Simple linear Regression:

Linear regression is a method we use to understand the relationship between two things.

It involves predicting a dependent variable (Y) based on a single independent variable (X).

E.g.:

* + - No. of. hours you study and the grades you get, has relationship.
    - Dengue patients’ blood pallets count, and no. of days stay in hospital has relationship.
    - Predicting a person's height or dress size based on their age.

The input given to simple linear regression may be in numerical, categorical, ordinal

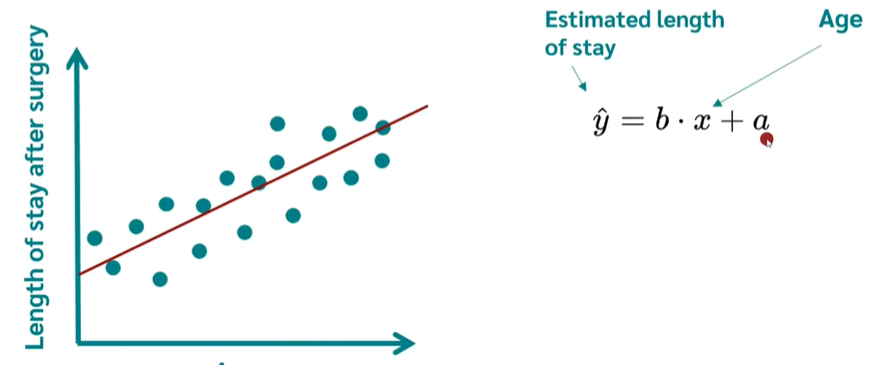
But output will be on metrics.

Linear regression helps us:

* Understand how two things are related.
* Make predictions based on data.
* Check how well our predictions match reality.
* Identify unusual points (outliers) that might need special attention.

Equation of linear regression:

Y = a + bx



Math Behind Linear Regression:

Assume we have x = 1, 2, 3, 4, 5 and y = 2, 4, 5, 4, 5

We want to fit linear regression line, to fit this data.

Steps to calculate a and b:

1. Calculate the means of x and y

A number with numbers and a line

Description automatically generated with medium confidence

A number with black text

Description automatically generated with medium confidence

1. Calculate the slope (b).

A mathematical equation with numbers

Description automatically generated with medium confidence



A number with numbers and symbols

Description automatically generated with medium confidence

A number with numbers and symbols

Description automatically generated with medium confidence

1. Calculating intercept, a,

A math equations on a white background

Description automatically generated

The linear regression equation for this example is:

y=2.2+0.6x

This means that for every unit increase in x, y is expected to increase by 0.6 units, starting from an initial value of 2.2 when x is 0.

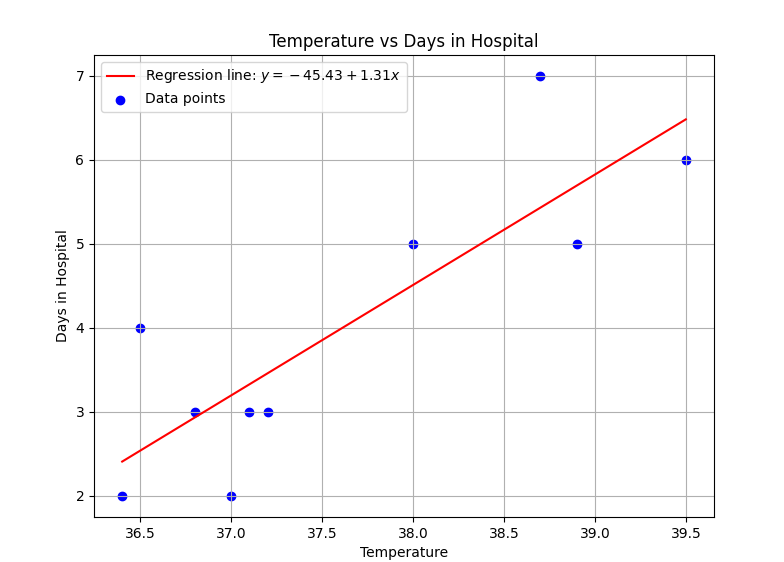
**Explanation of the Terms**

* **Dependent Variable (y)**: This is the variable you're trying to predict or explain.
  + **Example**: The future stock price.
* **Independent Variable (x)**: This is the variable you're using to make predictions.
  + **Example**: The date or time.
* **Intercept (a)**: This is the point where the regression line crosses the y-axis. It represents the expected value of y when x is 0.
  + **Example**: The stock price at the start of your data.
* **Slope (b)**: This represents the rate of change of y with respect to x. It shows how much y is expected to increase (or decrease) as x increases by one unit.

