Linear Correlation & Regression

Dr. Soe Moe Naing

Professor & Head

P & SM, UMM

Where are we now?

- Population and sample?
- Types of variable ?
- Descriptive statistics (scatter diagram)?
- Data summarization?
- Normal distribution?
- Hypothesis testing (“*t*” test)?

Introduction

- In analyzing data for the health sciences, it is frequently desirable to learn something about the relationship between two **numeric** variables
- E.g. blood pressure and age, height and weight, total family income and health care expenditures

Introduction

- can be examined using linear models (i.e. **regression** and **correlation**)
- two statistical techniques; although related, serve different purposes

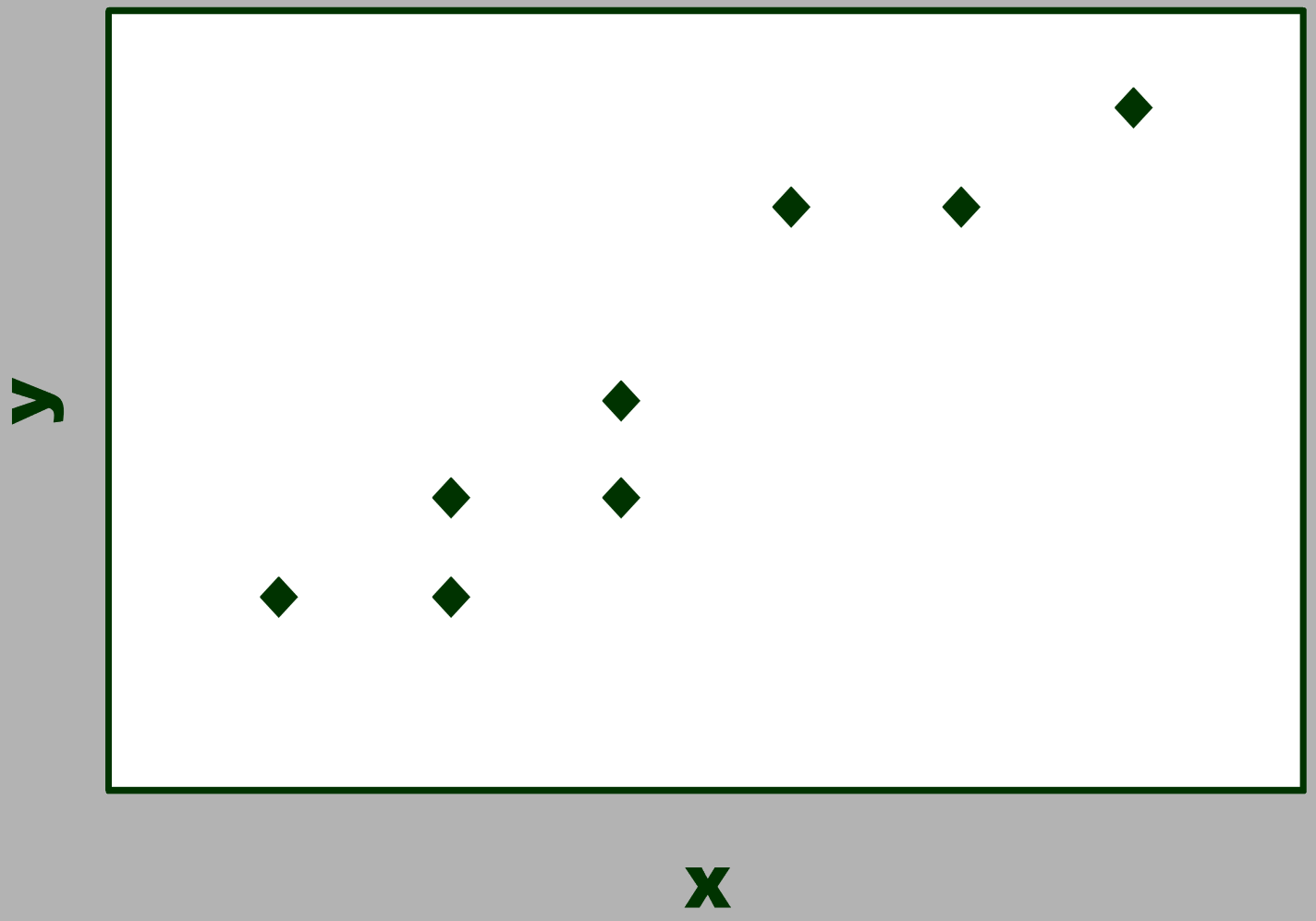
Correlation

- concerned with measuring the strength of the relationship between variables

Regression

- concerned with predicting or estimating the value of one variable corresponding to a given value of another variable

Scatter plot diagram

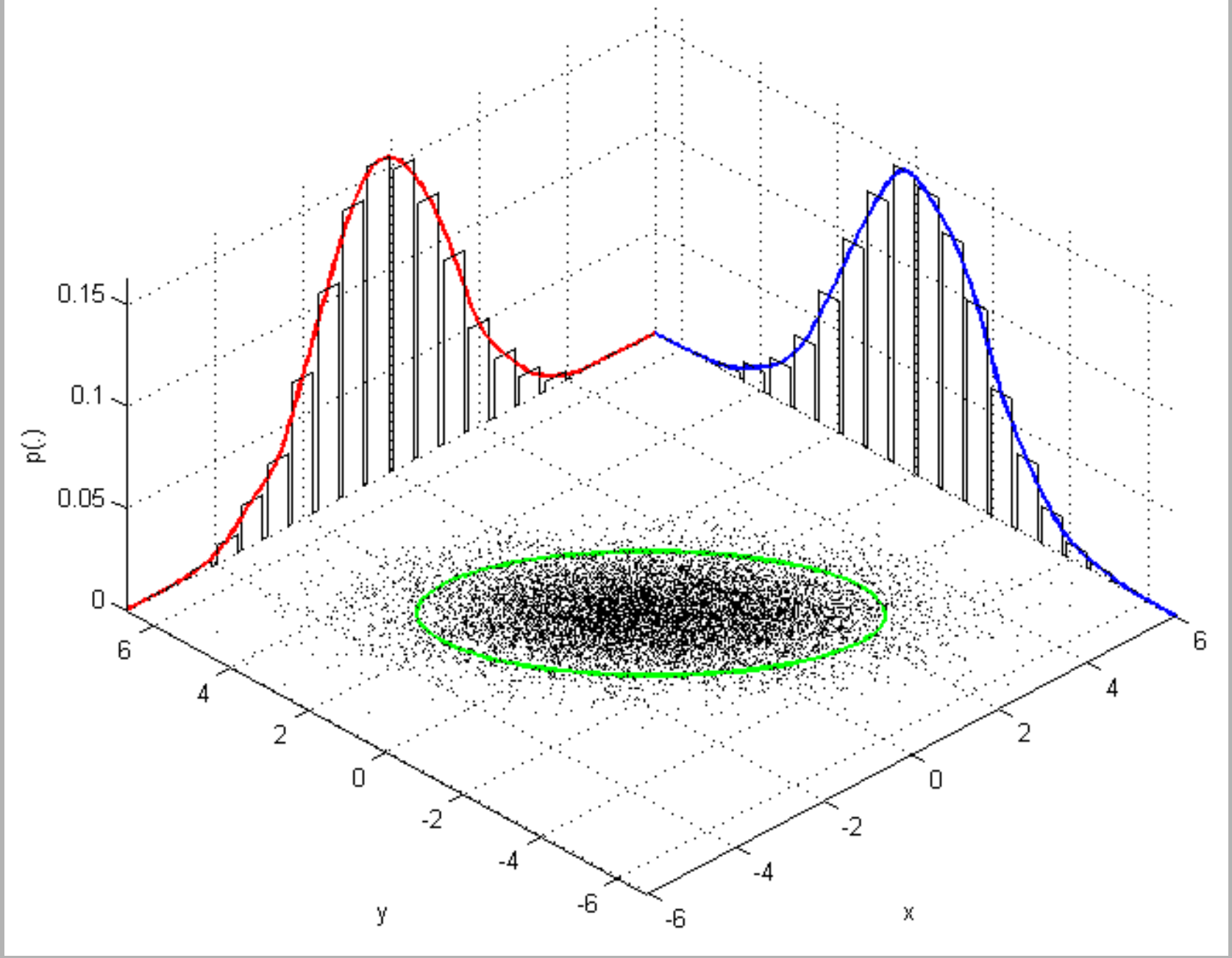


Correlation assumptions

1. For each value of X there is a normally distributed subpopulation of Y values.

2. For each value of Y there is a normally distributed subpopulation of X values.

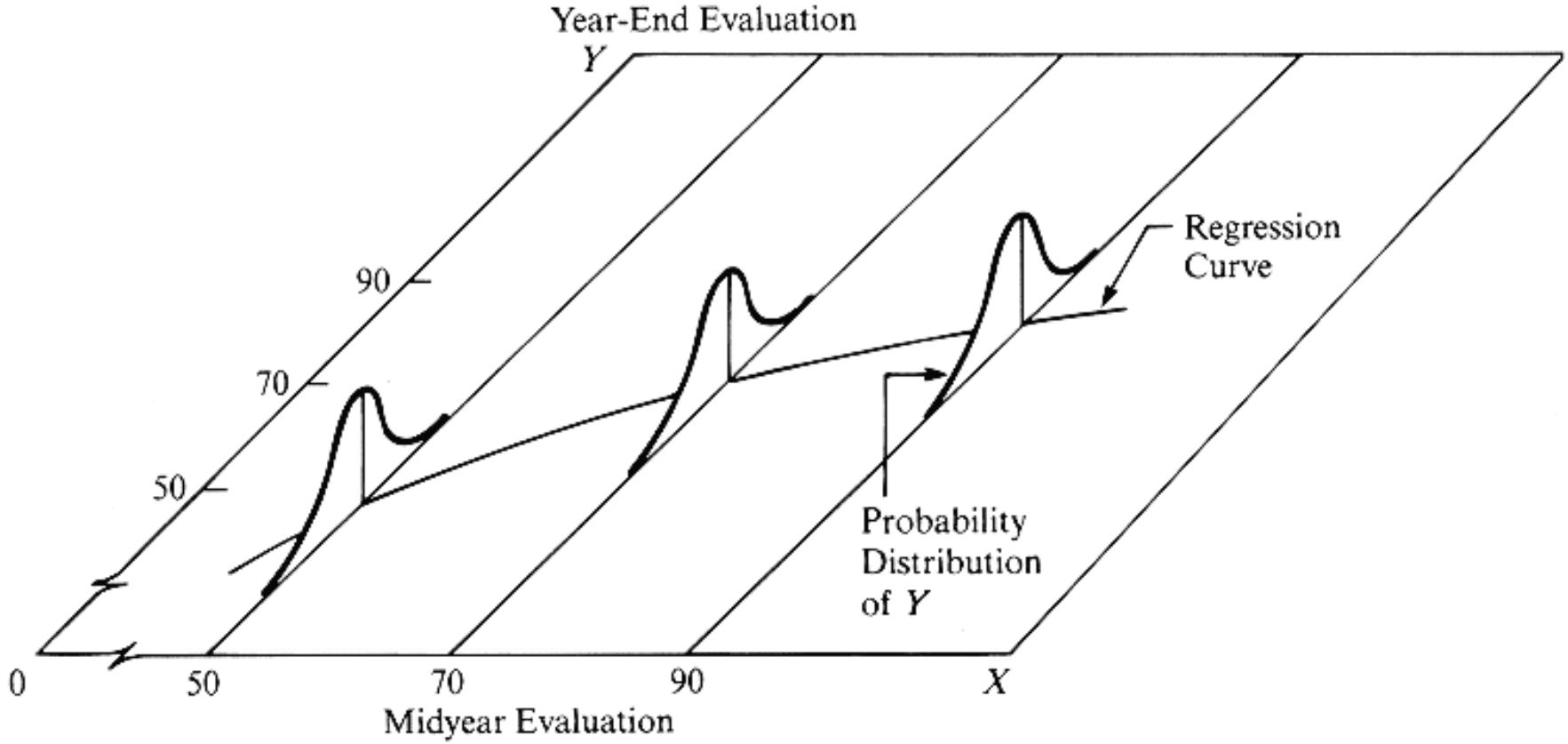
3. The joint distribution of X and Y is a normal distribution called the ***bivariate normal distribution***.


