# Correlation and Regression Analysis

Prof. Muhammad Abu-Salih

# Seeing Relationship and Predictive Analysis for Research

- Contents
- Correlation Analysis
- Regression Analysis for Postgraduate Research
- 
- Simple Regression Analysis
- 
- Multiple Regression Analysis
- Writing/Interpretation of Analysis
- Self Reflection

# Correlation Analysis

- Correlation deals with pairs of observations which are values of a bivariate variable. Also it deals with n-tuples which are values of n-dimensional variables.
- It is a statistical method which determines the relationship between variables, and measures the strength of that relationship.

# Correlation Cont'd

There are several types of correlation, parametric and non-parametric. The mostly used are:

Bivariate

Partial

Part

We limit our discussion to two types of bivariate correlation coefficients: Pearson or moments

Correlation coefficient which is used when data on both components of the variable are quantitative (interval or ratio), the second is Spearman's correlation coefficient which is used when data are ranked data (ordinal).

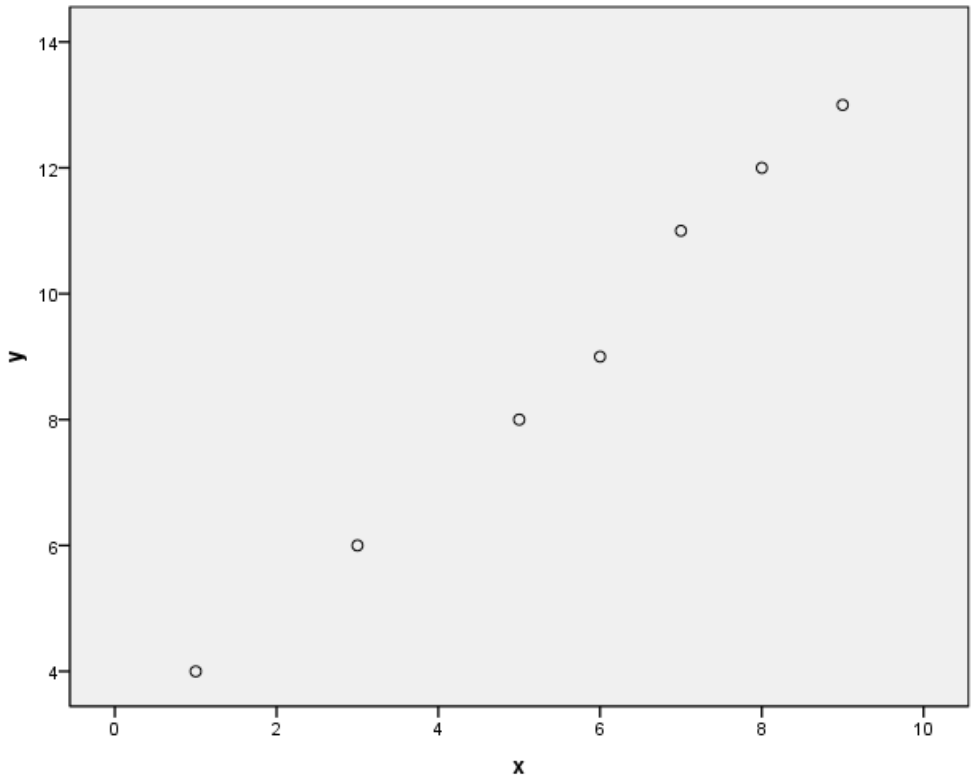
- In order to see what kind of relation exists between two variables we do the following:
- 1. Draw scatter diagram of Y on X. If the points of the scatter diagram can be confined between two parallel straight lines which are close to each other, we say there is a linear relationship between X and Y.

- If the direction of the scatter diagram is upward, then the relationship is direct, the slope is positive, and there is positive correlation between X and Y. If the direction is downward , then the relationship is negative, the slope is negative, whenever X increases Y decreases.

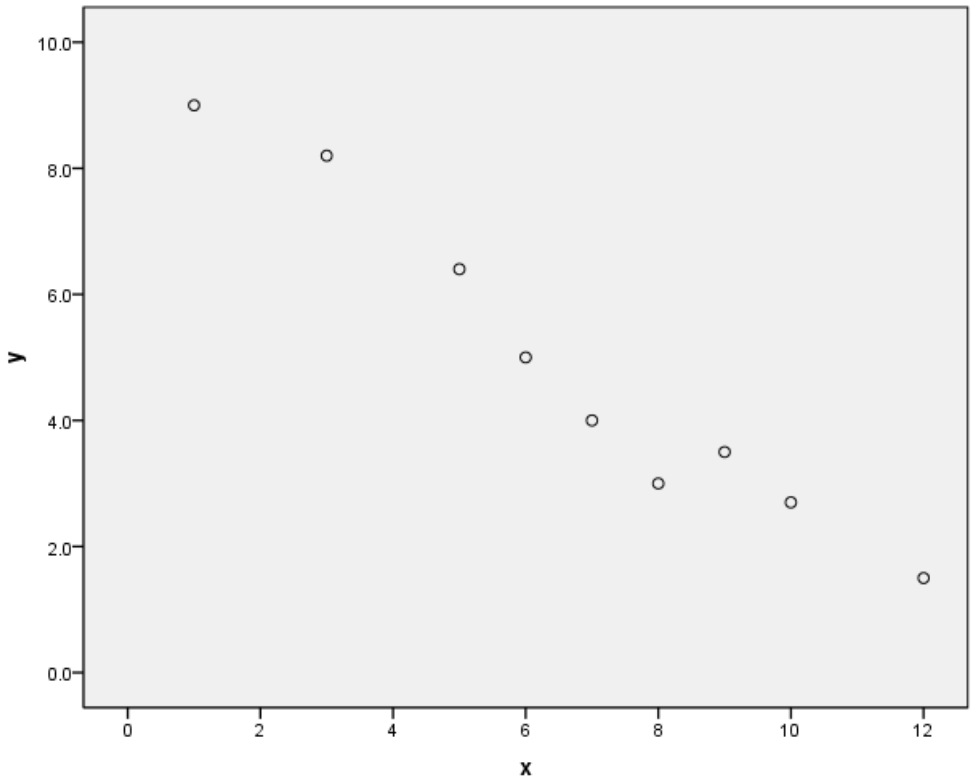
- If the points of the scatter diagram do not lie within two parallel lines, then the relationship between X and Y is not linear.
- The strength and direction of the relationship between two variables X and Y is measured by Pearson correlation or Spearman correlation coefficients.

