For Lectures 04 and 05
Association between Numerical Variables:
Scatterplot, Covariance, Correlation, Linear Regression

In this lecture:

This handout uses the **CEO Compensation 2008 Forbes.xlsx** data set.

For help with Excel, go to: http://office.microsoft.com/en-us/excel-help

← go back to Table of Contents

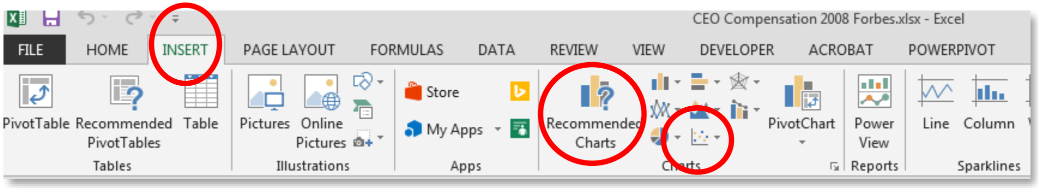
Graphical Technique

Scatterplot

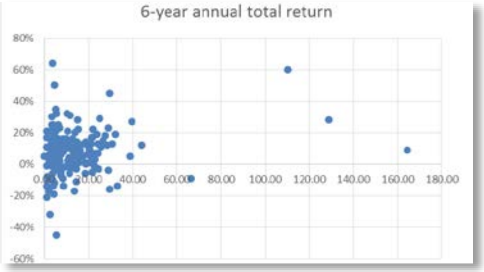
Create a scatterplot for *6-year average compensation* (Column Q) and *6-year annual total return* (Column R).

SCATTERPLOT – Example 1 - in basic Excel:

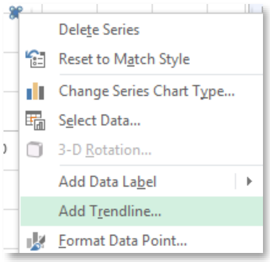
- Highlight columns Q and R. The first column should be the one that will be on the horizontal axis.
- **INSERT → Recommended Charts → All Charts → X Y (Scatter)**.



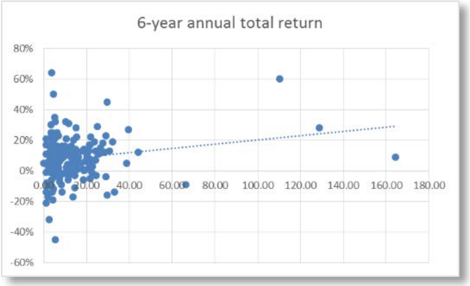
- Obtain the following scatterplot:



- To draw a straight line (a linear regression line) through the scatterplot, right-click on any point, select "**Add Trendline**", click OK.



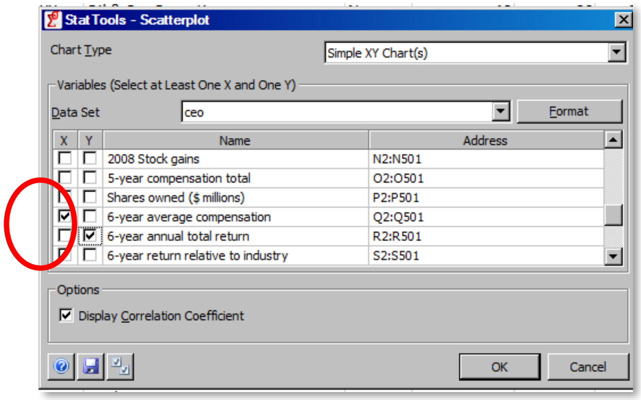
- The positive slope means a positive relationship between the two variables.



← go back to Table of Contents

SCATTERPLOT – Example 2 - in StatTools:

- Load **StatTools**. Highlight all data. Click on **Data Set Manager**, click Yes, rename data, click OK.
- Go to **StatTools → Summary Graphs → Scatterplot**.
- Click on *6-year average compensation* as X and *6-year annual total return* as Y. Click OK.



StatTools - Scatterplot

Chart Type: Simple XY Chart(s)

Variables (Select at Least One X and One Y)

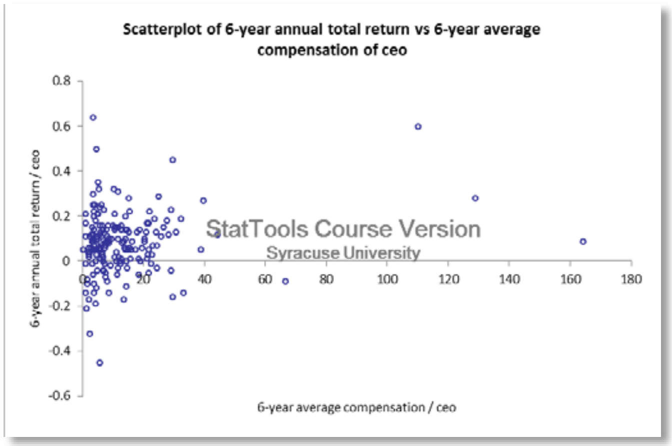
Data Set: ceo

X	Y	Name	Address
□	□	2008 Stock gains	N2:N501
□	□	5-year compensation total	O2:O501
□	□	Shares owned ($ millions)	P2:P501
☑	□	6-year average compensation	Q2:Q501
□	☑	6-year annual total return	R2:R501
□	□	6-year return relative to industry	S2:S501

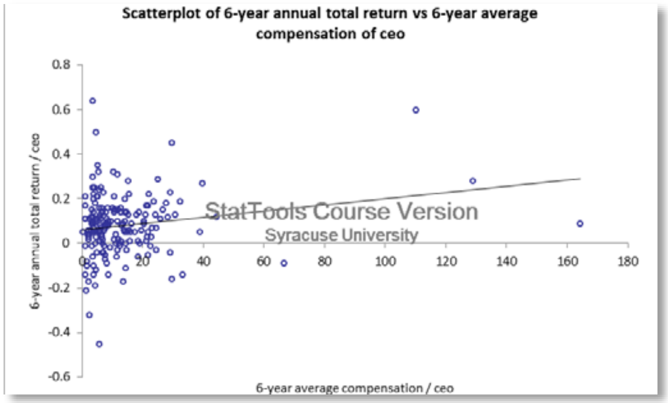
Options

☑ Display Correlation Coefficient

- Obtain the following scatterplot:



- To draw a straight line (a linear regression line) through the scatterplot, right-click on any point, select "**Add Trendline**", click OK.



