Week 13: Correlation, Regression

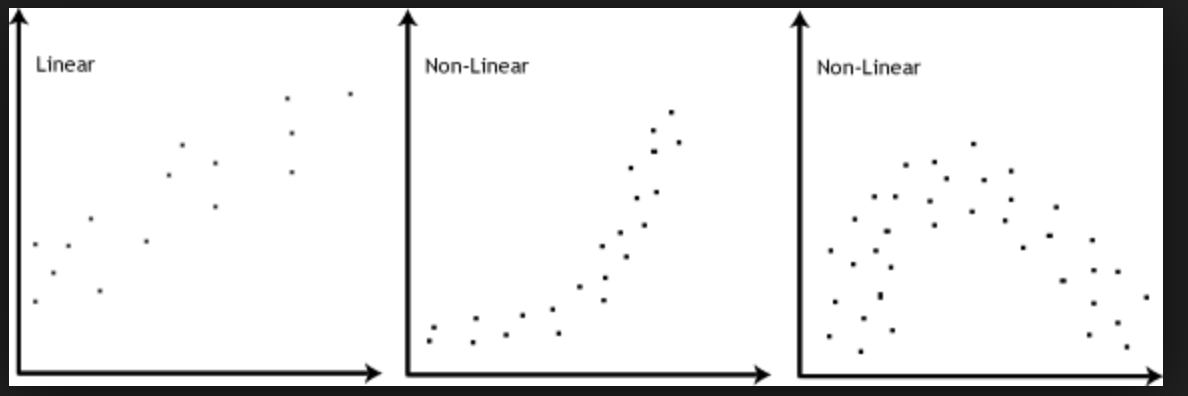
Data 8 Tutoring

# 1 Correlation

## Key Concepts

**Association**

* Refers to any relationship between two variables. It does not have to be linear. For instance, in the plots below, only the first graph demonstrates a linear association.



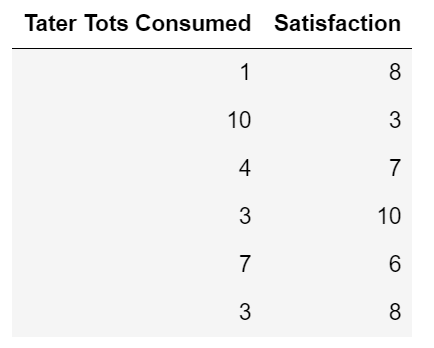
**Correlation**

* Denotes a linear association between two variables.
* The *correlation coefficient* quantitatively measures the strength as well as the direction of the linear relationship between two variables.
* The correlation coefficient is denoted as *r*, a number between -1 and 1
  + Strength: how clustered the scatter plot is around a straight line. If the plot is highly clustered, the *absolute value* of *r* is closer to 1
  + Direction: if y increases as x increases, *r* is positive. If y decreases as x increases, *r* is negative.

**Standard Units**

* Allows us to quantify the relationship between two variables on different scales
* Converting to Standard Units
  + variable\_su = (variable – mean) / SD

**A Formula for r**

* The average of the product of *x* and *y*, when both variables are measured in standard units.
* r = 1 if the scatter diagram is a perfect straight line sloping upwards, and r = −1 if the scatter diagram is a perfect straight line sloping downwards.
* r is *a number without units*. This is because it is computed with standard units, which have no units.
* r is *unaffected by changing the units* on either axis. This too is because r is based on standard units.
* r is *unaffected by switching the axes*. This is because it is the sum of products of standard units; xy = yx. More intuitively, since correlation is a measure of spread around a line, switching the axes won’t change the spread around the line.

## Practice Problems

**1.1** The following table, taters, depicts the number of tater tots a person has eaten, along with a number that quantifies their satisfaction, which is a number that goes from 0 to 10.

a) Complete the function standard\_units which takes in an array num\_array and returns the same array in standard units.

def standard\_units(num\_array):

arr\_mean = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

arr\_sd = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

return \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

def standard\_units(num\_array):

arr\_mean = np.mean(num\_array)

arr\_sd = np.std(num\_array)

return (num\_array-arr\_mean)/arr\_sd

b) Fill in the blanks to define a function correlation that finds the correlation from a table. It takes in three arguments: a table, tbl, and two column indices, x and y.

