Machine Learning in Python

Supervised Learning - Regression and Evaluation

Cristian A. Marocico, A. Emin Tatar

Center for Information Technology
University of Groningen

Wednesday, July 2nd 2025

Outline

Introduction to Regression

Regression Definition

Introduction to Regression

Regression — Definition

Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables.

Simple Linear Regression

Simple Linear Regression

Definition

Simple Linear Regression

Simple Linear Regression — Definition

Simple Linear Regression is a method to model the relationship between two variables by fitting a linear equation to observed data.

Simple Linear Regression

Simple Linear Regression — Definition

Simple Linear Regression is a method to model the relationship between two variables by fitting a linear equation to observed data. Mathematically:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where:

- y is the dependent variable (response).
- x is the independent variable (predictor).
- β_0 is the y-intercept (constant term).
- β_1 is the slope of the line (coefficient).
- ϵ is the error term (residuals).

Simple Linear Regression

A Simple Linear Regression Machine Learning model will learn the coefficients β_0 and β_1 from the training data to minimize the difference between the predicted values and the actual values.

Assumptions of Simple Linear Regression

- Linearity: The relationship between the independent and dependent variable is linear.

Assumptions of Simple Linear Regression

- Linearity: The relationship between the independent and dependent variable is linear.
- Independence: Observations are independent of each other.

Assumptions of Simple Linear Regression

- Linearity: The relationship between the independent and dependent variable is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: Constant variance of the error terms.

Assumptions of Simple Linear Regression

- Linearity: The relationship between the independent and dependent variable is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: Constant variance of the error terms.
- Normality: The residuals (errors) of the model are normally distributed.

Evaluation Metrics for Regression

Common Metrics for Regression

Definition

Evaluation Metrics for Regression

Common Metrics for Regression — Definition

Common metrics to evaluate regression models include:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R-squared (R^2)
- Adjusted R-squared

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) Definition

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) — Definition

Mean Absolute Error (MAE) is the average of the absolute differences between the predicted and the actual values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) — Definition

Mean Absolute Error (MAE) is the average of the absolute differences between the predicted and the actual values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

where:

- n is the number of observations.
- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) — Definition

Mean Absolute Error (MAE) is the average of the absolute differences between the predicted and the actual values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

where:

- n is the number of observations.
- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.

MAE is a linear score, which can be used when all errors are equally important; it is also less sensitive to outliers compared to MSE.

Mean Squared Error (MSE)

Mean Squared Error (MSE) Definition

Mean Squared Error (MSE)

Mean Squared Error (MSE) — Definition

Mean Squared Error (MSE) is the average of the squared differences between the predicted and the actual values:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Mean Squared Error (MSE)

Mean Squared Error (MSE) — Definition

Mean Squared Error (MSE) is the average of the squared differences between the predicted and the actual values:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

- n is the number of observations.
- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.

Mean Squared Error (MSE)

Mean Squared Error (MSE) — Definition

Mean Squared Error (MSE) is the average of the squared differences between the predicted and the actual values:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

- n is the number of observations.
- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.

MSE is more sensitive to outliers than MAE because it squares the errors, which can disproportionately affect the metric if there are large errors; however, it is useful when larger errors are more significant.

Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) Definition

Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) — Definition

Root Mean Squared Error (RMSE) is the square root of the average of the squared differences between the predicted and the actual values:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) — Definition

Root Mean Squared Error (RMSE) is the square root of the average of the squared differences between the predicted and the actual values:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where:

- n is the number of observations.
- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.

Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) — Definition

Root Mean Squared Error (RMSE) is the square root of the average of the squared differences between the predicted and the actual values:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where:

- n is the number of observations.
- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.

RMSE is in the same units as the dependent variable, making it interpretable; it is also sensitive to outliers, similar to MSE.

R-squared (R^2)

R-squared (R^2) — Definition

R-squared (R^2)

R-squared (R^2) — Definition

R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where:

- SS_{res} is the sum of squares of residuals (errors).
- SS_{tot} is the total sum of squares (variance of the dependent variable).

R-squared (R^2)

R-squared (R^2) Definition

R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

where:

- SS_{res} is the sum of squares of residuals (errors).
- SS_{tot} is the total sum of squares (variance of the dependent variable).

R^2 values range from 0 to 1, where:

- 0 indicates that the model does not explain any of the variability of the response data around its mean.
- 1 indicates that the model explains all the variability of the response data around its mean.

Adjusted R-squared

Adjusted R-squared

Definition

Adjusted R-squared

Adjusted R-squared — Definition

Adjusted R-squared adjusts the R^2 value for the number of predictors in the model, providing a more accurate measure when comparing models with different numbers of predictors:

$$\text{Adjusted } R^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-p-1}$$

where:

- n is the number of observations.
- p is the number of predictors in the model.

Adjusted R-squared

Adjusted R-squared — Definition

Adjusted R-squared adjusts the R^2 value for the number of predictors in the model, providing a more accurate measure when comparing models with different numbers of predictors:

$$\text{Adjusted } R^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-p-1}$$

where:

- n is the number of observations.
- p is the number of predictors in the model.

Adjusted R^2 can be negative, which indicates that the model is worse than a horizontal line (mean of the dependent variable); it is useful for comparing models with different numbers of predictors.