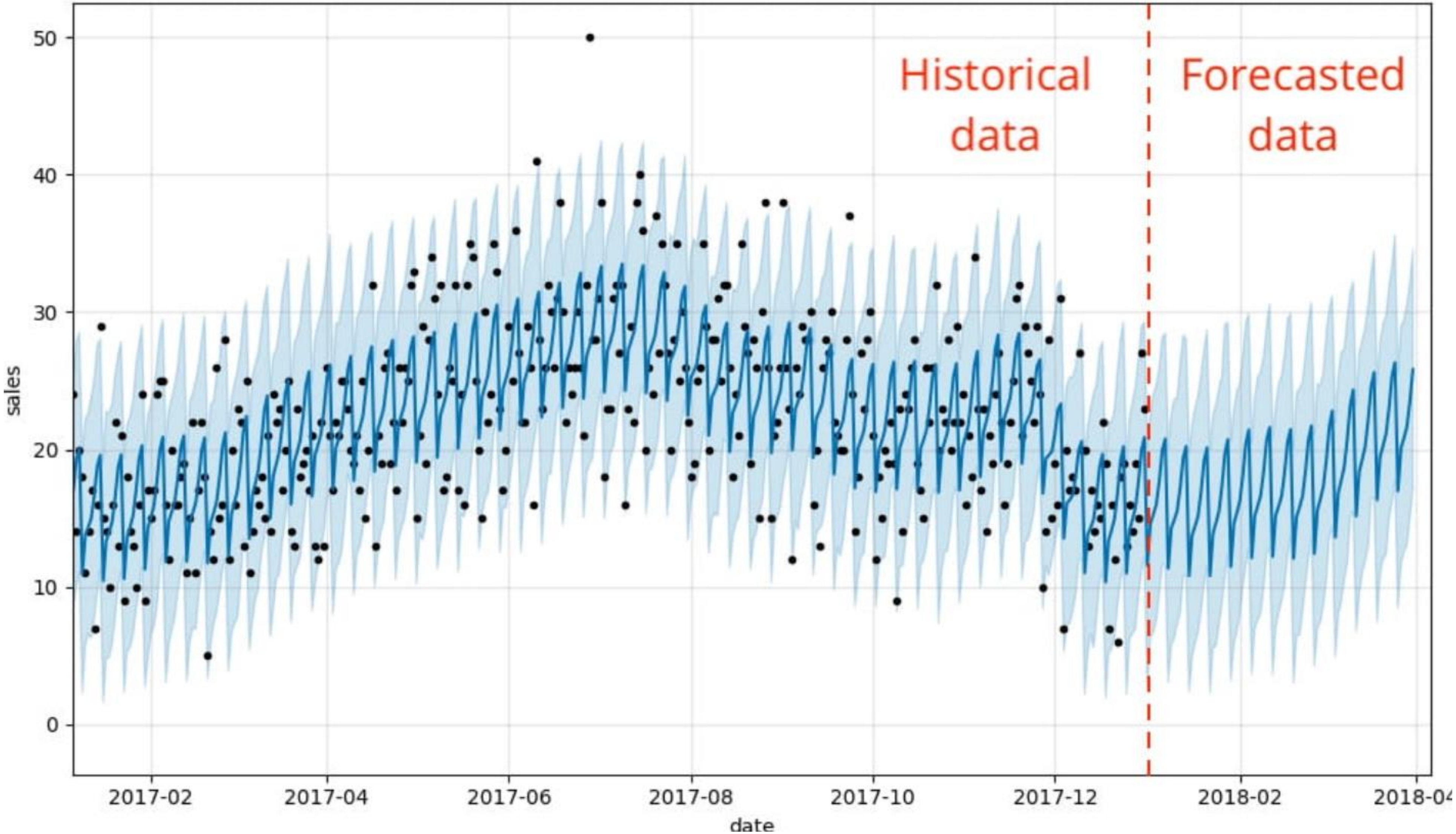
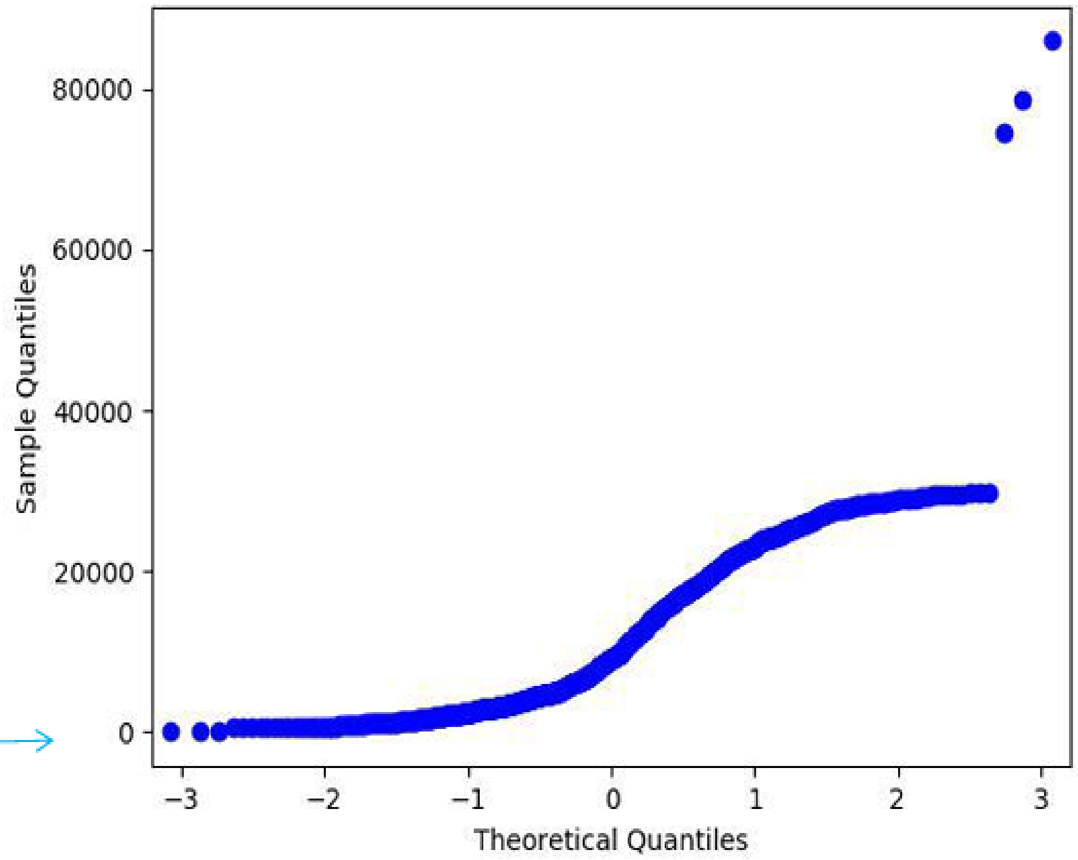
Part IV: Pre-Processing and Forecasting