# Examples of different relationships

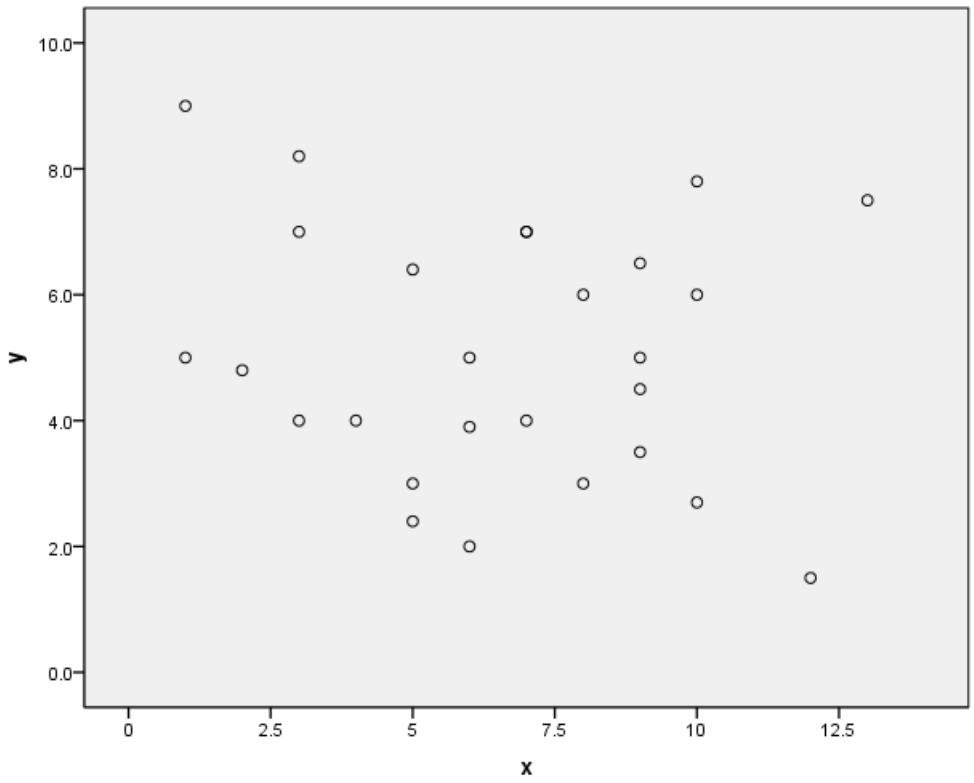
- 1. positive linear correlation


14
12
10
8
6
4
y
0
2
4
6
8
10
x

- 2. Negative Relationship

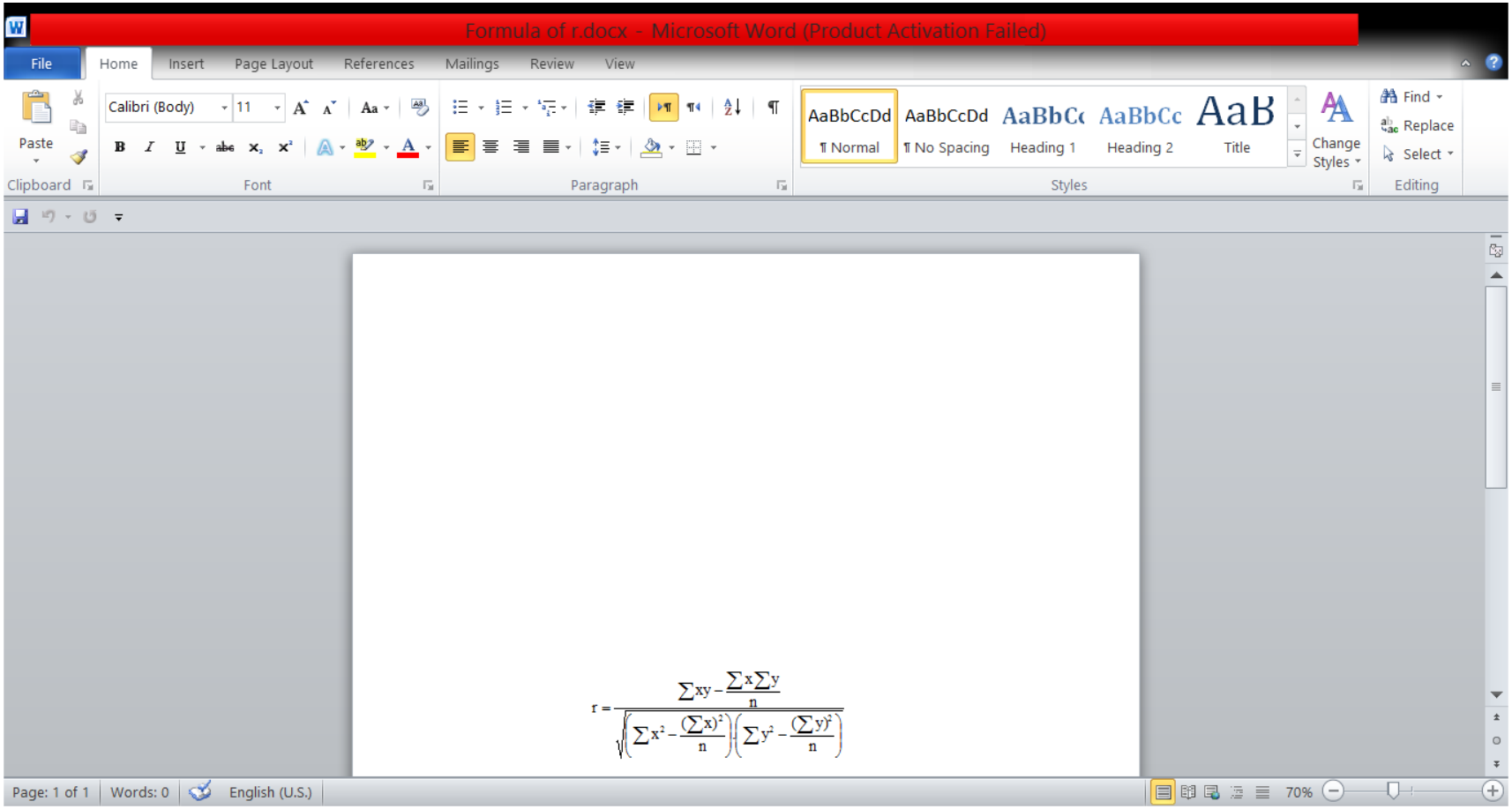

10.0
8.0
6.0
4.0
2.0
0.0
y
0
2
4
6
8
10
12
x

No linear relationship


10.0
8.0
6.0
4.0
2.0
0.0
y
0
2.5
5
7.5
10
12.5
x

# Pearson Correlation Coefficient

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

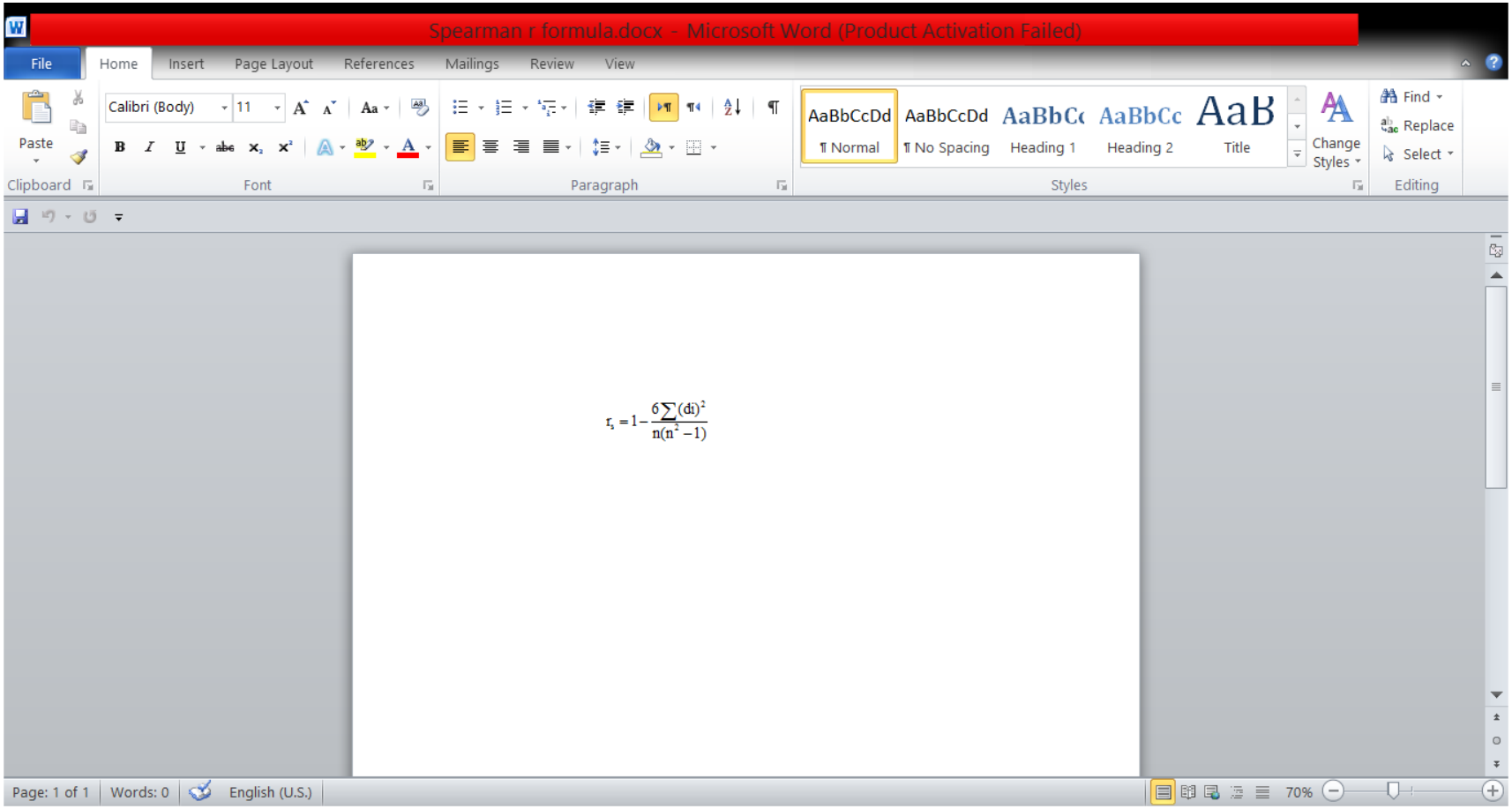

Formula of r.docx - Microsoft Word (Product Activation Failed)
File
Home
Insert
Page Layout
References
Mailings
Review
View
Paste
Clipboard
Calibri (Body)
11
Font
Paragraph
AaBbCcDd
¶ Normal
AaBbCcDd
¶ No Spacing
AaBbC(
Heading 1
AaBbCc
Heading 2
AaB
Title
Change Styles
Styles
Find
Replace
Select
Editing
$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$
Page: 1 of 1
Words: 0
English (U.S.)
70%

# Spearman Correlation Coefficient

- **It is a non-parametric measure of correlation.**
- **This procedure makes use of the two sets of ranks that may be assigned to the sample values of x and Y.**
- **Spearman Rank correlation coefficient could be computed in the following cases:**

*Both variables are quantitative.

- Both variables are qualitative ordinal.
- One variable is quantitative and the other is qualitative ordinal.


Spearman r formula.docx - Microsoft Word (Product Activation Failed)
File
Home
Insert
Page Layout
References
Mailings
Review
View
Paste
Clipboard
Calibri (Body)
11
Font
Paragraph
AaBbCcDd ¶ Normal
AaBbCcDd ¶ No Spacing
AaBbC( Heading 1
AaBbCc Heading 2
AaB Title
Change Styles
Styles
Find
Replace
Select
Editing
$r_s = 1 - \frac{6\sum(di)^2}{n(n^2-1)}$
Page: 1 of 1
Words: 0
English (U.S.)
70%

The computation of r using a calculator is very easy, but we use EXCEL or SPSS.

The properties of r

1. $-1 \leq r \leq 1$
2. Sign of r indicates the type of relationship, if r is positive ,relationship is direct, and if negative it is indirect.
3. Absolute value of r gives the strength of the relationship.
4. r does not mean causality
5. Using SPSS for correlation and its analysis will be covered with discussion of regression.

# Regression Analysis

- Regression is an important statistical technique used with the objective of exploring the relationship between a dependent variable (DV) usually called ‘response’ and one or more independent variables (IV’s), usually called predictors or explanatory variables.

# Linear Regression

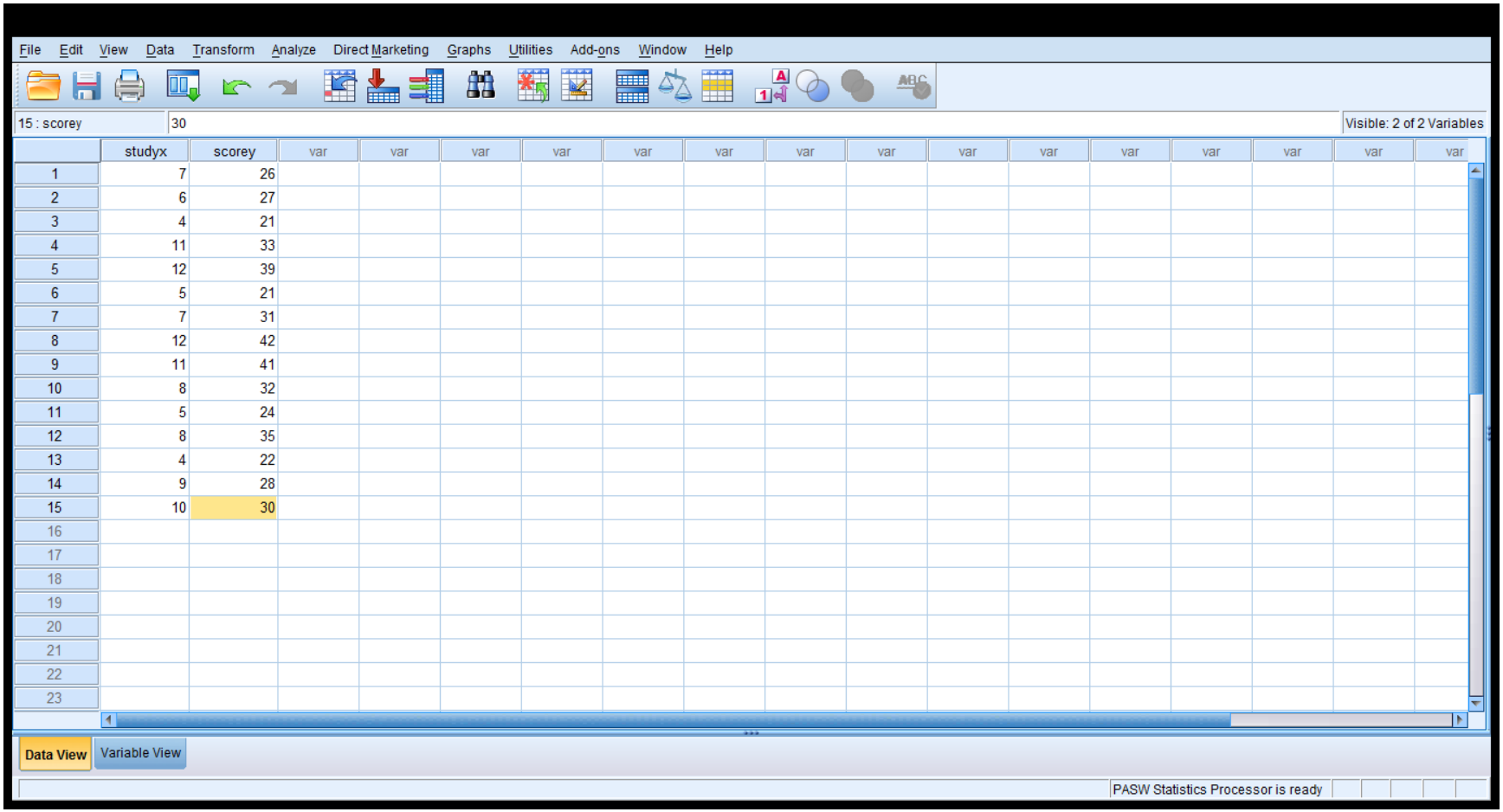
- Linear regression describes a linear relationship between a DV and IV's. It measures a causal relationship between these variables. The relationship will describe the response DV by a linear combination of the predictors.

# Simple Linear Regression (SLR)

- If the number of variables is two, one predictor (IV) and one response (DV) then we will be discussing simple linear regression. The data in such a case consist of a set of paired observations (Xi,Yi) , where Xi is the ith observation of IV and Yi is the ith observation of DV corresponding to Xi.
- The following table shows the number of hours of study X, before the exam, and the score , Y, corresponding to it, for 15 students chosen randomly

# Table 1:study hours X, and score Y

| studyX | scoreY |
|---|---|
| 7 | 26 |
| 6 | 27 |
| 4 | 21 |
| 11 | 33 |
| 12 | 39 |
| 5 | 21 |
| 7 | 31 |
| 12 | 42 |
| 11 | 41 |
| 8 | 32 |
| 5 | 24 |
| 8 | 35 |
| 4 | 22 |
| 9 | 28 |
| 10 | 30 |

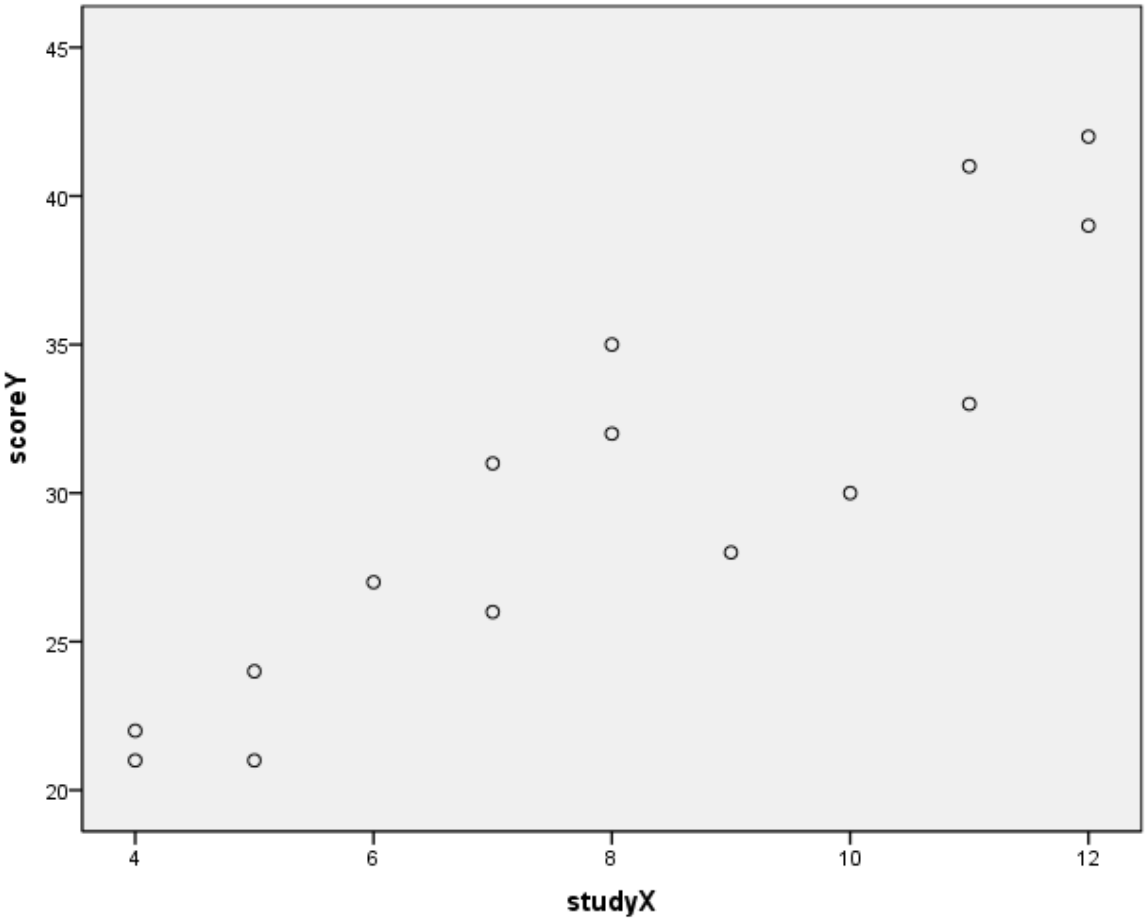

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
15 : scorey
30
Visible: 2 of 2 Variables
studyx scorey var var var var var var var var var var var var var var var
1 7 26
2 6 27
3 4 21
4 11 33
5 12 39
6 5 21
7 7 31
8 12 42
9 11 41
10 8 32
11 5 24
12 8 35
13 4 22
14 9 28
15 10 30
16
17
18
19
20
21
22
23
Data View
Variable View
PASW Statistics Processor is ready

- To check graphically if there is a linear relationship between X and Y we plot a scatter plot Y on X. If the scatter points lie within two parallel lines, close to each other, it is concluded that there is a linear relationship between the two variables. The direction of the parallel lines (the slope) determines if the relationship is positive, meaning, a direct relationship , or negative meaning inverse relationship.

# Scatter Diagram using SPSS

- Open SPSS. Click Type in data then OK
- Define your variables by clicking variable view
- Click Data view and enter your data
- The data set will be displayed on the screen.
- Click Graph and choose Legacy Dialogs then Scatter/Dot
- Choose Simple Scatter, Define
- Click Y (your dependent variable) and move it to Y-axis by clicking the arrow.
- Click X and move it to X-axis.
- Click OK.
- You can clearly see the type of relationship between X and Y.

# Scatter Diagram



- It is clear from the scatter diagram that there is a linear positive relationship between the IV

(X) and the DV (Y).

The strength and direction of the relationship is determined by correlation coefficients , which are discussed in another lecture.

Linear Regression is concerned in expressing the DV in a linear equation involving the IV’s.

