7. Tests of association and Linear Regression

In this chapter we consider

1. Tests of Association for 2 qualitative variables.
2. Measures of the strength of linear association between 2 quantitative variables.
3. Tests of association for 2 quantitative variables.
4. Linear Regression - Predicting one quantitative variable on the basis of another.

7.1 Chi-squared test for independence of qualitative variables

The data for these tests are contingency tables showing the relationship between 2 qualitative variables. For example, suppose we have the following information regarding hair and eye colour

	Red hair	Blonde hair	Dark hair	$\sum$
Blue eyes	30	90	30	150
Brown eyes	20	60	70	150
$\sum$	50	150	100	300

i.e. 30 people have red hair and brown eyes.

Test for independence

Let $n_{i,j}$ be the entry in the i-th row and j-th column of the contingency table. We wish to choose between the hypotheses

$$H_0 : \text{hair colour and eye colour are independent.}$$
$$H_A : \text{hair and eye colour are dependent.}$$

Row and column sums

The number of people in the sample with blue eyes is the sum of the entries in the first row (150).

The number of people in the sample with brown eyes is the sum of the entries in the second row (150).

The sum of all the entries is the number of individuals in the sample (300).

Probabilities of given events

The probability that an individual in the sample has blue eyes, P(blue), is the number of people with blue eyes divided by the number of people in the sample.

$$\text{i.e. } P(\text{blue}) = \frac{150}{300} = \frac{1}{2}$$

Similarly,

$$P(\text{brown}) = \frac{150}{300} = \frac{1}{2}$$

Probabilities of given events

The number of individuals with red hair is the sum of the number of entries in the first column (50). Arguing as above the probability the an individual in the sample has red hair is

$$P(\text{red}) = \frac{50}{300} = \frac{1}{6}$$

In a similar way,

$$P(\text{blond}) = \frac{150}{300} = \frac{1}{2}$$

$$P(\text{dark}) = \frac{100}{300} = \frac{1}{3}$$

Probabilities under the hypothesis of independence

If the traits are independent, then the probability that an individual has a given hair colour and given eye colour is the product of the two corresponding probabilities e.g.

$$\text{P(blond hair, blue eyes)} = \text{P(blond hair)P(blue eyes)}$$

In order to test whether two traits are independent, we need to calculate what we would expect to observe if the traits were independent.

The following calculations allow us to calculate what we expect to see under the null hypothesis of independence.

Probabilities under the hypothesis of independence

In general, let $n_{i,\bullet}$ be the sum of the entries in the i-th row and $n_{\bullet,j}$ be the sum of the entries in the j-th column.

Let n be the total number of observations (the sum of the entries in the cells).

Hence, here $n_{1,\bullet} = n_{2,\bullet} = 150$ (the sums of entries in first and second rows, respectively).

Also, $n_{\bullet,1} = 50, n_{\bullet,2} = 150, n_{\bullet,3} = 100$ (sum of entries in columns 1,2 and 3, respectively).

n=300 (i.e. there are 300 observations in total)

Probabilities under the hypothesis of independence

The probability of an entry being in the i-th row is $p_{i,\bullet} = \frac{n_{i,\bullet}}{n}$.

The probability of an entry being in the j-th column is $p_{\bullet,j} = \frac{n_{\bullet,j}}{n}$.

If the two traits considered are independent, then the probability of an entry being in the cell in the i-th row and j-th column is $p_{i,j}$, where

$$p_{i,j} = p_{i,\bullet} p_{\bullet,j} = \frac{n_{i,\bullet}}{n} \frac{n_{\bullet,j}}{n} = \frac{n_{i,\bullet} n_{\bullet,j}}{n^2}$$

Expected number of observations in a cell under H_0

Under the null hypothesis, we expect $e_{i,j}$ observations in cell (i, j), where

$$e_{i,j} = np_{i,j}$$

Hence, we can calculate the expected number of observations in each cell

$$e_{i,j} = np_{i,j} = \frac{n_{i,\bullet} n_{\bullet,j}}{n},$$

i.e. the expected number is the row sum times the column sum divided by the total number of observations.

Table of expected values

In the example considered the calculation is as follows:

	Red hair	Blond hair	Dark hair	$\sum$
Blue eyes	$\frac{150\times 50}{300} = 25$	$\frac{150\times 150}{300} = 75$	$\frac{150\times 100}{300} = 50$	150
Brown eyes	$\frac{150\times 50}{300} = 25$	$\frac{150\times 150}{300} = 75$	$\frac{150\times 100}{300} = 50$	150
$\sum$	50	150	100	300

Comparison of observations and expectations (expectations in brackets)

	Red hair	Blond hair	Dark hair	$\sum$
Blue eyes	30 (25)	90 (75)	30 (50)	150
Brown eyes	20 (25)	60 (75)	70 (50)	150
$\sum$	50	150	100	300

It should be noted that the sum of the expectations in a row is equal to the sum of the observations. An analogous result holds for the columns.

The test statistic

The test statistic is

$$T = \sum_{i,j} \frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}},$$

where the summation is carried out over all the cells of the contingency table.

The realisation of this statistic is labelled t. This is a measure of the distance of our observations from those we expect under H_0.

It should be noted that if the null hypothesis is true, then $n_{i,j}$ and $e_{i,j}$ are likely to be similar, hence the realisation t will tend to be close to 0 (by definition this realisation is non-negative). Large values of t indicate that the traits are dependent.

Calculation of the realisation of the test statistic

In this case,

$$t=\frac{(30-25)^2}{25}+\frac{(90-75)^2}{75}+\frac{(30-50)^2}{50}+$$
$$+\frac{(20-25)^2}{25}+\frac{(60-75)^2}{75}+\frac{(70-50)^2}{50}=24.$$

Distribution of the test statistic under H_0

Given the traits are independent, the test statistic has an approximate chi-squared distribution with $(r-1) \times (c-1)$ degrees of freedom, where r and c are the numbers of rows and columns, respectively.

It should be noted that this approximation is **reasonable if at least 5 observations are expected in each cell under H_0.**

Making conclusions

Since large values of t indicate that the traits are dependent, we reject the null hypothesis of independence at a significance level of α if

$$t > \chi^2_{(r-1)(c-1),\alpha},$$

where $\chi^2_{(r-1)(c-1),\alpha}$ is the critical value for the appropriate chi-squared distribution. These values can be read from Table 8 in the Murdoch and Barnes book.

Making conclusions

In this case at a significance level of 0.1%

$$\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{2,0.001} = 13.815.$$

Since $t = 24 > \chi^2_{2,0.001} = 13.815$, we reject the null hypothesis of independence.

Since we rejected at a significance level of 0.1%, we have very strong evidence that eye and hair colour are dependent.

Describing the nature of an association

In order to see the nature of the association between hair and eye colour, we should compare the observed and expected values.

Comparing the table of observed and expected values, it can be seen that dark hair is associated with brown eyes and blond hair with blue eyes (in both cases there are more observations than expected under the null hypothesis).

Simpsons Paradox

The nature of an association between two traits may change and even reverse direction when the data from several groups is combined into a single group.

For example, consider the following data regarding admissions to a university.

The Physics department accepted 60 applications from 100 males and 20 applications from 30 females.

The Fine Arts department accepted 10 applications from 100 males and 30 applications from 170 females.

Simpson's paradox

Hence, the Physics department accepted 60% of applications from males and 66.7% of applications from females. The Fine Arts department accepted 10% of applications from males and 17.6% of applications from females.

Hence, in both departments females are slightly more successful.

Simpson's paradox

Combining the departments, we obtain the following contingency table

	Accepted	Rejected	$\sum$
Male	70	130	200
Female	50	150	200
$\sum$	120	280	400

Combining the results females are less successful.

We now test the null hypothesis that acceptance is independent of sex. The alternative is that acceptance is associated with sex (i.e. the likelihood of acceptance depends on sex).

Simpson's paradox

The table of expectations is as follows

	Accepted	Rejected	$\sum$
Male	$\frac{200 \times 120}{400} = 60$	200-60=140	200
Female	120-60=60	280-140=140	200
$\sum$	120	280	400

Comparing the table of observations and expectations

	Accepted	Rejected	$\sum$
Male	70 (60)	130 (140)	200
Female	50 (60)	150 (140)	200
$\sum$	120	280	400

Simpson's paradox

The test statistic is

$$T = \sum_{i,j} \frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}}$$

The realisation is given by

$$t = \frac{(70-60)^2}{60} + \frac{(130-140)^2}{140} + \frac{(50-60)^2}{60} + \frac{(150-140)^2}{140} \approx 4.76$$

Simpson's paradox

Testing the null hypothesis at the 5% level, the critical value is

$$\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{1,0.05} = 3.841.$$

Since $t > \chi^2_{1,0.05} = 3.841$, we reject the null hypothesis.

We conclude that acceptance is associated with sex. Looking at the table which compares the observed and expected values, females are less likely to be accepted than males.

