# Linear Correlation & Regression

## Prerequisites

- Population and sample
- Types of variable
- Descriptive statistics (scatter diagram)
- Data summarization
- Normal distribution
- Hypothesis testing

## 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
  - 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

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

Example

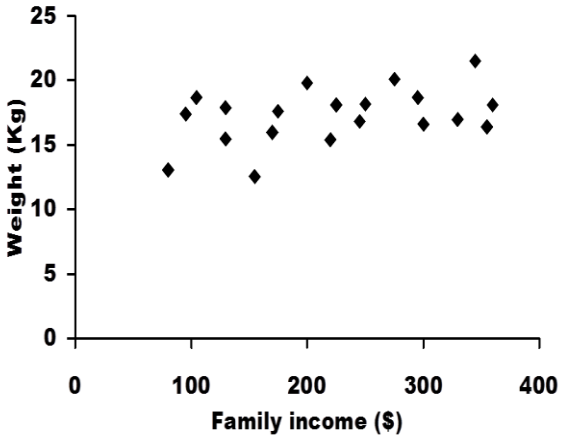
Table showing weights/ family incomes of 20 children 5 years of age;

| Family income in $/ Year (x) | Weight in Kg (y) | Family income in $/ Year (x) | Weight in Kg (y) | Family income in $/ Year (x) | Weight in Kg (y) | Family income in $/ Year (x) | Weight in Kg (y) |
|---|---|---|---|---|---|---|---|
| 130 | 15.5 | 300 | 16.6 | 225 | 18.1 | 170 | 16.0 |
| 200 | 19.8 | 360 | 18.1 | 95 | 17.4 | 250 | 18.2 |
| 345 | 21.5 | 105 | 18.7 | 130 | 17.9 | 355 | 16.4 |
| 245 | 16.8 | 80 | 13.1 | 330 | 17.0 | 220 | 15.4 |
| 155 | 12.6 | 275 | 20.1 | 295 | 18.7 | 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 showing weights and family incomes of 20 children 5 years of age**

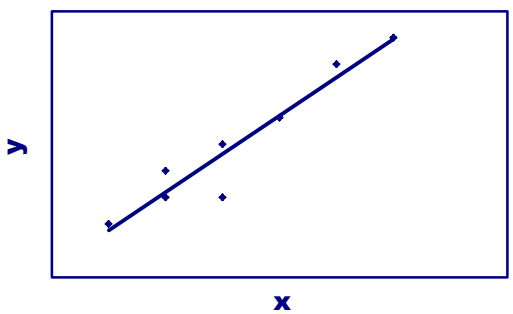


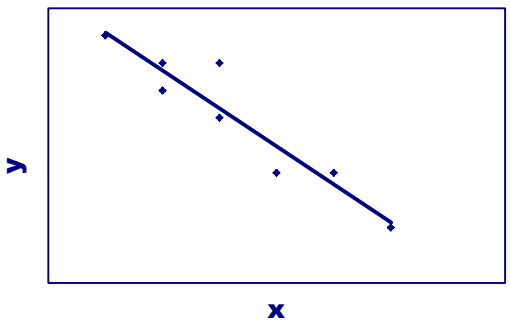

## Correlation coefficient

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

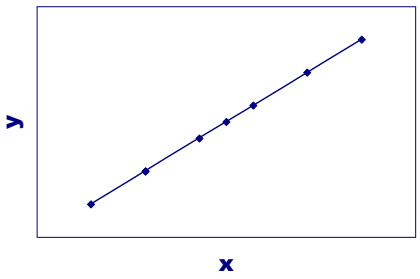
Properties of correlation coefficient

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.
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.
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.



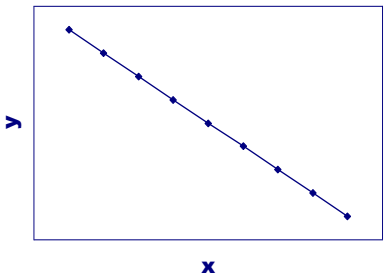



0 < (r) < (+ 1)




(r) = + 1

(- 1) < ( r) < 0




(r) = - 1

- 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 ($\rho$) = 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. ($\alpha = 0.05$)

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

- "*p*" value: $p > 0.05$ (0.07)
- Conclusion: Income and weight are not linearly correlated.