**Example**: How much the stock price changes per day.

**Why Do We Use Linear Regression?**

* **Understand Relationships**: To see how one thing affects another. Like, does more study time usually mean better grades?
* **Make Predictions**: To guess future outcomes. If you know the hours studied, you can predict the grade.
* **Check the Fit**: To see how well the line represents the data. We want the line to be as close to all the points as possible.
* **Find Outliers**: To spot unusual points that don’t fit the pattern. Like a super high grade with very few study hours.



**Understanding Outliers**

Outliers are points that are far from the rest of the data. They can affect the regression line a lot, making it less accurate. We use a method called standard deviation to find them.

**1. Finding Outliers**

* **Mean**: Average of all temperatures.
* **Standard Deviation**: Measure of how spread out the temperatures are.
* **Outliers**: Temperatures that are far from the mean (more than 2 standard deviations away).

What is mean by Residual?.

Residuals are the differences between the actual data points and the points on our regression line. If the line perfectly predicts the data, the residuals would all be zero. But usually, they're not, so we plot them to see if there are any patterns.

**Understanding Residuals with an Example**

Imagine you are a baker, and you want to predict how much flour you will need based on the number of cakes you plan to bake. You have some data from the past week:

* **Day 1**: 3 cakes, used 6 cups of flour
* **Day 2**: 5 cakes, used 10 cups of flour
* **Day 3**: 2 cakes, used 4 cups of flour
* **Day 4**: 4 cakes, used 9 cups of flour
* **Day 5**: 6 cakes, used 12 cups of flour

You notice a general pattern: each cake seems to need about 2 cups of flour. You decide to create a simple prediction model:

Flour needed=2×Number of cakes\text{Flour needed} = 2 \times \text{Number of cakes}Flour needed=2×Number of cakes

**Prediction vs. Reality**

Using your model, you predict:

* **Day 1**: 3 cakes -> 2×3=62 \times 3 = 62×3=6 cups of flour (perfect match)
* **Day 2**: 5 cakes -> 2×5=102 \times 5 = 102×5=10 cups of flour (perfect match)
* **Day 3**: 2 cakes -> 2×2=42 \times 2 = 42×2=4 cups of flour (perfect match)
* **Day 4**: 4 cakes -> 2×4=82 \times 4 = 82×4=8 cups of flour (off by 1 cup)
* **Day 5**: 6 cakes -> 2×6=122 \times 6 = 122×6=12 cups of flour (perfect match)

**What are Residuals?**

In the context of regression analysis, Residuals are the differences between what you predicted and what happened. They tell you how far off your predictions are:

* **Day 1**: Predicted 6, Actual 6 -> Residual = 6−6=0
* **Day 2**: Predicted 10, Actual 10 -> Residual = 10−10 = 0
* **Day 3**: Predicted 4, Actual 4 -> Residual = 4−4=0
* **Day 4**: Predicted 8, Actual 9 -> Residual = 9−8=1
* **Day 5**: Predicted 12, Actual 12 -> Residual = 12−12=0

### In fig 1,

* **Blue Points**: Show the actual data points.
* **Red Line**: Represents our prediction model.

### Why Residuals Matter

Residuals help us understand how good our model is:

* **Small Residuals**: Means our predictions are close to reality.
* **Large Residuals**: Indicates our model might be missing something.
* **Patterns in Residuals**: If residuals show a pattern (e.g., all positive or negative), it suggests our model might not be capturing some important aspect of the data.

### Residuals and Model Accuracy

1. **Definition**:
   * **Residual**: The difference between the actual observed value and the value predicted by the model. Mathematically, it's given by: Residual=Actual Value−Predicted Value
2. **Purpose**:
   * Residuals help us understand how well our model is performing. Small residuals mean our predictions are close to the actual values, indicating a good fit. Large residuals suggest our model is missing important information or patterns in the data.
3. **Analysing Residuals**:
   * **Random Distribution**: If the residuals are randomly distributed around zero (with no clear pattern), it indicates that the model is capturing the data well.
   * **Patterns**: If there is a pattern in the residuals (e.g., they systematically increase or decrease), it suggests that the model is not capturing some important aspect of the data.

### When Residuals are Too High

If the residuals are consistently high, it means the model's predictions are far from the actual values. This is a sign that the model might not be suitable for the data. In such cases, you might consider:

1. **Revisiting the Current Model**:
   * Check if the model has been properly fitted.
   * Ensure all relevant variables are included.
   * Look for data preprocessing issues (e.g., missing values, outliers).

2. **Trying Different Models**:

* **Linear Regression**: Good for data with a linear relationship between the variables.
* **Polynomial Regression**: Can capture non-linear relationships by adding polynomial terms.
* **Decision Trees**: Good for capturing complex interactions between variables.
* **Random Forests and Gradient Boosting**: Ensemble methods that often perform well on various types of data.
* **Support Vector Machines (SVM)**: Effective for classification and regression tasks with clear margins of separation.

