Assignment 3

Biostatistics 1:Introduction to biostatistics, 5BD000

December 2024

Introduction

Within this assignment you will analyse data on 487 births. The purpose is to establish if there is a causal effect of smoking during pregnancy on the risk of the child being born with low birth weight (below 2500 grams). You have been given two datasets, birthweight_original.csv and birthweight_analysisdata.csv. The first dataset is the raw dataset, and the second is a dataset that has been cleaned to do the analyses. The main difference between the datasets is that the some of the continuous covariates in the original data has been categorised into binary covariates in the analysis data. The covariates in the two datasets are described below. The data are based on the dataset clslowbwt.dta used in the book "An introduction to Stata for health researchers" (Juul and Frydenberg, 2010), but has been modified to suit this assignment.

Your task is to analyse the data using logistic regression, including the right covariates in the model to be able to estimate the total causal effect of smoking during pregnancy on the risk of low birth weight of the child. To do this you will have to create a DAG to decide the relationship between the variables based on current knowledge in the literature. Use for example the sources presented in the workshop by the libraries to find relevant information. For this assignment you can assume that there are no unmeasured covariates that are of importance, although, in reality that is unreasonable.

You should hand in a written report of your work as a pdf. The report should include descriptive statistics of all important covariates, a DAG for the specific research question (you can assume that there is no unmeasured confounding), a motivation for each arrow in the DAG (and if necessary the absence of an arrow), an explanation of which covariates to include in the model based on your chosen DAG, an output of the results from the final model as well as a summary of the final result with an interpretation. Also remember to include any references, and a description and motivation of any potential changes you've made to the data. This is an individual assignment, and should be uploaded in Canvas by January 14th 2025. The motivations and interpretation are more important than the final results, so clearly motivate the choices you make.

Variable name	Description
birth	Birth order of the child, e.g. if it is the woman's first birth the value will be 1
smoke	Smoking during pregnancy, 0=No, 1=Yes
age	Age of mother
bwt	Birth weight of child measured in grams
id	Sequence number randomly assigned to each woman
bp_sys	Systolic blood pressure before pregnancy, mmHg
preterm	Child born preterm, 0=No, 1=Yes
neocare	Child requiring neonatal care, 0=No, 1=Yes

Table 1: Codebook of dataset birthweight_original.csv

Variable name	Description
birth	Birth order of the child, e.g. if it is the woman's first birth the value will be 1
smoke	Smoking during pregnancy, 0=No, 1=Yes
age	Age of mother
low	Binary variable for low birth weight of child (below 2500g), 0=No, 1=Yes
id	Sequence number randomly assigned to each woman
high_bp	High systolic blood pressure before pregnancy (135 or above) , 0=No, 1=Yes
preterm	Child born preterm, 0=No, 1=Yes
neocare	Child requiring neonatal care, 0=No, 1=Yes

Table 2: Codebook of dataset birthweight_analysisdata.csv