Hint: Use the standard\_units function defined above!

def correlation(tbl, x, y):

su\_x = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)

su\_y = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)

return \_\_\_\_\_\_\_\_\_\_\_\_\_(\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)

def correlation(tbl, x, y):

su\_x = standard\_units(tbl.column(x))

su\_y = standard\_units(tbl.column(y))

return np.mean(su\_x\*su\_y)

c) Calculate r by using the correlation function.

t

correlation(\_\_\_\_\_\_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_\_\_\_\_)

correlation(Taters, “Tater Tots Consumed”, “Satisfaction”)

d) Suppose that we calculated a value of *r* to be equal to -0.879. What can you conclude about the association between the number of tater tots consumed and a person’s satisfaction?

The number of tater tots consumed and the person’s satisfaction show a strong negative linear relationship. More tater tots consumed is associated with less satisfaction, but consuming more tots does not cause lower levels of satisfaction. There isn’t necessarily a causal relationship between the two because correlation does not necessarily imply causation.

**1.2** True or False?

a. A high value of *r* shows that a change in *x* causes a change in *y*.

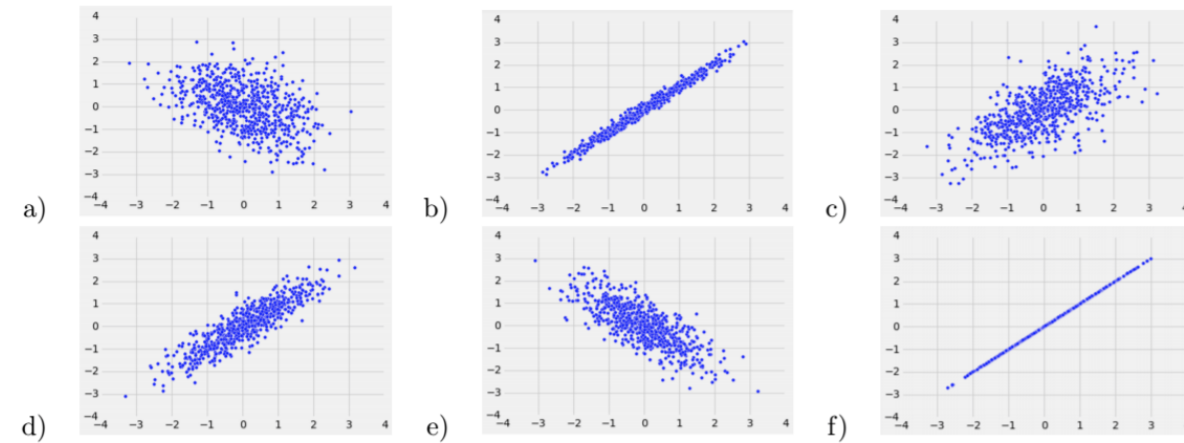
False, correlation doesn’t necessarily imply causation.

b. If we switch the axes of a plot, the correlation coefficient will not change.

True, we are still looking for the correlation between the same two variables.

c. Suppose that we calculated a value of *r* to be equal to .83. We should conclude that eating taters is indeed correlated with satisfaction.

True, we can determine that there exists a positive association.

**1.3** Answer the following questions about the plots below.

1. Order the scatter plots above from least correlated to most correlated.

A, C, E, D, B, F

1. Which plots have a positive correlation coefficient? Negative correlation coefficient?

Positive: D, B, C, F

Negative: A, E

# 2 Regression

## Key Concepts

**Correlation Coefficient**

* The equation for the data’s regression line can be calculated using *r*. Recall that the equation of a line is y = slope · x + intercept.

**Standard Units**

* Graphically, the scatterplot and regression line look the same whether x and y are in standard units or their original units.
* Calculating the regression line when *x* and *y* are in standard units:
  + estimate of *y* = *r* ⋅ the given *x*,where *x* and *y* are in standard units

**Calculating the regression line when *x* and *y* are in original units:**

1) Calculate the correlation coefficient, *r*.

2) Calculate the slope.

