Correlation and Regression

Dr. Md. Atiqul Islam
M.Sc. (SUST, BD), M.Sc. (UHasselt, BE), PhD (RuG, NL)

Professor
Department of Statistics
Jagannath University, Dhaka-1100
E-mail: atique@stat.jnu.ac.bd

September 22, 2025

Concept Covered

- Correlation Analysis
- Regression Analysis
- ANalysis Of VAriance (ANOVA)
- Kruskal-Wallis Test (Nonparametric ANOVA)

Correlation Analysis

- Karl Pearson's correlation coefficient (r)
- Spearman's rank correlation coefficient (rho, ρ)

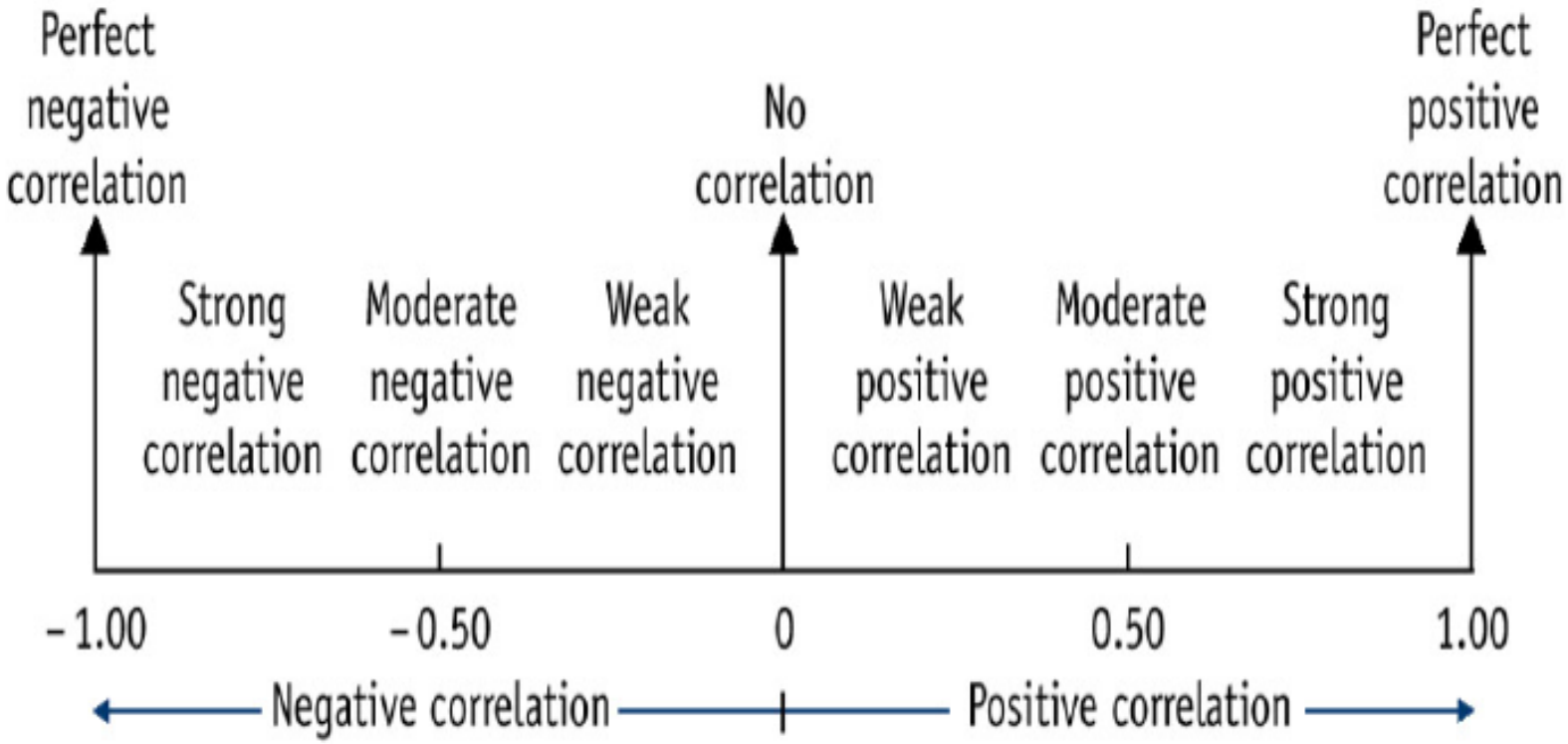
Correlation

- The degree of linear relationship existing between two or more variables observed from same entity is called correlation.
- Correlation coefficient is a quantitative measure of the **direction** and **strength** of the linear relationship between two numerically measured variables.

$$r_{xy} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}\right)}}$$

- This formula is also known as **Karl Pearson's correlation coefficient**.
- **Assumptions of r_{xy}**
  - ➡ Linearity
  - ➡ Bivariate normal distribution

Interpretation of Corr. Coeff.



Spearman's Rank Correlation

- When
  - The variables are **qualitative or binary** variables.
  - The population from which the observations are drawn is not **normal**.
  - To measure the association between two variables **without making any assumption** about their joint distribution.
- In these situations, rank each individual with respect to the two variables of interest.
- The coefficient of correlation between these ranks is known as the rank correlation or Spearman's rank correlation coefficient.

Spearman's Rank Correlation Coefficient

- **The Spearman's Rank Correlation Coefficient (ρ):**
- It is the nonparametric statistical measure used to study the strength of association between the two ranked variables.
- This method is applied to the ordinal set of numbers, which can be arranged in order, i.e. one after the other so that ranks can be given to each.
- It is defined as

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where, $\rho = Rank\ correlation\ coefficient$

$d = Difference\ of\ ranks$

$n = Number\ of\ observations$

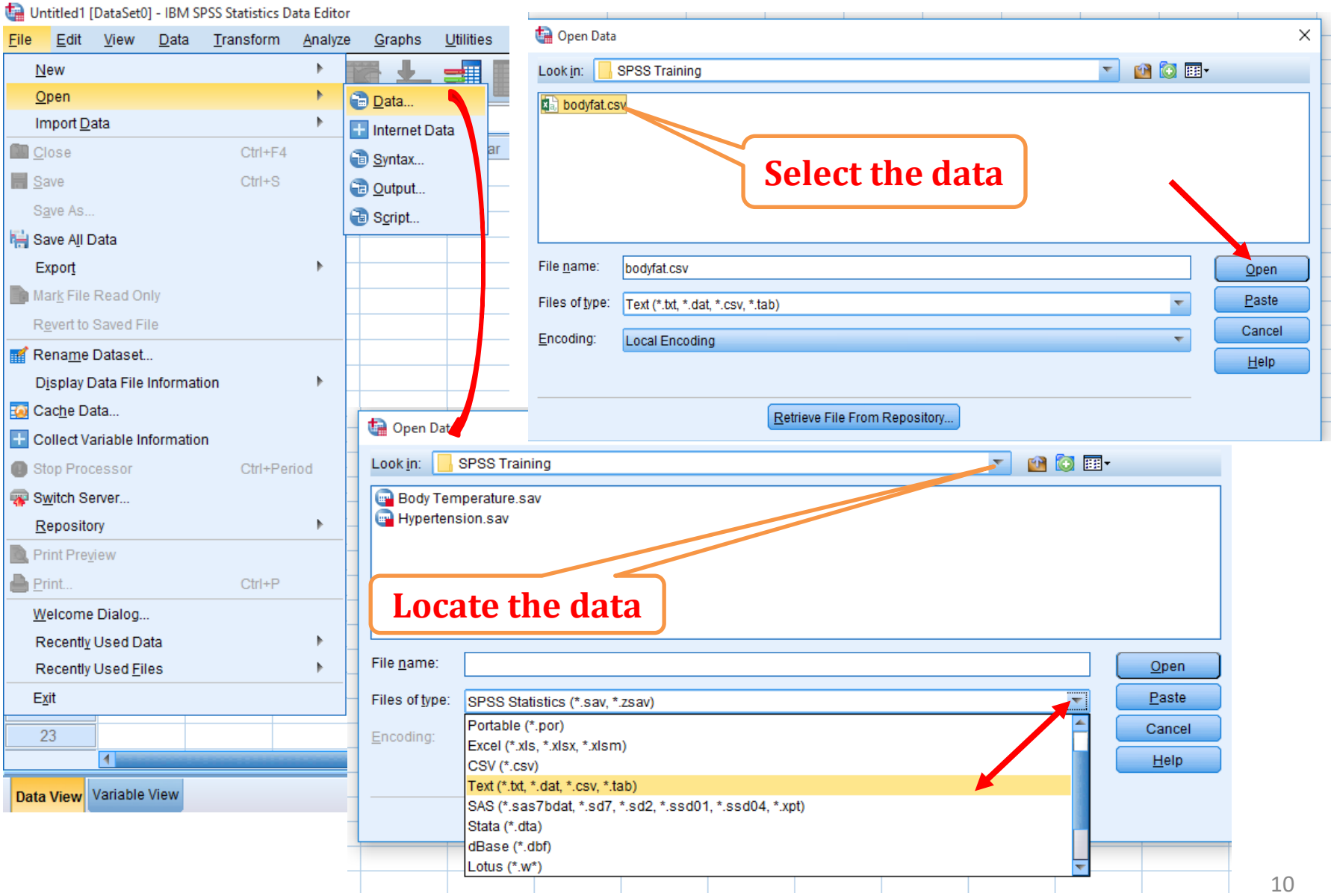
Correlation

- **Bodyfat data description (bodyfat.csv)**: A sample dataset containing the estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men.
- Some variables:
  - siri: percent body fat using Siri's equation: 495/Density – 450
  - Density: determined from underwater weighing (gm/cm**3)
  - age (years)
  - weight (lbs)
  - height (inches)
  - neck circumference (cm)

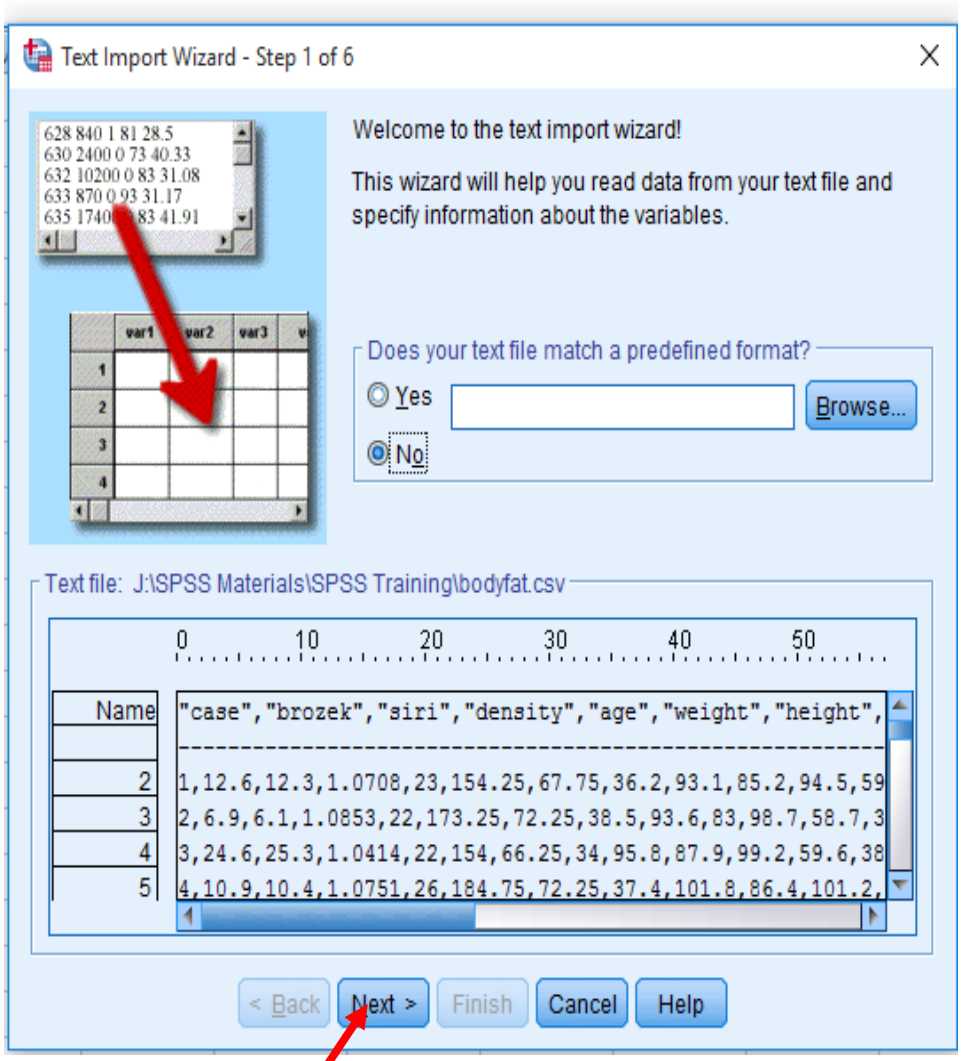
Correlation

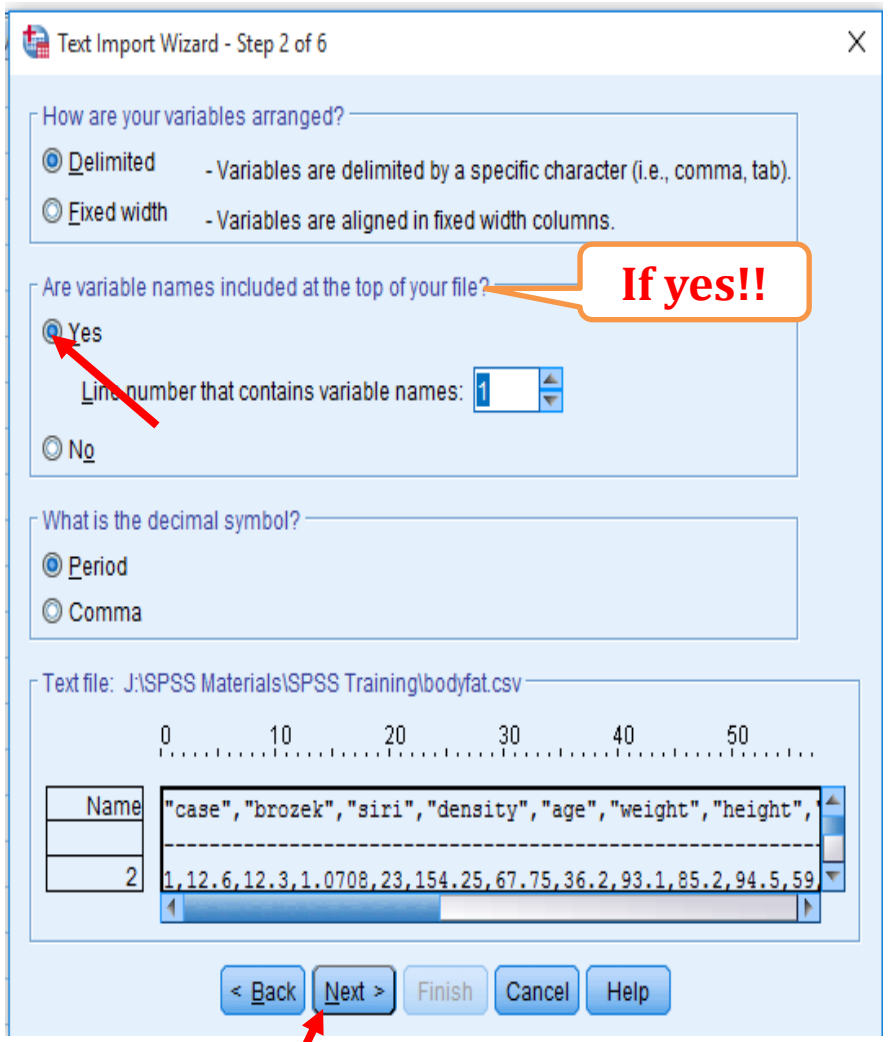
- First import bodyfat.csv data in SPSS.
- Q1: Calculate **BMI** in the bodyfat dataset and save the data in your convenient format.
- Q2: Draw a scatter plot of Siri and **neck**. Also, draw the fitted line on the scatter plot.
- Q3: Calculate the correlation coefficient of **Siri** and **neck** and comment on it.
- Q4: Draw a scatter plot of Siri and **BMI**. Also, draw the fitted line on the scatter plot.
  - What do you think about the plot? Calculate the correlation coefficient and comment on it.
  - Is there any unusual observation? If yes, how do you deal with it?
  - Do you think the unusual observation has an effect on the correlation coefficient?

