The Impact of Client Balance and Credit Default on Term Deposit Subscription, Controlling for Socioeconomic Factors

Alison Wong and Yehchan Yoo

Autumn 2025: BIOSTAT 531

Abstract

This study analyzes the Bank Marketing dataset from a Portuguese banking institution to quantify the effects of average yearly balance and credit default on term deposit subscription while controlling for missing socioeconomic data. Assuming a Missing at Random (MAR) mechanism based on observed data patterns, we utilized Complete Case (CC) analysis, Bayesian likelihood, and Multiple Imputation (MI) methods to handle missing data inside the dataset. We found that higher balances significantly increased subscription odds ($\approx$ 26% per 10,000€), while credit default reduced them ($\approx$ 38–41%). Notably, CC analysis overestimated the negative effect of default compared to the other two methods, hinting at potential bias from using CC analysis.

1 Introduction

For our final project, we aimed to analyze a bank's marketing campaign dataset to investigate the relationship between term deposit subscription and both average yearly balance and credit default, while accounting for auxiliary variables and missing data using missing data mechanisms.

We performed such analysis, as we believed that statistically modeling the term deposit subscription as a function of key financial indicators — average yearly balance and credit default — would yield actionable segmentation for a bank's future marketing campaigns, with auxiliary variables and missing data mechanisms included to ensure the reliability and precision of the estimated effects.

2 Dataset Overview

The dataset utilized for this project was the **Bank Marketing** dataset, sourced from the UCI Machine Learning Repository. It contained data related to direct marketing campaigns conducted by an unnamed Portuguese banking institution (Moro et al., 2014a). The dataset was collected by the institution from May 2008 to June 2013 and focused on the use of telemarketing phone calls for selling long-term deposits to a set list of clients; it was not specified how the bank obtained this list of clients (Moro et al., 2014b, p. 23).

2.1 Missingness Mechanism

The original dataset consisted of 41,188 observations and 17 variables. The original dataset had too many variables, especially unordered categorical variables, which caused computational and methodological issues with our research, as noted further in detail in our appendix at Section 7.1.

So, the initial data processing phase focused on dropping less important covariates, converting some multi-level categorical variables to binary variables, and properly processing any ordinal variable — leading to the list of variables seen in Table 1. As noted before, the detailed justification for selecting these variables can be found in the Appendix at Section 7.1.

Table 1: Final Feature Set Used for Modeling

Feature	Type	Description
Input Variables (Predictors)		
`age`	Numeric	Client's age
`jobUnemployed`	Binary	Whether the client is unemployed or not
`education`	Ordinal	Client's education level ("primary", "secondary", "tertiary")
`default`	Binary	Whether the client has credit in default.
`balance`	Numeric	Client's average yearly balance (euros)
`housing`	Binary	Whether the client has housing loan
`loan`	Binary	Whether the client has personal loan.
`contactCellular`	Binary	Whether the client was contacted via cellular (If the client was *not* contacted via cellular, the client was contacted via telephone)
`previous`	Numeric	Number of contacts performed before this campaign and for this client
Outcome Variable		
`y`	Binary	Whether the client subscribed to a term deposit

Within the dataset, the variables `job`, `education`, and `contact` had missing data. The variable `job` had 288 missing cases (0.64% of the observations), the variable `education` had 1,857 missing cases (4.11%), and the variable `contact` had 36,959 missing cases (28.8%). Running the `aggr()` function on the dataset, we did not find any particular non-general missingness pattern, as noted in the Appendix in Figure 1.

Also, through pairwise variable comparisons, we found that a noticeably higher proportion of clients with missing values of `contactCellular` had housing loans (71.9%) than others (49%) & did not subscribe to a term deposit (14.8%) than others (4.1%). Additionally, clients with missing values of `jobUnemployed` had a noticeably higher distribution of age (with a median of 47) than others (median of 39); this was also the case with clients with missing `education` values (with a median age of 39 for those without missing `education` values and 45 for the rest). Additionally, a noticeably higher proportion of clients with missing `education` values had housing loans (56.1%)

and personal loans (16.4%) than others (44.2% and 7.2%, respectively).

These observed trends led us to the assumption that the data was **missing at random (MAR)**. The bank appeared to have incomplete data for clients who already possess some form of financial engagement or demographic indicators of life stage. Specifically, the high rate of missing `contactCellular` values among clients with housing loans suggests that, for long-term mortgage clients, the bank may not have proactively recorded or updated the cellular contact method as it focused on home/landline contact or mailing address. As for the missing `education` and `jobUnemployed` values skewing towards clients with older ages, we believed that the older clients — potentially retired or from a time when such detailed demographic information was not standard bank procedure — may have incomplete records in the institution. The finding that clients with missing education also have higher rates of both housing and personal loans suggested that the loan history itself—an observed variable—is a predictor of which clients' non-essential demographic data (like education) is missing. The clients who were most likely to have missing education data might have been those who became customers and secured their mortgages or personal loans decades ago when the bank might have been primarily focused on income, assets, and credit history for loan approval, and collecting detailed, standardized education status was not a mandatory or routine part of the loan application process.

3 Methods

After the aforementioned data analysis process, we utilized **complete case (CC) analysis, Bayesian likelihood method, and multiple imputation (MI) method** to rigorously analyze the dataset.

We used CC analysis as a control method to compare the results of Bayesian and multiple imputation methods to. Among the likelihood-based methods, we chose the Bayesian method over the maximum likelihood method, as data augmentation within the Bayesian method allowed for uncertainty estimates that the Expectation-Maximization algorithm within the maximum likelihood method could not provide (Little, 2024, pp. 160-161). Last but not least, we also utilized the multiple imputation method, as it is the gold standard for handling missing data, despite potentially requiring the correct model specification to work properly (Erler et al., 2016).

Our outcome model was as follows:

$$\begin{aligned} \texttt{y} \sim\ & \texttt{contactCellular} + \texttt{jobUnemployed} + \texttt{education\_ordinal} \\ & + \texttt{balance} + \texttt{default} + \texttt{age} + \texttt{housing} + \texttt{loan} + \texttt{previous} \end{aligned}$$

Here, the top row of variables were ones with missing data, while the bottom row of variables were fully observed. The variables colored blue were the main predictors, while the other predictors were auxiliary.

3.1 Complete Case Analysis

The complete case analysis involved fitting a logistic regression model on only the data with complete cases in accordance with the outcome model. Such analysis involved the use of the `glm()` function in R.

3.2 Bayesian Likelihood

For the Bayesian likelihood method, we employed the **factored regression model (FRM)** under a **fully Bayesian (FB)** estimation framework, as implemented via the `mdmb::frm_fb` package in R. This methodology used Markov Chain Monte Carlo (MCMC) sampling to estimate the parameters of the joint distribution, which simultaneously accounted for the uncertainty introduced by the missing data and the model parameters. We ran this method for 3,000 iterations with 500 burn-ins and with a set maximum of 500 values that could be saved to the MCMC chain.

3.3 Multiple Imputation

For the multiple imputation method, we utilized the R package for Multivariate Imputation by Chained Equations or `mice` to impute the dataset multiple times and then fit the outcome model on the imputed datasets. The multiple imputation method used the outcome model for both imputation and analysis. We also ran the multiple imputation process for $m = 5$ imputations with a maximum of 5 iterations per imputation, as these were the default values for the numbers of imputations and of per-imputation iterations in the `mice` package (van Buuren & Groothuis-Oudshoorn, 2024). For reproducibility, we set a seed of 531 before processing with the multiple imputation method.

4 Results

The results of our research can be found in Table 2.

Table 2: Comparison of Log-Odds Estimates (β)

Predictor	**Method**	**Coef. (β)** (Log-Odds)	**SE**	**95% Interval**	
				Lower Bound	Upper Bound
`balance` (per 1€)	Bayesian	2.288×10^{-5}	3.958×10^{-6}	1.519×10^{-5}	3.105×10^{-5}
	MI	2.284×10^{-5}	3.943×10^{-6}	1.511×10^{-5}	3.056×10^{-5}
	CC	2.263×10^{-5}	4.293×10^{-6}	1.418×10^{-5}	3.104×10^{-5}
`default` (T vs. F)	Bayesian	−0.5058	0.1429	−0.7822	−0.2276
	MI	−0.4843	0.1464	−0.7712	−0.1975
	CC	−0.6880	0.1805	−1.0599	−0.3500

Note: The CC Analysis lost 14,304 observations due to listwise deletion.

