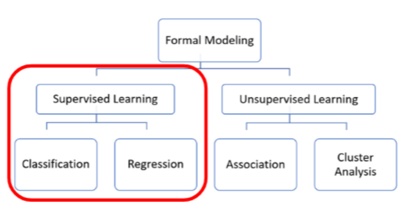
Linear Regression

Remember that in **supervised learning**, we build a model so that we can ***predict*** the value of one of the variables based on one or more of the other variables when new data is presented to us.

Supervised Machine Learning



If our objective is to predict the value of a ***categorical*** variable, this is called **classification**.

If our objective is to predict the value of a **numerical** variable, this is called **regression**.

Questions Related Regression

Are two or more variables related?

Age & Blood Pressure

Height & Shoe Size

Cigarettes a day & Life Expectancy

Marketing Budget & Sales

Absences & Final Test Score

BMI & Cholesterol

If so, what is the strength of the relationship?

Very strong, strong, weak?

What type of relationship exists?

Linear relationship, Exponential relationship?

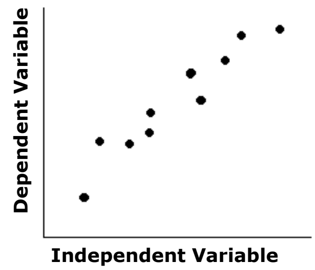
Given a strong relationship, how might we predict the value of one variable from the value of another?

Simple Regression vs. Multiple Regression

In a *simple relationship*, there are only two variables being studied. (Univariate Regression)

In *multiple relationships*, many variables are being studied. (Multivariate Regression)

Scatter Plots



A scatter plot is a **visual way** to describe the nature of the relationship between variables.

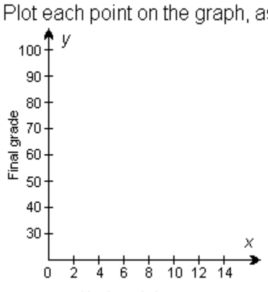
This is a 2D plot but 3D plots are also possible. However, if more than three variables are being considered, it is not possible to visualize the line.

Construct a scatter plot for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class. The data are shown here.

Student	Number of absences *x*	Final grade *y* (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

Draw and label the *x* and *y* axes.

Plot each point on the graph, as shown below.



Correlation is a statistical method used to determine **whether** a linear relationship between variables exists.

- The ***correlation coefficient*** computed from a set of data measures the **strength and direction** of a ***linear*** relationship between two variables.

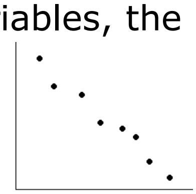
- The symbol we will use for the correlation coefficient is ***r***.

Correlation Coefficient

- The range of the correlation coefficient is from –1 to +1.

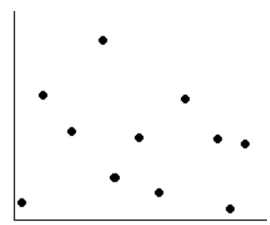
- If there is a ***strong positive linear relationship*** between the variables, the value of *r* will be close to +1.

- If there is a ***strong negative linear relationship*** between the variables, the value of *r* will be close to –1.

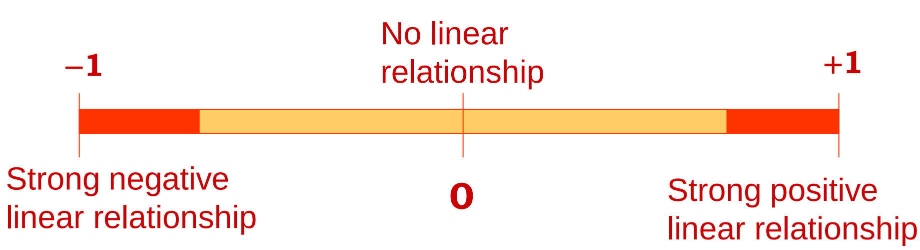


Correlation Coefficient (cont'd.)