**Regression**

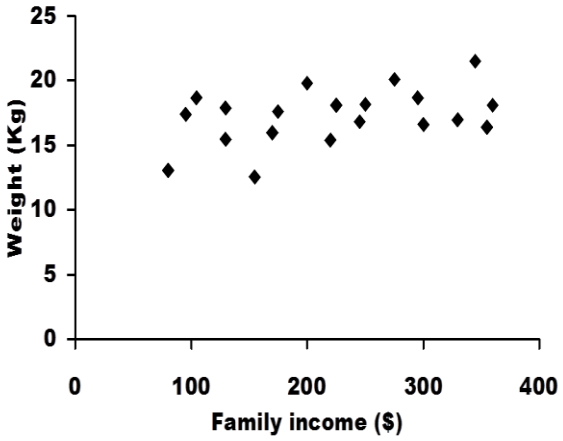
➢ Regression method of analysis is **to predict** the value of one variable (dependent) corresponding to a given value of another variable (independent)

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

- 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 showing 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:

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

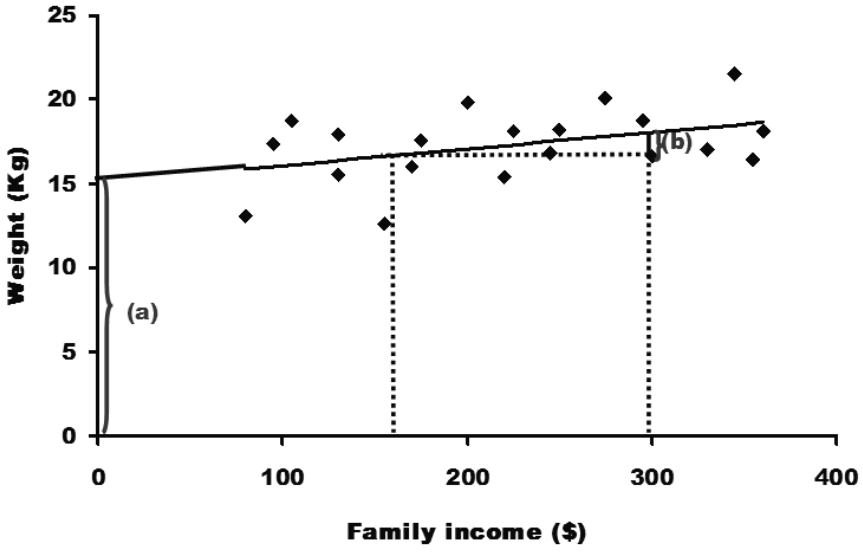
$\hat{y}$ = dependent variable

$a$ = intercept

$b$ = slope

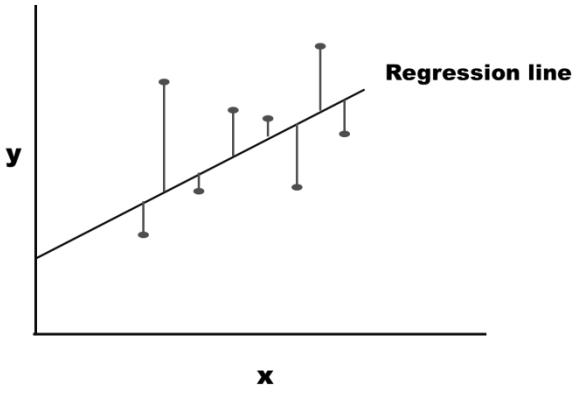
$\hat{x}$ = independent variable

**Figure showing the regression line for weights and family incomes of 20 children 5 years of age**



*** 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




Calculation of regression coefficient by LEAST SQUARES METHOD

$$y = a + bx$$

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

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

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

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

- 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, tell 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 outside 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.

