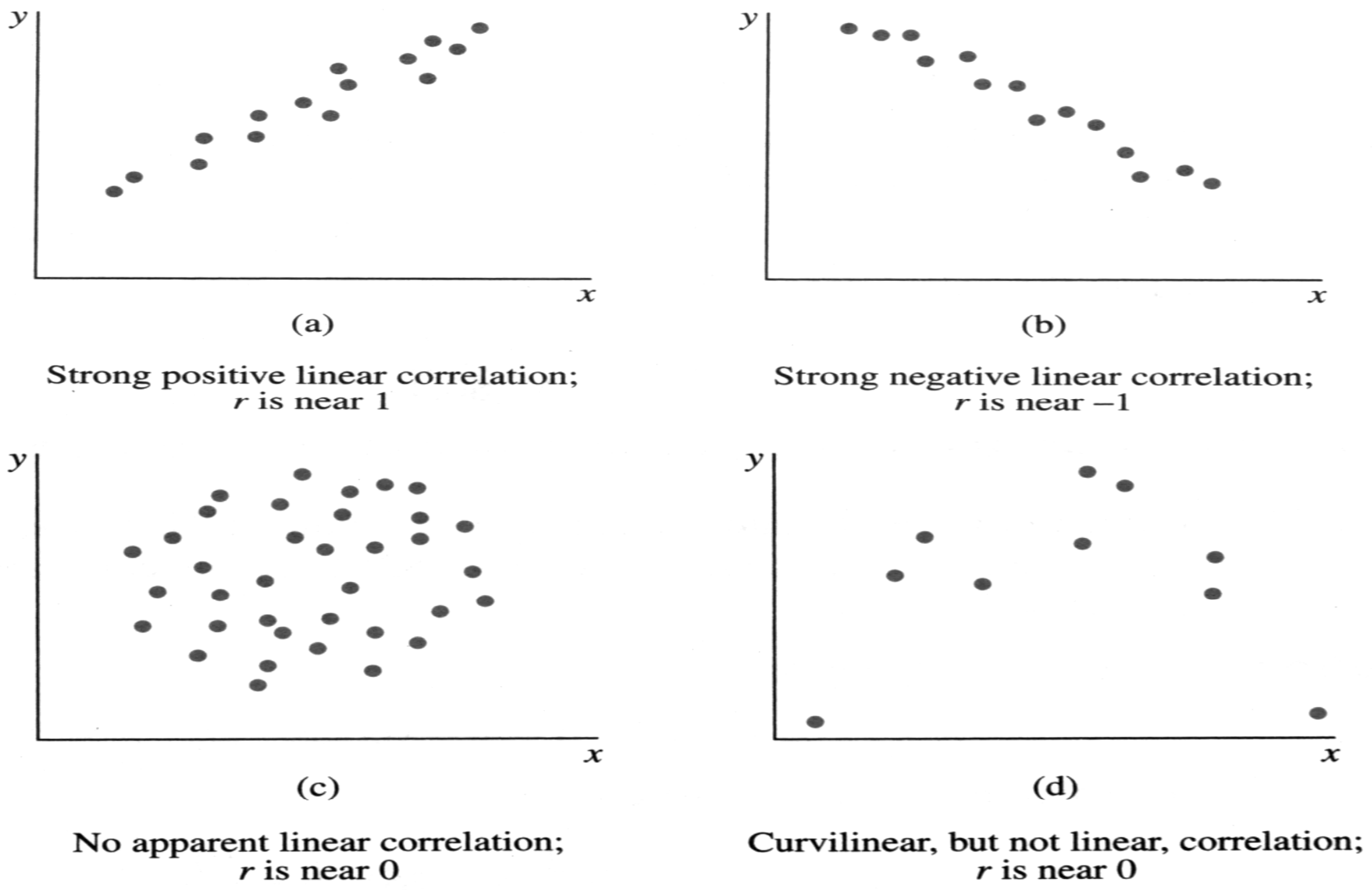
Formula for Computing Correlation

- Sample correlation is denoted by r.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where (x_i, y_i) is a pair of observations for the ith subject and n = sample size.

Scatter plots for Correlation



(a)

Strong positive linear correlation;
r is near 1

(b)

Strong negative linear correlation;
r is near -1

(c)

No apparent linear correlation;
r is near 0

(d)

Curvilinear, but not linear, correlation;
r is near 0

Properties of Pearson Correlation Coefficient

♦ r = 0 does not mean there is no relationship
it means there is **no linear** relationship.

♦ $-1 \leq r \leq 1$.

♦ r does not have any unit.

♦ r is invariant under transformations
-- location (i.e. added constant)
-- scale (i.e. multiplied constant)

