Handling Missing Data in Epidemiology Research:

Exploring Multiple Imputation as a Missing Data Strategy

Mackenzie Mueller

Advisors: Dr. Nicola Justice, Dr. Ksenija Simic-Muller

May 21, 2024

Abstract

Missing data in epidemiology and public health research is often poorly handled, poorly analyzed, or sometimes not even reported at all. This can significantly influence the results of these studies, often causing bias towards or against certain populations. In this paper, we will explore potential underlying factors for missing data, types of missing data, useful (and not-so-useful) strategies for addressing missing data, and analyze an example dataset from the National Health and Nutrition Examination Survey (NHANES) using R Statistical Software. For our main focus, we will explore Multiple Imputation as a missing data strategy—what it is, how it works, and how it can be applied to our NHANES dataset.

Contents

1 Introduction 3
- 1.1 Why Missing Data? 3
- 1.2 Water Insecurity 3

2 Background 3
- 2.1 Data Context 3
- 2.2 Data Cleaning 4
- 2.3 Codebook 4
- 2.4 Motivating Question 5
- 2.5 Exploring the Data 5

3 Missing Data 6
- 3.1 Types of Missing Data 6
- 3.2 Ad-hoc Strategies for Handling Missing Data 7
  - 3.2.1 Complete Cases 7
  - 3.2.2 Mean Imputation 8
  - 3.2.3 Regression Imputation 8
- 3.3 Research Question 9

4 Multiple Imputation 9

5 Linear Regression 10
- 5.1 Simple Linear Regression 10
- 5.2 Multiple Linear Regression 10

6 Multiple Imputation in Practice 11
- 6.1 Building a Model 11
- 6.2 Coefficients and Standard Errors 13
  - 6.2.1 Variance is Additive Proof 14
- 6.3 Final Model 15
  - 6.3.1 Interpretations of the coefficients 15
  - 6.3.2 Synthesis 16
  - 6.3.3 Confidence Intervals 16
  - 6.3.4 Synthesis 17

7 Visualizations 17

8 Toy Data Example of Multiple Imputation 18

9 Conclusion 20

10 Appendix 22

1 Introduction

1.1 Why Missing Data?

We refer to any missing cases, missing responses, “don’t know” responses, or refused responses as missing data. There is poor reporting, handling, and analysis of missing data, especially when it comes to public health and epidemiology research. The way a statistician chooses to address and handle missing data can affect the results of a model, a study, or statistical significance. The “quick-fix” solution of using only the complete cases in epidemiological research not only diminishes the size of the data by a sometimes unignorable amount, but also takes out a critical component of the narrative behind the data. “Why is this data missing? Why do certain groups have more missing data than others? How can we address this?” These are questions that every researcher should be asking before starting analysis on any dataset, but they are questions that often go overlooked or unanswered. It is important to learn about the ways that exist for handling missing data, how they work, and when to use them.

1.2 Water Insecurity

According to the National Library of Medicine (NLM), millions of people globally are water insecure. For the purpose of this paper, household water insecurity is defined as the inability to reliably access and benefit from safe and acceptable water [6]. This includes those who may avoid drinking tap due to its perceived risks, not necessarily actual risk. There is currently a growing number of US households who cannot afford water bills, or who rely on water with contaminants (such as lead). Per the NLM, water insecurity is an underappreciated risk to well-being in the United States.

Generally, water insecurity measurements are more common in low-income countries. But studies now show that better information about water intake in the United States is desperately needed. These studies can offer insight into the prevalence, causes, and consequences of water insecurity [6]. This includes potential water-nutrition insecurity linkages, since outside of the United States it has already been seen that greater water insecurity is associated with greater food insecurity, altered infant/young child feeding patterns, and greater psychological distress [6].

2 Background

2.1 Data Context

The National Health and Nutrition Examination Survey is a cross-sectional survey of the civilian, non-institutionalized population designed to assess the health and nutritional status of both adults and children in the US. It is a part of the National Center for Health Statistics, which is a part of the Centers for Disease Control & Prevention. NHANES began in the

1960s, and combines both interviews with patient physical examinations. Survey periods are two years, and have a changing focus on health and nutrition measurements to meet emerging needs [1].

NHANES examines around 5,000 people each year, and makes efforts to be nationally representative on race and ethnicity, socioeconomic status, and age within the US population by oversampling people aged 60 and over, African Americans, and Hispanic-identifying individuals. Participants complete an interview including questions about their demographics, socioeconomic status, dietary intake, and other health-related questions. The physical examination includes medical, dental, and physiological measurements, as well as laboratory testing. The participants are offered transportation to and from the mobile site where examinations are done, and later are compensated and given a report of their medical findings [1].

2.2 Data Cleaning

NHANES data exists in R through the package "nhanesA." Relevant variables were selected using the NHANES codebook: age, sex, race/ethnicity, patient education, reference education, federal income poverty ratio (fipr), and tap water.

There were two separate education variables for patient education in NHANES, one for patients under 20 years of age and one for patients over 20 years. These variables were combined to make one combined_education variable. The levels in the combined_education variable were then pooled to match the levels in the reference_education category. In the race/ethnicity category, the variables "Latino" and "Other Hispanic" were combined into a new variable new_race_ethnicity. While these are different ethnic identities, for the purpose of this project it made most sense to combine them into one category.

Indicator variables were created for each variable with factor levels (combined_education, reference_education, race_ethnicity, sex). The reference group for each variable is listed below (note R chooses reference groups alphabetically):

Variable	Reference Group
race_ethnicity	Hispanic/Latino
combined_education	College graduate or above
reference_education	Less than high school degree
sex	sex_female

Table 1

2.3 Codebook

NHANES contains questionnaires, datasets, and related documentation. The following are relevant variables:

Variable	Description
age	Patient age in years
sex	Sex of patient (M/F)
race_ethnicity	Patient race (Hispanic/Latino, Asian, Black, White, Multiracial/Other)
combined_education	Patient education level (Less than high school, Some college, College grad)
reference_education	Reference education level (Less than high school, Some college, College grad)
fipr	Federal income poverty ratio (0-5)
tap_water	Amount of tap water drank, day 1, grams

Table 2

Note that the variable fipr will be the most important predictor variable for this project, and that tap_water consumption on a given day is what we are ultimately trying to model.

Definition 1. fipr. A ratio of family income to poverty guidelines, where a 1 represents a family at the federal poverty line. (Range from 0-5) [1]

Note: A score under 1 means a family's income is lower than what the federal poverty line states (for a family of that size). A score of 2 represents a family whose income is double that of the federal poverty guidelines, etc. 5 is the maximum for this dataset, families whose income is more than 5 times greater than the federal poverty line were marked as a 5.

2.4 Motivating Question

Can `fipr` help predict `tap_water` consumption?

We are interested in exploring whether or not poverty is related to water security/insecurity in the US. Do lower/higher income areas have lower/higher access to potable drinking water? Does more research need to be done in the US on water insecurity? These are related questions that we should be asking ourselves throughout this paper.

2.5 Exploring the Data

The following are relevant data visualizations pertaining to some of the variable demographics we will be using for this project (see figures 1,2,3,4).

From figure 1 we can see that age somewhat skewed to the left for patients under age 20, but that there is a wider range of patients above age 20. Note that NHANES marks ages over 80 as 80.

Figure 2 illustrates that more patients identify as White, Hispanic/Latino, or Black when compared to Asian or Multiracial identifying individuals.

From figure 3 we can conclude that the majority of patients have an education level that is lower than a college degree.