Hence, using this data we might conclude that there is evidence that females are discriminated against.

Simpson's paradox

However, in both departments a female is more likely to be accepted.

On average, females are less likely to be accepted since they are more likely to apply to the fine arts department and the fine arts department rejects a higher proportion of applications.

Lurking variables

In this case, when we pool the data, department is what we call a hidden or lurking variable.

A hidden or lurking variable is a variable which may influence the observed results, but is not considered in the analysis.

For example, women's salaries are significantly lower on average than men's salaries.

1. Can this be used as evidence that women are discriminated against?
2. What are the lurking variables which may explain such a difference?
3. How should we test whether females are discriminated against with regard to salary?

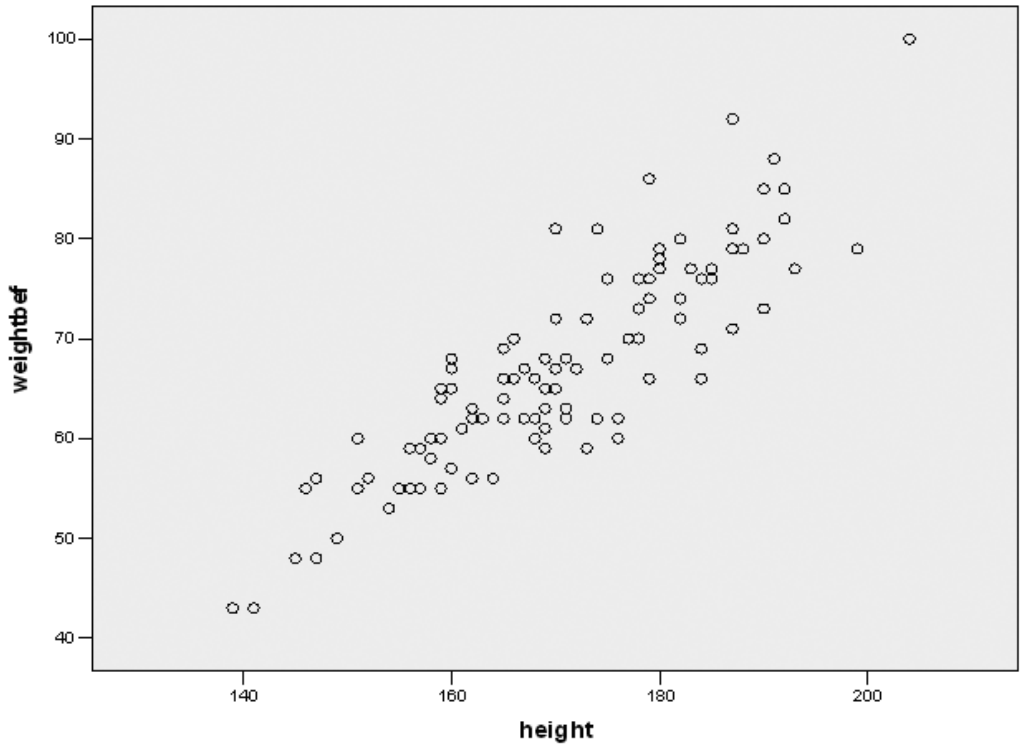
7.2 Regression and Prediction

Suppose we have n pairs of observations of quantitative variables $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

e.g. X_i and Y_i may represent the height and weight of the i-th person in a sample, respectively.

Scatter plots

The relationship between these two variables may be illustrated graphically using a scatter plot



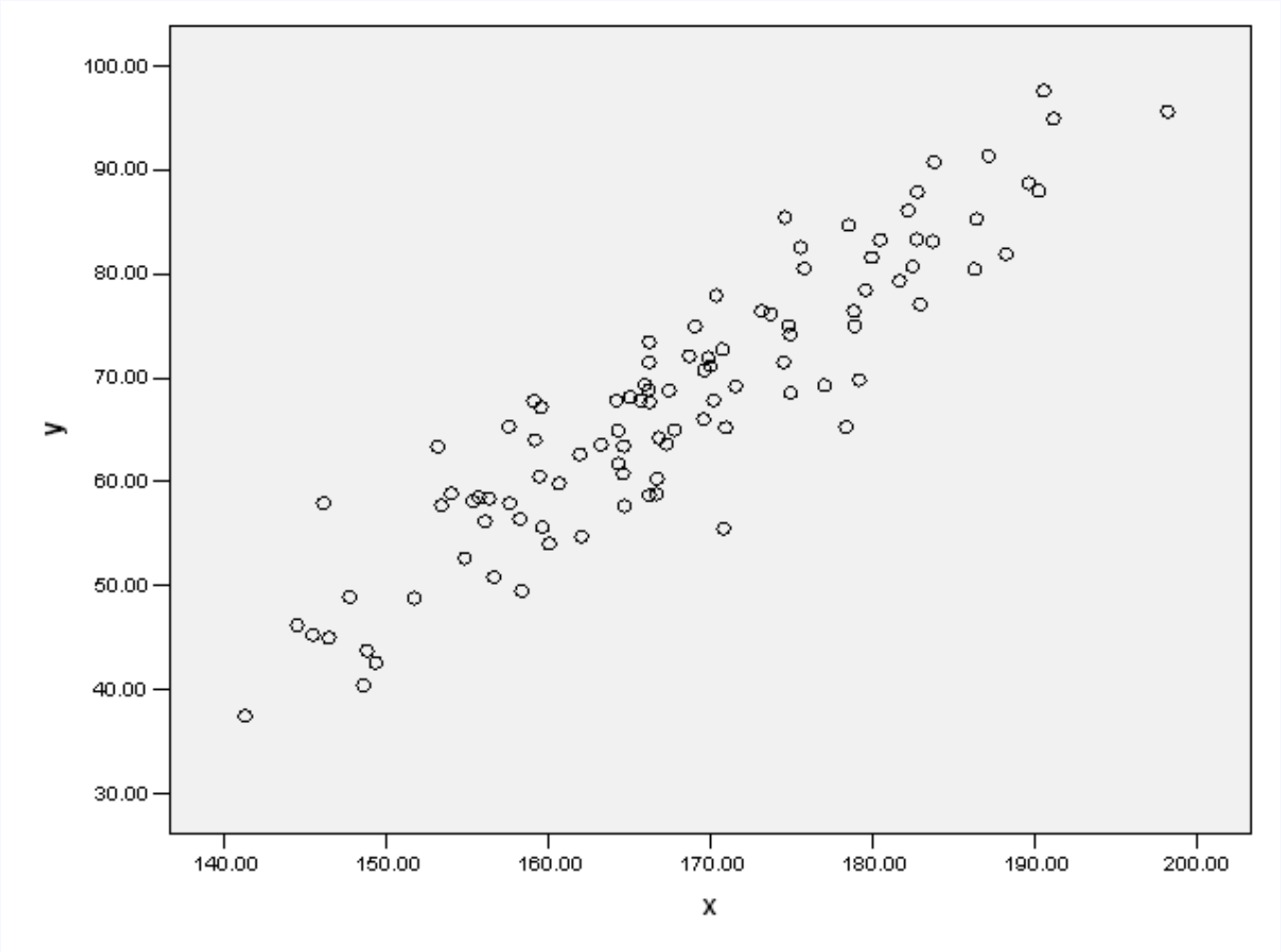
Types of relationship

We distinguish between 3 types of 'relationship'.