0.15
0.1
0.05
0
p(.)
6
4
2
0
-2
-4
-6
y
-6
-4
-2
0
2
4
6
x

Correlation assumptions

4. The subpopulations of Y values all have the same variance.

5. The subpopulations of X values all have the same variance.


Year-End Evaluation
Y
90
70
50
Regression
Curve
Probability
Distribution
of Y
0
50
70
90
X
Midyear Evaluation

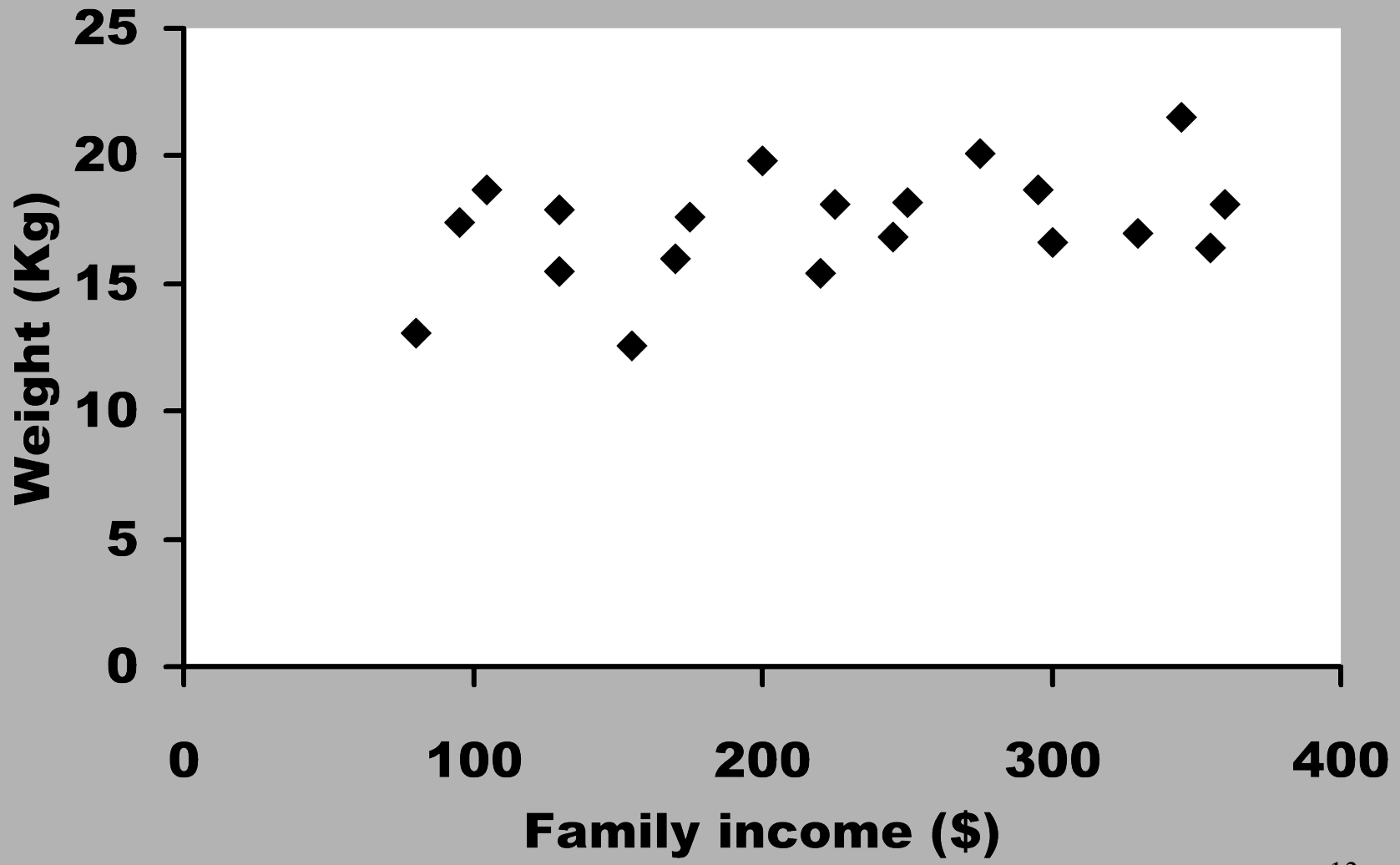
Weights/ family incomes of 20 children 5 years of age

Family income in $/ Year (x)	Weight in Kg (y)	Family income in $/ Year (x)	Weight in Kg (y)
130	15.5	225	18.1
200	19.8	95	17.4
345	21.5	130	17.9
245	16.8	330	17.0
155	12.6	295	18.7
300	16.6	170	16.0
360	18.1	250	18.2
105	18.7	355	16.4
80	13.1	220	15.4
275	20.1	175	17.6

The objective was to examine whether, for this sample of children, weight and family income were related.

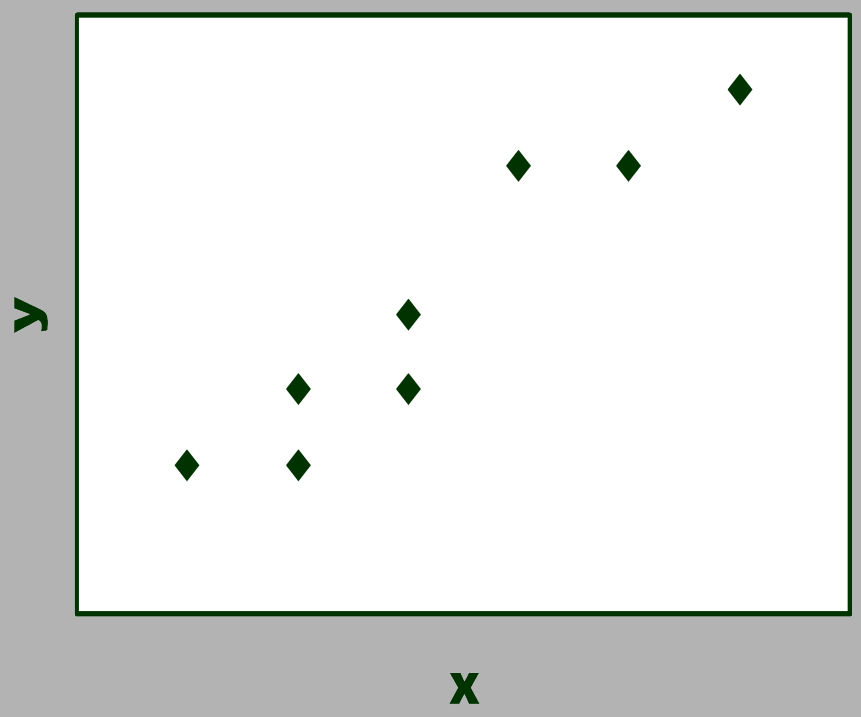
The following scatter diagram can be drawn:

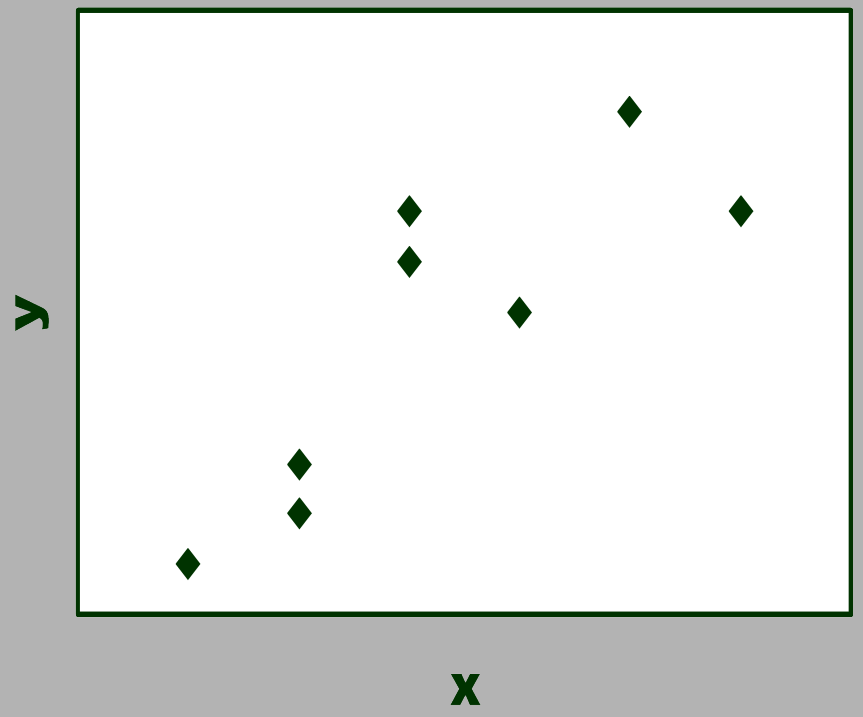
Figure 1. Weights and family incomes of 20 children 5 years of age.



CORRELATION COEFFICIENT

Consider the following two scatter diagrams:





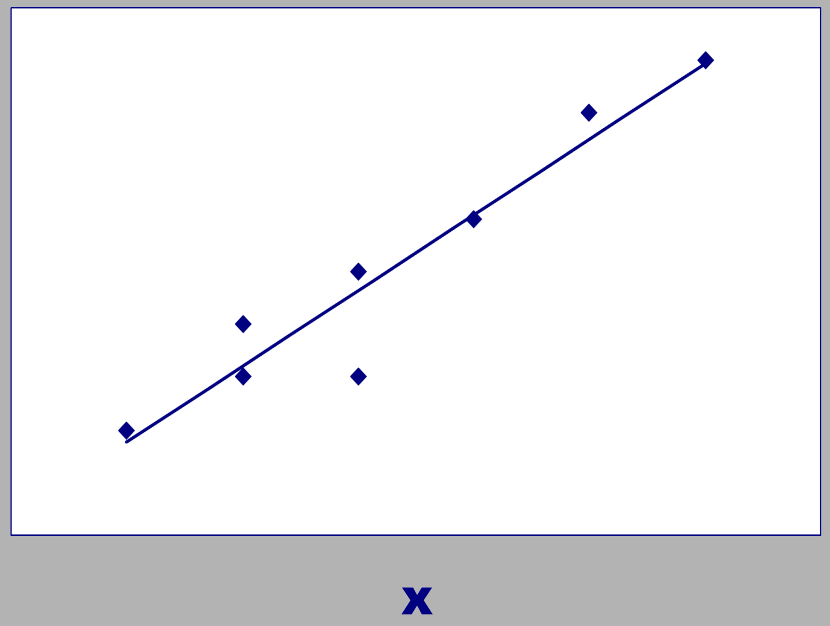
• The aim of PEARSON'S CORRELATION COEFFICIENT (r) is to measure the precision of the linear relationship between two variables.

The calculation of correlation coefficient (r):

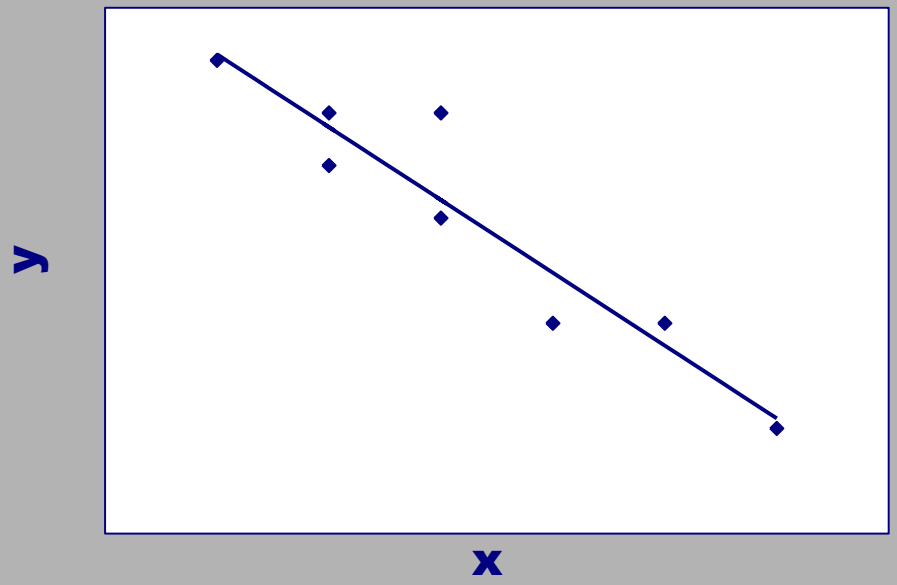
$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}}$$

