# Terms and definitions from Course 5

## A

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

**Accuracy**: The proportion of data points that were correctly categorized

**Adjusted R2**: A variation of R2 that accounts for having multiple independent variables present in a linear regression model

**Analysis of Variance (ANOVA)**: A group of statistical techniques that test the difference of means between three or more groups

**ANCOVA (Analysis of Covariance)**: A statistical technique that tests the difference of means between three or more groups while controlling for the effects of covariates, or variable(s) irrelevant to the test

## B

Backward elimination: A stepwise variable selection process that begins with the full model, with all possible independent variables, and removes the independent variable that adds the least explanatory power to the model

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

**Bias**: Simplifying the model predictions by making assumptions about the variable relationships

**Bias-variance trade-off**: Balance between two model qualities, bias and variance, to minimize overall error for unobserved data

**Binomial logistic regression**: A technique that models the probability of an observation falling into one of two categories, based on one or more independent variables

**Binomial logistic regression linearity assumption**: An assumption stating that there should be a linear relationship between each X variable and the logit of the probability that Y equals one

## C

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

**Chi-squared (χ²) Goodness of Fit Test**: A hypothesis test that determines whether an observed categorical variable follows an expected distribution

**Chi-squared (χ²) Test for Independence**: A hypothesis test that determines whether or not two categorical variables are associated with each other

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

**Confusion matrix**: A graphical representation of how accurate a classifier is at predicting the labels for a categorical variable

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

**Correlation**: Measures the way two variables tend to change together

## D

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

## E

**Errors**: In a regression model, the natural noise assumed to be in a model

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

**Extra Sum of Squares F-test**: A test that quantifies the difference between the amount of variance that is left unexplained by a reduced model that is explained by the full model

## F

**Feature selection**: (Refer to **variable selection**)

**Forward selection**: A stepwise variable selection process that begins with the null mode—with 0 independent variables—considers all possible variables to add; incorporates the independent variable that contributes the most explanatory power to the model

## H

**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

**Hypothesis testing**: A statistical procedure that uses sample data to evaluate an assumption about a population parameter

## I

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

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

**Interaction term**: The term that represents how the relationship between two independent variables is associated with changes in the mean of 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

**Likelihood**: The probability of observing the actual data, given some set of beta parameters

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

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

**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 (Y) based on one or more independent variables (X)

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

**Logit**: The logarithm of the odds of a given probability

**Log-Odds function**: (Refer to **logit**)

## M

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

**MANCOVA (Multivariate Analysis of Covariance)**: An extension of ANCOVA and MANOVA that compares how two or more continuous outcome variables vary according to categorical independent variables, while controlling for covariates

**MANOVA (Multivariate Analysis of Variance)**: An extension of ANOVA that compares how two or more continuous outcome variables vary according to categorical independent variables

**Maximum Likelihood Estimation (MLE)**: A technique for estimating the beta parameters that maximize the likelihood of the model producing the observed data

**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

**Multiple linear regression**: A technique that estimates the relationship between one continuous dependent variable and two or more independent variables

**Multiple regression**: (Refer to **multiple linear regression**)

## N

**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 residuals are normally distributed

**No multicollinearity assumption**: An assumption of simple linear regression stating that no two independent variables (Xi and Xj) can be highly correlated with each other

## 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

**One hot encoding**: A data transformation technique that turns one categorical variable into several binary variables

**One-Way ANOVA**: A type of statistical testing that compares the means of one continuous dependent variable based on three or more groups of one categorical variable

**Ordinary least squares estimation (OLS)**: A common way to calculate linear regression coefficients

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

**Overfitting**: When a model fits the observed or training data too specifically and is unable to generate suitable estimates for the general population

## P

**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

**Post hoc test**: An ANOVA test that performs a pairwise comparison between all available groups while controlling for the error rate

**Precision**: The proportion of positive predictions that were true positives

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

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

## R

**R2 (The Coefficient of Determination)**: The coefficient that measures the proportion of variation in the dependent variable, Y, explained by the independent variable(s), X

**Recall**: The proportion of positives the model was able to identify correctly

**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**)

**Regularization**: A set of regression techniques that shrinks regression coefficient estimates towards zero, adding in bias, to reduce variance

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

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

## S

**Scatterplot matrix**: A series of scatterplots that show 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

## T

**Two-Way ANOVA**: A type of statistical testing that compares the means of one continuous dependent variable based on three or more groups of two categorical variables

## V

**Variable selection**: The process of determining which variables or features to include in a given model

**Variance**: Model flexibility and complexity, so the model learns from existing data

**Variance inflation factors (VIF)**: Factors that quantify how correlated each independent variable is with all of the other independent variables