Glossary terms from module 2

Adjusted R^2: A variation of R^2 that accounts for having multiple independent variables present in a linear regression model

Best fit line: The line that fits the data best by minimizing some loss function or error

Causation: Describes a cause-and-effect relationship where one variable directly causes the other to change in a particular way

Confidence band: The area surrounding a line that describes the uncertainty around the predicted outcome at every value of X

Confidence interval: A range of values that describes the uncertainty surrounding an estimate

Correlation: Measures the way two variables tend to change together

Dependent variable (Y): The variable a given model estimates

Errors: The natural noise assumed to be in a regression

model

Hold-out sample: A random sample of observed data that is not used to fit the model

Homoscedasticity assumption: An assumption of simple linear regression stating that the variation of the residuals (errors) is constant or similar across the model

Independent observation assumption: An assumption of simple linear regression stating that each observation in the dataset is independent

Independent variable (X): A variable whose trends are associated with the dependent variable

Linear regression: A technique that estimates the linear relationship between a continuous dependent variable and one or more independent variables

Linearity assumption: An assumption of simple linear regression stating that each predictor variable (X_i) is linearly related to the outcome variable (Y)

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

Model assumptions: Statements about the data that must be true in order to justify the use of a particular modeling

technique

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

Negative correlation: An inverse relationship between two variables, where when one variable increases, the other variable tends to decrease, and vice versa.

Normality assumption: An assumption of simple linear regression stating that the residual values or errors are normally distributed

Ordinary least squares (OLS): A method that minimizes the sum of squared residuals to estimate parameters in a linear regression model

Outcome variable (Y): (Refer to **dependent variable**)

P-value: The probability of observing results as extreme as those observed when the null hypothesis is true

Positive correlation: A relationship between two variables that tend to increase or decrease together.

Predicted values: The estimated Y values for each X calculated by a model

R^2 (The Coefficient of Determination): Measures the proportion of variation in the dependent variable, Y,

explained by the independent variable(s), X

Residual: The difference between observed or actual values and the predicted values of the regression line

Scatterplot matrix: A series of scatter plots that demonstrate the relationships between pairs of variables

Simple linear regression: A technique that estimates the linear relationship between one independent variable, X, and one continuous dependent variable, Y

Slope: The amount that y increases or decreases per one-unit increase of x

Sum of squared residuals (SSR): The sum of the squared difference between each observed value and its associated predicted value

Terms and definitions from the previous module

A

Absolute values: (Refer to **observed values**)

C

Causation: A cause-and-effect relationship where one

variable directly causes the other to change in a particular way

D

Dependent variable (Y): The variable a given model estimates

E

Explanatory variable: (Refer to **independent variable**)

I

Independent variable (X): A variable whose trends are associated with the dependent variable

Intercept (constant B_0): The y value of the point on the regression line where it intersects with the y-axis

L

Line: A collection of an infinite number of points extending in two opposite directions

Linear regression: A technique that estimates the linear relationship between a continuous dependent variable and one or more independent variables

Link function: A nonlinear function that connects or links the dependent variable to the independent variables mathematically

Logistic regression: A technique that models a categorical dependent variable based on one or more independent variables

Loss function: A function that measures the distance between the observed values and the model's estimated values

M

Model assumptions: Statements about the data that must be true to justify the use of a particular modeling technique

N

Negative correlation: An inverse relationship between two variables, where when one variable increases, the other variable tends to decrease, and vice versa

O

Observed values: The existing sample of data, where each data point in the sample is represented by an observed value of the dependent variable and an observed value of the independent variable

Outcome variable: (Refer to **dependent variable**)

P

Positive correlation: A relationship between two variables that tend to increase or decrease together

Predictor variable: (Refer to **independent variable**)

R

Regression analysis: A group of statistical techniques that use existing data to estimate the relationships between a single dependent variable and one or more independent variables

Regression coefficient: The estimated betas in a regression model

Regression models: (Refer to **regression analysis**)

Response variable: (Refer to **dependent variable)**

S

Slope: The amount that y increases or decreases per one-unit increase of x