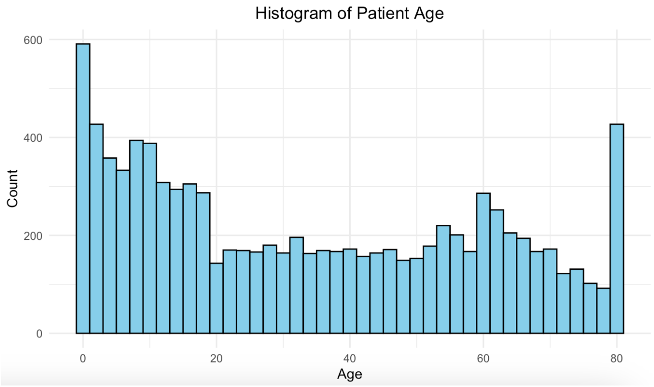


Figure 1

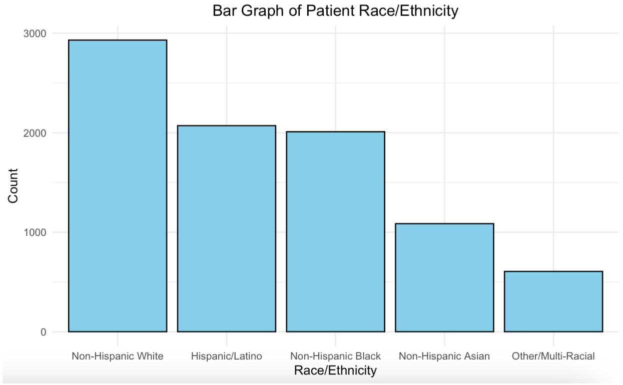


Figure 2

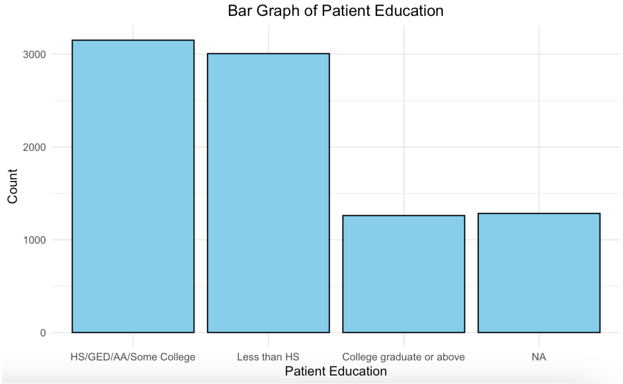


Figure 3

3 Missing Data

3.1 Types of Missing Data

There are three main types of missing data: MCAR, MAR, and MNAR.

Definition 2. MCAR. Data is said to be missing completely at random if the probability of being missing is the same for all cases [5]. For example, take a random sample of a population, where each person has the same likelihood of being selected. The (unobserved) people who were not selected are MCAR.

Definition 3. MAR. Data is said to be missing at random if the probability of being missing is the same only within groups defined by the *observed* data [5]. For example, if a random sample of a population depends on some known property, such as age or occupation, unobserved cases who do not fit that age or occupation are considered MAR.

Definition 4. MNAR. Data is said to be missing not at random if the probability of being missing varies for unknown reasons and depends on unobserved values [5]. For example, in a public opinion survey, those with weaker opinions may respond less frequently to certain questions. This is the trickiest type of missing data to work with.

3.2 Ad-hoc Strategies for Handling Missing Data

There are various ways to "handle" missing data in a quick way, each with disadvantages. These include the complete cases method, mean imputation, and regression imputation.

3.2.1 Complete Cases

The complete cases method omits all rows that contain a missing value in any category. This is also known as "listwise-deletion." This method leads to a reduced sample size, a loss of information from the dataset, and can cause biased results and invalid inferences [2]. Biased results can occur because it is often underrepresented and/or marginalized populations and/or participants at risk who have missing data. The complete cases method only works on data that is MCAR [5].

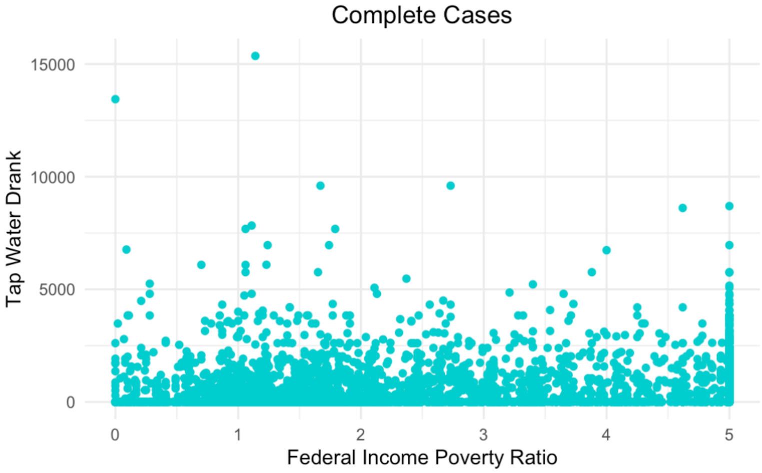


Figure 4

3.2.2 Mean Imputation

Mean imputation replaces missing data values with the mean of that variable. In this case, missing fipr values are imputed with the mean of fipr [5]. This weakens the relationship between variables, and leads to biased estimates and invalid inferences, even for data that is MCAR [2]. See figure 6, and note that the known cases (the "not imputed" cases) are the same as in 3.2.1. However, now there are imputed values (shown in red) where the complete cases method simply removed those cases.

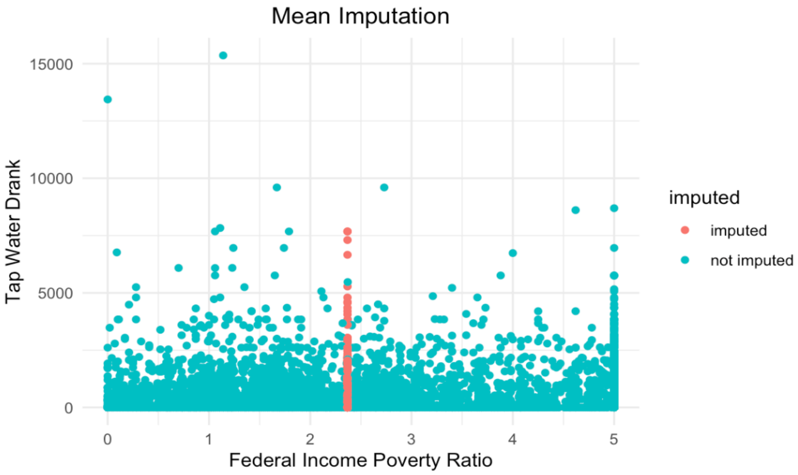


Figure 5

3.2.3 Regression Imputation

Regression imputation builds a model from the observed data first, then calculates predictions for missing values under the predicted model. The missing values are assigned the "most likely" values on the regression line [5]. In this example, a model to predict missing fipr values is built on the known cases of age and fipr. The missing values are then assigned the "most likely" values for fipr based on that person's age. The variable fipr can now be considered a complete variable, and could be used to plot patient fipr against tap water. Notice the imputed values, shown in red, are between 2 and 3 on the x-axis. This is because the model only predicted fipr values between 2 and 3 when using age as its predictor factor.

Regression imputation is not able to account for the variability of the data, and does not take any randomness into account. The model predicted is only a best fit model, which often causes an upward bias of the predicted cases. This method, like the others, leads to biased estimates and invalid inferences, even when the data is MCAR [2]. See figure 7, where the on the left, the imputed values shown in red have an upward bias and do not account for the robust variability of the known cases in blue. This translates to the plot on the right, where we can see that the imputed values do not accurately represent the variability of the data.