An Example

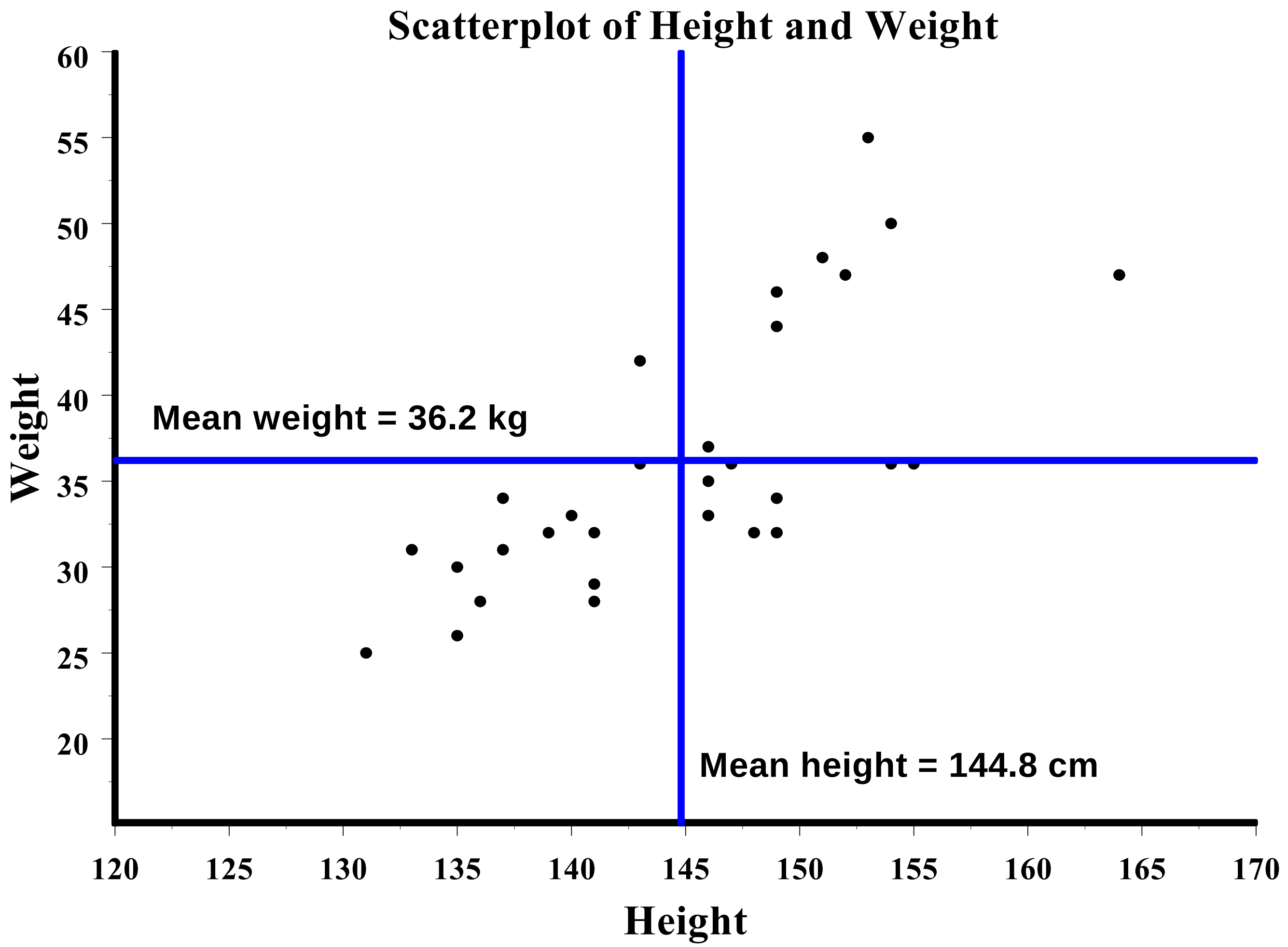
The heights (cm) and weights (kg) of 30 eleven-year-old girls attending Heaton Middle School, Bradford, were measured. (Statistics in Society, 1983).

Height	Weight	Height	Weight	Height	Weight
135	26	146	33	153	55
154	50	139	32	131	25
149	44	137	31	143	36
146	35	141	28	136	28
154	36	151	48	155	36
133	31	149	34	141	32
164	47	146	37	149	46
147	36	152	47	140	33
143	42	148	32	149	32
141	29	137	34	135	30

How are the two variables related?

Test the hypothesis: $H_0: \rho = 0$ vs. $H_A : \rho > 0$.

Scatter Plot of Height and Weight of 30 11-Year-Old Girls



Test of Hypothesis of Population Correlation Coefficient (ρ)

- Null hypothesis: $H_0 : \rho = 0$
- Alternative hypothesis:

One-Tailed Test	Two-Tailed Test
$H_a : \rho > 0$	$H_a : \rho \neq 0$
(or $H_a : \rho < 0$)	

- Test statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

where ***r*** is the sample correlation coefficient.

- Assumptions: (x_i, y_i) have a bivariate normal distribution and samples are independent.
- Under this assumption, the test statistic t has a Student's *t* distribution with (*n* - 2) degrees of freedom.

An Example

♦ Using the formula for *r* as described earlier, *r* = 0.74 suggesting a strong positive correlation between height and weight.

♦ For testing $H_0 : \rho = 0$, $H_A : \rho > 0$

the test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{(0.74237)\sqrt{30-2}}{\sqrt{1-(0.74237)^2}} = 5.86$$

♦ For one tailed test, p-value = Pr ($t > 5.86$) = 0.000001

♦ Using $\alpha = 0.05$ (type I error rate), we reject the null hypothesis that $H_0 : \rho = 0$ since p-value < 0.05.