Properties of correlation coefficient

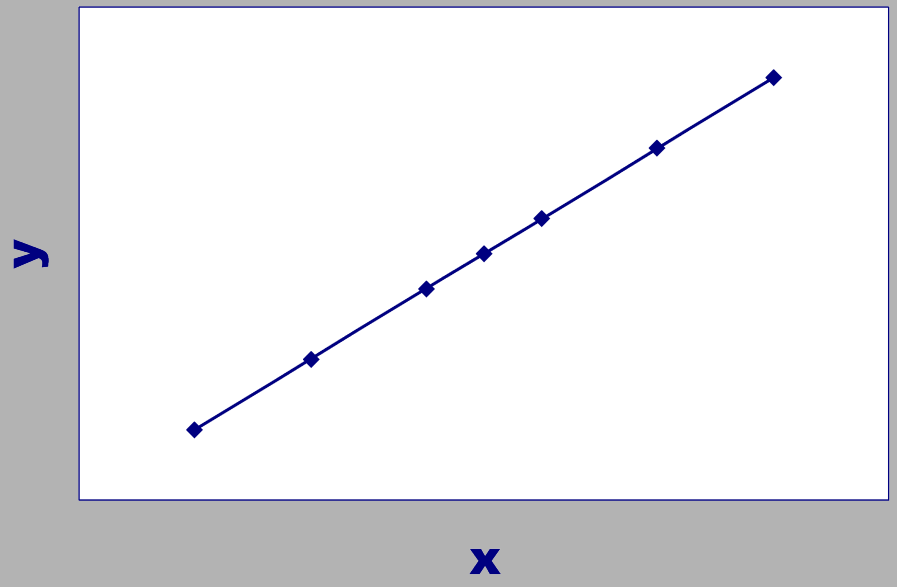
1. For any data set, (r) lies between (- 1) and (+ 1).
2. If (r) = (+ 1), or (- 1), the relationship is perfect, that is, all the points lie exactly on a line.
3. If (r) = (+ 1), variable y increases as x increases; if (r) = (- 1), variable y decreases as x increases



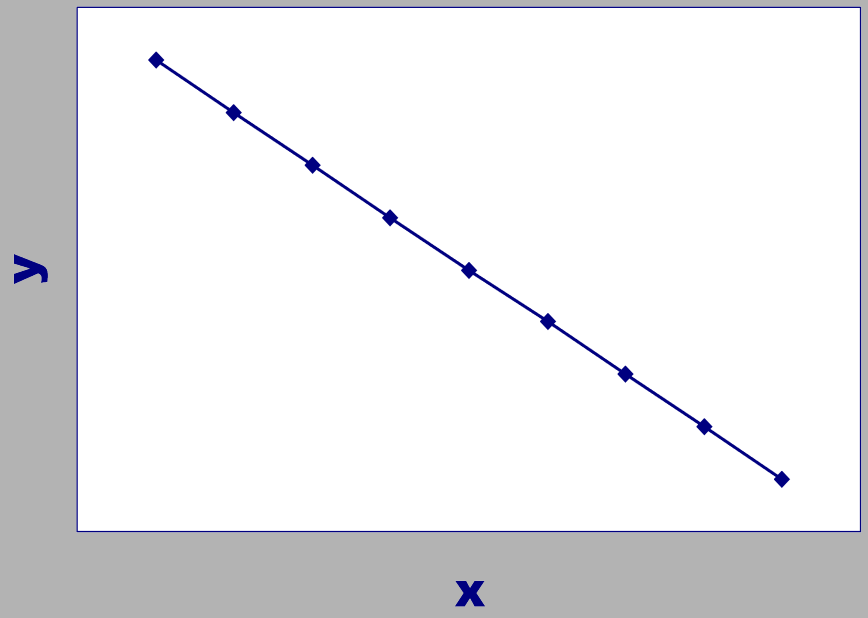
$0 < (r) < (+1)$



$(-1) < (r) < 0$



(r) = + 1

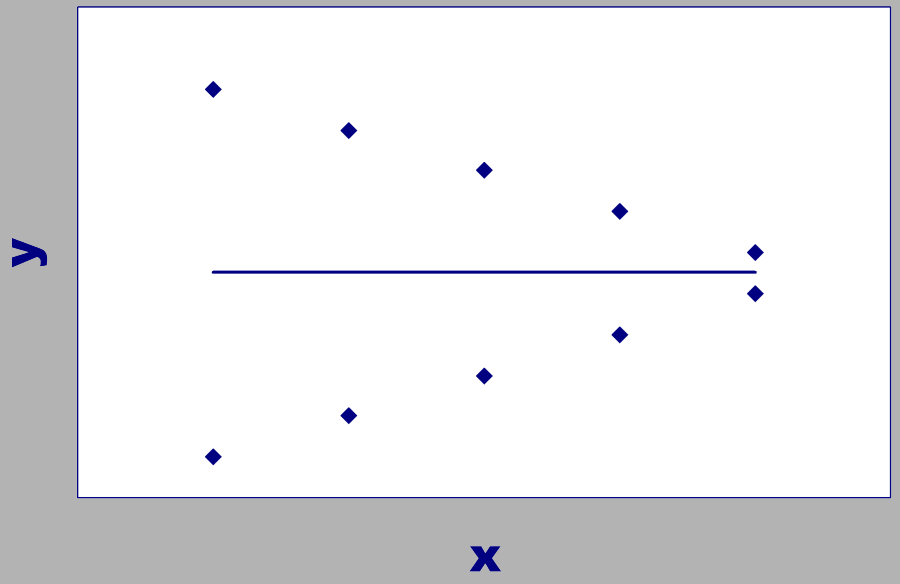


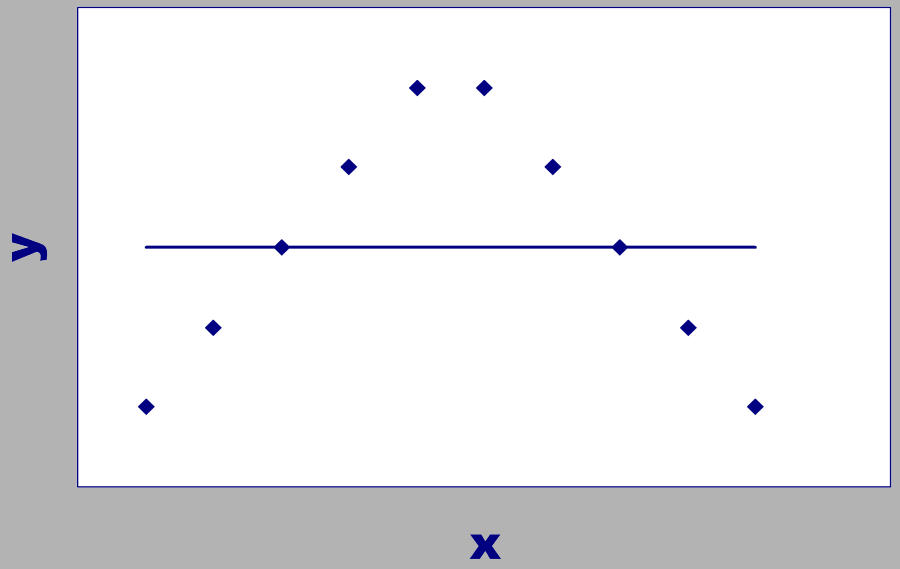
(r) = - 1

4. If (r) = 0, there is no linear relationship between y and x. This may mean that there is no relationship at all between the two variables

   (i.e. knowing x tells us nothing about the value of y).

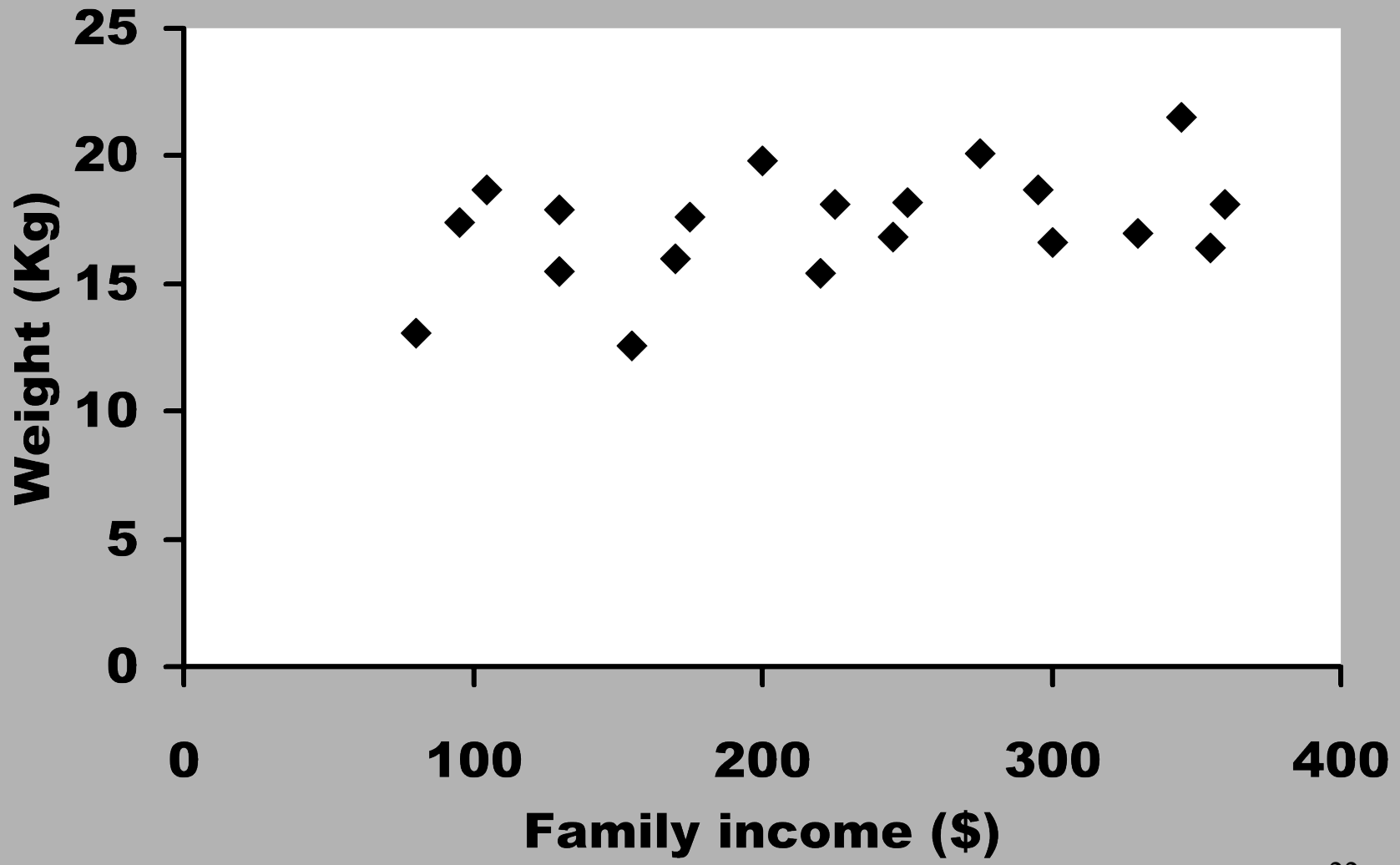
   However, we could also obtain (r) = 0 if there were a curved relationship between y and x.


y
x


y
x

5. A useful interpretation of (r) is that its square (r)^2 measures the proportion (%) of valid correlation against the variability (i.e. by chance) in variable y accounted for by the linear relationship with variable x.

Figure 1. Weights and family incomes of 20 children 5 years of age.



- From the example of weight and family income, the calculation gives;

  (r) = 0.414,

- it is possible (indicating the upwards sloping line), but a long way from 1 (indicating plenty of scatter about the line)

- Furthermore, calculation of $(r)^2$ gives;

$$r^2 = (0.414)^2 = 0.171 = 17\%,$$

which indicates the valid correlation between income and weight corresponds to 17% and variation by chance is 83%.