Data Import (1)

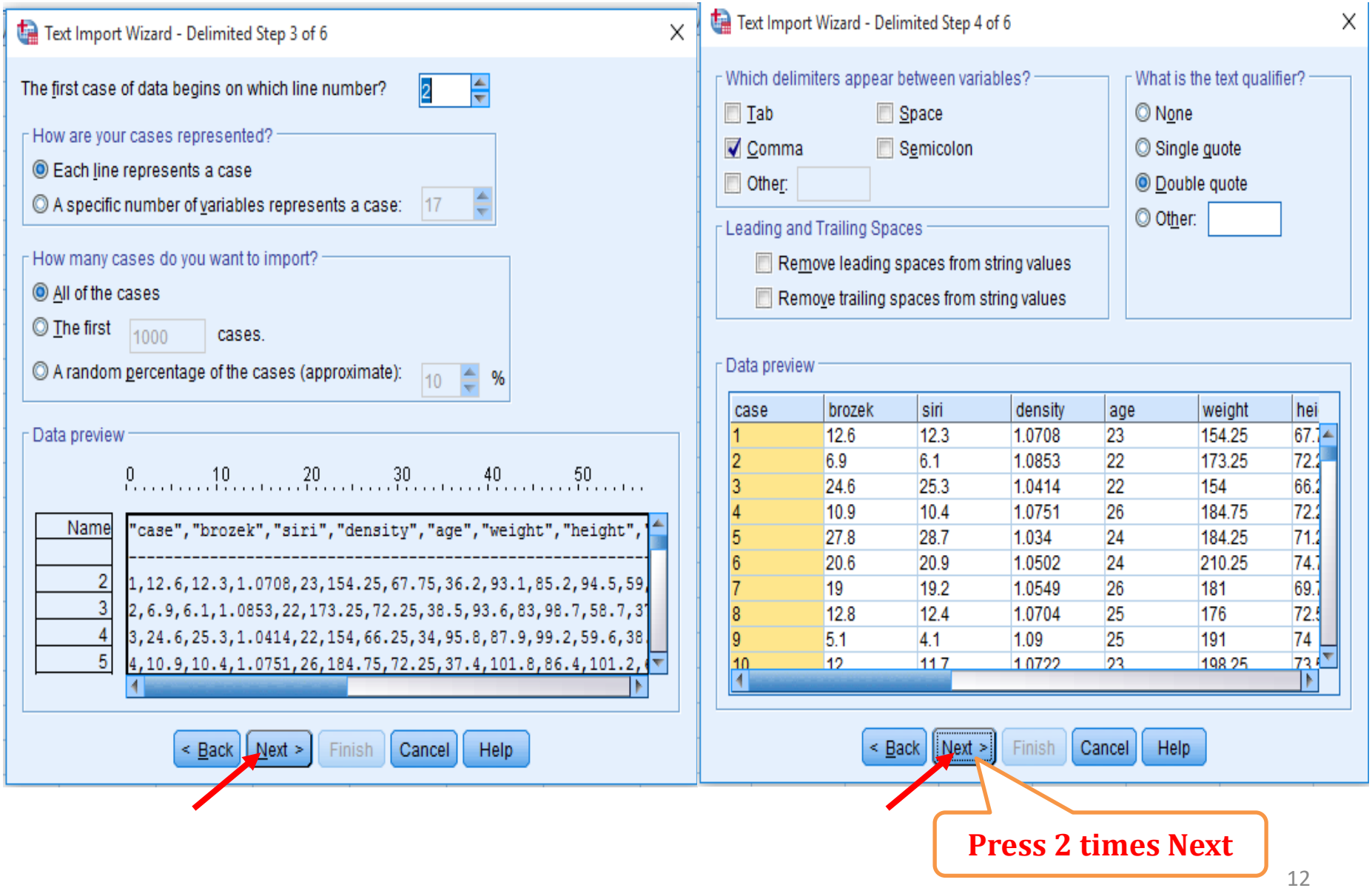


Data Import (2)

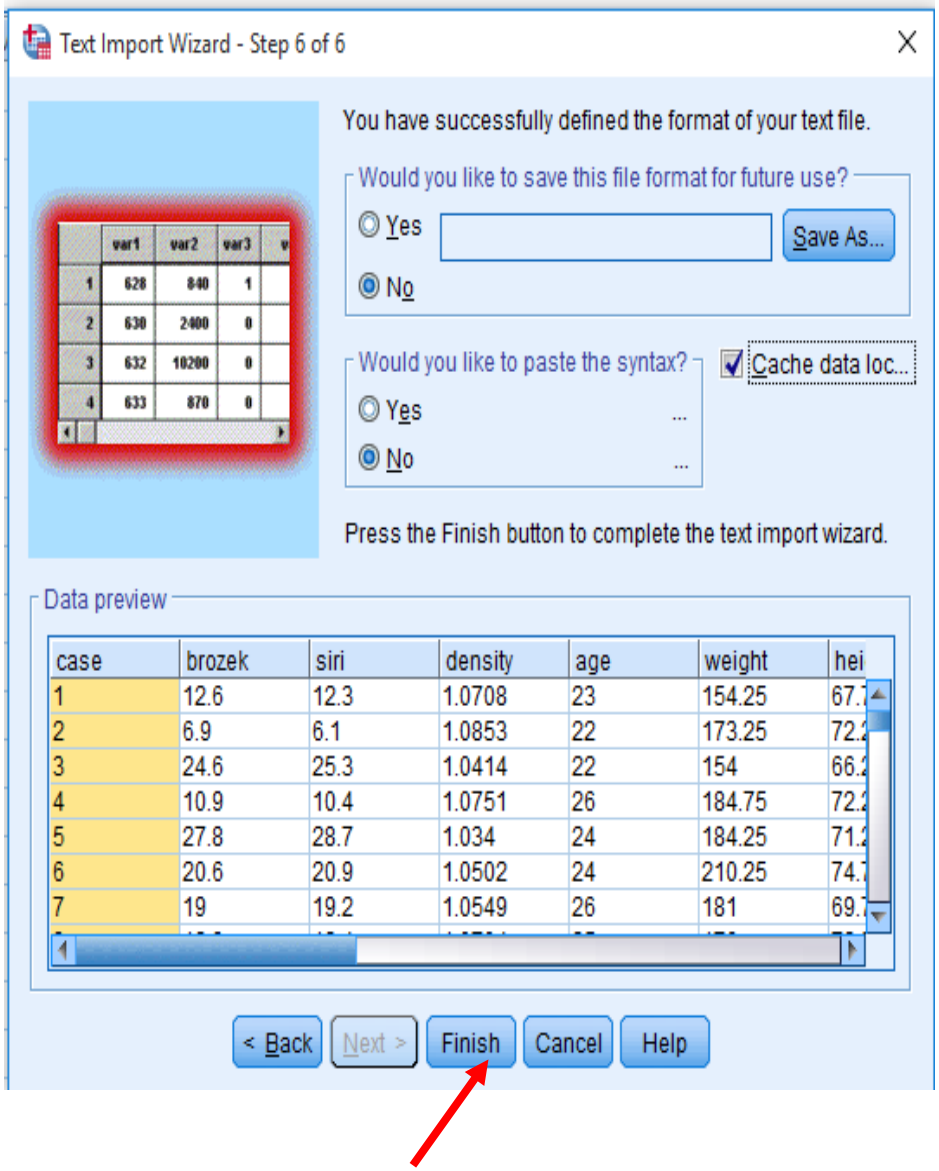


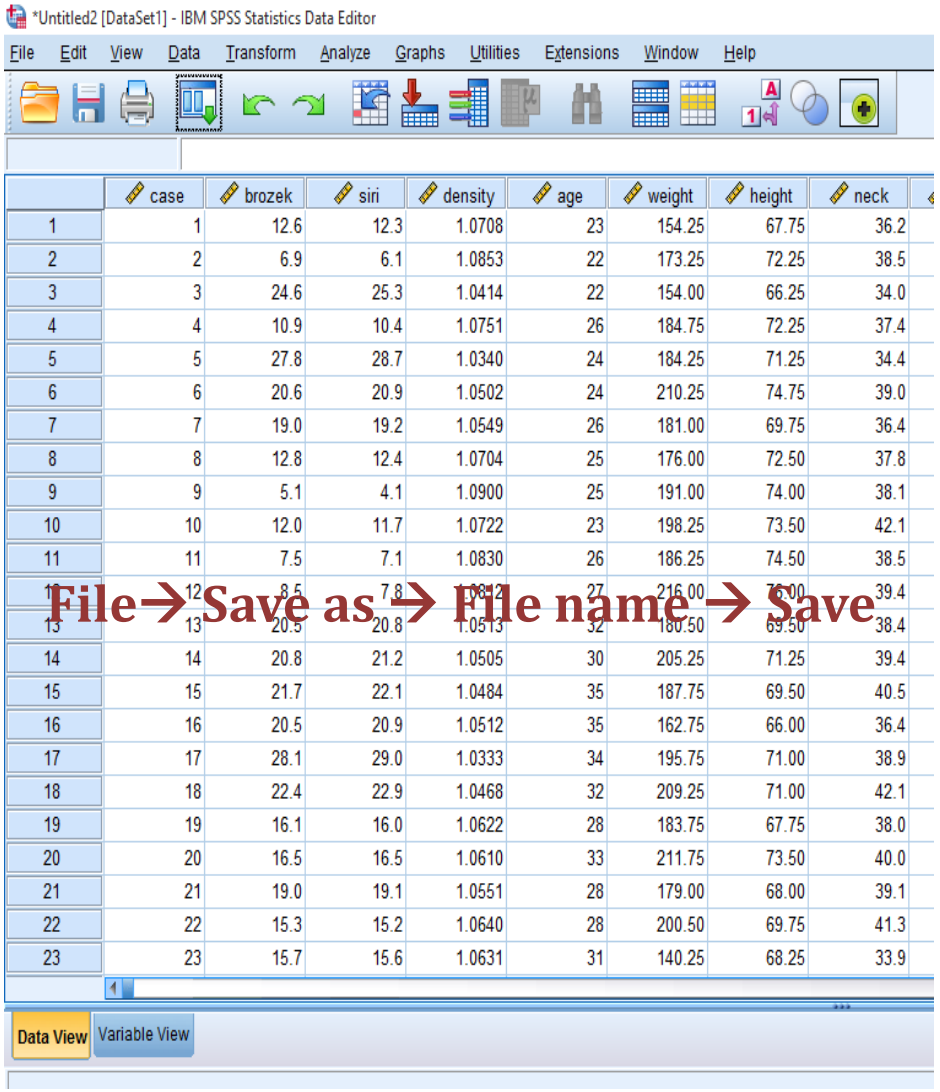


Data Import (3)



Data Import (4)

