Report

2023-06-02

## Scenario 1: Nonlinear Regression

In this scenario, we are studying the growth of a specific species of plant over time. We hypothesize that the plant’s growth follows a nonlinear pattern and is influenced by environmental factors such as temperature and humidity. We collect data on the plant’s height and the corresponding environmental conditions at regular intervals.

To generate synthetic data, we can create a function that incorporates a nonlinear growth pattern influenced by temperature and humidity. We can specify the range of temperature and humidity values and generate corresponding plant height values using the defined function. The synthetic data will consist of temperature, humidity, and plant height measurements over time.

# Generate synthetic data for nonlinear regression  
set.seed(123)  
  
# Create temperature and humidity values  
temperature <- seq(10, 30, by = 2)  
humidity <- seq(30, 70, by = 5)  
  
# Define a nonlinear growth function  
growth\_function <- function(temperature, humidity) {  
 height <- 5 + 0.5 \* temperature + 0.2 \* humidity + rnorm(1, mean = 0, sd = 1)  
 return(height)  
}  
  
# Generate synthetic data  
data <- expand.grid(temperature = temperature, humidity = humidity)  
data$height <- sapply(1:nrow(data), function(i) growth\_function(data[i, "temperature"], data[i, "humidity"]))  
  
# View the synthetic data  
head(data)

## temperature humidity height  
## 1 10 30 15.43952  
## 2 12 30 16.76982  
## 3 14 30 19.55871  
## 4 16 30 19.07051  
## 5 18 30 20.12929  
## 6 20 30 22.71506

## Scenario 2: Nested or Clustered Data Analysis

In this scenario, we are investigating the academic performance of students in different schools across various districts. We suspect that academic performance may vary not only between schools but also within schools, depending on factors such as teacher quality and resources. The data consists of students’ test scores, school information, and district information.

To generate synthetic data, we can create a hierarchical structure where students are nested within schools, and schools are nested within districts. We can assign random effects to each level to capture the variability between districts, schools, and individual students. The synthetic data will include student-level variables such as test scores, school-level variables such as teacher quality, and district-level variables such as average income.

# Generate synthetic data for nested or clustered data analysis  
set.seed(123)  
  
# Define number of districts, schools, and students  
num\_districts <- 5  
num\_schools\_per\_district <- 10  
num\_students\_per\_school <- 50  
  
# Create district, school, and student IDs  
district\_ids <- rep(1:num\_districts, each = num\_schools\_per\_district \* num\_students\_per\_school)  
school\_ids <- rep(rep(1:num\_schools\_per\_district, each = num\_students\_per\_school), times = num\_districts)  
student\_ids <- rep(1:num\_students\_per\_school, times = num\_schools\_per\_district \* num\_districts)  
  
# Generate random effects for districts, schools, and students  
district\_effects <- rnorm(num\_districts, mean = 0, sd = 1)  
school\_effects <- rnorm(num\_districts \* num\_schools\_per\_district, mean = 0, sd = 1)  
student\_effects <- rnorm(num\_districts \* num\_schools\_per\_district \* num\_students\_per\_school, mean = 0, sd = 1)  
  
# Generate synthetic data  
data <- data.frame(district = district\_ids, school = school\_ids, student = student\_ids,  
 district\_effect = district\_effects, school\_effect = school\_effects, student\_effect = student\_effects)  
data$test\_score <- 70 + 5 \* data$district\_effect + 2 \* data$school\_effect + 3 \* data$student\_effect + rnorm(nrow(data), mean = 0, sd = 5)  
  
# View the synthetic data  
head(data)

## district school student district\_effect school\_effect student\_effect  
## 1 1 1 1 -0.56047565 1.7150650 1.5164706  
## 2 1 1 2 -0.23017749 0.4609162 -1.5487528  
## 3 1 1 3 1.55870831 -1.2650612 0.5846137  
## 4 1 1 4 0.07050839 -0.6868529 0.1238542  
## 5 1 1 5 0.12928774 -0.4456620 0.2159416  
## 6 1 1 6 -0.56047565 1.2240818 0.3796395  
## test\_score  
## 1 68.20113  
## 2 73.81840  
## 3 73.02409  
## 4 73.17552  
## 5 71.99250  
## 6 65.46670

## Scenario 3: Regularization Methods

In this scenario, we are analyzing a dataset with multiple predictors to predict the risk of heart disease in patients. The dataset includes various demographic, lifestyle, and medical variables, such as age, gender, smoking status, blood pressure, and cholesterol levels. We want to develop a predictive model that effectively accounts for the complex relationships among the predictors.

To generate synthetic data, we can simulate a dataset with a mix of continuous and categorical predictors. We can specify the distribution and correlation structures among the predictors based on prior knowledge or assumptions. Additionally, we can introduce some noise to represent the variability in real-world data. The synthetic data will include multiple predictors and a binary response variable indicating the presence or absence of heart disease.

# Generate synthetic data for regularization methods  
set.seed(123)  
  
# Create predictor variables  
age <- runif(100, min = 20, max = 60)  
gender <- sample(c("Male", "Female"), size = 100, replace = TRUE)  
smoking\_status <- sample(c("Non-smoker", "Smoker"), size = 100, replace = TRUE)  
blood\_pressure <- rnorm(100, mean = 120, sd = 10)  
cholesterol <- rnorm(100, mean = 200, sd = 20)  
  
# Generate synthetic data  
data <- data.frame(age = age, gender = gender, smoking\_status = smoking\_status,  
 blood\_pressure = blood\_pressure, cholesterol = cholesterol)  
data$risk\_of\_heart\_disease <- ifelse((0.5 \* age + 0.2 \* blood\_pressure + 0.1 \* cholesterol + rnorm(100)) > 0, 1, 0)  
  
# View the synthetic data  
head(data)

## age gender smoking\_status blood\_pressure cholesterol  
## 1 31.50310 Male Smoker 127.8774 192.4879  
## 2 51.53221 Female Smoker 127.6904 188.7625  
## 3 36.35908 Female Smoker 123.3220 193.1217  
## 4 55.32070 Male Non-smoker 109.9162 201.8099  
## 5 57.61869 Female Smoker 118.8055 231.9702  
## 6 21.82226 Female Smoker 117.1960 198.2287  
## risk\_of\_heart\_disease  
## 1 1  
## 2 1  
## 3 1  
## 4 1  
## 5 1  
## 6 1

## Controlling for Random Effects in Nested or Clustered Data:

To control for random effects in nested or clustered data, we can employ mixed-effects models. These models allow us to account for the hierarchical structure of the data by incorporating random effects at different levels. In the scenario of student performance in different schools, we can include random intercepts for schools and districts, accounting for the variability between them. This approach considers the correlation within each level and provides more accurate estimates and inference.

Value of Mixed-Effect Models in Analyzing Nested or Crossed Data Structures: Mixed-effect models are valuable in analyzing nested or crossed data structures because they can handle the dependency and variability within different levels of the data. These models allow us to model the fixed effects (relationships between predictors and the response variable) while considering the random effects (variability between different levels). By accounting for random effects, we can estimate the effects of predictors more accurately and make valid statistical inferences. Mixed-effect models also provide insights into the variability between groups and allow for capturing the hierarchical nature of the data, providing a more comprehensive understanding of the underlying relationships.