Hypothesis testing for population correlation coefficient (ρ) = 0

- We wish to see if the sample value of (r) = 0.414 is of sufficient magnitude to indicate that, in the population, income and weight are correlated. (α = 0.05)

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

- **"*p*" value**

  $p > 0.05$ (0.07)

- **Conclusion**

  Income and weight are not linearly correlated

Regression

- Regression method of analysis is **to predict** the value of one variable (dependent) corresponding to a given value of another variable (independent)

Simple linear regression assumptions

1. Independent variable (x) is non-random variable (i.e. fixed measurement)
2. Variable (x) is measured without error (i.e. error is negligible)
3. Values of dependent variable y are normally distributed

4. Variances of sub-populations of (y) are all equal

5. There is assumption of linearity (i.e. mean values of sub-population (y) lie in a straight line)

6. (y) values are independent each other (i.e. value of (y_1) for (x_1) is independent to value of (y_2)for (x_2)

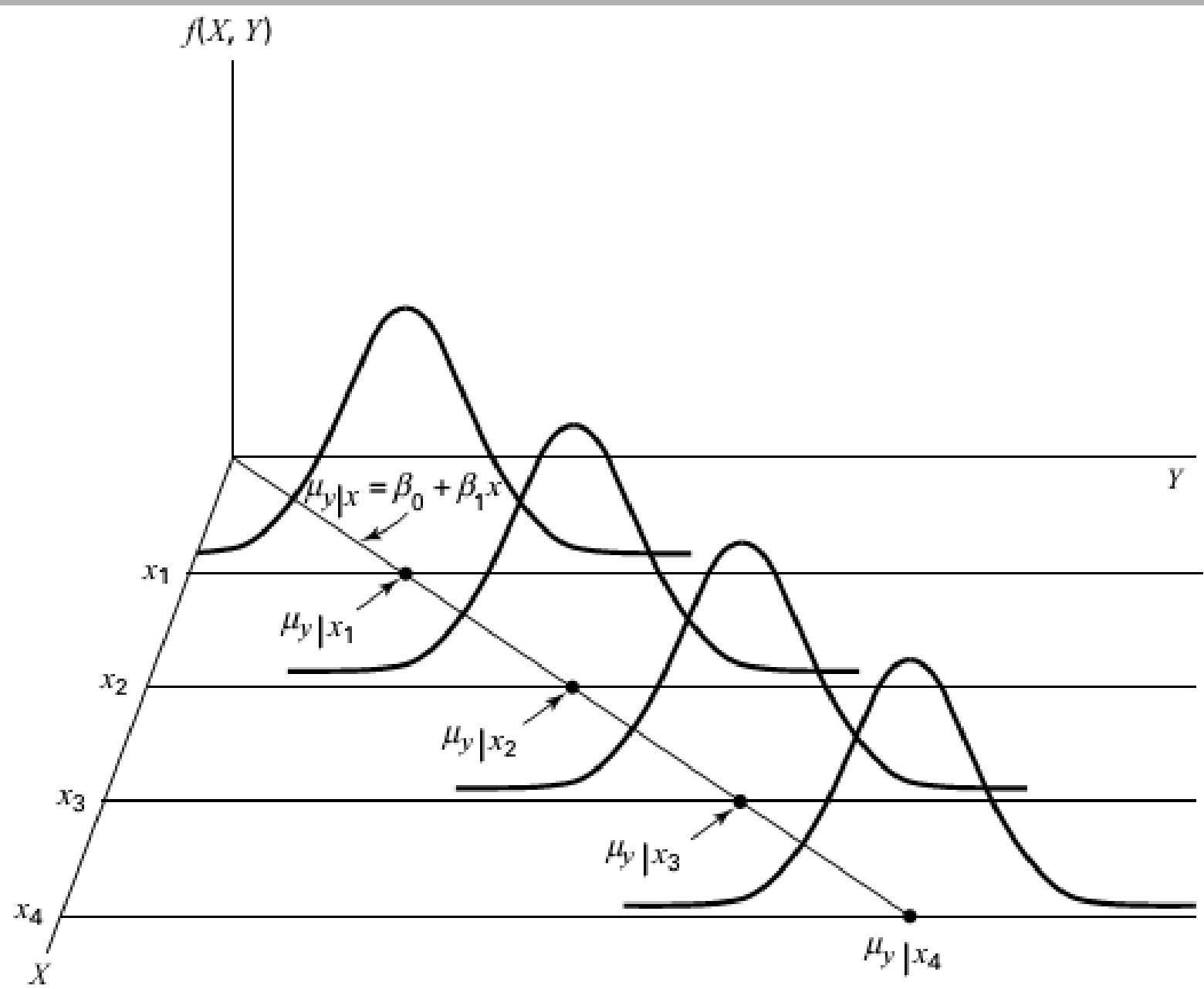


FIGURE 9.2.1 Representation of the simple linear regression model.

Weights/ family incomes of 20 children 5 years of age

Family income in $/ Year (x)	Weight in Kg (y)	Family income in $/ Year (x)	Weight in Kg (y)
130	15.5	225	18.1
200	19.8	95	17.4
345	21.5	130	17.9
245	16.8	330	17.0
155	12.6	295	18.7
300	16.6	170	16.0
360	18.1	250	18.2
105	18.7	355	16.4
80	13.1	220	15.4
275	20.1	175	17.6

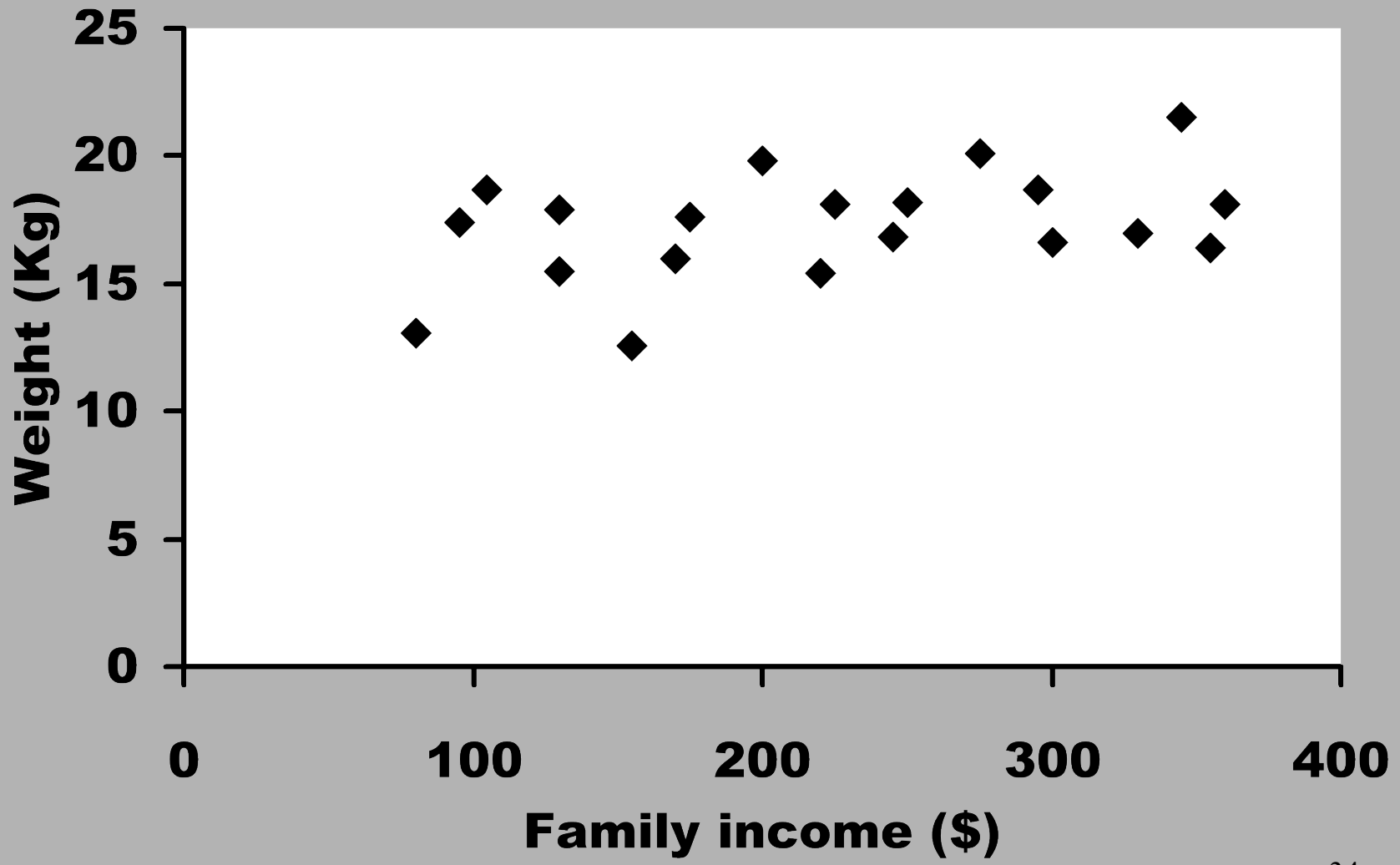
Scatter Diagram

- first step in examining the relationship between two variables, measured on the same subjects, is always to draw a “Scatter diagram”.

- In linear regression, there is a clear **dependent/ independent relationship** between the two numerical variables

- generally put the dependent variable on the vertical axis (the y-axis) and the independent variable in the horizontal axis (the x-axis).

Figure 1. Weights and family incomes of 20 children 5 years of age.



Fitting a regression Line

- In the scatter diagram there appears to be an upwards trend in weight, with increasing family income.
- Draw a line through the scatter of points, as a simple summary of the relationship between those variables.

Any straight line drawn on a graph can be represented by the regression equation:

$$\hat{y} = a + b\hat{x}$$

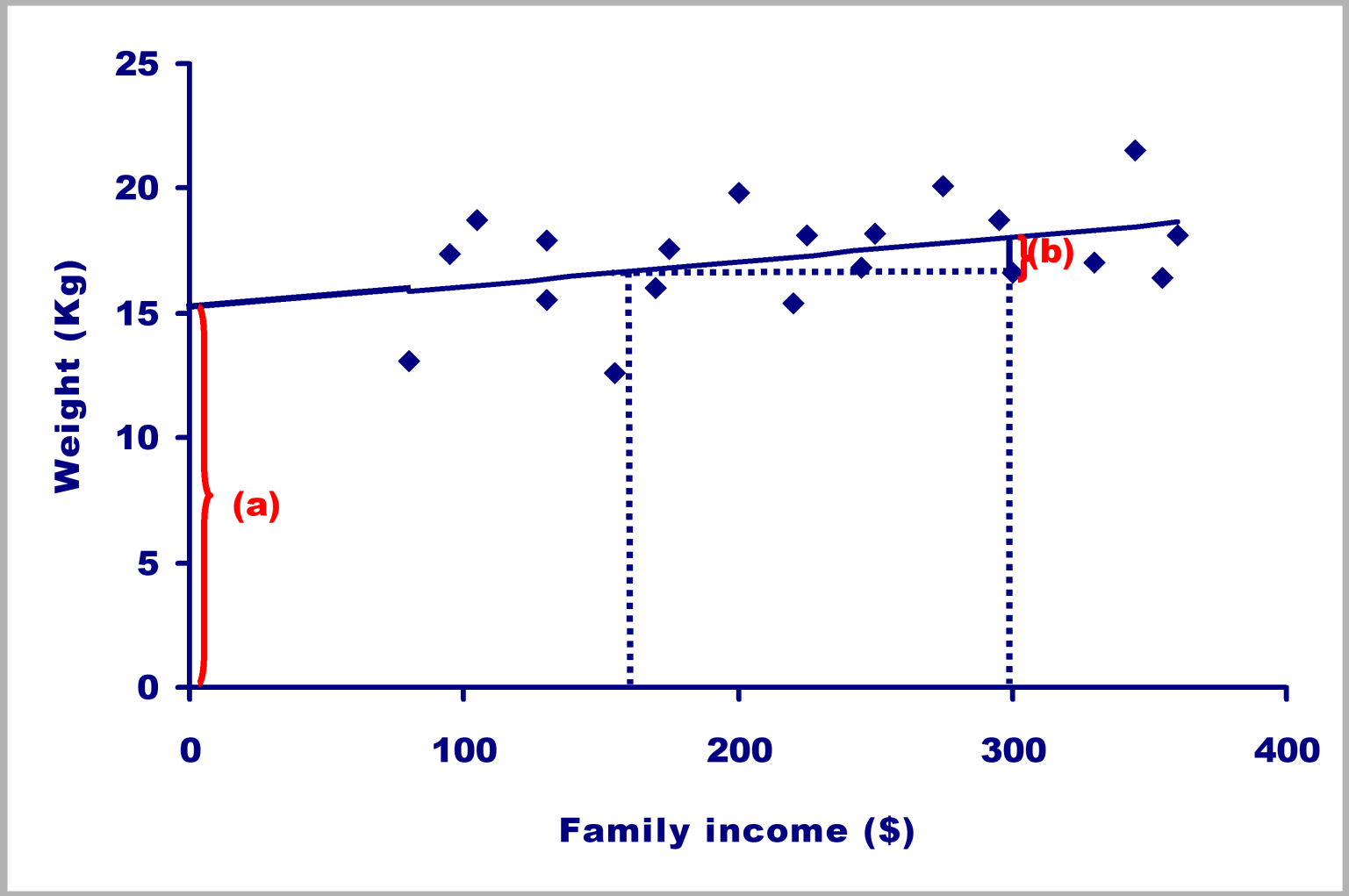
y = dependent variable

a = intercept

b = slope

x = independent variable

Figure 3. Regression line for weights and family incomes of 20 children 5 years of age.