Outline

- **Pre-Processing**
  - **Outlier Detection**
  - **Handling Missing Values**
- **Regression Analysis**
- **Forecasting**
- **Permutation Feature Importance**
- **Advanced Models**
  - **Generalized Additive Models (GAMs)**
  - **Neural Networks: Long Short-Term Memory Method (LSTM)**

Outliers



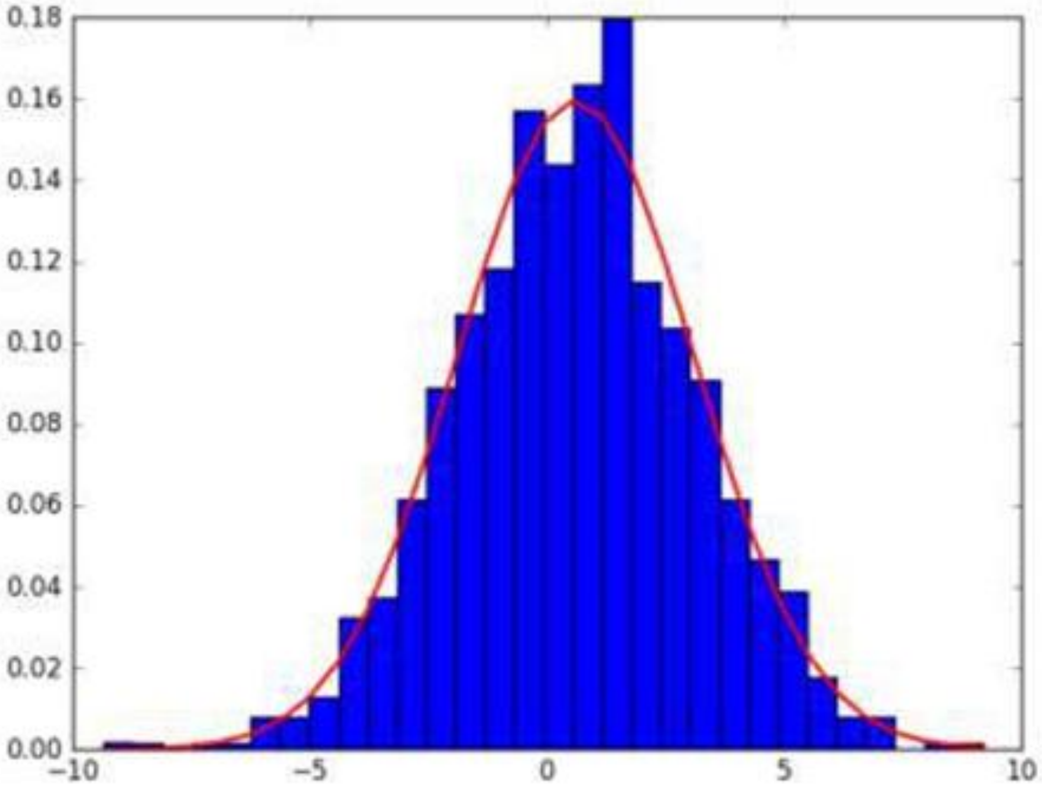
OUTLIERS

OUTLIERS

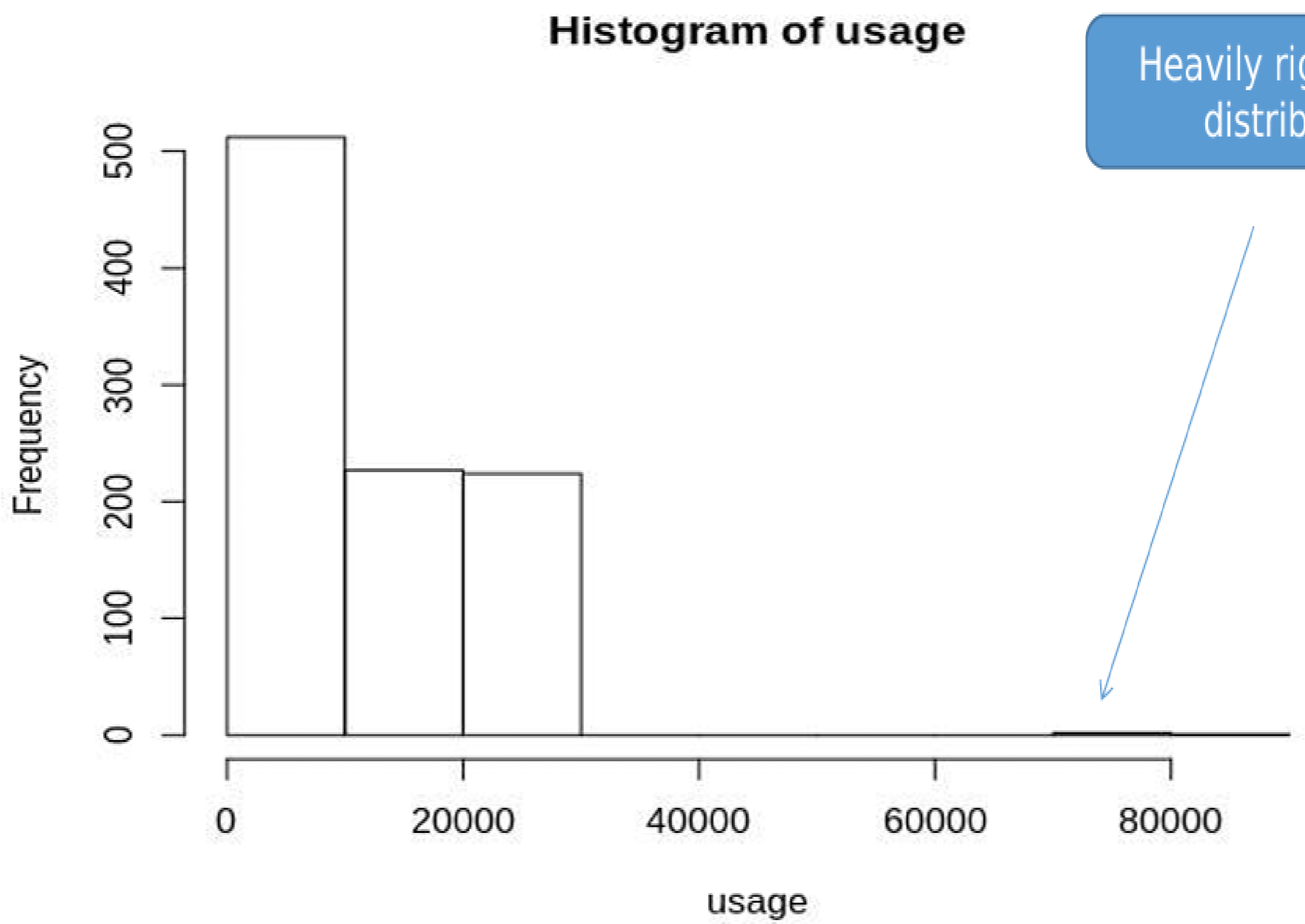
What is a normal distribution?

1) Mean = Median = Mode

2) Shape of bell curve is symmetric



A Skewed Distribution



Heavily right-tailed distribution

Is a normal distribution assumption important?

- Technically speaking, we do not NEED a normal distribution for regression analysis.
- We simply require a BLUE Estimator (Best Linear Unbiased Estimator).
- However, outliers can significantly change the shape of our distribution, and hence our overall results.

Effects of Outliers

- Skewing of mean and standard deviation
- Could significantly affect significance readings when generating regression analysis
- Can give us false readings on the magnitude of correlations

How to deal with outliers?

- Remove the outliers
- Normalize all data
- Weighting mechanism
- Keep the outliers

Missing data and sources of missing data?

University of Bremen

Missing data can arise from various places in data:

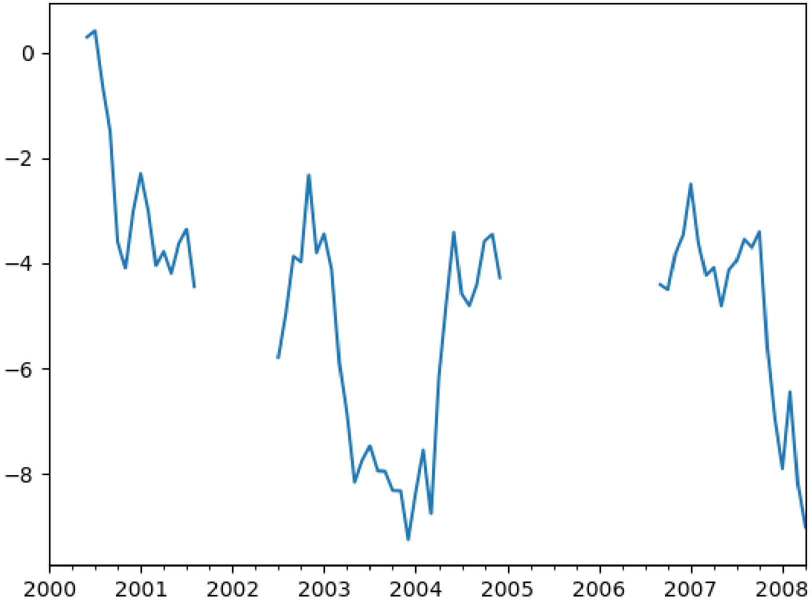
1) A survey was conducted and values were just randomly missed when being entered in the computer.

2) A respondent chooses not to respond to a question like `Have you ever recreationally used opioids?'.

3) You decide to start collecting a new variable (due to new actions: like a pandemic) partway through the data collection of a study.

4) You want to measure the speed of meteors, and some observations are just 'too quick' to be measured properly.

5) …



Handling Missing Data Using Imputation Methods

There are several different approaches to imputing missing values:

1) Impute the mean or median (quantitative) or most common class (categorical) for all missing values in a variable.

2) Create a new variable that is an indicator of missingness, and include it in any model to predict the response (also plug in zero or the mean in the actual variable).

3) Hot deck imputation: for each missing entry, randomly select an observed entry in the variable and plug it in.

4) Model the imputation: plug in predicted values from a model based on the other observed predictors.

5) …

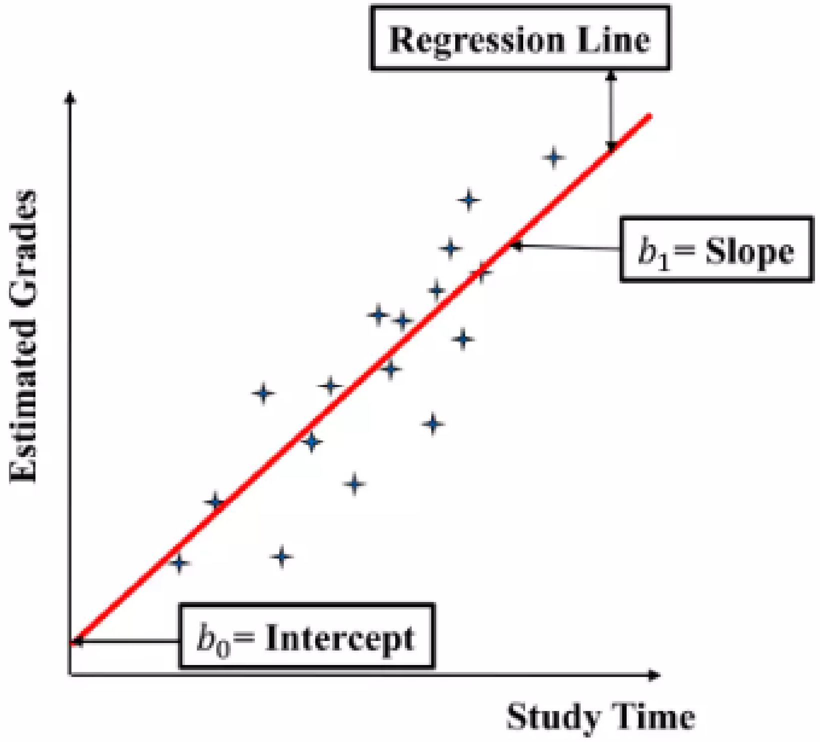
Regression Analysis

What is Regression Analysis?