- When there is ***no linear relationship*** between the variables or only a weak relationship, the value of *r* will be close to 0.



Correlation Coefficient (cont'd.)



Make a table.

Subject	Age x	Pressure y	xy	x^2	y^2
A	41	123			
B	42	124			
C	59	133			
D	61	144			
E	63	142			
F	77	155			

This is how we would find r by hand but we won't have to do that!

Find the value of xy, x^2, and y^2 and place these values in the corresponding columns of the table. The completed table.

Subject	Age x	Pressure y	xy	x^2	y^2
A	41	123	5043	1681	15129
B	42	124	5208	1764	15376
C	59	133	7847	3481	17689
D	61	144	8784	3721	20736
E	63	142	8946	3969	20736
F	77	155	11935	5929	24025
	$\Sigma x = 343$	$\Sigma y = 821$	$\Sigma xy = 47763$	$\Sigma x^2 = 20545$	$\Sigma y^2 = 113119$

Substitute in the formula and solve for *r*.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(6)(47763) - (343)(821)}{\sqrt{[(6)(20545) - (343)^2][(6)(113119) - (821)^2]}} = 0.971$$

(rounded to three decimal places)

The correlation coefficient suggests a strong positive relationship between age and blood pressure.

Possible Relationships Between Variables

- There is a ***direct*** *cause-and-effect relationship between the variables*: that is, *x* causes *y*.
- There is a ***reverse*** *cause-and-effect relationship between the variables*: that is, *y* causes *x*.
- The *relationship between the variable may be caused by a* ***third variable***: that is, *y* may appear to cause *x* but in reality *z* causes *x*.

  (a lurking variable)

Possible Relationships Between Variables

- There may be a ***complexity** of interrelationships among many variables*; that is, x may cause y but w, t, and z fit into the picture as well.

- The *relationship may be **coincidental***: although a researcher may find a relationship between x and y, common sense may prove otherwise.
  - A sample might coincidentally have a positive relationship between the number of siblings a person has and their IQ. But common sense tells us there is no relationship.

Interpretation of Relationships

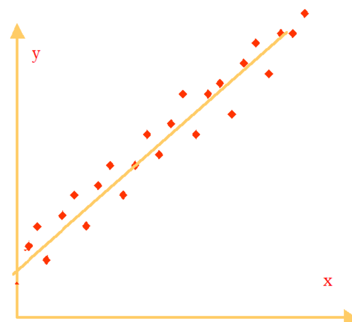
The researcher must **consider all possibilities** and select the appropriate relationship between the variables as determined by the study.

Correlation does not imply causation!

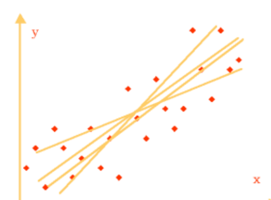
Regression Line

If the value of the correlation coefficient is close to 1 or -1, that means there is a linear relationship between the variables.

The next step, then, is to determine the equation of the ***regression line*** which is the data's line of best fit.



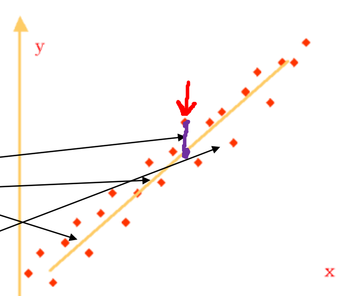
Which line do we use?



The Line of Best Fit

Best fit means that the sum of the squares of the vertical distance from each point to the line is at a minimum.