Here are also our calculations of the effects of `balance` and `default` on the odds:

- **Balance (Effect per 10,000€):**
  - **Bayesian** ($\text{OR} = \exp(2.288 \times 10^{-5} \times 10000) \approx 1.257$): $\sim 26\%$ increase in odds
  - **MI** ($\text{OR} = \exp(2.284 \times 10^{-5} \times 10000) \approx 1.257$): $\sim 26\%$ increase in odds
  - **CC** ($\text{OR} = \exp(2.263 \times 10^{-5} \times 10000) \approx 1.254$): $\sim 25\%$ increase in odds
- **Credit Default Status:**
  - **Bayesian (OR** $= \exp(-0.5058) \approx 0.603$**):** $\sim 41\%$ reduction in odds.
  - **MI (OR** $= \exp(-0.4843) \approx 0.616$**):** $\sim 38\%$ reduction in odds.
  - **CC (OR** $= \exp(-0.6880) \approx 0.502$**):** $\sim 50\%$ reduction in odds.

Overall, our results showed that `balance` had a statistically significant positive effect on term deposit subscription, as seen by the 95% intervals lying entirely above 0 for all three methods. That is, we found that wealthy clients (with higher balances) were more likely to subscribe, accounting for socioeconomic factors such as their ages, education levels, and existing loan statuses.

On the other hand, `default` had a statistically significant negative effect on the outcome, as seen by the 95% intervals lying entirely below 0 for all methods. That is, clients with credit in default were significantly less likely to subscribe when controlling for the same socioeconomic factors.

For the main predictors `balance` and `default`, we noticed that the CC method provided coefficient estimates that were noticeably different from the MI and Bayesian approaches, especially for `default`. Seeing that the data was assumed to be missing at random (MAR), we found that the CC method may be biased in its estimations of effects for both main predictors.

4.1 Diagnostics

For diagnostics, we created trace plots for both the Bayesian and multiple imputation methods. The trace plots for the Bayesian method mostly did not show concerning trends, though some graphs (e.g., `contactCellular ON (Intercept)` in the last graph (4/4)) did show signs of slow mixing and/or came out with low ESS and high $\hat{R}$ values. The trace plots from the multiple imputation method did not show any concerning trend. The trace plots for the Bayesian method can be found in the Appendix in Section 7.3.2, while the trace plot for the multiple imputation method can also be found in the Appendix in Section 7.3.3.

It should be noted that the multiple imputation method did lead to a warning that one observation was found with a fitted probability that was very close to 0 or 1. This observation notably had a value of 275 for the variable `previous`, which was a massive outlier with all other observations having values of less than 60 for `previous`. However, we did not drop this row, as we did not want to risk the information loss from dropping a case.

5 Limitations

While our missing data analysis and reasoning assumed that the data was missing at random (MAR), the possibility that the data was missing not at random (MNAR) remains. For instance, high-value clients might systematically have certain data fields left blank by the bank due to data collection protocols. As the UCI Machine Learning Repository and the paper associated with the dataset did not include much information on how the bank collected the personal information of its clients and how the bank created its list of clients for marketing campaigns, we did find that the data could be reasoned to be MNAR. However, due to time and computational restraints, we were unable to perform sensitivity analysis for our models.

Also, due to computational constraints of the `mdmb` package, we are unable to use nominal variables that were provided in the dataset (e.g., job categories, marital status), which could improve the precision of the estimate. Similarly, due to time and computational constraints, we were unable to fit models that could take into account potential interactions in our approaches (e.g. using `smcfcs`), and we did not use particularly high numbers of iterations for our Bayesian and MI approaches.

6 Conclusion

Our analysis confirms that financial stability is a robust predictor of term deposit subscription: higher yearly balances increased the odds of subscription by approximately 26% per 10,000€, while credit default reduced the odds by roughly 40% according to the Bayesian and MI approaches. Crucially, the CC analysis exaggerated the negative impact of credit default compared to the Bayesian and MI approaches, potentially illustrating the bias introduced by listwise deletion when data is missing at random. Consequently, accurate bank marketing strategies must rely on proper missing data techniques to avoid distorting the risk profiles of potential subscribers.

More information on the code and the output can be found in Sections 7.2 and 7.3, respectively.

References

Erler, N. S., Rizopoulos, D., Rosmalen, J. v., Jaddoe, V. W. V., Franco, O. H., & Lesaffre, E. M. E. H. (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full bayesian approach. *Statistics in Medicine*, *35*(17), 2955–2974. https://doi.org/https://doi.org/10.1002/sim.6944

Little, R. J. (2024). Missing data analysis. *Annual Review of Clinical Psychology*, *20*, 149–173. https://doi.org/10.1146/annurev-clinpsy-080822-051727

Moro, S., Cortez, P., & Rita, P. (2014a). Bank Marketing. https://doi.org/10.24432/C5K306

Moro, S., Cortez, P., & Rita, P. (2014b). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. https://doi.org/10.1016/j.dss.2014.03.001

van Buuren, S., & Groothuis-Oudshoorn, K. (2024). mice: Multivariate imputation by chained equations [R package version 3.17.0, retrieved from RDocumentation]. https://www.rdocumentation.org/packages/mice/versions/3.17.0/topics/mice

7 Appendix

7.1 Variable Selection

Several variables were intentionally excluded from the Bank Marketing dataset to ensure a simplified feature set and adhere to specific methodological constraints imposed by the missingness modeling approach, as including all the variables in the original dataset led to the breakdown of both the MI and Bayesian methods in R. We especially removed multi-level categorical variables or simplified them to binary variables, as it was difficult to handle these variables both analytically and computationally.

The `marital` variable was a nominal categorical variable with three distinct levels: "married", "divorced", and "single". It was removed because (`mdmb`) does not readily handle non-ordinal, multi-level categorical variables without transforming them via techniques like one-hot encoding. Since this variable could not be easily converted into a single binary variable of primary interest, it was excluded to simplify the feature space and maintain consistency with the required variable types (numeric, binary, or ordinal).

A set of variables related to the last contact of the current campaign and results from previous campaigns were also dropped due to the extreme complexity in handling their inherent missingness structures: `day`, `month`, `duration`, `pdays`, and `poutcome`. These were deemed too challenging to handle robustly within the scope and constraints of the project, potentially leading to non-trivial data cleaning solutions and unreliable signals.

The `campaign` variable, which recorded the number of contacts performed during the current campaign, was also removed. It was considered highly **redundant** in the presence of the `previous` variable (number of contacts performed before this campaign). The `previous` variable provides a clearer signal of the client's past interaction history, and including both highly correlated count variables was unnecessary and could have potentially introduced multicollinearity issues for our regression models.

7.2 Code

It should be noted that this section only includes R code needed to reproduce our research, not the code output. The output can be found in Section 7.3.

7.2.1 Importing Libraries and Loading Dataset

```r
# Import libraries
library(tidyverse)
library(VIM)
library(mdmb)
library(mice)
library(mlogit)
```

```
library(smcfcs)
library(mitools)

# Load dataset
df <- read.csv("bank-full.csv", sep = ";")
```

Now, the dataset puts the unknown values as the "unknown" string. We will replace them with NA for easier handling.

```
bank <- df %>%
  mutate(across(where(is.character), ~ na_if(., "unknown")))
```

From here, we need to clean the data in two different ways: one for the Bayesian likelihood method and one for missingness analysis, complete case analysis, and multiple imputation. This is due to differences between `mice` (for multiple imputation) and `mdmb` (for Bayesian likelihood method) in how the ordinal and binary variables are handled.

