MODULE 3

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

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Two ways to handle categorical data:

- One hot encoding
  Data transformation technique that turns one categorical variable into several binary variables
- Label encoding

Multiple linear regression assumptions:

- Linearity
- Independent Observations
- Normality
- Homoscedasticity
- **No multicollinearity assumption**
  No two independent variables (x and y) can be highly correlated with each other
  Variance Inflation Factors (VIF)
  Quantifies how correlated each independent variable is with all of the other independent variables
  How to handle data with multicollinearity:
  - Drop one or more variables that have high multicollinearity
  - Create new variables using existing data

Interaction term
A term that represents how the relationship between two independent variables is associated with changes in the mean of the dependent variable

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_i (X_1 * X_2) + \varepsilon$$

Where β_i is coefficient associated with interaction term

Data professionals divide the sample data into **two categories**:

- **training data**
  - used to build the model
- **test data**
  - used to evaluate the model's performance
- The method are called **holdout sampling**. Enables data professionals to evaluate how a model performs on data it has not experienced yet
- To identify overfitting

Underfitting

- Multiple regression model fails to capture the underlying pattern in the outcome variable
- Has low R^2 value
- Reasons for the model underfit:
  - The independent variables might not have strong relationship with the outcome variable
  - The sample dataset is too small (prevents the model to learn the relationship between predictors and outcome)

Overfitting

- Multiple regression model fits the observed or training data too specifically, and is unable to generate suitable estimates for the general population

- Its performance is worse when evaluated using the unseen test data
- Tends to occur when a model is too complex or incorporates too many variables
- R^2 will increase when more independent variables are added to the model. High R^2 value is not enough to indicate that the model will perform well

Adjusted R^2

- A variation of the R^2 regression evaluation metric that penalizes unnecessary explanatory variables
- Used to compare models of varying complexity
  - determine if you should add another variable or not
- R^2 is more easily interpretable
  - determine how much variation in the dependent variable is explained by the model

Model that underfits data is described as having **high bias** and model that does not perform well on new data is described as having **high variance**.

The phenomenon is known as **bias versus variance tradeoff**.

The tradeoff is a dilemma faced when building machine learning model because an ideal model should have low bias and low variance

Variable selection

The process of determining which variables or features to include in a given model

- **Forward elimination**
  Begins will 0 independent variables and considers all possible variables to add
  Incorporates the independent variable that contributes the most explanatory power
- **Backward elimination**
  Begins with full model and removes the independent variable that adds the least explanatory power
- Requires threshold to determine when to add or remove variables

Extra-sum-of-squares F-test (based on p-value)

Quantifies the difference between the amount of variance that is left unexplained by a reduced model that is explained by the full model

Bias-variance tradeoff

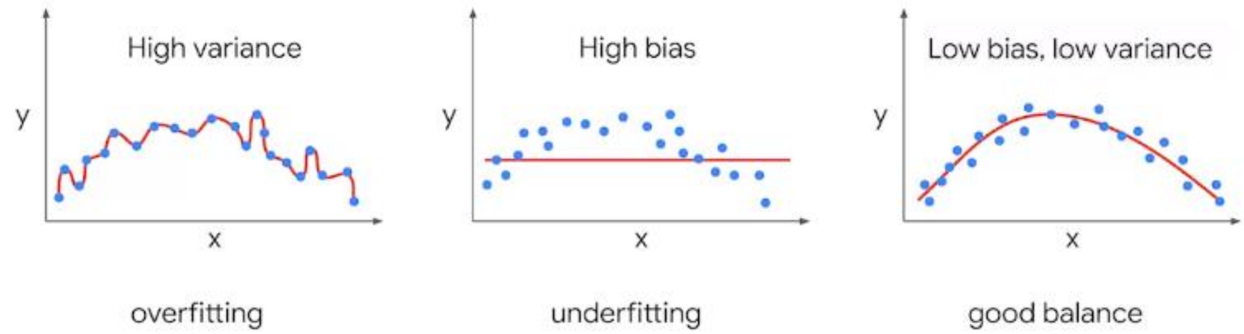
Balance between bias and variance to minimize overall error for unobserved data

Bias

- Simplifies the model predictions by making assumptions about the variable relationships
- Highly biased model oversimplify the relationship, underfitting the observed data and generate inaccurate estimates

Variance

- Model flexibility and complexity
- High variance overfit the observed data and generate inaccurate estimates for unseen data



Regularization

- A set of regression techniques that shrinks regression coefficient estimates toward zero, adding bias to reduce variance
- Avoid the risk of overfitting
- Types of regularized regression:
    - Lasso regression
    - Ridge regression
    - Elastic-net regression