# Types of relations

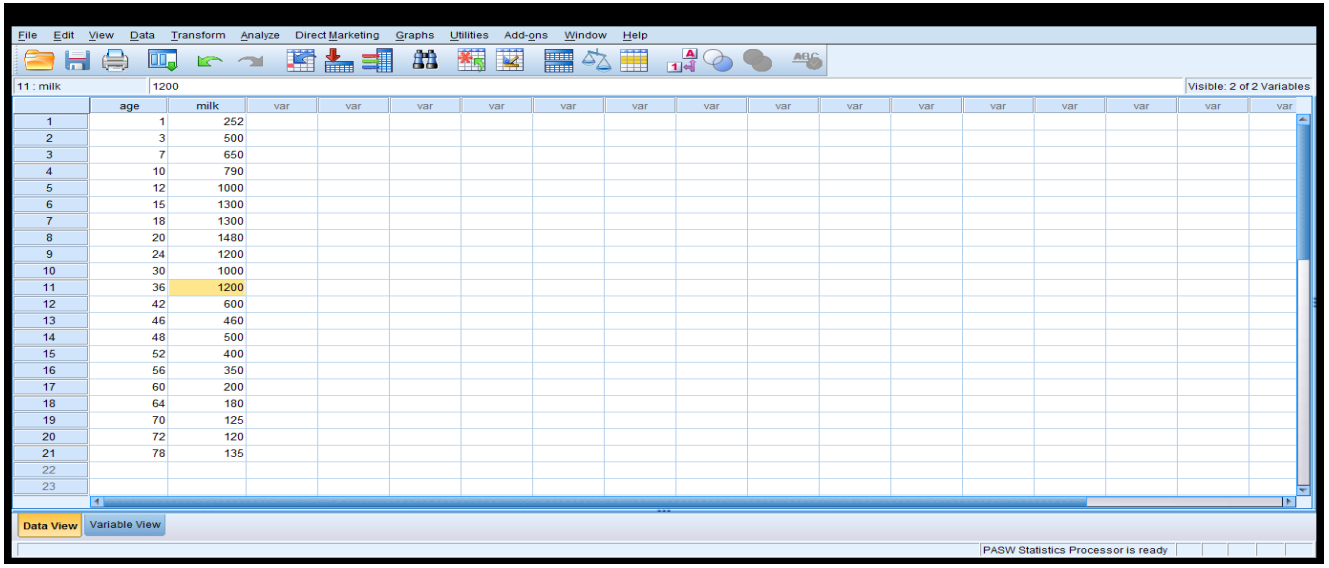
- The following are some types of relationships between DV and IV variables.
- 1.Linear
- 2.Exponential
- 3.Quadratic
- 4.Non-linear
- 5.Logistic
- 6.Others

# Example 2

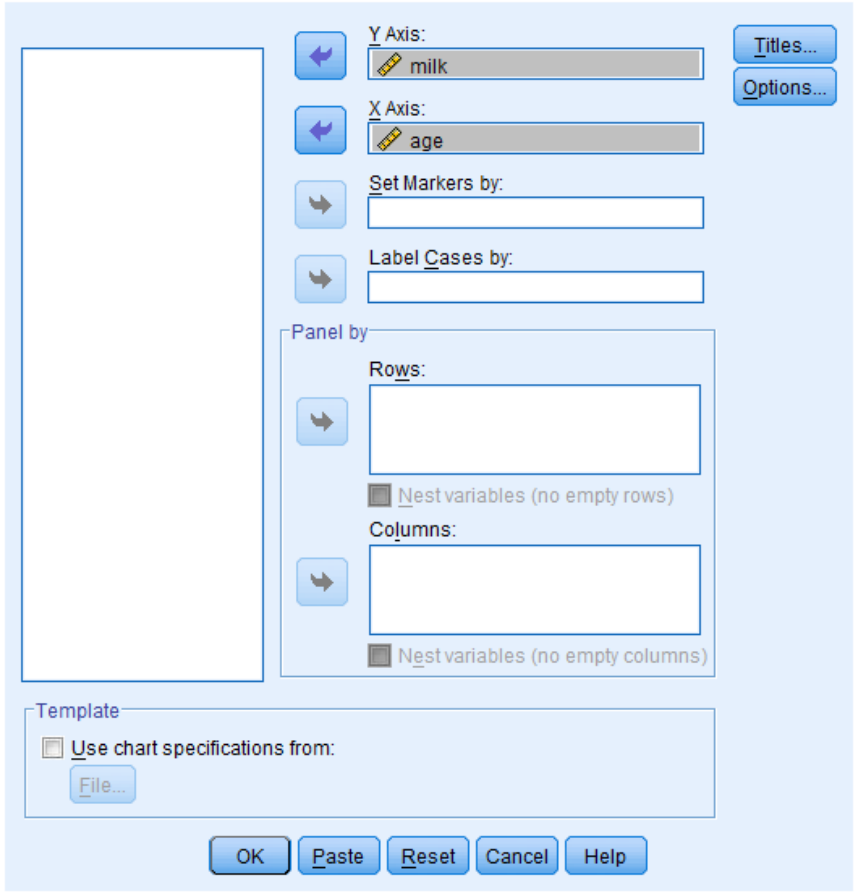
- Daily consumption Y in mg. by a sample of children of age in months X

# Example 2

- Daily consumption of milk



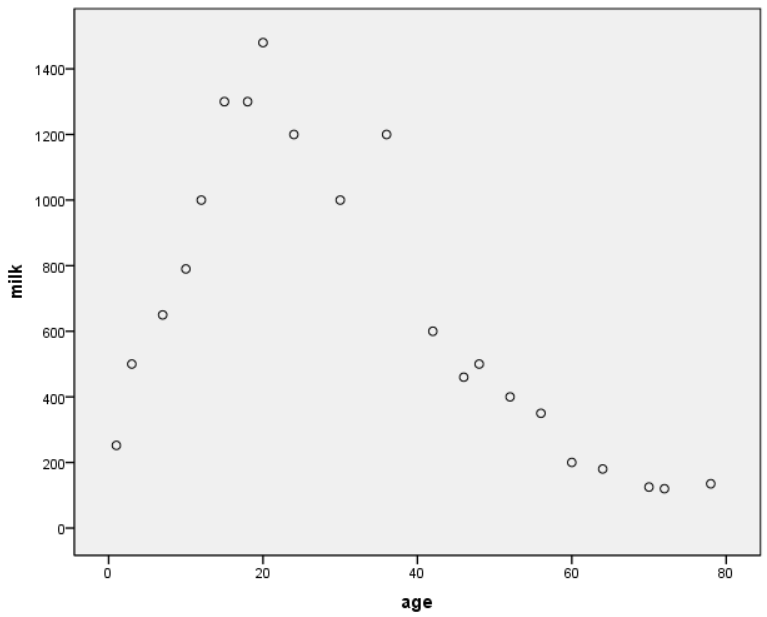

| | age | milk |
|---|---|---|
| 1 | 1 | 252 |
| 2 | 3 | 500 |
| 3 | 7 | 650 |
| 4 | 10 | 790 |
| 5 | 12 | 1000 |
| 6 | 15 | 1300 |
| 7 | 18 | 1300 |
| 8 | 20 | 1480 |
| 9 | 24 | 1200 |
| 10 | 30 | 1000 |
| 11 | 36 | 1200 |
| 12 | 42 | 600 |
| 13 | 46 | 460 |
| 14 | 48 | 500 |
| 15 | 52 | 400 |
| 16 | 56 | 350 |
| 17 | 60 | 200 |
| 18 | 64 | 180 |
| 19 | 70 | 125 |
| 20 | 72 | 120 |
| 21 | 78 | 135 |
| 22 | | |
| 23 | | |


Y Axis:
milk
X Axis:
age
Set Markers by:
Label Cases by:
Titles...
Options...
Panel by
Rows:
Nest variables (no empty rows)
Columns:
Nest variables (no empty columns)
Template
Use chart specifications from:
File...
OK
Paste
Reset
Cancel
Help

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

Output
- Log
- Graph
  - Title
  - Notes
  - Active Dataset
  - Scatter of milk age

[DataSet0]




PASW Statistics Processor is ready

- It is clear that the relationship between Y and X is not linear. A quadratic relationship may be more proper.

# Simple Linear Regression Equation

- The linear equation for regression of Y on X is assumed to be:
- $Y_i = \alpha + \beta X_i + \varepsilon_i$
- Where i is the ith observation of X and Y, and $\varepsilon_i$ is the error term.$\alpha$ is the intercept, $\beta$ is the slope of the equation.
- $\alpha$ and $\beta$ will be estimated by the least square method which fits a straight line to the scatter diagram points in such a way that the sum of squares of errors is minimized. i.e. minimize:

$Q(\alpha,\beta) = \Sigma\varepsilon*\varepsilon$

# Linear Regression Cont'd

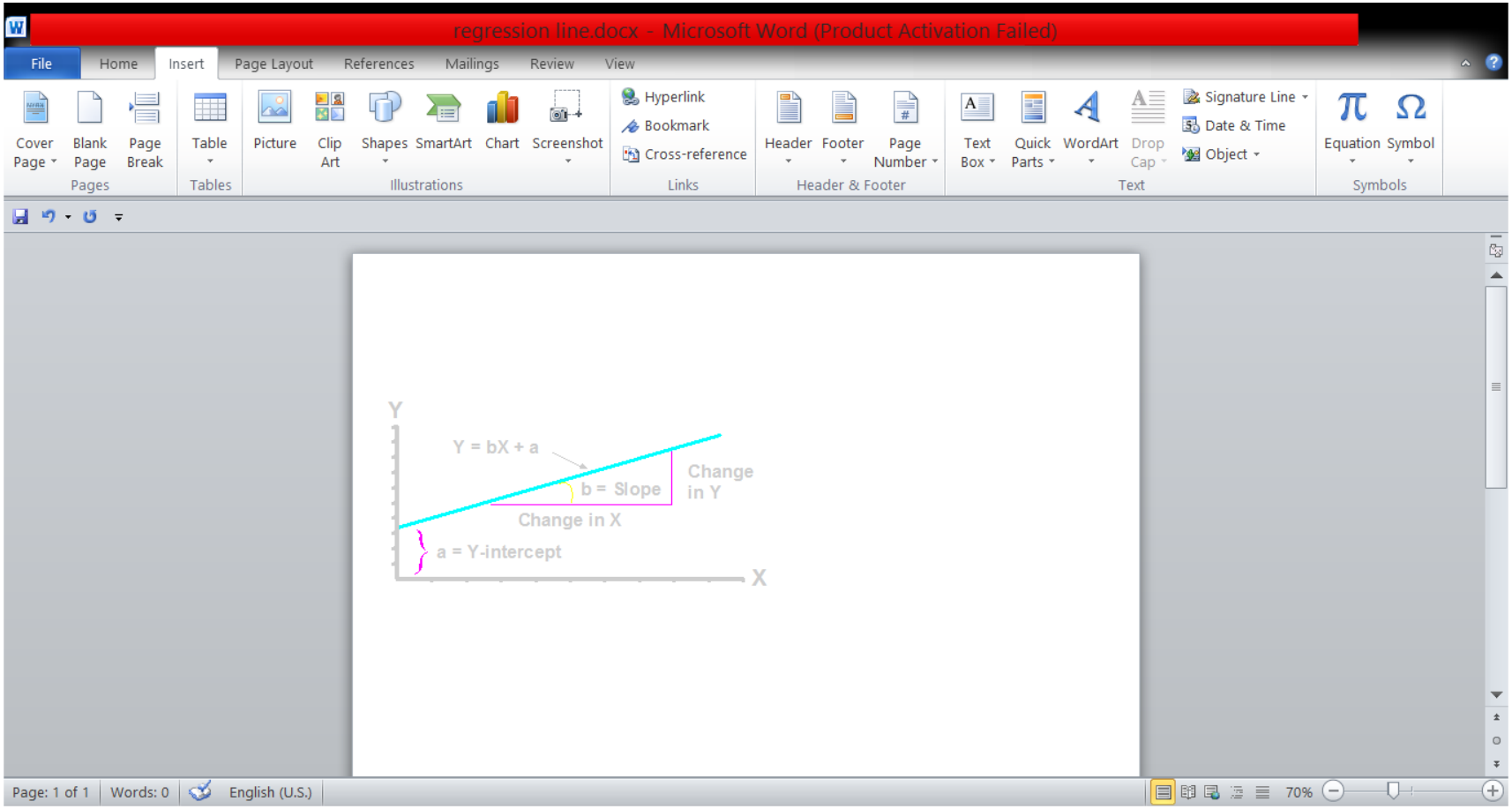
- Linear Regression Equation
  - Response = $\alpha + \beta * \text{explanatory} + \varepsilon$
  - $\alpha$ is the intercept
    - the value of the response variable when the explanatory variable is 0
  - $\beta$ is the slope
    - For each 1 unit increase in the explanatory variable, the response variable increases by $\beta$
- As mentioned earlier, $\alpha$ and $\beta$ are most often found using least squares estimation.

- We use $\Sigma$ (Y – Y hat)$^2$ …and minimize that . There is a simple, elegant formula for “discovering” the line that minimizes the sum of squared errors

$$\mathbf{b=}\ \frac{\Sigma((X - Xbar)(Y - Ybar))}{\Sigma(X - Xbar)^2}$$

- a = Ybar - bXbar  Then, regression equation is : Y hat = a + bX

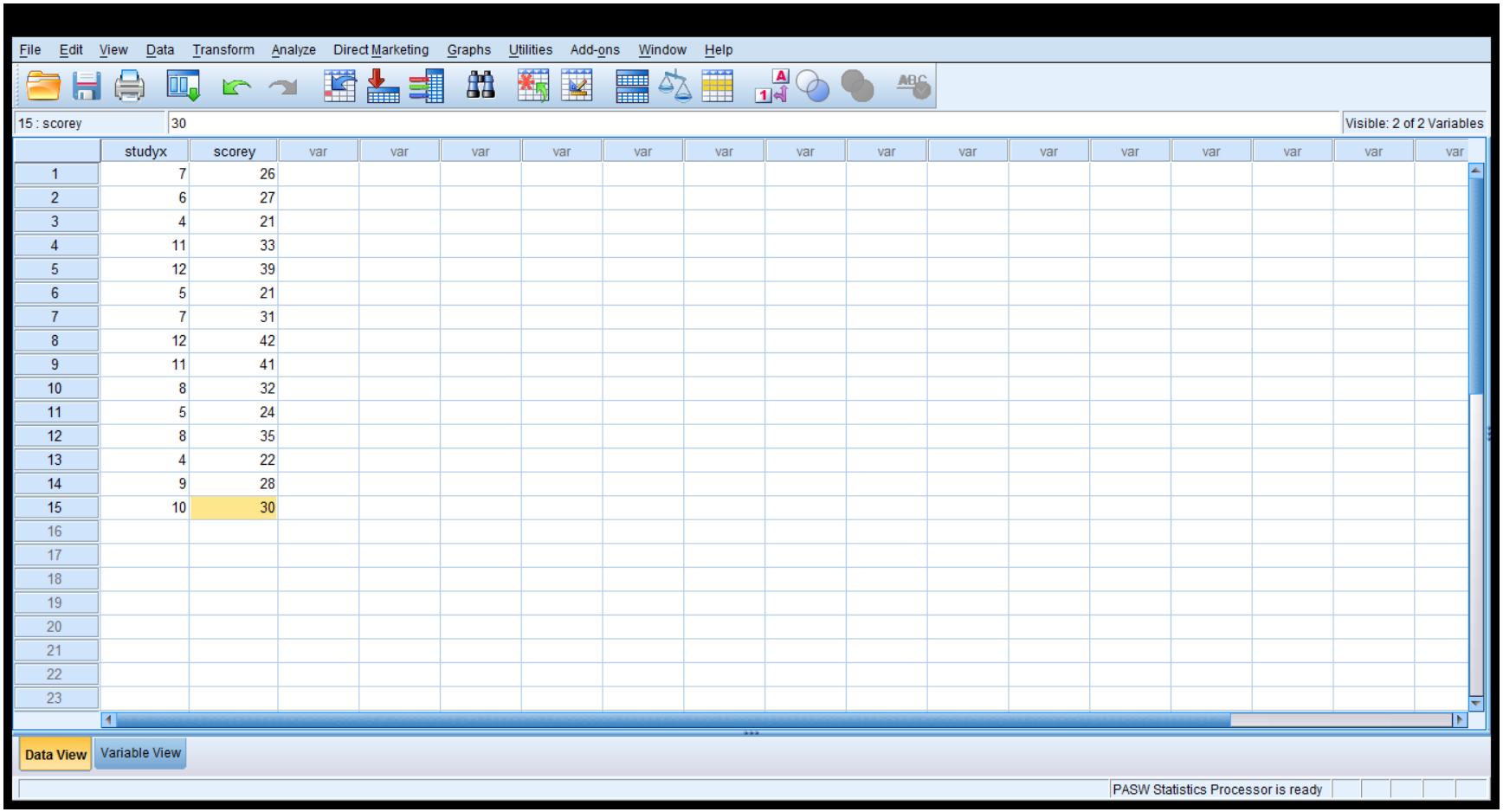
- This is the method of least squares, it gives our least squares estimate and indicates why we call this technique “ordinary least squares” or OLS regression