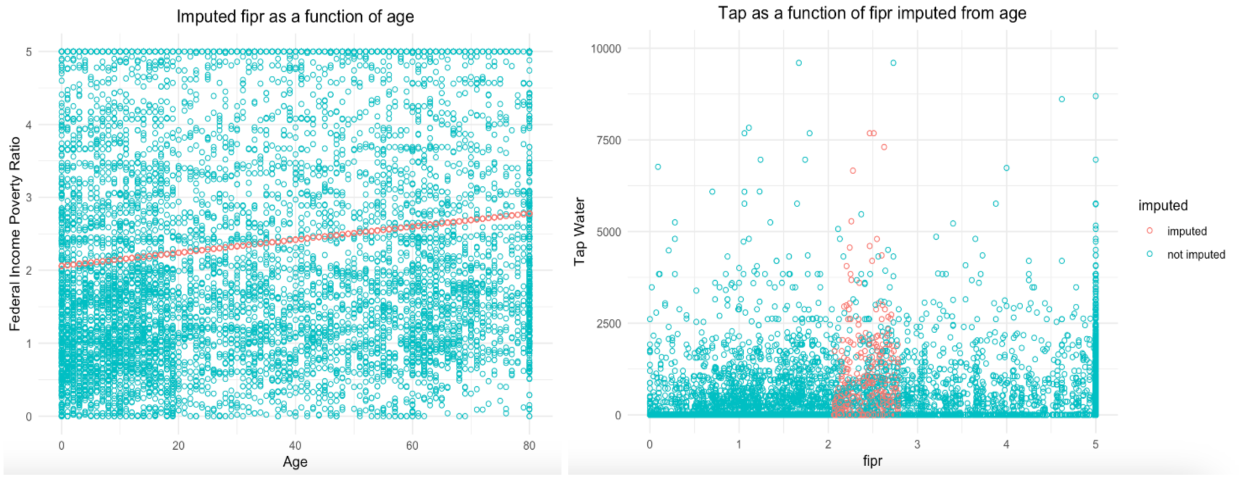


Figure 6

3.3 Research Question

The previous strategies for handling missing data all lead to biased estimates and invalid inferences, so what can be done to more effectively handle missing data? This paper will now explore the question: **how can multiple imputation as a missing data strategy be used to handle missing data?**

4 Multiple Imputation

Multiple imputation (MI) is a missing data strategy that provides a way of dealing with the uncertainty of the missing values themselves [5]. MI is able to do this because rather than imputing only one value for each missing value, it imputes several values that later get pooled back together. Generally 5-10 imputations take place [4]. For this project, we will impute missing values 5 times.

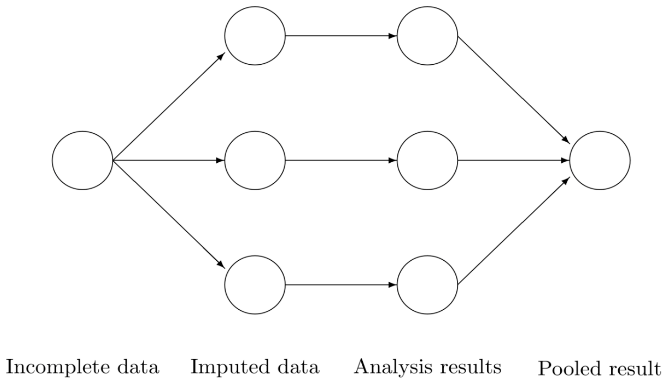


Figure 7: *Flexible Imputation of Missing Data*, van Buuren, 2018.

The flowchart above represents the general steps of MI. We begin with an incomplete dataset, impute the missing values g times, in this case g is 3, perform statistical analysis on the imputed data, and then average the imputed values to end with one pooled result. These steps will be explained in more detail in the following sections, but first we pause to review basic bivariate linear regression.

5 Linear Regression

The following definitions and explanations are from [3].

To understand the process of the model we will be creating and the steps of MI, we first begin by reviewing regression. Regression is used to estimate the association between an outcome value (Y) and one or more predictor variables ($X_1, X_2, X_3, ...$). There are two types of linear regression: simple linear regression and multiple linear regression. For our model we will focus on multiple linear regression, but an understanding of simple linear regression is needed first.

5.1 Simple Linear Regression

Definition 5. Simple Linear Regression (SLR) is used when there is only one continuous predictor value (X). The equation is as follows:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where

- Y is the outcome.
- β_0 is the intercept.
- β_1 is the slope.
- ϵ is the error term that represents the part of Y that cannot be explained by the predictor variable.

In simple linear regression, we have one predictor variable that we believe is related to the outcome variable. Commonly, the predictor variable is continuous. A **continuous** variable is a type of quantitative variable that can take any value within a certain measurable range. Examples include: height, weight, age, distance, and temperature.

5.2 Multiple Linear Regression

Definition 6. Multiple Linear Regression is used when there is more than one predictor value ($X_1, X_2, X_3, ...$) in the model, and the equation is as shown:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$$

where

- Y is the outcome variable.
- B_0 is the constant term.
- $X_1, X_2, ..., X_k$ are the predictor variables.
- $\beta_1, \beta_2, ..., \beta_k$ are the weights of the predictor variables.
- ϵ is the random error term that represents the part of Y that cannot be explained by the predictor variables.

6 Multiple Imputation in Practice

For the purpose of this project, we are predicting tap water consumption based on age, ref_some_college, ref_college_grad, com_some_college, com_less_hs, race_asian, race_black, race_white, race_other, sex_female, and fipr, but fipr has missing values. Note there are 11 variables due to the creation of indicator variables (see section 2.2). We will illustrate MI in handling the fipr missing values.

6.1 Building a Model

We begin by building a multiple regression model in R to predict a model for fipr based on the complete cases of the following variables:

```
fipr_model <- lm(fipr ~ reference_education + combined_education +
                    new_race_ethnicity + age + sex)
```

We will use this model to impute our missing fipr values 5 times, each time adding different random noise ϵ to the regression output. ϵ is drawn from the standard deviation of the residuals in `fipr_model` that assumes a normal distribution[2]:

$$\epsilon \sim N\left(0, \sigma^2\right)$$

based on $n - m$ complete cases:

$$\sigma^2 = \frac{\sum_{i=m+1}^{n} \left(Y - \hat{Y}\right)^2}{n - m}.$$

This is the variance we see in the complete cases across our dataset, so we are using the same variance to model the variance in the imputed cases. A normal distribution is shown in Fig 9 as a reminder that the normal distribution is bell-shaped and symmetric:

Using `fipr_model` 5 times with different random noise added to each regression output produces 5 complete datasets that all have the original fipr values if they are not missing,

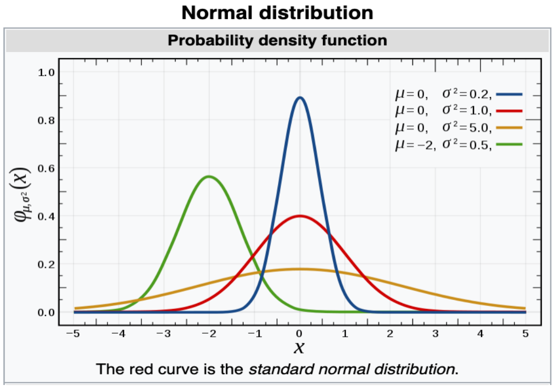


Figure 8

but have different imputed fipr values if fipr was originally missing. The `ifelse()` function in R is used to assign missing values an imputed value, and to keep existing values the same. The process is as follows:

- If `fipr` is missing, replace NA value with

  $$\beta_0 + \beta_1 x_1 + ... + \beta_{11} x_{11} + \epsilon_i$$

  where

  - $\beta_1, \beta_2, ..., \beta_{11}$ are the regression coefficients
  - $x_1, x_2, ...x_{11}$ are the indicator variables
  - $\epsilon_i \sim N(0, \sigma^2)$ is the random noise

- Else, use given `fipr` value.

To keep `fipr` values within the 0-5 range (to match the range of the complete cases), the `pmin()` and `pmax()` functions were used. Negative imputed values were recoded as 0 and values greater than 5 were recoded to a 5. As aforementioned, there are 11 regression coefficients due to the creation of indicator variables (see 2.2).

