DATA 8

Fall 2020

Tutoring Section 13

Machine Learning: Correlation, Regression

Slides by Kevin Miao

Logistics

- Vibe Check:
  - How stressed/relaxed do you feel?
  - 2.5 weeks left of classes! How prepared do you feel for the last stretch of the class/semester?
- **Project 3**, movie recommendations, has been released!

As always, let me know if you have any questions about anything.

EOS Evaluation

Data 8 Tutor Evaluation Form

https://tinyurl.com/d8tutfeedback

Today

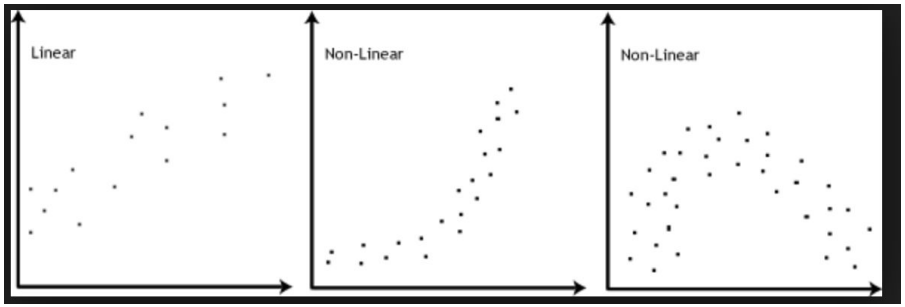
- Correlation Coefficient
- Regression
- Errors: (R)MSE

Worksheet

Link: **https://tinyurl.com/d8tutweek13**

Associations

- **Association**
  - Any type of relationship between two variables
    - Could be **linear**, **non-linear**



  - In this class, we will only focus on **linear relationships**

Correlation

- **Correlation**
  - **Goal**: How do we quantify a linear relationship?
    - **Correlation coefficient, r**
      - Strength
      - Direction
  - Calculation
    - **Mean** of the **product** of **x** and **y** in **standard units**
  - Does our correlation coefficient change if …
    - We swap our axes x, y?
    - We convert our x units from say inches to centimeters?
  - What is the range of our correlation coefficient?

Q1.1a

Tater Tots Consumed	Satisfaction
1	8
10	3
4	7
3	10
7	6
3	8

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 ____________________________________
```

Q1.1bcd

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 _________________(_____________________________________)
```

c) Calculate r by using the correlation function.

```
correlation(_____________, _____________, _____________)
```

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?

Q1.2

1.2 True or False?

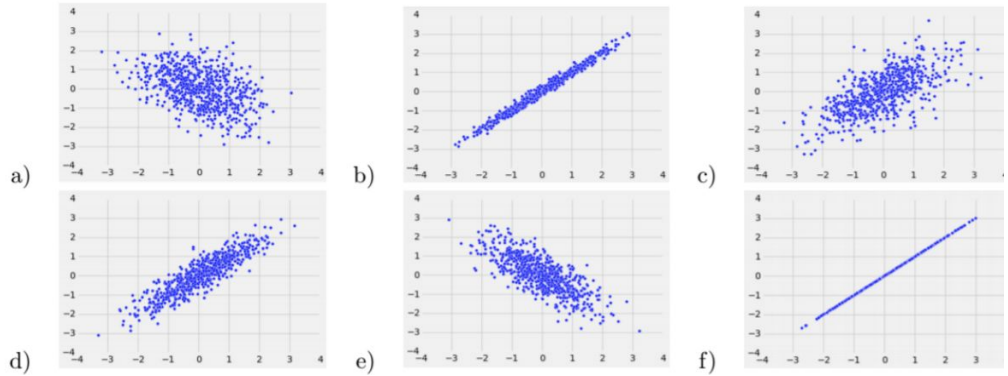
a. A high value of r shows that a change in x causes a change in y.

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

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.

Q1.3

1.3 Answer the following questions about the plots below.



a. Order the scatter plots above in from least correlated to most correlated.

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

Regression

- **Objective:** We want to predict a **number** based on given parameters.
  - **Linear Regression**
    - We know that the relationship between our variables and the number we want to predict has a linear shape.
  - **Calculating the formula/line that predicts the numbers**
    - Calculate the correlation coefficient
      - Mean of the product of x and y in standard units
    - Calculate the slope
      - Slope = r * (SD_Y/SD_X)
    - Calculate the intercept by plugging in the means of x and y

Q2.2

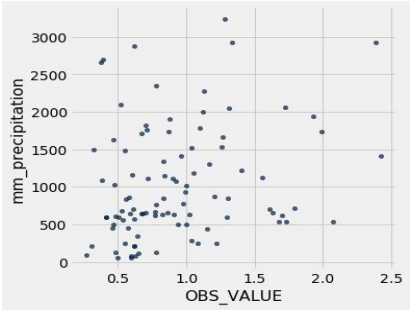
Practice Problems

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).