7.2.2 Data Processing for Missingness Analysis, Complete Case Analysis, and Multiple Imputation

Here, we start by converting binary variables into Boolean values.

```
to_binary <- function(x) {
  x == "yes"
}
bank$y       <- to_binary(bank$y)
bank$default <- to_binary(bank$default)
bank$housing <- to_binary(bank$housing)
bank$loan    <- to_binary(bank$loan)
```

Now, since binary variables are much easier to handle than categorical variables with multiple levels, we can convert `job` to binary variable `jobUnemployed` by setting it to `TRUE` if `job` is `"unemployed"` and `FALSE` otherwise. Note that this is done while preserving NA values.

```
# 1. contact (contactCellular)
bank$contactCellular <- (bank$contact == "cellular")

# 2. job (jobUnemployed)
bank$jobUnemployed <- (bank$job == "unemployed")

# 3. Feature: education
```

```r
# ordinal factor
bank$education_ordinal <- factor(bank$education,
                                 levels = c("primary", "secondary", "tertiary"),
                                 ordered = TRUE)
```

Then, we will only select the columns we need for our analysis. The processed data for the missingness analysis, complete case analysis, and multiple imputation are stored in `final_df`.

```r
vars_to_keep <- c("balance", "default", "age", "jobUnemployed",
                  "education_ordinal", "housing", "loan", "contactCellular",
                  "previous", "y")
final_df <- bank[, vars_to_keep]
```

7.2.3 Data Processing for Bayesian Likelihood

For the Bayesian likelihood method, we tried to convert binary and ordinal variables to integers, as we found that `mdmb` package works well with data that is in computationally numeric form even when some of the data is binary or ordinal.

```r
# Return bank back to its previous state
# when "Unknown" was just converted to NA
bank <- df %>%
  mutate(across(where(is.character), ~ na_if(., "unknown")))

# Convert to Numeric/Integer
to_binary <- function(x) {
  as.numeric(x == "yes")
}

# Binary
bank$y       <- to_binary(bank$y)
bank$default <- to_binary(bank$default)
bank$housing <- to_binary(bank$housing)
bank$loan    <- to_binary(bank$loan)
```

From here, we can do further processing to convert multi-level categorical variables to binary variables `contactCellular` and `jobUnemployed`. We also need to process the ordinal variable `education_ordinal`.

```r
# 1. Contact
# cellular = 1
# telephone = 0
# unknown or NA in original data = NA (R's default for non-matching conditions)
c_char <- as.character(bank$contact)
bank$contactCellular <- ifelse(c_char == "cellular", 1,
                               ifelse(c_char == "telephone", 0, NA))

# 2. job (jobUnemployed)
# binary (1=Unemployed, 0=Others, NA=Missing original job data)

j_char <- as.character(bank$job)
bank$jobUnemployed <- ifelse(j_char == "unemployed", 1,
                             ifelse(is.na(j_char), NA, 0))

# 3. Feature: education
# Ordinal Numeric (0=Primary, 1=Secondary, 2=Tertiary, NA=Unknown/Missing)
e_char <- as.character(bank$education)
bank$education_ordinal <- ifelse(e_char == "primary", 0,
                                 ifelse(e_char == "secondary", 1,
                                        ifelse(e_char == "tertiary", 2, NA)))
```

Then, we will only select the columns we need for our analysis. The processed data for the Bayesian likelihood method is stored in `final_df_Bayesian`.

```r
vars_to_keep <- c("balance", "default", "age", "jobUnemployed",
                  "education_ordinal", "housing", "loan", "contactCellular",
                  "previous", "y")
final_df_Bayesian <- bank[, vars_to_keep]
```

7.2.4 Missingness Analysis

First, we can look at both how many and what proportion of missing values there are in each variable.

```r
# Check # of missing values per column
sapply(final_df, function(x) sum(is.na(x)))
# Check % of missingness per column
sapply(final_df, function(x) round(sum(is.na(x)) / length(x) * 100, 2))
# Visualize missingness pattern
```

```r
aggr(final_df, numbers = TRUE, prop = TRUE)
```

From here, we can check for relationships between missing variables and non-missing variables with data visualizations. The code below defines a function that compares each pair of a missing variable and a fully observed variable using boxplots for plotting the numeric fully observed variable to the missingness of the missing variable and bar plots for plotting the binary fully observed variable to the missingness of the missing variable. Summary statistics are also printed for each comparison.

```r
# Define fully observed columns
fully_observed_columns <- c("age", "default", "balance", "housing",
                            "loan", "previous", "y")

# Function for checking missingness mechanism
check_missingness_mechanism <- function(data, column_with_missingness) {
  # Create missingness indicator
  data <- data %>%
    mutate(missing_indicator = is.na(.data[[column_with_missingness]]))

  for (col in fully_observed_columns) {
    if (col != column_with_missingness) {
      if (is.numeric(data[[col]])) {
        # Boxplot for numeric columns
        p <- ggplot(data, aes(x = missing_indicator, y = .data[[col]])) +
          geom_boxplot() +
          labs(title = paste("Boxplot of", col, "by missingness in",
                             column_with_missingness),
               x = paste("Missingness in", column_with_missingness),
               y = col)
        print(p)
        # Summary statistics
        summary_stats <- data %>%
          group_by(missing_indicator) %>%
          summarise(
            count = n(),
            mean = mean(.data[[col]], na.rm = TRUE),
            median = median(.data[[col]], na.rm = TRUE),
            sd = sd(.data[[col]], na.rm = TRUE)
          )
        print(summary_stats)
      } else {
```

```
        # Bar plot of % of TRUE value for binary variables
        p <- ggplot(data, aes(x = missing_indicator, fill = .data[[col]])) +
          geom_bar(position = "fill") +
          labs(title = paste("Proportion of", col, "by missingness in",
                             column_with_missingness),
               x = paste("Missingness in", column_with_missingness),
               y = "Proportion") +
          scale_y_continuous(labels = scales::percent)
        print(p)
        # Summary statistics
        summary_stats <- data %>%
          group_by(missing_indicator) %>%
          summarise(
            count = n(),
            proportion_TRUE = mean(.data[[col]] == TRUE, na.rm = TRUE)
          )
        print(summary_stats)
      }
    }
  }
}

# Missing variable #1: jobUnemployed
check_missingness_mechanism(final_df, "jobUnemployed")

# Missing variable #2: education
check_missingness_mechanism(final_df, "education_ordinal")

# Missing variable #3: contactCellular
check_missingness_mechanism(final_df, "contactCellular")
```

7.2.5 Missing Data Method 1: Complete Case Analysis

We can perform a complete case analysis using logistic regression as follows.

```
# Run complete case analysis
fit.cc <- glm(y ~ contactCellular + jobUnemployed + education_ordinal +
                balance + default + age + housing + loan + previous,
              data = final_df,
```

```
              family = binomial)

# Regression summary
print("Complete Case Analysis Summary:")
summary(fit.cc)

# Confidence intervals
confint(fit.cc)
```

7.2.6 Missing Data Method 2: Bayesian Likelihood

The factored regression used for the Bayesian likelihood method decomposes the joint probability distribution of all variables into a sequence of conditional models, where the variables with missing data (X) are modeled sequentially before the final analysis model (Y).

Let Y be the binary outcome (client subscribed a term deposit). Let Z be the vector of the fully observed covariates:

$$Z = \{\text{age, balance, default, housing, loan, previous}\}$$

Let the variables with missing data be:

- X_{edu}: `education_ordinal` (Ordinal: $0, 1, 2$)
- X_{job}: `jobUnemployed` (Binary: $0, 1$)
- X_{cont}: `contactCellular` (Binary: $0, 1$)

The joint probability of the observed data, conditional on Z, is factored as follows, defining the imputation order:

$$\begin{aligned} p(Y, X_{\text{cont}}, X_{\text{job}}, X_{\text{edu}}, Z) = & \underbrace{p(Y|X_{\text{cont}}, X_{\text{job}}, X_{\text{edu}}, Z)}_{\text{Scientific Model (Logistic Regression)}} \cdot \\ & \underbrace{p(X_{\text{cont}}|X_{\text{job}}, X_{\text{edu}}, Z)}_{\text{Auxiliary Model 3 (Logistic Regression)}} \cdot \\ & \underbrace{p(X_{\text{job}}|X_{\text{edu}}, Z)}_{\text{Auxiliary Model 2 (Logistic Regression)}} \cdot \\ & \underbrace{p(X_{\text{edu}}|Z)}_{\text{Auxiliary Model 1 (Ordinal Regression)}} \cdot \\ & \underbrace{p(Z)}_{\text{Distribution of Fully Observed Covariates}} \end{aligned}$$

This complete data likelihood is sampled using MCMC. The scientific model estimates the effect of all predictors on the term deposit subscription (Y), while the three sequential auxiliary models ensure that the posterior distribution reflects the relationships between all variables, including the missing covariates.

