MICROSOFT EXCEL - CORRELATIONS & REGRESSIONS: HINTS

To compute the mean for HbA1c use AVERAGE predefines function and comma as separator.
To compute correlation coefficient: 2 ways are possible:

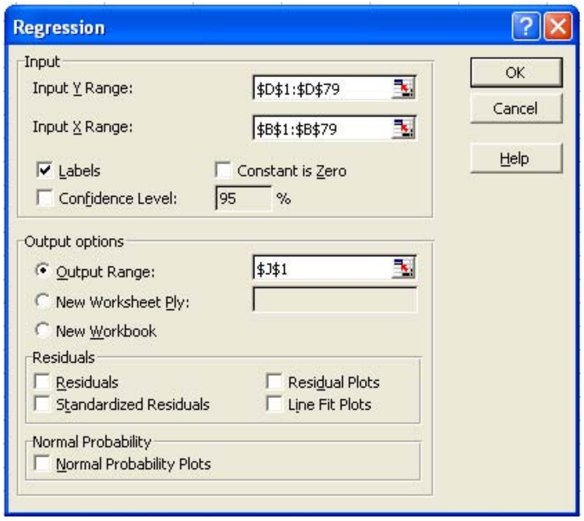
- CORREL predefine function: to Array1 select the range for first variable (DO NOT select labels; e.g. m-HbA1c) and to Array2 select the range for second variable (e.g. Measure of growth).
- **[Tools – Data Analysis - Correlation]**: to the *Input range* select the cells where the data for the quantitative variables are (include into selection the label of variable) and choose *Labels* in first row.

To interpret correlation coefficient (Colton rules for interpreting the correlation coefficient values):

- Correlation coefficient between -0.25 and +0.25 = little or no relationship;
- Correlation coefficient between 0.25 and 0.50 (or - 0.25 and - 0.50) = weak to acceptable degree of association;
- Correlation coefficient between 0.50 and 0.75 (or - 0.50 and - 0.75) = moderate to good association;
- Correlation coefficient higher than 0.75 (or lower than - 0.75) = a very good level of association.

To perform a simple linear regression analysis:

- **[Tools – Data Analysis - Regression]**
- In the Regression window:
  - Input Y Range: select the range of dependent variable
  - Input X Range: select the range of independent variable
  - Click on Labels
  - Output range: click on one cell in the Simple Regression sheet (to the right of the data).



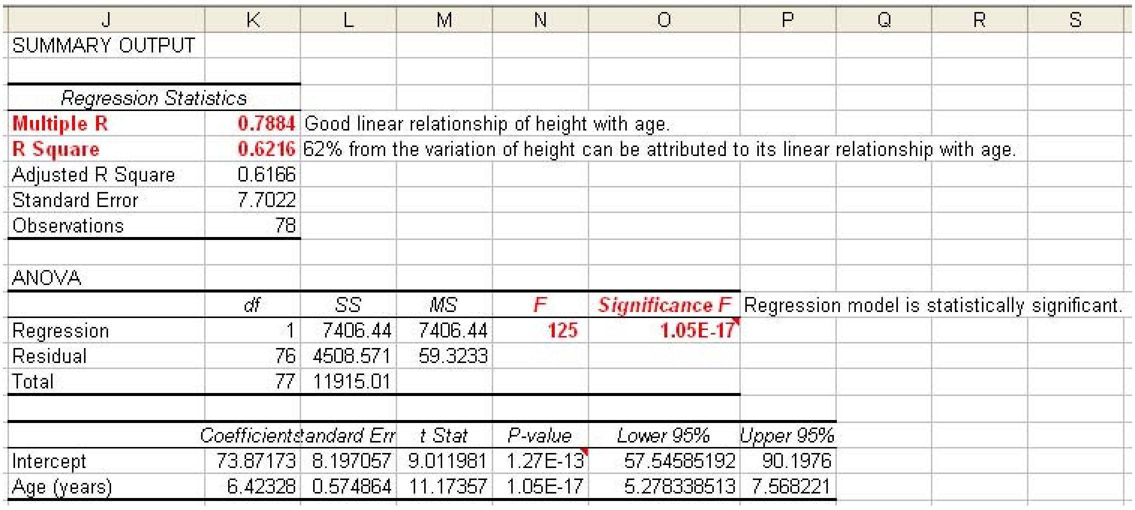
Input Y Range: $D$1:$D$79
Input X Range: $B$1:$B$79
☑ Labels □ Constant is Zero
□ Confidence Level: 95 %