- ✓ Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable**(s)** (predictor).
- ✓ This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.
- ✓ For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Population Regression Line

Example



Population regression function =

$$\hat{y} = b_0 + b_1 x$$

$\hat{y}$ = **Estimated Grades**
x = **Study Time**
b_0= **Intercept**
b_1= **Slope**

University of Bremen

Types of Regression Analysis

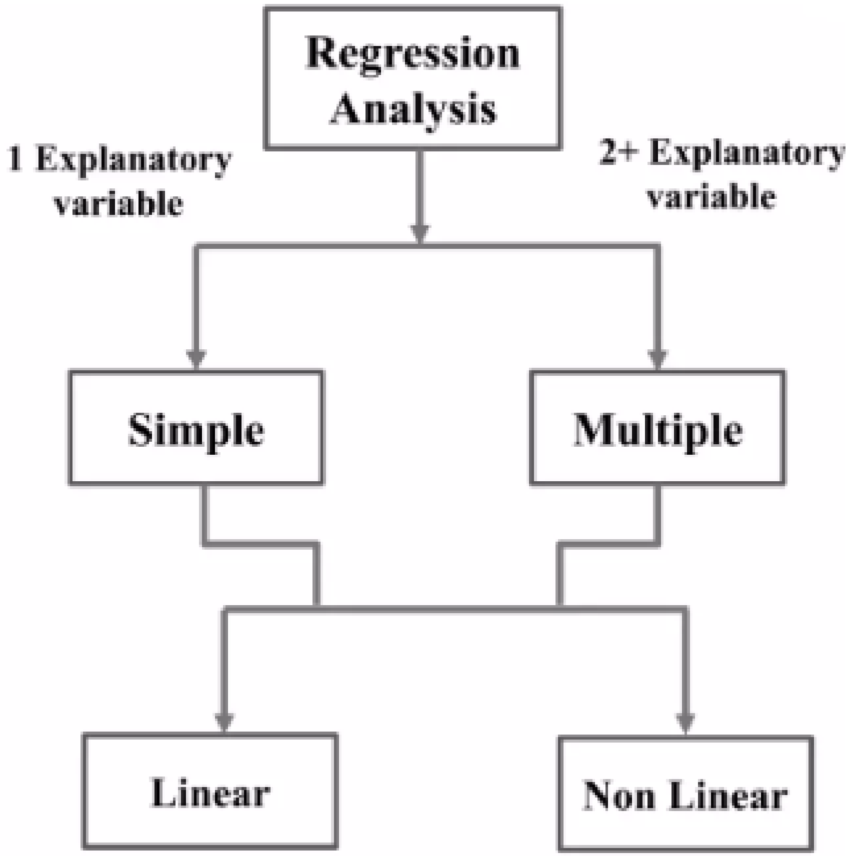
Types of regression analysis:

Regression analysis is generally classified into two kinds: simple and multiple.

Simple Regression:

It involves only two variables: dependent variable , explanatory (independent) variable.

A regression analysis may involve a **linear** model or a **nonlinear** model.
The term linear can be interpreted in two different ways:
1. Linear in variable
2. Linearity in the parameter



Simple Linear Regression Model

Simple linear regression model is a model with a single regressor x that has a linear relationship with a response y.

Simple linear regression model:

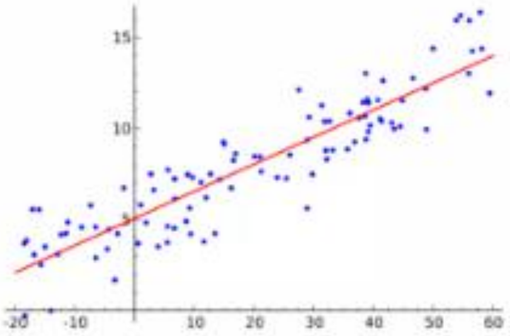
Intercept **Slope** **Random error component**

$$y = b_0 + b_1 x + \varepsilon$$

Response variable **Regressor variable**

$$\boxed{\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} \quad ; \quad \hat{b}_1 = \frac{\sum_{i=1}^{n}(yi-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(xi-\bar{x})^2}}$$

In this technique, the dependent variable is continuous and random variable, independent variable(s) can be continuous or discrete but it is not a random variable, and nature of regression line is linear.



