Homework 2

February 9, 2022

(I). Consider the "Pima Indians Diabetes Database" (`https://www.kaggle.com/uciml/pima-indians-diabetes-database`). This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. All patients here are females at least 21 years old of Pima Indian heritage. The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables include: Pregnancies, Glucose, Blood pressure, Skin thickness, Insulin, BMI, Diabetes pedigree function and Age. The data includes 768 observations.

1. For the variable "Pregnancies",can we conclude that the mean of the population from which the sample was drawn is greater than 8?

2. By repeatedly drawing multiple random samples from the variable "Pregnancies", please plot the sampling distribution (histogram and boxplots) of the proposed test statistic in part 1.

3. For the variable "Outcome", can we claim that the proportion of diabetes women in the population is different from 0.5.

4. Repeat step 2 for the variable and the test statistic used in part 3.

(II). Suppose we observe data (Y_i, X_i, Z_i), $i = 1, \ldots, n$, from the following regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

where ϵ_i's are random. Find the expressions for $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\beta}_2$. Show that the sum of residuals $\sum_{i=1}^n \epsilon_i$ and the sum of errors $\sum_{i=1}^n \widehat{\epsilon}_i$ are equal to zeroes.