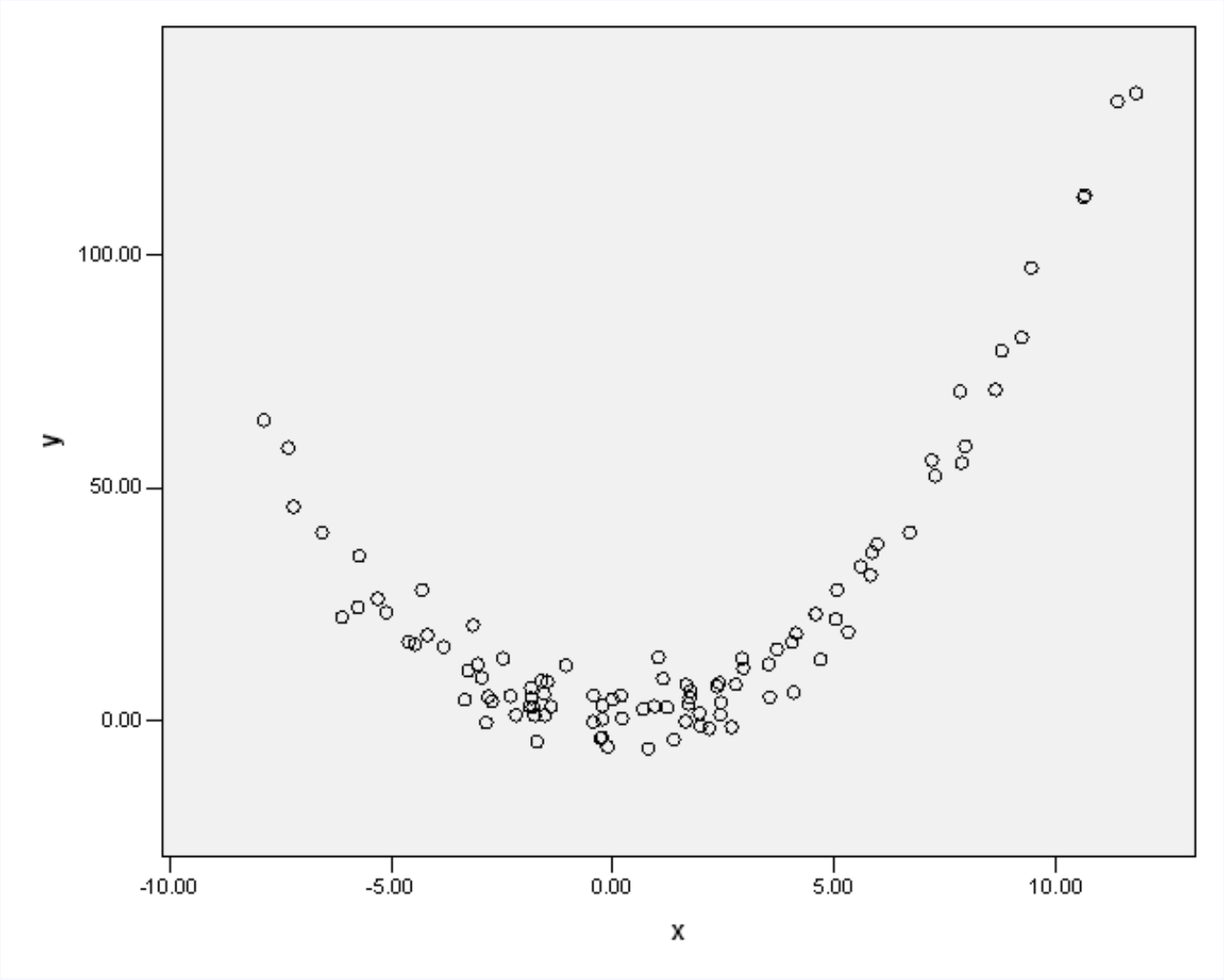
1. Linear relationship.
2. Non-linear (curvilinear) relationship. In this case, exponential growth and seasonal variation will be important types of relationship.
3. No relationship (independence)

These relationships are illustrated on the following slides.

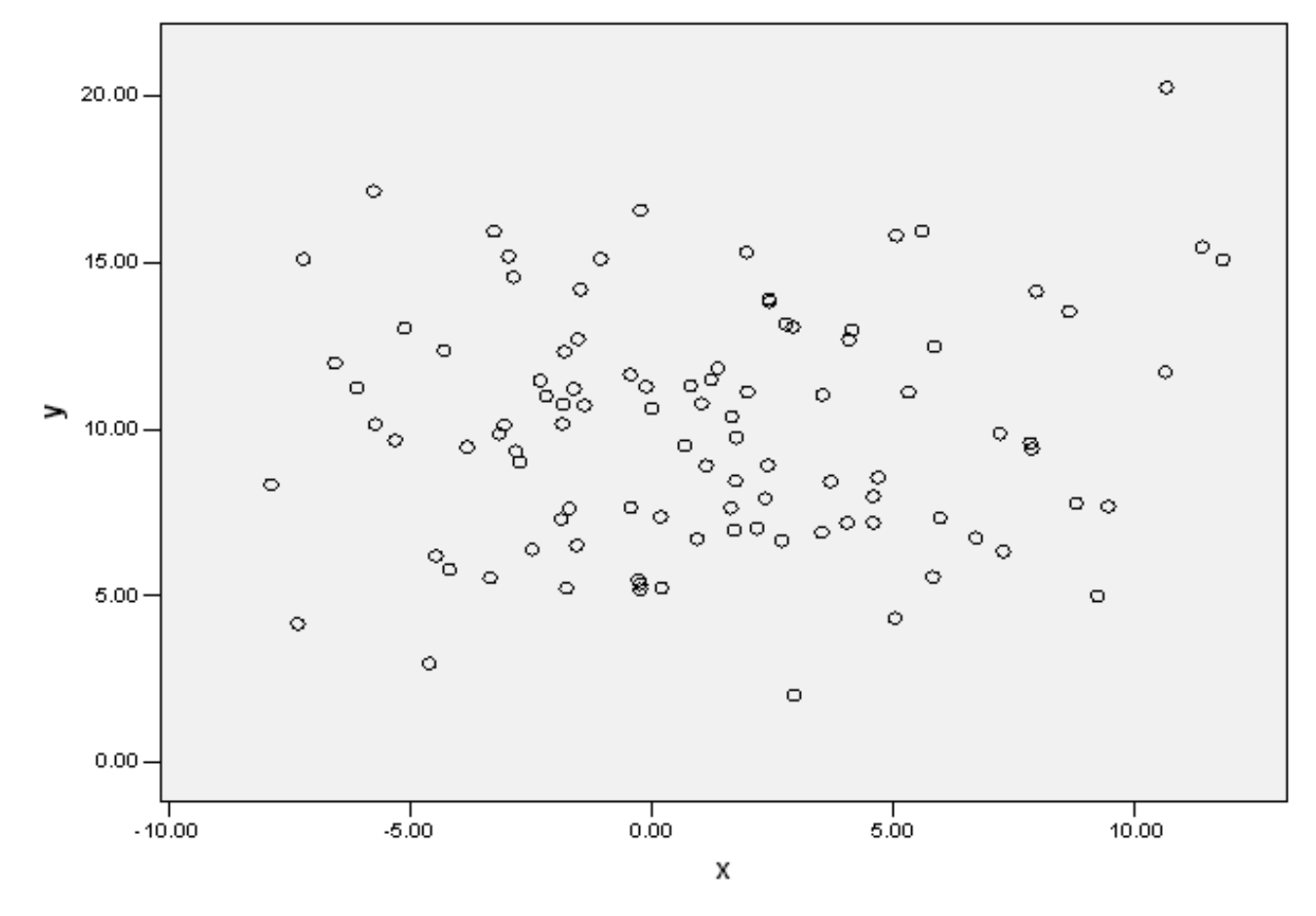
A linear relationship



A non-linear relationship



Independent traits



Does association imply cause and effect?

It should be noted that a relationship (association) between two variables does not necessarily mean that there is a cause-effect mechanism at work between the variables.

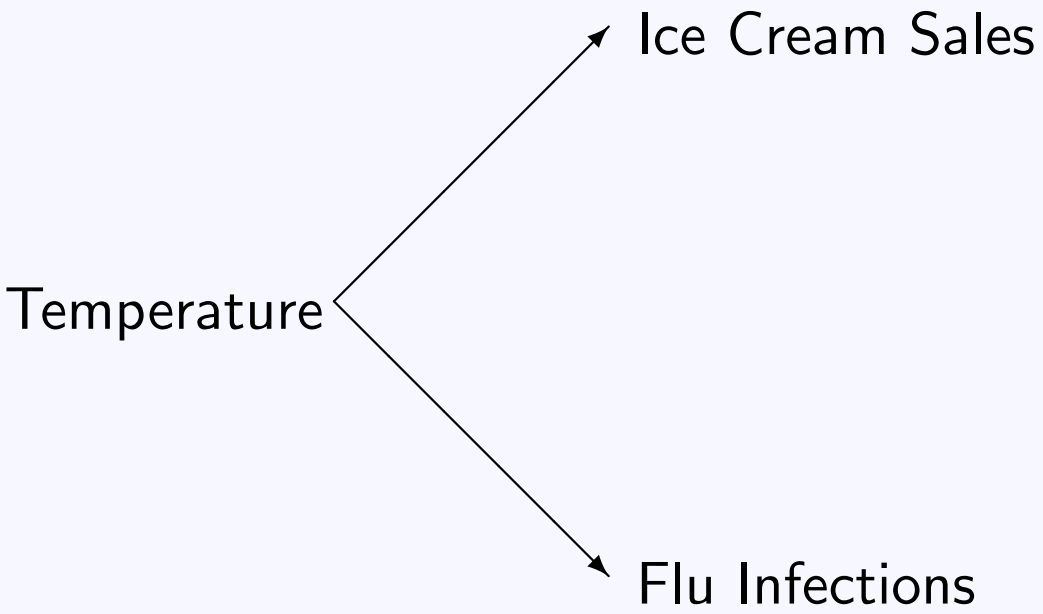
Also, correlation or regression analysis can only detect associations. They cannot determine what is the cause and what is the effect.

For example, there is a negative correlation between flu infections and the sales of ice cream.

Of course this does not mean that ice cream gives protection against being infected by flu. This correlation arises because as the temperature increases, ice cream sales increase and the number of flu infections decreases.