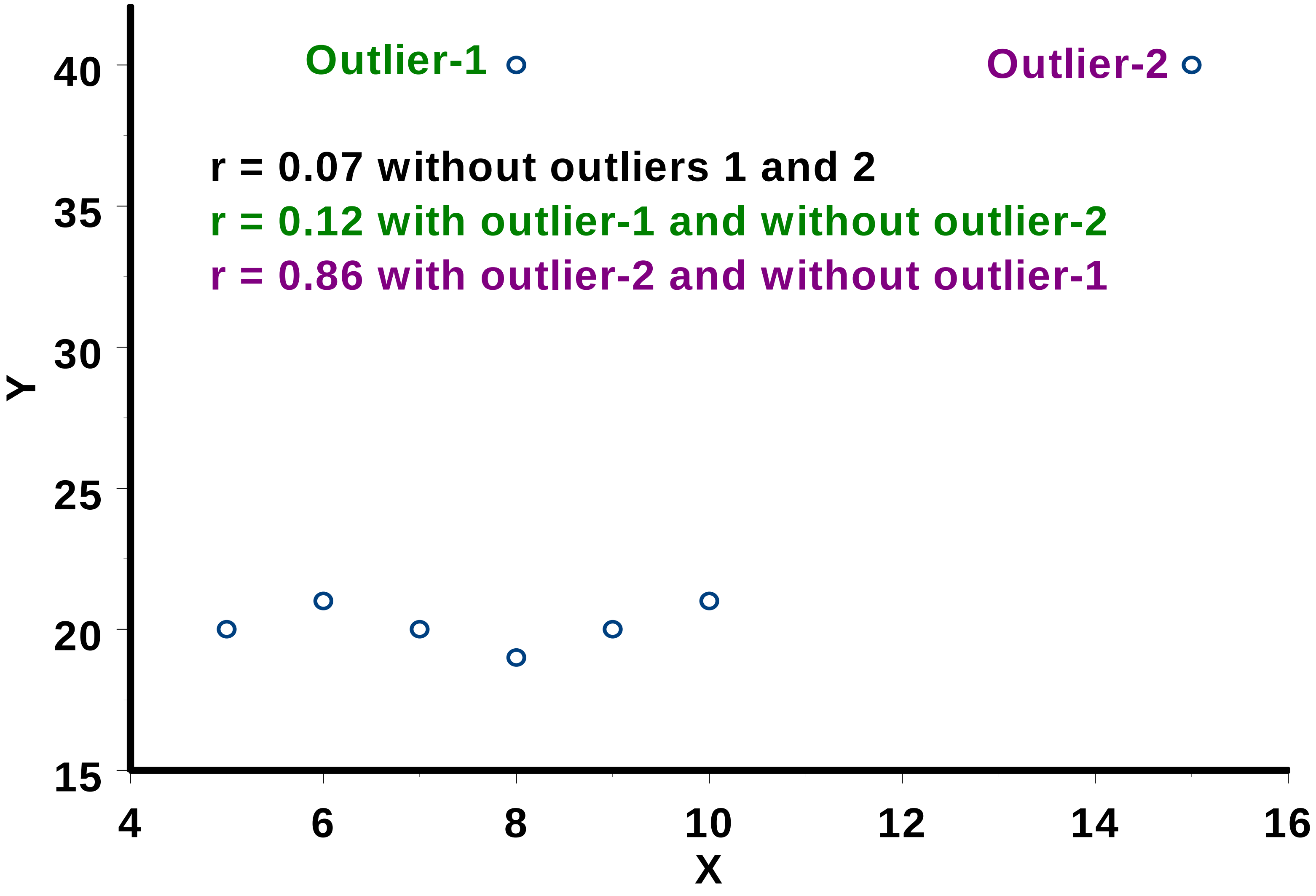
Interpretation of the size of r

- ♦ r, the sample correlation coefficient, can be squared to form the statistic called the coefficient of determination

- ♦ In the previous example *r* was 0.74, so $r^2 = 0.55$

- ♦ $r^2 = 0.55$ means that 55% of the variation in weight may be accounted for by a linear association with height.