First, we need to define the scientific/outcome model and auxiliary models for the missing variables.

```
# observed predictors
obs_preds <- paste(c("balance", "default", "age",
                     "housing", "loan", "previous"),
                   collapse = " + ")

# A. Scientific Model (The Final Step: Y)
model.y <- list(
  "model" = "logistic",
  "formula" = as.formula(paste("y ~ contactCellular + jobUnemployed +
                               education_ordinal +",
                               obs_preds))
)

# B. Auxiliary Models (The Missing Variables)
model.contact <- list(
  "model" = "logistic",
  "formula" = as.formula(paste("contactCellular ~ jobUnemployed +
                               education_ordinal +", obs_preds))
)

model.job <- list(
  "model" = "logistic",
  "formula" = as.formula(paste("jobUnemployed ~ education_ordinal +",
                               obs_preds))
)

model.education <- list(
  "model" = "oprobit",
  "formula" = as.formula(paste("education_ordinal ~",
                               obs_preds))
)

predictor.models <- list(
```

```
  "contactCellular" = model.contact,
  "jobUnemployed" = model.job,
  "education_ordinal" = model.education
)
```

Now, we can do the Fully Bayesian Factored Regression using the `mdmb` package.

```
print("Running Fully Bayesian Factored Regression...")

fit.fb <- mdmb::frm_fb(
  dat = final_df_Bayesian,
  dep = model.y,
  ind = predictor.models,
  verbose = TRUE,
  iter = 3000,
  burnin = 500,
  Nsave = 1000
)

print("Bayesian Model Summary:")
summary(fit.fb)

# Extract Posterior Means (Coefficients)
print("Posterior Means for Y:")

n_models <- length(fit.fb$coef)
print(fit.fb$coef[[n_models]])
```

We can also perform diagnostics by drawing trace plots.

```
# Plot the trace plots in grid
quartz(width = 8.5, height = 11)
par(mfrow = c(3, 2))
plot(fit.fb)
```

7.2.7 Missing Data Method 3: Multiple Imputation

Last but not least, we can run multiple imputation using `mice`.

```r
set.seed(531)  # For reproducibility
# Performing multiple imputation
mice_imputed <- mice(final_df, m = 5, method = make.method(final_df),
                     maxit = 5, seed = 531)

# Fit logistic regression model on each imputed dataset and pool results
imputation_list <- imputationList(complete(mice_imputed, "all"))
fit <- with(imputation_list, glm(y ~ age + jobUnemployed + education_ordinal +
                                   default + balance + housing + loan +
                                   contactCellular + previous,
                                 family = binomial))
pooled_results <- MIcombine(fit)
confint_res <- confint(pooled_results)
print("Summary:")
summary(pooled_results)
print("Pooled Results Confidence Intervals:")
print(confint_res)
```

Afterwards, we can run diagnostics. Here, we create a trace plot. Also, the code indicated at least one observation with a fitted probability that was very close to 0 or 1, so we look at one imputed dataset to see what such observation(s) might be.

```r
## Diagnostics
# 0. Trace plots
plot(mice_imputed)

# 1. Extract the first imputed dataset
check_data <- complete(mice_imputed, 1)

# 2. Fit the standard model on this single dataset
fit_diag <- glm(y ~ age + jobUnemployed + education_ordinal + default + balance +
                housing + loan + contactCellular + previous,
                data = check_data, family = binomial)

# 3. Calculate predicted probabilities
probs <- predict(fit_diag, type = "response")

# 4. Find the "Separated" cases (Probabilities extremely close to 0 or 1)
# "Numerically 0 or 1" usually means < 1e-8 or > 1 - 1e-8
extreme_rows <- which(probs < 1e-8 | probs > (1 - 1e-8))
```

```r
if(length(extreme_rows) > 0) {
  message(paste("Found", length(extreme_rows),
                "observations with probability 0 or 1."))

  # 5. Print the values for the suspected variables for these people
  print(check_data[extreme_rows, c("y", "balance", "previous", "housing", "loan")])

  # Compare them to the average person in the dataset
  message("\n--- For Comparison: Averages for the whole dataset ---")
  print(summary(check_data[, c("balance", "previous")]))
} else {
  message("No numerical 0/1 found in this specific imputed set.")
}

# 6. Visual check for outliers vs Outcome
par(mfrow=c(1,2))
boxplot(balance ~ y, data = check_data, main = "Balance by Outcome",
        col = "lightblue")
boxplot(previous ~ y, data = check_data, main = "Previous Contacts by Outcome",
        col = "lightgreen")
```

