

AAIL100: Introduction to Data Science

Ragesh Jaiswal, CSE, IIT Delhi

Prediction

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: *Based on a person's internet profile, which websites are likely to interest them?*
- We start with one-variable prediction:
 - *Predicting the value of one variable (**response**) based on the value of another (**predictor**)*
 - Let us build a predictor of a child's height based on their parents' average height.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: *Based on a person's internet profile, which websites are likely to interest them?*
- We start with one-variable prediction:
 - *Predicting the value of one variable (**response**) based on the value of another (**predictor**)*
 - Let us build a predictor of a child's height based on their parents' average height.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: *Is there a measure for linear association?*

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: *Based on a person's internet profile, which websites are likely to interest them?*
- We start with one-variable prediction:
 - *Predicting the value of one variable (**response**) based on the value of another (**predictor**)*
 - Let us build a predictor of a child's height based on their parents' average height.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: *Is there a measure for linear association?*
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: *Based on a person's internet profile, which websites are likely to interest them?*
- We start with one-variable prediction:
 - *Predicting the value of one variable (**response**) based on the value of another (**predictor**)*
 - Let us build a predictor of a child's height based on their parents' average height.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: *Is there a measure for linear association?*
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.
 - Quantitative: **Correlation coefficient** or simply **correlation** r , has the following desirable properties:
 - $-1 \leq r \leq 1$
 - $r = 1$ if the scatter plot is a perfect straight line sloping upwards and $r = -1$ if the scatter plot is a perfect straight line sloping downwards.

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: *Based on a person's internet profile, which websites are likely to interest them?*
- We start with one-variable prediction:
 - *Predicting the value of one variable (**response**) based on the value of another (**predictor**)*
 - Let us build a predictor of a child's height based on their parents' average height.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: *Is there a measure for linear association?*
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.
 - Quantitative: **Correlation coefficient** or simply **correlation** r , has the following desirable properties:
 - $-1 \leq r \leq 1$
 - $r = 1$ if the scatter plot is a perfect straight line sloping upwards and $r = -1$ if the scatter plot is a perfect straight line sloping downwards.
 - Candidate measure: *r is the average of the product of the two variables in standard units. This is also known as **Pearson correlation**.*

Prediction

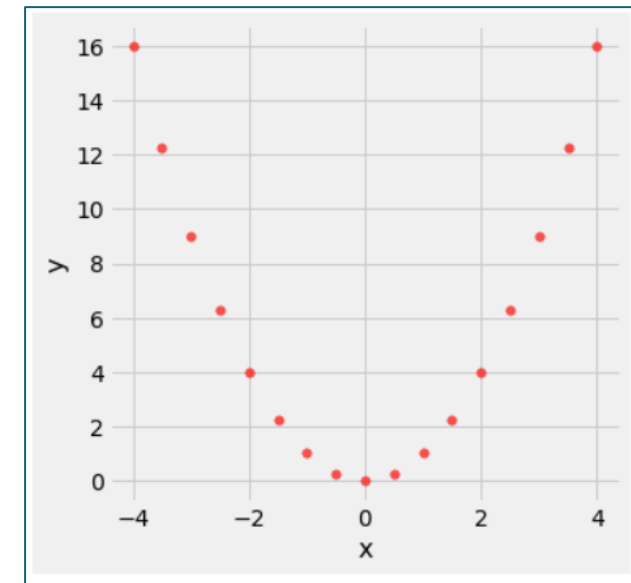
- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: Based on a person's internet profile, which websites are likely to interest them?
- We start with one-variable prediction:
 - Predicting the value of one variable (**response**) based on the value of another (**predictor**)
 - Let us build a predictor of a child's height based on their parents' average height.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: Is there a measure for linear association?
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.
 - Quantitative: **Correlation coefficient** or simply **correlation** r , has the following desirable properties:
 - $-1 \leq r \leq 1$
 - $r = 1$ if the scatter plot is a perfect straight line sloping upwards and $r = -1$ if the scatter plot is a perfect straight line sloping downwards.
 - Candidate measure: r is the average of the product of the two variables in standard units. This is also known as **Pearson correlation**.