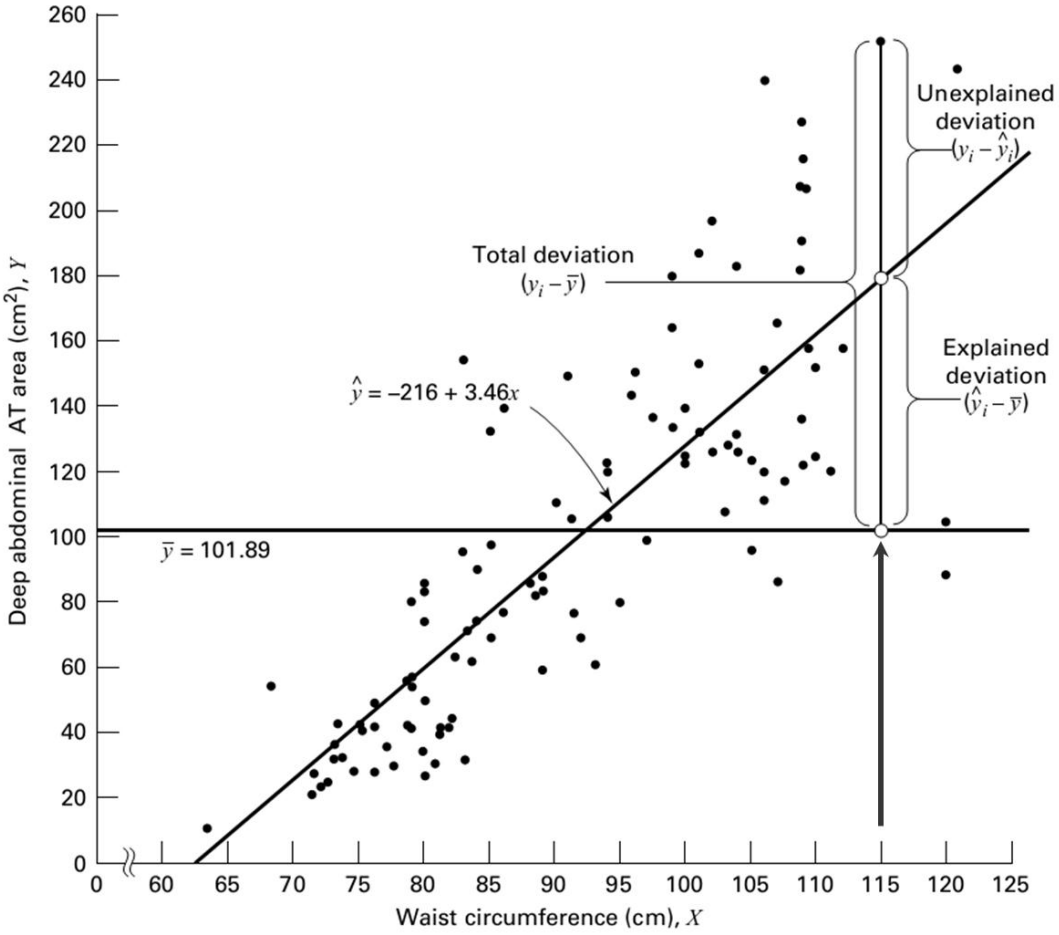



**FIGURE 9.4.4** Scatter diagram showing the total, explained, and unexplained deviations for a selected value of *Y*, Example 9.3.1.

$$\underset{\text{total deviation}}{(y_i - \bar{y})} = \underset{\text{explained deviation}}{(\hat{y}_i - \bar{y})} + \underset{\text{unexplained deviation}}{(y_i - \hat{y}_i)}$$

$$\underset{\substack{\text{total}\\\text{sum}\\\text{of squares}}}{\sum (y_i - \bar{y})^2} = \underset{\substack{\text{explained}\\\text{sum}\\\text{of squares}}}{\sum (\hat{y}_i - \bar{y})^2} + \underset{\substack{\text{unexplained}\\\text{sum}\\\text{of squares}}}{\sum (y_i - \hat{y}_i)^2}$$

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)

- Previous Example: $r^2 = 0.171 = 17\%$

**Hypothesis testing for population regression coefficient ($\beta$) = 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 ($\alpha = 0.05$)

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

- **"*p*" value**

  $p > 0.05$ (0.07)

- **Conclusion**

  Income and weight are not linearly correlated

**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

Regression equation

- $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 \ldots + \beta_kX_k$

  ($\beta_0$ is the Y intercept; $\beta_1$ is the slope for independent variable $X_1$)

- 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
- ✓ $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 \ldots + \beta_kX_k$