7.3 Output

7.3.1 Missingness Pattern

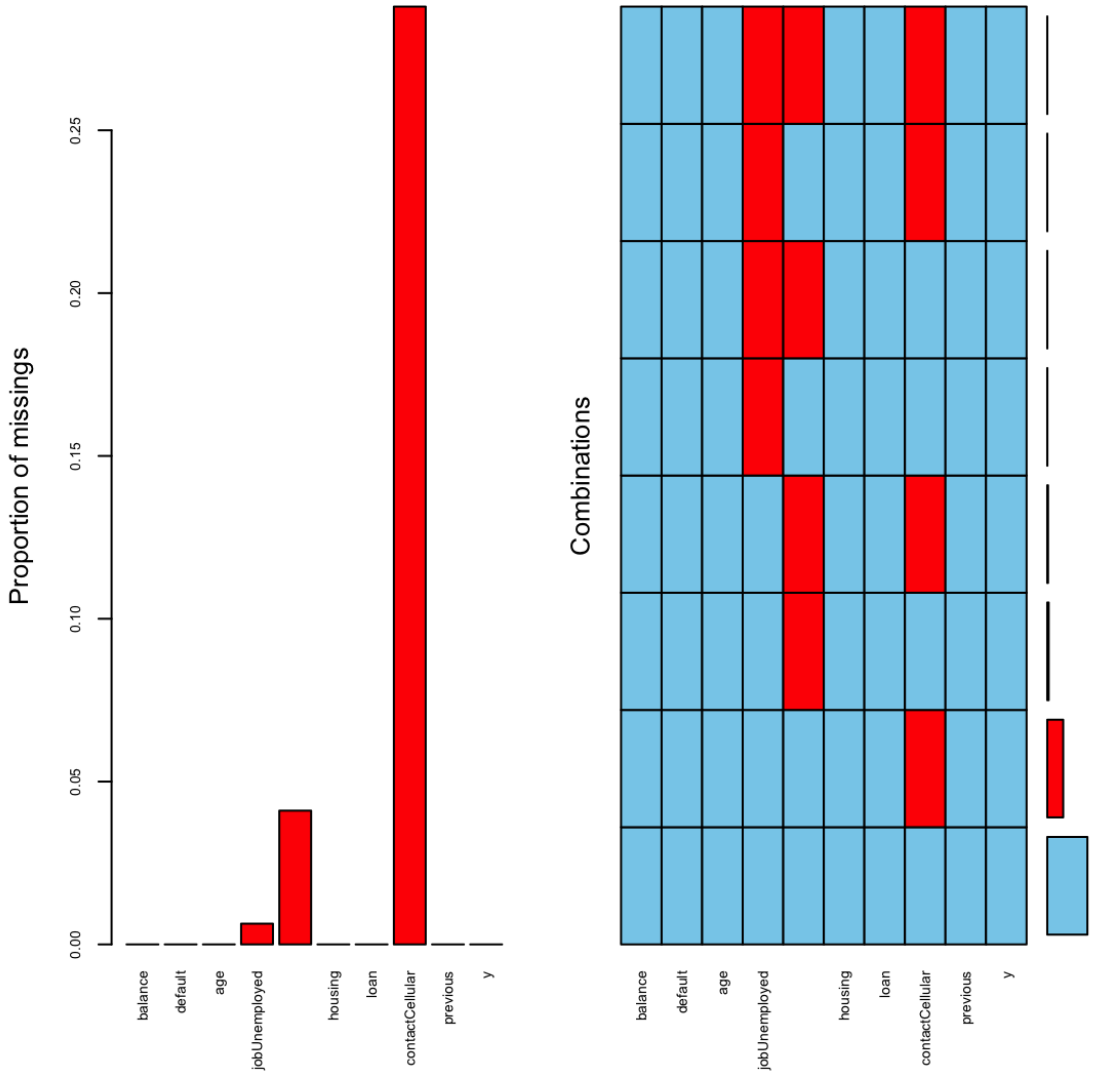

Figure 1: Missingness Pattern of the Bank Marketing Dataset

7.3.2 Trace Plots from Bayesian Analysis

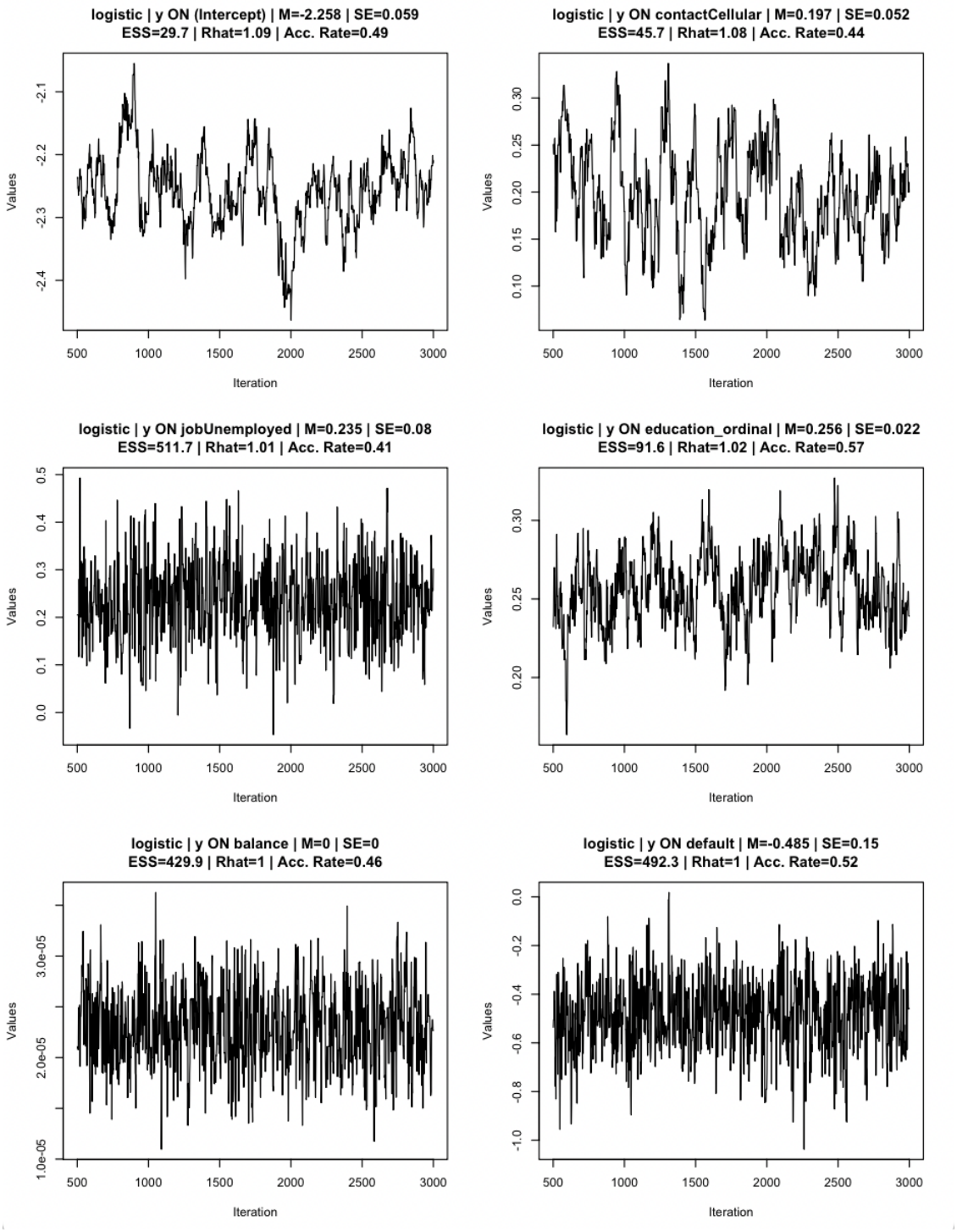

Figure 2: Trace Plots for Bayesian Method (1/5)

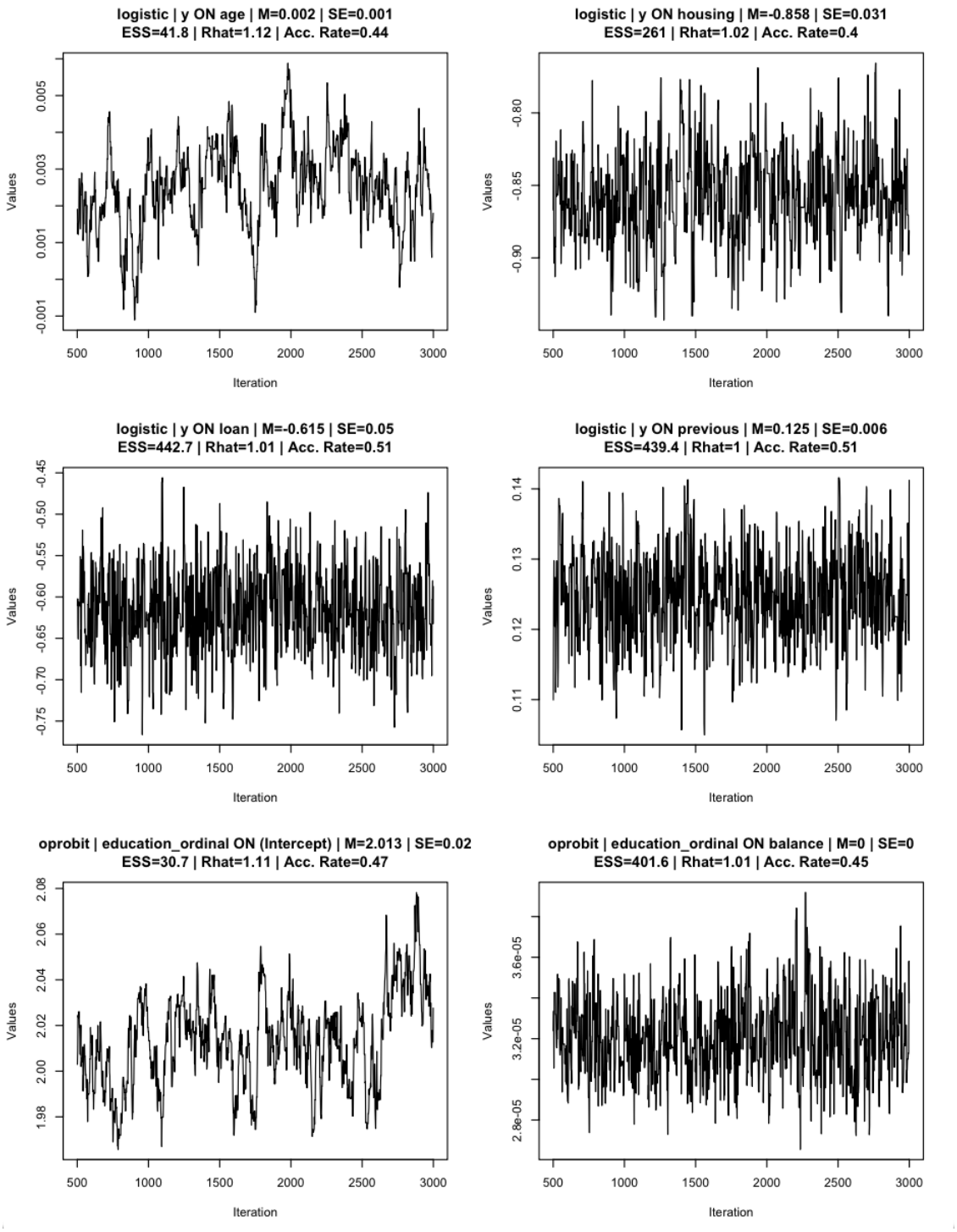

Figure 3: Trace Plots for Bayesian Method (2/5)