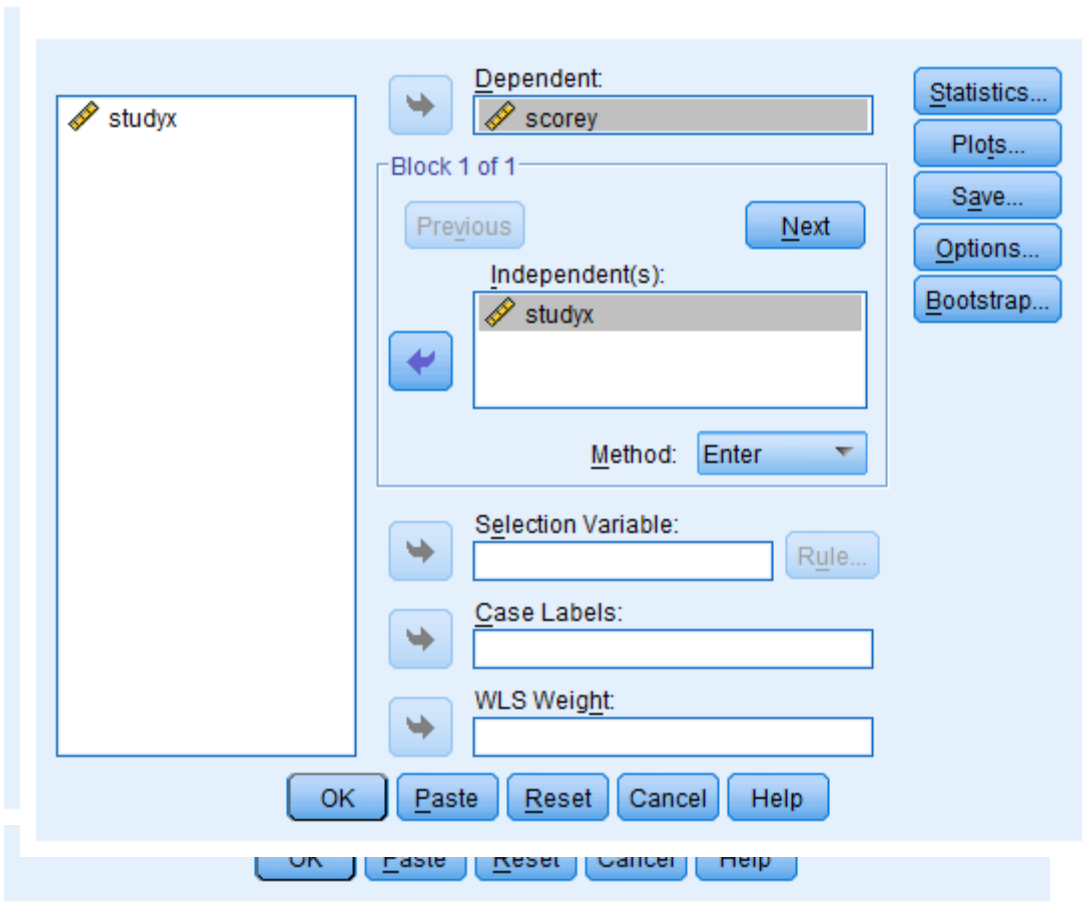

regression line.docx - Microsoft Word (Product Activation Failed)
File
Home
Insert
Page Layout
References
Mailings
Review
View
Cover Page
Blank Page
Page Break
Pages
Table
Tables
Picture
Clip Art
Shapes
SmartArt
Chart
Screenshot
Illustrations
Hyperlink
Bookmark
Cross-reference
Links
Header
Footer
Page Number
Header & Footer
Text Box
Quick Parts
WordArt
Drop Cap
Signature Line
Date & Time
Object
Text
Equation
Symbol
Symbols
Y
Y = bX + a
b = Slope
Change in Y
Change in X
a = Y-intercept
X
Page: 1 of 1
Words: 0
English (U.S.)
70%

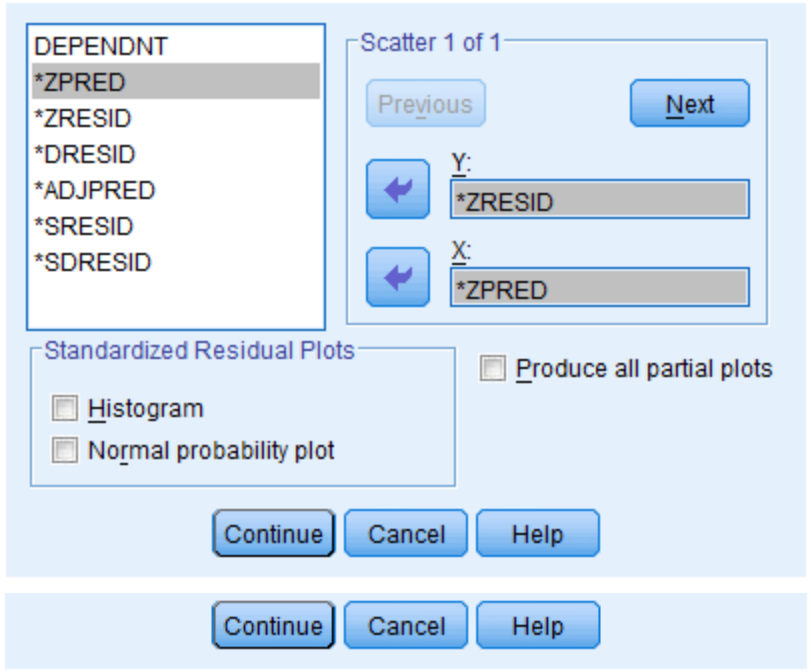
- R square
  - Is the improvement obtained by using X (and drawing a line through the conditional means) in getting as near as possible to everybody's value for Y over just using the mean for Y alone.
    - Falls between 0 and 1
    - Of 1 means an exact fit (and there is no variation of scores around the regression line)
    - Of 0 means no relationship (and as much scatter as in the original Y variable and a flat regression line through the mean of Y)
  - Would be the same for X regressed on Y as for Y regressed on X
    - Can be interpreted as the percentage of variability in Y that is explained by X.

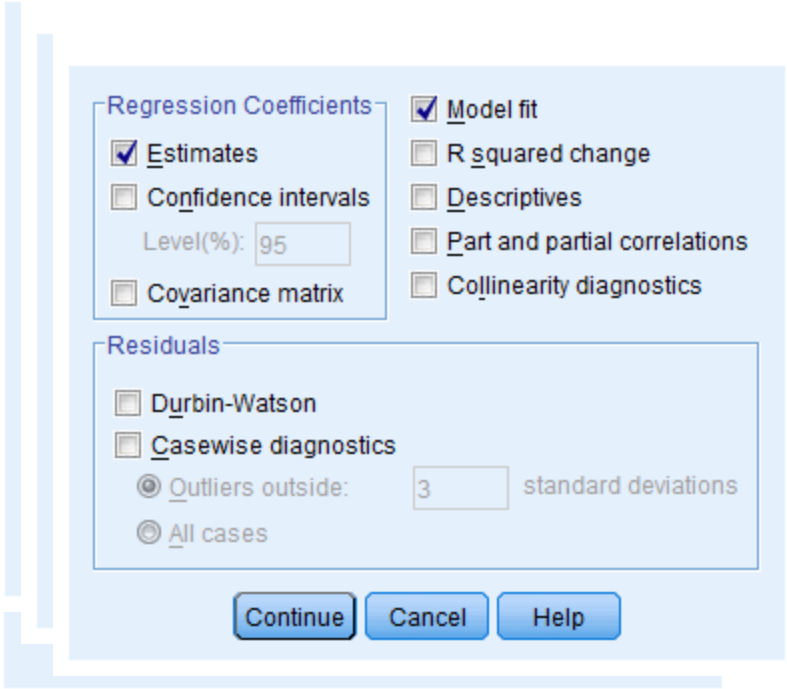
# Example 1 Continued

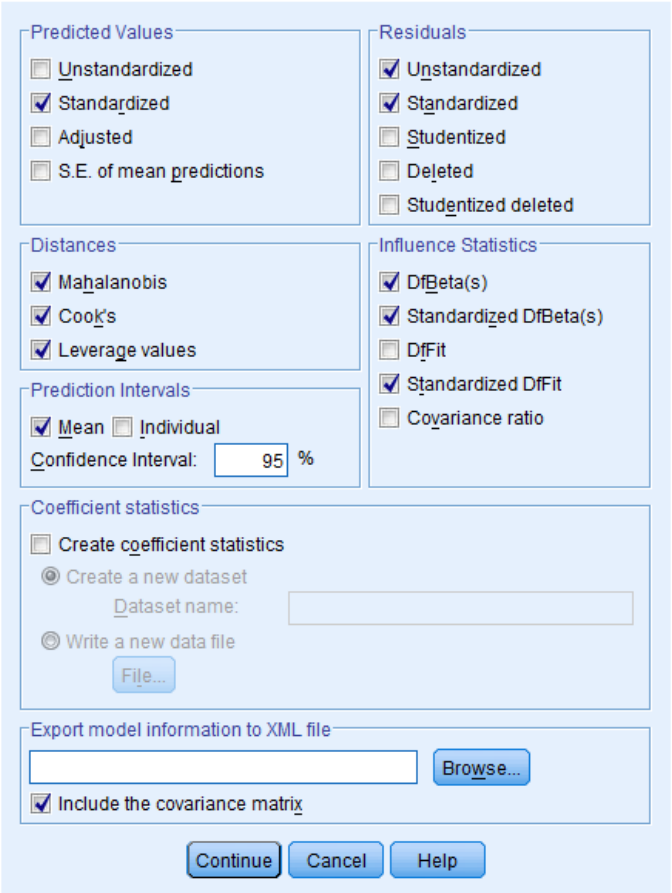
- Use SPSS to make regression analysis for example 1
- Open SPSS >existing data >OK
- Click Analyze> regression >linear> transfer Y under dependent variable and X under independent variables
- Click options and choose the items you need>Continue
- Click Save > choose items you need > Continue
  Click OK
- ALL DIALOG BOXES ARE SHOWN ON THE SCREEN DURING WORKSHOP.

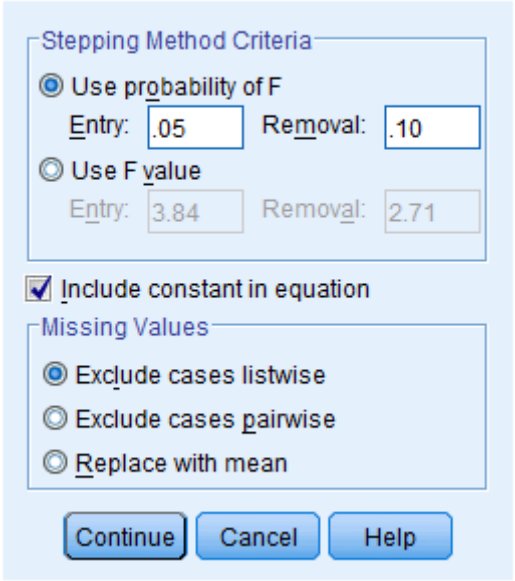

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
15 : scorey
30
Visible: 2 of 2 Variables
studyx scorey var var var var var var var var var var var var var var var
1 7 26
2 6 27
3 4 21
4 11 33
5 12 39
6 5 21
7 7 31
8 12 42
9 11 41
10 8 32
11 5 24
12 8 35
13 4 22
14 9 28
15 10 30
16
17
18
19
20
21
22
23
Data View
Variable View
PASW Statistics Processor is ready