3. **Cross-Validation**:

* Use techniques like cross-validation to evaluate the performance of different models on the data and choose the one that performs best.

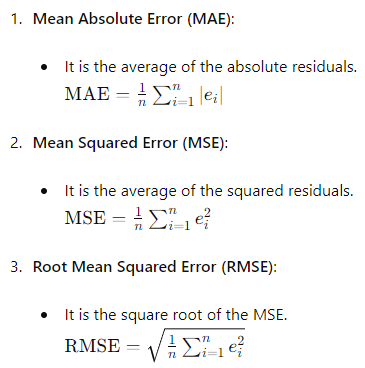
**Residuals are used to calculate metrics:**

**Residuals** are not metrics themselves, but they are a fundamental component **used** in calculating metrics **that evaluate the performance of regression models**.

they are crucial in **calculating performance metrics** that provide insights into the model's accuracy.

If yi is the observed value and y^i is the predicted value for the iii-th observation, the residual ei​ is:

ei=yi − y^i



A white background with black text

Description automatically generated

A screenshot of a math equation

Description automatically generated

These metrics help us understand how well the model is performing by summarizing the residuals into a single value that is easier to interpret.

So, **Metrics**: Use residuals to evaluate model performance (e.g., MAE, MSE, RMSE).

helping to assess the accuracy of the model's predictions.

### What Does "Fit" mean?

When we talk about a model being "fitted" to data, we mean that the model has been trained to understand the relationship between the input data (features) and the output data (targets).

### Fitting a Model

1. **Collect Data**:
   * You have a set of data points. For example, in our baker's example, the number of cakes and the amount of flour used.
2. **Choose a Model**:
   * You decide on a model that you think will best describe the relationship between your data. For simplicity, let's say we choose a linear regression model.
3. **Train the Model**:
   * This is the process of fitting the model. During training, the model looks at the data and tries to find the best parameters (like the slope and intercept in linear regression) that describe the relationship.
   * The goal is to minimize the differences between the actual values and the predicted values. This process is called "optimization."

### Properly Fitted Model

A model is properly fitted when it has learned the relationship between the input and output data well. This means it can make accurate predictions on new, unseen data.

### Signs of a Properly Fitted Model

1. **Low Residuals**:
   * The differences between the actual values and the predicted values (residuals) are small.
2. **Random Residual Pattern**:
   * The residuals are randomly distributed around zero. There are no obvious patterns, indicating that the model is capturing the data's underlying structure well.

### Underfitting and Overfitting

1. **Underfitting**:
   * This happens when the model is too simple to capture the underlying pattern of the data. It results in high residuals and a poor fit.
2. **Overfitting**:
   * This occurs when the model is too complex and captures not just the underlying pattern but also the noise in the data. While it might perform well on training data, it fails to generalize to new data.

Summary,

* **Fitted Model**: A model that has learned the relationship between the input and output data.
* **Properly Fitted**: The model accurately captures the relationship, leading to small, randomly distributed residuals.
* **Underfitting/Overfitting**: Problems that occur when the model is too simple or too complex, respectively.

**Methods to capture the outliers,**

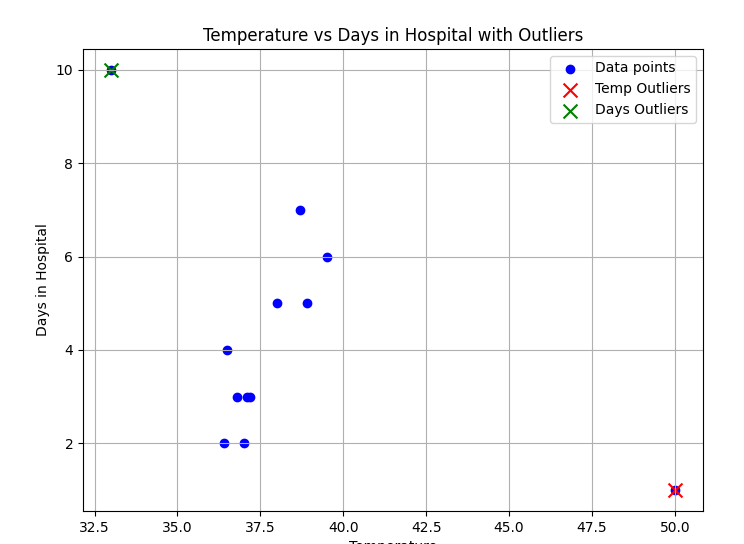
### Using Interquartile Range (IQR) to Detect Outliers

We Use the IQR method to identify outlier indices based on residuals.

