**Data cleaning for column: “diabetes”**

**Lingyu Zhou**

# Read in dataset

First imports the raw data from the target file under proper directory into a R data frame object named df, which allows for subsequent data manipulation and analysis.

```{R}

df<-read.csv("..\\Handout\\final\_cardiac\_data.csv")

```

# Describe df$diabetes

Secondly, conduct the initial exploratory data analysis by generating summary statistics and a histogram visualizing the distribution of the diabetes variable.

```{r}

summary(df$diabetes)

```

[OP]

Then, a frequency table with the useNA="ifany" option tabulates any missing values encoded as NaN. Table shows 72 null values are presented in this dataset, which constitutes around 3% of the data in “diabetes” column.

```{r}

hist(df$diabetes)

table(df$diabetes,useNA = "ifany")

```

```{r}

sum(is.na(df$diabetes)) /length(df$diabetes)

```

[OP]

Above provides an overview of the variable's values and missing data.

# Cleaning1-Categorize into 0, 1, and NaN

The diabetes variable undergoes careful coding to prepare it for binary logistic regression modeling. Values of 3, indicating borderline diabetes, are recoded to 1 given common knowledge that borderline is closer to a diabetic diagnosis than healthy. Values of 9 representing unknown diabetes status are then treated first as NaN’s that are about to be processed later. Finally, values of 2 are recoded to 0 to transform the coding into a strict binary indicator more amenable for regression analysis and intuitive interpretation.

```{r}

df$diabetes[df$diabetes == 3] <- 1

df$diabetes[df$diabetes == 9] <- NaN

df$diabetes[df$diabetes == 2] <- 0

df\_f<-na.omit(df)

```

After all of above, with the diabetes variable cleaned, rows with remaining NaN values are removed and original dataset are overwritten.

```{r}

df

```

[OP]

#Cleaning2-Fit prediction model for missing values

A binary logistic regression model is estimated with diabetes as the outcome and education and gender as predictors judging from the previous correlation plot (plot[]). This allows the predictions of missing data in column “diabetes”.

```{r}

diab\_pr<-glm(diabetes~ educ + gender, data=df\_f, family=binomial)

```

```{r}

summary(diab\_pr)

```

[OP]

As we can see, both variable are highly statistically significant from Wald P-value or from p-value from LR test.

# Cleaning2-Predict values based on other non-missing values

For rows where diabetes is NaN, the fitted logistic model is used to generate predicted probabilities of diabetes based on the known data from “gender” and “education” in the same row, where both values are free from missing in any row so that it’s safe to do the prediction.

```{r}

table(df$educ,useNA = "ifany")

```

[OP]

```{r}

table(df$gender,useNA = "ifany")

```

[OP]

Then based on a clinically informed threshold of 0.5, these probabilities are used to impute 0 or 1, filling in the missing values.

```{r}

rows\_with\_na <- is.na(df$diabetes)

predicted\_probabilities <- predict(diab\_pr, newdata = df[rows\_with\_na, ], type = "response")

df$diabetes[rows\_with\_na] <- ifelse(predicted\_probabilities > 0.5, 1, 0)

```

By leveraging the relationships between the observed data, this allows missing values to be imputed in a principled manner for subsequent analysis.

```{r}

head(df,10)

```

[OP]

Overall, the coding follows standard data cleaning methodology to prepare the diabetes variable, applies subject matter knowledge, builds an appropriate prediction model, and uses it to fill in missing values.