**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**

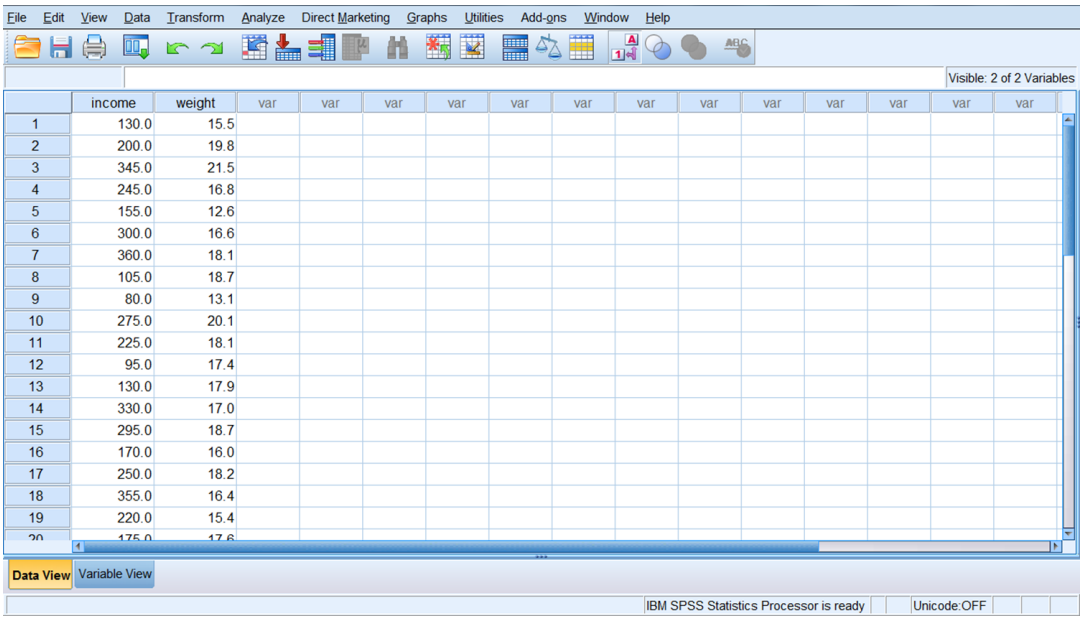
- ✓ Forward stepwise begins with forward selection followed by a test for backward elimination
- ✓ Backward stepwise begins with backward elimination followed by a test for forward selection
- ✓ deciding if predictors should be added or deleted from a regression model on the basis of a fixed decision rule