Cause-effect diagrams

The relationship considered may be illustrated by the following cause-effect diagram



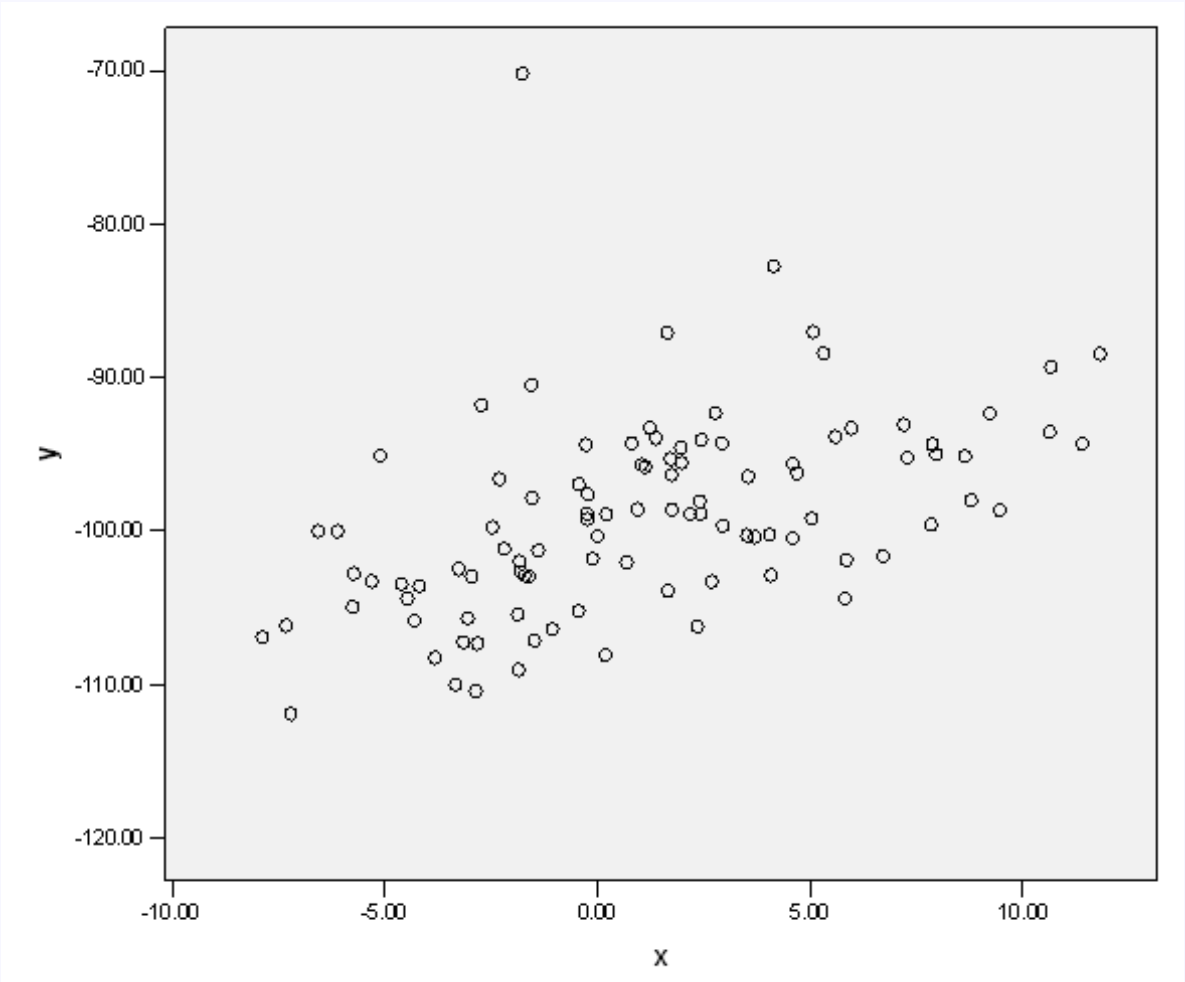
i.e. the weather affects both ice cream sales and the number of flu infections.

Criteria for establishing causation

1. An association should be found in numerous studies which consider various factors (this eliminates the possibility that a lurking variable is the cause).
2. The cause must precede the effect.
3. Established physical laws: i.e. A force causes an acceleration. If I heat a gas, it will cause an increase in pressure (this is related to 1).

Outliers

A scatter plot may indicate that certain observations are outliers. Such observations may result from a mistake in recording the observations (see diagram).



Outliers

Analysis should be carried out both with and without outliers (this indicates how robust the analysis is to possible mistakes).

7.2.1 Correlation coefficients

A coefficient of the correlation between two variables X and Y, denoted $r(X, Y)$, is a measure of the strength of the relationship between X and Y.

Both X and Y should be quantitative variables.

Properties of a correlation coefficient

1. $-1 \leq r(X, Y) \leq 1$.
2. If $r(X, Y) = 0$, then there is no general tendency for Y to increase as X increases (this does not mean that there is no relationship between X and Y).
3. A correlation coefficient is independent of the units of measurement.

Properties of the correlation coefficient

4. If $r(X, Y) > 0$, then there is a tendency for Y to increase when X increases.
5. If $r(X, Y) < 0$, then there is a tendency for Y to decrease as X increases.
6. Values of $r(X, Y)$ close to 1 or -1 indicate a strong relationship between X and Y (note that values close to 0 do not necessarily indicate a weak relationship between X and Y).

Estimation of the population correlation coefficient

In practice, we estimate the population correlation coefficient based on a sample using either Pearson's or Spearman's correlation coefficient.

This correlation coefficient will have some distribution centred around the appropriate population (i.e. true) correlation coefficient.

Use of Pearson's correlation coefficient

We should use Pearson's correlation coefficient when

1. The distributions of both variables are close to normal (we can check this by looking at the histograms for both of the variables).
2. The relationship between the variables can be assumed to be linear (we can check this by looking at a scatter plot for the pair of variables).

Otherwise, we should use Spearman's correlation coefficient.

$r_P(X, Y)$ and $r_S(X, Y)$ will be used to denote Pearson's and Spearman's correlation coefficient between X and Y, respectively.

An additional property of Pearson's correlation coefficient

In addition to the 6 general properties listed above, Pearson's correlation coefficient has the following property

If $|r_P(X, Y)| = 1$, then the relationship between X and Y is an exact linear relationship, i.e. $Y = a + bX$.

In this case, if $r_P(X, Y) = 1$, then $b > 0$ (i.e. a positive relationship) and if $r_P(X, Y) = -1$, then $b < 0$ (i.e. a negative relationship).

Pearson's correlation coefficient should not be used to measure the strength of an obviously non-linear relationship.

Properties of Spearman's correlation coefficient

In addition to the 6 general properties listed above, Spearman's correlation coefficient has the following properties.

1. Let R_i and S_i be the ranks of the observations X_i and Y_i, respectively (i.e. $R_i = k$ if X_i is the k-th smallest of the X observations). Spearman's correlation coefficient is defined to be Pearson's correlation coefficient between the ranks R and S.
2. If $r_S(X, Y) = 1$, then there is a monotonically increasing relationship between X and Y and if $r_S(X, Y) = -1$, then there is a monotonically decreasing relationship between X and Y.

The calculation of these correlation coefficients is numerically intensive and hence will only be calculated using the SPSS package.