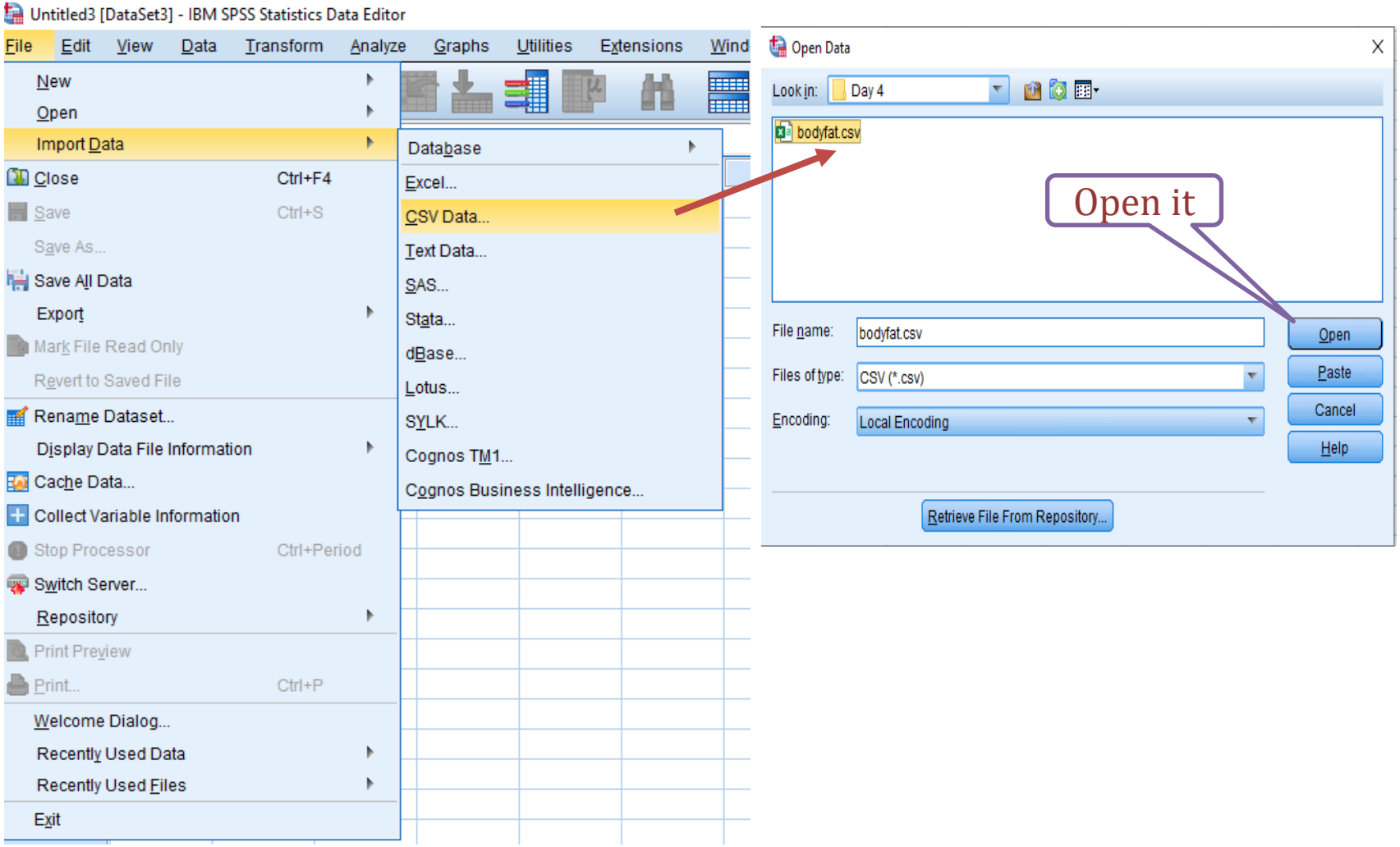




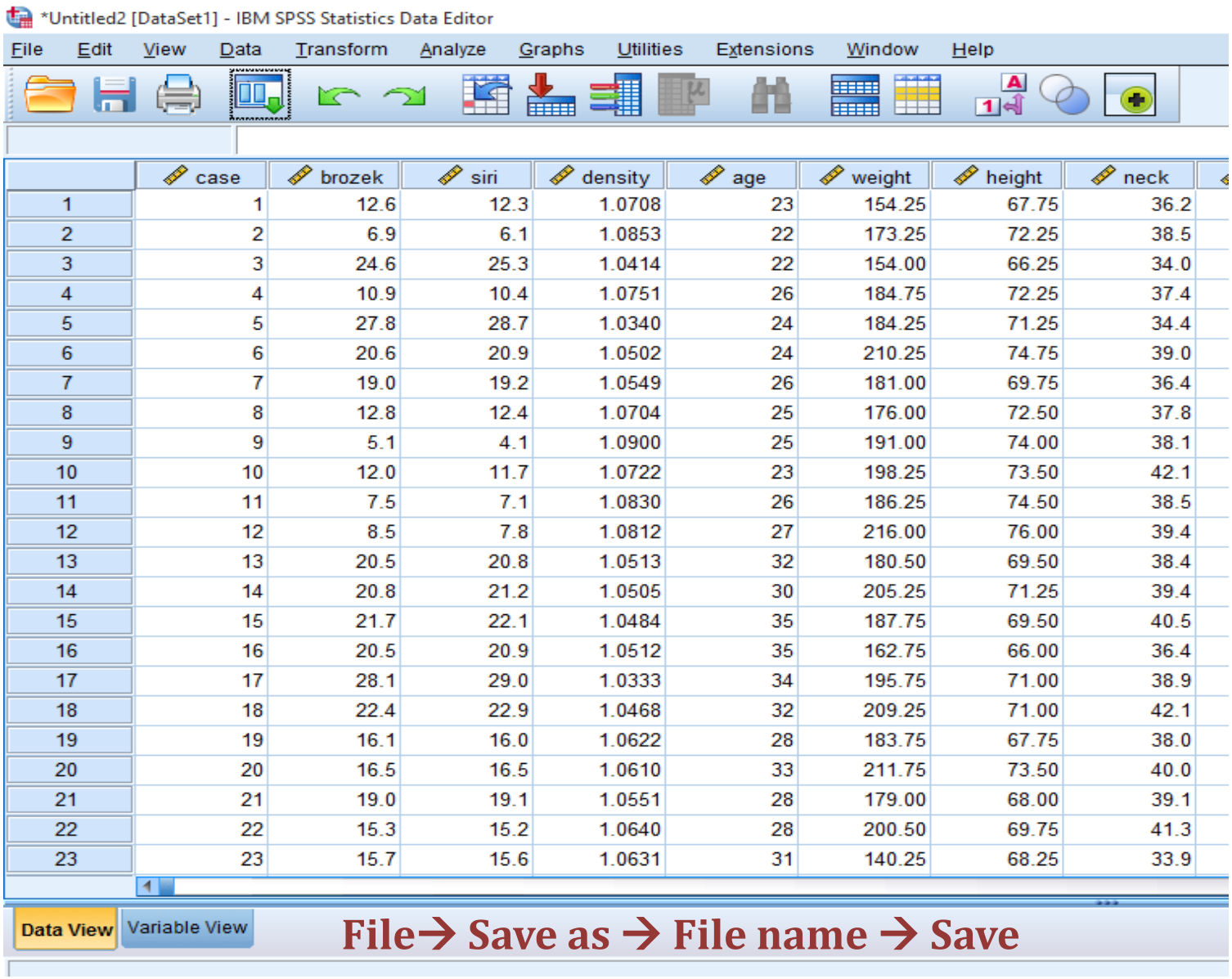
	case	brozek	siri	density	age	weight	height	neck
1	1	12.6	12.3	1.0708	23	154.25	67.75	36.2
2	2	6.9	6.1	1.0853	22	173.25	72.25	38.5
3	3	24.6	25.3	1.0414	22	154.00	66.25	34.0
4	4	10.9	10.4	1.0751	26	184.75	72.25	37.4
5	5	27.8	28.7	1.0340	24	184.25	71.25	34.4
6	6	20.6	20.9	1.0502	24	210.25	74.75	39.0
7	7	19.0	19.2	1.0549	26	181.00	69.75	36.4
8	8	12.8	12.4	1.0704	25	176.00	72.50	37.8
9	9	5.1	4.1	1.0900	25	191.00	74.00	38.1
10	10	12.0	11.7	1.0722	23	198.25	73.50	42.1
11	11	7.5	7.1	1.0830	26	186.25	74.50	38.5
12	12	8.5	7.8	1.0812	27	216.00	74.00	39.4
13	13	20.5	20.8	1.0513	32	180.50	69.50	38.4
14	14	20.8	21.2	1.0505	30	205.25	71.25	39.4
15	15	21.7	22.1	1.0484	35	187.75	69.50	40.5
16	16	20.5	20.9	1.0512	35	162.75	66.00	36.4
17	17	28.1	29.0	1.0333	34	195.75	71.00	38.9
18	18	22.4	22.9	1.0468	32	209.25	71.00	42.1
19	19	16.1	16.0	1.0622	28	183.75	67.75	38.0
20	20	16.5	16.5	1.0610	33	211.75	73.50	40.0
21	21	19.0	19.1	1.0551	28	179.00	68.00	39.1
22	22	15.3	15.2	1.0640	28	200.50	69.75	41.3
23	23	15.7	15.6	1.0631	31	140.25	68.25	33.9

File→Save as→ File name→ Save

Data Import: Simple way



Data Import: Body Fat



	case	brozek	siri	density	age	weight	height	neck
1	1	12.6	12.3	1.0708	23	154.25	67.75	36.2
2	2	6.9	6.1	1.0853	22	173.25	72.25	38.5
3	3	24.6	25.3	1.0414	22	154.00	66.25	34.0
4	4	10.9	10.4	1.0751	26	184.75	72.25	37.4
5	5	27.8	28.7	1.0340	24	184.25	71.25	34.4
6	6	20.6	20.9	1.0502	24	210.25	74.75	39.0
7	7	19.0	19.2	1.0549	26	181.00	69.75	36.4
8	8	12.8	12.4	1.0704	25	176.00	72.50	37.8
9	9	5.1	4.1	1.0900	25	191.00	74.00	38.1
10	10	12.0	11.7	1.0722	23	198.25	73.50	42.1
11	11	7.5	7.1	1.0830	26	186.25	74.50	38.5
12	12	8.5	7.8	1.0812	27	216.00	76.00	39.4
13	13	20.5	20.8	1.0513	32	180.50	69.50	38.4
14	14	20.8	21.2	1.0505	30	205.25	71.25	39.4
15	15	21.7	22.1	1.0484	35	187.75	69.50	40.5
16	16	20.5	20.9	1.0512	35	162.75	66.00	36.4
17	17	28.1	29.0	1.0333	34	195.75	71.00	38.9
18	18	22.4	22.9	1.0468	32	209.25	71.00	42.1
19	19	16.1	16.0	1.0622	28	183.75	67.75	38.0
20	20	16.5	16.5	1.0610	33	211.75	73.50	40.0
21	21	19.0	19.1	1.0551	28	179.00	68.00	39.1
22	22	15.3	15.2	1.0640	28	200.50	69.75	41.3
23	23	15.7	15.6	1.0631	31	140.25	68.25	33.9

BMI Calculation

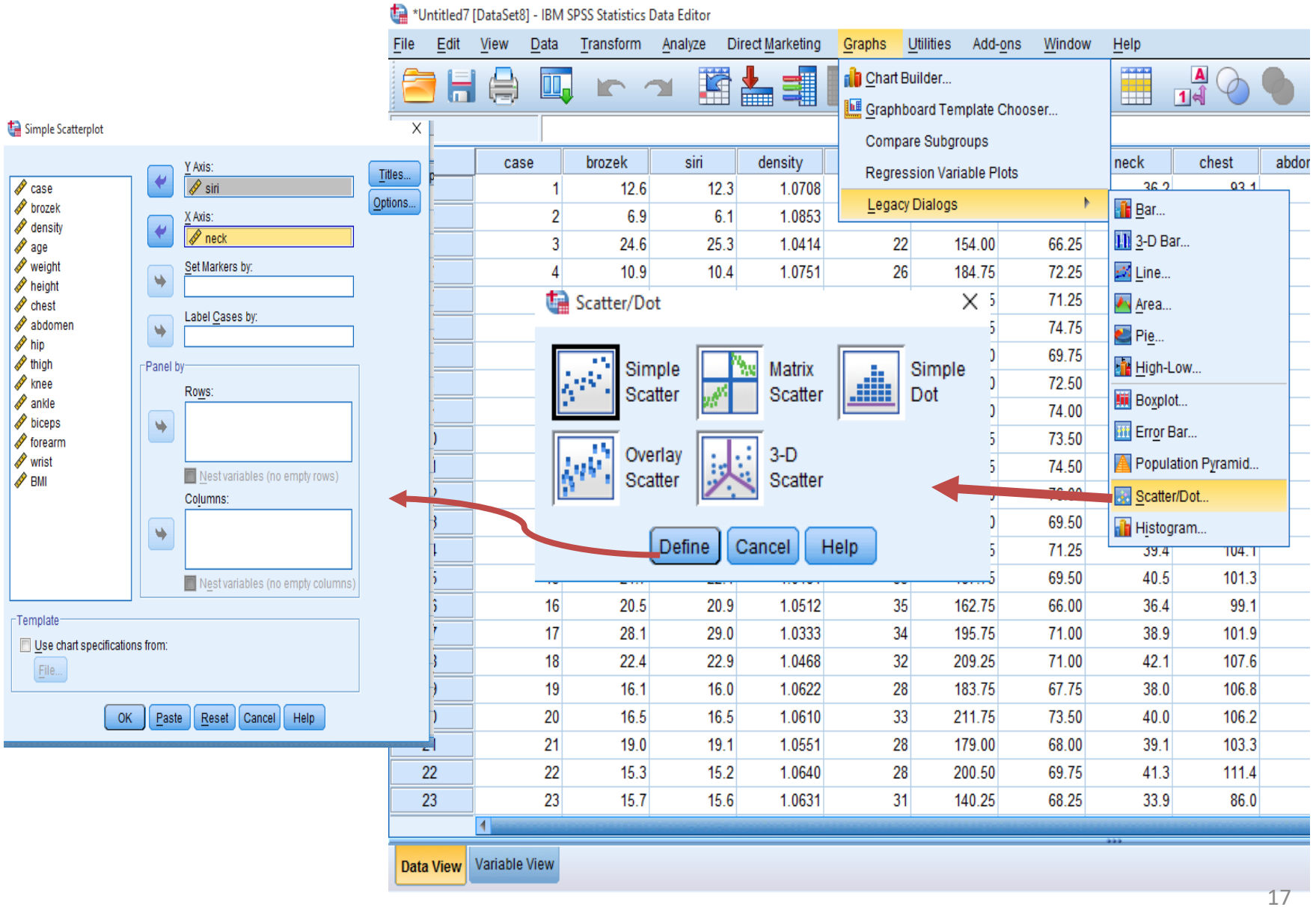
- Q1:
- To calculate BMI, let's open the Syntax window in SPSS.

FILE -> Open -> Syntax

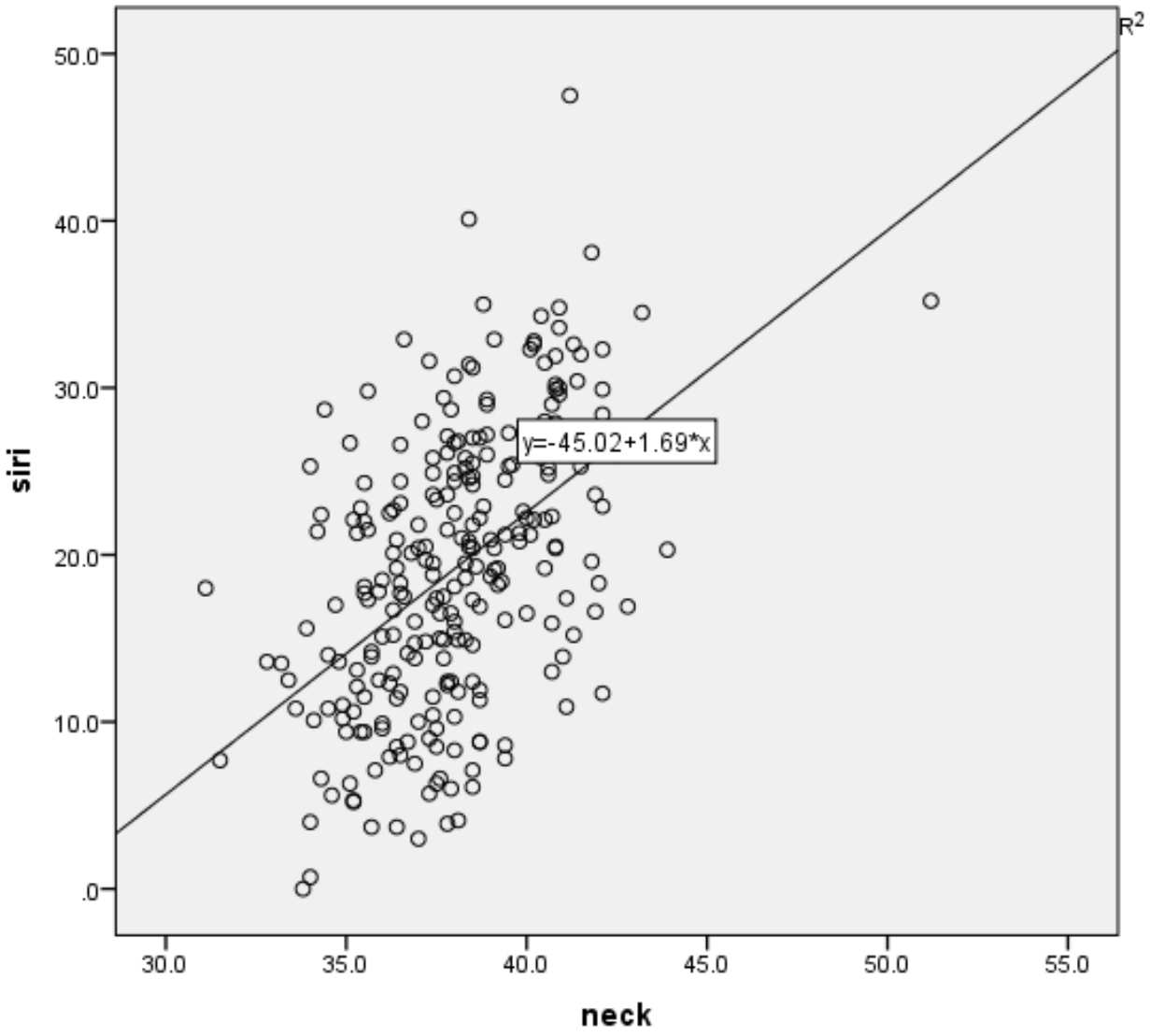
- Write down in the Syntax window:

```
Compute BMI = (weight*703/height**2).
EXECUTE.
```

Correlation (Q2): Scatter plot

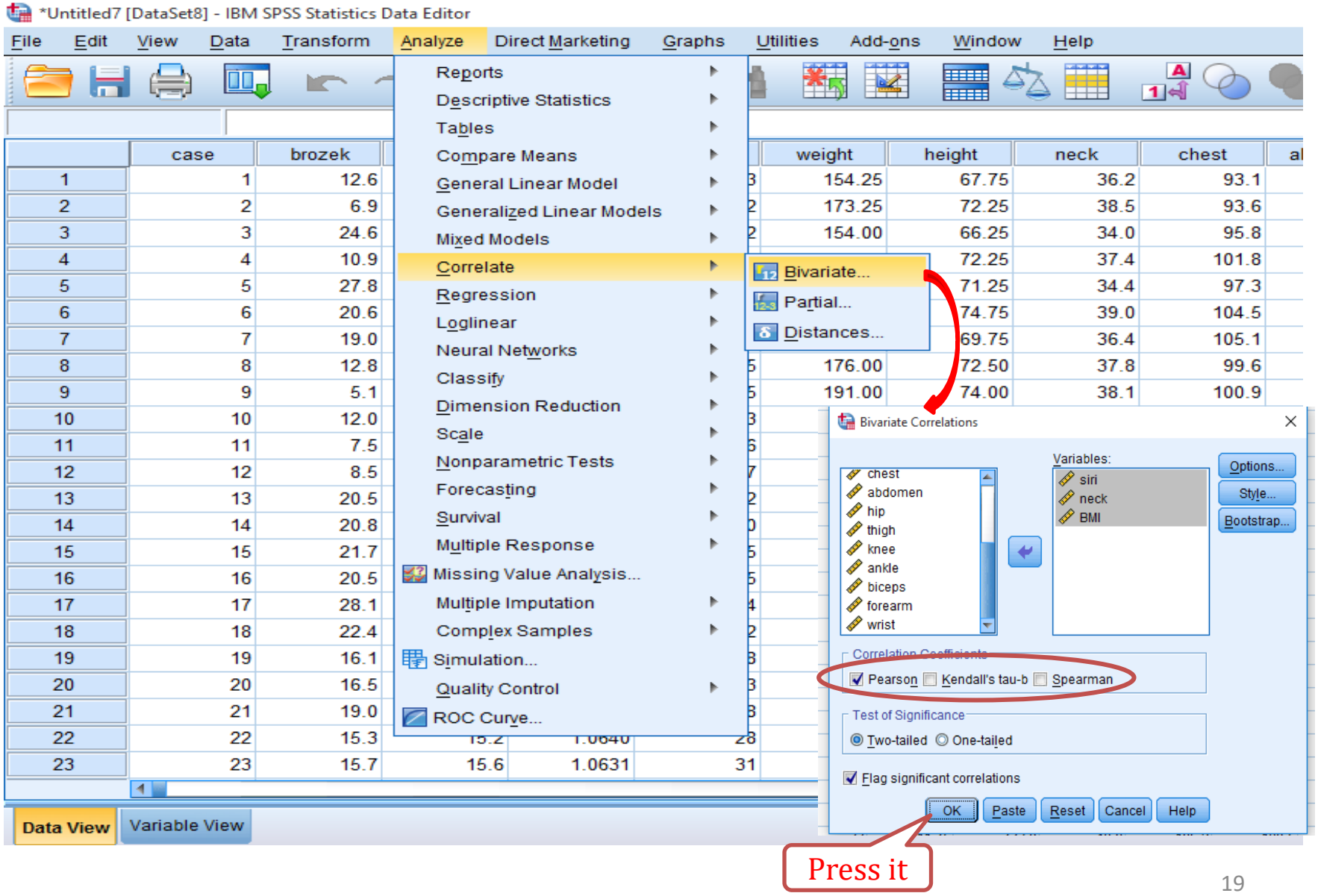


Correlation (Q2)