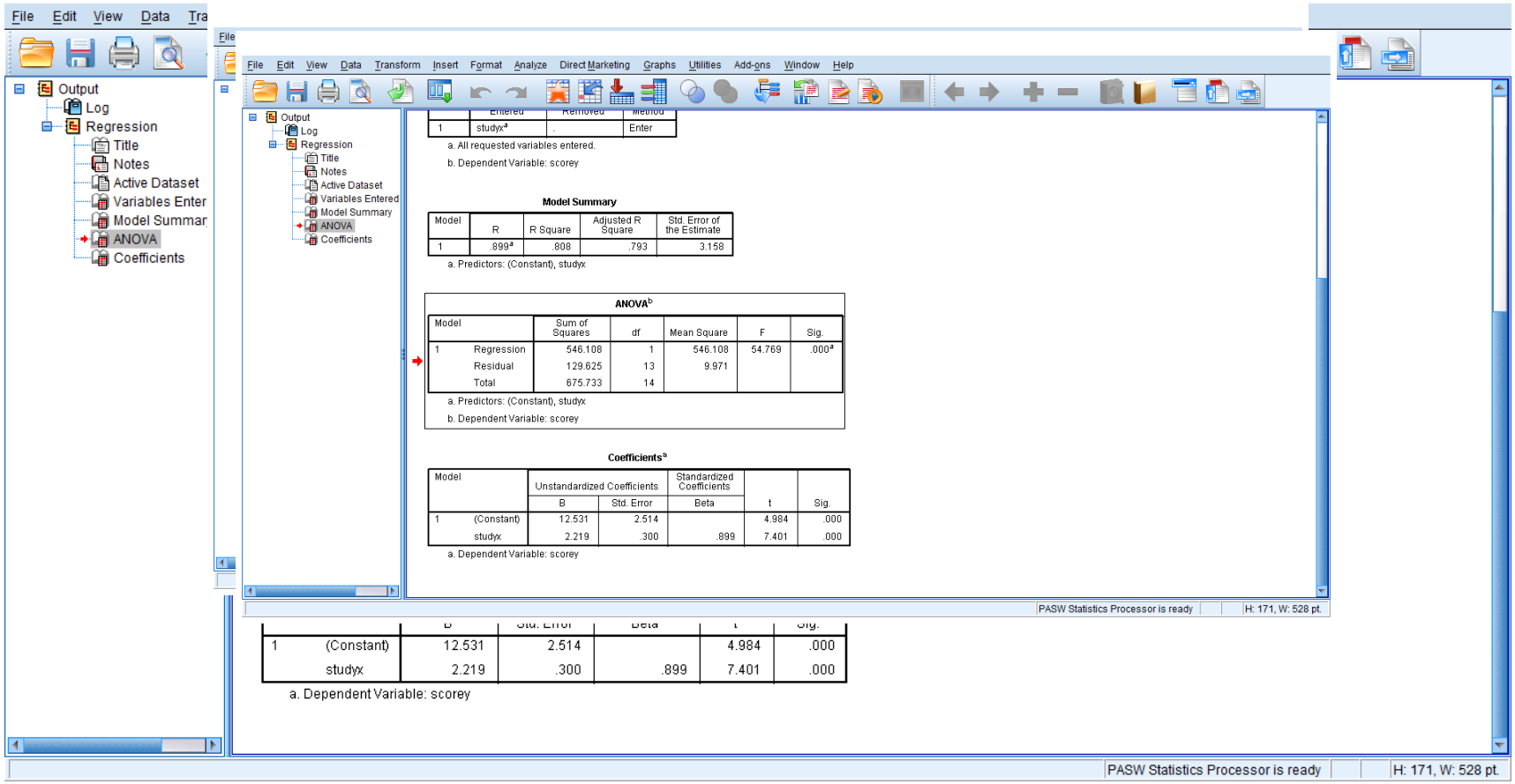

studyx
Dependent:
scorey
Block 1 of 1
Previous
Next
Independent(s):
studyx
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
OK
Paste
Reset
Cancel
Help
Statistics...
Plots...
Save...
Options...
Bootstrap...


DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
Scatter 1 of 1
Previous
Next
Y:
*ZRESID
X:
*ZPRED
Standardized Residual Plots
Histogram
Normal probability plot
Produce all partial plots
Continue
Cancel
Help
Continue
Cancel
Help


Regression Coefficients
Estimates
Confidence intervals
Level(%): 95
Covariance matrix
Model fit
R squared change
Descriptives
Part and partial correlations
Collinearity diagnostics
Residuals
Durbin-Watson
Casewise diagnostics
Outliers outside: 3 standard deviations
All cases
Continue
Cancel
Help


Predicted Values
Unstandardized
Standardized
Adjusted
S.E. of mean predictions
Residuals
Unstandardized
Standardized
Studentized
Deleted
Studentized deleted
Distances
Mahalanobis
Cook's
Leverage values
Influence Statistics
DfBeta(s)
Standardized DfBeta(s)
DfFit
Standardized DfFit
Covariance ratio
Prediction Intervals
Mean
Individual
Confidence Interval:
95
%
Coefficient statistics
Create coefficient statistics
Create a new dataset
Dataset name:
Write a new data file
File...
Export model information to XML file
Browse...
Include the covariance matrix
Continue
Cancel
Help


Stepping Method Criteria
Use probability of F
Entry: .05
Removal: .10
Use F value
Entry: 3.84
Removal: 2.71
Include constant in equation
Missing Values
Exclude cases listwise
Exclude cases pairwise
Replace with mean
Continue
Cancel
Help


File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Log
Regression
Title
Notes
Active Dataset
Variables Entered
Model Summary
ANOVA
Coefficients
a. All requested variables entered.
b. Dependent Variable: scorey
Model Summary
Model | R | R Square | Adjusted R Square | Std. Error of the Estimate
1 | .899[a] | .808 | .793 | 3.158
a. Predictors: (Constant), studyx
ANOVA[b]
Model | | Sum of Squares | df | Mean Square | F | Sig.
1 | Regression | 546.108 | 1 | 546.108 | 54.769 | .000[a]
Residual | 129.625 | 13 | 9.971
Total | 675.733 | 14
a. Predictors: (Constant), studyx
b. Dependent Variable: scorey
Coefficients[a]
Model | | Unstandardized Coefficients B | Unstandardized Coefficients Std. Error | Standardized Coefficients Beta | t | Sig.
1 | (Constant) | 12.531 | 2.514 | | 4.984 | .000
studyx | 2.219 | .300 | .899 | 7.401 | .000
a. Dependent Variable: scorey
PASW Statistics Processor is ready
H: 171, W: 528 pt.

# Assumptions on Linear Regression

- **The regression model is based on the following assumptions.**
- **1-The relationship between *X* and *Y* is linear.**
- **2-The expected value of the error term is zero**
- **3-The variance of the error term is constant for all the values of the independent variable, *X*. This is the assumption of homoscedasticity.**
- **4-There is no autocorrelation in the error terms.**
- **5-The independent variable is uncorrelated with the error term.**
- **6-The error term is normally distributed.**

# These assumptions can be written as: ε's are iid Normal with mean zero and fixed variance )

- If the assumptions are not met, the estimates of, α, β and their standard deviations, and estimates of $R^2$ will be incorrect

- Maybe possible to do transformations to either the explanatory or response variable to make the relationship linear

# Correlation Coefficient and R Square

- Correlation indicates the strength and direction of the linear relationship between two quantitative variables
  - Values between -1 and +1
- $R^2$ is the fraction of the variability in the response that can be explained by the linear relationship with the explanatory variable
  - Values between 0 and +1
- $\text{Correlation}^2 = R^2$

# Simple Linear Regression Analysis

- WE will analyze the example mentioned earlier:
- Check for normality *P-P Plot
- Kolmogorov –Smirnov Test
- Interpretation of results
- Histogram of residuals with Normal Curve
- Plots: Standardized Residuals vs Standardized predicted values
- Individual predicted values.

| One-Sample Kolmogorov-Smirnov Test | | |
|---|---|---|
| | | y |
| N | | 22 |
| Normal Parameters[a,b] | Mean | 75.27 |
| | Std. Deviation | 12.166 |
| Most Extreme Differences | Absolute | .115 |
| | Positive | .077 |
| | Negative | -.115 |
| Kolmogorov-Smirnov Z | | .538 |
| Asymp. Sig. (2-tailed) | | .934 |
| a. Test distribution is Normal. | | |
| b. Calculated from data. | | |

# K-S Test

K-S test is used to test the null hypothesis:

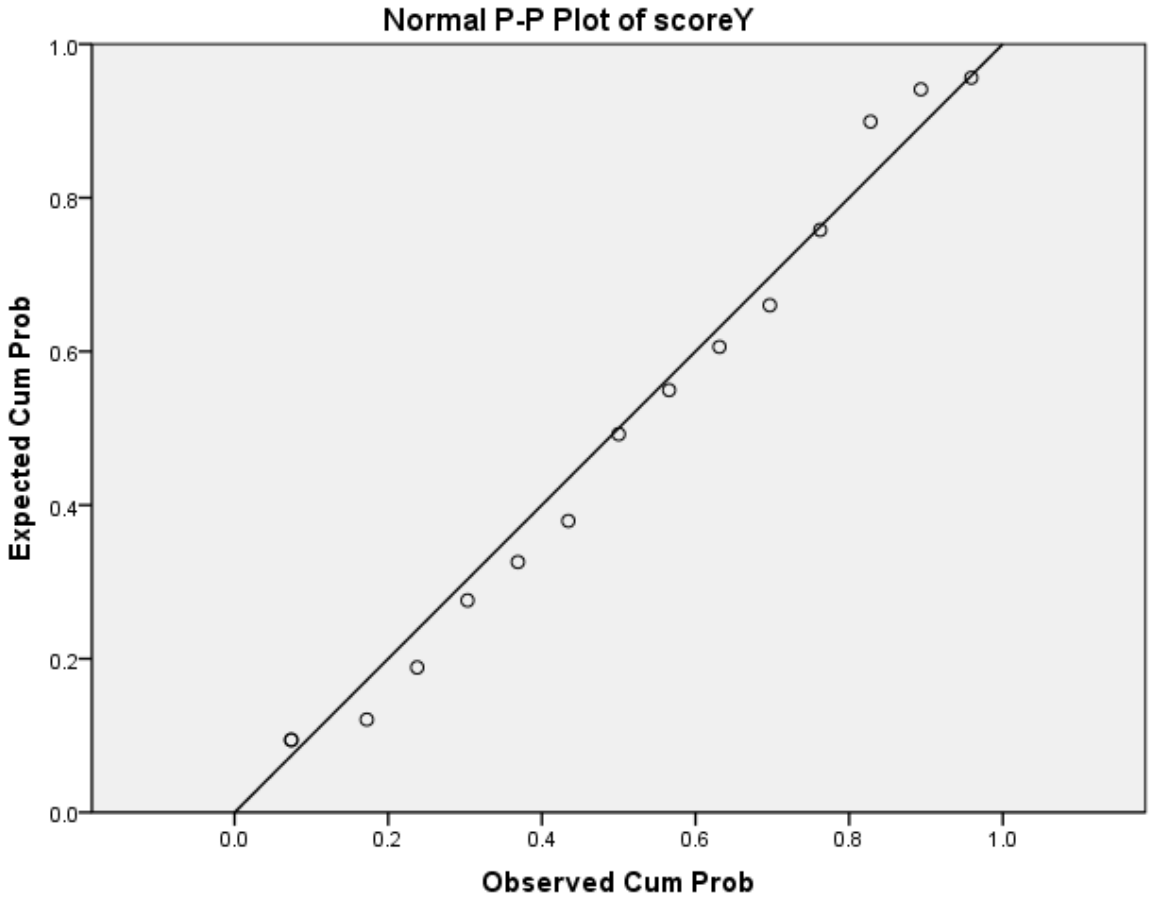
Ho : The sample is drawn from population with Normal distribution.

As seen from output, value of test statistic is K-S Z =0.538,whose absolute value is $< 1.96$ , the z-tabulated value.

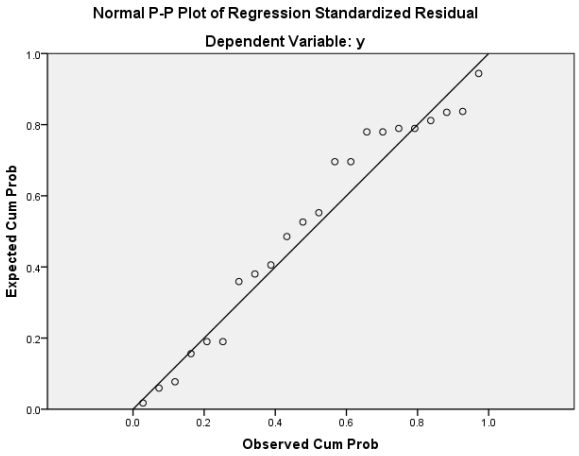
Also significance is 0.934 which is greater than 0.05

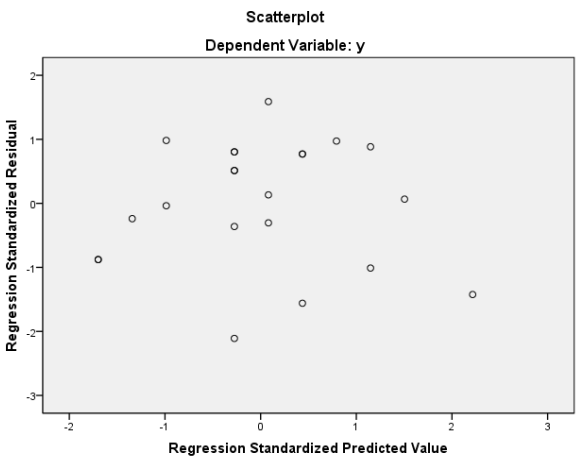
This means that data do not give significant evidence to reject Ho.

# Normal P –P Plot of score Y



- The P-P plot shows that points  lie on a straight line or very close to it. This indicates that Y-values come from normal distribution.


Normal P-P Plot of Regression Standardized Residual
Dependent Variable: y
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0


Scatterplot
Dependent Variable: y
Regression Standardized Residual
Regression Standardized Predicted Value
-3
-2
-1
0
1
2
3

- Again ,the plot of regression standardized residuals against  standardized predicted values does not show a pattern of residuals.
- The points are almost equally scattered around the horizontal line . This indicates that values are random and residuals have fixed variance.

- **Model Summary[b]**
- R=0.899
- R square=0.808
- Adjusted R square=0.793
- R square 0.808 indicates that the model is adequate, the time of study is a good predictor of the score of student on the exam. It indicates that about 81% of the total variance is accounted for (explained) by the model, where X is the predictor of Y.

# ANOVA for Regression

| Residuals Statistics[a] | | | | | |
|---|---|---|---|---|---|
| | Minimum | Maximum | Mean | Std. Deviation | N |
| Predicted Value | 58.03 | 97.78 | 75.27 | 10.156 | 22 |
| Std. Predicted Value | -1.698 | 2.216 | .000 | 1.000 | 22 |
| Standard Error of Predicted Value | 1.469 | 3.628 | 1.981 | .616 | 22 |
| Adjusted Predicted Value | 59.37 | 101.56 | 75.50 | 10.377 | 22 |
| Residual | -14.480 | 10.906 | .000 | 6.699 | 22 |
| Std. Residual | -2.109 | 1.589 | .000 | .976 | 22 |
| Stud. Residual | -2.163 | 1.626 | -.015 | 1.028 | 22 |
| Deleted Residual | -15.227 | 11.429 | -.229 | 7.465 | 22 |
| Stud. Deleted Residual | -2.409 | 1.702 | -.032 | 1.069 | 22 |
| Mahal. Distance | .007 | 4.910 | .955 | 1.291 | 22 |
| Cook's Distance | .000 | .545 | .060 | .115 | 22 |
| Centered Leverage Value | .000 | .234 | .045 | .061 | 22 |

# Multiple Regression

- Multiple regression is an extension of simple regression , whereby there are several IV's (predictors), say p instead of one.
- The goal of MR is to predict a response (DV) based on several predictors. In Multiple Linear Regression, the end result is to develop a linear equation of response Y on a linear combination of several X's (IV's).

