Study of Missing Data and Imputation*

Zheng(Zeb) Yang

March 3, 2024

Table 1: The missing data of orginal dataset of bill length of penguins

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Gentoo	Biscoe	NA	NA	NA	NA	NA	2009

Table 2: The all missing data of bill length of penguins

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Dream	NA	18.5	193	3500	female	2009
Adelie	Dream	NA	17.9	193	4250	male	2009
Gentoo	Biscoe	NA	NA	NA	NA	NA	2009
Chinstrap	Dream	NA	17.3	198	3675	female	2009

Table 3: Comparing the imputed values of bill length for penguins and the overall mean bill length

species	Island	Input mean	Multiple imputation	Actual
Adelie	Torgersen	43.93805	40.7	NA
Adelie	Dream	43.93805	37.9	36.8
Adelie	Dream	43.93805	39.6	39.7
Gentoo	Biscoe	43.93805	50.6	NA
Chinstrap	Dream	43.93805	45.7	49.8

1 Methodology and Result

First of all, I summarized and filtered the dataset. As you can see in Table 1, I found that there were two missing value of bill length of penguins in the original dataset. Then, I set the seed and randomly selected three rows of the dataset of penguins and removed the bill length values to make the dataset has some missing at random value. Therefore, now, as you can see in Table 2, I have five missing data of bill length of penguins in total. Then, I imputed the mean of observations without missing data. To impute the mean, I constructed a second dataset with the observations with missing data removed. I then computed the mean of the bill length, and imputed that into the missing values in the original dataset. The other way to impute the missing data is using multiple imputation. Now, I have the imputed dataset and as you can see in the

*Code and data are available at: https://github.com/iloveyz12/Penguins.git, Acknowledge to the review of Bernice Bao

Table 3, the missing bill length of penguins have been imputed by imputed mean and multiple imputation. The column of actual value in Table 3 is the actual bill length of penguin, so if there is a value of actual value, then it means that this row is selected randomly to be removed.

The simulation was conducted using the statistical programming language R (R Core Team 2020). The data is from data source palmerpenguins(Horst, Hill, and Gorman 2020). To further enable the analysis I employed the use of the package of knitr(Xie 2023), mice(van Buuren and Groothuis-Oudshoorn 2011) and dplyr(Wickham et al. 2023).

2 Analysis

Compare the imputed value with the actual value, I have found that the value of multiple imputation is closer to the actual data than imputing mean in our case. However, none of thease approaches can be naively imposed. For instance, the one Chinstrap penguin's bill length in Dream Island should be 49.8mm. Imputing mean across all species and islands would result in an estimate of 43.94mm and multiple imputation results in an estimate of 45.7, which are both too low.

References

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://doi.org/10.5281/zenodo.3960218.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in r." *Journal of Statistical Software* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.