- Properties of r :
 - r does not have any units since it is based on standard units.
 - r is unaffected by switching the axis.
- Important points:
 - Correlation only measures association. It does not imply causation.

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: Based on a person's internet profile, which websites are likely to interest them?
- We start with one-variable prediction:
 - Predicting the value of one variable (**response**) based on the value of another (**predictor**)
 - Let us build a predictor of a child's height based on their parents' average height.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: Is there a measure for linear association?
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.
 - Quantitative: **Correlation coefficient** or simply **correlation** r , has the following desirable properties:
 - $-1 \leq r \leq 1$
 - $r = 1$ if the scatter plot is a perfect straight line sloping upwards and $r = -1$ if the scatter plot is a perfect straight line sloping downwards.
 - Candidate measure: r is the average of the product of the two variables in standard units. This is also known as **Pearson correlation**.

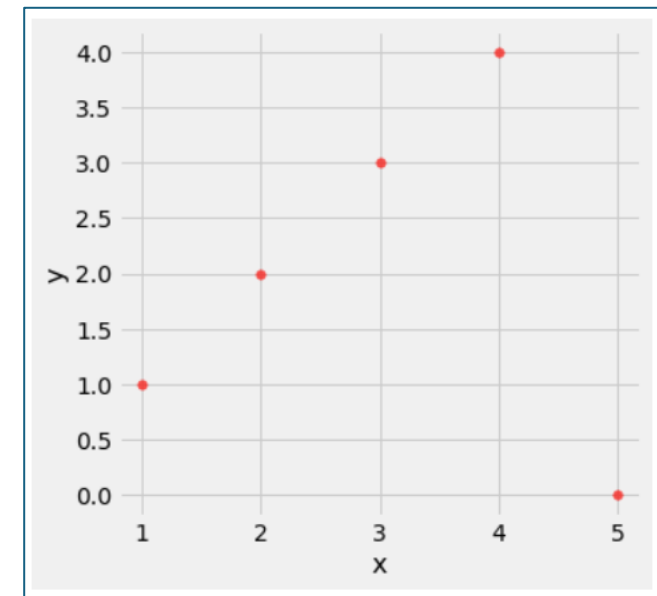
- Properties of r :
 - r does not have any units since it is based on standard units.
 - r is unaffected by switching the axis.
- Important points:
 - Correlation only measures association. It does not imply causation.
 - Correlation measures linear association only.
 - Example: $y = x^2$



Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: Based on a person's internet profile, which websites are likely to interest them?
- We start with one-variable prediction:
 - Predicting the value of one variable (**response**) based on the value of another (**predictor**)
 - Let us build a predictor of a child's height based on their parents' average height.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: Is there a measure for linear association?
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.
 - Quantitative: **Correlation coefficient** or simply **correlation** r , has the following desirable properties:
 - $-1 \leq r \leq 1$
 - $r = 1$ if the scatter plot is a perfect straight line sloping upwards and $r = -1$ if the scatter plot is a perfect straight line sloping downwards.
 - Candidate measure: r is the average of the product of the two variables in standard units. This is also known as **Pearson correlation**.

- Properties of r :
 - r does not have any units since it is based on standard units.
 - r is unaffected by switching the axis.
- Important points:
 - Correlation only measures association. It does not imply causation.
 - Correlation measures linear association only.
 - Example: $y = x^2$
 - Correlation is sensitive to outliers



Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: Based on a person's internet profile, which websites are likely to interest them?
- We start with one-variable prediction:
 - Predicting the value of one variable (**response**) based on the value of another (**predictor**)
 - Let us build a predictor of a child's height based on their parents' average height.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: Is there a measure for linear association?
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.
 - Quantitative: **Correlation coefficient** or simply **correlation** r , has the following desirable properties:
 - $-1 \leq r \leq 1$
 - $r = 1$ if the scatter plot is a perfect straight line sloping upwards and $r = -1$ if the scatter plot is a perfect straight line sloping downwards.
 - Candidate measure: r is the average of the product of the two variables in standard units. This is also known as **Pearson correlation**.