← go back to Table of Contents

Numerical Techniques

Covariance and correlation

COVARIANCE AND CORRELATION – Example 1:

- Computing covariance between *6-year average compensation* (Column Q) and *6-year annual total return* (Column R).

$$\boxed{\mathrm{cov}(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}}$$

=covariance.s(Q:Q,R:R). The answer is: 0.4888.

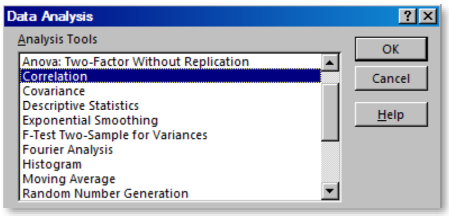
- Computing correlation coefficient between *6-year average compensation* (Column Q) and *6-year annual total return* (Column R).

$$\boxed{r=\frac{\mathrm{cov}(X,Y)}{s_x\cdot s_y}}$$

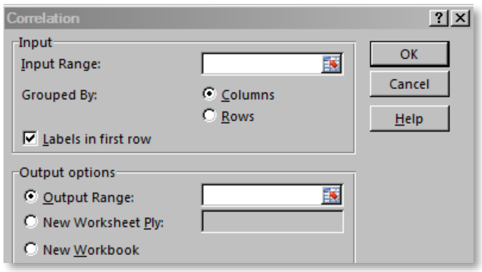
=correl(Q:Q,R:R). The answer is: 0.1880.

COVARIANCE AND CORRELATION – Example 2:

- Create a table of correlations, using **Data Analysis ToolPak.**
- If the Data Analysis add-in is not activated, go to **FILE** → **Options** → **Add-Ins** → **Analysis ToolPak** → **GO** → click on **Analysis ToolPak** → **OK**.
- Highlight the columns for each pair of variables of which you want to compute correlation coefficients.
- Note: The columns must be adjacent to each other; all columns must contain numerical variables.
- For example, highlight Column G (*years as company CEO*) through Column R (*6-year annual total return*).
- **DATA** → **Data Analysis** → **Correlation**:

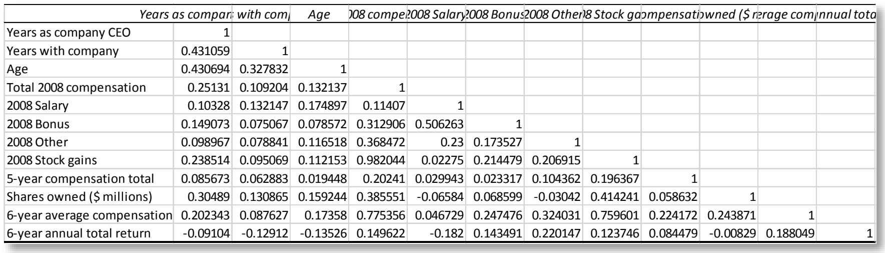


- In "**Input Range**" enter columns G through R (or you can just type **G:R**).
- Put a check mark next to "**Labels in First Row**."
- Specify output range.
- Click OK.



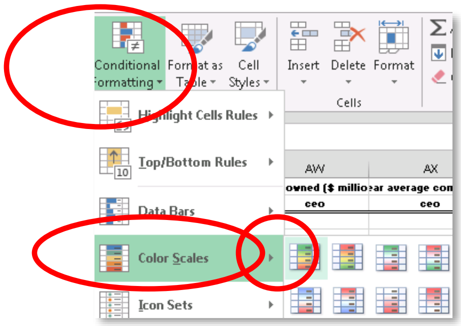
← go back to Table of Contents

- The following correlation table will appear:

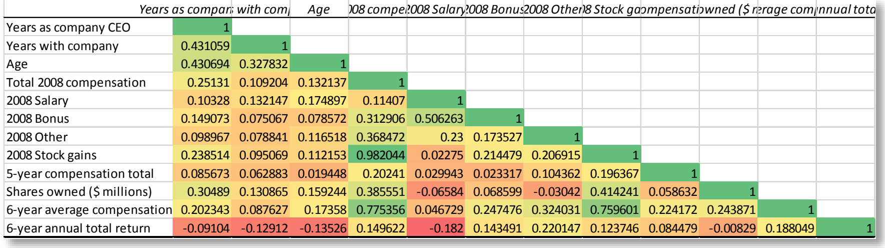


	Years as compan	with com	Age	)08 compe	2008 Salary	2008 Bonus	2008 Other	8 Stock ga	ompensati	wned ($ n	erage com	nnual tota
Years as company CEO	1											
Years with company	0.431059	1										
Age	0.430694	0.327832	1									
Total 2008 compensation	0.25131	0.109204	0.132137	1								
2008 Salary	0.10328	0.132147	0.174897	0.11407	1							
2008 Bonus	0.149073	0.075067	0.078572	0.312906	0.506263	1						
2008 Other	0.098967	0.078841	0.116518	0.368472	0.23	0.173527	1					
2008 Stock gains	0.238514	0.095069	0.112153	0.982044	0.02275	0.214479	0.206915	1				
5-year compensation total	0.085673	0.062883	0.019448	0.20241	0.029943	0.023317	0.104362	0.196367	1			
Shares owned ($ millions)	0.30489	0.130865	0.159244	0.385551	-0.06584	0.068599	-0.03042	0.414241	0.058632	1		
6-year average compensation	0.202343	0.087627	0.17358	0.775356	0.046729	0.247476	0.324031	0.759601	0.224172	0.243871	1	
6-year annual total return	-0.09104	-0.12912	-0.13526	0.149622	-0.182	0.143491	0.220147	0.123746	0.084479	-0.00829	0.188049	1

- We can apply **conditional formatting** to the table to better see high and low correlations. Highlight the table area, then go to **HOME → Conditional Formatting → Color Scales →** select the first option.