When Not to Use r

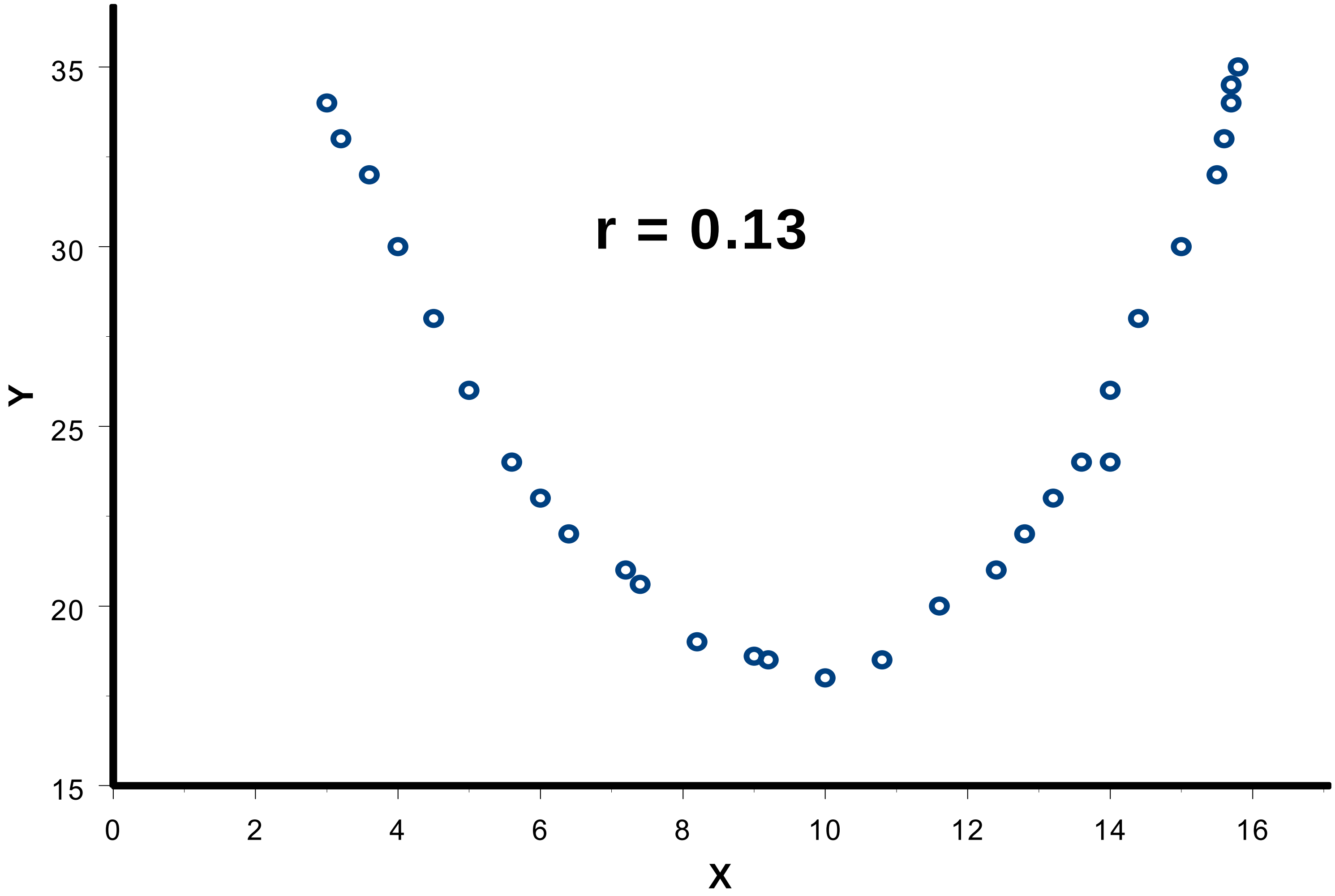


r = 0.07 without outliers 1 and 2

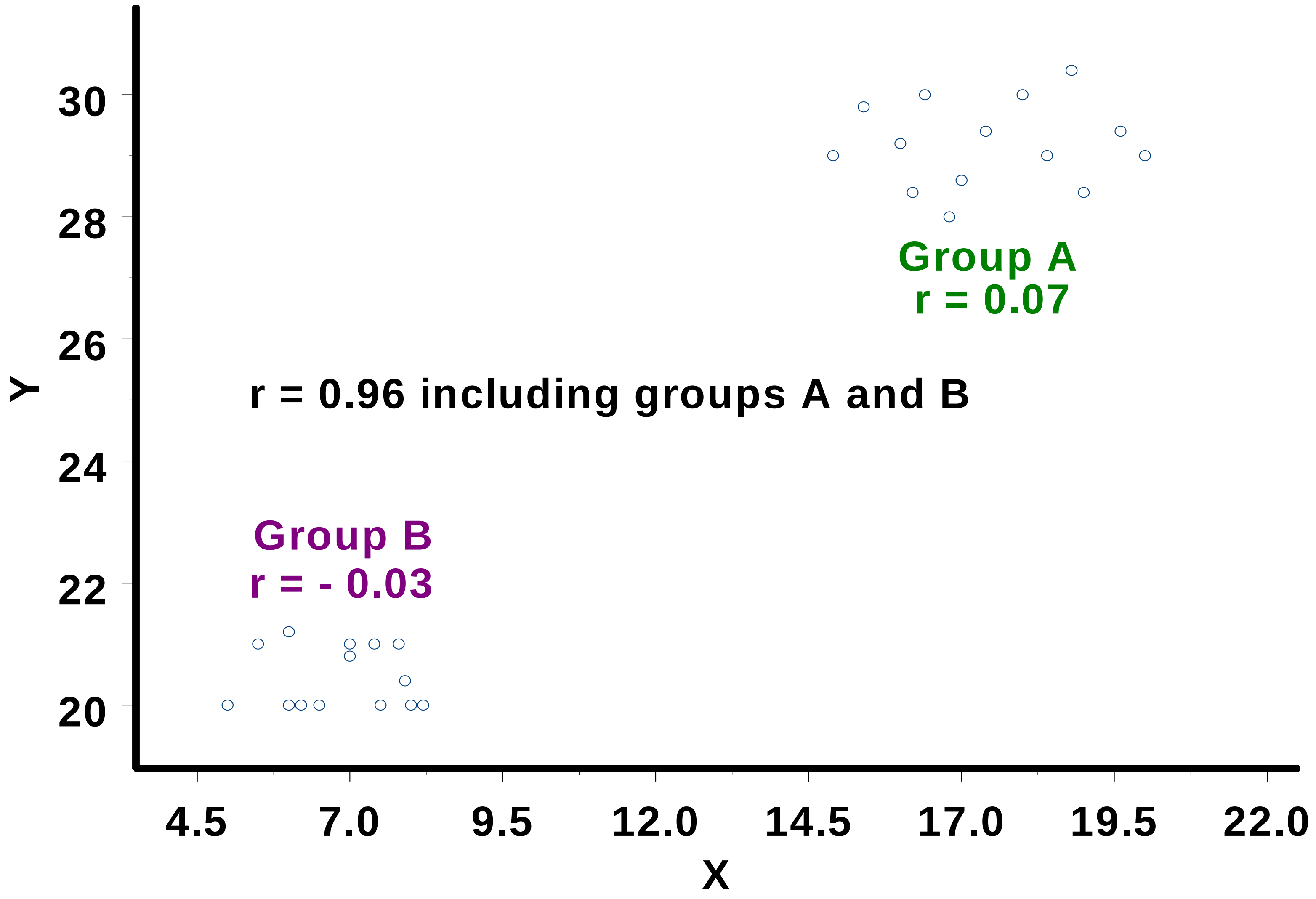
r = 0.12 with outlier-1 and without outlier-2

r = 0.86 with outlier-2 and without outlier-1

When Not to Use r



When Not to Use r



Spearman Rank Correlation (r_s)

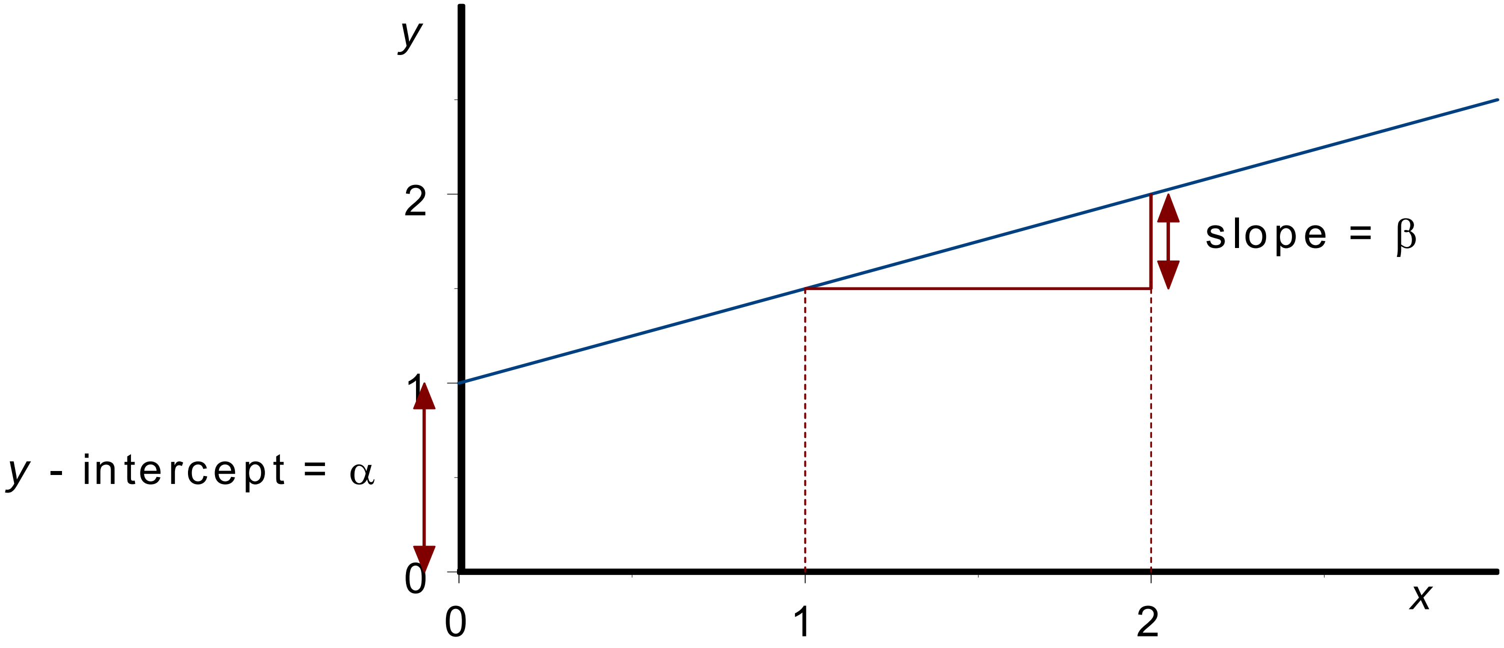
- ♦ Rank observations separately for each of the two variables
- ♦ Use the average ranks of the tied values
- ♦ Calculate Pearson coefficient using ranks in place of actual data

- ♦ Use r_s when
  - --- distributions of the variables are skewed
  - --- the joint distribution of two variables is far from normal
  - --- the data contain extreme values
  - --- the variables are ordinal

● Simple Linear Regression

A simple linear probabilistic model

♦ Use the deterministic model $y = \alpha + \beta x$ where α is the y-intercept, the value of y when $x = 0$ and β is the slope of the line, as shown in Figure below.