This creates 5 different complete datasets across education, race/ethnicity, sex, age, and now also **fipr**. These predictor variables can be used to model tap water consumption. 5 tap water models are now created from the complete variables and from the newly imputed fipr variables (fipr1 through fipr5):

```
tap1 <- lm(reference_education + combined_education + new_race_ethnicity +
                                age + sex + fipr1)

                                      ⋮

tap5 <- lm(reference_education + combined_education + new_race_ethnicity +
                                age + sex + fipr5)
```

6.2 Coefficients and Standard Errors

The following equations are from [2]. From the tap water models we acquire coefficient estimates and their standard errors (SE):

- $\hat{\beta}_0^\ell, \hat{\beta}_1^\ell, \hat{\beta}_2^\ell, ..., \hat{\beta}_k^\ell$ for $\ell = 1...g$
- $SE(\hat{\beta}_0^\ell), SE(\hat{\beta}_1^\ell), ..., SE(\hat{\beta}_k^\ell)$ for $\ell = 1...g$

where

- k represents each of our 11 predictor variables, and
- ℓ represents each of our 5 imputed datasets.

To obtain the coefficient estimates of our final model, we average each of the regression coefficients across our five imputed tap water models as follows:

$$\tilde{\beta}_j \equiv \frac{\sum_{\ell=1}^{g} \hat{\beta}_j^\ell}{g}$$

where

- $\tilde{\beta}_j$ represents each pooled coefficient estimate.
- g represents the number of imputed datasets, in our case, $g = 5$.

To obtain the standard errors of the estimated coefficients, we must take into consideration both the *within* imputation and *between* imputation variation.

The within imputation component represents the standard error that nearly any model is going to have, since very rarely is a model expected to perfectly predict the data. Here we average the standard errors:

$$V_j^{(W)} \equiv \frac{\sum_{\ell=1}^{g} SE^2\left(\beta_j^\ell\right)}{g}$$

The between imputation component represents the variation in the regression coefficients between each imputed dataset's estimate, and must also be taken into consideration:

$$V_j^{(B)} \equiv \frac{\sum_{\ell=1}^{g} \left(\beta_j^\ell - \tilde{\beta}_j\right)^2}{g-1}$$

The within and between imputation components can be added together as shown below, multiplying the between imputation component by $\frac{g+1}{g}$ to prevent a biased model, and then the square root is taken to obtain the standard error of each regression coefficient. This works because of the additive property of variance (see proof).

$$\tilde{SE}\left(\tilde{\beta}_j\right) \equiv \sqrt{V_j^{(W)} + \frac{g+1}{g} V_j^{(B)}}$$

6.2.1 Variance is Additive Proof

Theorem 1. *For two independent random variables, the variance is additive. [2]*

Proof. We will show that for two independent random variables, the variance is additive:

$$V(X+Y) = V(X) + V(Y)$$

Per a convenient property of variance, let:

$$V(X) = E(X^2) - [E(X)]^2.$$

Then

$$V(X) + V(Y) = E(X)^2 - [E(X)]^2 + E(Y)^2 - [E(Y)]^2$$

And

$$V(X+Y) = E(X+Y)^2 - [E(X+Y)]^2.$$

Using algebra and the fact that expectation is a linear operator:

$$V(X+Y) = E(X^2 + 2XY + Y^2) - E(X+Y)E(X+Y)$$

$$V(X+Y) = E(X^2) + 2E(XY) + E(Y^2) - [E(X) + E(Y)][E(X) + E(Y)]$$

$$V(X+Y) = E(X^2) + 2E(XY) + E(Y^2) - [[E(X)]^2 + 2E(X)E(Y) + [E(Y)]^2]$$

Rearranging, we see that:

$$V(X+Y) = E(X^2) - [E(X)]^2 + E(Y^2) - [E(Y)]^2 + 2E(XY) - 2E(X)E(Y)$$

Which simplifies to:

$$V(X+Y) = E(X^2) - [E(X)]^2 + E(Y^2) - [E(Y)]^2.$$

Per the same convenient property of variance as above,

$$V(X) = E(X^2) - [E(X)]^2$$

$$V(Y) = E(Y^2) - [E(Y)]^2$$

and so we have shown that:

$$V(X+Y) = V(X) + V(Y).$$

□

6.3 Final Model

After obtaining the regression coefficients and their respective standard errors, we are left with a final model:

$$\hat{tap} = \texttt{570.1 + (-86.3)ref\_some\_college + (-62.9)ref\_college\_grad +} \\ \texttt{(-164.9)com\_some\_college + (-361.9)com\_less\_hs + (184.1)race\_asian + (17.2)race\_black} \\ \texttt{+ (286.9)race\_white + (254.0)race\_other + (-0.8)age + (-12.9)sex\_female + (38.8)fipr}$$

6.3.1 Interpretations of the coefficients

Our final model predicts that when all other variables are held constant...

- A person whose reference person has some college education will drink 86.3 grams less tap water than a person from the reference group for this category (less than high school degree).
- A person whose reference person has a college degree or higher will drink 62.9 grams less tap water compared to a person from the reference group of this category.
- A person with some college education will drink 164.9 grams less tap water compared to a person from the reference group of this category (college degree or higher).
- A person with less than a high school degree will drink 361.9 grams less tap water compared to a person from the reference group of this category.
- A person who identifies as Asian will drink 184.1 grams more tap water compared to a person from the reference group of this category (Hispanic/Latino).
- A person who identifies as Black will drink 17.2 grams more tap water compared to a person from the reference group of this category.
- A person who identifies as White will drink 286.9 grams more tap water compared to a person from the reference group of this category.
- A person who identifies as “Other” or Multiracial will drink 254 grams more tap water compared to a person from the reference group of this category.
- For each one unit increase in age there will be a 0.8 gram decrease in tap water consumption.
- A person who identifies as female will drink 12.9 grams tap water less than a person who identifies as male (the reference group for this category).
- For each one unit increase in fipr there will be a 38.8 gram increase in tap water consumption.

6.3.2 Synthesis

Our interpretations suggest that on average, more educated people drink more tap water (this was seen in patient education). Our interpretations also suggest that on average, people who identify as Asian, Black, White, or Multiracial are expected to drink more tap water than Hispanic/Latino identifying individuals. Lastly, our interpretations suggest that on average, families with higher incomes (higher fipr) drink more tap water.

6.3.3 Confidence Intervals

Note that there is uncertainty around each of our regression coefficients. The standard errors and confidence intervals for each regression coefficient were calculated by hand. To calculate a confidence interval, the degrees of freedom must first be found as follows [2]:

$$df_j = (g-1)\left(1+\frac{g}{g+1} * \frac{V_j^{(W)}}{V_j^{(B)}}\right)^2$$

A 95% confidence interval is then calculated [2]:

$$\beta_j = \tilde{\beta_j} \pm t_{.025,df_j}\tilde{SE}\left(\tilde{\beta_j}\right)$$

Below are the standard errors and confidence intervals for each regression coefficient:

- The standard error for `ref_some_college` is 37.09, with a 95% confidence interval between -159.02 and -13.65.
- The standard error for `ref_college_grad` is 493.52, with a 95% confidence interval between -1030 and 904.
- The standard error for `com_some_college` is 44.6, with a 95% confidence interval between -252.29 and -77.46.
- The standard error for `com_less_hs` is 47.56, with a 95% confidence interval between -455.16 and -268.72.
- The standard error for `race_asian` is 44.64, with a 95% confidence interval between 96.64 and 271.63.
- The standard error for `race_black` is 35.59, with a 95% confidence interval between -52.51 and 87.
- The standard error for `race_white` is 33.26, with a 95% confidence interval between 221.75 and 352.13.

- The standard error for `race_other` is 53.34, with a 95% confidence interval between 149.44 and 358.53.