Double click on the graph and click the "Add Fit line at Total" Symbol

Correlation (Q3)



Correlation Matrix

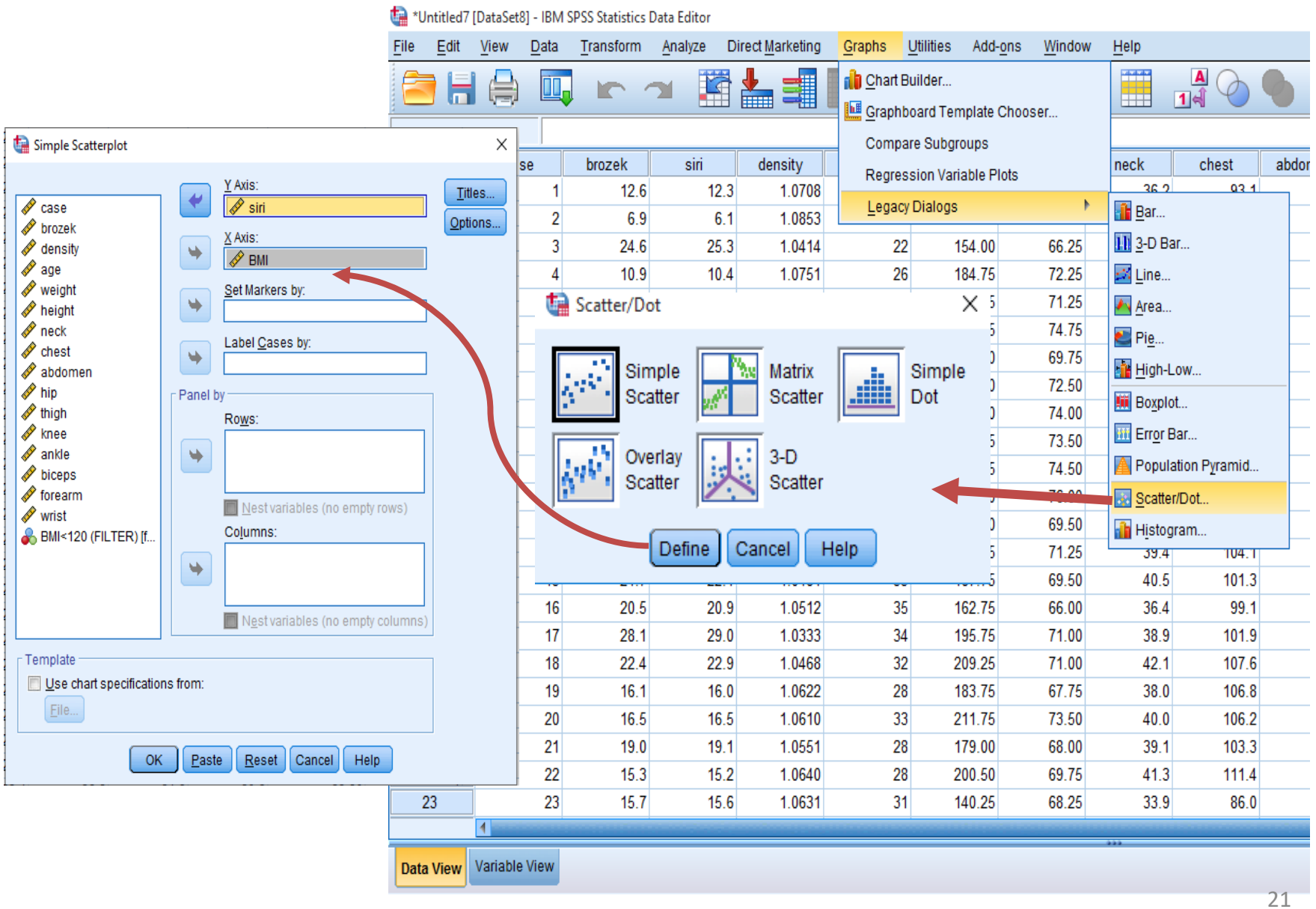
Correlations

		siri	neck	BMI
siri	Pearson Correlation	1	.491**	.371**
	Sig. (2-tailed)		.000	.000
	N	252	252	252
neck	Pearson Correlation	.491**	1	.266**
	Sig. (2-tailed)	.000		.000
	N	252	252	252
BMI	Pearson Correlation	.371**	.266**	1
	Sig. (2-tailed)	.000	.000	
	N	252	252	252

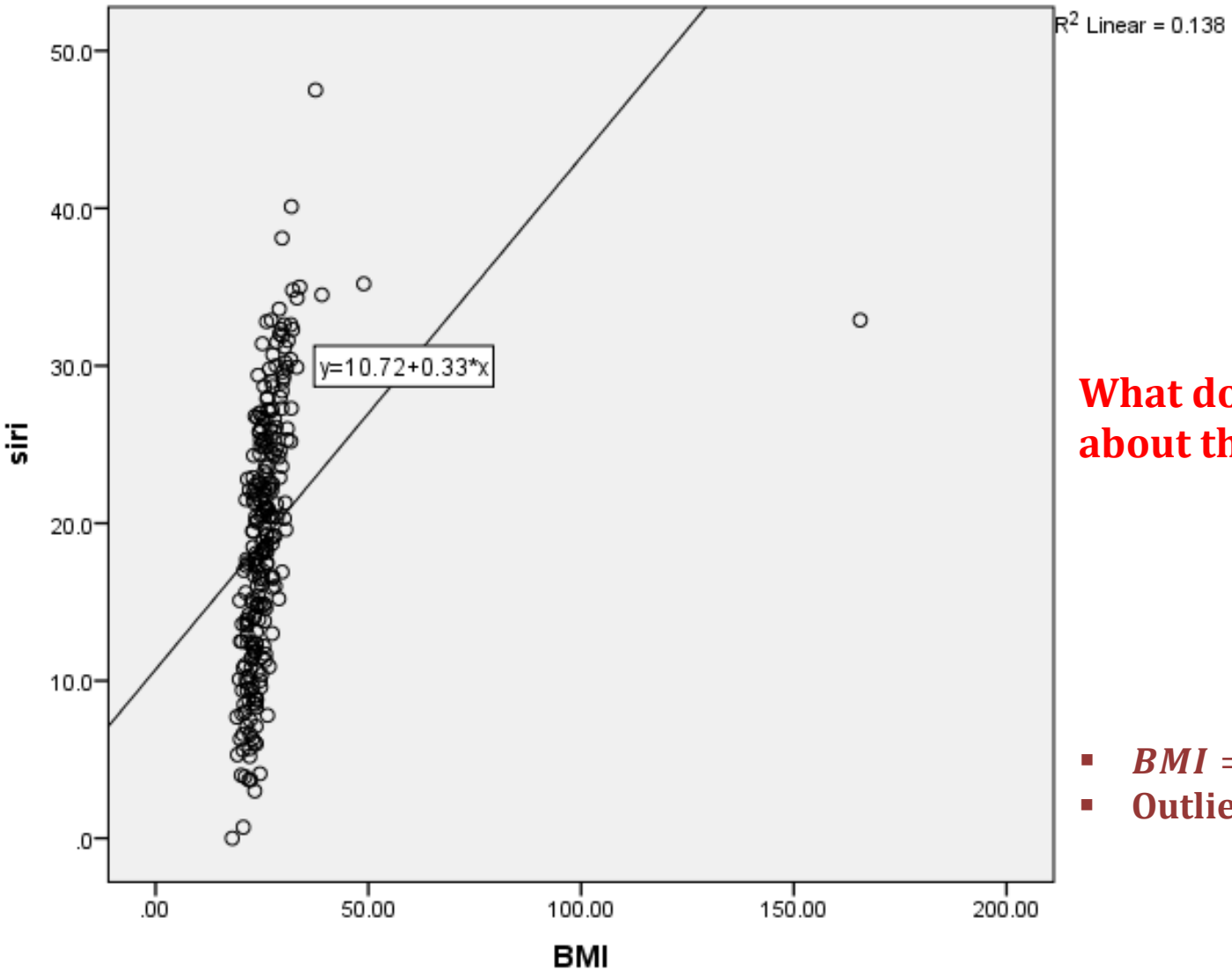
**. Correlation is significant at the 0.01 level (2-tailed).

- There is a moderate positive correlation ($r = 0.49, p < 0.001$) between *siri* and *neck circumference*. Also, a weak positive relationship is observed between *BMI* and *Siri* ($r = 0.37, p < 0.001$); and *BMI* and *neck* ($r = 0.266, p < 0.001$). Both relationships are statistically significant.

Correlation (Q4): Scatter plot



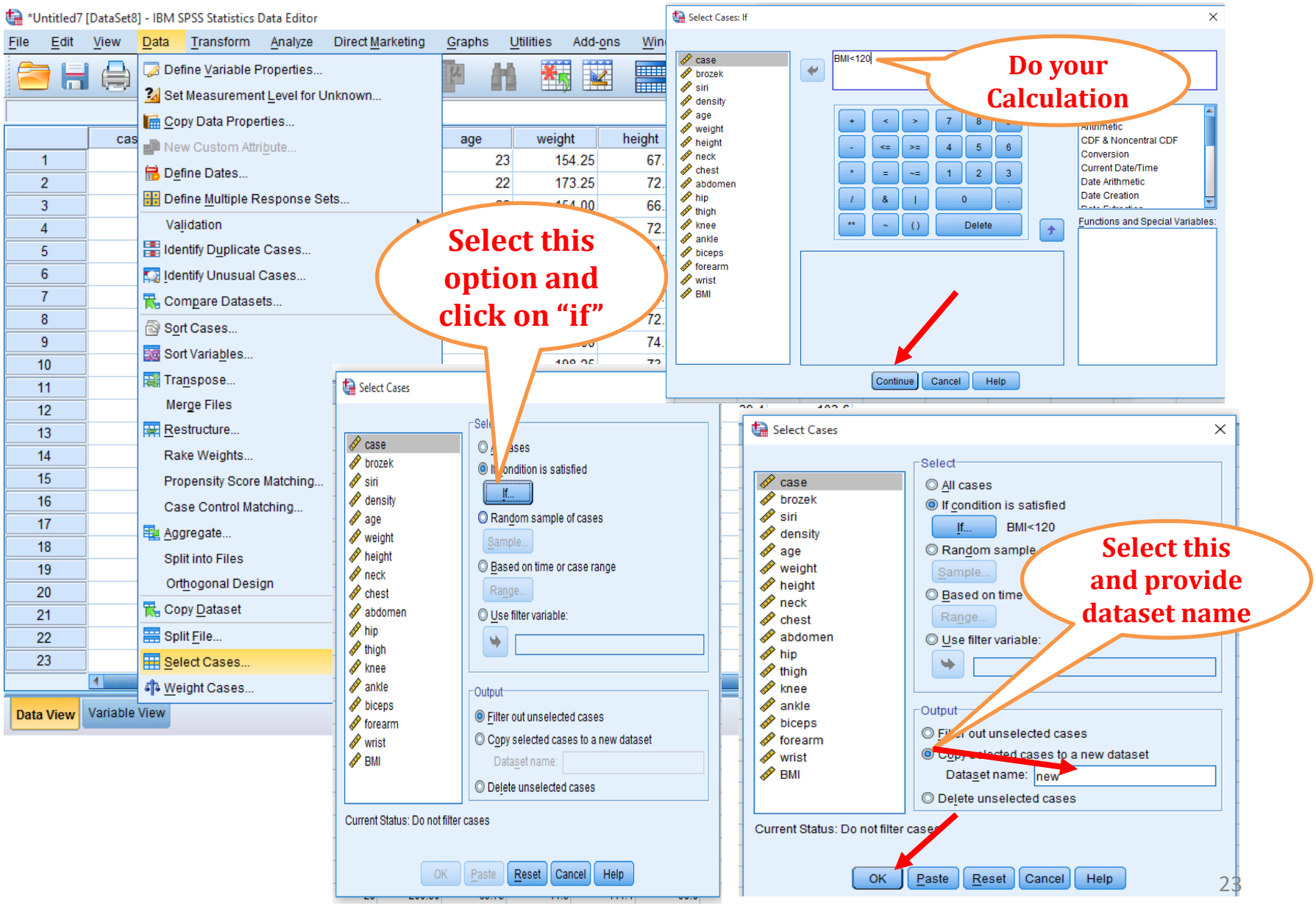
Correlation (Q4): Scatter plot



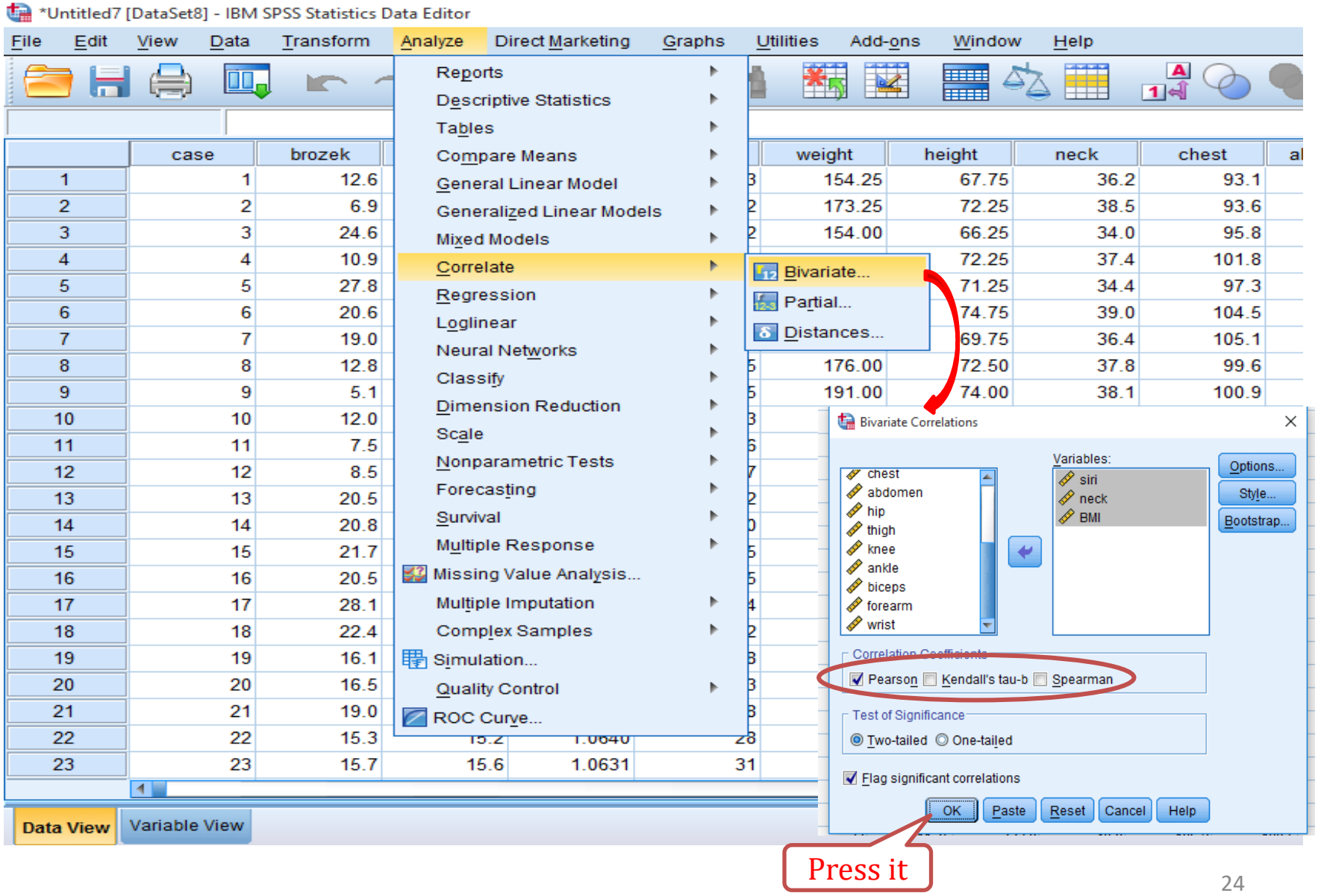
What do you think about the plot?