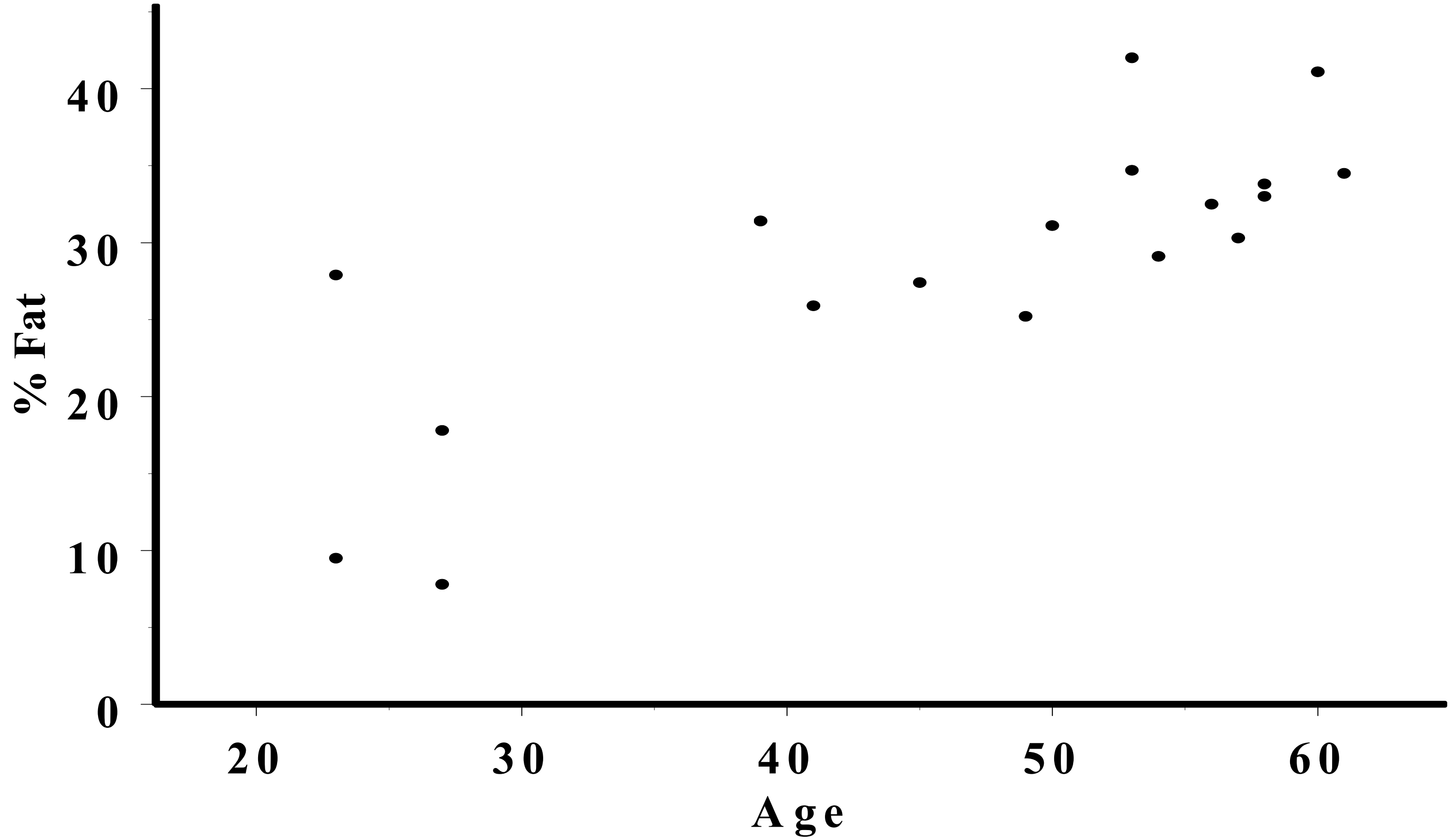
A Case Study

The data come from a study investigating a new method of measuring body composition, and give the body fat percentage (% fat), age and sex for 18 normal adults aged between 23 and 61 years (*American Journal of Clinical Nutrition*, 40, 834-839).

Age	% Fat	Sex	Age	% Fat	Sex
23	9.5	M	23	27.9	F
27	7.8	M	27	17.8	M
39	31.4	F	41	25.9	F
45	27.4	M	49	25.2	F
50	31.1	F	53	34.7	F
53	42.0	F	54	29.1	F
56	32.5	F	57	30.3	F
58	33.0	F	58	33.8	F
60	41.1	F	61	34.5	F

Scatterplot of the case study data

A particular response y is described using the probabilistic model $y = \alpha + \beta\, x + \varepsilon$ *(error term)*



● Assumptions about the random error:

Error (*ε)* terms

- ♦ Are independent in the probabilistic sense.
- ♦ Have a mean of 0 and a common variance equal to σ^2 .
- ♦ Have a normal probability distribution.

● Fitting the linear model

♦ We can use sample information to estimate the values of ***α*** and ***β***, which are the coefficients of the line of means

♦ These estimates are used to form the best-fitting line for a given set of data, called the least squares line or regression line.