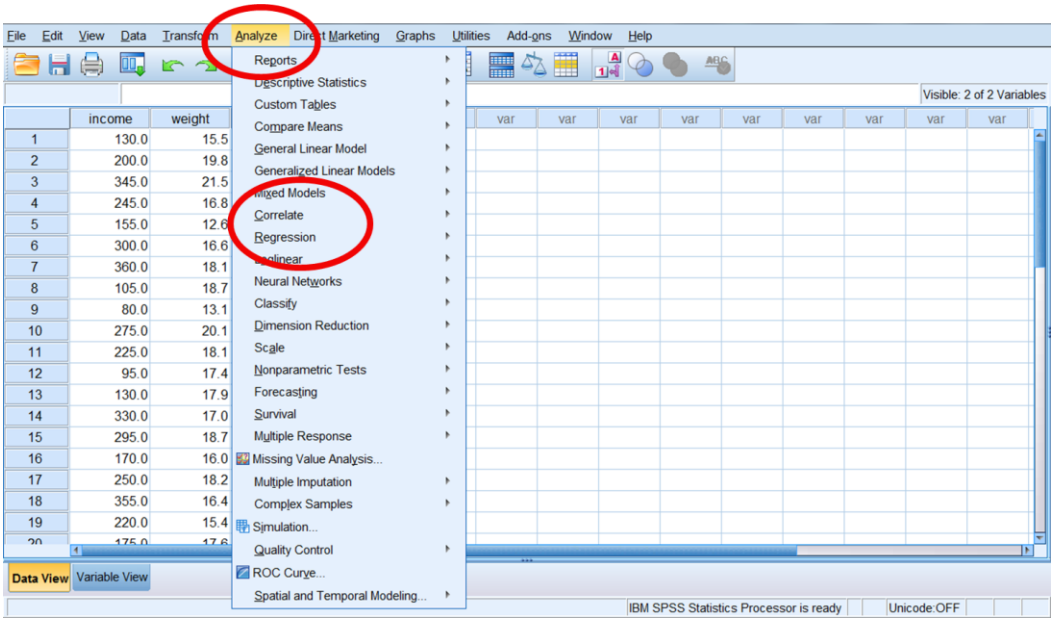
**Data management**

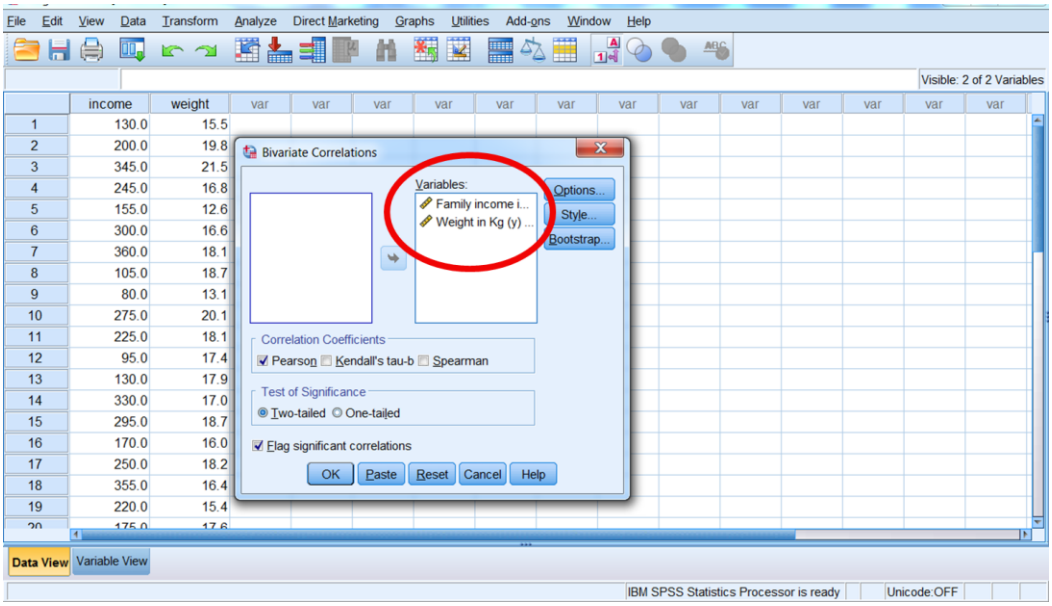
Raw data

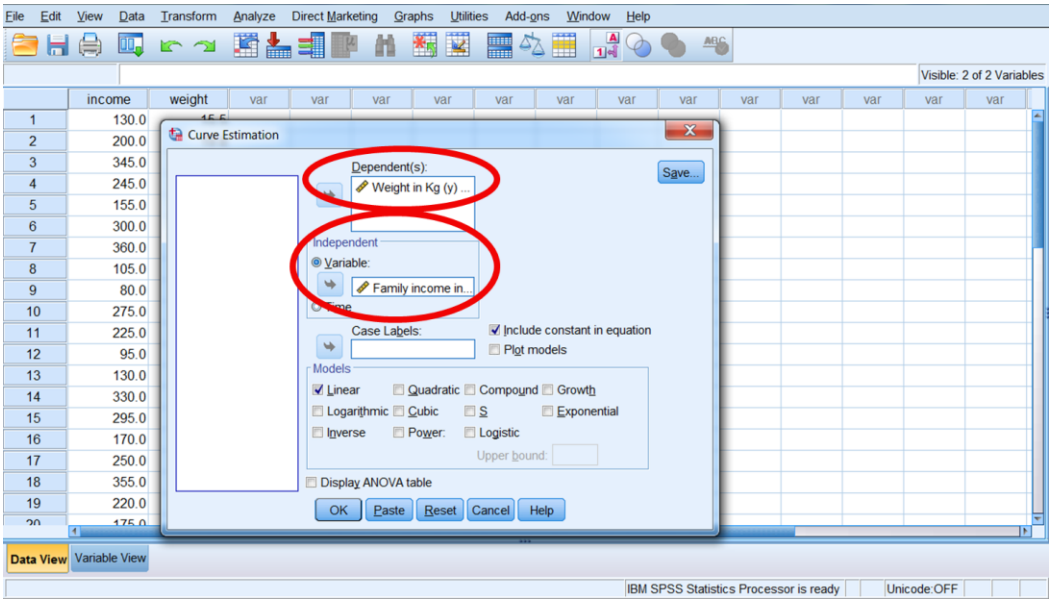
| Family income in $/ Year (x) | Weight in Kg (y) | Family income in $/ Year (x) | Weight in Kg (y) | Family income in $/ Year (x) | Weight in Kg (y) | Family income in $/ Year (x) | Weight in Kg (y) |
|---|---|---|---|---|---|---|---|
| 130 | 15.5 | 300 | 16.6 | 225 | 18.1 | 170 | 16.0 |
| 200 | 19.8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