- The table of correlations will now look like this:

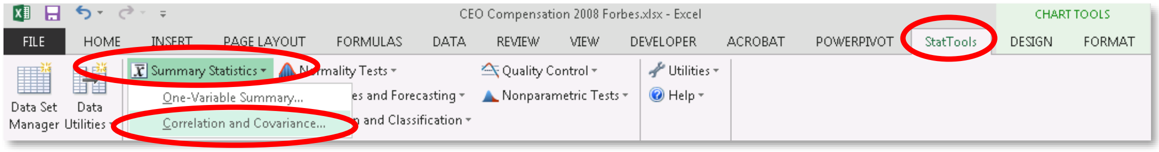


	Years as compan	with com	Age	)08 compe	2008 Salary	2008 Bonus	2008 Other	8 Stock ga	ompensati	wned ($ n	erage com	nnual tota
Years as company CEO	1											
Years with company	0.431059	1										
Age	0.430694	0.327832	1									
Total 2008 compensation	0.25131	0.109204	0.132137	1								
2008 Salary	0.10328	0.132147	0.174897	0.11407	1							
2008 Bonus	0.149073	0.075067	0.078572	0.312906	0.506263	1						
2008 Other	0.098967	0.078841	0.116518	0.368472	0.23	0.173527	1					
2008 Stock gains	0.238514	0.095069	0.112153	0.982044	0.02275	0.214479	0.206915	1				
5-year compensation total	0.085673	0.062883	0.019448	0.20241	0.029943	0.023317	0.104362	0.196367	1			
Shares owned ($ millions)	0.30489	0.130865	0.159244	0.385551	-0.06584	0.068599	-0.03042	0.414241	0.058632	1		
6-year average compensation	0.202343	0.087627	0.17358	0.775356	0.046729	0.247476	0.324031	0.759601	0.224172	0.243871	1	
6-year annual total return	-0.09104	-0.12912	-0.13526	0.149622	-0.182	0.143491	0.220147	0.123746	0.084479	-0.00829	0.188049	1

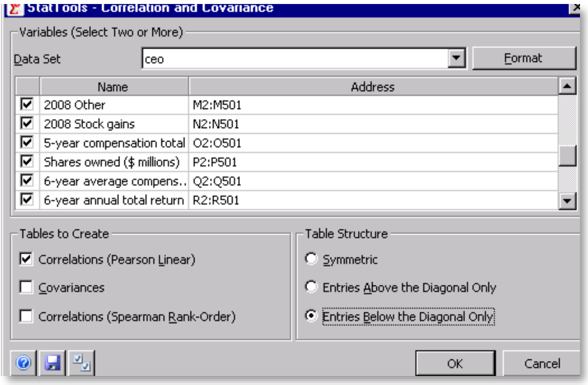
← go back to Table of Contents

COVARIANCE AND CORRELATION – Example 3:

- Create a table of correlations using **StatTools**:
- Make sure StatTools is running. Highlight your data, go to **StatTools → Data Set Manager →** click Yes → rename data, click OK.
- **StatTools → Summary Statistics → Correlation and Covariance**.



- A small screen will appear. Click on all variables in columns G through R.
- On bottom left, click "**Correlations**." On bottom right, click "**Entries below the Diagonal only**" (or you can pick Symmetric if you like).
- Click OK.


StatTools - Correlation and Covariance

Variables (Select Two or More)

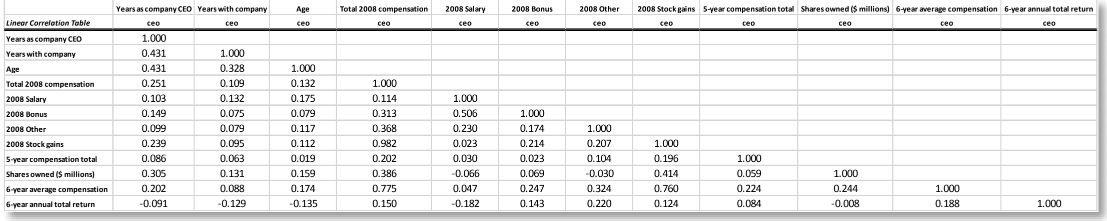
Data Set: ceo [Format]

	Name	Address
☑	2008 Other	M2:M501
☑	2008 Stock gains	N2:N501
☑	5-year compensation total	O2:O501
☑	Shares owned ($ millions)	P2:P501
☑	6-year average compens..	Q2:Q501
☑	6-year annual total return	R2:R501

Tables to Create
☑ Correlations (Pearson Linear)
□ Covariances
□ Correlations (Spearman Rank-Order)

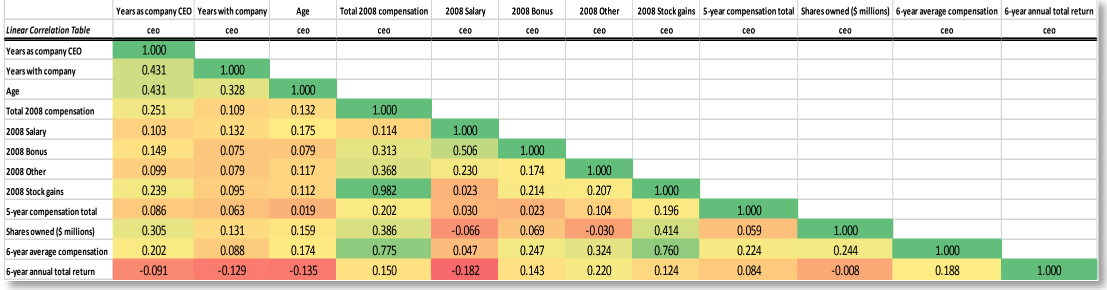
Table Structure
○ Symmetric
○ Entries Above the Diagonal Only
● Entries Below the Diagonal Only