$$E(y|x) = \alpha + \beta x$$

Figure 3 : Linear probabilistic model

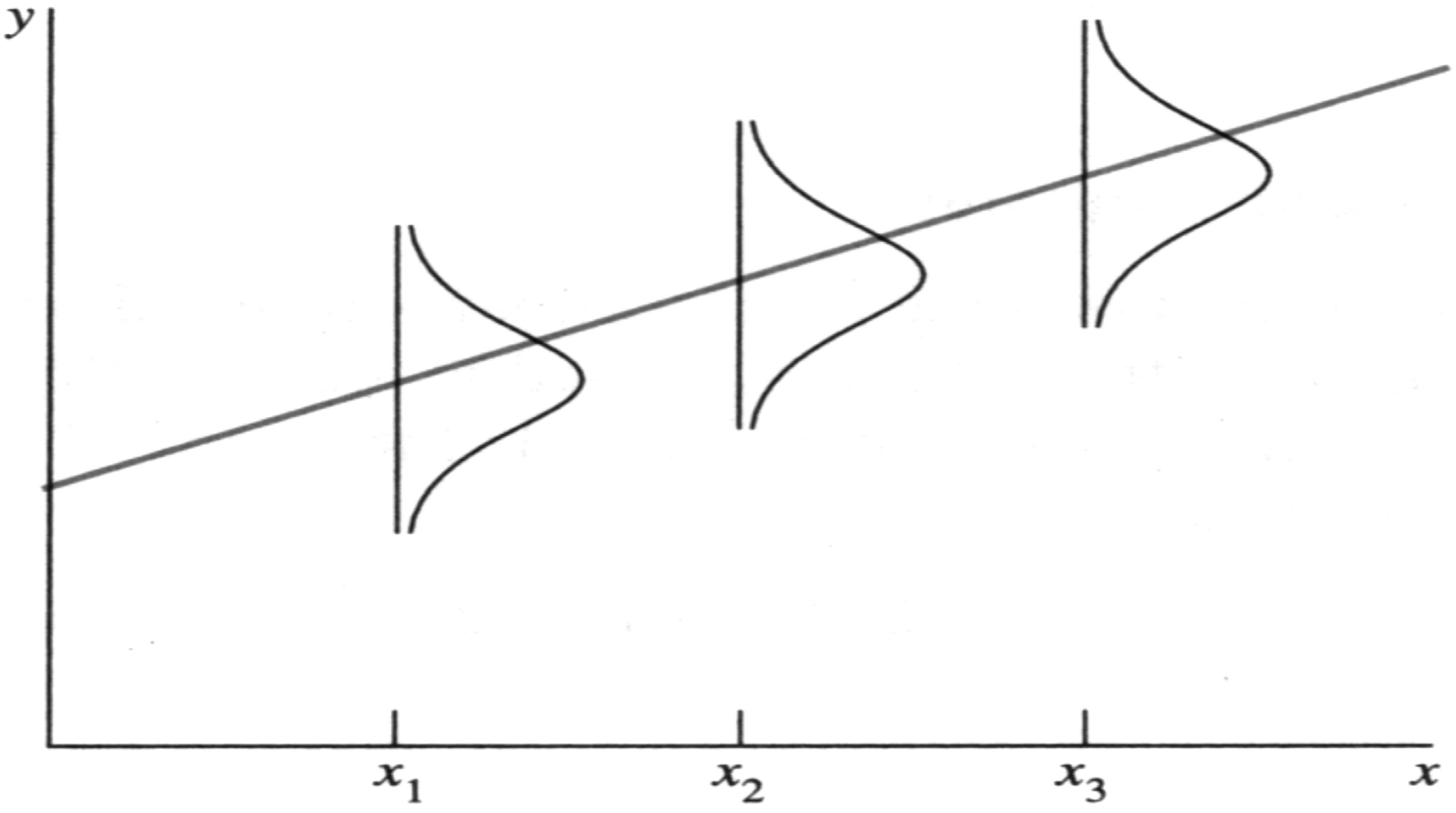
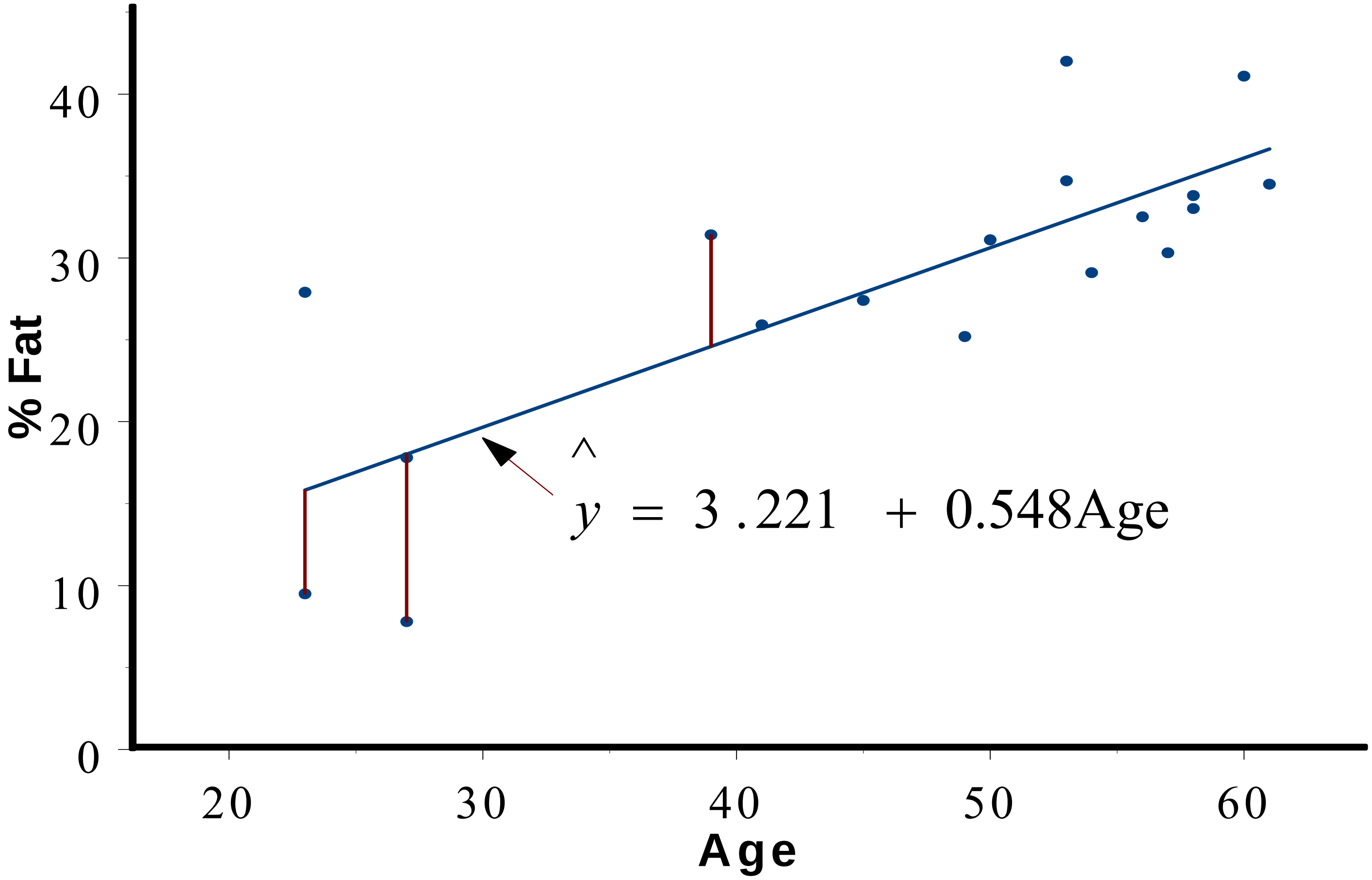


Figure : Graph of the least square line and data points

Principle of Least Squares:

The line that minimizes the sum of squares of the deviations (e.g., vertical lines in the figure) of the observed values of *y* from those predicted is the best-fitting line.



$$\hat{y} = 3.221 + 0.548\text{Age}$$

- **The fitted model:**

  % Fat = 3.221 + 0.548 x Age

  where, 3.221 is an estimate for α and 0.548 is an estimate for β for the population model: E (% Fat) = α + β Age.