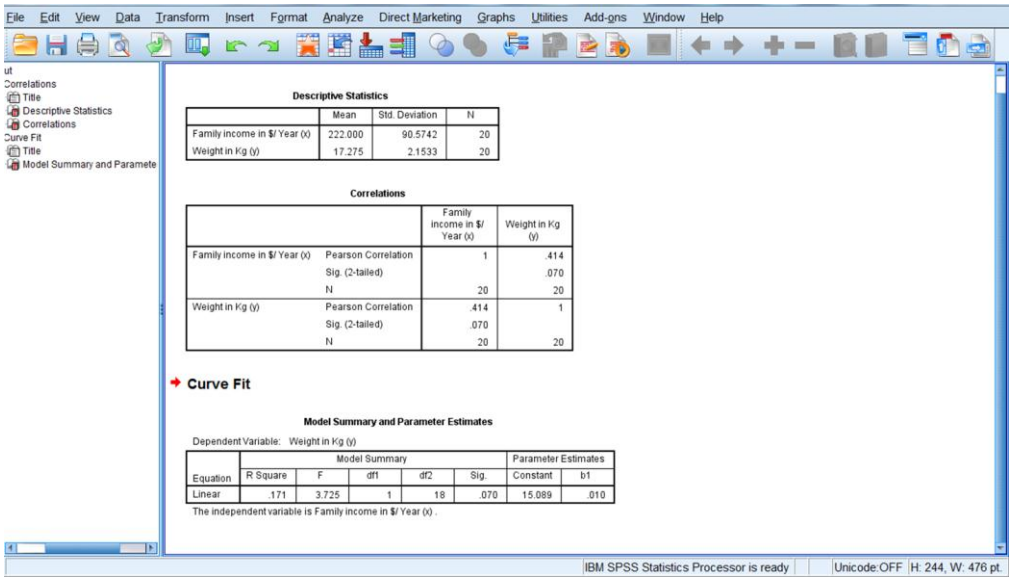
Example: **SPSS**

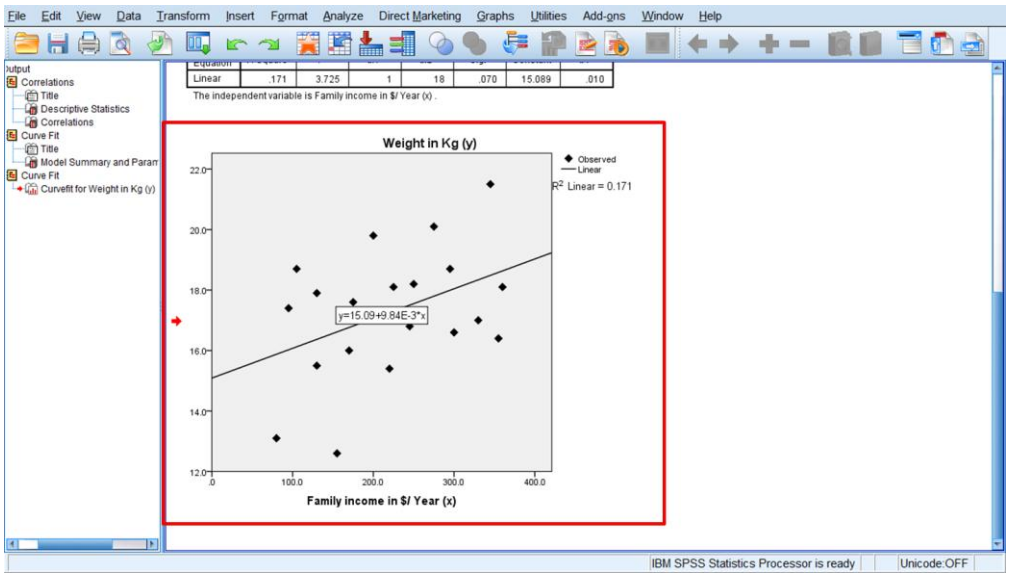



File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Custom Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric 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
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:OFF


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


Curve Estimation
Dependent(s):
Weight in Kg (y) ...
Save...
Independent
Variable:
Family income in...
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 presentation

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