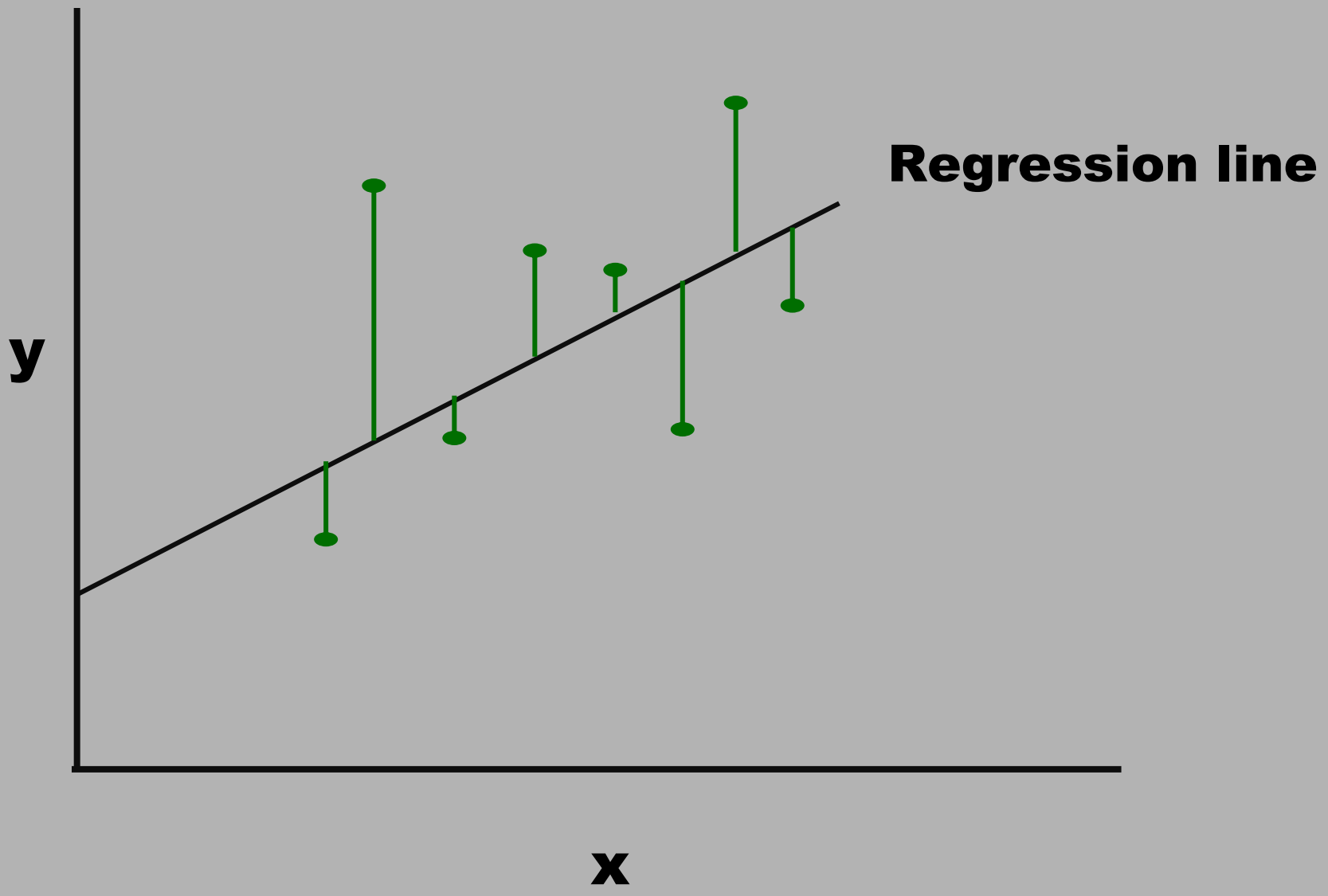


- **The fitted line is called the LINEAR REGRESSION of weight on family income.**

The **LEAST SQUARES METHOD** gives the "best" line.

- The sum of the squared vertical deviations of the observed data points (y_i) from the least-squares line is smaller than the sum of the squared vertical deviations of the data points from any other line

Regression line



Calculation of regression coefficient by LEAST SQUARES METHOD:

$$y = a + bx$$

$$b = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$$a = \bar{y} - b\bar{x}$$

$\bar{y}$ = mean value of y variable

$\bar{x}$ = mean value of x variable

Calculation of regression coefficient in our example:

a = 15.09 *b* = 0.00984

So, the equation of our fitted line is;

y = 15.09 + 0.00984 x

- The intercept ‘*a*’ (also called the constant), tells us the value of y when the value of x is “0”

- So in our example, weight of child is 15.09 Kg when family income is 0 $

- The slope '*b*', called the **REGRESSION COEFFICIENT**, tells us the increase in the average value of y corresponding to a unit increase in x

- E.g. mean weight increased by 0.00984 Kg (or about 10 grams) for each increase of $ 1 in family income (or 1 Kg weight gain per $ 100 increase)

caution:

- It is dangerous to extrapolate the regression line out side the range of the data. In our example, extrapolating the line to an income of $ 2000 per year would yield an estimated mean weight of 34.8 Kg, which is of course absurd.

The Coefficient of Determination

- It is calculated to evaluate the strength of linear regression equation

- Compares the scatter of regression line ($\hat{y}$) about ($\bar{y}$) line with scatter of observed (y) measurements about ($\bar{y}$) line

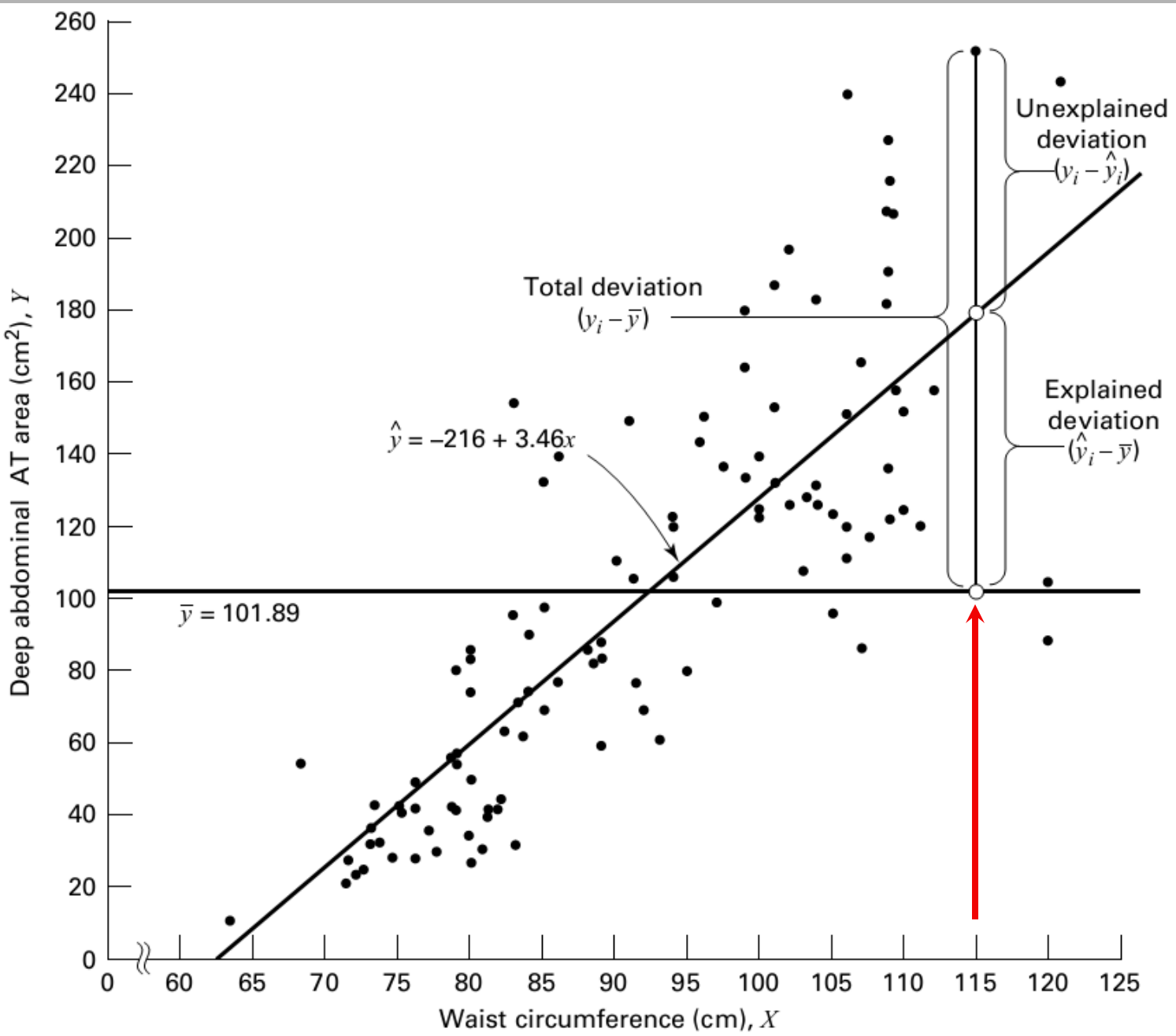


FIGURE 9.4.4 Scatter diagram showing the total, explained, and unexplained deviations for a selected value of Y, Example 9.3.1.

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

total deviation	explained deviation	unexplained deviation

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

total sum of squares	explained sum of squares	unexplained sum of squares

SST = total sum of square

SSR = explained sum of square

SSE = unexplained sum of square

$$r^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{SSR}{SST}$$

The Coefficient of Determination

- The sample correlation coefficient is the square root of the sample coefficient of determination (v.v. the sample coefficient of determination is square of the sample correlation coefficient)

- E.g. $r^2 = (0.414)^2 = 0.171 = 17\%$

Hypothesis testing for population regression coefficient (β) = 0

- We wish to see if the sample value of (b) = 0.00984 is of sufficient magnitude to indicate that change in income would change the weight (α = 0.05)

$$t = \frac{\hat{\beta}_1 - (\beta_1)_0}{s_{\hat{\beta}_1}}$$

- **"*p*" value**

  $p > 0.05$ (0.07)

- **Conclusion**

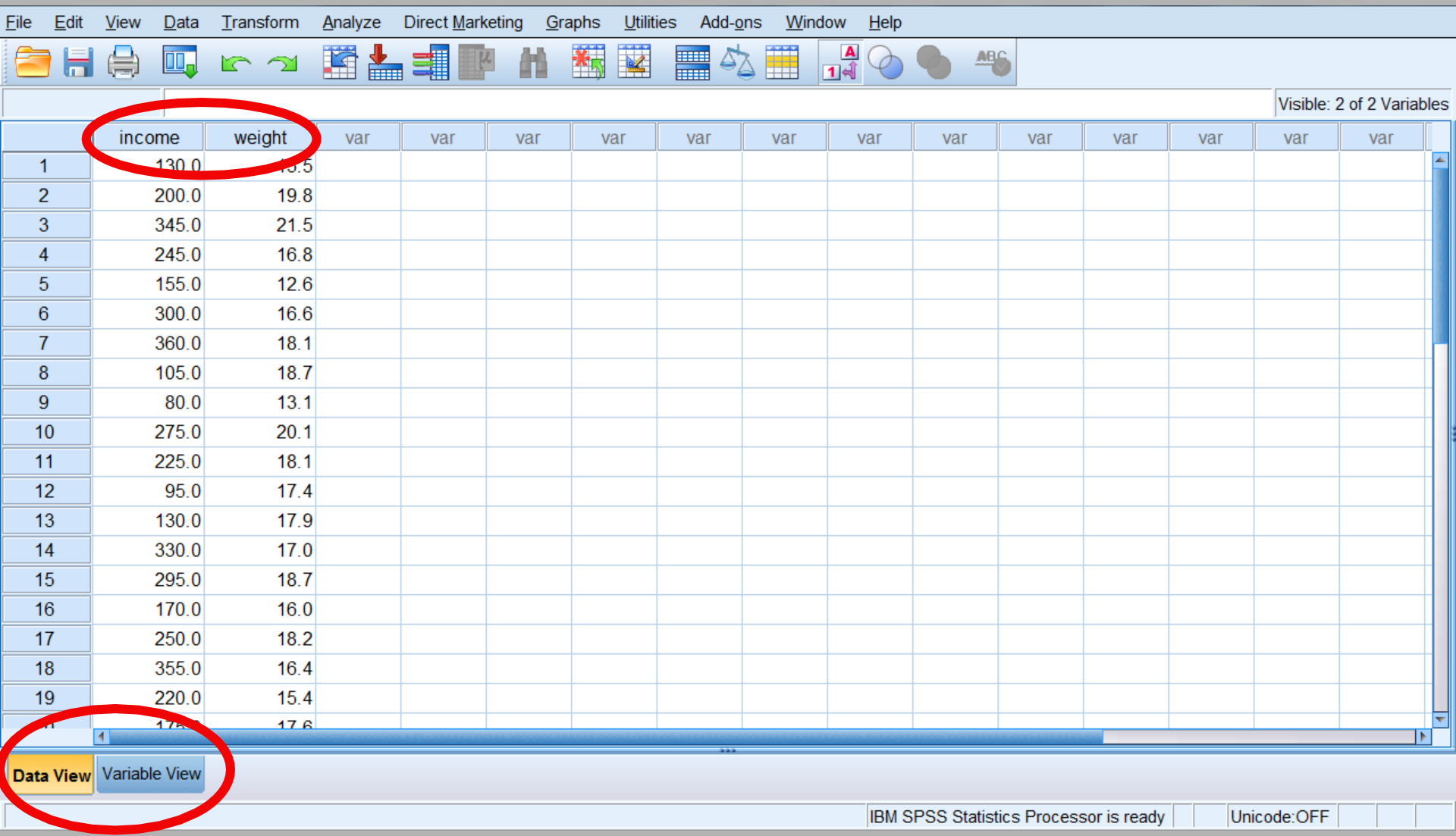
  Income and weight are not linearly correlated