- $BMI = 165.60$
- **Outlier**

Correlation (Q4)



Correlation (Q4)



Correlation: After Removing Outlier

Correlations

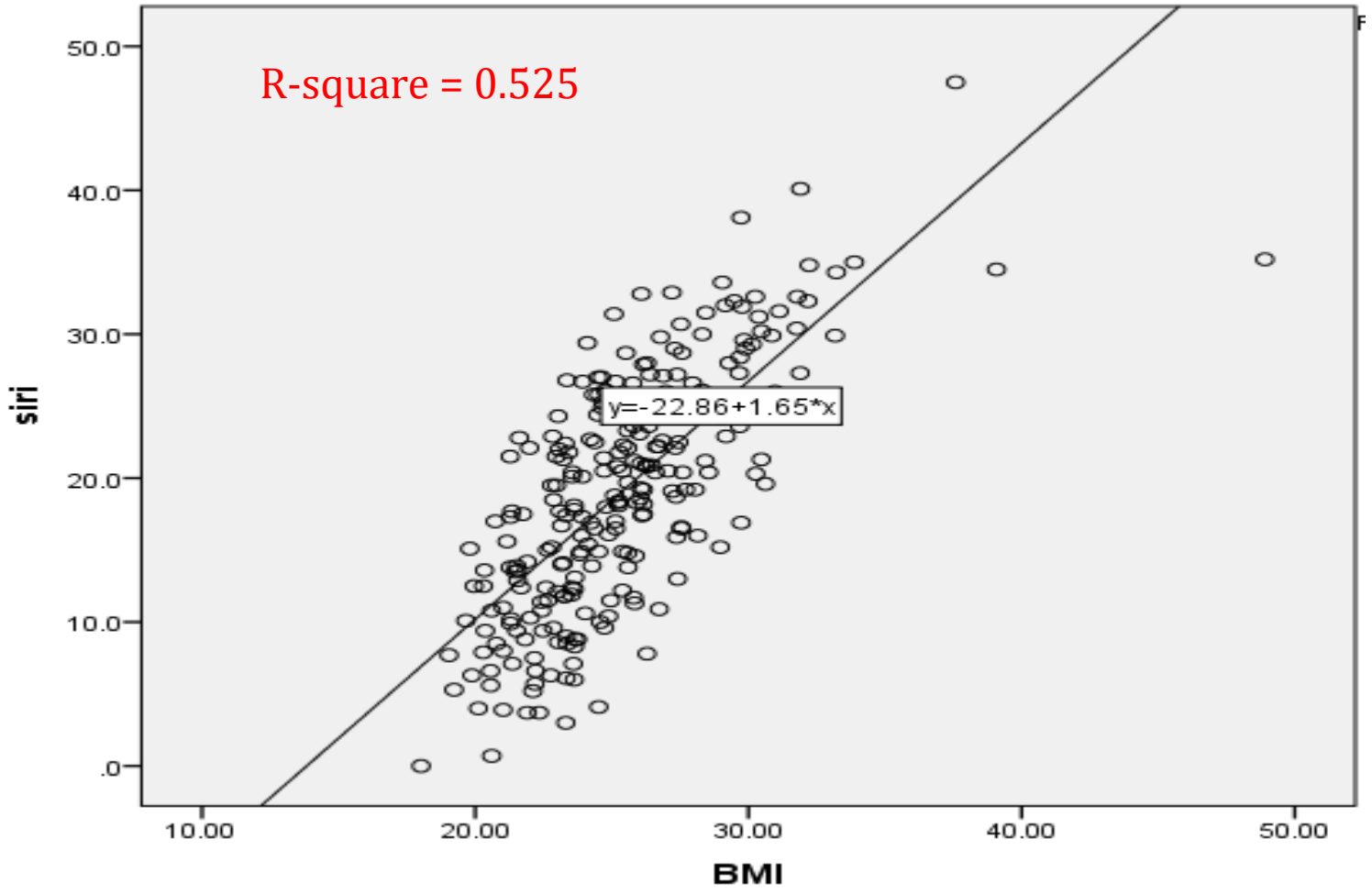
		siri	neck	BMI
siri	Pearson Correlation	1	.497**	.725**
	Sig. (2-tailed)		.000	.000
	N	251	251	251
neck	Pearson Correlation	.497**	1	.785**
	Sig. (2-tailed)	.000		.000
	N	251	251	251
BMI	Pearson Correlation	.725**	.785**	1
	Sig. (2-tailed)	.000	.000	
	N	251	251	251

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

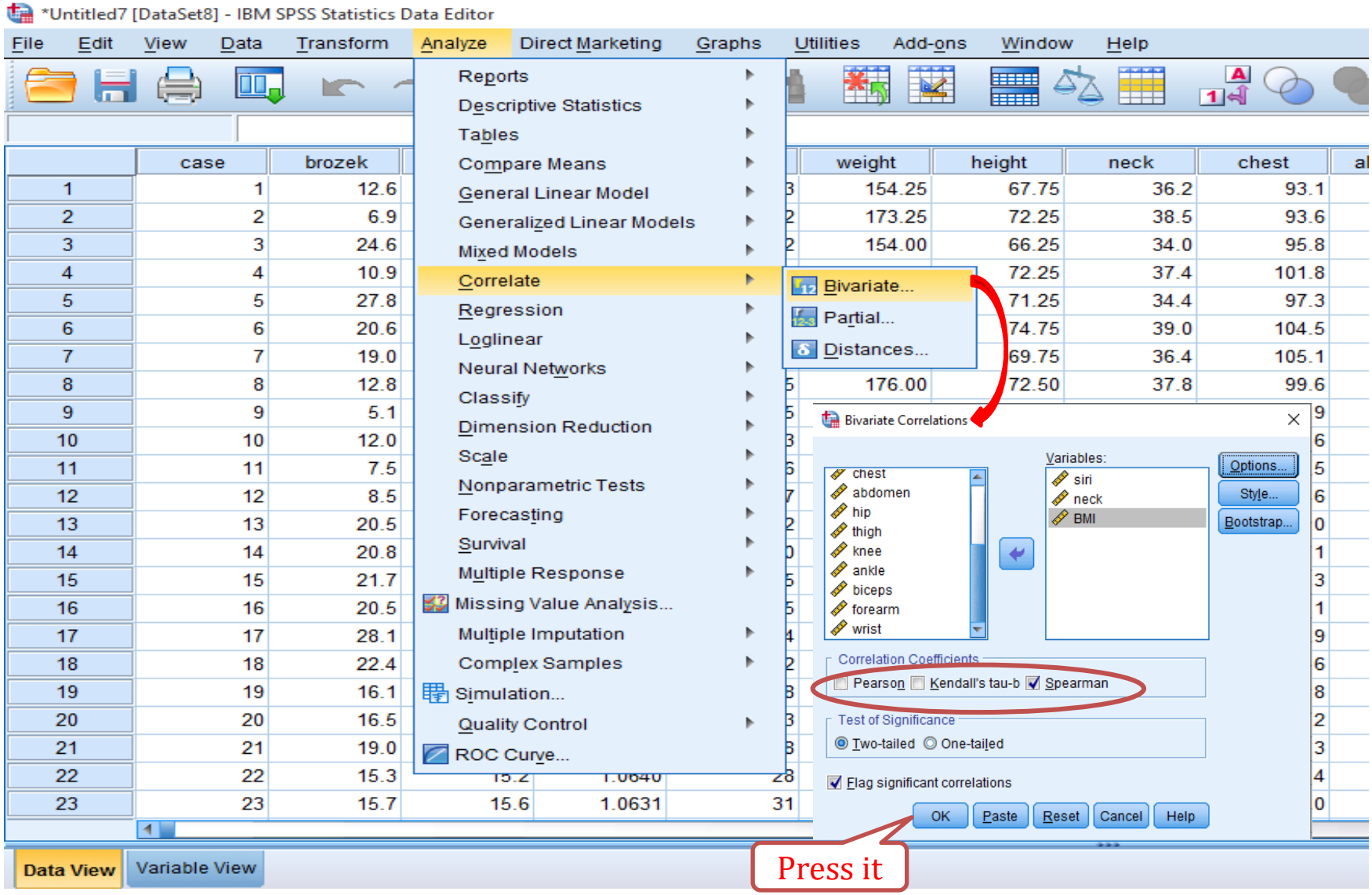
Sample size reduced

Scatter Plot (Q4)

- Use new data “Dataset name” to generate the Scatter plot of Siri and BMI



Rank Correlation Coefficient



Rank Correlation Coefficient

Correlations

			siri	neck	BMI
Spearman's rho	siri	Correlation Coefficient	1.000	.491**	.727**
		Sig. (2-tailed)	.	.000	.000
		N	252	252	252
	neck	Correlation Coefficient	.491**	1.000	.767**
		Sig. (2-tailed)	.000	.	.000
		N	252	252	252
	BMI	Correlation Coefficient	.727**	.767**	1.000
		Sig. (2-tailed)	.000	.000	.
		N	252	252	252

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

- There is a moderate positive correlation ($\rho = 0.491, p < 0.001$) between *siri* and *neck circumference*. Also, a strong positive relationship ($\rho = 0.727, p < 0.001$) is observed between *BMI* and *Siri*; and *BMI* and *neck* ($\rho = 0.266, p < 0.001$). Both relationships are statistically significant.

Regression Analysis

- Regression analysis is a statistical technique to study the cause-and-effect relationship of two or more variables observed from the same entity.
- More precisely, it is a technique to analyze and to estimate the influence of the independent variable on the dependent variable.
- After setting up a hypothesis, how can we examine the relationship between the DV and the IV?

Variables Used in Regression

- The variable being predicted is called the **dependent variable**.
- The variable(s) being used to predict the value of the dependent variable is called an **independent variable**.
- **Dependent** or **Response** or **Outcome** or **Target** or **Explained** variable (***Y***):
  - Variable we want to predict or explain.
- **Independent** or **Predictor** or **Covariate** or **Feature** or **Explanatory** variable (***X***):
  - Attempts to explain the response.
- *Independent variable causes a change in Dependent Variable and it is not possible that Dependent Variable could cause a change in Independent Variable.*
- In regression:
  - ***X*** is used to predict or explain outcome ***Y***.

Simple Linear Regression Model

- The simple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Where,

  - The variable Y is regarded as the response, outcome, or dependent variable.
  - The variable X is regarded as the predictor, explanatory, or independent variable.
  - β_0 is the intercept (parameter) of the regression model measuring the value of Y in absence of X.
  - β_1 is the slope or regression coefficient of Y on X (parameter) which measures the rate of change of Y for a unit change in the value of X.
  - ε_i is the random error term.

- The above equation is also called two variable regression model.

Multiple Linear Regression Analysis

- The multiple linear regression model is

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \ldots + \beta_p X_{ip} + \varepsilon_i}$$

Where,

- Y_i is the outcome for i
- β_0 is the intercept
- $\beta_1, \beta_2 \ldots, \beta_p$ are the slopes/regression coefficients
- $X_{i1}, X_{i2,\ldots,} X_{ip}$ are the predictors for i
- ε_i is the residual variation for i

$$\varepsilon_i \sim N(0, \sigma^2)$$

Regression Assumptions: LINE

- The dependent variable Y is linearly related to X.
- $Y's$ are independent of each other.
- Distribution of Y is normal.
- $Var(Y)$ does not depend on X.
- The error term ε_i is normally and identically and independently distributed with mean zero and variance σ^2, i. e. ,

$$\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

Linearity

Independence

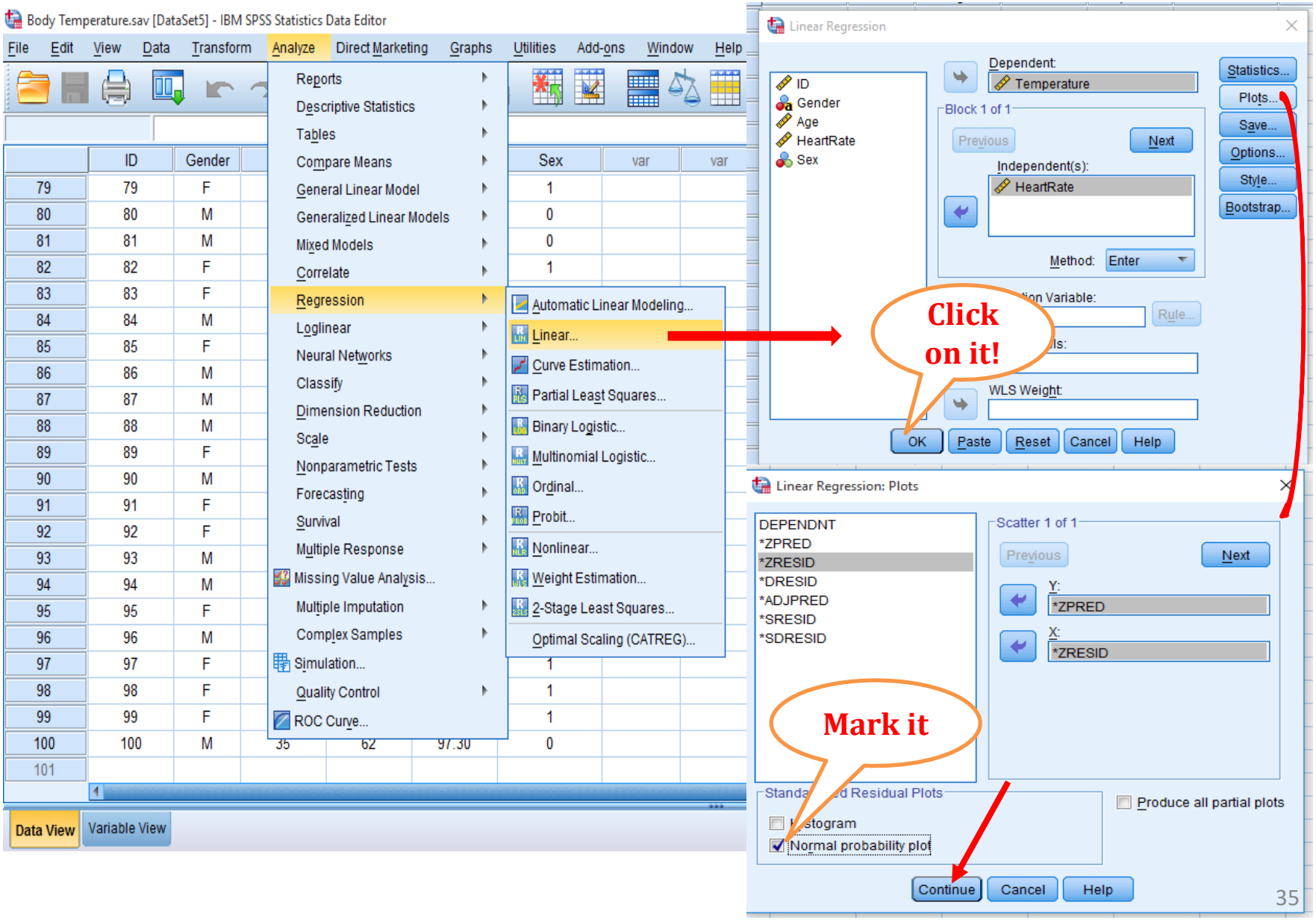
Normality

Equal variance

Example: Regression Analysis

- First import Body Temperature data (.sav) data in SPSS.
- The data represent the 100 people with heart rate and body temperature.
- Q1: What is the effect of heart rate on body temperature?
- Q2: What percentage of the variation of the body temperature is explained by heart rate?
- Q3: Check the diagnostics of the model.
- Q4: Based on the estimated regression equation, predict the body temperature if the heart rate of a person is 92.
- Q5: Identify the determinants of heart rate. **(CW)**

Simple Linear Regression Analysis



Result

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.448[a]	.200	.192	.86001

a. Predictors: (Constant), HeartRate

b. Dependent Variable: Temperature

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18.168	1	18.168	24.564	.000[b]
	Residual	72.482	98	.740		
	Total	90.650	99			

a. Dependent Variable: Temperature

b. Predictors: (Constant), HeartRate

??