● Output Range: $J$1
○ New Worksheet Ply:
○ New Workbook

Residuals
□ Residuals □ Residual Plots
□ Standardized Residuals □ Line Fit Plots

Normal Probability
□ Normal Probability Plots

To interpret the results of the simple linear regression:



J	K	L	M	N	O	P	Q	R	S
SUMMARY OUTPUT									
Regression Statistics									
Multiple R	**0.7884**	Good linear relationship of height with age.							
R Square	**0.6216**	62% from the variation of height can be attributed to its linear relationship with age.							
Adjusted R Square	0.6166								
Standard Error	7.7022								
Observations	78								
ANOVA									
	df	*SS*	*MS*	*F*	*Significance F*	Regression model is statistically significant.			
Regression	1	7406.44	7406.44	**125**	**1.05E-17**				
Residual	76	4508.571	59.3233						
Total	77	11915.01							
	Coefficients	*Standard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*			
Intercept	73.87173	8.197057	9.011981	1.27E-13	57.54585192	90.1976			
Age (years)	6.42328	0.574864	11.17357	1.05E-17	5.278338513	7.568221			

Reading the output of regression analysis:

- Regression statistics:
  - Multiple R: is the correlation between the predictor variable(s) and the criterion variable (for one variable represent the Pearson correlation coefficient, expressing the linear relationship between weight and cranial perimeter);
  - R square (the coefficient of determination): It represents the proportion of variation in Y that is explained by its linear relationship with X;
  - Adjusted R Squared: provides a better estimation of R2;
  - Standard error: is the standard error of the estimate and is interpreted as the average error in predicting Y by means of the regression equation;
  - Observations: refers to the number of subjects included in the analysis.
- ANOVA: Regression analysis includes a test of the hypothesis that the slop of the regression line is equal to 0. If the slope is significantly different from), then it can be conclude that there is a statistically significant linear relationship between weight and cranial perimeter:
  - Regression: this component represents the variation in weight that is explained by its linear relationship with cranial perimeter;
  - Residual: residual variation represent the variation in weight that is not explained by cranial perimeter;
  - Total: refers to “total variation”.
  - For each source of variation, the output provides degrees of freedom (df) and sum of squares (SS). The F value is obtained dividing the mean square (MS) regression by MS residual. The significance of F is the probability (P-value) associated with the obtained value of F.
  - Coefficients: The information provided at the bottom of the output refers to the coefficients in the regression equation.
  - Intercept: the intercept is 73.87. The t-Stat refers to a test of the hypothesis that the intercept is significantly different from zero. The P-value is the probability associated with obtained t statistic. The 95% confidence interval boundaries are applied to form the 95% CI around the intercept.
  - Age (years): The slop of the regression line is 6.42. The t-Stat refers to a test of the hypothesis that the slope is significantly different from zero. The p-value is the probability associated with the obtained t statistic.

To create a Scatter:

- **[Insert – Chart… – Scatter]**
- Chart Wizard - Step 1 of 4-Chart Type:

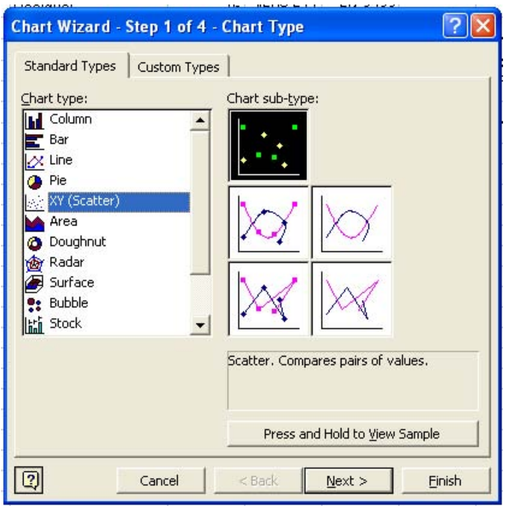
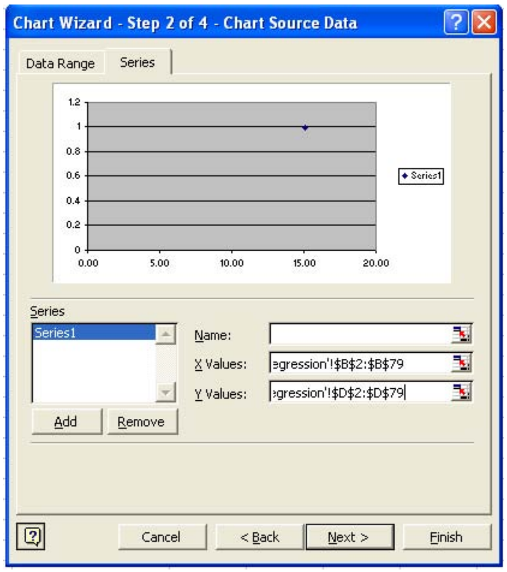


Chart type: XY (Scatter)

Scatter. Compares pairs of values.

- Chart Wizard - Step 2 of 4-Chart Type: select the range corresponding to Age for X Values and to Height for Y Values

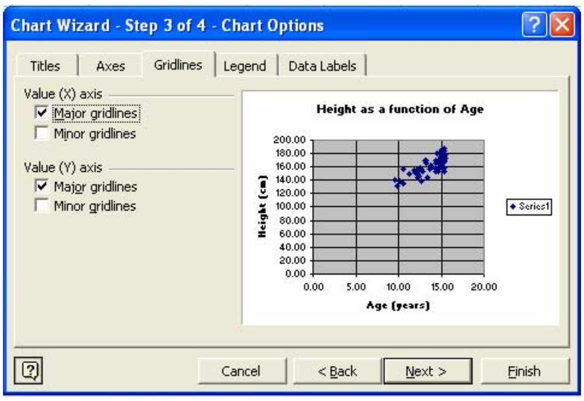


Name:

X Values: =gression'!$B$2:$B$79

Y Values: =gression'!$D$2:$D$79

- Chart Wizard - Step 3 of 4-Chart Type.
  - Chart title: Height as a function of Age
  - Value (X) axis: Age (years)
  - Value (Y) axis: Height (cm)
  - Click the Gridlines tab at the top of the dialog box. In the Value (X) axis section, click in the box to the left of Major gridlines so that these gridlines will appear in the scatterplot.



Value (X) axis
☑ Major gridlines
☐ Minor gridlines

Value (Y) axis
☑ Major gridlines
☐ Minor gridlines

- Click the Legend tab at the top of the dialog box and remove the name of series (click in the box to the left of Show legend to remove the check mark that appears there):