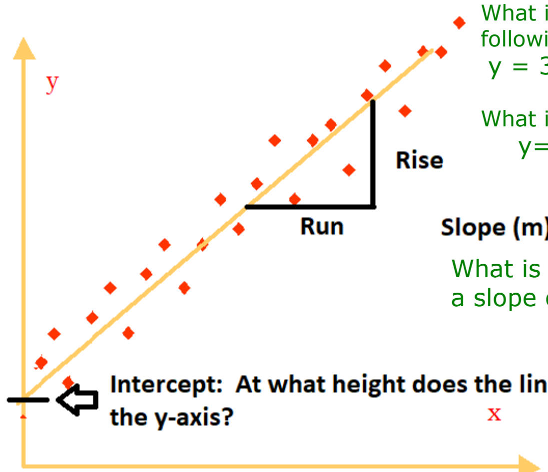
We square each of these distances and add them up. The line for which this amount is at a minimum is our line of **best fit**.



Remember the slope-intercept form of a line?

$$y = mx + b$$

Slope Intercept



Examples:
What is the slope and intercept of the following line?
$y = 3x - 2$ Slope: 3 Intercept: -2

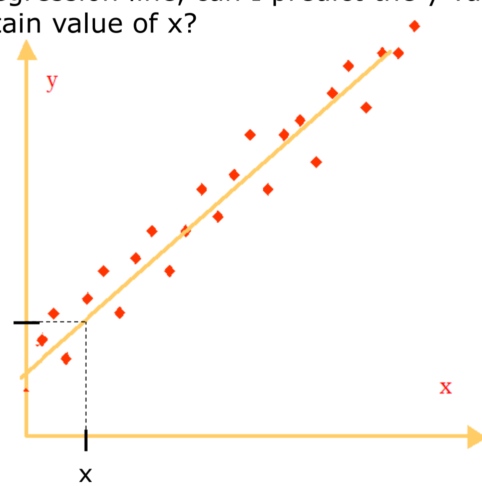
What is the slope and intercept of
$y=-0.5x + 1.42$?
Slope: -0.5 Intercept: 1.42

Slope (m) = Rise/Run

What is the equation of the line with a slope of 5 and an intercept of 3?
$y = 5x + 3$

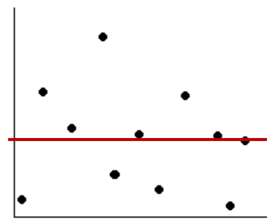
Once we know the regression line, we can use it to make predictions.

Given my regression line, can I predict the y value if I'm given a certain value of x?



I predict this will be the y value. →
We would call this value y'.

If there is not a significant relationship between the two variables, a regression line will be useless.



The data in the following table was obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class. Find the equation of the regression line.

Student	Number of absences *x*	Final grade *y* (%)
A	6	86
B	2	95
C	15	43
D	9	64
E	12	57
F	5	94
G	8	73

We can find the regression line using R, Python, a graphing calculator, Knime, etc.

It is y' = -4.389x + 108.884

Given this regression line, predict what the final grade might be for a student missing

a. 10 days of class.

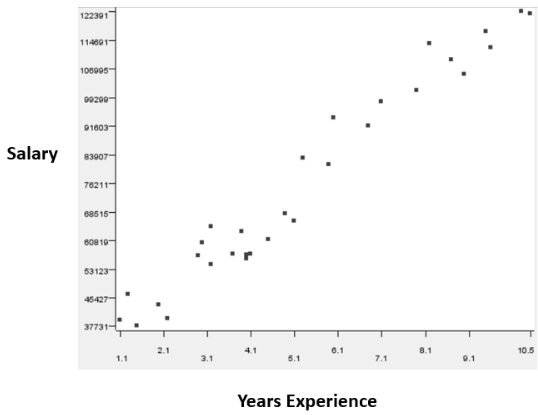
y' = -4.389(10) + 108.884 = 64.994 (round to 65)

b. 4 days of class.

y' = -4.389(4) + 108.884 = 91.328 (round to 91)

Let’s look at some data that shows the **years of experience** of an employee along with the **salary** of that employee.