# Multiple Linear Regression

- Use more than one explanatory variable to explain the variability in the response variable

- Regression Equation

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \quad -$$

- $\beta_j$ is the change in the response variable (Y) when $X_j$ increases by 1 unit and all the other explanatory variables remain fixed

# Signs of Mullticolinearity

Signs of multicollinearity include:

- 1) none of the t-ratios of the coefficients are statistically significant, but the F-test for the equation as a whole is significant;
- 2) adding an additional independent variable to the equation radically changes either the size or the sign (plus or minus) of the coefficients associated with the other independent variables
- If multicollinearity is discovered, the researcher may drop one of the two variables that are highly correlated, or simply leave them in and note that multicollinearity is present.

# EXAMPLE

- For explanation, consider example 11.3.1 p.478 on job performance, Daniel (1995) , whereby ,it is intended to develop a model for job performance (JOBPER) of nurses based on their personal characteristics consisting of six (IV's).
- X1=assertiveness (ASRV)
- X2=enthusiasm (ENTH)

# Example (Cont'd)

- X3=ambition (AMB)
- X4=communication skills (COMM)
- X5=problem solving skills (PROB)
- X6= initiative(INIT)

SPSS will be used to build the required model through several steps:

# Step One

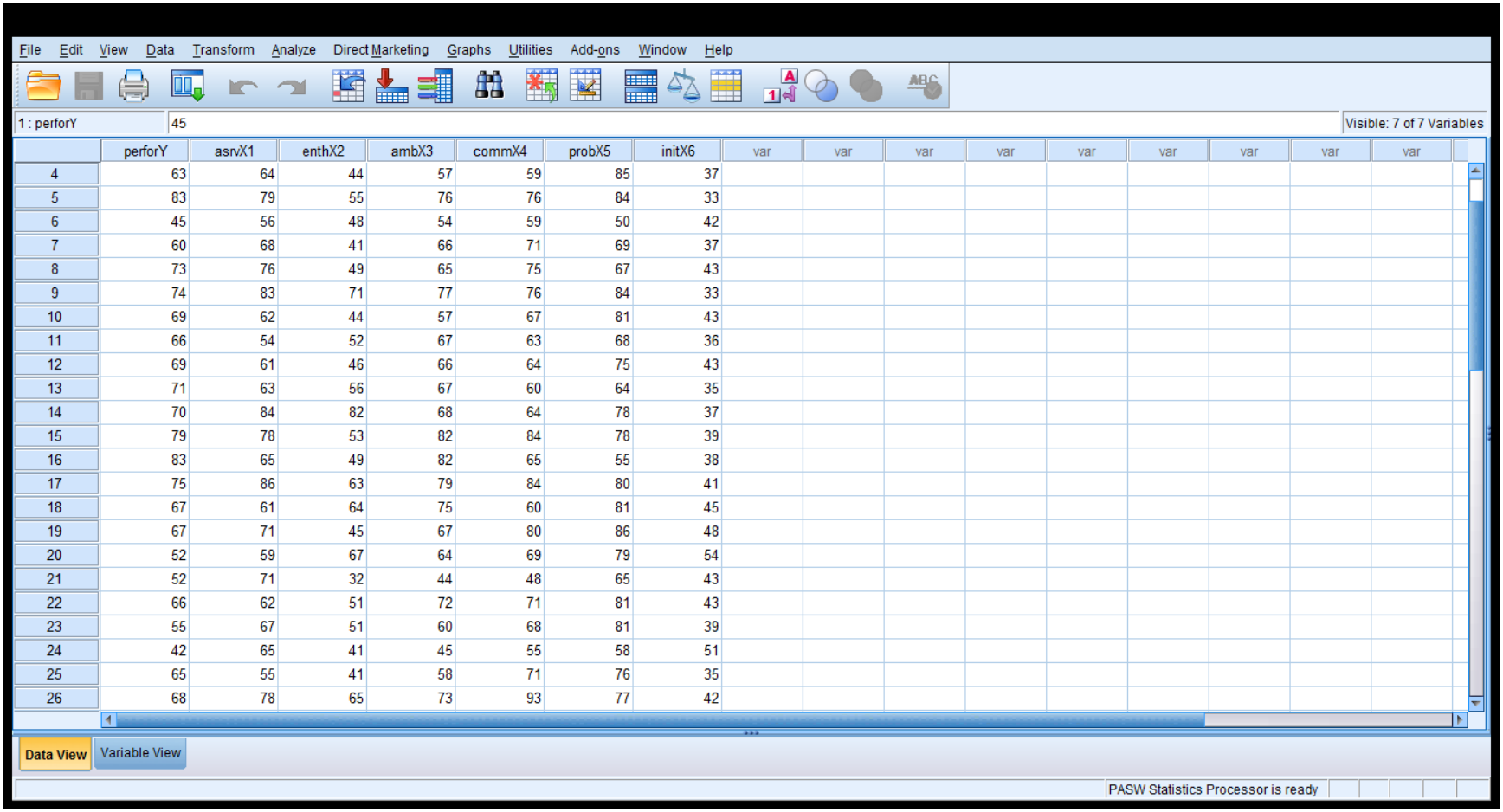
- 1.Open SPSS from short cut on desktop, or :
- Click start> Programs>SPSS for windows> SPSS 18 for windows>Type in data>OK.

2. A Data Editor will appear. Click Variable View. A new screen will appear with the following words:

Name Type Width Decimals Label values Missing columns Align Measure

# Step one (cont’d)

- Be sure your variables type is numeric, and determine number of decimals needed.
- 3.Click Data View and enter your data. The result will be as follows:




| | perforY | asrvX1 | enthX2 | ambX3 | commX4 | probX5 | initX6 | var | var | var | var | var | var | var | var | var |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 63 | 64 | 44 | 57 | 59 | 85 | 37 | | | | | | | | | |
| 5 | 83 | 79 | 55 | 76 | 76 | 84 | 33 | | | | | | | | | |
| 6 | 45 | 56 | 48 | 54 | 59 | 50 | 42 | | | | | | | | | |
| 7 | 60 | 68 | 41 | 66 | 71 | 69 | 37 | | | | | | | | | |
| 8 | 73 | 76 | 49 | 65 | 75 | 67 | 43 | | | | | | | | | |
| 9 | 74 | 83 | 71 | 77 | 76 | 84 | 33 | | | | | | | | | |
| 10 | 69 | 62 | 44 | 57 | 67 | 81 | 43 | | | | | | | | | |
| 11 | 66 | 54 | 52 | 67 | 63 | 68 | 36 | | | | | | | | | |
| 12 | 69 | 61 | 46 | 66 | 64 | 75 | 43 | | | | | | | | | |
| 13 | 71 | 63 | 56 | 67 | 60 | 64 | 35 | | | | | | | | | |
| 14 | 70 | 84 | 82 | 68 | 64 | 78 | 37 | | | | | | | | | |
| 15 | 79 | 78 | 53 | 82 | 84 | 78 | 39 | | | | | | | | | |
| 16 | 83 | 65 | 49 | 82 | 65 | 55 | 38 | | | | | | | | | |
| 17 | 75 | 86 | 63 | 79 | 84 | 80 | 41 | | | | | | | | | |
| 18 | 67 | 61 | 64 | 75 | 60 | 81 | 45 | | | | | | | | | |
| 19 | 67 | 71 | 45 | 67 | 80 | 86 | 48 | | | | | | | | | |
| 20 | 52 | 59 | 67 | 64 | 69 | 79 | 54 | | | | | | | | | |
| 21 | 52 | 71 | 32 | 44 | 48 | 65 | 43 | | | | | | | | | |
| 22 | 66 | 62 | 51 | 72 | 71 | 81 | 43 | | | | | | | | | |
| 23 | 55 | 67 | 51 | 60 | 68 | 81 | 39 | | | | | | | | | |
| 24 | 42 | 65 | 41 | 45 | 55 | 58 | 51 | | | | | | | | | |
| 25 | 65 | 55 | 41 | 58 | 71 | 76 | 35 | | | | | | | | | |
| 26 | 68 | 78 | 65 | 73 | 93 | 77 | 42 | | | | | | | | | |

# Using SPSS

- WE use the same steps used in Example 1, predicting score Y from study hours X

# Step Two

To begin analysis, it is necessary to check the following: •

- 1.Scatter Plots:

Click Graphs> Legacy Dialogs >Scatter/Dot >Matrix Scatter > Define > transfer all variables by clicking on each variable then on arrow

2.Normality of errors

This is done by several methods:

a. Kolmogorv-Smirnov Test

Click Analyze> nonparametric > one sample tests >Legacy Dialogs>1-sample (K-S)>move Y to the dialog box> normal under test distribution>OK.

- From the output check if Sig. is > chosen level of significance, usually $\alpha=0.05$.If so, then errors are normally distributed. If Sig <0.05, reject normality. See result

# Normal Plot

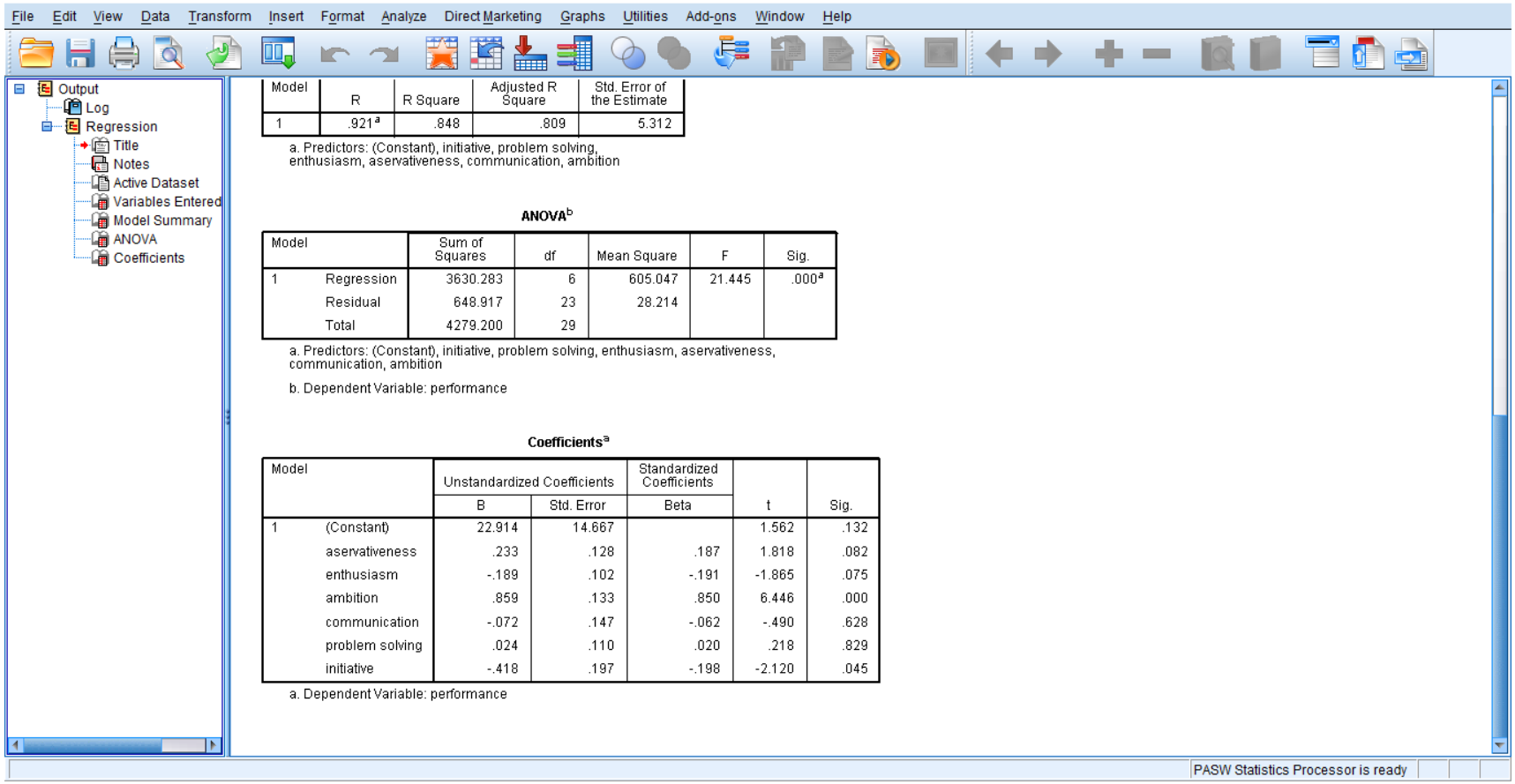
- b. Normal Plot
- To use normal plot for checking normality, do the following Click Analyze > Descriptive Statistics > P-P plots
- Move DV (perforY) to Variables box
- Be sure that Normal is in the box under Test Distribution
- Click OK.
- You get graph of: Expected Cum Prob  vs Observed Cum Prob
- If the points lie on a straight line, then conclude normality of y observations implying normality of error terms.

# Standard Regression

- Standard regression would be used to answer the following questions:
- A. What is the size of the overall relationship between the predicted variable and all of the predictor variables.
- B. How much each of the predictors uniquely contribute to that relationship. In standard regression all the predictor variables are entered at the same time. This means doing regression using all variables by choosing ENTER.
- IN our example: Do the following:
- Click Analyze > Regression >Linear >Move DV(perforY ) under Dependent > Move all the six X's under Independent(s)
- Use Method: Enter
- Click Statistics and from dialog box, choose: Estimates, Model fit, R squared change, Partial correlations, Collinearity diagnostics, then click Continue

# Standard Regression Cont'd

- Click Plots and from linear regression: Plots box ,choose ZRESID and transfer it  under Y, and ZPRED under X.
- Choose Normal probability plot , and click Continue.
- Click OK.




| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .921[a] | .848 | .809 | 5.312 |

a. Predictors: (Constant), initiative, problem solving, enthusiasm, aserativeness, communication, ambition

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 3630.283 | 6 | 605.047 | 21.445 | .000[a] |
| | Residual | 648.917 | 23 | 28.214 | | |
| | Total | 4279.200 | 29 | | | |

a. Predictors: (Constant), initiative, problem solving, enthusiasm, aserativeness, communication, ambition

b. Dependent Variable: performance

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 22.914 | 14.667 | | 1.562 | .132 |
| | aserativeness | .233 | .128 | .187 | 1.818 | .082 |
| | enthusiasm | -.189 | .102 | -.191 | -1.865 | .075 |
| | ambition | .859 | .133 | .850 | 6.446 | .000 |
| | communication | -.072 | .147 | -.062 | -.490 | .628 |
| | problem solving | .024 | .110 | .020 | .218 | .829 |
| | initiative | -.418 | .197 | -.198 | -2.120 | .045 |

a. Dependent Variable: performance

- INTERPRET RESULTS AS YOU DID IN EXAMPLE 1
- FOR EXAMPLE:
- R=0.91, R square=0.848
- This means that model is good , 84.8% of the total Variance is explained by the six IV's.
- F value is 21.445> f()= and sig.is0.000< level of significance 0.05
- This means that there is significant effect of the IV's on the DV.