- The following table of correlations will appear:



	Years as company CEO	Years with company	Age	Total 2008 compensation	2008 Salary	2008 Bonus	2008 Other	2008 Stock gains	5-year compensation total	Shares owned ($ millions)	6-year average compensation	6-year annual total return
Linear Correlation Table	ceo	ceo	ceo	ceo	ceo	ceo	ceo	ceo	ceo	ceo	ceo	ceo
Years as company CEO	1.000											
Years with company	0.431	1.000										
Age	0.431	0.328	1.000									
Total 2008 compensation	0.251	0.109	0.132	1.000								
2008 Salary	0.103	0.132	0.175	0.114	1.000							
2008 Bonus	0.149	0.075	0.079	0.313	0.506	1.000						
2008 Other	0.099	0.079	0.117	0.368	0.230	0.174	1.000					
2008 Stock gains	0.239	0.095	0.112	0.982	0.023	0.214	0.207	1.000				
5-year compensation total	0.086	0.063	0.019	0.202	0.030	0.023	0.104	0.196	1.000			
Shares owned ($ millions)	0.305	0.131	0.159	0.386	-0.066	0.069	-0.030	0.414	0.059	1.000		
6-year average compensation	0.202	0.088	0.174	0.775	0.047	0.247	0.324	0.760	0.224	0.244	1.000	
6-year annual total return	-0.091	-0.129	-0.135	0.150	-0.182	0.143	0.220	0.124	0.084	-0.008	0.188	1.000

- We can apply **conditional formatting** to the table to better see high and low correlations. Highlight the table area, then go to **HOME → Conditional Formatting → Color Scales →** select the first option.
- The correlations table will now look like this:



	Years as company CEO	Years with company	Age	Total 2008 compensation	2008 Salary	2008 Bonus	2008 Other	2008 Stock gains	5-year compensation total	Shares owned ($ millions)	6-year average compensation	6-year annual total return
Linear Correlation Table	ceo	ceo	ceo	ceo	ceo	ceo	ceo	ceo	ceo	ceo	ceo	ceo
Years as company CEO	1.000											
Years with company	0.431	1.000										
Age	0.431	0.328	1.000									
Total 2008 compensation	0.251	0.109	0.132	1.000								
2008 Salary	0.103	0.132	0.175	0.114	1.000							
2008 Bonus	0.149	0.075	0.079	0.313	0.506	1.000						
2008 Other	0.099	0.079	0.117	0.368	0.230	0.174	1.000					
2008 Stock gains	0.239	0.095	0.112	0.982	0.023	0.214	0.207	1.000				
5-year compensation total	0.086	0.063	0.019	0.202	0.030	0.023	0.104	0.196	1.000			
Shares owned ($ millions)	0.305	0.131	0.159	0.386	-0.066	0.069	-0.030	0.414	0.059	1.000		
6-year average compensation	0.202	0.088	0.174	0.775	0.047	0.247	0.324	0.760	0.224	0.244	1.000	
6-year annual total return	-0.091	-0.129	-0.135	0.150	-0.182	0.143	0.220	0.124	0.084	-0.008	0.188	1.000