- The standard error for `age` is 0.65, with a 95% confidence interval between -2.074 and 0.474.

- The standard error for `sex_female` is 23.44, with a 95% confidence interval between -58.78 and 33.08.

- The standard error for `fipr` is 8.27, with a 95% confidence interval between 22.4 and 54.99.

6.3.4 Synthesis

Because 0 was not included in their confidence intervals, we can conclude that ref_some_college, com_some_college, com_less_hs, race_asian, race_white, race_other, and fipr are significant predictors of tap water consumption. Because 0 was included in the confidence interval for ref_college_grad, race_black, age, and sex, we do not have enough evidence to conclude whether these are significant predictors of tap water consumption.

7 Visualizations

Figures 10 and 11 show two plots that help visualize our model prediction for patient race/ethnicity and patient education, respectively. Note that the median age is fixed on both cross-sectional graphs.

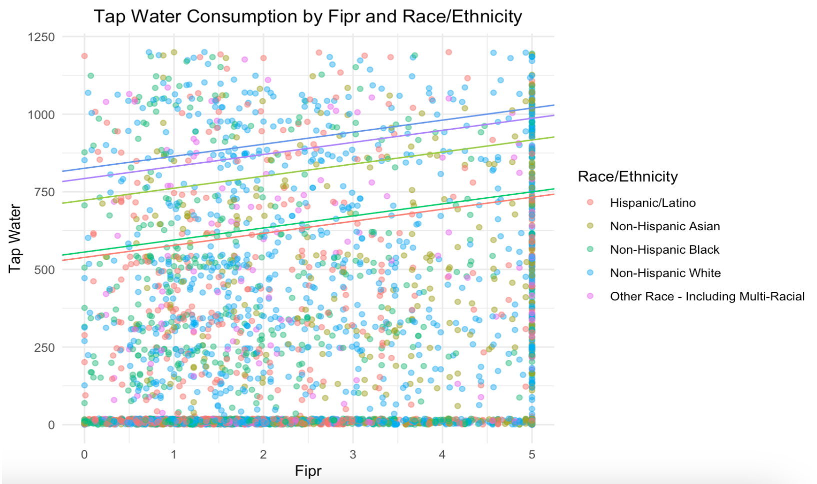


Figure 9

Note that the slope of each of the lines in Figure 9 is 38.8, which comes directly from our tap water model in 6.3. When all other variables are held constant, patients who self identify as White, Multiracial/Other, Asian, and Black will drink 286.9g, 254g, 184.1g, and 17.2g more tap water respectively than the reference group for this category: Hispanic/Latino.

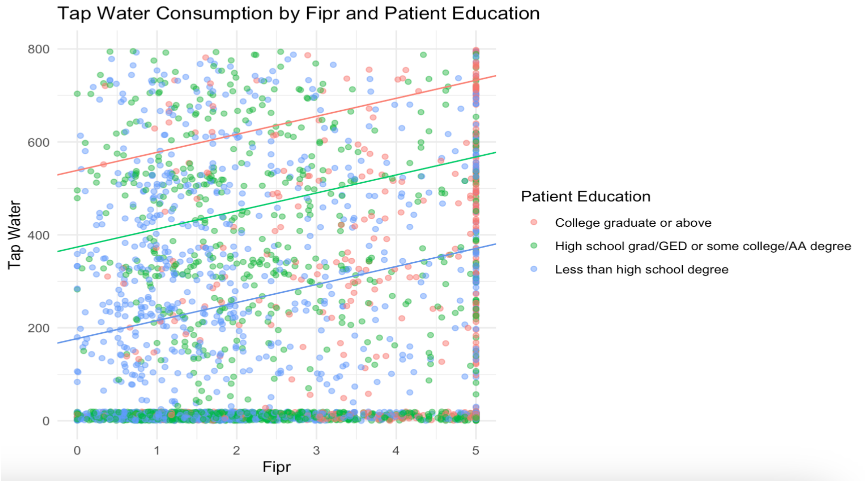
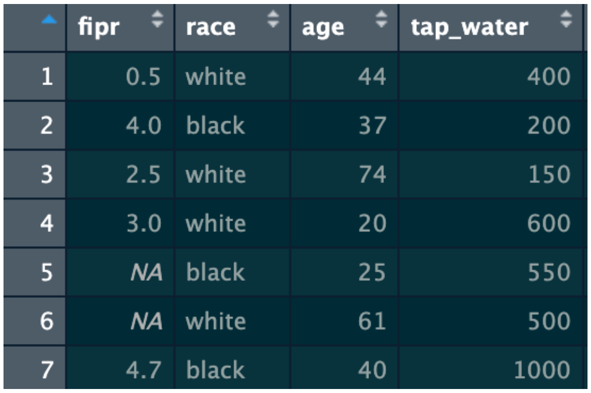


Figure 10

Again, the slope of these lines is 38.8. Figure 10 shows that when all other variables are held constant, patients who have some college education and patients who have less than a high school education will drink 164.9g and 361.9g less tap water when compared to the reference group for this category: college graduate or above.

8 Toy Data Example of Multiple Imputation

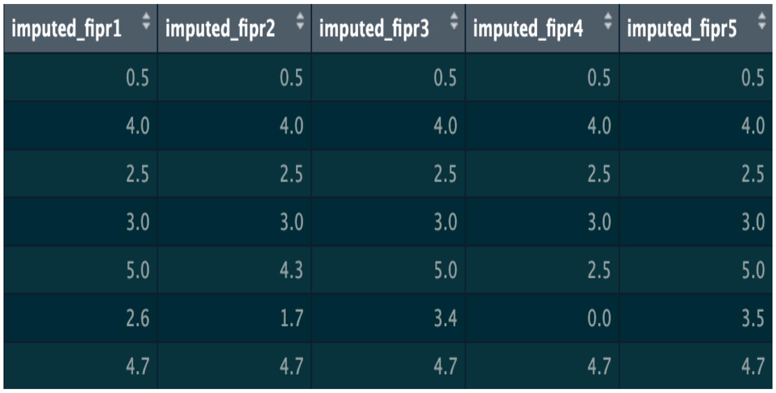
To demonstrate how multiple imputation works with our model on a smaller scale, we will look at a toy data example. Say we have 7 people, with two of them having missing fipr values.



	fipr	race	age	tap_water
1	0.5	white	44	400
2	4.0	black	37	200
3	2.5	white	74	150
4	3.0	white	20	600
5	*NA*	black	25	550
6	*NA*	white	61	500
7	4.7	black	40	1000

Figure 11

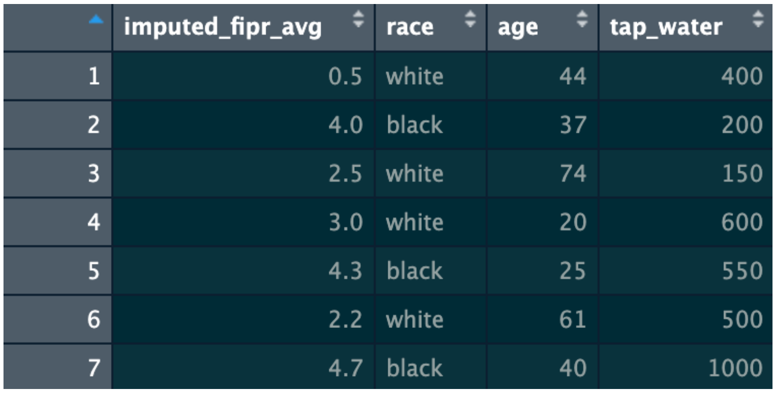
Our goal is to fill persons 5 and 6 missing fipr cases with plausible fipr values using multiple imputation. We begin with a model to predict fipr based on the complete cases of the other predictor variables (in this case we just have race and age). We impute 5 times using this regression model, each time adding random noise to the regression output. We now have 5 imputed datasets, where the complete cases stay the same and only the missing values are imputed.