??

Which one will we report?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	92.391	1.201		76.900	.000
	HeartRate	.081	.016	.448	4.956	.000

a. Dependent Variable: Temperature

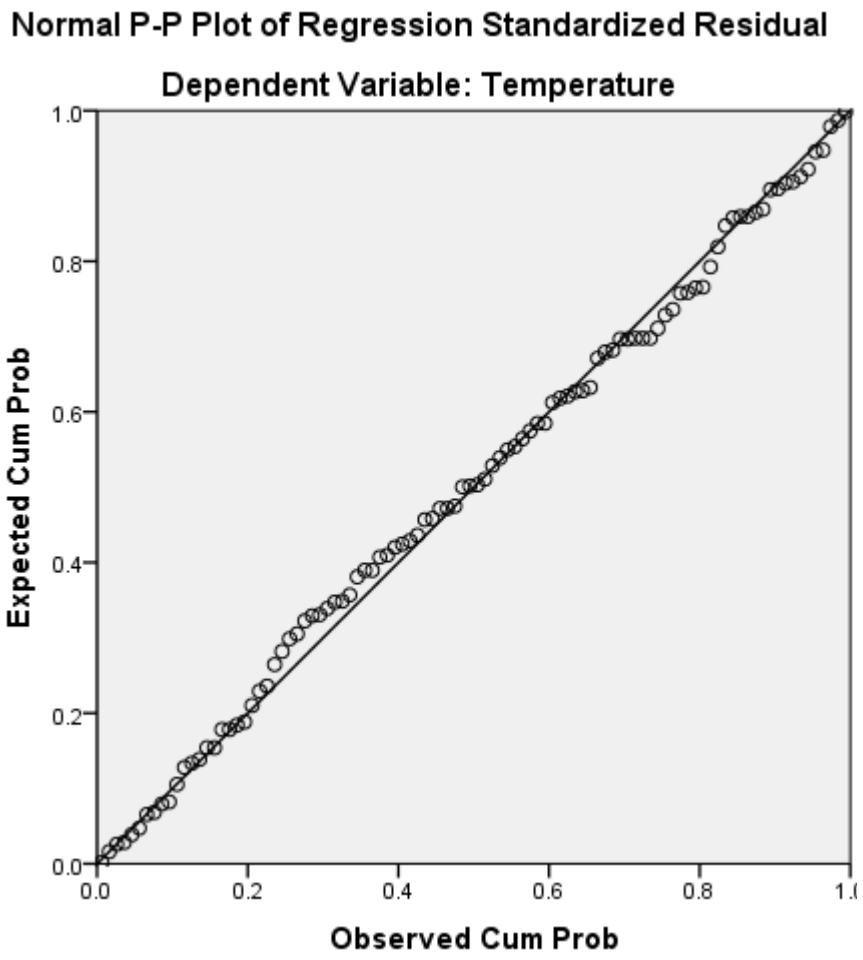
Standardized or Unstandardized

- Unstandardized β: Heart Rate $(\beta) = 0.081$ meaning that for every unit (*every beat per minute*) increase in heart rate, the body temperature (*on average*) will be increase by 0.081 unit (degree Fahrenheit).
- Standardized $\boldsymbol{\beta}$: Heart Rate $(\beta) = 0.448$ indicates that a change of one standard deviation in the *heart rate* results in a 0.448 standard deviations increase in the *body temperature*.
- Procedure to Calculate Standardized β:

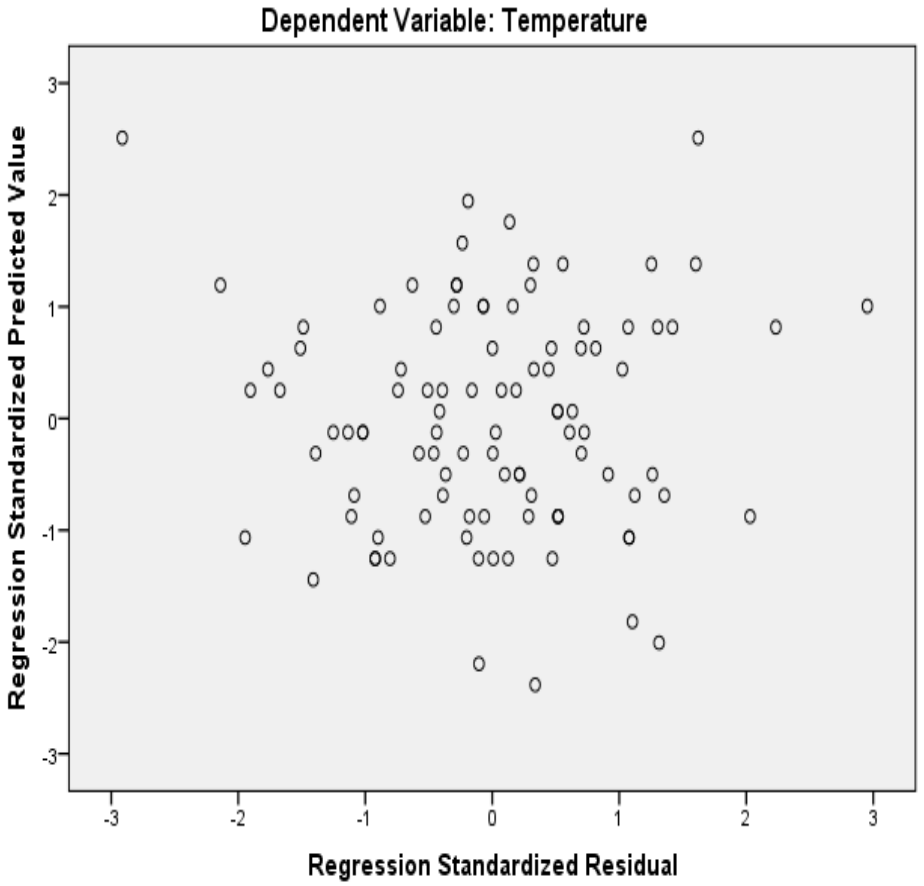
$$Standardized\ value = \frac{Each\ value\ of\ the\ variable}{Std.\ deviation\ of\ the\ variable}$$

- Re-run the regression: the desired standardized regression coefficients will be obtained.

Model Diagnostic



Constant variance



Multiple Linear Regression Analysis

- The multiple linear regression model is

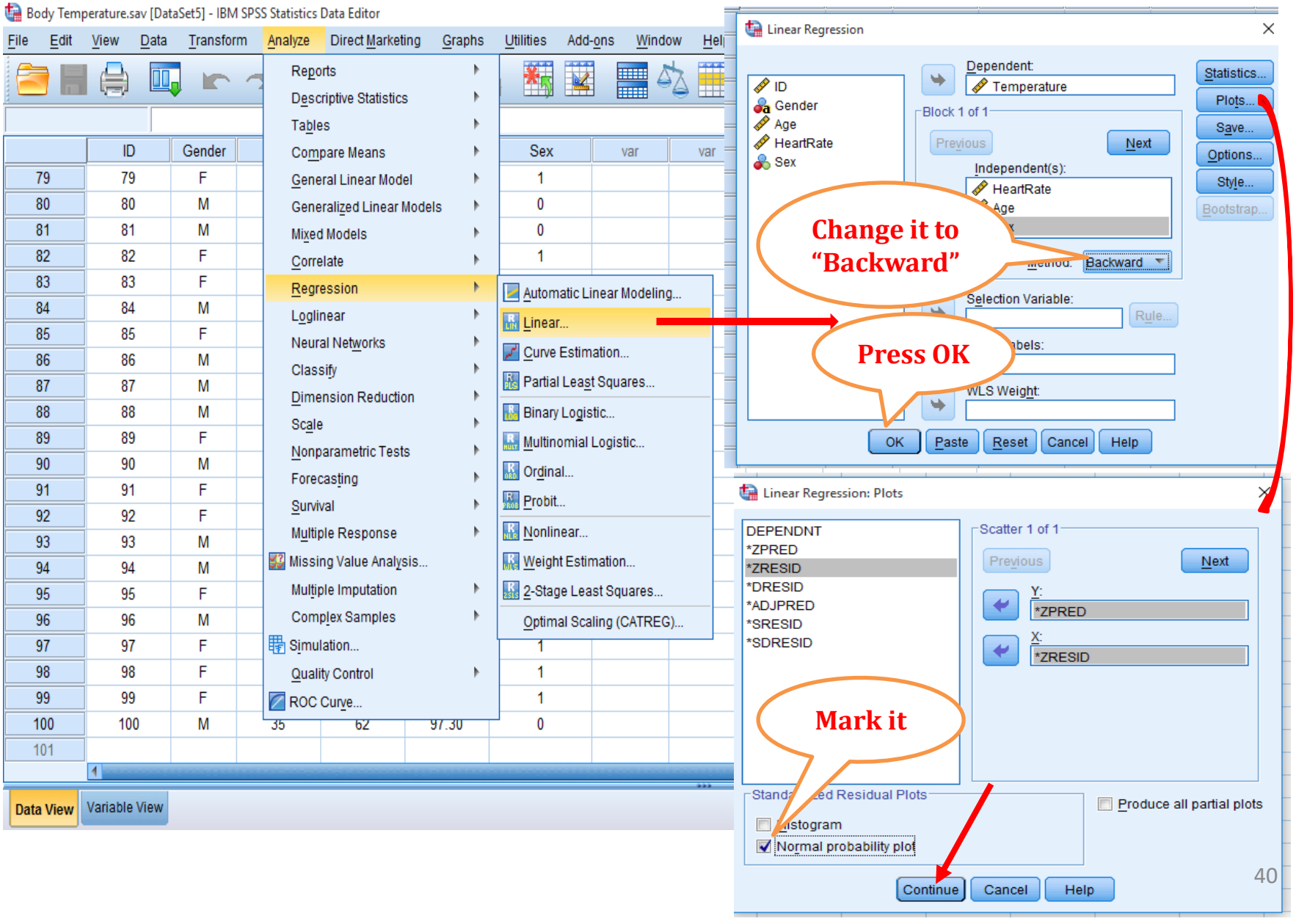
$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \ldots + \beta_p X_{ip} + \varepsilon_i}$$

Where,

- Y_i is the outcome for i
- β_0 is the intercept
- $\beta_1, \beta_2 \ldots, \beta_p$ are the slopes/regression coefficients
- $X_{i1}, X_{i2,\ldots,} X_{ip}$ are the predictors for i
- ε_i is the residual variation for i

$$\varepsilon_i \sim N(0, \sigma^2)$$

Multiple Linear Regression Analysis



Multiple Linear Regression Analysis

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.506[a]	.256	.233	.83810

a. Predictors: (Constant), Sex, Age, HeartRate

b. Dependent Variable: Temperature

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23.218	3	7.739	11.018	.000[b]
	Residual	67.432	96	.702		
	Total	90.650	99			

a. Dependent Variable: Temperature

b. Predictors: (Constant), Sex, Age, HeartRate

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	92.890	1.238		75.023	.000	90.432	95.347		
	HeartRate	.085	.016	.473	5.344	.000	.054	.117	.987	1.013
	Age	-.026	.013	-.177	-1.996	.049	-.052	.000	.989	1.012
	Sex	.292	.168	.153	1.739	.085	-.041	.625	.997	1.003

a. Dependent Variable: Temperature

The general rule of thumb:
VIF > 4 warrant further investigation

VIF > 10 are signs of serious multicollinearity

Tolerance < 0.10 means multicollinearity

Model building

- How do we decide if an explanatory variable has to be included in the model or not?
  - Because of the theory
  - P-value < 0.05
- Method of variable selection
  - Simple linear regression for all variables
  - Pre-selection on large alpha (0.25) [Hosmer–Lemeshow criteria]
    - Default in SPSS: Enter
    - Better: Backward by hand
    - Stepwise methods save time, but in general not recommended.
- Model evaluation: (adjusted) R^2

Model building

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.506[a]	.256	.233	.83810

a. Predictors: (Constant), Sex, Age, HeartRate

b. Dependent Variable: Temperature

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23.218	3	7.739	11.018	.000[b]
	Residual	67.432	96	.702		
	Total	90.650	99			

a. Dependent Variable: Temperature

b. Predictors: (Constant), Sex, Age, HeartRate

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	92.890	1.238		75.023	.000	90.432	95.347		
	HeartRate	.085	.016	.473	5.344	.000	.054	.117	.987	1.013
	Age	-.026	.013	-.177	-1.996	.049	-.052	.000	.989	1.012
	Sex	.292	.168	.153	1.739	.085	-.041	.625	.997	1.003

a. Dependent Variable: Temperature

Problem!!!

The general rule of thumb: Variance Inflation Factor (VIF)

VIF > 4 warrant further investigation.

VIF > 10 are signs of serious multicollinearity.

Tolerance < 0.10 means multicollinearity

Model building

Stepwise method

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	92.391	1.201		76.900	.000	90.006	94.775		
	HeartRate	.081	.016	.448	4.956	.000	.048	.113	1.000	1.000
2	(Constant)	93.152	1.242		75.023	.000	90.688	95.617		
	HeartRate	.084	.016	.466	5.216	.000	.052	.116	.989	1.011
	Age	-.027	.013	-.181	-2.020	.046	-.053	.000	.989	1.011

a. Dependent Variable: Temperature

Regression: CW

- Research Question: Identify the determinants of heart rate.

ANOVA

- The one-way ANOVA is used to test the claim that two or more population means are equal.
- This is an extension of the two independent samples *t*-test.
- The outcome variable must be continuous, and the independent variable must have more than two groups.
- An ANOVA uses the *F* test – comparison of variance.
- Reveals if a difference is present.
  - Does NOT reveal which sample means are different.
  - If there is no significance found, the test is concluded.
  - If there is a difference, do analyze a Post-hoc test (Bonferroni, Tukey, or Scheffe test).

Assumptions in ANOVA

- Conditions or Assumptions
  - The data are randomly sampled
  - The variances of each sample are assumed equal
  - The residuals are normally distributed

Hypothesis in ANOVA

- The null hypothesis is that the means are all equal

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

- The alternative hypothesis is that at least one of the means is different

$$H_1: at\ least\ one\ of\ them\ differs.$$

Why not Multiple Pairwise t-tests?

- With $t-test$, only 2 means at a time are compared
  - F test compares all means at once
- The more $t-tests$ that are conducted, more likely to find significant differences by chance alone (Type-I error).
- Many means to compare so more tests required.
  - For 10 means, 45 $t-tests$ required
- ANOVA greatly reduces Type-I errors as the number of treatments to compare increases.