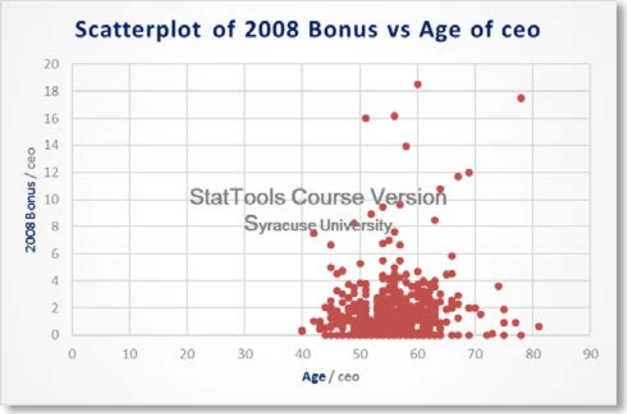
← go back to Table of Contents

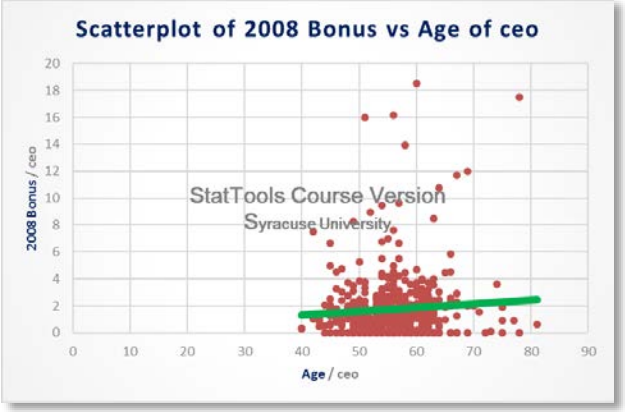
Linear regression

Simple linear regression

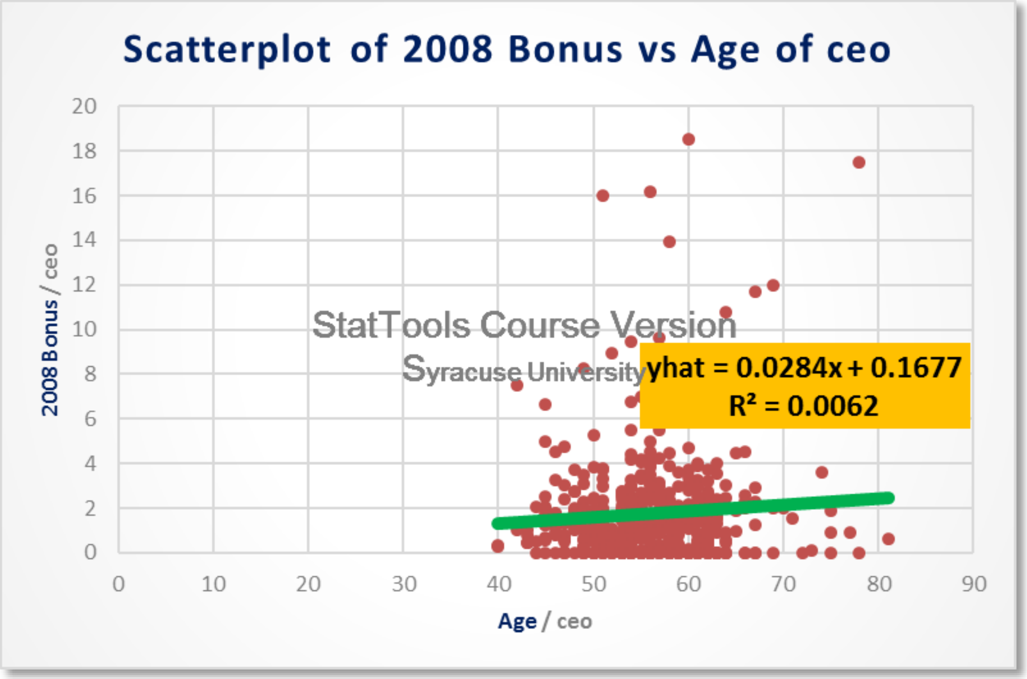
SIMPLE LINEAR REGRESSION – Example 1 – using scatterplot trendline:

- Suppose we want to predict (explain) CEOs' bonuses using CEOs' ages. The scatterplot would look like this (left). Right-click on any point, select **Add Trendline** (right).





- Right-click on the trend line, select **Format Trendline**. On the bottom of **Trendline Options**, click on **Display Equation on chart** and **Display R-squared value on chart**. Change "y" in the equation to "yhat".



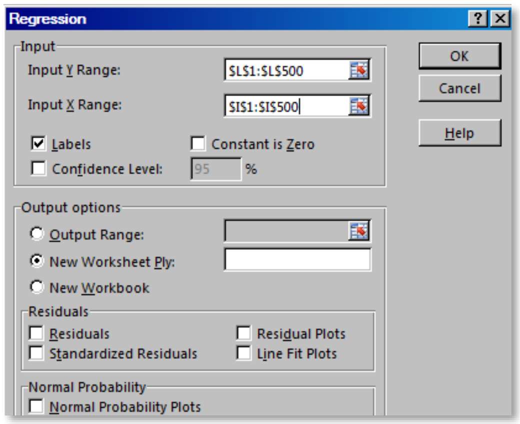
← go back to Table of Contents

SIMPLE LINEAR REGRESSION – Example 2 – using Excel commands:

- Suppose we want to predict (explain) CEOs' bonuses using CEOs' ages. Here, *Age* (Column I) is the explanatory variable, and *Bonus* (Column L) is the dependent variable.
- Use the following Excel commands to compute intercept, slope, and R-squared:
    - Intercept: **=INTERCEPT(y range, x range)**
    - Slope: **=SLOPE(y range, x range)**
    - R^2: **=correl(y range, x range)^2**
- *Note*: You CANNOT use these commands to compute intercept, slope, and R^2 for multiple regressions. These only work for simple regressions (one predictor).

SIMPLE LINEAR REGRESSION – Example 3 – using Data Analysis ToolPak:

- Suppose we want to predict (explain) CEOs' bonuses using CEOs' ages. Here, *Age* (Column I) is the explanatory variable, and *Bonus* (Column L) is the dependent variable.
- **DATA → Data Analysis → Regression.**
- Select appropriate ranges for Y (dependent variable: *Bonus*) and X (explanatory variable: *Age*). If labels are included in the ranges, then put a check mark in "Labels." Click OK.

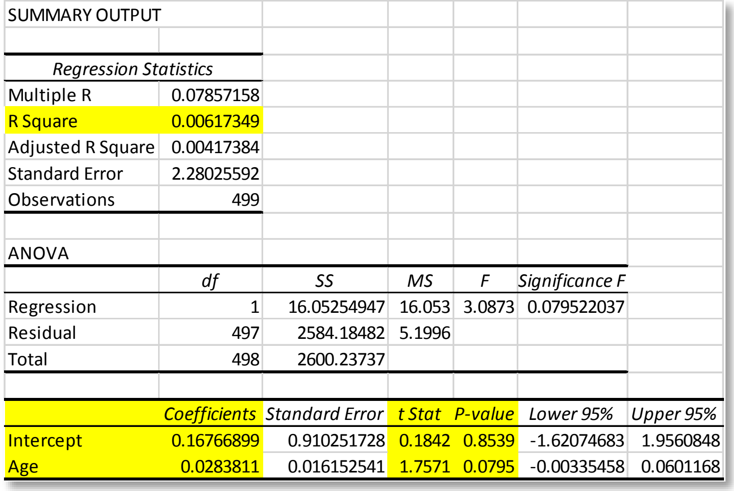


Input
Input Y Range: $L$1:$L$500
Input X Range: $I$1:$I$500
☑ Labels □ Constant is Zero
□ Confidence Level: 95 %