A test of association based on the correlation coefficient

We can test the null hypothesis

$$H_0 : r(X, Y) = 0$$

against

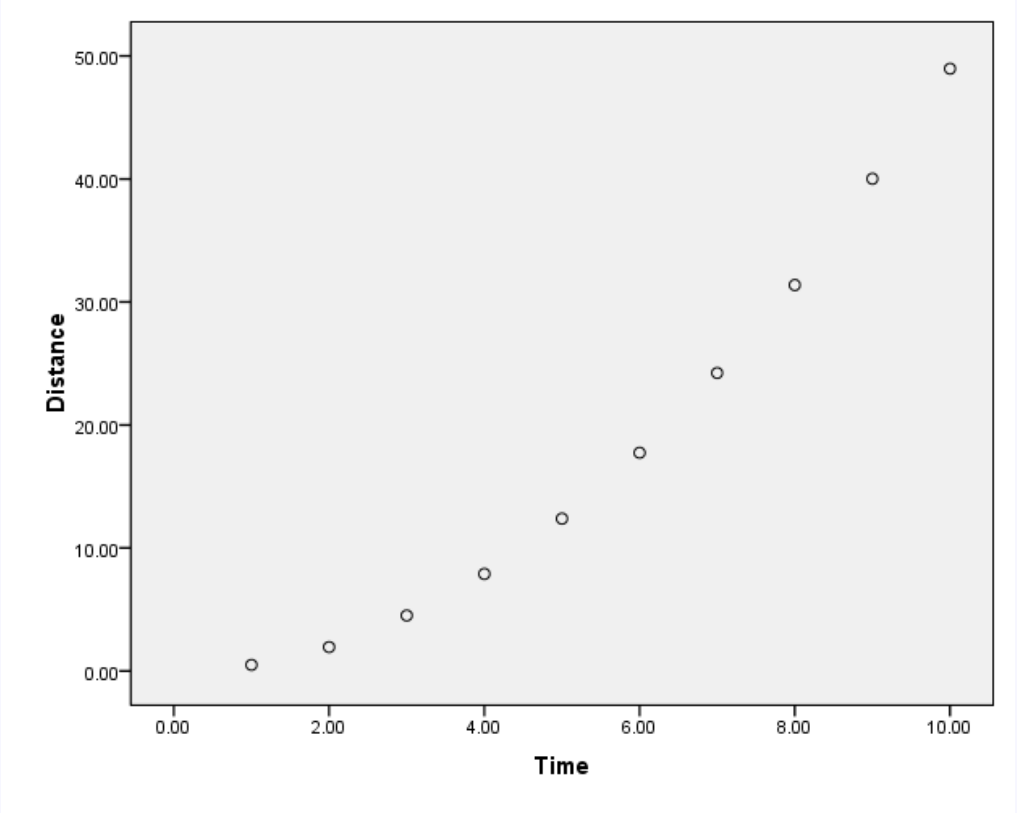
$$H_A : r(X, Y) \neq 0.$$

H_0 is rejected if the correlation coefficient used is significantly different from 0.

We only carry out these tests with the aid of SPSS. The test carried out corresponds to the correlation coefficient used.

Example 7.1

The following graph illustrates the distance travelled by a ball rolling down a hill in metres according to time measured in seconds.



Example 7.1

The relationship is clearly non-linear and hence the appropriate measure of correlation is Spearman's.

From the Correlate menu in SPSS, we highlight the option to calculate Spearman's correlation coefficient (Pearson's correlation coefficient is calculated by default).

The results are illustrated on the next slide

Example 7.1

Correlations

		Time	Distance
Time	Pearson Correlation	1.000	.976**
	Sig. (2-tailed)		.000
	N	10.000	10
Distance	Pearson Correlation	.976**	1.000
	Sig. (2-tailed)	.000	
	N	10	10.000

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

			Time	Distance
Spearman's rho	Time	Correlation Coefficient	1.000	1.000**
		Sig. (2-tailed)	.	.
		N	10	10
	Distance	Correlation Coefficient	1.000**	1.000
		Sig. (2-tailed)	.	.
		N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

Example 7.1

Note that any correlation coefficient between X and itself is by definition equal to one.

Spearman's correlation coefficient between time and distance is equal to 1. This is due to the fact that as time increases, the distance always increases, i.e. the relationship is monotonic and positive.

Note that Pearson's correlation coefficient is close to one (0.976). Any positive, monotonic relationship will give a large positive value for Pearson's correlation coefficient.

Hence, a Pearson's correlation coefficient close to, but not equal to, 1 or (-1) is not evidence of a linear relationship.

7.3 Linear Regression

Linear regression is used to predict the value of a **response variable** Y given the value of a predictor variable X.

It should be noted that the association between X and Y does not need to be a cause-effect relationship, but if there is such a relationship X should be defined to be the cause.

Linear regression

Linear regression assumes that the 'real' relation between the response variable y and the predictor variable x is of the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$\beta_0 + \beta_1 x_i$ is called the linear component of the model and ϵ_i is the random component.

This will be called the regression model. It is assumed that the random components are normally distributed with mean 0 (this distribution does not depend on x).

Linear regression

The real (population) regression line is given by

$$Y = \beta_0 + \beta_1 X.$$

This gives the average value of the response variable given the value of the predictor variable i.e. when $X = k$ the average value of the response variable is $\beta_0 + \beta_1 k$.

The response variable will in general show variation around this mean value.

Interpretation of the parameter β_1

β_1 is the slope of the real regression line. A positive slope indicates that the response variable increases as the predictor variable increases e.g. if $\beta_1 = 2$ a unit increase in the predictor variable, causes on average an increase of 2 units in the response variable (note that β_0 and β_1 are sensitive to the units used).

A negative slope indicates that the response variable decreases as the predictor variable increases e.g. if $\beta_1 = -5$ a unit increase in the predictor variable, causes on average an decrease of 5 units in the response variable.

Interpretation of the parameter β_0

β_0 is called the intercept [the graph of the real regression line crosses the y-axis at $(0, \beta_0)$].

If the predictor variable x can take the value 0, then when $x = 0$, on average $y = \beta_0$.

Estimation of the regression line

In practice, we estimate the parameters β_0 and β_1 by $\hat{\beta}_0$ and $\hat{\beta}_1$. The estimate of the regression line is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

Given the predictor variable takes value x_i, if we did not know the value of the response variable, we would estimate it to be

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Estimation of the regression line

It should be noted that this estimate is most accurate for observations of x which are close to the median observation of x.

Such estimation should not be used for a value of x outside the range observed.

Errors in estimation

The difference between an observation of the response variable and its estimate using the regression line is called the residual and is denoted r_i i.e.

$$r_i = y_i - \hat{y}_i,$$

y_i is the i-th observation of the response variable and $\hat{y}_i$ is its estimate based on the regression line.

If $r_i > 0$, then use of the regression line results in underestimation of y_i.

If $r_i < 0$, then use of the regression line results in overestimation of y_i.

7.3.1 Estimation of the regression parameters

The parameters β_0 and β_1 are estimated using $\hat{\beta}_0$ and $\hat{\beta}_1$, such that

1. $\sum_{i=1}^{n} r_i^2$ is minimised (i.e. the mean square error from estimating the observations of the response variable is minimised).
2. $\sum_{i=1}^{n} r_i = 0$ (i.e. given the model is correct, there is no tendency to under- or overestimate the observations of the response variable).

Test of association based on regression analysis

Given our estimate of the slope of the regression line we can carry out a test of the hypothesis

$$H_0 : \beta_1 = 0$$

against

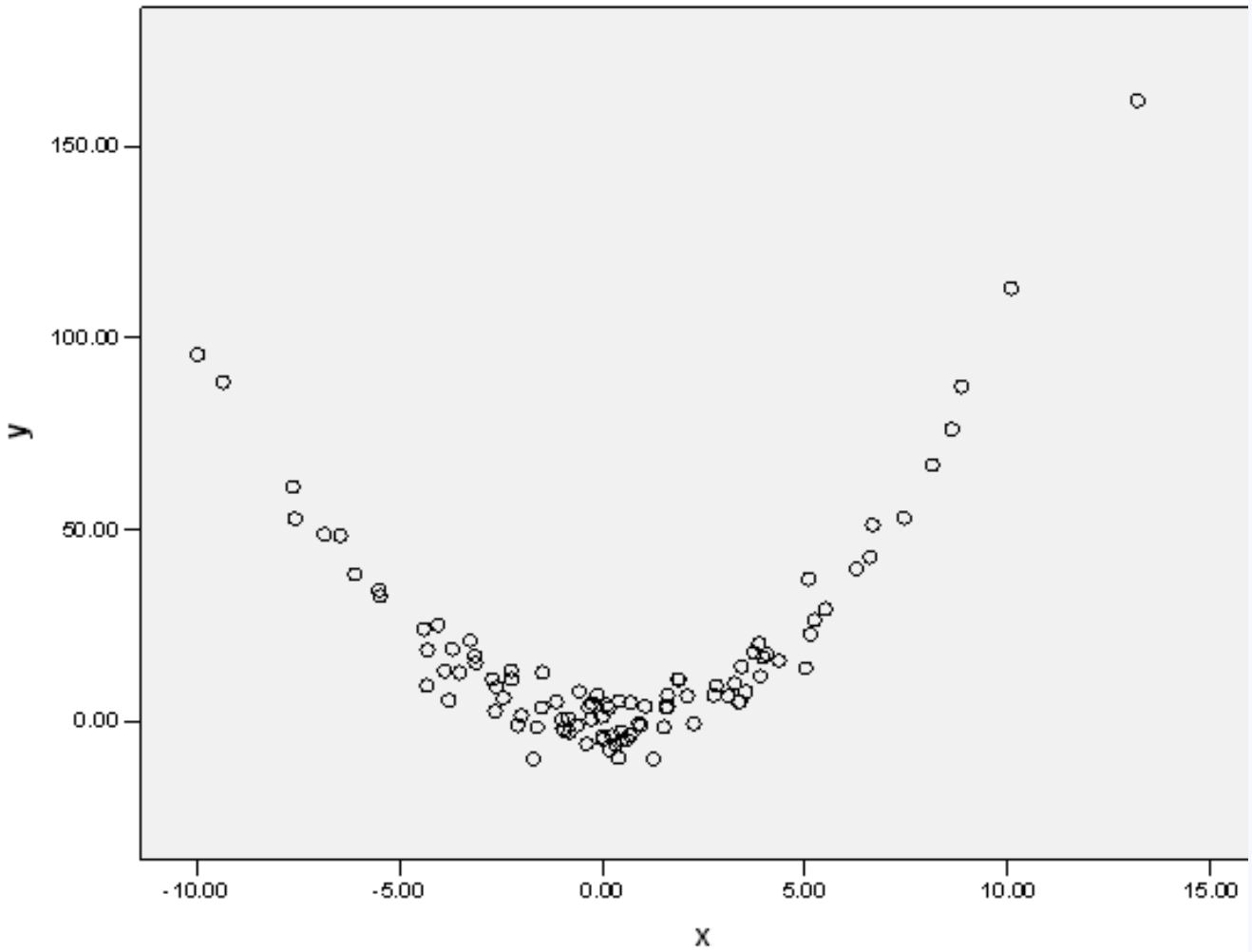
$$H_A : \beta_1 \neq 0.$$

In words, H_0 may be interpreted as the hypothesis that there is no linear relation between X and Y (as X varies Y does not on average vary).