Data management

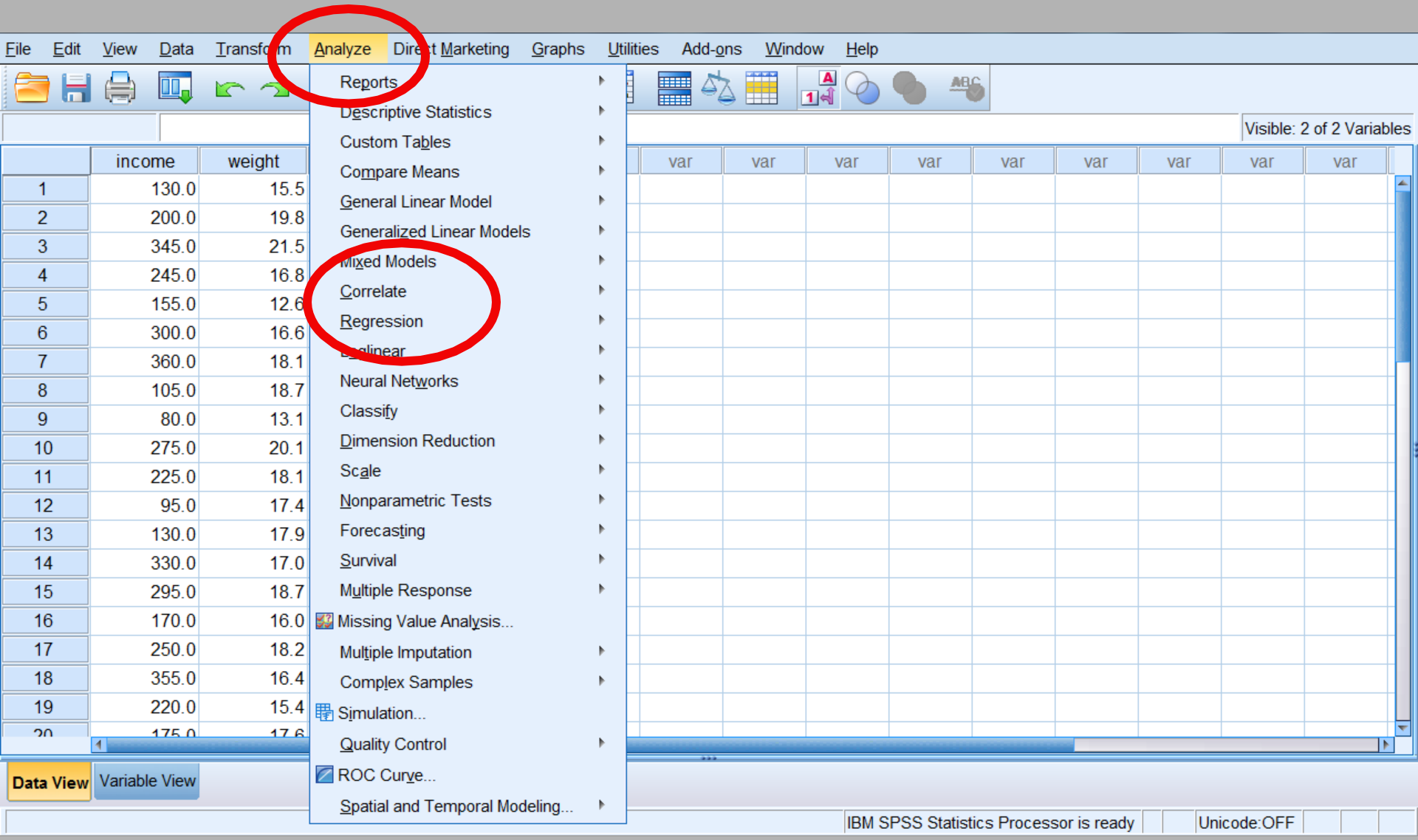
Raw data

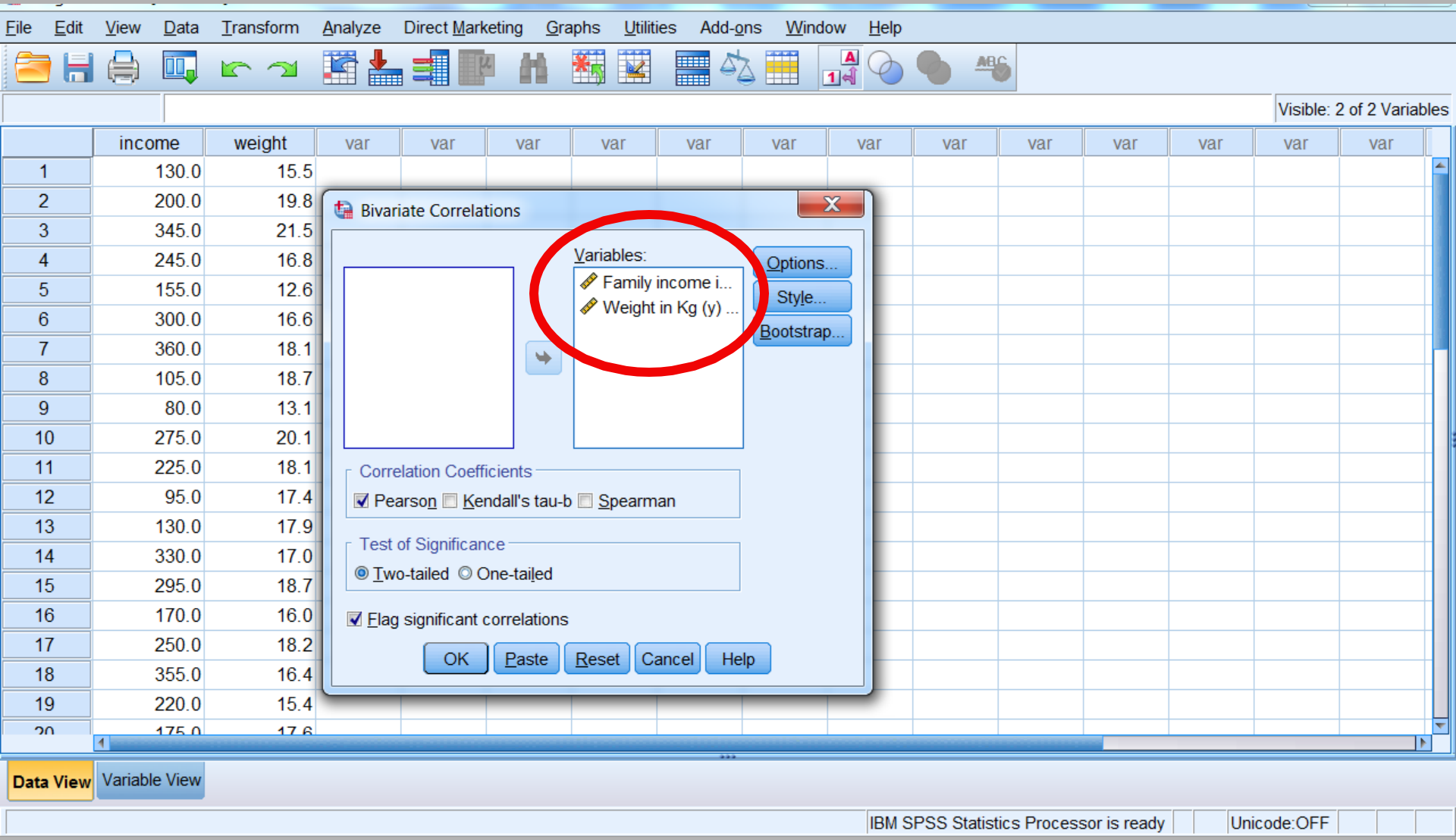
Family income in $/ Year (x)	Weight in Kg (y)	Family income in $/ Year (x)	Weight in Kg (y)
130	15.5	225	18.1
200	19.8	95	17.4
345	21.5	130	17.9
245	16.8	330	17.0
155	12.6	295	18.7
300	16.6	170	16.0
360	18.1	250	18.2
105	18.7	355	16.4
80	13.1	220	15.4
275	20.1	175	17.6