- Where is the line given r ?
- How does the line give a predictor?

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: Based on a person's internet profile, which websites are likely to interest them?
- We start with one-variable prediction:
 - Predicting the value of one variable (**response**) based on the value of another (**predictor**)
 - Let us build a predictor of a child's height based on their parents' average height.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: Is there a measure for linear association?
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.
 - Quantitative: **Correlation coefficient** or simply **correlation** r , has the following desirable properties:
 - $-1 \leq r \leq 1$
 - $r = 1$ if the scatter plot is a perfect straight line sloping upwards and $r = -1$ if the scatter plot is a perfect straight line sloping downwards.
 - Candidate measure: r is the average of the product of the two variables in standard units. This is also known as **Pearson correlation**.

- Where is the line given r ?
 - The line through $(0, 0)$ with slope r passes through the “core” of the data points in the scatter plot in the standard units.
- How does the line give a predictor?
 - A good predictor for y based on x , would consider the points in the scatter plot with similar values of x and take the average of their y values.

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: Based on a person's internet profile, which websites are likely to interest them?
- We start with one-variable prediction:
 - Predicting the value of one variable (**response**) based on the value of another (**predictor**)
 - Let us build a predictor of a child's height based on their parents' average height.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: Is there a measure for linear association?
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.
 - Quantitative: **Correlation coefficient** or simply **correlation** r , has the following desirable properties:
 - $-1 \leq r \leq 1$
 - $r = 1$ if the scatter plot is a perfect straight line sloping upwards and $r = -1$ if the scatter plot is a perfect straight line sloping downwards.
 - Candidate measure: r is the average of the product of the two variables in standard units. This is also known as **Pearson correlation**.

- Where is the line given r ?
 - The line through $(0, 0)$ with slope r passes through the “core” of the data points in the scatter plot in the standard units.
- How does the line give a predictor?
 - A good predictor for y based on x , would consider the points in the scatter plot with similar values of x and take the average of their y values.

- $\frac{\bar{y}_i - \mu_Y}{\sigma_Y} = r \cdot \frac{x_i - \mu_X}{\sigma_X}$, for every i .
- So,
$$\bar{y}_i = r \cdot \frac{\sigma_Y}{\sigma_X} \cdot x_i + \left(\mu_Y - r \cdot \mu_X \cdot \frac{\sigma_Y}{\sigma_X} \right)$$


Slope


Intercept

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: Based on a person's internet profile, which websites are likely to interest them?
- We start with one-variable prediction:
 - Predicting the value of one variable (**response**) based on the value of another (**predictor**)
 - Let us build a predictor of a child's height based on their parents' average height.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: Is there a measure for linear association?
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.
 - Quantitative: **Correlation coefficient** or simply **correlation** r , has the following desirable properties:
 - $-1 \leq r \leq 1$
 - $r = 1$ if the scatter plot is a perfect straight line sloping upwards and $r = -1$ if the scatter plot is a perfect straight line sloping downwards.
 - Candidate measure: r is the average of the product of the two variables in standard units. This is also known as **Pearson correlation**.

- Where is the line given r ?
 - The line through $(0, 0)$ with slope r passes through the “core” of the data points in the scatter plot in the standard units.
 - The slope being ≤ 1 is called **regression** to the mean.
- How does the line give a predictor?
 - A good predictor for y based on x , would consider the points in the scatter plot with similar values of x and take the average of their y values.

$$\frac{\bar{y}_i - \mu_Y}{\sigma_Y} = r \cdot \frac{x_i - \mu_X}{\sigma_X}, \text{ for every } i.$$

• So,

$$\bar{y}_i = r \cdot \frac{\sigma_Y}{\sigma_X} \cdot x_i + \left(\mu_Y - r \cdot \mu_X \cdot \frac{\sigma_Y}{\sigma_X} \right)$$

Slope

Intercept

End