Regression Analysis

Example

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
3	4			10	6

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(yi-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(xi-\bar{x})^2} = \frac{6}{10} = 0.6$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x} = 2.2$$

$$\boxed{\hat{y} = 2.2 + .6x}$$

Forecasting in Time Series Analysis

Overview: Forecasting is a fundamental task in time series analysis. Given a stationary time series X_t, with a mean of zero and a sample $X_{1:n}$, there are two main types of forecasting:

- **In-Sample Prediction**: Predicts values at observed time points t_0(where $1 \leq t_0 \leq n$).
  - The predicted values are called fitted values.
  - Used to evaluate the model against existing data.
- **Out-of-Sample Forecasting:** Predicts future values beyond the observed data. Important for planning and decision-making.

Forecasting Using an Innovations Algorithm

Definition

We denote the forecast of X_{n+h} as $\hat{X}_n(h)$ (h is called the *lead time*) and call it the *h-step-ahead predictor(forecast)*.

error as $e_n(h) = X_{n+h} - \hat{X}_n(h)$

For a causal ARMA(p, q) model with mean zero, the one-step-ahead predictors are recursively obtained by

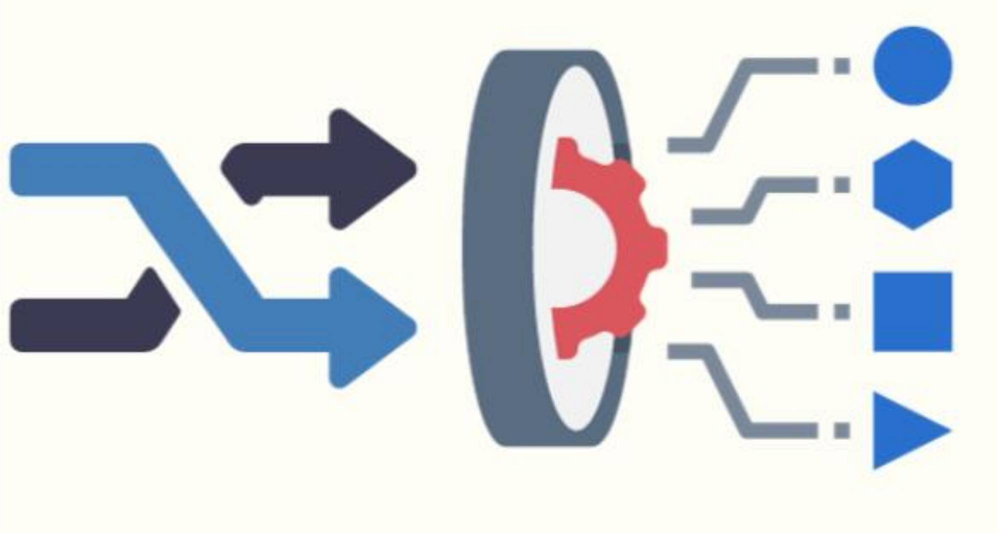
$$\hat{X}_{n+1} = \begin{cases} 0, & n = 0 \\ \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), & 1 \le n < \max(p, q) \\ \sum_{i=1}^{p} \varphi_i X_{n+1-i} + \sum_{j=1}^{q} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), & n \ge \max(p, q) \end{cases}$$

then for all $h \ge 1$, the h-step- ahead predictor is

$$\hat{X}_n(h) = \sum_{i=1}^{p} \varphi_i \hat{X}_n(h-i) + \sum_{j=h}^{q} \theta_{n+h-1,j}(X_{n+h-j} - \hat{X}_{n+h-j})$$

Permutation Feature Importance

- A technique used in machine learning to assess the **importance of different features** in a predictive model.
- Measures how model performance changes when a feature's values are randomly shuffled while keeping other variables unchanged.
- Model Agnostic: Works with both interpretable models (like linear regression) and black-box models (like neural networks).



