# Coefficient for Interquantile range (IQR coefficient)

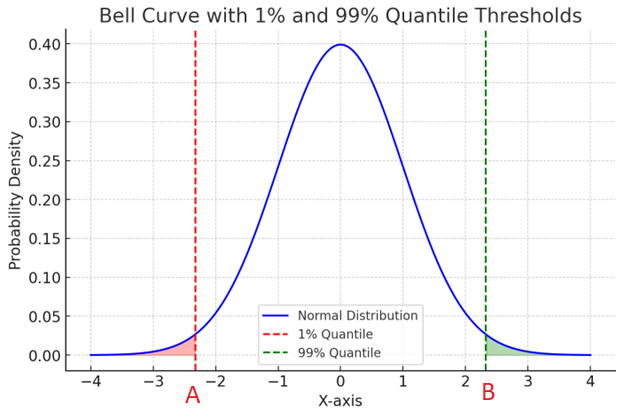
The **Coefficient of Interquartile Range (IQR Coefficient)** is a statistical measure that quantifies the relative dispersion of a dataset. It is a **normalized measure of spread**, useful for comparing variability across datasets with different units or scales.

**Formula:**

IQR\_range = B - A

Upper\_limit = B + IQR\_coef \* IQR\_range

Lower\_limit = A - IQR\_coef \* IQR\_range



**Interpretation:**

* A **higher IQR coefficient** increases the upper and lower limits, thus reduces the number of outliers.
* A **lower IQR coefficient** reduces the upper and lower limits, thus increases the number of outliers.

# Variance Inflation Factor (VIF)

The **Variance Inflation Factor (VIF)** is used to detect multicollinearity in multiple regression models. It quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors.

The formula for VIF is:



Where:

* VIF\_i ​ is the variance inflation factor for the i-th predictor.
* R\_i^2 is the coefficient of determination from the regression of predictor i on all other predictors.

**Interpretation:**

* A **VIF** value of 1 indicates no correlation between the i-th predictor and the other predictors.
* A **VIF** value greater than 1 indicates increasing correlation.
* Generally, a VIF above **5-10** suggests a problematic amount of collinearity.

# Logistic regression interpretation

## Odds ratio

The **odds ratio (OR)** is a measure used to describe the strength and direction of the association between an independent variable (predictor) and the dependent binary outcome.

**Odds Ratio (OR) Calculation:**

The odds ratio for a predictor Xj​ in a logistic regression model is derived from the **log-odds** (logit) coefficients (βj​):

OR = exp(βj)

Where:

* βj​ is the coefficient for predictor Xj​.
* exp(βj)​ is the exponentiation of the coefficient.

**Interpretation of the Odds Ratio:**

1. **OR = 1:** No effect. The predictor Xj​ has no impact on the odds of the outcome.
   * For example, if βj=0, then exp(0)=1, meaning the predictor has no influence on the odds of the outcome.
2. **OR > 1:** Positive effect. A one-unit increase in the predictor increases the odds of the event occurring (outcome = 1).
   * For example, if βj=0.5, then exp(0.5)≈1.65, meaning a one-unit increase in Xj​ increases the odds of the outcome by 65%.
3. **OR < 1:** Negative effect. A one-unit increase in the predictor decreases the odds of the event occurring (outcome = 1).
   * For example, if βj=−0.5, then exp(−0.5)≈0.61, meaning a one-unit increase in Xj​ decreases the odds of the outcome by 39%.

## ROC Curve (Receiver Operating Characteristic Curve)

**Definition:**

The ROC curve plots the **True Positive Rate (TPR, or Sensitivity)** against the **False Positive Rate (FPR)** at various classification thresholds.

* **True Positive Rate (TPR) / Recall:**



Measures how many actual positives are correctly identified.

* **False Positive Rate (FPR):**



Measures how many actual negatives are incorrectly classified as positive.

**Interpretation:**

* The **closer the ROC curve is to the top-left corner**, the better the model.
* The **Area Under the Curve (AUC-ROC)** measures overall performance (higher is better).
  + **AUC = 0.5** → Random guessing.
  + **AUC = 1.0** → Perfect model.

**Best Use Case:**

* **Balanced datasets** (where both positive and negative classes are equally represented).
* Evaluating overall model performance across all classification thresholds.

## PR Curve (Precision-Recall Curve)

**Definition:**

The PR curve plots **Precision** against **Recall** at different thresholds.

* **Precision (Positive Predictive Value):**



Measures how many predicted positives are actually positive.

* **Recall (Sensitivity or TPR):**



Measures how many actual positives are correctly predicted.

**Interpretation:**

* The **higher the PR curve**, the better the model at distinguishing positives.
* **Area Under the Precision-Recall Curve (AUC-PR)** is a better indicator of performance when **positives are rare**.

**Best Use Case:**

* **Imbalanced datasets** (e.g., fraud detection, rare disease detection).
* When **false positives** are less concerning than false negatives.
* More relevant when the **positive class is more important** than the negative class.

## F1 Score: A Balanced Metric for Classification Performance

**F1 Score** is a performance metric that balances **Precision** and **Recall**, making it useful when dealing with imbalanced datasets. It provides a single score that captures both false positives and false negatives.

**Formula for F1 Score:**

The F1 score is the **harmonic mean** of Precision and Recall:



Where:

* **Precision (Positive Predictive Value):**



Measures how many predicted positives are actually correct. Out of all transactions flagged as fraud, how many are actually fraud?

* **Recall (Sensitivity or True Positive Rate):**



Measures how many actual positives are correctly predicted. Out of all fraud cases, how many did we catch?