H_A may be interpreted as the hypothesis that there is a relation between X and Y such that as X rises the mean value of Y either systematically rises or systematically falls.

H_0 is not equivalent to the hypothesis that there is no relationship between X and Y (see diagram on the next slide).

A non-linear relationship with non-significant regression slope



Interpretation of the results of a test of association

As in the interpretation of a correlation coefficient, a significant relationship between two variables does not necessarily imply a cause-effect relationship.

A lurking variable may be the cause of such a relationship.

For example, weekly sales of ice cream will be negatively correlated with the number of flu infections, since hot weather leads to increased ice cream sales, but fewer flu infections.

We only consider such tests in SPSS.

Example 7.2

We consider the data on height X and weight before studies Y given in lab1.sav.

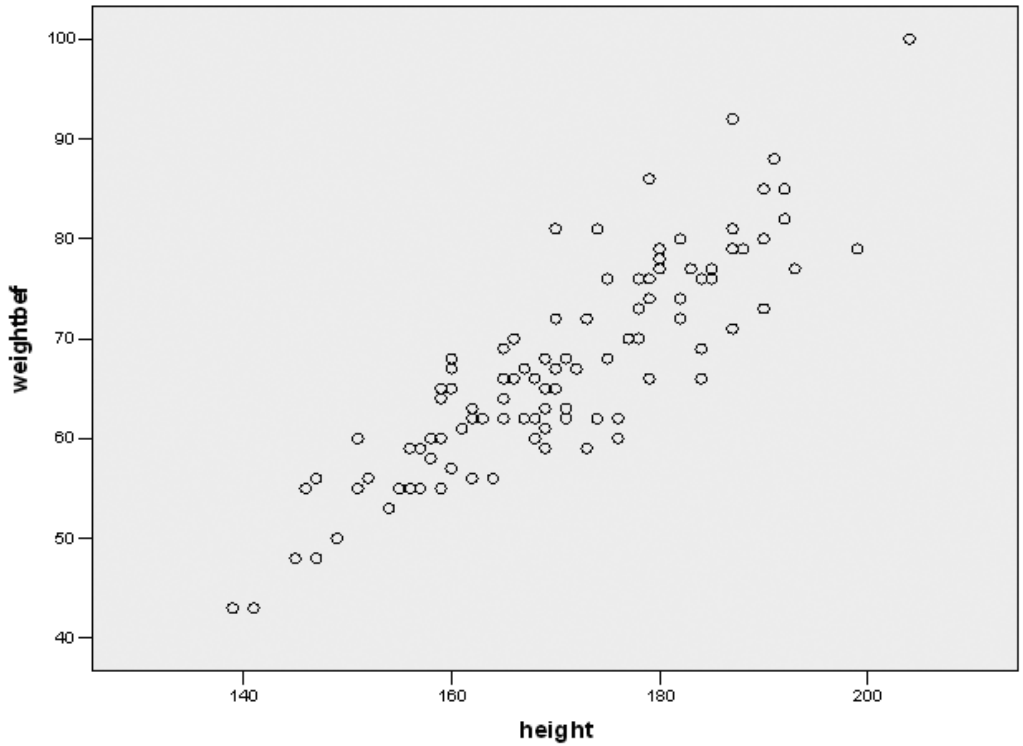
A scatter plot indicates that there is a clear positive relationship between height and weight. (as height increases, on average weight increases).

From the scatter plot it seems that the assumption of a linear relationship between X and Y is reasonable.

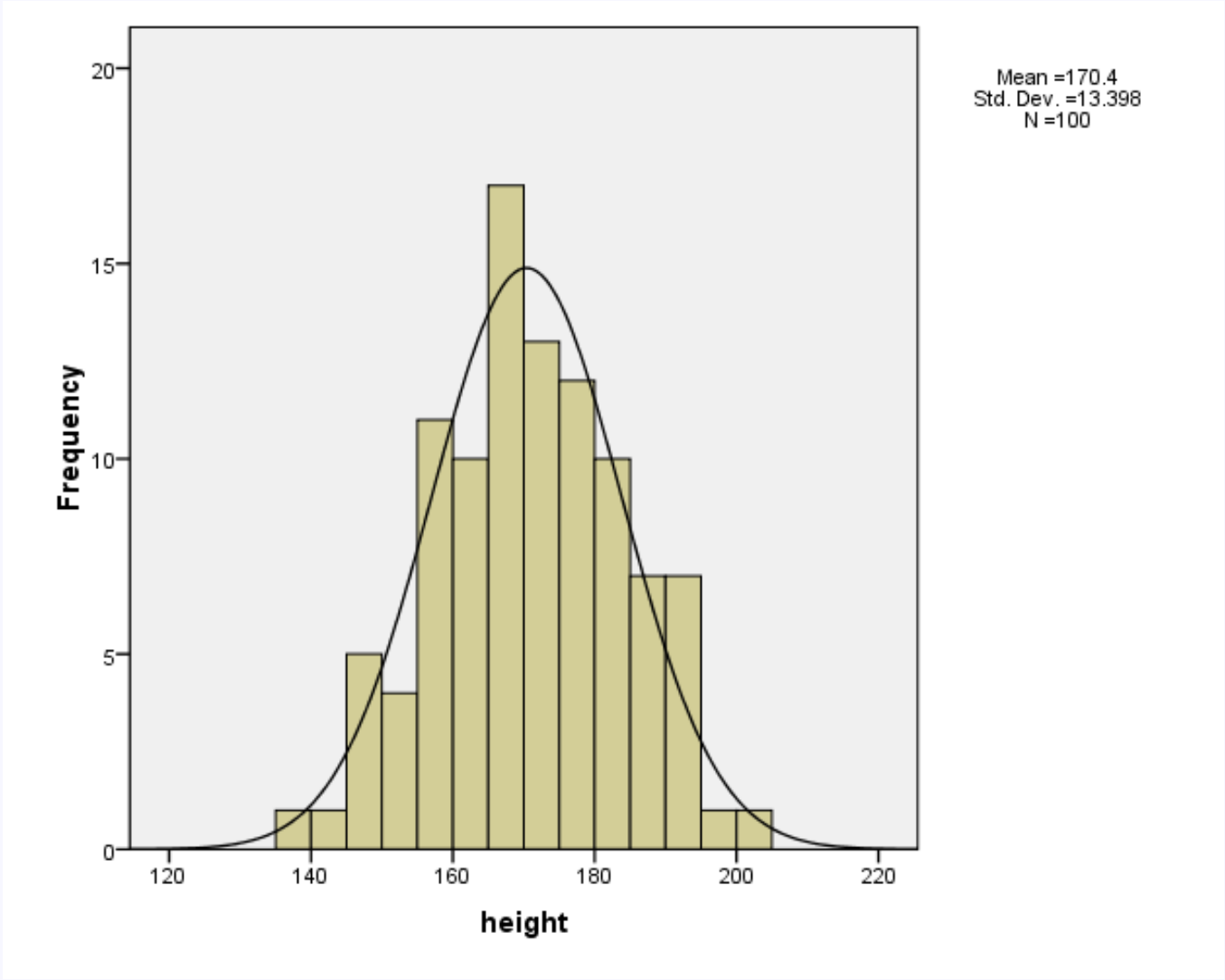
The histograms for height and weight show that the distribution of both of these variables is similar to the normal distribution.

It follows that Pearson's correlation coefficient will be a good measure of the strength of the relationship between height and weight.