How Permutation Feature Importance Works

- **Train the Model**: Use original features to train the model.
- **Evaluate Performance**: Measure model performance (e.g., accuracy or mean squared error).
- **Permute Feature Values**: Randomly shuffle a feature’s values.
- **Reevaluate Performance**: Measure performance with the permuted feature.
- **Calculate Importance**: The performance drop indicates the feature's importance.

Importance Metric and Randomness

- **Metric Selection:** Use metrics like accuracy, ROC-AUC (classification), or mean squared error (regression).
- **Randomness:** Results can vary due to the shuffling process.
- **Mitigation:** Shuffle each feature multiple times and average the importance values to get reliable results.

Applications of Permutation Feature Importance

- **Model Inspection:**

1) Explains model decisions.
2) Identifies critical features impacting output.

- **Feature Selection:**

1) Helps select features with higher importance.

Generalized Additive Models (GAMs)

Definition: GAMs extend Generalized Linear Models (GLMs) by allowing the relationship between predictors and the response variable to be modeled as a sum of smooth functions.

Key Features:

- **Flexibility:** Combines linear and non-linear effects.
- **Additive Structure:** Response variable Y modeled as:

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_n(X_n) + \epsilon$$

Applications:

- Commonly used in fields like ecology, epidemiology, and economics for modeling complex relationships.

Advantages and Considerations of GAMs

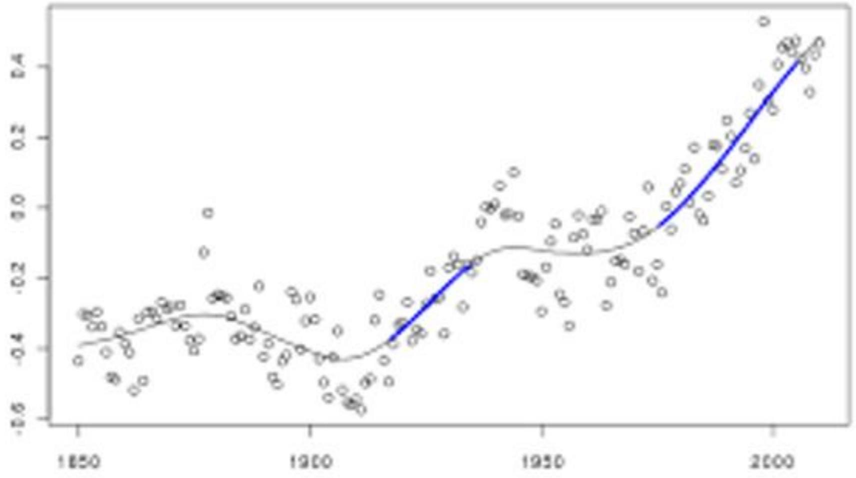
University of Bremen

Advantages:

- **Interpretability:** Each smooth term fif_ifi can be interpreted individually.
- **Handling Non-linearity:** Captures non-linear relationships without specifying a particular form.
- **Flexibility in Modeling:** Can incorporate various types of data (e.g., continuous, categorical).

Considerations:

- **Overfitting Risk:** Complex models with many smooth terms may overfit data.
- **Computational Intensity:** Requires more computational resources compared to GLMs.
- **Selection of Smoothing Parameters:** Choosing the right smoothing parameter is critical and can affect model performance.



Long Short-Term Memory (LSTM) Method

What is LSTM?

- LSTM stands for **L**ong **S**hort-**T**erm **M**emory.
- It is a type of **Recurrent Neural Network (RNN)** designed to capture long-term dependencies in sequential data.
- Capable of modeling longer term dependencies by having **memory cells** and **gates** that controls the information flow along with the memory cells.