Output options
○ Output Range:
● New Worksheet Ply:
○ New Workbook

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

Normal Probability
□ Normal Probability Plots



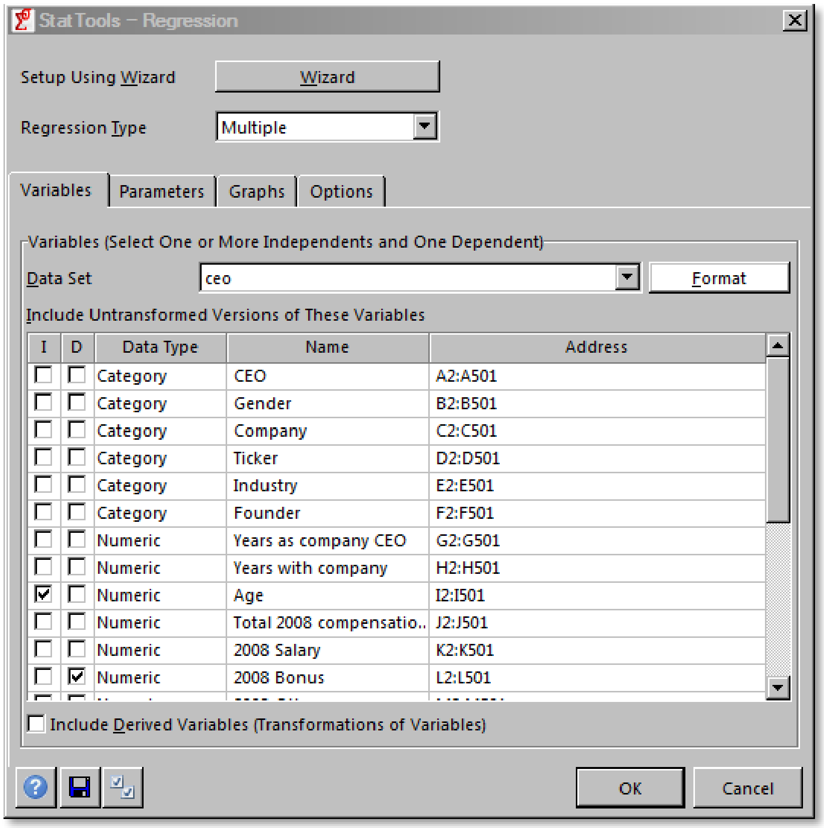
SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.07857158					
R Square	0.00617349					
Adjusted R Square	0.00417384					
Standard Error	2.28025592					
Observations	499					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	16.05254947	16.053	3.0873	0.079522037	
Residual	497	2584.18482	5.1996			
Total	498	2600.23737				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.16766899	0.910251728	0.1842	0.8539	-1.62074683	1.9560848
Age	0.0283811	0.016152541	1.7571	0.0795	-0.00335458	0.0601168

- *Note*: Excel will give you an error message if: (1) the columns are of different lengths, (2) there are missing values.

← go back to Table of Contents

SIMPLE LINEAR REGRESSION – Example 4 – using StatTools:

- **StatTools → Regression and Classification → Regression.**
- Specify: I = *Age*, D = *2008 Bonus*. "I" stands for independent (or explanatory) variable, and "D" stands for dependent variable. Click OK.



StatTools – Regression

Setup Using Wizard: Wizard

Regression Type: Multiple

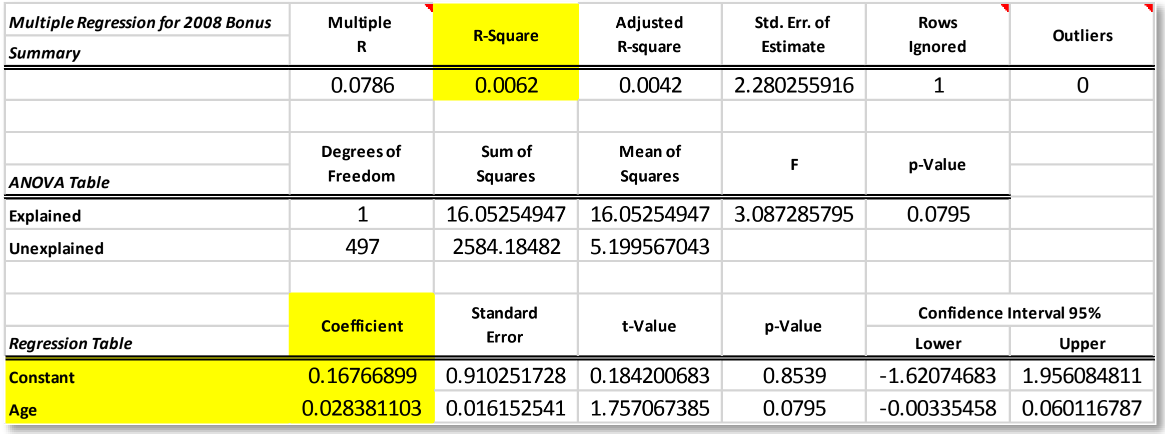
Variables (Select One or More Independents and One Dependent)

Data Set: ceo

Include Untransformed Versions of These Variables

I	D	Data Type	Name	Address
□	□	Category	CEO	A2:A501
□	□	Category	Gender	B2:B501
□	□	Category	Company	C2:C501
□	□	Category	Ticker	D2:D501
□	□	Category	Industry	E2:E501
□	□	Category	Founder	F2:F501
□	□	Numeric	Years as company CEO	G2:G501
□	□	Numeric	Years with company	H2:H501
☑	□	Numeric	Age	I2:I501
□	□	Numeric	Total 2008 compensatio..	J2:J501
□	□	Numeric	2008 Salary	K2:K501
□	☑	Numeric	2008 Bonus	L2:L501

□ Include Derived Variables (Transformations of Variables)



Multiple Regression for 2008 Bonus *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.0786	0.0062	0.0042	2.280255916	1	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	1	16.05254947	16.05254947	3.087285795	0.0795	
Unexplained	497	2584.18482	5.199567043			
	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95%	
Regression Table					Lower	Upper
Constant	0.16766899	0.910251728	0.184200683	0.8539	-1.62074683	1.956084811
Age	0.028381103	0.016152541	1.757067385	0.0795	-0.00335458	0.060116787