Here is the scatterplot for this data:



https://s3.us-west-2.amazonaws.com/public.gamelab.fun/dataset/salary_data.csv

YearsExperience	Salary
1.1	39343
1.3	46205
1.5	37731
2	43525
2.2	39891
2.9	56642
3	60150
3.2	54445
3.2	64445
3.7	57189
3.9	63218
4	55794
4	56957
4.1	57081
4.5	61111
4.9	67938
5.1	66029
5.3	83088
5.9	81363
6	93940
6.8	91738
7.1	98273
7.9	101302
8.2	113812
8.7	109431
9	105582
9.5	116969
9.6	112635
10.3	122391
10.5	121872

It appears that there is a linear relationship between the two variables.

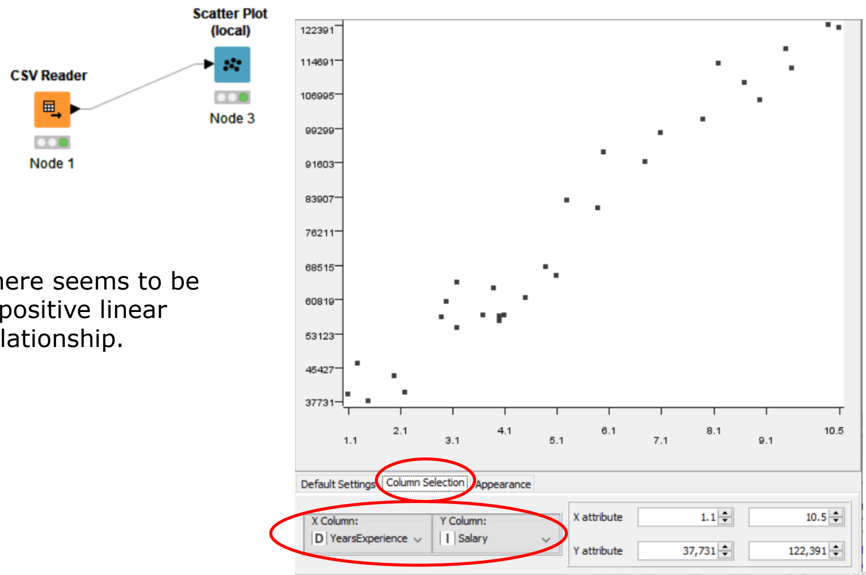
Let’s see if we can use KNIME to

1. create a scatterplot,
2. determine the correlation coefficient,
3. find the equation that describes the line of best fit,
4. and use the line of best fit to make a prediction.

Work along with me using the file salary_data.csv on Canvas.

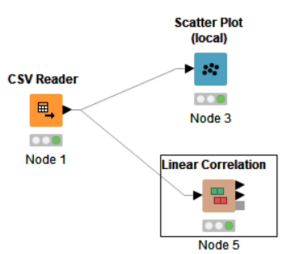
YearsExperience	Salary
1.1	39343
1.3	46205
1.5	37731
2	43525
2.2	39891
2.9	56642
3	60150
3.2	54445
3.2	64445
3.7	57189
3.9	63218
4	55794
4	56957
4.1	57081
4.5	61111
4.9	67938
5.1	66029
5.3	83088
5.9	81363
6	93940
6.8	91738
7.1	98273
7.9	101302
8.2	113812
8.7	109431
9	105582
9.5	116969
9.6	112635
10.3	122391
10.5	121872

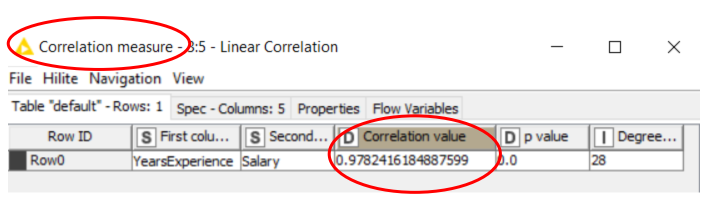
1. Create a scatterplot.



There seems to be a positive linear relationship.