imputed_fipr1	imputed_fipr2	imputed_fipr3	imputed_fipr4	imputed_fipr5
0.5	0.5	0.5	0.5	0.5
4.0	4.0	4.0	4.0	4.0
2.5	2.5	2.5	2.5	2.5
3.0	3.0	3.0	3.0	3.0
5.0	4.3	5.0	2.5	5.0
2.6	1.7	3.4	0.0	3.5
4.7	4.7	4.7	4.7	4.7

Figure 12

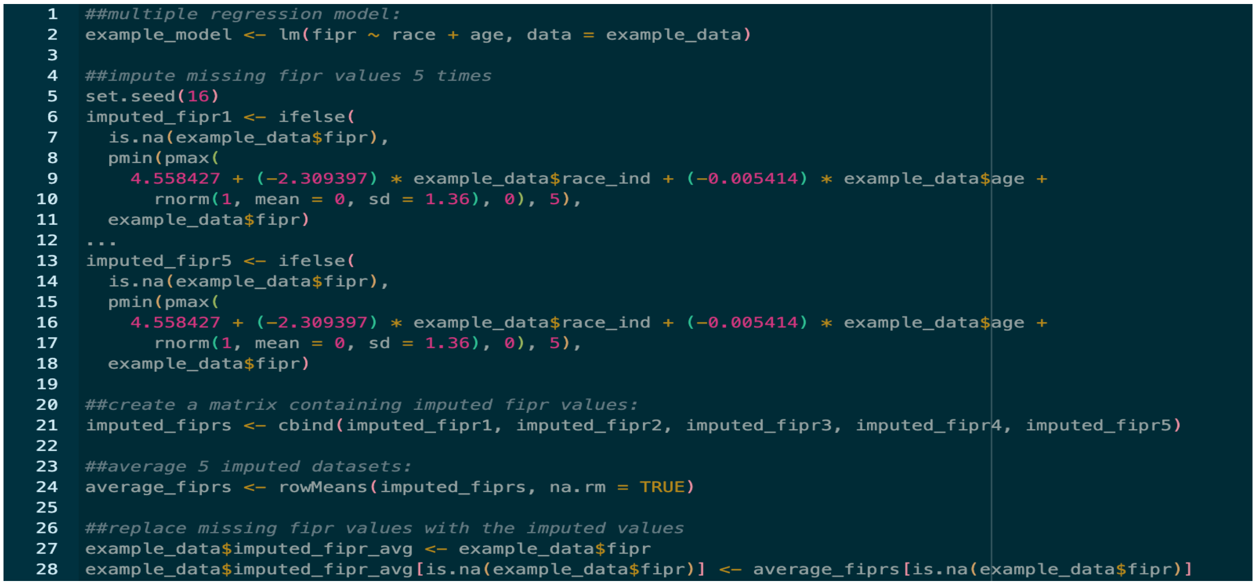
We can now average the imputed fipr values for persons 5 and 6, and will end a pooled result in a now "complete" dataset.



	imputed_fipr_avg	race	age	tap_water
1	0.5	white	44	400
2	4.0	black	37	200
3	2.5	white	74	150
4	3.0	white	20	600
5	4.3	black	25	550
6	2.2	white	61	500
7	4.7	black	40	1000

Figure 13

See Figure 14 for the code used for the toy data example.



```r
##multiple regression model:
example_model <- lm(fipr ~ race + age, data = example_data)

##impute missing fipr values 5 times
set.seed(16)
imputed_fipr1 <- ifelse(
  is.na(example_data$fipr),
  pmin(pmax(
    4.558427 + (-2.309397) * example_data$race_ind + (-0.005414) * example_data$age +
      rnorm(1, mean = 0, sd = 1.36), 0), 5),
  example_data$fipr)
...
imputed_fipr5 <- ifelse(
  is.na(example_data$fipr),
  pmin(pmax(
    4.558427 + (-2.309397) * example_data$race_ind + (-0.005414) * example_data$age +
      rnorm(1, mean = 0, sd = 1.36), 0), 5),
  example_data$fipr)

##create a matrix containing imputed fipr values:
imputed_fiprs <- cbind(imputed_fipr1, imputed_fipr2, imputed_fipr3, imputed_fipr4, imputed_fipr5)

##average 5 imputed datasets:
average_fiprs <- rowMeans(imputed_fiprs, na.rm = TRUE)

##replace missing fipr values with the imputed values
example_data$imputed_fipr_avg <- example_data$fipr
example_data$imputed_fipr_avg[is.na(example_data$fipr)] <- average_fiprs[is.na(example_data$fipr)]
```

Figure 14

9 Conclusion

Multiple imputation (MI) is a more effective way of imputing missing data compared to Ad-hoc strategies such as Complete Cases, Mean Imputation, and Regression Imputation. MI is able to deal with the uncertainty of the missing values themselves by imputing several values for each missing case, rather than just one. These imputed values are drawn from a plausible assumed normal distribution based on the model's standard deviation, with random noise added to account for the variability of the missing data.

In response to our motivating question, we can conclude that fipr is a statistically significant predictor of tap water consumption because 0 was not included in the confidence interval. Additionally, we can conclude that ref_some_college, com_some_college, com_less_hs, race_asian, race_white, and race_other are also significant predictors of tap water consumption. Because 0 was included in the confidence interval for ref_college_grad, race_black, age, and sex, we do not have enough evidence to conclude whether these are significant predictors of tap water consumption.

Water Insecurity is very much an issue in the US, though many people might think it is not. Theoretically, all areas of the US have adequate drinking water. In reality, this is not the case. Low-income and marginalized neighborhoods do not have the same quality of tap water that higher income neighborhoods do, and are more likely to experience water insecurity due to numerous factors such as old pipes corroding, water sources being switched with no notice, and in extreme cases such as in Flint, Michigan, lead poisoning. While research on water insecurity in the US has a long way to go, NHANES provides a start to being able to dive deeper into the demographics and socioeconomic status of those who tend to avoid tap water more than others.

References

[1] National Center for Health Statistics. About the national health and nutrition examination survey.

[2] John Fox. Applied Regression Analysis and Generalized Linear Models. SAGE Publications, 2008.

[3] Ramzi W. Nahhas. Introduction to Regression Methods for Public Health Using R. bookdown, 2024.

[4] Mailman School of Public Health. Missing data and multiple imputation. 2024.

[5] Stef van Buuren. Flexible Imputation of Missing Data, Second Edition. CRC Press, 2018.

[6] Sera L Young and Joshua D Miller. Water insecurity in the united states: Quantifying an invisible crisis. The Journal of nutrition, 152(2), 2022.