Post-hoc pairwise comparison tests

Rule of Thumb

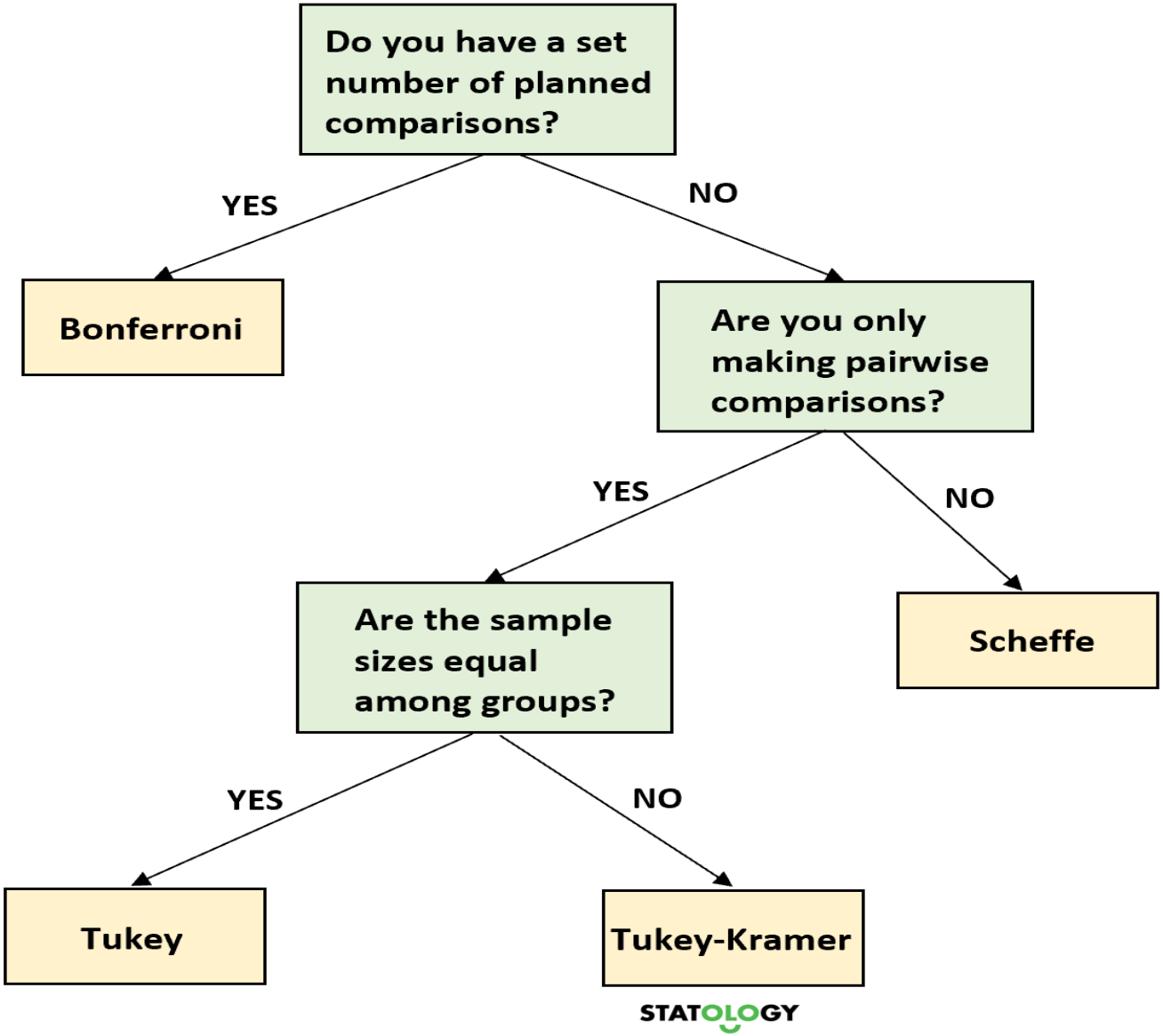
- Use the Tukey test when samples are equal in size.
- Use Scheffe's test if samples differ in size.
- These two tests are similar to a $t-test$ in that they do pairwise comparisons, but are adjusted so that they account for the fact that more than 2 means are compared.

Post-hoc pairwise comparison tests

Rule of Thumb

- Bonferroni correction is used to reduce the chances of obtaining false-positive results (type I errors).
- To perform a Bonferroni correction, divide the significance level (α) by the number of comparisons being made. For example, if 10 hypotheses are being tested, the new critical value would be $new\ \alpha = \frac{\alpha}{10}, e.g., new\ \alpha = \frac{0.05}{10} = 0.005.$
- Bonferroni has more power when the number of comparisons is small, whereas Tukey is more powerful when testing large numbers of means.
- Tukey's HSD is only applied to situations where you want to examine all possible pairwise comparisons, whereas Bonferroni correction can be applied to any set of hypothesis tests.

Which Post-hoc Test should we use?



Exercise

- *Sample_Dataset_2014.sav*
- Question: To test if there is a statistically significant difference in *sprint time* with respect to *smoking status*.
- Sprint time will serve as the dependent variable, and smoking status will act as the independent variable.
- First, import the data into SPSS.
  - Hint: ANOVA

Hypothesis in ANOVA

- The null hypothesis is that the means are all equal

$$H_0: \mu_{nonsmoker} = \mu_{past\ smoker} = \mu_{current\ smoker}$$

- The alternative hypothesis is that at least one of the means is different

$$H_1: at\ least\ one\ of\ them\ differs.$$

In SPSS: ANOVA

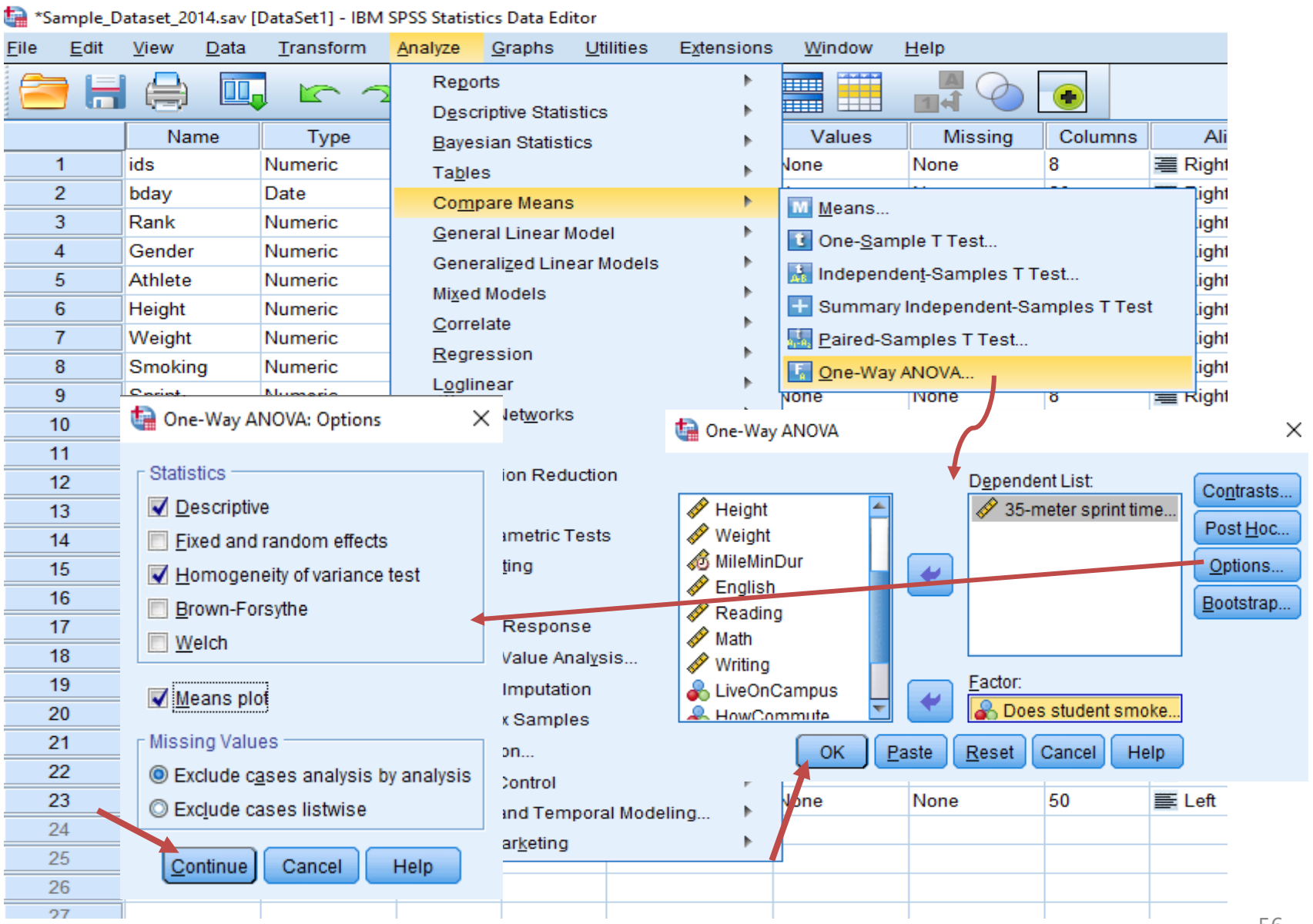
- To run a One-Way ANOVA in SPSS

  click Analyze > Compare Means > One-Way ANOVA.

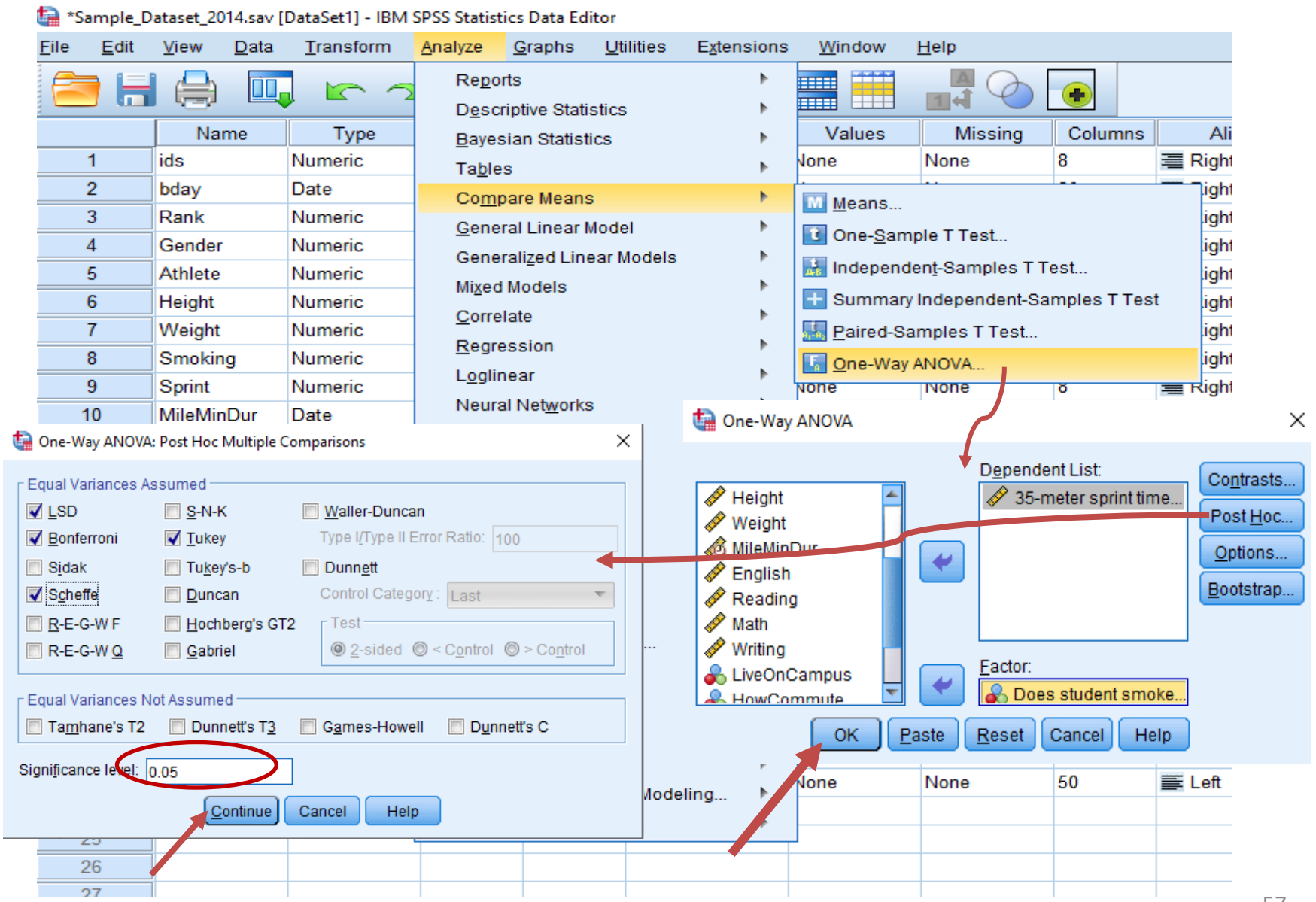
- Syntax:

```
ONEWAY Sprint BY Smoking
  /STATISTICS DESCRIPTIVES HOMOGENEITY
  /PLOT MEANS
  /MISSING ANALYSIS
  /POSTHOC=TUKEY SCHEFFE LSD BONFERRONI ALPHA(0.05).
```

ANOVA


*Sample_Dataset_2014.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Reports
Descriptive Statistics
Bayesian Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Means...
One-Sample T Test...
Independent-Samples T Test...
Summary Independent-Samples T Test
Paired-Samples T Test...
One-Way ANOVA...
Name Type Values Missing Columns
1 ids Numeric None None 8 Right
2 bday Date
3 Rank Numeric
4 Gender Numeric
5 Athlete Numeric
6 Height Numeric
7 Weight Numeric
8 Smoking Numeric
One-Way ANOVA: Options
Statistics
Descriptive
Fixed and random effects
Homogeneity of variance test
Brown-Forsythe
Welch
Means plot
Missing Values
Exclude cases analysis by analysis
Exclude cases listwise
Continue Cancel Help
One-Way ANOVA
Height
Weight
MileMinDur
English
Reading
Math
Writing
LiveOnCampus
Dependent List:
35-meter sprint time...
Factor:
Does student smoke...
Contrasts...
Post Hoc...
Options...
Bootstrap...
OK Paste Reset Cancel Help
23 None None 50 Left

ANOVA


*Sample_Dataset_2014.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Reports
Descriptive Statistics
Bayesian Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Means...
One-Sample T Test...
Independent-Samples T Test...
Summary Independent-Samples T Test
Paired-Samples T Test...
One-Way ANOVA...
Name Type Values Missing Columns
1 ids Numeric
2 bday Date
3 Rank Numeric
4 Gender Numeric
5 Athlete Numeric
6 Height Numeric
7 Weight Numeric
8 Smoking Numeric
9 Sprint Numeric
10 MileMinDur Date
One-Way ANOVA
Dependent List:
35-meter sprint time...
Factor:
Does student smoke...
Contrasts...
Post Hoc...
Options...
Bootstrap...
OK Paste Reset Cancel Help
Height
Weight
MileMinDur
English
Reading
Math
Writing
LiveOnCampus
One-Way ANOVA: Post Hoc Multiple Comparisons
Equal Variances Assumed
LSD
Bonferroni
Sidak
Scheffe
R-E-G-W F
R-E-G-W Q
S-N-K
Tukey
Tukey's-b
Duncan
Hochberg's GT2
Gabriel
Waller-Duncan
Type I/Type II Error Ratio: 100
Dunnett
Control Category: Last
Test
2-sided < Control > Control
Equal Variances Not Assumed
Tamhane's T2
Dunnett's T3
Games-Howell
Dunnett's C
Significance level: 0.05
Continue Cancel Help
Modeling...
None None 50 Left
26

ANOVA

Descriptives

35-meter sprint time (seconds)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Nonsmoker	261	6.41149	1.251783	.077483	6.25891	6.56406	4.503	9.597
Past smoker	33	6.83533	1.024415	.178328	6.47209	7.19858	4.889	8.549
Current smoker	59	7.12092	1.083500	.141060	6.83855	7.40328	5.295	9.475
Total	353	6.56968	1.233839	.065671	6.44053	6.69884	4.503	9.597

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
35-meter sprint time (seconds)	Based on Mean	2.415	2	350	.091
	Based on Median	2.322	2	350	.100
	Based on Median and with adjusted df	2.322	2	343.190	.100
	Based on trimmed mean	2.349	2	350	.097

ANOVA

ANOVA

35-meter sprint time (seconds)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	26.788	2	13.394	9.209	.000
Within Groups	509.082	350	1.455		
Total	535.870	352			

- Since $p < 0.05$, the null hypothesis is rejected. Therefore, we may conclude that there is a significant difference in *sprint time* with respect to *smoking status*.