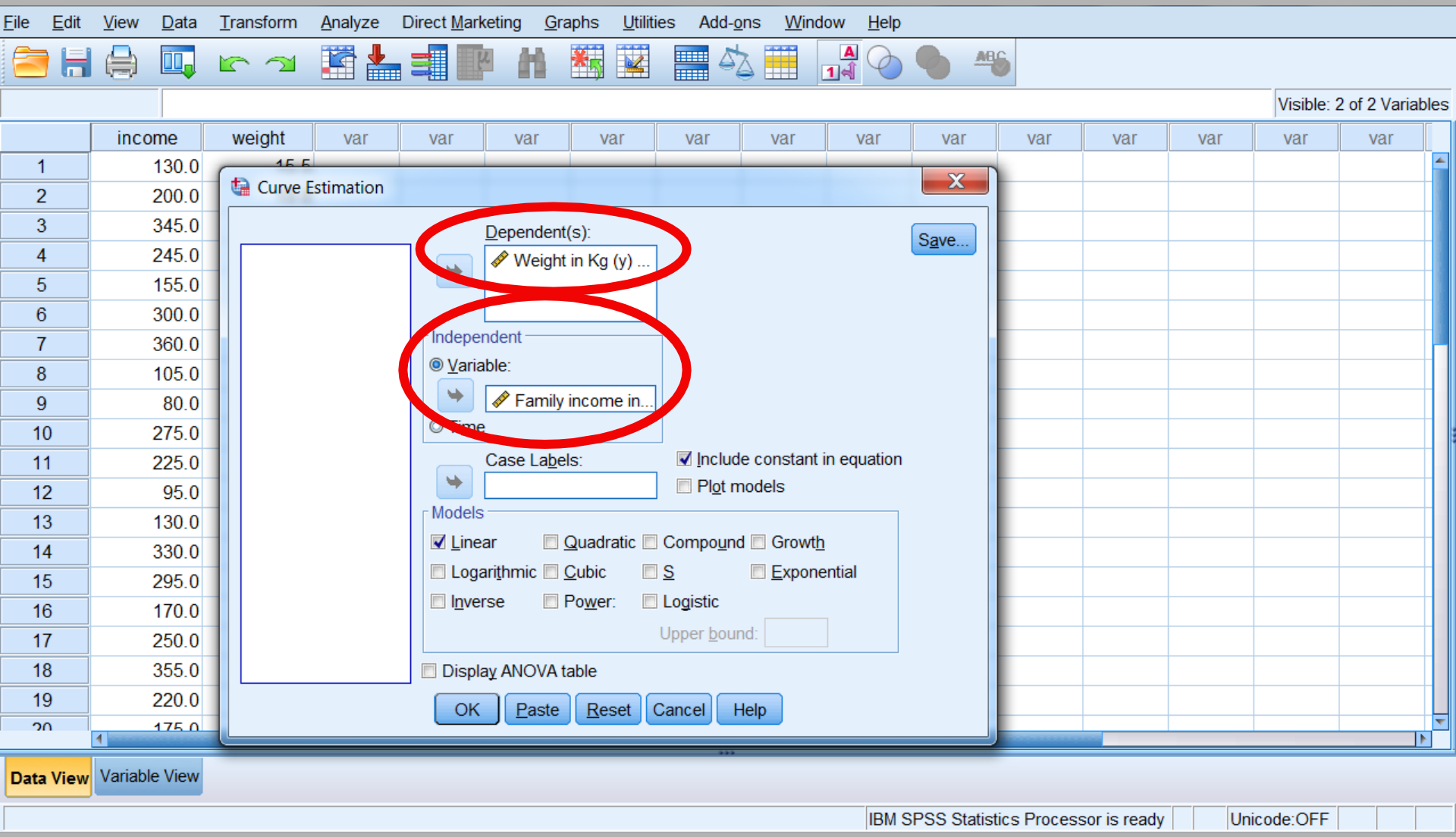
Example: SPSS

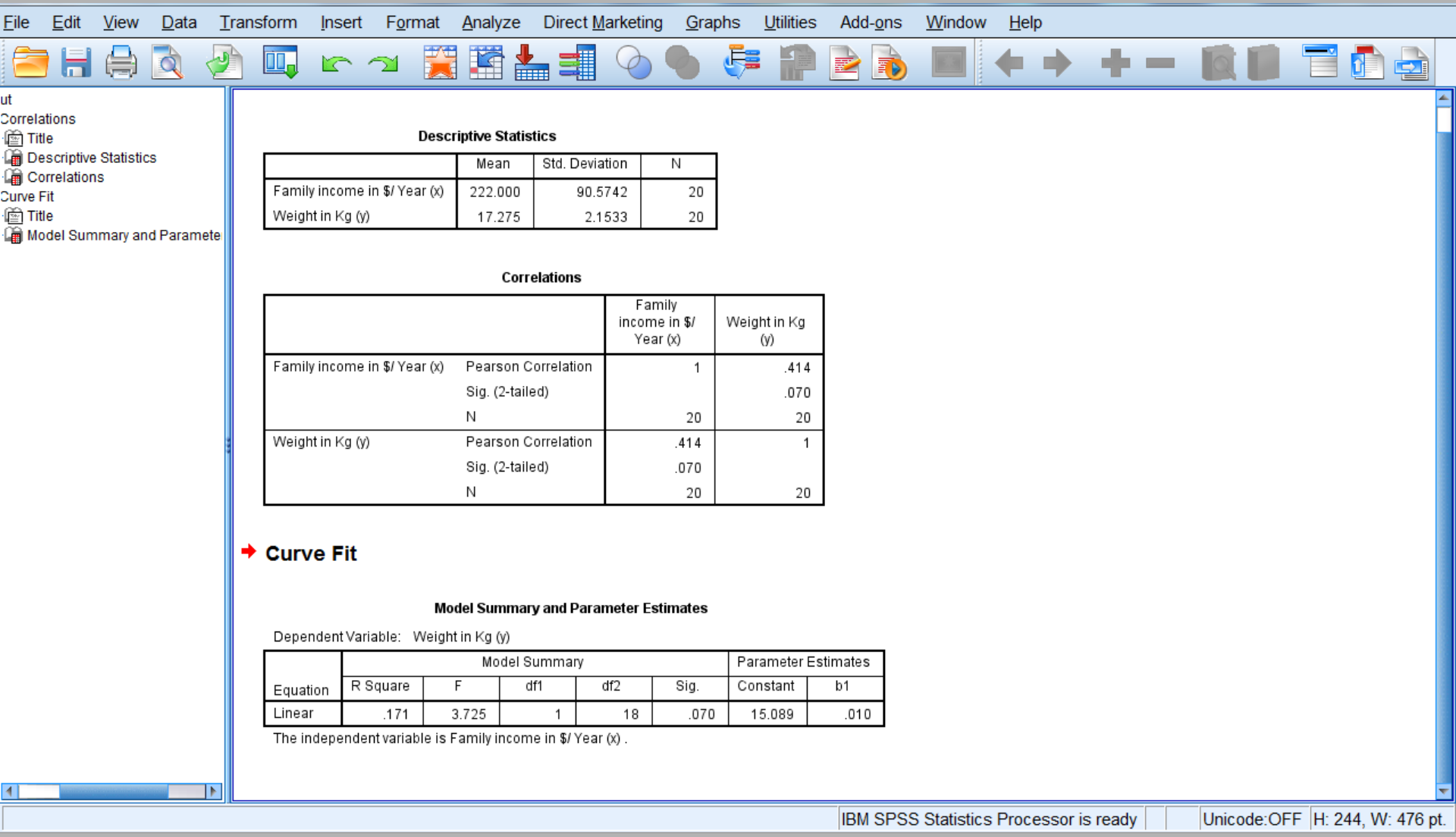


	income	weight	var	var	var	var	var	var	var	var	var	var	var	var	var
1	130.0	19.5													
2	200.0	19.8													
3	345.0	21.5													
4	245.0	16.8													
5	155.0	12.6													
6	300.0	16.6													
7	360.0	18.1													
8	105.0	18.7													
9	80.0	13.1													
10	275.0	20.1													
11	225.0	18.1													
12	95.0	17.4													
13	130.0	17.9													
14	330.0	17.0													
15	295.0	18.7													
16	170.0	16.0													
17	250.0	18.2													
18	355.0	16.4													
19	220.0	15.4													


File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Custom Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
ROC Curve...
Spatial and Temporal Modeling...
Visible: 2 of 2 Variables
income weight
1 130.0 15.5
2 200.0 19.8
3 345.0 21.5
4 245.0 16.8
5 155.0 12.6
6 300.0 16.6
7 360.0 18.1
8 105.0 18.7
9 80.0 13.1
10 275.0 20.1
11 225.0 18.1
12 95.0 17.4
13 130.0 17.9
14 330.0 17.0
15 295.0 18.7
16 170.0 16.0
17 250.0 18.2
18 355.0 16.4
19 220.0 15.4
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:OFF


File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
income weight
1 130.0 15.5
2 200.0 19.8
3 345.0 21.5
4 245.0 16.8
5 155.0 12.6
6 300.0 16.6
7 360.0 18.1
8 105.0 18.7
9 80.0 13.1
10 275.0 20.1
11 225.0 18.1
12 95.0 17.4
13 130.0 17.9
14 330.0 17.0
15 295.0 18.7
16 170.0 16.0
17 250.0 18.2
18 355.0 16.4
19 220.0 15.4
Bivariate Correlations
Variables:
Family income i...
Weight in Kg (y) ...
Options...
Style...
Bootstrap...
Correlation Coefficients
Pearson Kendall's tau-b Spearman
Test of Significance
Two-tailed One-tailed
Flag significant correlations
OK Paste Reset Cancel Help
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:OFF


File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
income
weight
var
1 130.0
2 200.0
3 345.0
4 245.0
5 155.0
6 300.0
7 360.0
8 105.0
9 80.0
10 275.0
11 225.0
12 95.0
13 130.0
14 330.0
15 295.0
16 170.0
17 250.0
18 355.0
19 220.0
Curve Estimation
Dependent(s):
Weight in Kg (y) ...
Save...
Independent
Variable:
Family income in...
Time
Case Labels:
Include constant in equation
Plot models
Models
Linear
Quadratic
Compound
Growth
Logarithmic
Cubic
S
Exponential
Inverse
Power:
Logistic
Upper bound:
Display ANOVA table
OK
Paste
Reset
Cancel
Help
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:OFF


File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
ut
Correlations
Title
Descriptive Statistics
Correlations
Curve Fit
Title
Model Summary and Paramete
Descriptive Statistics
Mean Std. Deviation N
Family income in $/ Year (x) 222.000 90.5742 20
Weight in Kg (y) 17.275 2.1533 20
Correlations
Family income in $/ Year (x) Weight in Kg (y)
Family income in $/ Year (x) Pearson Correlation 1 .414
Sig. (2-tailed) .070
N 20 20
Weight in Kg (y) Pearson Correlation .414 1
Sig. (2-tailed) .070
N 20 20
Curve Fit
Model Summary and Parameter Estimates
Dependent Variable: Weight in Kg (y)
Model Summary
Parameter Estimates
Equation R Square F df1 df2 Sig. Constant b1
Linear .171 3.725 1 18 .070 15.089 .010
The independent variable is Family income in $/ Year (x) .
IBM SPSS Statistics Processor is ready
Unicode:OFF
H: 244, W: 476 pt.

Correlations

		Family income in $/ Year (x)	Weight in Kg (y)
Family income in $/ Year (x)	Pearson Correlation	1	.414
	Sig. (2-tailed)		.070
	N	20	20
Weight in Kg (y)	Pearson Correlation	.414	1
	Sig. (2-tailed)	.070	
	N	20	20

Curve Fit

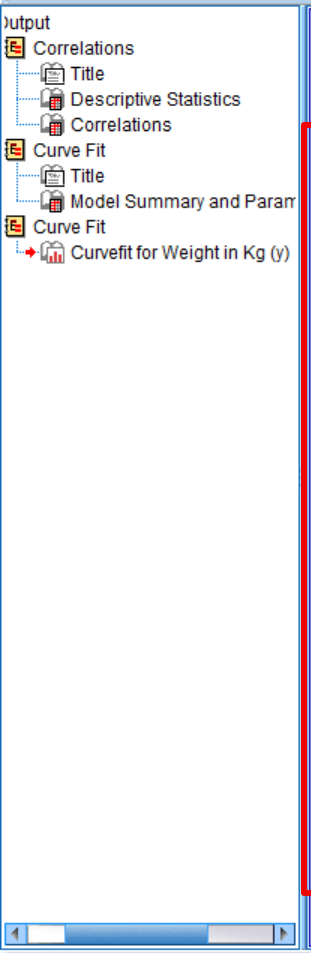
Model Summary and Parameter Estimates

Dependent Variable: Weight in Kg (y)

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.171	3.725	1	18	.070	15.089	.010

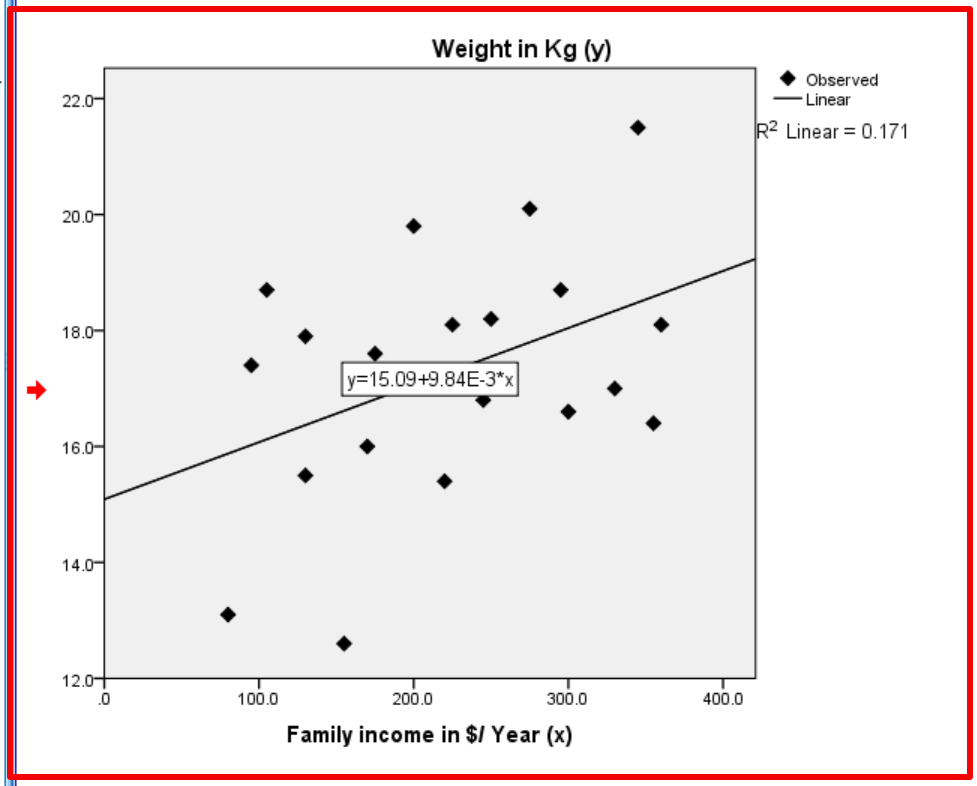
The independent variable is Family income in $/ Year (x).

File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help



Equation							
Linear	.171	3.725	1	18	.070	15.089	.010

The independent variable is Family income in $/ Year (x) .



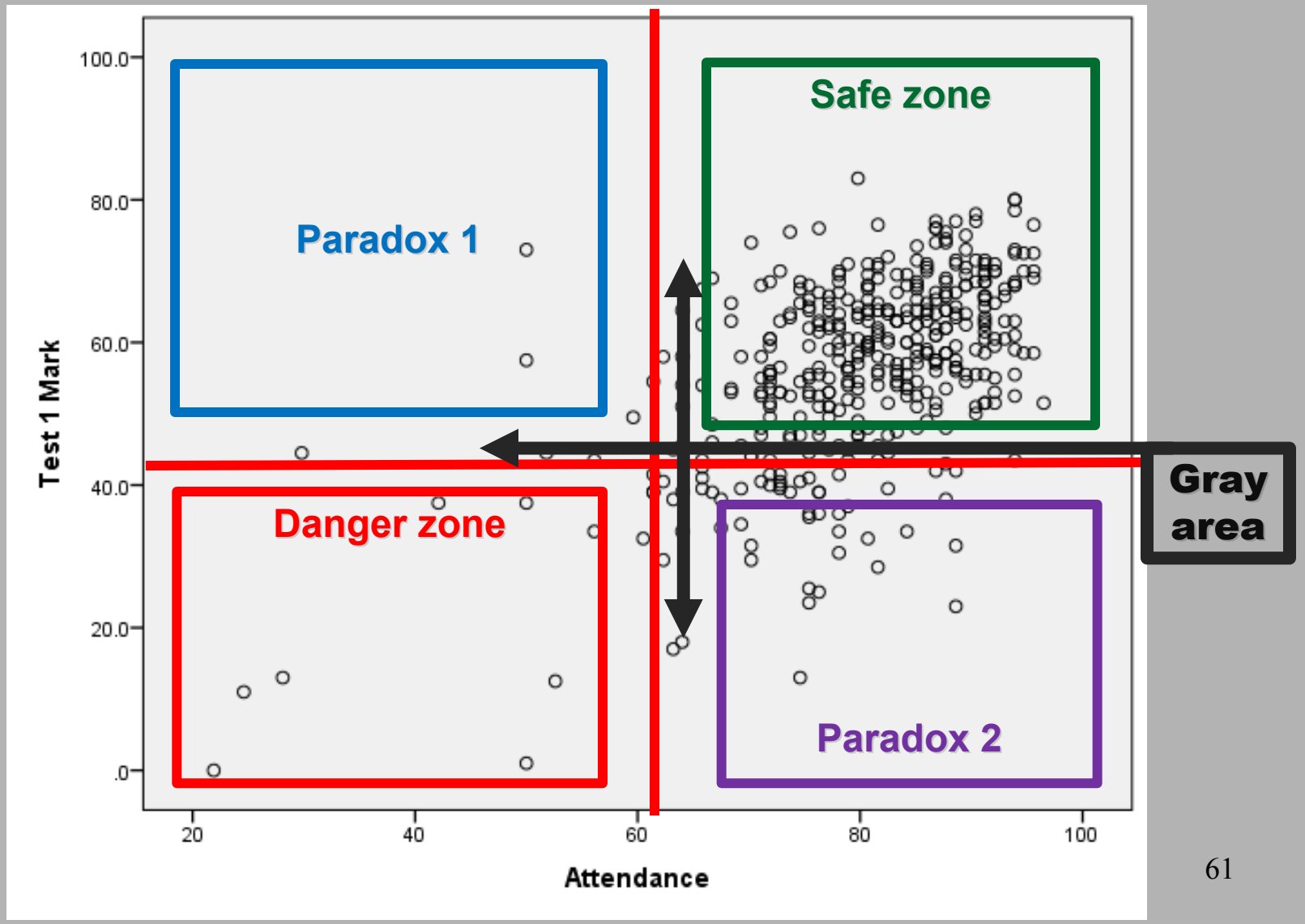
IBM SPSS Statistics Processor is ready Unicode:OFF

