Risk of Osteoporosis Based on Dietary Vitamin D and Calcium (NHANES 2017-2018)

A. Renganathan & Y.Baldonado

2023-04-07

Abstract

Introduction

Logistic Regression is used when trying to model a response that is binary, not continuous, and has an outcome of 0/1 or yes/no. Unlike typical regression models, logistic regression transforms the model into a probability model for the outcome, in other words the probability that something will occur based on the predictors in the model. For this project, we are trying to assess the risk of Osteoporosis diagnosis based on a variety of predictors. Logistic regression is also a great method for modeling our question as it allows us to test both continuous and categorical predictors against a binary response.

Objectives

The objective of this analysis is to identify which dietary factors affect risk of osteoporosis diagnosis in older adults.

Methodology

Data Source

For this project, we will be using the National Health and Nutrition Examination Survey (NHANES) provided by the Center for Disease Control (CDC) to obtain our data. NHANES is program of study that has been ongoing since the 1960's and houses health and nutritional information for adults and children in the United States. The survey includes a wide range of information from Demographics data, Dietary data, Laboratory data, Questionnaire data and Limited Access data.

R Studio

Using the R package RNHANES, we were able to extract the pertinent information needed for our analysis directly from the website. From the package, we extracted Demographic, Dietary, Laboratory, and Questionnaire data from the years 2017-2018. Specifically, the following were pulled:

- Age, Gender and Race/Ethnicity from the demographic files
- Calcium Intake, Vitamin D intake, Calcium Vitamin Supplementation, Vitamin D Supplementation from the Dietary files

- Dairy and Nondairy milk consumption, whether participants have fractured one of their wrists before, and whether they had Osteoporosis from the Questionnaire files After some data manipulation and removing unnecessary information, we were left with 3,069 observations. The continuous predictors were Age, Calcium Intake, Vitamin D intake, Calcium Vitamin Supplementation, Vitamin D Supplementation and the categorical predictors were Gender, Race/Ethnicity, Dairy milk consumption, Nondairy milk consumption, Fractured Wrist, and Osteoporosis.

Data Management

Survey Weighting

Analysis

Exploratory Analysis

We first ran a simple logistic regression to test each variable against our response Osteoporosis to see which variables were individually significant. We found that Age, Gender (Female), Calcium Intake, and Calcium Supplementation were significant.

Descriptive Statistics

Model Assumptions

Independent Observations

Multicollinearity

Linearity of Continuous Variables Across Strata

Outliers

Variable Selection

Multivariate Logistic Regression

Initial Model

Subset Models

Best Fit Model

Results

Interpretation of everything here.

Conclusions

Limitations

Future Research

References