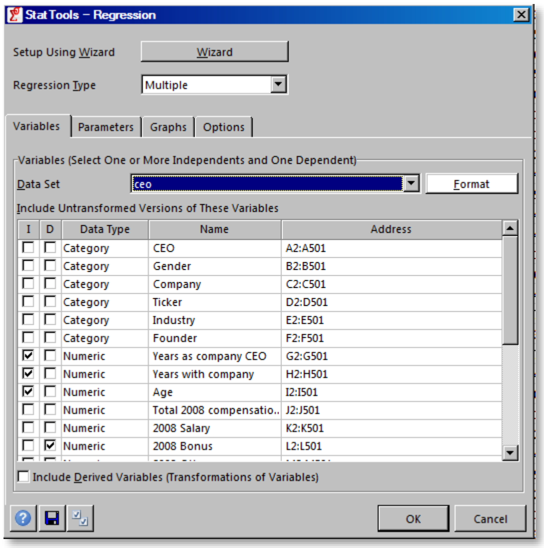
← go back to Table of Contents

Multiple linear regression

Suppose we want to see how *Bonus* (Column L) is explained (predicted) by 3 variables:

1) *Years as company CEO* (Column G),
2) *Years with company* (Column H),
3) *Age* (Column I)
   – all in one regression model.

- The steps are the same as before (Example 3 and Example 4 above). You need to include multiple X variables instead of one.
- *Note 1*: If using **Data Analysis ToolPak**, the range of X variables should include columns that are all next to each other (in our case, Columns G, H, I). Otherwise, Excel will give you an error message. If the X variables are not all next to each other, rearrange them. You do not have to worry about this when using **StatTools** – you simply check the boxes next to the variables that you want to include in your model.
- *Note 2*: If using **Data Analysis ToolPak**, your X and Y variables cannot contain missing values. If they do, then you first need to clean your data and eliminate all rows that contain missing values. You do not have to worry about this when using **StatTools** – rows with missing observations will be automatically dropped.
- *Note 3*: For multiple regression models, we must look at "Adjusted R-Square" to evaluate the model.
- Using StatTools: **StatTools → Regression and Classification → Regression**:



Regression Type: Multiple

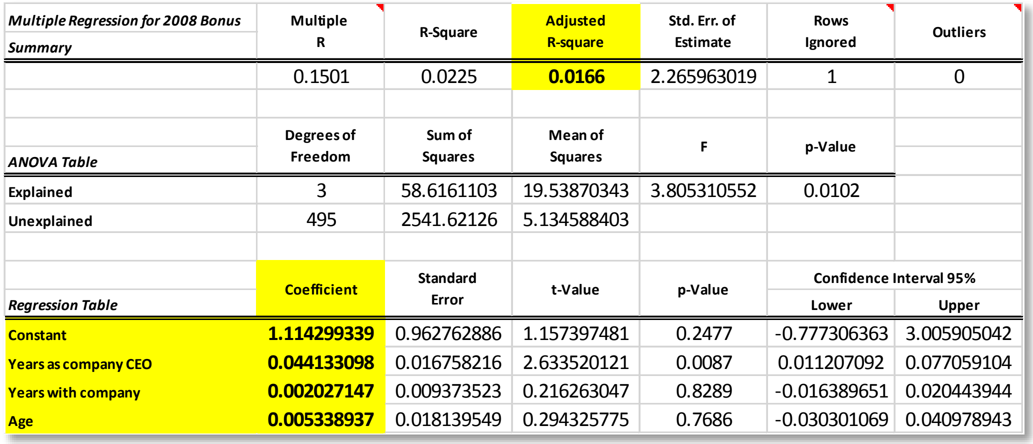
Variables (Select One or More Independents and One Dependent)

Data Set: ceo

Include Untransformed Versions of These Variables

I	D	Data Type	Name	Address
□	□	Category	CEO	A2:A501
□	□	Category	Gender	B2:B501
□	□	Category	Company	C2:C501
□	□	Category	Ticker	D2:D501
□	□	Category	Industry	E2:E501
□	□	Category	Founder	F2:F501
☑	□	Numeric	Years as company CEO	G2:G501
☑	□	Numeric	Years with company	H2:H501
☑	□	Numeric	Age	I2:I501
□	□	Numeric	Total 2008 compensatio..	J2:J501
□	□	Numeric	2008 Salary	K2:K501
□	☑	Numeric	2008 Bonus	L2:L501

□ Include Derived Variables (Transformations of Variables)