- **Interpretation of the fitted regression coefficients:**

  3.221 is the intercept of the fitted line or mean value of % fat for age = 0. However, we do not predict % fat for ages that are beyond the range of the age included in the sample.

  0.548 is the slope of the fitted line. This means for one year increase in age, percent of body fat increases by 0.548 on average.

● Testing the usefulness of the linear regression model

♦ In considering linear regression, you may ask:

♪ Is the independent variable x useful in predicting the response variable y ?

♦ We answer this question by testing the following hypothesis:

$$H_0 : \beta = 0 \qquad \text{versus} \qquad H_A : \beta \neq 0$$

♦ Test statistic: $t = \dfrac{b-\beta}{\text{S.E.}(b)} = \dfrac{b-\beta}{\sqrt{\text{MSE} \,/ \sum (x_i - \bar{x})^2}}$

where $\text{MSE} = \sum (y_i - \hat{y})^2 / (n-2)$

the test statistic t has a t-distribution with (n-2) df.

● Example (case study)

♦ Determine whether there is a significant linear relationship between age and percent fat for the case study data. Test at the 5% level of significance.

♦ The hypotheses to be tested are

$$H_0 : \beta = 0 \qquad \text{versus} \qquad H_A : \beta \neq 0$$

and the observed value of the test statistic is calculated as

$$t = \frac{b - 0}{S.E.(b)} = \frac{0.548 - 0}{0.10558} = 5.19$$

with (n - 2) = 16 degrees of freedom.

♦ p-value = Pr (t > 5.19 or t < -5.19) = 2 x Pr (t > 5.19) = 0.00009

♦ Conclusion: Since p-value < 0.05, we reject the null hypothesis at α=0.05 and conclude that there is a significant linear relationship between age and fat amount.

A (1 - α)100% Confidence Interval (CI) for β :

♦ $b \pm t_{a/2}$ SE(b) where $t_{a/2}$ is based on (n - 2) degrees of freedom and

♦ 95% CI : $$= 0.548 \pm 2.12\sqrt{\frac{33.10}{2970}}$$
$$= 0.548 \pm 0.224$$

♦ The resulting 95% confidence interval is 0.324 to 0.772. Since the interval does not contain 0, we can conclude that the true value of β is not 0, and we can reject the null hypothesis $H_0 : \beta = 0$ in favor of $H_a : \beta \neq 0$, a conclusion that agrees with the findings in the previous example.

♦ Furthermore, the confidence interval estimate indicates that there is an increase of from as little as 0.324 to as much as 0.772 in percent fat for each 1-year increase in age.

● How much of the total variation in Y is explained by the linear relation with X

♦ Total sum of squares (Total SS) can be partitioned into two components, sum of squares due to regression (SSR) and sum of squares due to error (SSE), i.e.,

$$\text{Total SS} = \text{SSR} + \text{SSE}$$

♦ R^2, the coefficient of determination is the proportion of the total variation that is explained by the linear regression of y on x, where

$$R^2 = \frac{\textbf{SSR}}{\textbf{Total SS}} = \frac{\textbf{891.8736}}{\textbf{1421.5378}} = \textbf{0.63}$$ (for the previous example)

♦ Also note that the correlation coefficient between age and percent fat (r) is 0.79209. So $r^2 = (0.79209)^2 = 0.63$, the same as SSR / Total SS.

♦ Since $r^2 = 0.63$, 63% of the total variation in percent fat (*y*) is explained by the linear regression of percent fat on age (*x*).

● Estimation and prediction using the fitted line

♦ Once we conclude that the slope (β) is significantly different from zero, we can use the fitted regression line for one of two purposes:

♪ Estimating the average value of *y* for a given value of *x*

♪ Predicting a particular value of *y* for a given value of *x*

♦ Example: The estimated average percent fat when age (*x*) = 50 is

$$\begin{aligned}\hat{y} &= a + bx \\ &= 3.221 + 0.548 \times 50 \\ &= 30.6\end{aligned}$$

Some Warnings

♦ **Extrapolation**

♪ Do not use the fitted line to predict y for values of *x* that are not included within the range of the fitted data.

♦ Causality

♪ A significant regression implies that a relationship exists and that it may be possible to predict one variable with another.

♪ However, this in no way implies that one variable causes the other variable.

Simple Linear Regression

● Assumption for ε

♦ Error terms

♪ Are independent in the probabilistic sense.