Post-hoc Tests

Multiple Comparisons

Dependent Variable: 35-meter sprint time (seconds)

	(I) Does student smoke?	(J) Does student smoke?	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	Nonsmoker	Past smoker	-.423847	.222821	.140	-.94831	.10061
		Current smoker	-.709429*	.173856	.000	-1.11864	-.30022
	Past smoker	Nonsmoker	.423847	.222821	.140	-.10061	.94831
		Current smoker	-.285582	.262163	.521	-.90264	.33148
	Current smoker	Nonsmoker	.709429*	.173856	.000	.30022	1.11864
		Past smoker	.285582	.262163	.521	-.33148	.90264
Scheffe	Nonsmoker	Past smoker	-.423847	.222821	.165	-.97160	.12391
		Current smoker	-.709429*	.173856	.000	-1.13681	-.28205
	Past smoker	Nonsmoker	.423847	.222821	.165	-.12391	.97160
		Current smoker	-.285582	.262163	.553	-.93005	.35888
	Current smoker	Nonsmoker	.709429*	.173856	.000	.28205	1.13681
		Past smoker	.285582	.262163	.553	-.35888	.93005
LSD	Nonsmoker	Past smoker	-.423847	.222821	.058	-.86208	.01439
		Current smoker	-.709429*	.173856	.000	-1.05136	-.36750
	Past smoker	Nonsmoker	.423847	.222821	.058	-.01439	.86208
		Current smoker	-.285582	.262163	.277	-.80119	.23003
	Current smoker	Nonsmoker	.709429*	.173856	.000	.36750	1.05136
		Past smoker	.285582	.262163	.277	-.23003	.80119
Bonferroni	Nonsmoker	Past smoker	-.423847	.222821	.174	-.95985	.11216
		Current smoker	-.709429*	.173856	.000	-1.12765	-.29121
	Past smoker	Nonsmoker	.423847	.222821	.174	-.11216	.95985
		Current smoker	-.285582	.262163	.830	-.91623	.34506
	Current smoker	Nonsmoker	.709429*	.173856	.000	.29121	1.12765
		Past smoker	.285582	.262163	.830	-.34506	.91623

*. The mean difference is significant at the 0.05 level.

Kruskal-Wallis Test

- The nonparametric test makes no assumptions about the distribution of the data (e.g., ***normality***).
- More than 2 groups – a nonparametric alternative to the *one way ANOVA*.
- The Kruskal-Wallis test (also called the H test) is a nonparametric test that uses ranks of sample data from three or more independent populations.
- It is used to test the null hypothesis that the independent samples come from populations with the equal medians.
- **Hypothesis**

H_0: *The samples come from populations with equal medians*.

H_1: *The samples come from populations with medians that are not all equal*.

Kruskal-Wallis Test: Requirements

- Need at least three independent samples, all of which are randomly selected.
- Each sample has at least 5 observations.
- There is no requirement that the populations have a normal distribution or any other particular distribution.

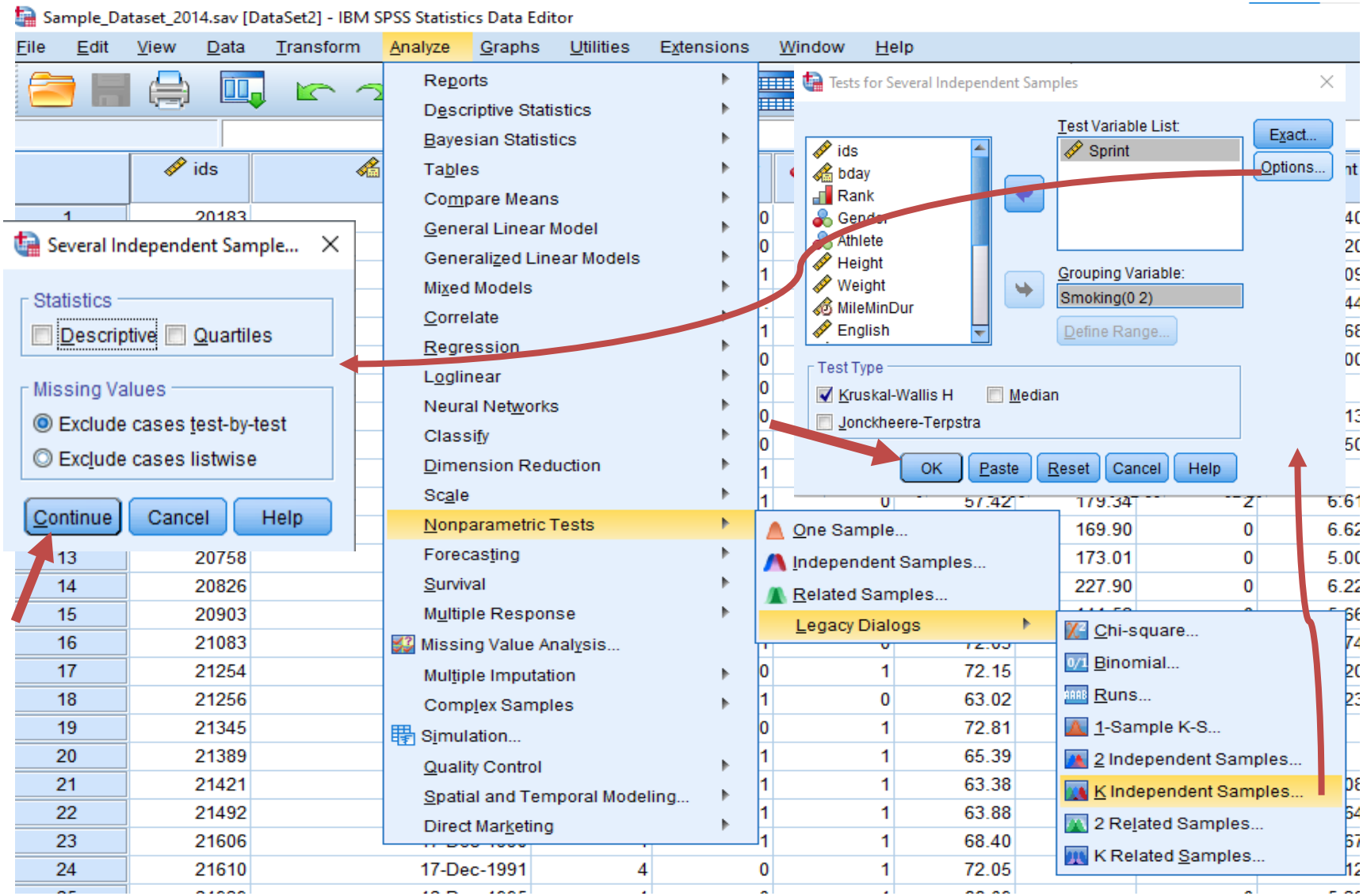
Example: Sample data 2014

▷ **Hypothesis**

H_0: *The populations of sprinters with sprint time come from the three groups of smokers with equal medians.*

H_1: *The three populations medians are not all equal.*

K-W Test: Sample Data set



K-W test

Kruskal-Wallis Test

Ranks

	Smoking	N	Mean Rank
Sprint	Nonsmoker	261	163.50
	Past Smoker	33	202.12
	Current Smoker	59	222.66
	Total	353	

Test Statistics[a,b]

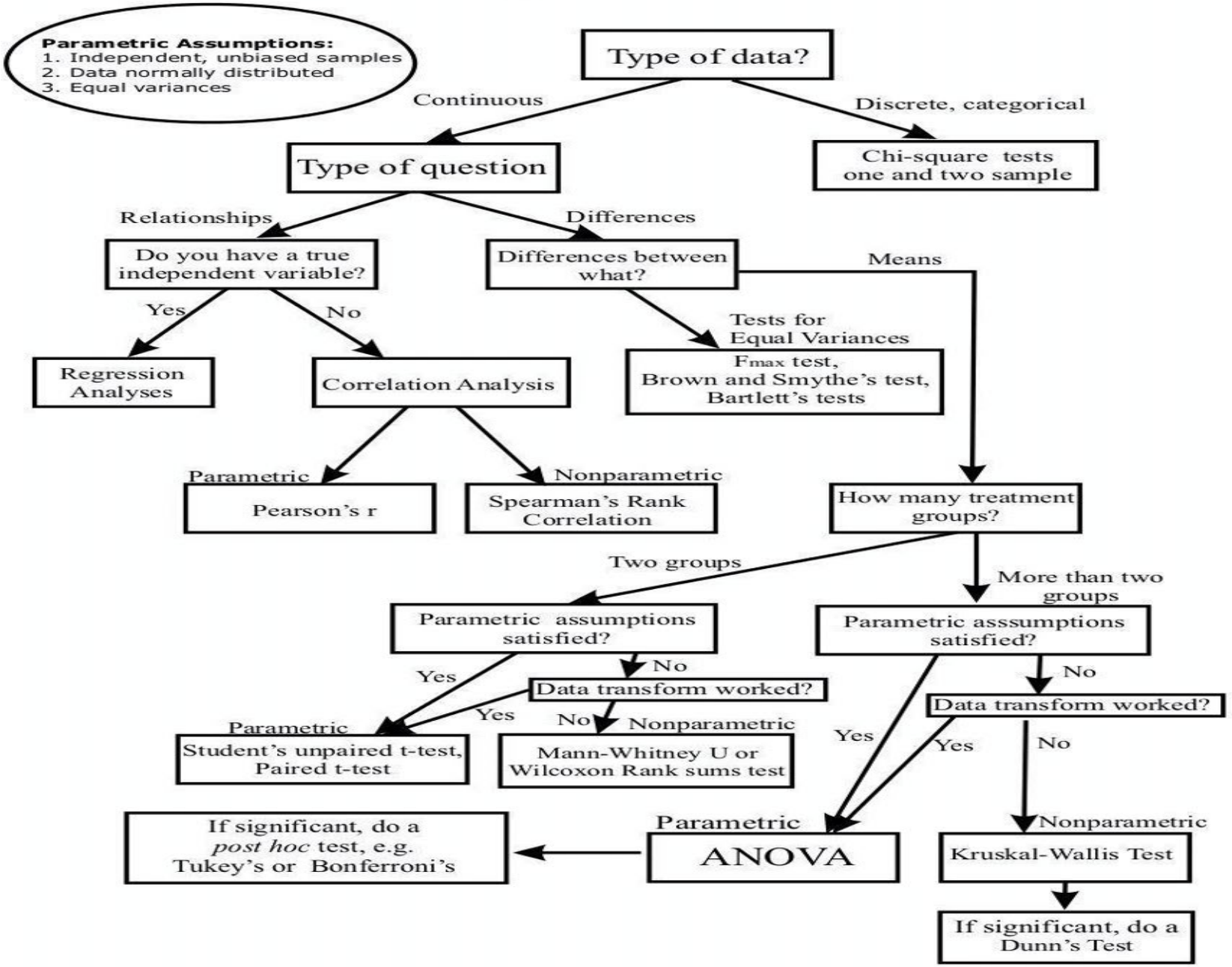
	Sprint
Kruskal-Wallis H	18.379
df	2
Asymp. Sig.	.000

a. Kruskal Wallis Test

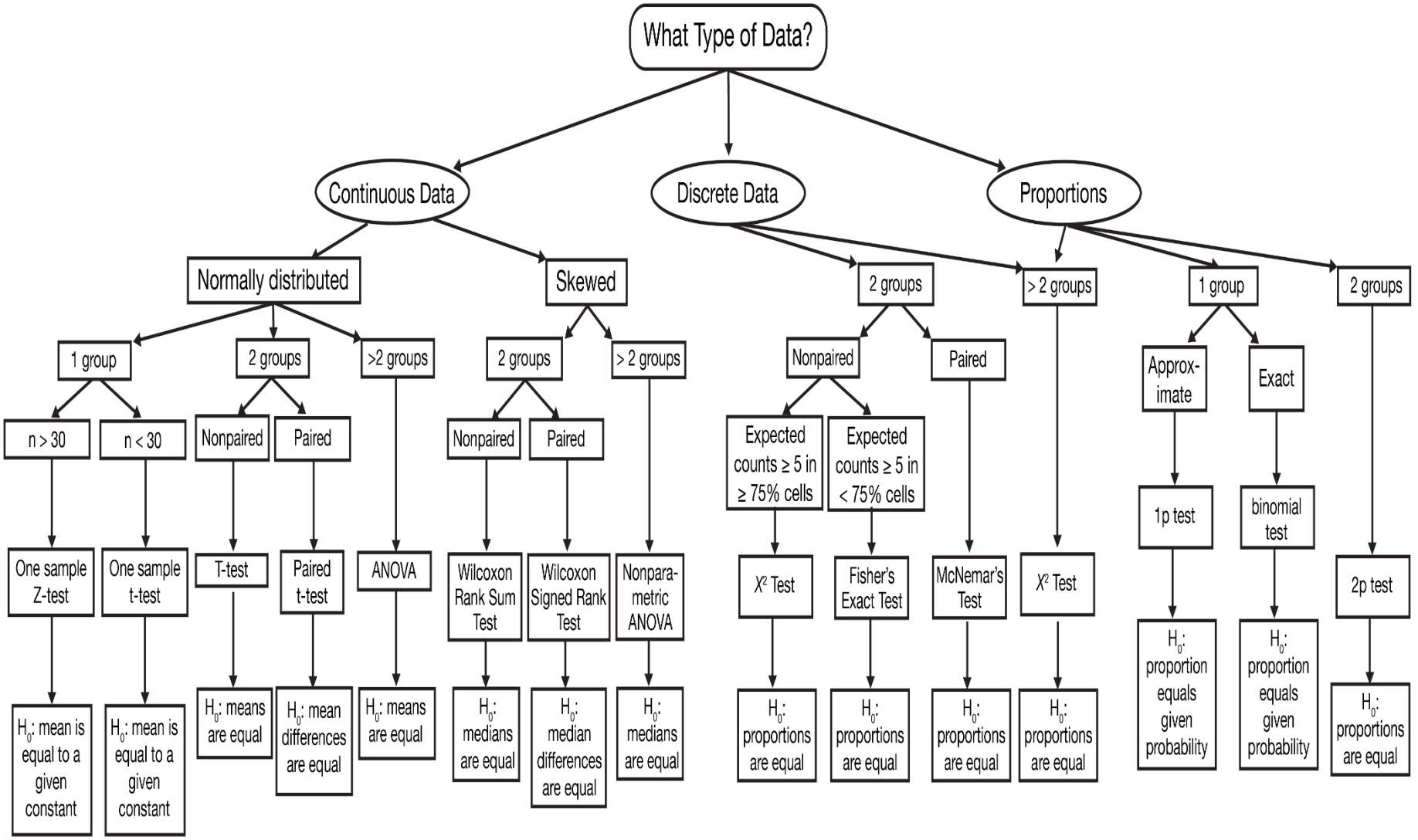
b. Grouping Variable: Smoking

- Since $p < 0.05$, the null hypothesis is rejected. Therefore, we may conclude that there is a significant difference in *sprint time* with respect to *smoking status*.

Flow Chart for Selecting Commonly Used Statistical Tests



Flow chart: Which test statistics should we use?



Thank You!