Introduction to Predictive Modeling

Overview

1. What is predictive modeling?
2. Linear Models
3. Parameter Estimation

Motivation

You are a Data Scientist

- working in healthcare. You are tasked with building a predictive model to extract insights using medical records to predict patients' outcomes.
- working for an NBA team. The team is about to play in an NBA finals game and you are tasked with building a model to predict the probability of a win based on several other variables.

What is predictive modeling?

Predictive modeling is the process of developing a mathematical tool or model that generates an accurate prediction about a random quantity of interest.

- In predictive modeling we are interested in predicting a random variable, namely the **response** variable, typically denoted by Y, from a set of related variables, namely **predictors** or **regressors**, $X_1, X_2, \cdots, X_p$.The focus is on learning what is the probabilistic model that relates Y with $X_1, X_2, \cdots, X_p$ and use that acquired knowledge for predicting Y given an observation of $X_1, X_2, \cdots, X_p$.

Usefulness of models

"All models are wrong, some are useful." George Box, 1976.

Many models are never used, for several reasons including:

- it was not deemed relevant to make predictions in the setting envisioned by the authors
- potential users of the model did not trust the relationships, weights, or variables used to make the predictions
- the variables necessary to make the predictions were not routinely available

Linear Models

Multiple Linear Model

The multiple linear model is a simple but useful statistical model. In short, it allows us to analyze the (assumed) linear relation between a response Y and multiple predictors $X_1, X_2, \cdots, X_p$ in the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

where ε is a random variable and β_i are unknown parameters.

- The simplest case is when $p = 1$, known as the Simple Linear Regression:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

The linear regression model can also be expressed in a matrix form.

Simple Linear Model

For an individual observation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where

- β_0 is the population y-intercept (interpreted as the initial value)
- β_1 is the population slope (interpreted as the is the increment in the mean of Y for an increment of one unit in X
- ε_i is the error or deviation of y_i from the line $\beta_0 + \beta_1 x_i$

Parameter Estimation

Loss Function

A **loss function** is a real-valued function of two variables, $\mathrm{L}(\theta, a)$, where θ is a parameter and a is a real number. The interpretation is that the Data Scientist losses $\mathrm{L}(\theta, a)$ if the parameter equals θ and the estimate equals a.

- The squared Error Loss Function: $\mathrm{L}(\theta, a) = (\theta - a)^2$.
- The Absolute Error Function: $\mathrm{L}(\theta, a) = |\theta - a|$.

Simple Linear Model: Parameter Estimation

For an individual observation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Three approaches to consider:

- Least Square: Minimizing the sum of square of residuals
- Maximum likelihood estimation
- Bootstrapping

Simple Linear Model: Least Square Parameter Estimation

For an individual observation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Goal is to minimize the square of the residual of the i-th observation. That is, minimize

$$SSR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Simple Linear Model: Maximum likelihood estimation

For an individual observation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Assume $\varepsilon_i \sim N(0, \sigma^2)$, which means $y_i \mid x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
- Maximize the likelihoodness of obtaining β_0 and β_1 given the observed data.

Simple Linear Model: Bootstrapping

For an individual observation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Generate Bootstrap samples from the data
- Fit regression model for each bootstrap sample
- Find summaries of the estimates, including confidence intervals