- The table: Coefficients gives the estimated coefficients (the slopes) of the individual IV's , and whether they are significant or not.
- Also it gives the standardized values (betas) which give the strength of each IV, for example ,the strongest effect is for ambition and it = 0.850

# Stepwise Regression

- In stepwise regression , not all the predictors may end up to be included in the model. Predictor variables are entered into the regression equation one at a time based upon statistical criteria.
- Stepwise regression answers a different question, namely, what is the best combination of independent (predictor) variables would predict the dependent variable? Of course the answer to the two questions a and b of standard regression would be answered, but for the variables entered in the best combination.

# Stepwise Regression Cont'd

- At each step in the analysis the predictor variable which contributes the most to the prediction equation is entered first. The criteria used is the value of the multiple correlation, R. The process is continued only if additional variables add statistically. The process is stopped when no more predictors add significantly to R. SPSS requires specification of values to Pin and Pout.
- Pin= probability allowing a variable to enter.
- Pout=probability letting a variable out .

# Stepwise Linear Regression

- Begin as we did in standard MR but choose Method Stepwise. Continue as in standard Mr with an extra step: Click options > Determine Entry value (Pin) and Removal value (Pout). Removal value should be larger than Entry value.Then click :include constant in equation >Continue> OK.

# Regression Diagnostics

- Methods for identifying problems in your multiple regression analysis -- a good idea for any multiple regression analysis
- Can help identify
  - violation of assumptions
  - outliers and overly influential cases—cases you might want to delete or transform
    - important variables you've omitted from the analysis

# Interpretation of Regression Coefficients

- Regression coefficients in multiple regression (unstandardized and standardized) are considered **partial regression coefficients** because each coefficient is calculated after controlling for the other predictors in the model.
- Tests of regression coefficients represent a test of the unique contribution of that variable in predicting y over and above all other predictor variables in the model.

# Three Classes of MR Diagnostic Statistics

.1. **Distance** -- detects outliers in the dependent variable and assumption violations -- primary measure is the residual (Y-Y^) or standardized residual (i.e., put in terms of z scores) or studentized residual (i.e., put in terms of t-scores)

2. **Leverage --** identifies potential outliers in the independent variables -- primary measure is the leverage statistic or "hat" diagnostic

# Three Classes Cont'd

3. **Influence** -- combines distance and leverage to identify unusually influential observations (i.e., observations or cases that have a big influence on the MR equation is *Cook's D*

(the measure we will use

- Cook's D is a measure of how much $MS_{Residual}$ would change if a particular case were excluded from the calculations of the regression coefficients.

# Distance

- Analyze residuals
- Pay attention to standardized or studentized residuals > 2.5; shouldn't be more than 5% of cases
- Tells you which cases are not predicted well by regression analysis -- you can learn from this in itself
- Necessary to test MR assumptions
  - homoscedasticity
  - normality of errors

# Distance

- Unstandardized Residuals
  - The difference between an observed value and the value predicted by the model. The mean is 0.
- Standardized Residuals
  - The residual divided by an estimate of its standard error. Standardized residuals have a mean of 0 and a standard deviation of 1.
- Studentized Residuals
  - The residual divided by an estimate of its standard deviation that varies from case to case, depending on the leverage of each case's predictor values in determining model fit. They have a mean of 0 and a standard deviation slightly larger than 1.

# Distance

- Deleted Residuals
  - The residual for a case that is excluded from the calculation of the regression coefficients. It is the difference between the value of the dependent variable and the adjusted predicted value.
- Studentized Deleted Residuals
  - It is a studentized residual with the effect of the observation deleted from the standard error. The residual can be large due to distance, leverage, or influence. The mean is 0 and the variance is slightly greater than 1.

# Our Example

- Open your Data in SPSS
- Regress Y on the six independent variables
- Under statistics select estimates, covariance matrix, and model fit.
- Save predicted values unstandardized and save all residuals (unstandardized, standardized, Studentized, deleted, and Studentized deleted)
- Okay

# Example Cont'd

- Interpret b's and betas.  Compare betas with correlations.
  - Zero order correlations
  - Validity coefficients
- Why is the standard error of estimate different from the standard deviation of unstandardized residuals?
- Note case wise diagnostics compared to saved values.

# Influence Statistics

- Cook's D
  - A measure of how much $MS_{Residual}$ would change if a particular case were excluded from the calculations of the regression coefficients. A large Cook's D indicates that excluding a case from computation of the regression statistics changes the coefficients substantially.
- Dfbeta(s)
  - The difference in beta value is the change in the regression coefficient that results from the exclusion of a particular case. A value is computed for each term in the model, including the constant.
- Standardized Dfbeta(s)
  - Standardized difference in the beta value.The change in the regression coefficient that results from the exclusion of a particular case. You may want to examine cases with absolute values greater than 2/sqrt(N), where N is the number of cases.

# Collinearity

- Identifying the Source(s) of Collinearity
  - Tolerance
  - Variance Inflation Factor
  - Condition Indices and Variance Proportions
- Handling Collinearity

- Collinearity
  - We want the predictors to be highly correlated with the dependent variable.
  - We do not want the predictors to be highly correlated with each other.
  - Collinearity occurs when a predictor is “too” highly correlated with one or more of the other predictors.
- Impact of Collinearity
  - The regression coefficients are very sensitive to minor changes in the data.
  - The regression coefficients have large standard errors, which lead to low power for the predictors.
  - In the extreme case, singularity, you cannot calculate the regression equation.

- Tolerance =1-R square

The amount of overlap between the predictor and all other remaining predictors.

- The degree of instability in the regression coefficients.

- Tolerance values less than 0.10 are often considered to be an indication of collinearity.

# Variance Inflation Factor (VIF)

- The VIF tells us:
  - The degree to which the standard error of the predictor is increased due to the predictor's correlation with the other predictors in the model.

- *VIF* values greater than 10 (or, Tolerance values less than 0.10) are often considered to be an indication of collinearity.

# Collinearity diagnostics

- Eigen value -- the amount of total variation that can be explained by one dimension among the variables -- when several are close to 0, this indicates high multicollinearity. Ideal situation: all 1's.
- Condition index --square root of the ratio of each eigen value to each successive eigen value; > 15 indicates possible problem and > 30 indicates serious problem with multicollinearity.

# Collinearity diagnostics (cont'd.)

- Variance proportions -- proportion of each variable explained by a given dimension -- multicollinearity can be a problem when a dimension explains a high proportion of variance in more than one variable. The proportions of variance for each variable shows the "damage" multicollinearity does to estimation of the regression coefficient for each.

# Example

- Regress predicted variable Y on the six independent variables.
- Under statistics select collinearity diagnostics
- Okay

- Examine tolerance, VIF, and collinearity diagnostics

# Collinearity in practice

- When is collinearity a problem?
  - When you have predictors that are VERY highly correlated (>.7).
  - Multicollinearity  refers to the presence of highly intercorrelated variables in regression models, and its effect is to invalidate some of the basic assumptions underlying their estimation.

# Signs of Multicollinearity

- 1.Large standard error on the regression coefficients which leads to making estimates unstable and low t-values.
- 2.Drastic changes in regression estimates after only minor data revision.
- 3.Extreme correlations between pairs of predictors.
- 4.Omitting a variable from the equation results smaller regression errors.
- 5.A good fit may not provide good forecasts.

# Recommended Steps in Diagnosing Multicollinearity

- 1.Correlation Matrix: look for high pairwise correlation. Not enough.
- 2.VIF> 10 a sign of multicllinearity
- 3.Condition Indices (CN) :CN between 30-100 then moderate to strong multicollinearity, if combined with at least 2 high Variance proportion numbers (0.5 or more) .

# 3 Cases

- 1. Only one near dependency present

Only one CN>30 and two or more variables have Variance Proportion >0.5

2.Competing dependencies

Two or more of CN >30 and close to each other. Cummulate Variance proportions over these rows(with CN>30) and chick the ones >0.5

3.Dominating dependencies when high CN exist. Need auxiliary regression.

# Handling Collinearity

- Combine the information contained in your predictors, linear combinations (mean of z-scored predictors), factor analysis, SEM.
- Delete some of the predictors that are too highly correlated.
- Collect additional data…in the hope that additional data will reduce the collinearity.

# Comments

Since the variance of any regression coefficient depends on regression residual error, sample size, and the extent of multicollinearity, we look for:

1. Improvement of precision measurement of any variable.

- 2.Check if model specification can be improved. Omit a variable, make transformation,
- 3.Increase sample size
- 4.Replace a variable with another less correllated with current set of IV's.
- 5. Aggregation or averaging of highly intercorrellated variables.
- 6.Drop a redundant variable.

# Intercept

- A collinearity with the constant term may occur because linear combination of two or more variables are essentially constant. In such a case, do regression without intercept and examine std.dev. Or CV of the variables.

# Example

- Work out the Job Performance example and apply the diagnostics.

# Logistic Regression

- **The purpose of logistic regression**
- 
- The crucial limitation of linear regression is that it cannot deal with DV's that are dichotomous and categorical.

. Many interesting variables in  in the real life are dichotomous:  for  example,  a patient under surgery may live or die, a person mat pass or fail, a product may pass or fail quality control, a person may get(Coronary Artery Disease) CAD or not,

an employee may be promoted or not. A range of

regression techniques have been developed for analyzing data with categorical dependent variables, including logistic regression .

- Logistical regression is regularly used rather than Discriminatory analysis when there are only two categories. It is easy to use with SPSS and requires few assumptions, there is a mixture of numerical and categorical IV's
- Thus, in instances where

the independent variables are categorical, or a mix of continuous and categorical, and the

DV is categorical, logistic regression is necessary

- Since the dependent variable is dichotomous we cannot predict a numerical value for it

using logistic regression.

* So the usual regression least squares deviations criteria for best fit approach of minimizing error around the line of best fit is inappropriate. Instead, logistic regression employs binomial probability theory in which there are only two values to predict.

- Predict the probability of response=1 rather than 0, i.e. the event/person belongs to one group

rather than the other. Logistic regression forms a best fitting equation or function using the

maximum likelihood method, which maximizes the probability of classifying the observed

data into the appropriate category given the regression coefficient.

- Like ordinary regression, logistic regression provides a coefficient ‘b’, which measures

each IV’s partial contribution to variations in the DV. The goal is to correctly predict

the category of outcome for individual cases using the most parsimonious model.

- To accomplish this goal, a model (i.e. an equation) is created that includes all predictor
- variables that are useful in predicting the response variable

- There are two main uses of logistic regression:
- The first is the prediction of group membership. Since logistic regression calculates the probability of success over the probability of failure, the results of the analysis are in the form of an odds ratio.
- Logistic regression also provides knowledge of the relationships and strengths among the variables (e.g. marrying the boss's daughter puts you at a higher probability for job promotion than undertaking fi ve hours unpaid overtime each week).

- **Assumptions of logistic regression**
- Logistic regression does not assume a linear relationship between the dependent and independent variables.
- The dependent variable must be a dichotomy (2 categories).
- The independent variables need not be interval, nor normally distributed, nor linearly related, nor of equal variance within each group.
- The categories (groups) must be mutually exclusive and exhaustive; a case can only be in one group and every case must be a member of one of the groups.

- Larger samples are needed than for linear regression because maximum likelihood coefficients are large sample estimates. A minimum of 50 cases per predictor is recommended.

# LOGISTIC REGRESSION

- ***Logistic regression***. *Determines the impact of multiple independent variables presented simultaneously to predict membership of one or other of the two dependent variable categories.*

- Logistic regression analysis (LRA) extends the techniques of multiple regression analysis to research situations in which the outcome variable is categorical.
- In practice, situations involving categorical outcomes are quite common. In the setting of evaluating an educational program, for example predictions may be made for the dichotomous outcome of success/failure or improved/not-improved. Similarly, in a medical setting, an outcome to be present or abscent.

- The focus of this presentation is on situations in which the outcome variable is dichotomous, although extension of the techniques of LRA to outcomes with three or more categories (e.g., improved, same, or worse) is possible (see, for example, Hosmer
- & Lemeshow, 1989, Chapter 8). In this section, we review the multiple regression model and, then, present the model for LRA.

- The model for logistic regression analysis assumes that the outcome variable, Y, is categorical (e.g., dichotomous), but LRA does not model this outcome variable directly. Rather, LRA is based on probabilities associated with the values of Y.

- For simplicity, and because it is the case most commonly encountered in practice, we assume that Y is dichotomous, taking on values of 1 (i.e., the positive outcome, or success) and 0 (i.e., the negative outcome, or failure). In theory, the hypothetical, population proportion of cases for which Y = 1 is defined as p = P(Y =1). Then, the theoretical proportion of cases for which Y = 0 is 1 - p = P(Y = 0). In the absence of other information, we would estimate p by the sample proportion of cases for which Y = 1.

- However, in the regression context, it is assumed that there is a set of predictor variables, X1,...,Xp, that are related to Y and, therefore, provide additional information for predicting Y. For theoretical, mathematical reasons, LRA is based on a linear model

for the natural logarithm of the odds (i.e., the log-odds) in favor of Y = 1

# LOGISTIC REGRESSION ANALYSIS

- logistic regression with a binary DV is mainly used for the following reasons:
- If you use linear regression, the predicted values will become greater than one and less than zero if you move far enough on the X-axis. Such values are theoretically inadmissible.
- One of the assumptions of regression is that the variance of Y is constant across values of X (homoscedasticity). This cannot be the case with a binary variable, because the variance is PQ which varies with the value of P with maximum value =0.25..
- The significance testing of the *b* weights rest upon the assumption that errors of prediction (Y-Y') are normally distributed. Because Y only takes the values 0 and 1, this assumption is pretty hard to justify, even approximately. Therefore, the tests of the regression weights are suspect if you use linear regression with a binary DV.

