AssignmentB

Martijn Koster, William Schaafsma, Martijn van Dam, Victor Hovius

3/21/2022

Contents

1 Introduction

The matter of interest for this assignment will be the impact that incomplete data (**observed data**) have on our inferences compared to the inferences we make with complete data (**true data**). To investigate the effect that missing values have on model inferences, we will build a random multiple regression model.

Firstly, we provide descriptive statistics and correlations. In Table 1 and Table 2 we compare the head of the observed data and the true data. Additionally, in Table 3 the means and variances are compared. With regard to correlations, we present two correlations matrices; one for the observed data in Table **??** and the other for the true data in Table **??**.

Secondly, we present our multiple regression model in Table **??**. Our model consists of the outcome variable: *active heart rate* and the predictors: *age* and *smoke*. We also included an interaction effect between *bmi* and *sex*. The first three columns reflect the observed data, whereas the latter reflect the true data.

The research question we try to answer in accordance with our model is: *What impact do missing values have on an "active heart rate" model inference?*

Thirdly, we start by inspecting the missing values. Then, we try to find out where the missing values occur. In **??** we begin by giving a global overview of the missingness. Then, in **??** we compare the distributions for the observed data and the missing values.

Lastly, we perform t-tests on the variables containing missing values to check the type of missingness, either MNAR, MAR, or MCAR. We also provide plots here to visualize where the missing values occur.

2 Observed vs True data

In this section we will compare the observed with the true data set.

Table 1: First Five Cases of Observed Data

age	smoke	sex	intensity	active	rest	height	weight	bmi
42	no	female	high	NA	75	NA	NA	22.4
31	NA	male	low	NA	62	NA	NA	23.8
36	no	male	low	109	76	182	78.0	23.5
31	no	female	low	78	62	164	53.9	20.0
42	no	male	low	NA	66	189	NA	23.4

Table 2: First Five Cases of True Data

age	smoke	sex	intensity	active	rest	height	weight	bmi
42	no	female	high	94	75	161	58.1	22.4
31	no	male	low	86	62	184	80.6	23.8
36	no	male	low	109	76	182	78.0	23.5
31	no	female	low	78	62	164	53.9	20.0
42	no	male	low	103	66	189	83.6	23.4

2.1 Descriptives

Obviously, neither the mean nor the variance of the variables age and rest changed since they have no missing values.

The mean of active is also almost entirely unaffected. The variance of active changed a bit in the observed data, but this difference is simply due to sampling variability (we have deleted about 40% of the observations). The missing values in active are MCAR, so we would not expect any substantial changes in the marginal distribution of active.

The mean of height is also almost entirely unaffected. The variance of active changed a bit in the observed data, but this difference is simply due to sampling variability (we have deleted about 30% of the observations). The missing values in height are MCAR, so we would not expect any substantial changes in the marginal distribution of height.

The mean of weight is also almost entirely unaffected. The variance of active changed a bit in the observed data, but this difference is simply due to sampling variability (we have deleted about 57% of the observations). The missing values in weight are MCAR, so we would not expect any substantial changes in the marginal distribution of weight.

The mean of bmi is also almost entirely unaffected. The variance of active changed a bit in the observed data, but this difference is simply due to sampling variability (we have deleted about 30% of the observations). The missing values in bmi are MCAR, so we would not expect any substantial changes in the marginal distribution of bmi.

Furthermore, the variance of the variables age and rest are unaffected in the observed data set. The variance of the variable active in the observed data set is .01 lower than the true data set, and thus almost entirely unaffected. However the variables height, weight, and bmi have greater variance in the true data set than the observed data set. This implies that the missingness causes an underestimation of the variance.

2.2 Categorical variables

The categorical variables in both data sets are *smoke*, *sex* and *intensity*. *Smoke* and *sex* both have two levels (“no” and “yes” for smoke and “male” and “female” for sex), while *intensity* has three levels (“low”, “moderate”, and “high”).

Table 3: Means and variances in true and observed dataset

Variables	M obs	M true	var obs	var true	N obs	N true
Age	38.52	38.52	149.73	149.73	306	306
Active	92.58	93.13	383.05	383.04	183	306
Rest	69.83	69.83	120.78	120.78	306	306
Height	174.50	173.99	100.66	105.29	214	306
Weight	73.91	73.58	260.26	274.85	132	306
Bmi	24.11	24.06	12.91	13.38	213	306

Note.
obs = Observed Dataset, true = True Dataset

Table 4: proportion table of categorical in observed and true data

sex	smoke	intensity
male	no	high
female	no	high
male	yes	high
female	yes	high
male	no	moderate
female	no	moderate
male	yes	moderate
female	yes	moderate
male	no	low
female	no	low
male	yes	low
female	yes	low

Note. makecell[l]On the left side of the table the proportions of the observed data are shown, whereas the proportions of t

Table 5: proportion table of categorical in observed and true data

sex	smoke	intensity
male	no	high
female	no	high
male	yes	high
female	yes	high
male	no	moderate
female	no	moderate
male	yes	moderate
female	yes	moderate
male	no	low
female	no	low
male	yes	low
female	yes	low

Note. makecell[l]On the left side of the table the proportions of the observed data are shown, whereas the proportions of t

Despite differences in the number of observed values between the data sets, differences between groups remain unchanged. For example, there are more males than females in both data sets and more non-smokers than smokers. Also, in both data sets, more males reported smoking than females. The most frequently reported workout intensity for both males and females in the two data sets is moderate, followed by low and high.