COUNTRY	OBS_VALUE	mm_precipitation
Albania	0.55	1485
Algeria	0.27	89

... (89 rows omitted)

Expression	Values
np.average(water.column('OBS_VALUE'))	0.919016
np.std(water.column('OBS_VALUE'))	0.464763
np.average(water.column('mm_precipitation'))	1010.4
np.std(water.column('mm_precipitation'))	752.475
correlation(water, 'OBS_VALUE', 'mm_precipitation')	0.262079



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.

Q2.3

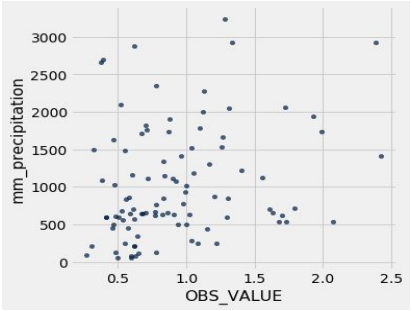
Practice Problems

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).

COUNTRY	OBS_VALUE	mm_precipitation
Albania	0.55	1485
Algeria	0.27	89

... (89 rows omitted)

Expression	Values
np.average(water.column('OBS_VALUE'))	0.919016
np.std(water.column('OBS_VALUE'))	0.464763
np.average(water.column('mm_precipitation'))	1010.4
np.std(water.column('mm_precipitation'))	752.475
correlation(water, 'OBS_VALUE', 'mm_precipitation')	0.262079



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?

Errors

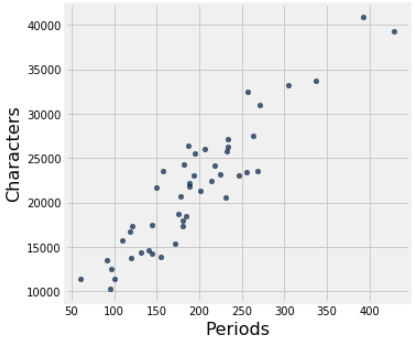
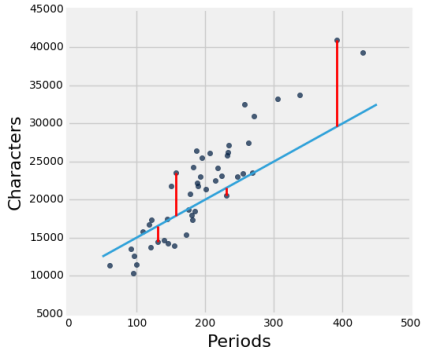
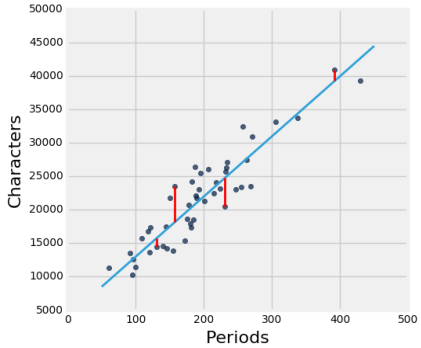
- **Context:** In Data 8, we provide you a lot of statistical knowledge, but in traditional Machine Learning Engineering side, we will approach ML problems through this perspective.
- **Set-up:** You have a problem, you want to predict something, define a model (Linear Regression), define an error (RMSE/MSE) and minimize it. This gives you a model (line) with the lowest error possible given your data points.
- **Here:**
  - **Root Mean Squared Errors** (looks like SD)
    - **Square root** of the **average** of **squared errors**
    - $\sqrt{(error^2_{point\ 1} + \cdots + error^2_{point\ n})/n}$
    - Error = actual y – predicted y

RMSE/MSE/Linear Regression Facts

- **Statistics perspective** vs **Computer Science perspective**
  - The line calculated with correlation coefficients is the same line that minimizes the error. In other words, the linear regression line is the line that is the best!
- **Why do we pick RMSE?**
  - It is completely in your right to substitute the RMSE by another loss function such as the absolute loss. It provides different assumptions.
- **What does the minimize function do?**
  - Imagine it takes the derivate, sets it to 0 and calculates the parameters for which the maximum is attained.
- **What happens if we run minimize on MSE instead of RMSE?**
  - MSE does not change the shape of the graph and will not affect the outputted line.

You don't have to know what's in grey.

Examples

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

Q3.1-3.2

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):
    residual = ________________________________________
    squared_residuals = ____________________________________________
    squared_resid_avg = ____________________________________________
    return ____________________________________________
```

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