10 Appendix

```
install.packages(”nhanesA”)
library(nhanesA)
install.packages(ggplot2)
library(ggplot2)
library(dplyr)

##load dietary data
nhanesTables(data_group=’DIET’, year=2018)
diet<−nhanes(’DR1TOT_J’)
diet

##take out relevant variables
water<− diet[c(”DR1_320Z”,”DR1_330Z”,”DR1BWATZ”,”DR1TWSZ”)]
names(water)<− c(”plain_water”, ”tap_water”,
”bottled_water”,”tap_water_source”)

##load demographics 2018
nhanesTables(data_group=’DEMO’, year=2018)
demo2<−nhanes(’DEMO_J’)

##take out relevant variables
demo<−demo2[c(”RIDAGEYR”,”RIAGENDR”,”RIDRETH3”,”DMDHREDZ”,
”DMDEDUC2”,”DMDEDUC3”,”DMDHHSIZ”,”INDFMPIR”)]
names(demo)<− c(”age”,”sex”,”race_ethnicity”,”reference_education”,
”adult_education”,”kid_education”,”ppl_in_household”,”fipr”)

##exploring demographics of patients:
table(demo$race_ethnicity)
hist(demo$age)

ggplot(data=demo, aes(x=age))+
  geom_histogram(binwidth=2, fill=”skyblue”, color=”black”)+
  labs(x=”Age”, y=”Count”, title=”Histogram of Patient Age”)+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))


ggplot(data=demo, aes(x=sex))+
  geom_bar(fill=”skyblue”,color=”black”)+
  labs(x=”Sex”, y=”Count”, title=”Bar Graph of Patient Sex”)+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

ggplot(data=demo, aes(x=race_ethnicity))+
  geom_bar(fill=”skyblue”,color=”black”)+
  labs(x=”Race/Ethnicity”, y=”Count”,
  title=”Bar Graph of Patient Race/Ethnicity”)+
  scale_x_discrete(labels = c(”Mexican American”, ”Other Hispanic”,
  ”Non–Hispanic White”, ”Non–Hispanic Black”, ”Non–Hispanic Asian”,
  ”Other/Multi–Racial”))+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

ggplot(data=water, aes(x=tap_water_source))+
  geom_bar(fill=”skyblue”,color=”black”)+
  labs(x=”Tap Water Source”, y=”Count”, title=”Patient Tap Water Source”)+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

##combine diet & demo datasets by identifier
water$identifier<–diet$SEQN
demo$identifier<–demo2$SEQN

combined<–merge(demo, water, by=”identifier”)
combined

##combining patient education categories:
combined$adult_education <– as.character(combined$adult_education)
combined$new_patient_education <– combined$adult_education
combined$new_patient_education[combined$adult_education
%in% c(”Less than
9th grade”, ”9–11th grade (Includes 12th grade with no diploma)”)] <–
”Less than high school degree”
combined$new_patient_education[combined$adult_education
%in% c(”High school
graduate/GED or equivalent”, ”Some college or AA degree”)] <– ”High school
grad/GED or some college/AA degree”
combined$new_patient_education[combined$adult_education
%in% c(”College graduate or above”)] <– ”College
graduate or above”
combined$new_patient_education[combined$adult_education
%in% c(”Refused”,
"Don't Know")] <- NA

table(combined$new_patient_education)
combined$new_patient_education <- as.factor(combined$new_patient_education)


# Recode kid_education variable
combined$new_kid_education <- recode_factor(combined$kid_education,
 "Never attended / kindergarten only" = "Less than high school degree",
 "1st grade" = "Less than high school degree",
 "2nd grade" = "Less than high school degree",
 "3rd grade" = "Less than high school degree",
 "4th grade" = "Less than high school degree",
  "5th grade" = "Less than high school degree",
"6th grade" = "Less than high school degree",
"7th grade" = "Less than high school degree",
"8th grade" = "Less than high school degree",
"9th grade" = "Less than high school degree",
"Less than 9th grade" = "Less than high school degree",
"10th grade" = "Less than high school degree",
"11th grade" = "Less than high school degree",
"12th grade, no diploma" = "Less than high school degree",
"High school graduate" = "High school grad/GED or some college/AA degree",
"GED or equivalent" = "High school grad/GED or some college/AA degree",
"More than high school" = "High school grad/GED or some college/AA degree",
"Refused" = NA_character_, # Set to NA if needed
"Don't Know" = NA_character_ # Set to NA if needed
)

table(combined$new_kid_education)
table(combined$new_patient_education)
table(combined$adult_education)
combined$adult_education[combined$adult_education %in% c("Refused",
"Don't Know")] <- NA

### combine new_kid_education with new_patient_education
combined$new_kid_education <- as.character(combined$new_kid_education)
combined$new_patient_education <- as.character(combined$new_patient_education)
combined$combined_education <- ifelse(is.na(combined$new_patient_education),
combined$new_kid_education, combined$new_patient_education)
table(combined$combined_education, useNA = "always")
combined$combined_education <- as.factor(combined$combined_education)
```

```
# Assign combined education to new_patient_education column
combined$new_patient_education <- combined$combined_education

#visualizing variables:
##bar graph of patient race/ethnicity (figure 2)
ggplot(data=combined, aes(x=reorder(new_race_ethnicity,
-table(new_race_ethnicity)[new_race_ethnicity])))+
  geom_bar(fill=”skyblue”,color=”black”)+
  labs(x=”Race/Ethnicity”, y=”Count”, title=”Bar Graph of Patient
  Race/Ethnicity”)+
  scale_x_discrete(labels = c(”Non-Hispanic White”,”Hispanic/Latino”,
  ”Non-Hispanic Black”, ”Non-Hispanic Asian”, ”Other/Multi-Racial”))+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

##bar graph of patient education (figure 3)
ggplot(data=combined, aes(x=reorder(combined_education,
-table(combined_education)[combined_education])))+
  geom_bar(fill=”skyblue”,color=”black”)+
  labs(x=”Patient Education”, y=”Count”,
  title=”Bar Graph of Patient Education”)+
  scale_x_discrete(labels = c(”HS/GED/AA/Some College”, ”Less than HS”,
  ”College graduate or above”, ”NA”))+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

##bar graph of reference education
ggplot(data=combined, aes(x=reorder(reference_education,
-table(reference_education)[reference_education])))+
  geom_bar(fill=”skyblue”,color=”black”)+
  labs(x=”Reference Education”, y=”Count”, title=
  ”Bar Graph of Reference Education”)+
  scale_x_discrete(labels = c(”HS/GED/AA/Some college”, ”College degree or
  above”, ”Less than HS”, ”NA”))+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))


####Pooling variables together
combined$race_ethnicity <- as.character(combined$race_ethnicity)
combined$race_ethnicity <- as.character(combined$race_ethnicity)
```

```
combined$new_race_ethnicity <- ifelse(combined$race_ethnicity %in%
c("Mexican American", "Other Hispanic"),
"Hispanic/Latino", combined$race_ethnicity)
table(combined$new_race_ethnicity)


##mean imputation model done by hand (figure 5)
combined$fipr_mean<-ifelse(is.na(combined$fipr),mean(combined$fipr,
na.rm = TRUE), combined$fipr)
combined$imputed<- ifelse(is.na(combined$fipr), "imputed", "not imputed")

ggplot(data=combined, aes(x = fipr_mean, y = tap_water, color = imputed))+
  geom_point()+
  labs(x="Federal Income Poverty Ratio", y="Tap Water Drank", title=
  "Mean Imputation")+
  #geom_smooth(method = "lm", color = "black") + # Add regression line
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

##complete cases model (figure 4):
ggplot(cleaned_data, aes(x = fipr, y = tap_water))+
  geom_point(color = "cyan3")+
  labs(x="Federal Income Poverty Ratio", y="Tap Water Drank", title=
  "Complete Cases")+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

###creation of indicator variables:

subset_data_complete$ref_some_college <- ifelse(subset_data_complete$
reference_education == "High school grad/GED or some college/AA degree", 1, 0)