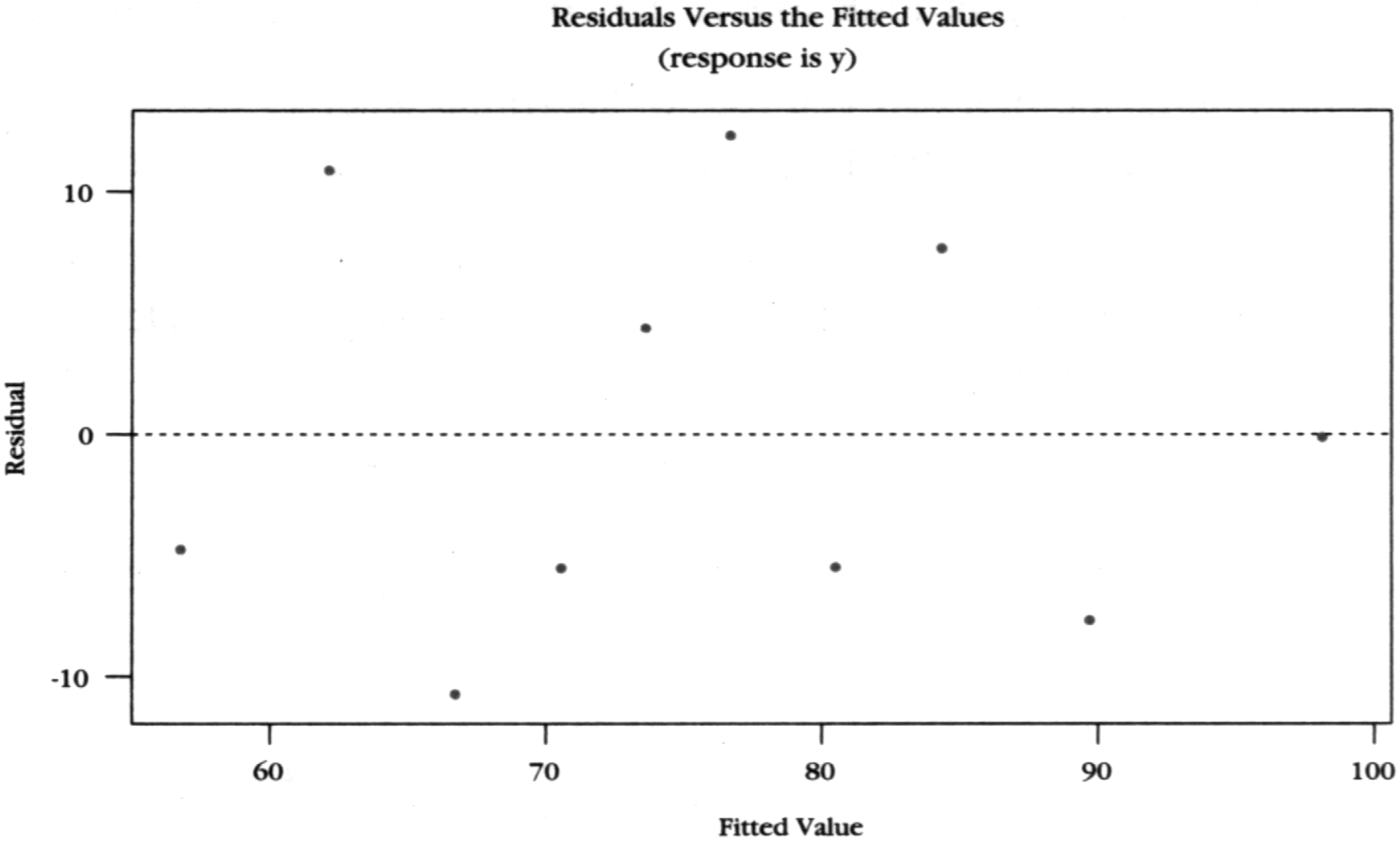
♪ Have a mean of 0 and a common variance equal to σ^2 .

♪ Have a normal probability distribution.

♦ We can use the plot of residuals versus fit to check for a constant variance as well as to make sure that the linear model is in fact adequate.

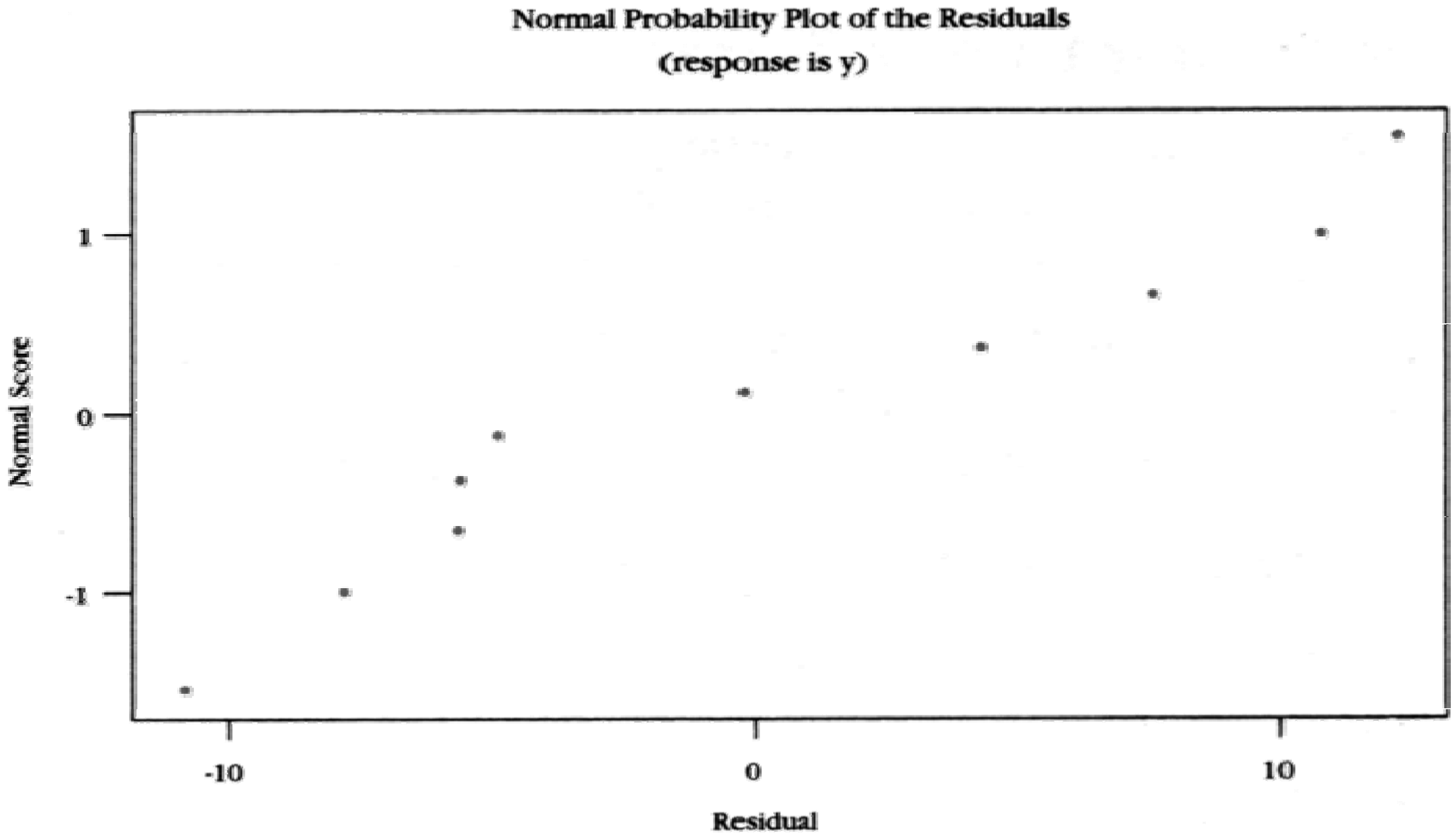
Constant Variance Assumption: Plot of Fitted Values vs. Residuals

Note : There is no tendency for the residuals to increase or decrease systematically with the fitted values, meaning constant variance



Normality Assumption

- Normal probability plot: residual vs. expected value of that residual if it had come from a normal distribution. If the graph shows almost a straight line, this means, y-values follow a normal distribution.


Normal Probability Plot of the Residuals
(response is y)
Normal Score
Residual