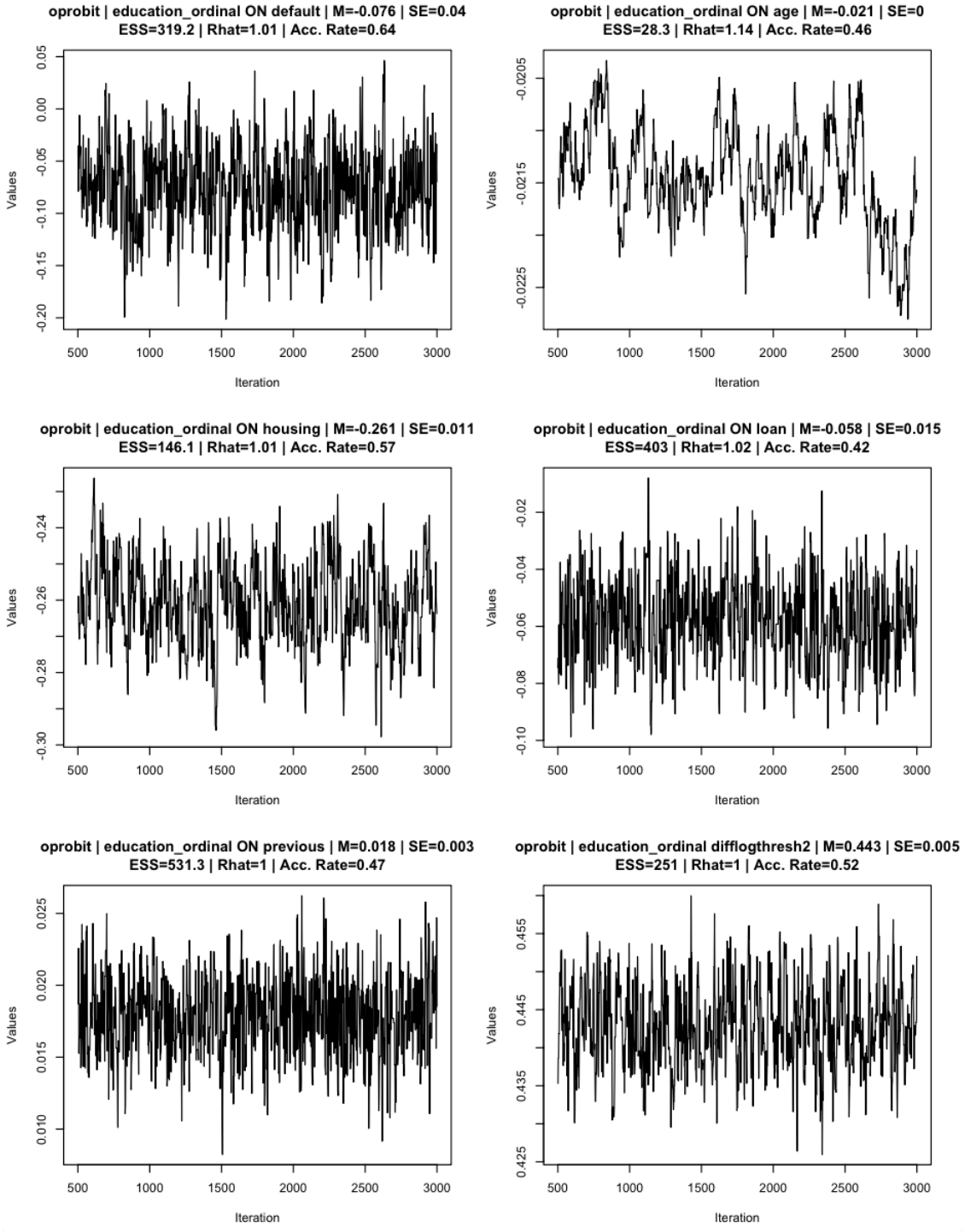

Figure 4: Trace Plots for Bayesian Method (3/5)

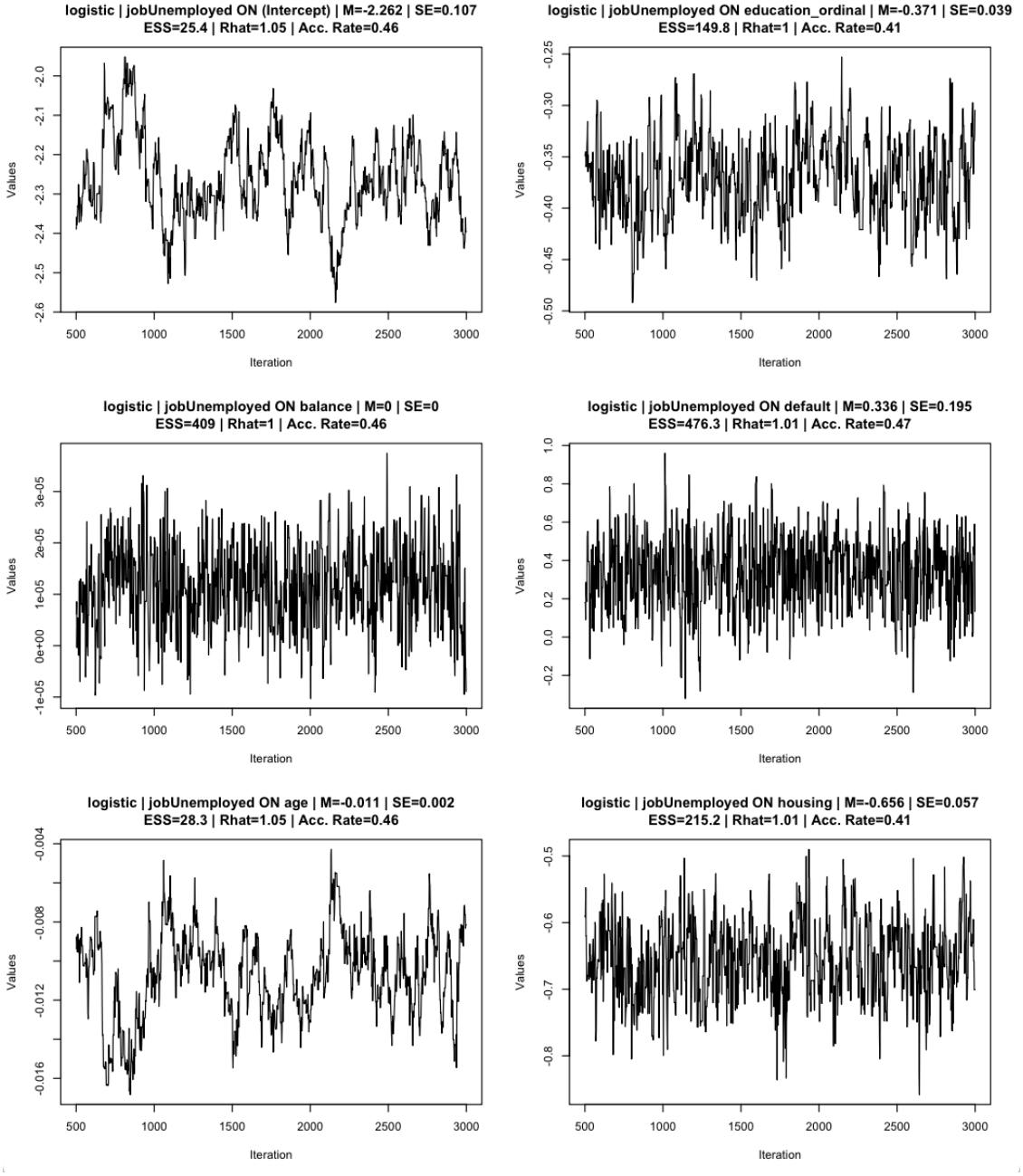

Figure 5: Trace Plots for Bayesian Method (4/5)