2. Determine the correlation coefficient.

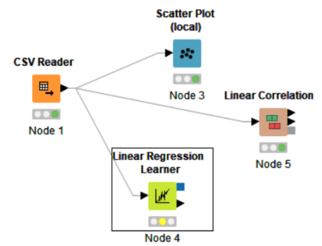


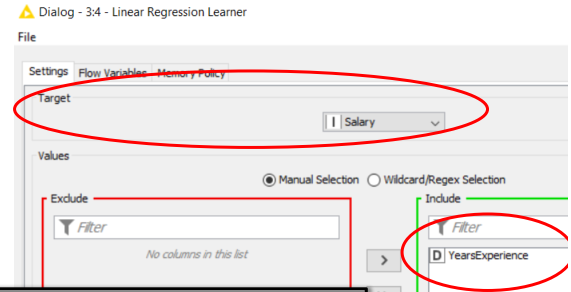


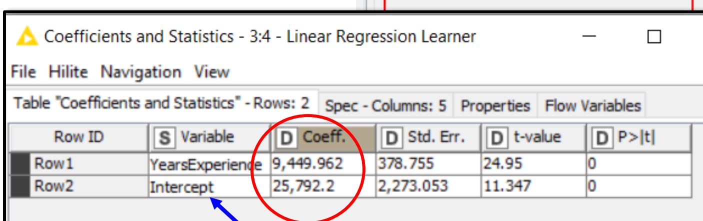
Row ID	S First colu...	S Second...	D Correlation value	D p value	I Degree...
Row0	YearsExperience	Salary	0.9782416184887599	0.0	28

The correlation coefficient is close to 1 so there is indeed a strong positive linear relationship.

3. Find the equation that describes the line of best fit.







Coefficients and Statistics - 3:4 - Linear Regression Learner

File Hilite Navigation View

Table "Coefficients and Statistics" - Rows: 2 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	S Variable	D Coeff.	D Std. Err.	D t-value	D P>|t|
Row1	YearsExperience	9,449.962	378.755	24.95	0
Row2	Intercept	25,792.2	2,273.053	11.347	0

y' = mx + b

The line is
y' = 9449.962x + 25792.2 where
x is the number of years experience.

4. Use the line of best fit to make a prediction.

The line is y' = 9449.962x + 25792.2.

What would we predict the salary to be of a person who has 5 years of experience?

y' = **9449.962*5 + 25792.2**
= 73,042.01

So we would say that we would predict that the salary would be $73,042 (rounded).

4. Use the line of best fit to make a prediction.

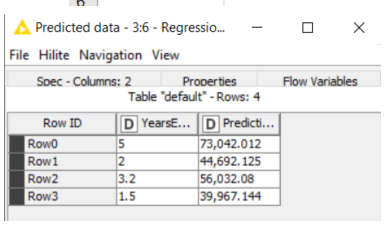
How would we do this in KNIME?

We'll create a CSV file with only years of experience.

We were wanting a prediction for someone with 5 years of experience earlier, so we'll start with that but put some other values in there too while we're at it.

JustYearsExperience.csv

	A
1	**YearsExperience**
2	5
3	2
4	3.2
5	1.5
6	

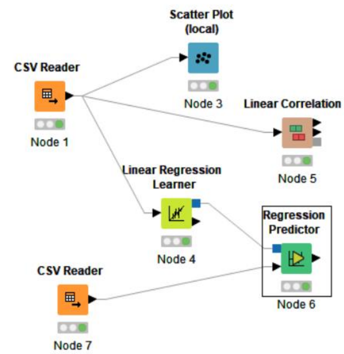

Predicted data - 3:6 - Regressio...

File Hilite Navigation View

Spec - Columns: 2 Properties Flow Variables

Table "default" - Rows: 4

Row ID	D YearsE...	D Predicti...
Row0	5	73,042.012
Row1	2	44,692.125
Row2	3.2	56,032.08
Row3	1.5	39,967.144



Linear Regression Lab

Due next week