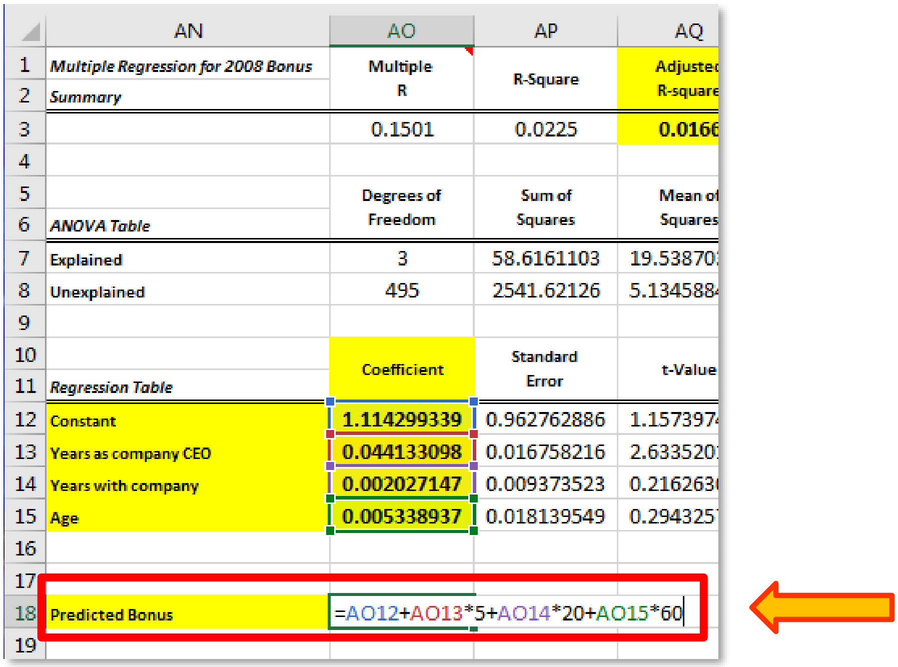


Multiple Regression for 2008 Bonus *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.1501	0.0225	**0.0166**	2.265963019	1	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	3	58.6161103	19.53870343	3.805310552	0.0102	
Unexplained	495	2541.62126	5.134588403			
					Confidence Interval 95%	
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Lower	Upper
Constant	**1.114299339**	0.962762886	1.157397481	0.2477	-0.777306363	3.005905042
Years as company CEO	**0.044133098**	0.016758216	2.633520121	0.0087	0.011207092	0.077059104
Years with company	**0.002027147**	0.009373523	0.216263047	0.8289	-0.016389651	0.020443944
Age	**0.005338937**	0.018139549	0.294325775	0.7686	-0.030301069	0.040978943

← go back to Table of Contents

- Suppose you want to **predict** the bonus of a 60 year old CEO who has been the CEO of the company for 5 years, has been with the company for a total of 20 years. Here:
    1) *Age* = 60,
    2) *Years as CEO* = 5,
    3) *Years with Company* =20.

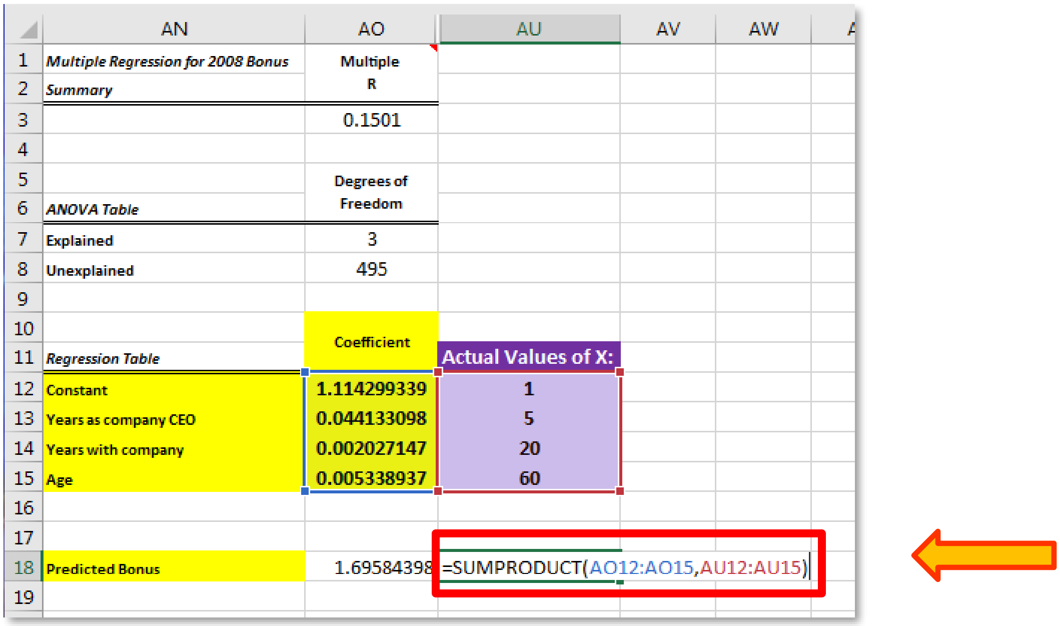
    We need to plug these values into the regression equation to obtain predicted bonus:



	AN	AO	AP	AQ
1	*Multiple Regression for 2008 Bonus*	Multiple R	R-Square	Adjustec R-square
2	*Summary*			
3		0.1501	0.0225	**0.0166**
4				
5		Degrees of Freedom	Sum of Squares	Mean of Squares
6	*ANOVA Table*			
7	Explained	3	58.6161103	19.53870
8	Unexplained	495	2541.62126	5.134588
9				
10		Coefficient	Standard Error	t-Value
11	*Regression Table*			
12	Constant	**1.114299339**	0.962762886	1.157397
13	Years as company CEO	**0.044133098**	0.016758216	2.633520
14	Years with company	**0.002027147**	0.009373523	0.216263
15	Age	**0.005338937**	0.018139549	0.294325
16				
17				
18	Predicted Bonus	=AO12+AO13*5+AO14*20+AO15*60		
19				

<u>Answer</u>: $1.6958 million.

- There is a faster way to obtain the predicted value. The following approach is especially useful when you have many coefficients in the list. Use the **SUMPRODUCT( )** command in Excel. It returns the sum of the product terms (1 * intercept + each X value * each coefficient):



	AN	AO	AU	AV	AW
1	*Multiple Regression for 2008 Bonus*	Multiple R			
2	*Summary*				
3		0.1501			
4					
5		Degrees of Freedom			
6	*ANOVA Table*				
7	Explained	3			
8	Unexplained	495			
9					
10		Coefficient			
11	*Regression Table*		**Actual Values of X:**		
12	Constant	**1.114299339**	**1**		
13	Years as company CEO	**0.044133098**	**5**		
14	Years with company	**0.002027147**	**20**		
15	Age	**0.005338937**	**60**		
16					
17					
18	Predicted Bonus	1.69584398	=SUMPRODUCT(AO12:AO15,AU12:AU15)		
19					

<u>Answer</u>: $1.6958 million.

← go back to Table of Contents