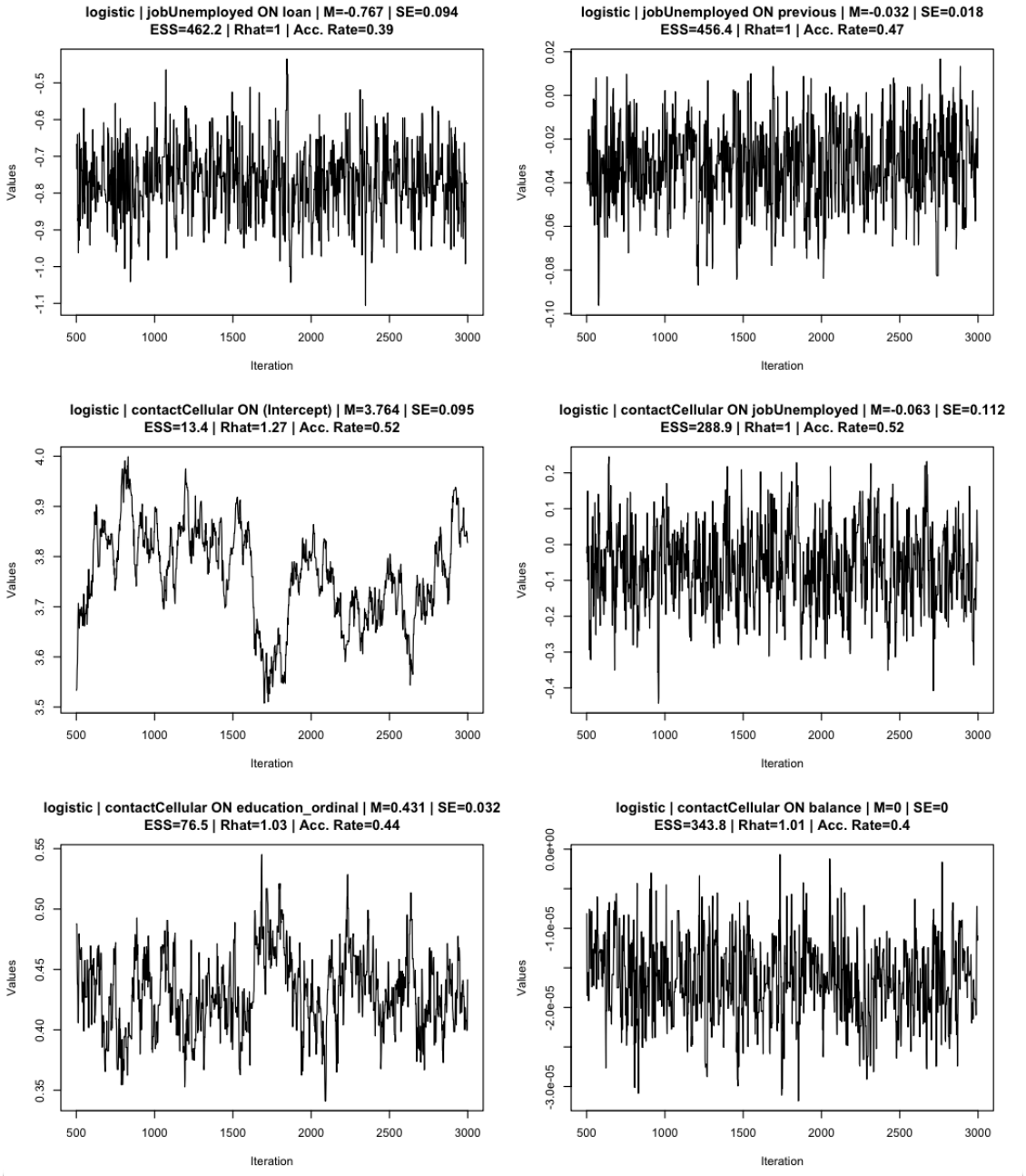

Figure 6: Trace Plots for Bayesian Method (5/5)

7.3.3 Trace Plots from Multiple Imputation

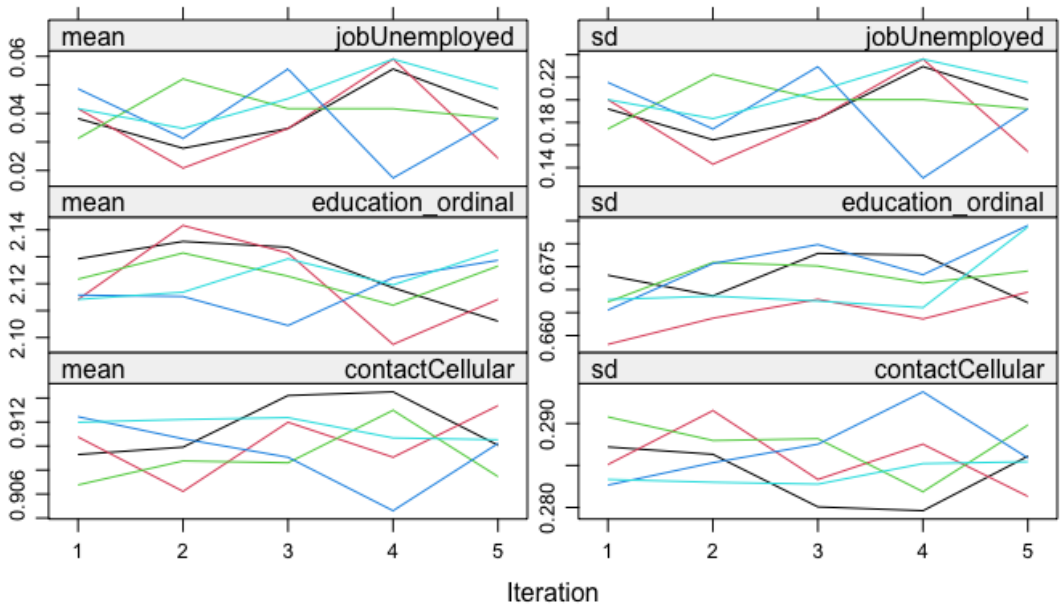

Figure 7: Trace Plots for Multiple Imputation

7.3.4 CC Analysis Output

```
Call:
glm(formula = y ~ contactCellular + jobUnemployed + education_ordinal +
    balance + default + age + housing + loan + previous, family = binomial,
    data = final_df)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.931e+00  9.775e-02 -19.754  < 2e-16 ***
contactCellularTRUE  2.465e-01  6.241e-02   3.950 7.82e-05 ***
jobUnemployedTRUE    2.059e-01  8.571e-02   2.402 0.016284 *
education_ordinal.L  2.777e-01  4.030e-02   6.892 5.50e-12 ***
education_ordinal.Q -1.340e-02  2.988e-02  -0.448 0.653908
balance              2.263e-05  4.293e-06   5.271 1.36e-07 ***
defaultTRUE         -6.880e-01  1.805e-01  -3.811 0.000138 ***
age                  4.541e-03  1.524e-03   2.979 0.002891 **
housingTRUE         -7.847e-01  3.531e-02 -22.223  < 2e-16 ***
```

```
loanTRUE            -6.892e-01  5.544e-02 -12.431  < 2e-16 ***
previous             9.725e-02  6.302e-03  15.433  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25699  on 30906  degrees of freedom
Residual deviance: 24542  on 30896  degrees of freedom
  (14304 observations deleted due to missingness)
AIC: 24564

Number of Fisher Scoring iterations: 5

                           2.5 %        97.5 %
(Intercept)         -2.123394e+00 -1.740181e+00
contactCellularTRUE  1.255097e-01  3.702022e-01
jobUnemployedTRUE    3.533443e-02  3.715082e-01
education_ordinal.L  1.992372e-01  3.572196e-01
education_ordinal.Q -7.211069e-02  4.504783e-02
balance              1.418242e-05  3.104022e-05
defaultTRUE         -1.059872e+00 -3.500219e-01
age                  1.549206e-03  7.524944e-03
housingTRUE         -8.540437e-01 -7.156322e-01
loanTRUE            -7.991626e-01 -5.817745e-01
previous             8.492162e-02  1.096272e-01
```