subset_data_complete$ref_college_grad <- ifelse(subset_data_complete$
reference_education == "College graduate or above", 1, 0)
subset_data_complete$com_some_college <- ifelse(subset_data_complete$
combined_education == "High school grad/GED or some college/AA degree", 1, 0)
subset_data_complete$com_less_hs <- ifelse(subset_data_complete$
combined_education == "Less than high school degree", 1, 0)
subset_data_complete$race_asian <- ifelse(subset_data_complete$
new_race_ethnicity == "Non-Hispanic Asian", 1, 0)
subset_data_complete$race_black <- ifelse(subset_data_complete$
new_race_ethnicity == "Non-Hispanic Black", 1, 0)
subset_data_complete$race_white <- ifelse(subset_data_complete$
new_race_ethnicity == "Non-Hispanic White", 1, 0)
subset_data_complete$race_other <- ifelse(subset_data_complete$
new_race_ethnicity == "Other Race - Including Multi-Racial", 1, 0)
subset_data_complete$sex_female <- ifelse(subset_data_complete$
sex == "Female", 1, 0)


##fipr imputed using complete cases and different random noise:
set.seed(18)
fipr1 <- ifelse(
  is.na(subset_data_complete$fipr),
  pmin(pmax(
    1.937103821 + (0.712372698) * subset_data_complete$ref_some_college +
      (1.692130115) * subset_data_complete$ref_college_grad +
      (-0.524241462) * subset_data_complete$com_some_college +
      (-0.666407519) * subset_data_complete$com_less_hs +
      (0.357892051) * subset_data_complete$race_asian +
      (-0.233977902) * subset_data_complete$race_black +
      (0.105956535) * subset_data_complete$race_white +
      (0.017010060) * subset_data_complete$race_other +
      (-0.102803632) * subset_data_complete$sex_female +
      rnorm(dim(subset_data_complete)[1], mean = 0, sd = 1.388), 0), 5),
  subset_data_complete$fipr)

fipr2 <- ifelse(
  is.na(subset_data_complete$fipr),
  pmin(pmax(
    1.937103821 + (0.712372698) * subset_data_complete$ref_some_college +
      (1.692130115) * subset_data_complete$ref_college_grad +
      (-0.524241462) * subset_data_complete$com_some_college +
      (-0.666407519) * subset_data_complete$com_less_hs +
      (0.357892051) * subset_data_complete$race_asian +
      (-0.233977902) * subset_data_complete$race_black +
      (0.105956535) * subset_data_complete$race_white +
      (0.017010060) * subset_data_complete$race_other +
      (-0.102803632) * subset_data_complete$sex_female +
      rnorm(dim(subset_data_complete)[1], mean = 0, sd = 1.388), 0), 5),
  subset_data_complete$fipr)


fipr3 <- ifelse(
  is.na(subset_data_complete$fipr),
  pmin(pmax(
    1.937103821 + (0.712372698) * subset_data_complete$ref_some_college +
      (1.692130115) * subset_data_complete$ref_college_grad +
      (-0.524241462) * subset_data_complete$com_some_college +
      (-0.666407519) * subset_data_complete$com_less_hs +
      (0.357892051) * subset_data_complete$race_asian +
      (-0.233977902) * subset_data_complete$race_black +
      (0.105956535) * subset_data_complete$race_white +
      (0.017010060) * subset_data_complete$race_other +
      (-0.102803632) * subset_data_complete$sex_female +
      rnorm(dim(subset_data_complete)[1], mean = 0, sd = 1.388), 0), 5),
  subset_data_complete$fipr)


fipr4 <- ifelse(
  is.na(subset_data_complete$fipr),
  pmin(pmax(
    1.937103821 + (0.712372698) * subset_data_complete$ref_some_college +
      (1.692130115) * subset_data_complete$ref_college_grad +
      (-0.524241462) * subset_data_complete$com_some_college +
      (-0.666407519) * subset_data_complete$com_less_hs +
      (0.357892051) * subset_data_complete$race_asian +
      (-0.233977902) * subset_data_complete$race_black +
      (0.105956535) * subset_data_complete$race_white +
      (0.017010060) * subset_data_complete$race_other +
      (-0.102803632) * subset_data_complete$sex_female +
      rnorm(dim(subset_data_complete)[1], mean = 0, sd = 1.388), 0), 5),
  subset_data_complete$fipr)


fipr5 <- ifelse(
  is.na(subset_data_complete$fipr),
  pmin(pmax(
    1.937103821 + (0.712372698) * subset_data_complete$ref_some_college +
      (1.692130115) * subset_data_complete$ref_college_grad +
      (-0.524241462) * subset_data_complete$com_some_college +
      (-0.666407519) * subset_data_complete$com_less_hs +
      (0.357892051) * subset_data_complete$race_asian +
      (-0.233977902) * subset_data_complete$race_black +
      (0.105956535) * subset_data_complete$race_white +
      (0.017010060) * subset_data_complete$race_other +
      (-0.102803632) * subset_data_complete$sex_female +
    rnorm(dim(subset_data_complete)[1], mean = 0, sd = 1.388), 0), 5),
  subset_data_complete$fipr)


tap1 <- lm(tap_water ~ reference_education + combined_education +
new_race_ethnicity + age + sex + fipr1, data = subset_data_complete)
tap1
tap2 <- lm(tap_water ~ reference_education + combined_education +
new_race_ethnicity + age + sex + fipr2, data = subset_data_complete)
summary(tap2)
tap3 <- lm(tap_water ~ reference_education + combined_education +
new_race_ethnicity + age + sex + fipr3, data = subset_data_complete)
summary(tap3)
tap4 <- lm(tap_water ~ reference_education + combined_education +
new_race_ethnicity + age + sex + fipr4, data = subset_data_complete)
summary(tap4)
tap5 <- lm(tap_water ~ reference_education + combined_education +
new_race_ethnicity + age + sex + fipr5, data = subset_data_complete)
summary(tap5)

coef(tap1)
coef(tap2)
coef(tap3)

### table of different coefficients:

coef_df <- as.data.frame(coef(tap1))
str(coef_df)

coef_df$tap2 <- coef(tap2)
coef_df$tap3 <- coef(tap3)
coef_df$tap4 <- coef(tap4)
coef_df$tap5 <- coef(tap5)

coefficients(tap4)
str(anova(tap4))

stderr_df <- as.data.frame(summary(tap1)$coefficients[, 2])
stderr_df$tap2 <- summary(tap2)$coefficients[, 2]
stderr_df$tap3 <- summary(tap3)$coefficients[, 2]
stderr_df$tap4 <- summary(tap4)$coefficients[, 2]
stderr_df$tap5 <- summary(tap5)$coefficients[, 2]
```

```
###fipr & patient education to predict tap water (figure 11)
 ggplot(subset_data_complete , aes(x = imputed_fipr ,
y = jitter(tap_water , factor=10), color = combined_education)) +
  geom_point(alpha=0.5) +
  geom_abline(intercept=538.8, slope=38.8, colour= ”salmon”)+
  geom_abline(intercept=373.9, slope=38.8, colour=”springgreen3”)+
  geom_abline(intercept=176.9, slope=38.8, colour=”cornflowerblue”)+
  #geom_smooth(method = ”lm”, se = FALSE, aes(group = combined_education)) +
  #geom_smooth(aes(y=predicted_tap , group = combined_education))+
  labs(title = ”Tap Water Consumption by Fipr and Patient Education”,
      x = ”Fipr”,
      y = ”Tap Water”,
      color = ”Patient Education”) +
  ylim(0,800)+
  theme_minimal()

 colors()
## fipr & race to predict tap water(figure 10)
 ggplot(subset_data_complete , aes(x = imputed_fipr ,
y = jitter(tap_water , factor=10), color = new_race_ethnicity)) +
  geom_point(alpha=0.5) +
  geom_abline(intercept=538.8, slope=38.8, colour=”salmon”)+
  geom_abline(intercept=722.9, slope=38.8, colour=”olivedrab3”)+
  geom_abline(intercept=556, slope=38.8, colour=”springgreen3”)+
  geom_abline(intercept=825.7, slope=38.8, colour=”cornflowerblue”)+
  geom_abline(intercept=792.8, slope=38.8, colour=”mediumpurple1”)+
  #geom_smooth(method = ”lm”, se = FALSE, aes(group = new_race_ethnicity)) +
  #geom_smooth(aes(y = predicted_tap , group = new_race_ethnicity))+
  labs(title = ”Tap Water Consumption by Fipr and Race/Ethnicity”,
      x = ”Fipr”,
      y = ”Tap Water”,
      color = ”Race/Ethnicity”) +
  ylim(0,1200)+
  theme_minimal()+
  theme(plot.title = element_text(hjust=0.5))
```