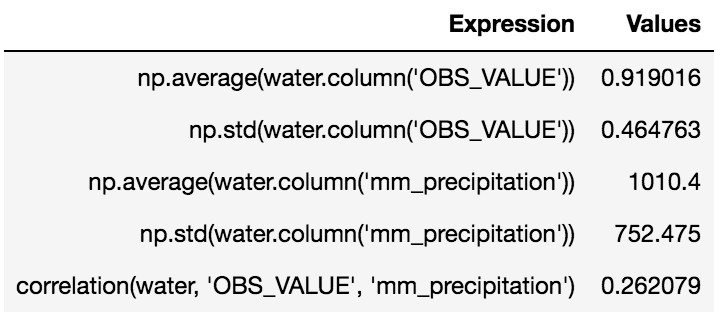
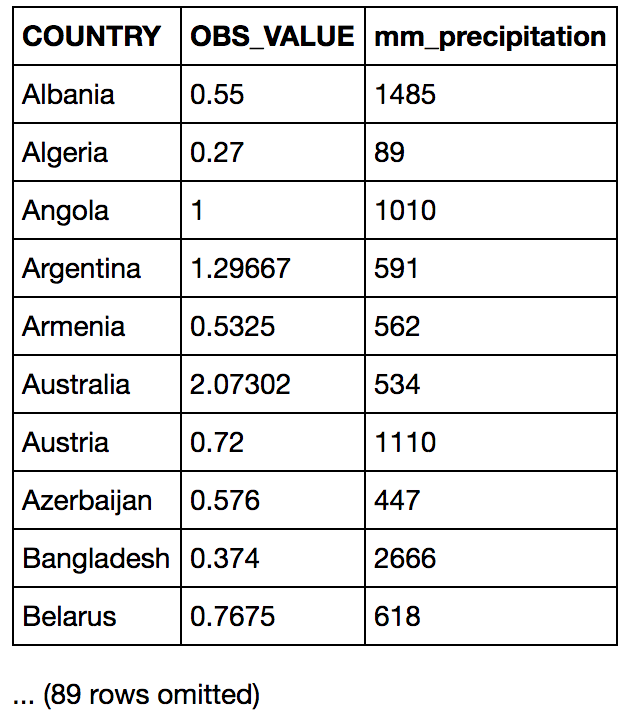
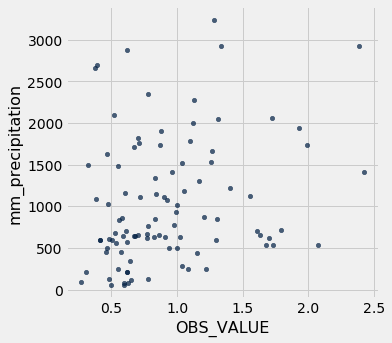
slope of the regression line

3) Calculate the intercept.

intercept of the regression line

## Practice Problems

The water table contains one row per country with data from 2014. The OBS\_VALUE column represents the approximate price ranking of a 1.5 liter bottle of mineral water in that country, and the mm\_precipitation column represents the average precipitation in that country (in millimeters).



**2.1** What is the value of correlation(water, ‘mm\_precipitation’, ‘OBS\_VALUE’)?

0.262079; the correlation(x, y) is the same as correlation(y, x) because correlation measures the strength of a linear relationship (if it exists) between two variables. Order doesn’t matter!

**2.2** Write an equation for the regression line of the data in the water table, using OBS\_VALUE as *y* using the mm\_precipitation as *x*.

**2.3** Using the regression line equation above, what would we expect the OBS\_VALUE to be in 2014 for a country that had an average of 700 mm of precipitation?

y=0.00016(700) + 0.76= 0.87

# 3 Root Mean Squared Errors (RMSE)

## Key Concepts

* Root Mean Squared Error is the square root of the average of the squared errors
* RMSE, where *n* is the number of points in our dataset and each

|  |  |  |
| --- | --- | --- |
| Original Plot | Poor Fit (High RMSE) | Best Fit (Lowest RMSE) |
|  |  |  |

## Practice Problems

**3.1** Write a function that returns the RMSE of an array of observed values if the predicted values are given by an array. The two arrays have the same length.

def RMSE(observed, predicted):

residuals = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

squared\_residuals = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

squared\_resid\_avg = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

return \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

def RMSE(observed, predicted):

residuals = observed - predicted

squared\_residuals = residuals\*\*2

squared\_resid\_avg = np.mean(squared\_residuals)

return np.sqrt(squared\_resid\_avg)

**3.2** In the calculation of root mean squared error, why is it important for us to square the residual before taking the sum?

Because if we don’t square them, then the sum of the differences will be zero. The positives and the negatives will cancel each other out. Squaring our residuals get rid of the negatives, giving us a better way of quantifying error.