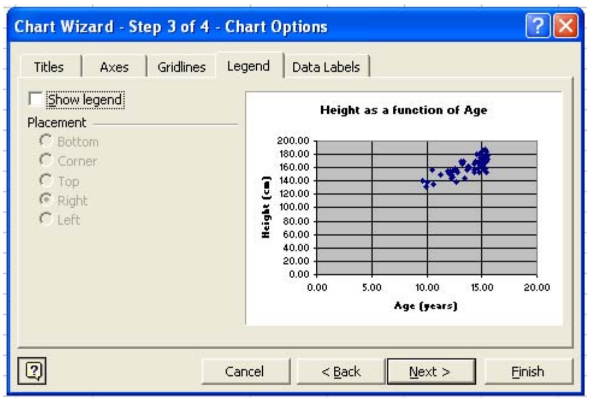


Chart Wizard - Step 3 of 4 - Chart Options

Titles | Axes | Gridlines | Legend | Data Labels

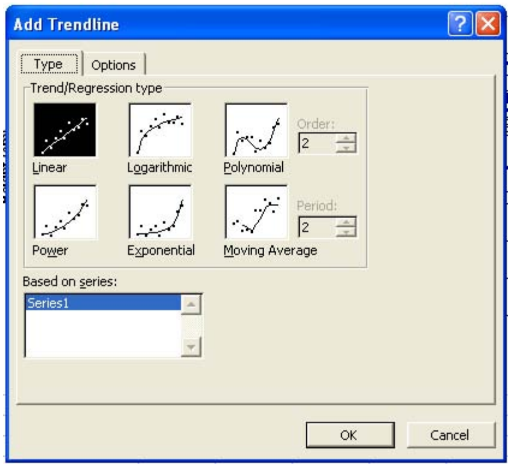
□ Show legend

Placement

○ Bottom
○ Corner
○ Top
● Right
○ Left

- Chart Wizard - Step 4 of 4-Chart Type. For example, we would like to display the chart in the same sheet. So, select *As object in* and click Finish.

- Add Trendline on chart:
  - Select the data series for the trendline by clicking one of its markers;
  - Right-click and choose Add Trendline from the shortcut menu;
  - In the Add Trendline dialog box, pick a trend/regression type as is show in the image bellow:



Add Trendline

Type | Options

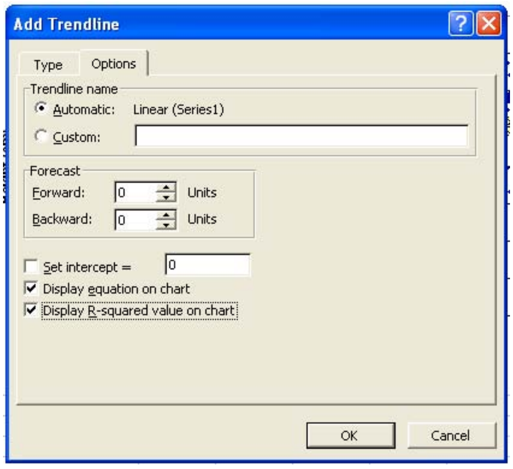
Trend/Regression type

Linear (selected) | Logarithmic | Polynomial | Order: 2

Power | Exponential | Moving Average | Period: 2

Based on series: Series1

  - Click the Options tab of the Add Trendline dialog box and change options to display equation and R-squared value on the chart. Select Display equation on chart and Display R-squared value on chart.



Add Trendline

Type | Options

Trendline name

● Automatic: Linear (Series1)

○ Custom:

Forecast

Forward: 0 Units

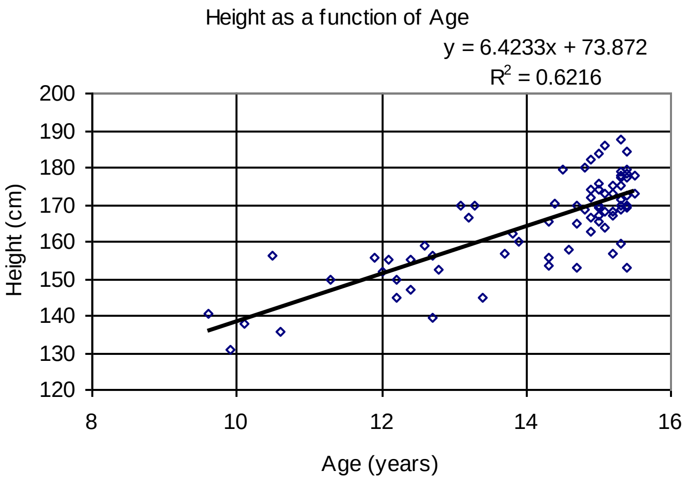
Backward: 0 Units

☐ Set intercept = 0

☑ Display equation on chart

☑ Display R-squared value on chart

- Your chart will be as in the image bellow:



To perform multiple linear regression analysis:

- The steps are the same as for simple linear regression with the exception:
    - Input X range: age, height and weight. These three variables must be in consecutive columns in order to can be selected in one selection.
    - Input Y range: HbA1c.