Data presentation

- depends on purpose of presentation
  - ? advocacy presentation
  - ? journal article publication
  - ? submission of thesis paper

Advocacy presentation (example)

Correlation (Attendance/ Test 1 Mark)



Journal article publication **(example)**

Table 1 Pearson correlation coefficient among variables

	GFFI	$\dot{V}O_2$ max
Age (years)	−0.38**	−0.61**
Total-cholesterol (mg/dL)	−0.35**	−0.25**
LDL-cholesterol (mg/dL)	−0.06	−0.06
HDL-cholesterol (mg/dL)	0.22	0.18*
Triglycerides (mg/dL)	−0.43**	−0.30**
Uric Acid (mg/dL)	−0.24**	−0.22*
Body Mass Index (kg/m^2)	−0.22*	−0.28**
Systolic Blood Pressure (mmHg)	−0.43**	−0.37**
Diastolic Blood Pressure (mmHg)	−0.34**	−0.20*
Nitrite (μM)	0.31**	0.16
T-BARS (μM)	0.15	0.13

* $p<0.05$ and **$p<0.01$.

Journal article publication **(example)**

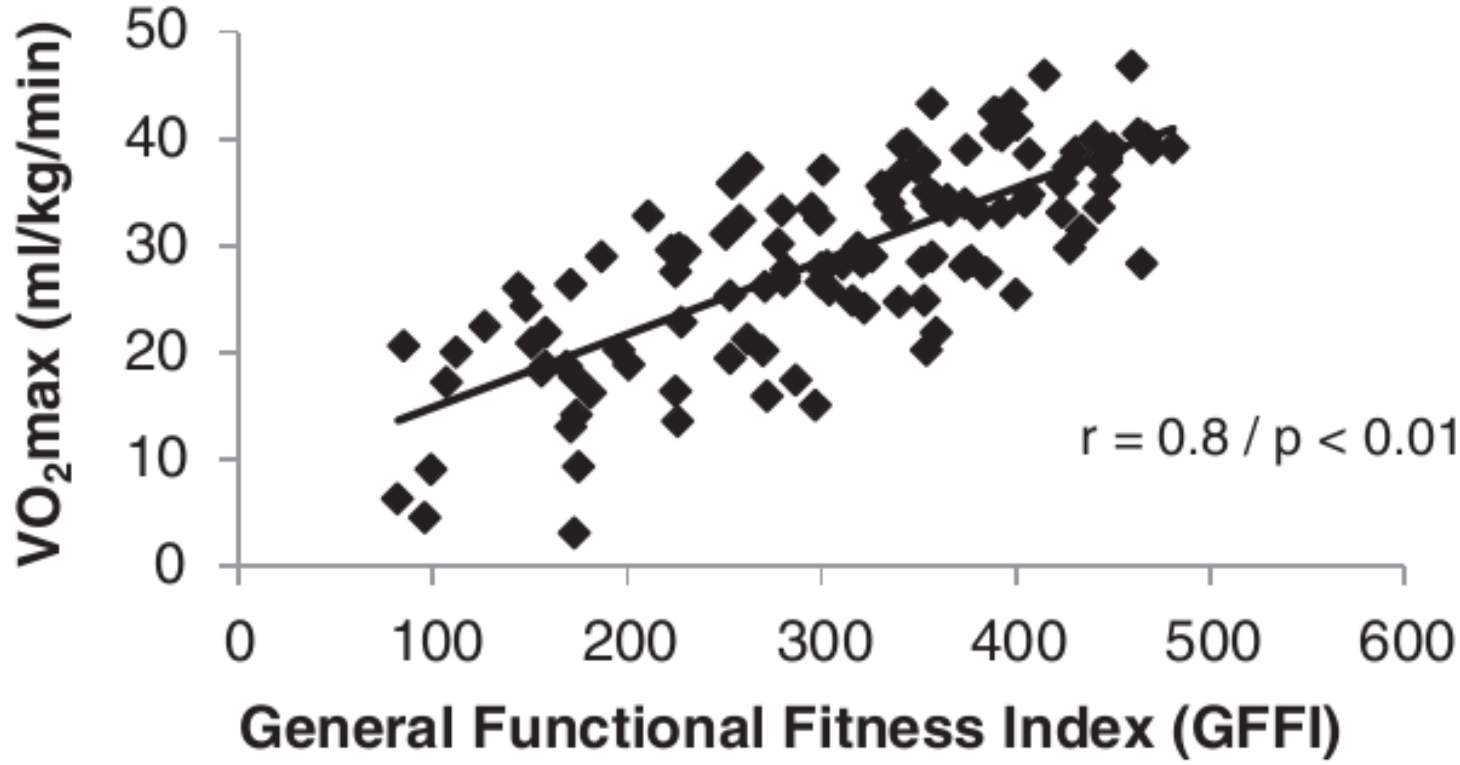


Figure 1 Pearson correlation coefficient between general functional fitness index (GFFI) and $\dot{V}O_2max$. *Correlation is significant at the 0.01 level.

Submission of thesis paper (example)

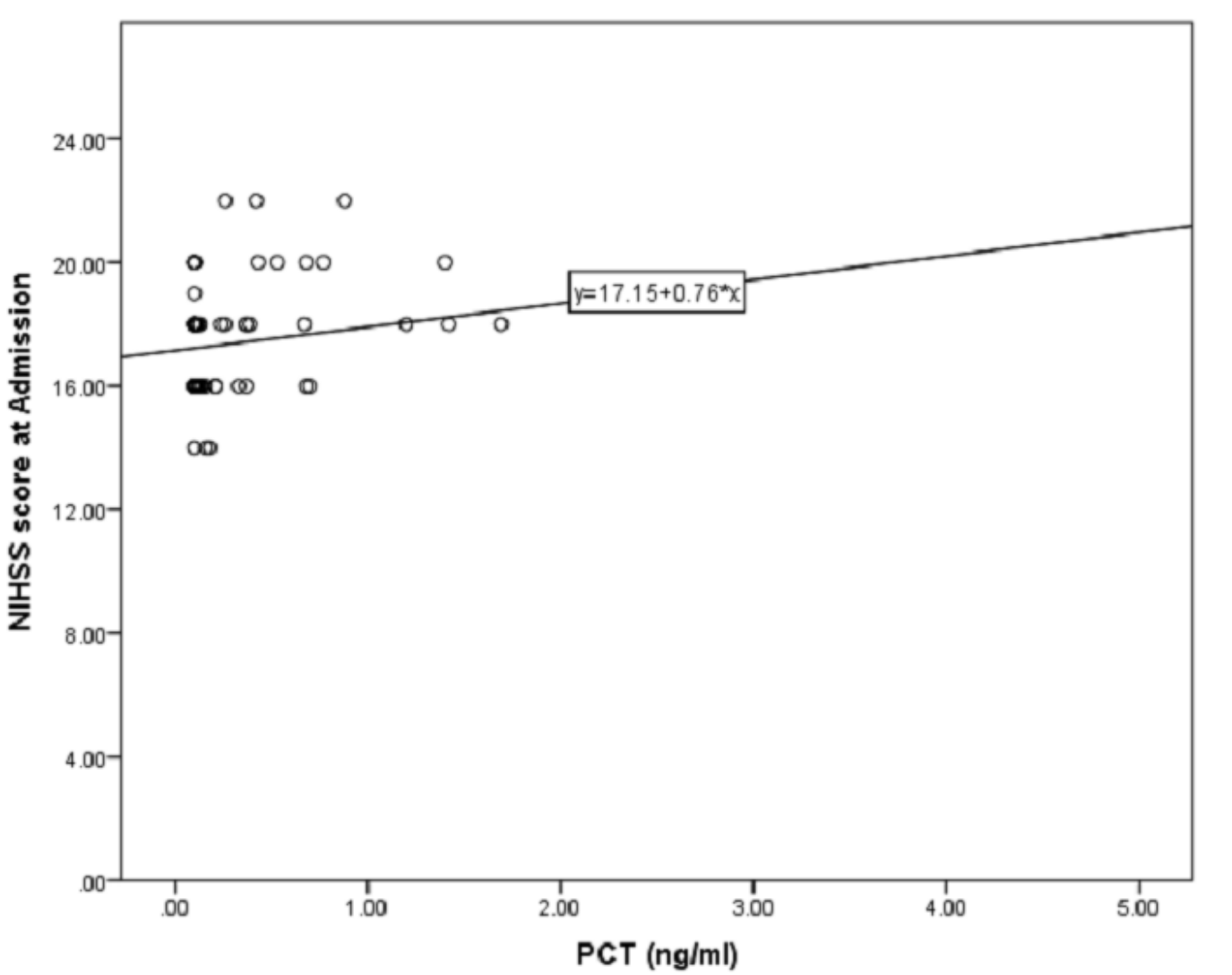
6.4. Correlation between Serum Procalcitonin Level and Initial Severity of Acute Ischemic Stroke

In this study on 70 patients with acute ischaemic stroke, there was a positive correlation between admission serum procalcitonin level and initial severity of acute ischemic stroke i.e. admission NIHSS score (r value = +0.47, P< 0.001). The strength of correlation was moderate.

Table 4. Correlation between serum procalcitonin level and admission NIHSS score

Correlation between Serum procalcitonin level and admission NIHSS score	Pearson correlation coefficient = r value	P value
	+ 0.47	<0.001

Submission of thesis paper (example)



r= +0.47

p= <0.001

Multiple regression

- To examine the interaction of multiple independent variables on one dependent variable
- Multiple regression fits an equation that predicts one variable (the dependent variable, Y) from two or more independent (X) variables.

- Three reasons of application

✓ to find an equation that best predicts dependent variable with multiple independent variables

✓ to find out which variable has the largest influence on dependent variable

✓ to find out interaction of only one of the independent variables, but analysis needs to adjust for differences in other variables.

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \ldots +$

(β_0 is the Y intercept; β_1 is the slope)

- when considering one X variable, other Xs are fixed or adjusted
- used for adjusted multivariate statistical analysis

Different methods are;

- ✓enter method
- ✓ forward method
- ✓ backward method
- ✓ stepwise method

Enter method

- ✓ enter all independent variables

Forward method

- ✓ procedure start with empty model
- ✓ deciding if a new single predictor should be added to a regression model
- ✓ in each forward step, one variable is added to give the single best improvement for regression model

Backward method

- ✓ procedure start with full model
- ✓ deciding if a single predictor should be deleted to a regression model
- ✓ in each forward step, one variable is deleted to give the further improvement for regression model

Stepwise method

✓ Forward stepwise begins with forward selection followed by a test for backward elimination

✓ Backward stepwise begins with backward elimination followed by a test for forward selection

✓ deciding if predictors should be added or deleted from a regression model on the basis of a fixed decision rule

Data interpretation

Time for;

- Exercise

Exercise

Task: interpret the results of given paper (**ONLY** relevant results on correlation or regression) and present your understandings.

- 5 minutes presentation for each group
- 5 minutes discussion after presentation by two groups with same exercise
- Exercise (1) for Group 1 and 2
- Exercise (2) for Group 3 and 4

Thanks