Supervised Learning-Regression

Determining coefficient, meaning and significance of coefficients.

In a regression analysis, coefficients represent the relationship between independent variables (predictors) and the dependent variable (outcome). Their interpretation depends on the type of regression model:

1. Linear Regression Coefficients: These represent the change in the dependent variable for a one-unit change in the independent variable, holding other variables constant.

   For example, if the coefficient of a variable (e.g., X1) is 3, it means that for every one-unit increase in X1, the dependent variable Y increases by 3 units.

2. Logistic Regression Coefficients: These coefficients (often referred to as log-odds) represent the effect of predictors on the log-odds of the outcome. The coefficients can be transformed into odds ratios for easier interpretation in probability terms.

3. Polynomial or Non-linear Models: The coefficients may represent different powers of the independent variables and can indicate curvature or more complex relationships between predictors and the outcome.

In regression models, the coefficients (also known as **parameters**) quantify the relationship between the predictors and the response variable.

Linear Regression

The equation of a simple linear regression model can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Where:

- Y = dependent (response) variable
- $X_1, X_2, \ldots, X_n$ = independent (predictor) variables
- β_0 = intercept (value of Y when all X's are zero)
- $\beta_1, \beta_2, \ldots, \beta_n$ = coefficients corresponding to each predictor X
- ϵ = error term (accounts for random variability)

The **coefficients** ($\beta_1, \beta_2, \ldots$) provide the expected change in the response Y for a one-unit increase in the corresponding predictor, assuming other predictors remain constant.

Logistic Regression

Logistic regression is used for binary outcomes (e.g., success/failure, yes/no). The model can be written as:

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

Where:

- P is the probability of the event occurring (e.g., success).
- The coefficients here represent the change in the **log-odds** of the outcome for a one-unit change in the predictor.

To make these coefficients interpretable in terms of probability, you can exponentiate them to obtain the **odds ratio**:

$$\text{Odds Ratio} = e^{\beta}$$

This represents the multiplicative change in the odds for a one-unit increase in the predictor.

The process of determining coefficients involves **estimating** them from the data. This is typically done through **optimization techniques** that minimize the error between predicted and actual values.

Ordinary Least Squares (OLS) for Linear Regression

OLS is the most common method used to estimate coefficients in linear regression. It aims to minimize the **sum of squared residuals (errors)**:

$$\text{Minimize} \sum (Y_{\text{actual}} - Y_{\text{predicted}})^2$$

The solution to this minimization problem gives the **best-fit line** that represents the relationship between predictors and the dependent variable.

Maximum Likelihood Estimation (MLE) for Logistic Regression

In logistic regression, the likelihood of the observed data is maximized. The coefficients are estimated by finding the values that maximize the likelihood of observing the data given the model.

The process of determining coefficients involves **estimating** them from the data. This is typically done through **optimization techniques** that minimize the error between predicted and actual values.

Ordinary Least Squares (OLS) for Linear Regression

OLS is the most common method used to estimate coefficients in linear regression. It aims to minimize the **sum of squared residuals (errors)**:

$$\text{Minimize} \sum (Y_{\text{actual}} - Y_{\text{predicted}})^2$$

The solution to this minimization problem gives the **best-fit line** that represents the relationship between predictors and the dependent variable.

Maximum Likelihood Estimation (MLE) for Logistic Regression

In logistic regression, the likelihood of the observed data is maximized. The coefficients are estimated by finding the values that maximize the likelihood of observing the data given the model.

Key Points of Coefficients:

- **Magnitude: Indicates the strength of the relationship between the predictor and the outcome.**
- **Sign: Indicates the direction of the relationship (positive or negative).**
- **Statistical significance: Tells whether the relationship is likely to be genuine or due to chance.**