- The logistic curve relates the independent variable, X, to the rolling mean of the DV, P ( ). The formula which can do so may be

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

where P is the probability of a 1 (the proportion of 1s, the mean of Y), *e* is the base of the natural logarithm (about 2.718) and *a* and *b* are the parameters of the model. The value of *a* yields P when X is zero, and *b* adjusts how quickly the probability changes with changing X a single unit (we can have standardized and unstandardized b weights in logistic regression, just as in ordinary linear regression). Because the relation between X and P is nonlinear, *b* does not have a straightforward interpretation in this model as it does in ordinary linear regression

$$odds = \frac{P}{1 - P}$$

- In logistic regression, the dependent variable is a *logit*, which is the natural log of the odds, that is,
-

$$\log(\text{odds}) = \text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

- So a logit is a log of odds and odds are a function of P, the probability of a 1. In logistic regression, we find
- logit(P) = a + bX,
- Which is assumed to be linear, that is, the log odds (logit) is assumed to be linearly related to X, our IV. So there's an ordinary regression hidden in there.

- We could talk about odds instead. Of course, we like to convert odds to a simple probability: probabilities more than odds. To get there (from logits to probabilities), we first have to take the log out of both sides of the equation.

- Then we have:

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

- . If log odds are linearly related to X, then the relation between X and P is nonlinear.

# EXAMPLE

- The following gives data regarding potential risk factors for Coronary Artery Disease(CAD). The risk factors on which data were collected are: gender, age, systolic blood pressure, diastolic blood pressure, obesity (three levels),smoking (three levels), and cholesterol level.
- Data in file:Untitled3MEDICAL  CAD.sav

# SPSS

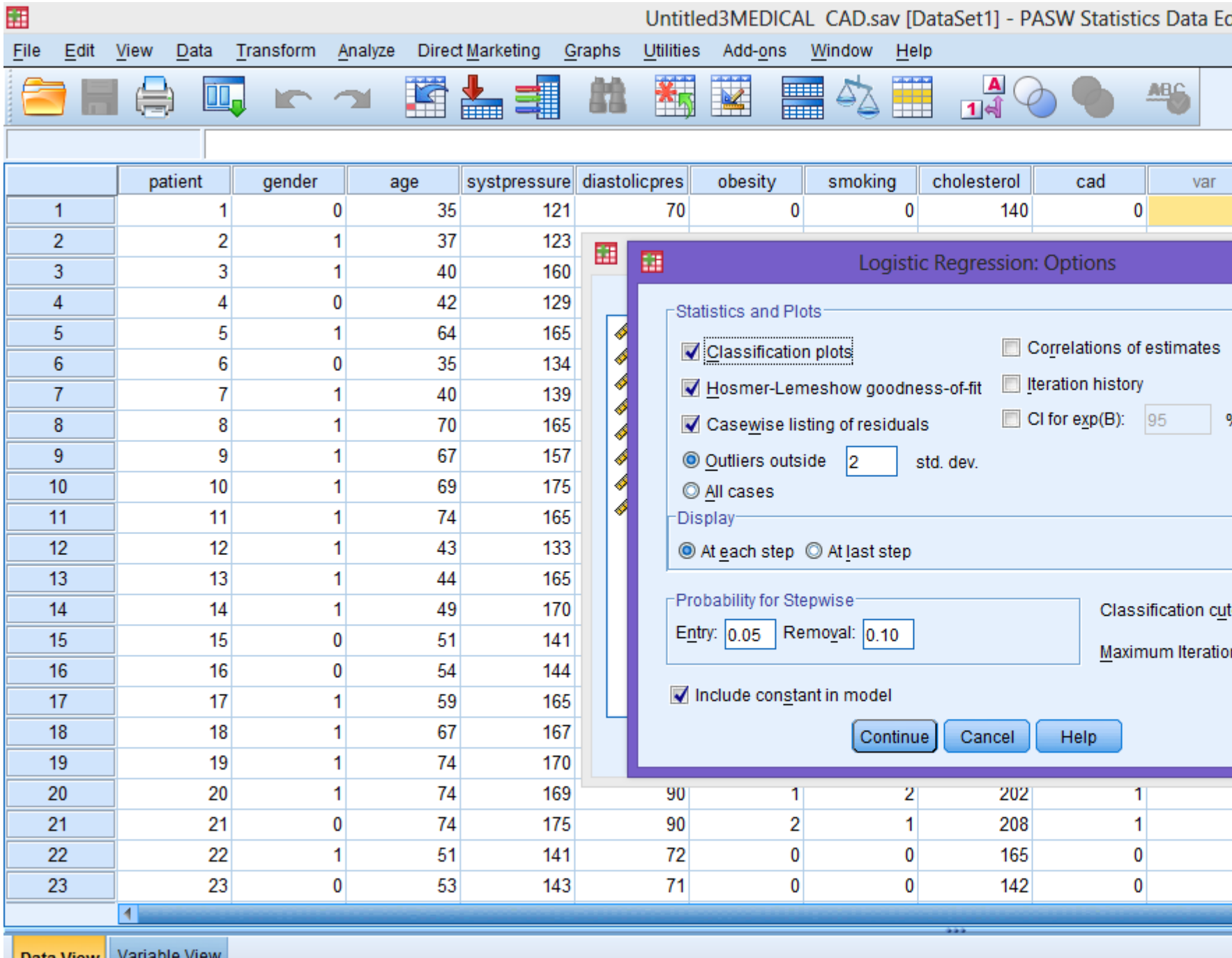
- The first analysis we shall do is to check if any one individual IV has significant effect on our ability to predict the response variable. This is done by simple logistic regression. We begin with gender as predictor and CAD as response.
- Open SPSS, get data from file:

Untitled3MEDICAL  CAD.sav

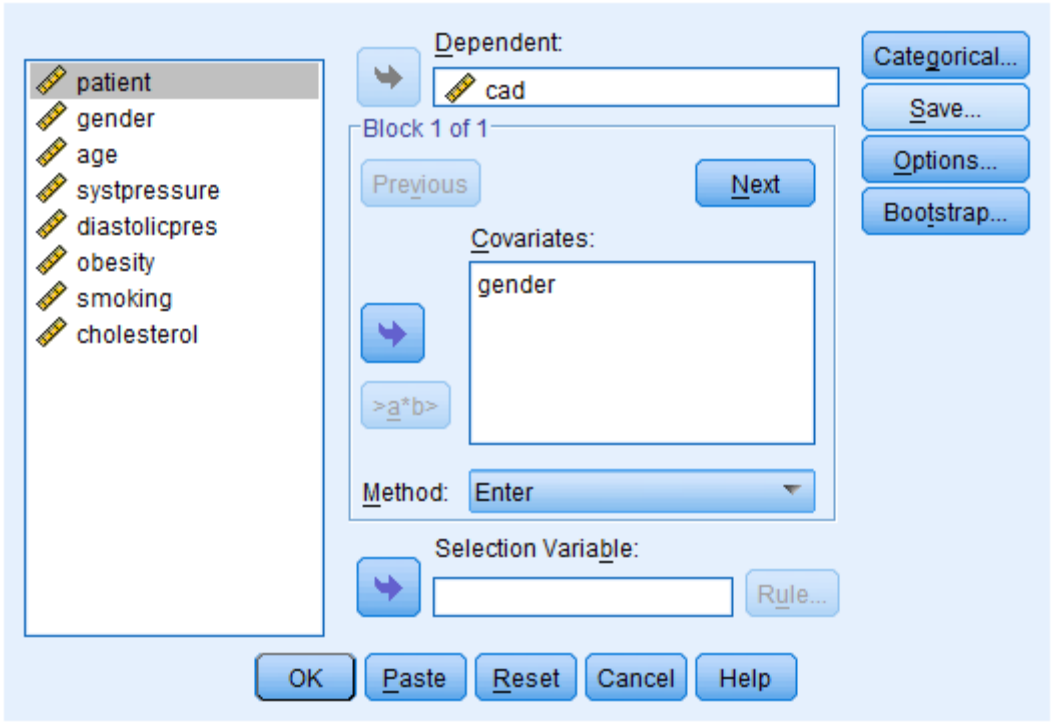
- Click Analyze> Regression > Binary Logistic

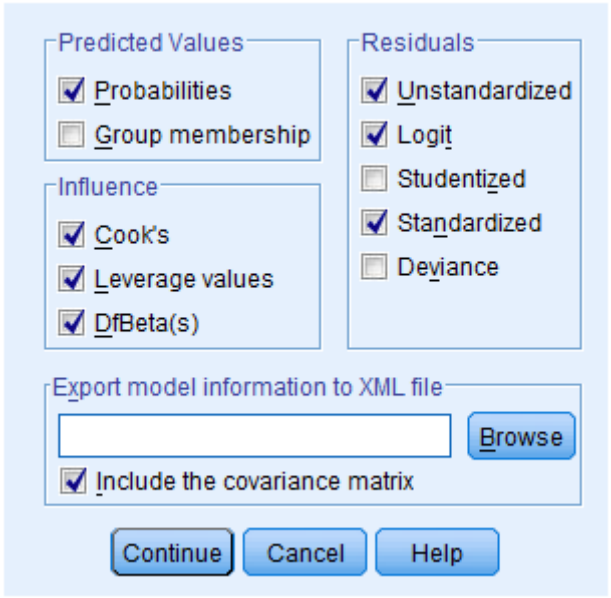
# SPSS

- Click options > click the necessary options required ,like classification plots , etc.
- Click continue
- See dialog box next slide
- Click save and click the necessary boxes.
- Click continue
- Click OK.
- You get the output.




| | patient | gender | age | systpressure | diastolicpres | obesity | smoking | cholesterol | cad | var |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 35 | 121 | 70 | 0 | 0 | 140 | 0 | |
| 2 | 2 | 1 | 37 | 123 | | | | | | |
| 3 | 3 | 1 | 40 | 160 | | | | | | |
| 4 | 4 | 0 | 42 | 129 | | | | | | |
| 5 | 5 | 1 | 64 | 165 | | | | | | |
| 6 | 6 | 0 | 35 | 134 | | | | | | |
| 7 | 7 | 1 | 40 | 139 | | | | | | |
| 8 | 8 | 1 | 70 | 165 | | | | | | |
| 9 | 9 | 1 | 67 | 157 | | | | | | |
| 10 | 10 | 1 | 69 | 175 | | | | | | |
| 11 | 11 | 1 | 74 | 165 | | | | | | |
| 12 | 12 | 1 | 43 | 133 | | | | | | |
| 13 | 13 | 1 | 44 | 165 | | | | | | |
| 14 | 14 | 1 | 49 | 170 | | | | | | |
| 15 | 15 | 0 | 51 | 141 | | | | | | |
| 16 | 16 | 0 | 54 | 144 | | | | | | |
| 17 | 17 | 1 | 59 | 165 | | | | | | |
| 18 | 18 | 1 | 67 | 167 | | | | | | |
| 19 | 19 | 1 | 74 | 170 | | | | | | |
| 20 | 20 | 1 | 74 | 169 | 90 | 1 | 2 | 202 | 1 | |
| 21 | 21 | 0 | 74 | 175 | 90 | 2 | 1 | 208 | 1 | |
| 22 | 22 | 1 | 51 | 141 | 72 | 0 | 0 | 165 | 0 | |
| 23 | 23 | 0 | 53 | 143 | 71 | 0 | 0 | 142 | 0 | |


patient
gender
age
systpressure
diastolicpres
obesity
smoking
cholesterol
Dependent:
cad
Block 1 of 1
Previous
Next
Covariates:
gender
>a*b>
Method:
Enter
Selection Variable:
Rule...
Categorical...
Save...
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help


Predicted Values
Probabilities
Group membership
Influence
Cook's
Leverage values
DfBeta(s)
Residuals
Unstandardized
Logit
Studentized
Standardized
Deviance
Export model information to XML file
Browse
Include the covariance matrix
Continue
Cancel
Help

- **Dependent Variable Encoding**
- Original Value Internal Value
- DIDN'T GET CAD 0
- HAD CAD 1

- Go to output
- Step 0: Beginning Block

IT gives the classification table. It gives the percentage of correct predictions without taking any of the predictors into consideration.

- Variables in the equation

It gives the coefficient of the constant, i.e. no predictor is taken in the analysis, B=0.516. Also it gives the odds ratio when no predictor is used (exp(B)=1.675

- WALD Statistic = B/S.e.(B) and it is z-distributed, or if it is squared then it is chi-square with df=1.It indicates the variable is significant if Wald statistic lies in the rejection region.
- -2logliklihood measures how well the model predicts the response. If large, prediction is poor, if small the model is good.
- Usually we use Chi-square=-2LLR-(-2LLn)

R:restricted model

n:null model

LL: natural log of the likelihood function

If value is> the tabulated value of chi-square, sig. is < level of significance then the variable is considered to be a significant predictor in the equation analogous to significance of B. It is the likelihood ratio test statistic.

- Cox and Snell R square: It can be interpretd like R square in Multiple Regression, but cannot reach 1.
- Nagel Kerke : can reach 1.

# Multiple Logistic Regression

- It is a generalization of simple logistic regression, whereby the DV is binary and k IV's are used as predictors. The IV's may be a combination of categorical and interval variables.
- How to combine the information of several significant explanatory variables?

WE use what is called Prognostic index

PI =sum of B*X for the k explanatory variables

If PI is large (> a cut point) we predict Y=1.

# Example:CAD problem using MLR BY SPSS

- The steps of SPSS used for MLR are the steps we used for simple logistic regression. Of course we transfer the number of explanatory we like to study.
- Open SPSS, choose type in data or existing data, OK
- Click Analyze>Regression> Binary logistic.
- Transfer cad under dependent
- Transfer the explanatory variables under Covariates.
- Continue like Simple Logistic Regression.

- Click Options and on the Dialog box ,choose the options needed for analysis as we did in the simple logistic case. Click continue.
- Click Save and on the Dialog box ,choose the options needed for analysis as we did in the simple logistic case.Click continue.
- Click OK
- Interpret printout.

# EXERCISE (1)

- 1. Get data from file conference .sav JOB PERFORMANCE.
- 2. Regress job performance on all IV's and check ones are significant.
- 3. Check assumptions

# EXERCISE (2)

- Consider data from file admission data .There are seven variables:

Gender, gpa in university, score on qualifying exam., interview score, civil service score, admission for master program, and get job.

- 1. Regress civil score on all variabes except admission and get job.Check Assumptions.
- 2. Perform logistic regression taking (a) admission as response (b) get job as response.
- 3. find odd ratios in each case.

# EXERCISE (3)

- Get data from file untitled 3 Medical CAD.
- Perform logistic regression using cad as response variable and IV's cholesterol level.
- Interpret your findings.