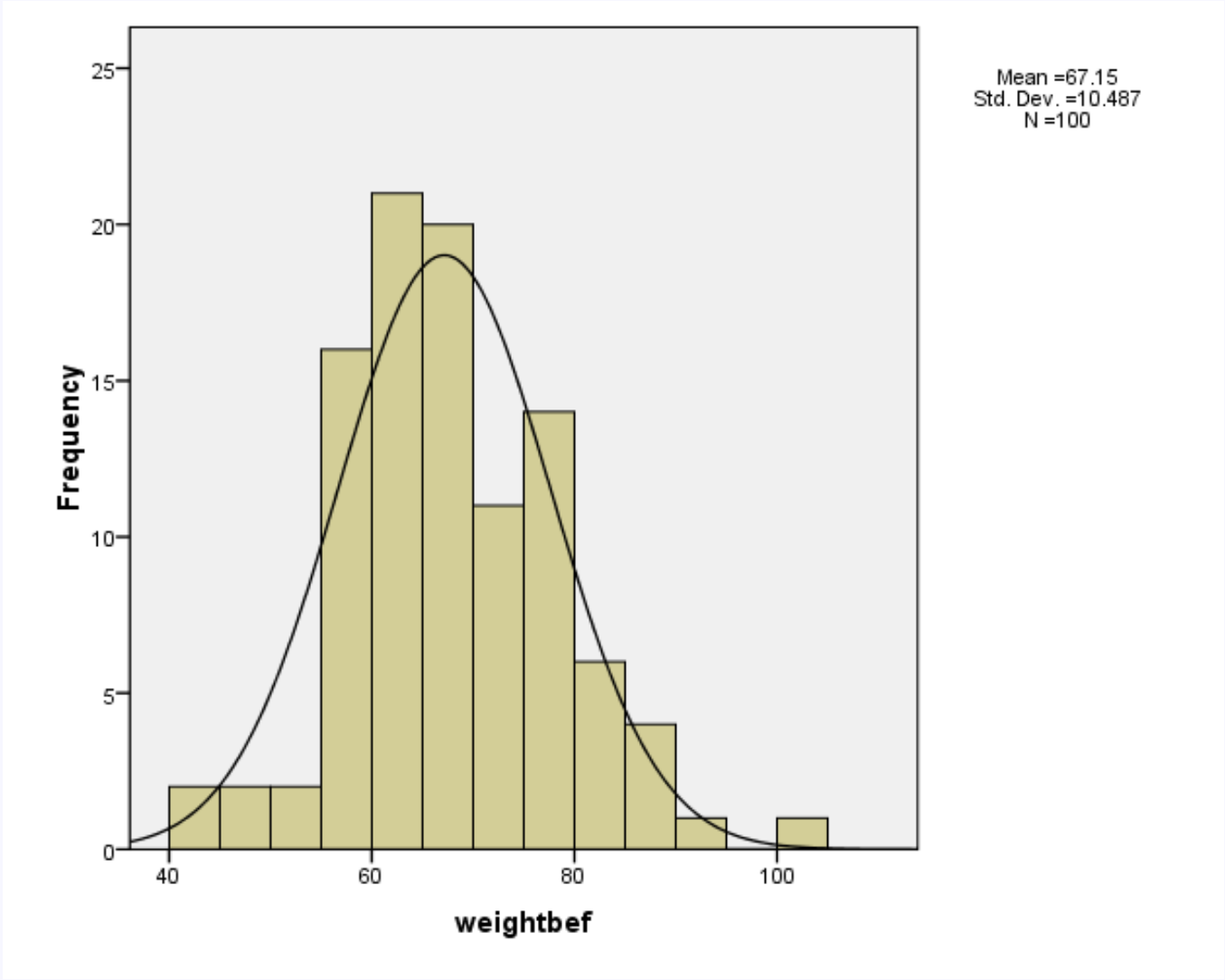
Scatter plot of the relationship between height and weight



Histogram of height



Histogram of weight



Example 7.2

There do not seem to be any significant outliers. The observation in the top right corner would be reasonably close to the straight line which best fits the data.

An observation in the top left or bottom right corner would be an outlier.

The Pearson correlation coefficient is 0.871.

Since this coefficient is large and positive, it indicates a strong positive relationship between the two variables (i.e. as height increases, weight on average increases).

Interpretation of the results from SPSS

SPSS gives the following regression coefficients together with associated statistics for the relationship of height (in cms) with weight (in kgs)

Model	Unstandardised Coefficients			
	B	Std. Error	*t*	sig.
(Constant)	-48.97	6.650	-7.365	.000
height	.681	.039	17.516	.000

The estimated regression model is
Weight = -48.97+ 0.681 Height.

The coefficients are read from the *B* column. $\hat{\beta}_0 = -48.97$, $\hat{\beta}_1 = 0.681$.

Interpretation of the results from SPSS

Weight = -48.97+ 0.681 Height.

i.e. an increase in height of 1cm leads to an average increase in weight of 0.681kg.

The sig. value in the bottom right hand corner (0.000) gives the p-value for the test

$$H_0 : \beta_1 = 0$$

against

$$H_A : \beta_1 \neq 0.$$

Interpretation of the results from SPSS

H_0 states that there is no linear relationship between height and weight.

Since $p = 0.000 < 0.001$, we reject H_0 at a significance level of 0.1% . We have very strong evidence that there is a relationship between height and weight.

Since $\hat{\beta}_1 > 0$ this indicates a positive relationship between height and weight (i.e. as height increases weight increases).

Interpretation of the results from SPSS

We can use the estimated regression line to estimate the average weight for a given height.

e.g. The average weight of people of height 170cm can be approximated by substituting $X = 170$ into the regression equation

$$\begin{aligned}\hat{Y} &= -48.975 + 0.681X \\ &= -48.975 + 0.681 \times 170 = 66.8kg\end{aligned}$$

This is a clearly reasonable approximation. Note 170 is very close to the average height.

Interpretation of the results from SPSS

Using this regression line to estimate the average weight of 50cm high individuals

$$\begin{aligned}\hat{Y} &= -48.975 + 0.681X \\ &= -48.975 + 0.681 \times 50 = -14.9kg\end{aligned}$$

This is clearly a nonsensical estimate.

This is due to the fact that 50cm is well outside the range of heights we observe. In such cases approximation using the regression line will be inappropriate.

Note that although the relationship may be reasonably linear over the range observed, it may not be linear over a larger range. If the relationship is clearly non-linear, such approximation will always be inappropriate.

7.4 Transforming data in order to obtain a linear relationship - Exponential growth

Assume that there is a non-linear relationship between X and Y.

It may be possible to find a function $U = g(Y)$ such that there is a linear relationship between U and X.

In this case, we can carry out linear regression for U on X, obtaining a model $U = g(Y) = a + bX$.

Transforming this equation, we can find the regression equation for Y as some non-linear function of X.

Exponential Growth

Exponential growth is a very important relationship in the financial sector, e.g. the value of investments and prices.

We can model the value of an investment, Y, as an exponential function of time, X.

Such a model states that if the value of an investment at time 0 is β_0, then its expected value at time X is of the form $Y = \beta_0 e^{\beta_1 X}$, where β_1 is the expected interest rate on the investment.

Taking logarithms, we obtain $\ln Y = \ln(\beta_0) + \beta_1 X$. Hence, there is a linear relationship between X and $U = \ln Y$.

Example 7.3 - Exponential Regression

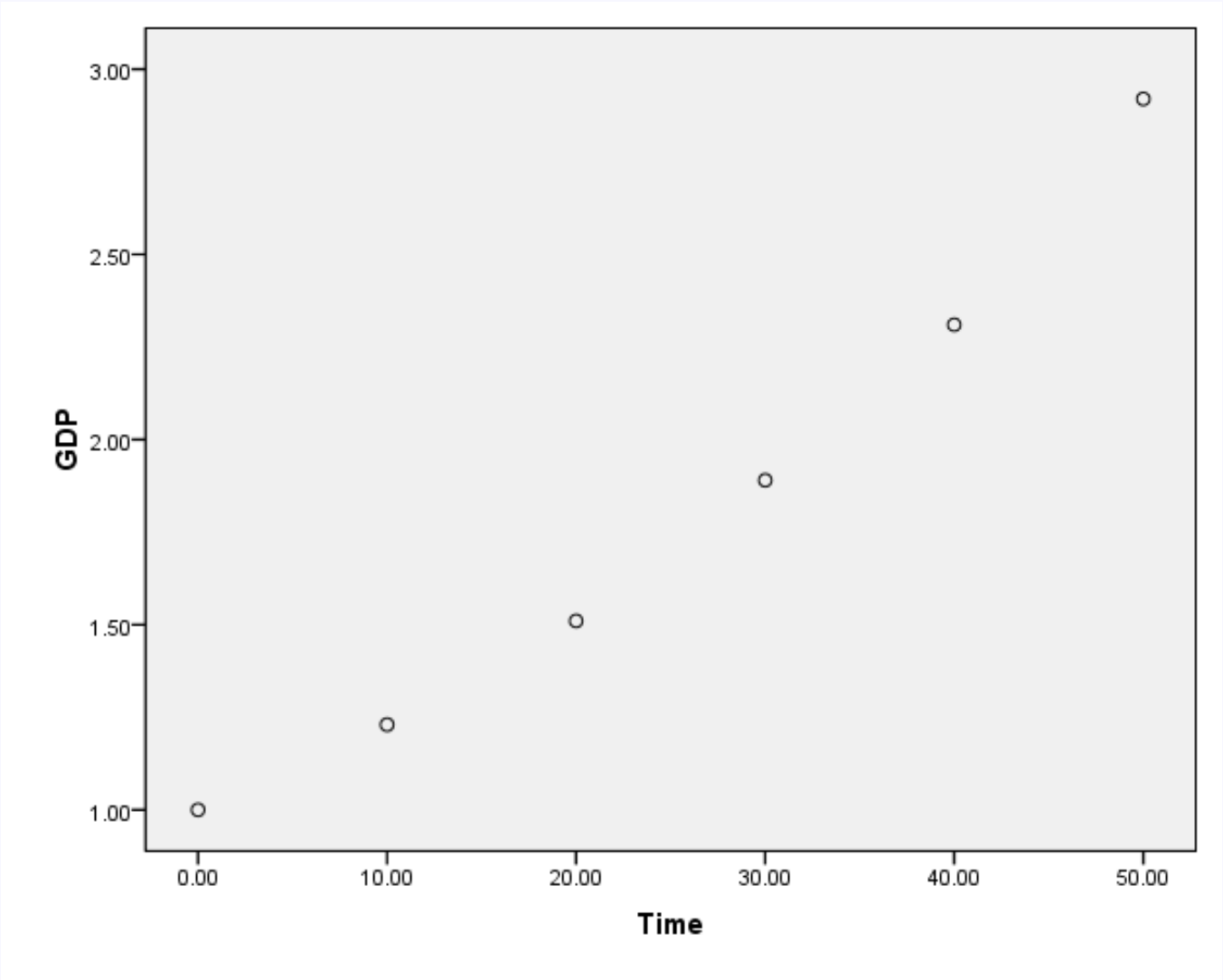
The nominal annual GDP of Ireland relative to GDP in 1950 (year 0) is given below.