The IQR method identifies outliers as observations that fall below

Q1−1.5×IQR or above Q3+1.5×IQR, where Q1 and Q3 are the first and third quartiles, respectively, and IQR is the interquartile range.

how outliers are detected using quantiles and the interquartile range (IQR) method



### Understanding Quantiles:

Quantiles divide a dataset into equal parts. The median, for example, divides the data into two halves: the lower half and the upper half. Here’s a quick rundown of some common quantiles:

* **Median (50th percentile)**: Divides the data into two equal parts.
* **First Quartile (Q1, 25th percentile)**: Divides the lowest 25% of data from the rest.
* **Third Quartile (Q3, 75th percentile)**: Divides the lowest 75% of data from the rest.

### Interquartile Range (IQR):

The IQR is a measure of statistical dispersion, or spread, of a dataset. It is calculated as:

IQR=Q3−Q1

Where:

* Q1 is the first quartile (25th percentile).
* Q3 is the third quartile (75th percentile).

The IQR gives us an idea of how spread out the middle 50% of data is. A larger IQR indicates that the data points are more spread out from the median.

### Detecting Outliers Using IQR:

Outliers are data points that significantly differ from other observations in the dataset. The IQR method helps us identify potential outliers:

1. **Calculate IQR**:
   * First, sort the dataset from smallest to largest.
   * Find Q1 and Q3 by locating the values that mark the 25th and 75th percentiles of the sorted data.
2. **Identify Potential Outliers**:
   * Any data points below Q1−1.5×IQR or above Q3+1.5×IQR considered potential outliers.
   * Here, 1.5×IQR is a rule of thumb multiplier. It determines the range beyond which data points are flagged as potential outliers.

 Adjust the multiplier (e.g., 1.5 in 1.5 \* IQR) depending on your specific needs and dataset characteristics.

* + This method assumes normally distributed residuals, which is an underlying assumption of the IQR method for outlier detection.

1. **Mark Outliers in a Plot**:
   * Once outliers are identified, they can be visually marked in a scatter plot or any other graphical representation of the data.

Percentiles are calculated based on the position of a specific value within a sorted dataset. Here’s a brief explanation of how percentiles are calculated:

1. **Sort the Data**: Arrange the dataset in ascending order from smallest to largest.
2. **Determine the Position**: To find the percentile P, calculate its position n using the formula: n=P / 100×(N+1) Where:
   * PPP is the percentile you want to find (e.g., 25th percentile, 50th percentile, etc.).
   * NNN is the total number of data points in the dataset.
3. **Find the Value**: If n is an integer, the P th percentile is the value at position n in the sorted dataset. If n is not an integer, interpolate between the values at positions ⌊n⌋ to find the exact percentile value.

### Example:

Let's calculate the 50th percentile (which is the median) for the dataset: {65, 72, 68, 75, 83, 92, 105}.

1. **Sort the Data**: {65, 68, 72, 75, 83, 92, 105}.
2. **Calculate Position n**: n=50 / 100×(7+1)=0.5×8=4 So, n=4 (an integer).
3. **Find the Value**: The 50th percentile (median) is the value at position n=4n = 4n=4 in the sorted dataset, which is 757575.

Therefore, the 50th percentile (median) of the dataset {65, 72, 68, 75, 83, 92, 105} is 757575.

Percentiles provide a way to understand the distribution of data by dividing it into hundred equal parts, each representing the percentage of data below a given value.

### Example Calculation for a Dataset of Size 8

Consider a dataset: {12, 15, 17, 20, 22, 25, 28, 30}.

1. **Sort the Data**: {12, 15, 17, 20, 22, 25, 28, 30}.
2. **Calculate Position n**: To find the percentile P, calculate its position n: n=P / 100×(N+1)
3. For example:
   * 25th percentile (P=25): n=25 / 100× (8+1) =0.25×9=2.25. The 25th percentile position n is between the 2nd and 3rd values (since it's not an integer), so you would interpolate to find the exact value.
   * 50th percentile (Median, P=50): n=50 / 100× (8+1) =0.5×9=4.5. The median is between the 4th and 5th values.
   * 75th percentile (P=75): n=75 / 100× (8+1) =0.75×9=6.75. The 75th percentile position is between the 6th and 7th values.
4. **Interpolation (if needed)**:
   * For percentiles where n is not an integer, interpolate between the values at positions ⌊n⌋ to find the exact percentile value.

### Key Points:

* When n is an integer, the percentile value is simply the value at that position in the sorted dataset.
* When n is not an integer, interpolate between the values to find the exact percentile value.

The calculation of percentiles for a dataset of size 8 follows the same methodology as for any dataset size, adjusting N (the number of data points) accordingly in the formula to determine the position n of the percentile. This approach allows for the effective partitioning of data into specified percentiles to understand its distribution and characteristics.