Derive an appropriate regression equation for GDP as a function of time in years.

Year	0	10	20	30	40	50
GDP	1	1.23	1.51	1.89	2.31	2.92

The graph on the next slide shows a scatter plot of GDP (y-axis) as a function of time (x-axis).

Example 7.3 - Exponential Regression - Scatter Plot of Data



Example 7.3 - Exponential Regression

It can be seen that the relationship between GDP and time is not linear (it is upwards curving).

This suggests that the relationship between time and GDP is exponential (often nominal monetary values show exponential growth).

For this reason, we take the logarithms of the dependent variable (GDP).

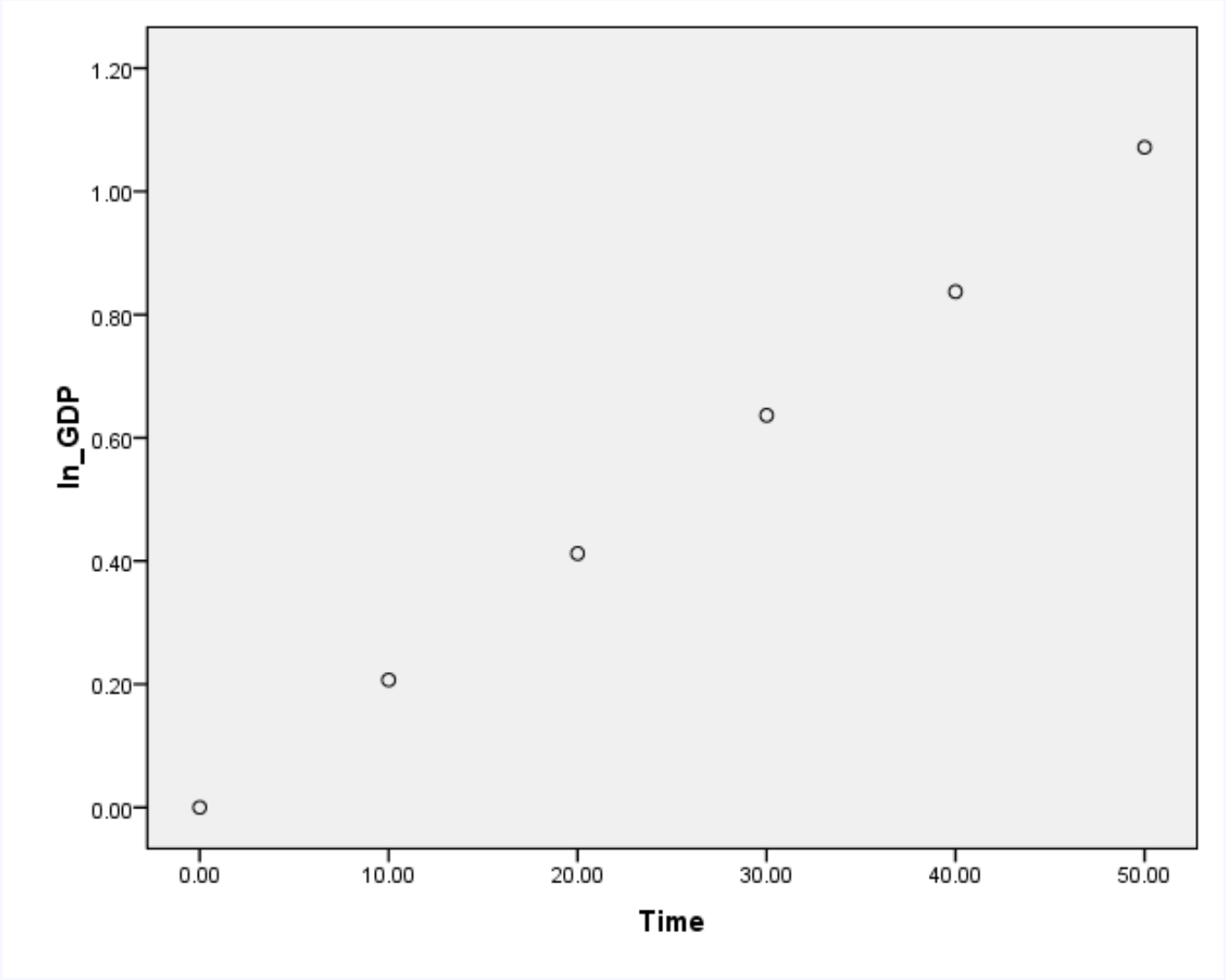
This gives us the table on the next slide.

Example 7.3 - Exponential Regression

Year	0	10	20	30	40	50
GDP	1	1.23	1.51	1.89	2.31	2.92
ln(GDP)	0	0.21	0.41	0.64	0.84	1.07

The graph on the next slide shows a scatter plot of ln(GDP) (y-axis) as a function of time (x-axis).

Example 7.3 - Exponential Regression



Example 7.3 - Exponential Regression

This scatterplot shows that it is reasonable to assume that the relationship between time and ln(GDP) is linear.

Hence, we now carry out linear regression to find an expression for ln(GDP) as a function of time.

The results of the analysis from SPSS are given on the next slide.

Example 7.3 - Exponential Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.006	.007		-.960	.391
	Time	.021	.000	1.000	97.423	.000

a. Dependent Variable: ln_GDP

Example 7.3 - Exponential Regression

The model for ln(GDP) as a function of time is

$$\ln(GDP) = -0.006 + 0.021 time.$$

Since we want a model for GDP, we invert the operation of taking logs by exponentiating. It follows that

$$\begin{aligned} GDP &= \exp(-0.006 + 0.021 time) \\ &= \exp(-0.006)\exp(0.021 time) \\ &= 0.994 \exp(0.021 time) \end{aligned}$$

Example 7.3 - Exponential Regression

Using this equation we can estimate relative GDP in 2010 (year 60).

$$GDP = 0.994 \exp(0.021 \times 60) = 3.504.$$

Each year GDP increases by a factor of $\exp(\beta_1)$.

Thus, our estimate of the annual growth rate per unit time (here a year) is $\exp(\beta_1) - 1$.

Hence, the average growth rate per annum is estimated to be $e^{0.021} - 1 = 0.0212$, i.e. 2.12% per annum.

Example 7.3 - Exponential Regression

Note that since GDP is given relative to GDP in 1950 (year 0), ln(GDP) in year 0 is **by definition** zero.

In such cases, we can leave out the intercept (constant) in the regression equation for ln(GDP).

In SPSS this can be done by clicking on the options menu in the regression window and unchecking the option "Include constant in the equation".

Example 7.3 - Exponential Regression

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	Time	.021	.000	1.000	172.668	.000

a. Dependent Variable: ln_GDP

b. Linear Regression through the Origin

Example 7.3 - Exponential Regression

The model obtained in this way is

$$\ln(GDP) = 0.021 time.$$

Exponentiating, we obtain

$$GDP = \exp(0.021 time).$$

The estimate of the relative GDP in 2010 (year 60) is given by

$$GDP = \exp(0.021 \times 60) = 3.525.$$