7.3.5 Bayesian Analysis Output

```
------------------------------------------------------------------
mdmb 1.9-22 (2024-07-15 17:26:54)
R version 4.3.2 (2023-10-31) aarch64, darwin20 | nodename=Alisons-MacBook-Pro-9.lo
cal | login=root
Date of Analysis: 2025-11-24 16:17:01.568635
Time difference of 43.21928 mins
Computation Time: 43.21928
```

```
Call:
mdmb::frm_fb(dat = final_df, dep = model.y, ind = predictor.models,
    verbose = TRUE, iter = 3000, burnin = 500, Nsave = 1000)

----------------------------------------------------------------
Number of observations = 45211

Number of iterations = 2999
Number of burnin iterations = 500

Number of estimated parameters = 35
----------------------------------------------------------------
Descriptive Statistics (Imputed Values)
            variable N_obs N_miss         M        SD
10                 y 45211      0    0.1170    0.3214
3    contactCellular 32191  13020    0.9100    0.2862
7      jobUnemployed 44923    288    0.0291    0.1680
5  education_ordinal 43354   1857    1.1477    0.6656
2            balance    NA     NA 1362.2721 3044.7658
4            default    NA     NA    0.0180    0.1330
1                age    NA     NA   40.9362   10.6188
6            housing    NA     NA    0.5558    0.4969
8               loan    NA     NA    0.1602    0.3668
9           previous    NA     NA    0.5803    2.3034
----------------------------------------------------------------
Predictor Matrix
                 y contactCellular jobUnemployed education_ordinal balance default
y                0               1             1                 1       1       1
contactCellular  0               0             1                 1       1       1
jobUnemployed    0               0             0                 1       1       1
education_ordinal 0              0             0                 0       1       1
balance          0               0             0                 0       0       0
default          0               0             0                 0       0       0
age              0               0             0                 0       0       0
housing          0               0             0                 0       0       0
loan             0               0             0                 0       0       0
previous         0               0             0                 0       0       0
                 age housing loan previous
y                  1       1    1        1
```

```
contactCellular     1       1    1        1
jobUnemployed       1       1    1        1
education_ordinal   1       1    1        1
balance             0       0    0        0
default             0       0    0        0
age                 0       0    0        0
housing             0       0    0        0
loan                0       0    0        0
previous            0       0    0        0
-----------------------------------------------------------------
Model Equations
 * Model 4: mdmb::logistic_regression( y ~ contactCellular + jobUnemployed +
 education_ordinal + balance + default + age + housing + loan + previous )
 * Model 3: mdmb::oprobit_regression( education_ordinal ~ balance + default
 + age + housing + loan + previous )
 * Model 2: mdmb::logistic_regression( jobUnemployed ~ education_ordinal + balance
 + default + age + housing + loan + previous )
 * Model 1: mdmb::logistic_regression( contactCellular ~ jobUnemployed +
 education_ordinal + balance + default + age + housing + loan + previous )
-----------------------------------------------------------------
Estimated Parameters

***********************************
Model 4: mdmb::logistic_regression( y ~ contactCellular + jobUnemployed +
education_ordinal + balance + default + age + housing + loan + previous )

   idparm dv                   parm ON     est     se      p lower95 upper95
1      26  y       y ON (Intercept)  1 -2.3026 0.0888 0.0000 -2.4683 -2.1390
2      27  y   y ON contactCellular  1  0.2240 0.0539 0.0000  0.1074  0.3256
3      28  y     y ON jobUnemployed  1  0.2368 0.0815 0.0048  0.0808  0.4038
4      29  y y ON education_ordinal  1  0.2573 0.0217 0.0000  0.2146  0.2986
5      30  y           y ON balance  1  0.0000 0.0000 0.0000  0.0000  0.0000
6      31  y           y ON default  1 -0.5058 0.1429 0.0000 -0.7822 -0.2276
7      32  y               y ON age  1  0.0029 0.0015 0.0256  0.0002  0.0057
8      33  y           y ON housing  1 -0.8556 0.0328 0.0000 -0.9202 -0.7945
9      34  y              y ON loan  1 -0.6121 0.0519 0.0000 -0.7223 -0.5121
10     35  y          y ON previous  1  0.1250 0.0060 0.0000  0.1127  0.1365

***********************************
```

```
Model 3: mdmb::oprobit_regression( education_ordinal ~ balance + default + age +
housing + loan + previous )

  idparm                dv                                parm ON     est     se      p
1     18 education_ordinal education_ordinal ON (Intercept)  1  2.0210 0.0231 0.0000
2     19 education_ordinal     education_ordinal ON balance  1  0.0000 0.0000 0.0000
3     20 education_ordinal     education_ordinal ON default  1 -0.0759 0.0409 0.0752
4     21 education_ordinal         education_ordinal ON age  1 -0.0216 0.0005 0.0000
5     22 education_ordinal     education_ordinal ON housing  1 -0.2614 0.0114 0.0000
6     23 education_ordinal        education_ordinal ON loan  1 -0.0571 0.0153 0.0000
7     24 education_ordinal    education_ordinal ON previous  1  0.0178 0.0029 0.0000
8     25 education_ordinal education_ordinal difflogthresh2  0  0.4428 0.0053 0.0000
  lower95 upper95
1  1.9826  2.0738
2  0.0000  0.0000
3 -0.1560  0.0073
4 -0.0227 -0.0207
5 -0.2825 -0.2389
6 -0.0881 -0.0259
7  0.0121  0.0234
8  0.4324  0.4531

************************************
Model 2: mdmb::logistic_regression( jobUnemployed ~ education_ordinal + balance +
default + age + housing + loan + previous )

  idparm            dv                                 parm ON     est     se      p
1     10 jobUnemployed       jobUnemployed ON (Intercept)  1 -2.3809 0.1303 0.0000
2     11 jobUnemployed jobUnemployed ON education_ordinal  1 -0.3548 0.0412 0.0000
3     12 jobUnemployed           jobUnemployed ON balance  1  0.0000 0.0000 0.1744
4     13 jobUnemployed           jobUnemployed ON default  1  0.3427 0.1935 0.0864
5     14 jobUnemployed               jobUnemployed ON age  1 -0.0085 0.0026 0.0000
6     15 jobUnemployed           jobUnemployed ON housing  1 -0.6364 0.0589 0.0000
7     16 jobUnemployed              jobUnemployed ON loan  1 -0.7688 0.1010 0.0000
8     17 jobUnemployed          jobUnemployed ON previous  1 -0.0314 0.0176 0.0528
  lower95 upper95
1 -2.6168 -2.1113
2 -0.4299 -0.2698
3  0.0000  0.0000
```

```
4 -0.0490  0.7120
5 -0.0139 -0.0037
6 -0.7507 -0.5267
7 -0.9634 -0.5741
8 -0.0677  0.0004

************************************
Model 1: mdmb::logistic_regression( contactCellular ~ jobUnemployed +
education_ordinal + balance + default + age + housing + loan + previous )

  idparm              dv                                parm ON     est     se
1      1 contactCellular       contactCellular ON (Intercept)  1  3.7424 0.0827
2      2 contactCellular     contactCellular ON jobUnemployed  1 -0.0828 0.1102
3      3 contactCellular contactCellular ON education_ordinal  1  0.4350 0.0325
4      4 contactCellular           contactCellular ON balance  1  0.0000 0.0000
5      5 contactCellular           contactCellular ON default  1  0.5451 0.2017
6      6 contactCellular               contactCellular ON age  1 -0.0453 0.0014
7      7 contactCellular           contactCellular ON housing  1  0.1668 0.0399
8      8 contactCellular              contactCellular ON loan  1  0.1037 0.0596
9      9 contactCellular          contactCellular ON previous  1 -0.0023 0.0074
       p lower95 upper95
1 0.0000  3.5916  3.9356
2 0.4560 -0.2992  0.1388
3 0.0000  0.3661  0.4935
4 0.0016  0.0000  0.0000
5 0.0112  0.1582  0.9315
6 0.0000 -0.0483 -0.0425
7 0.0000  0.0821  0.2427
8 0.0928 -0.0125  0.2201
9 0.6976 -0.0153  0.0126

------------------------------------------------------------------------
MCMC Algorithm Informations

   idparm model                dv                                parm Nsampled
1      26     4                 y                       y ON (Intercept)     1250
2      27     4                 y                  y ON contactCellular     1250
3      28     4                 y                    y ON jobUnemployed     1250
4      29     4                 y                y ON education_ordinal     1250
```

```
5      30     4                 y                         y ON balance     1250
6      31     4                 y                         y ON default     1250
7      32     4                 y                             y ON age     1250
8      33     4                 y                         y ON housing     1250
9      34     4                 y                            y ON loan     1250
10     35     4                 y                       y ON previous     1250
11     18     3 education_ordinal     education_ordinal ON (Intercept)     1250
12     19     3 education_ordinal         education_ordinal ON balance     1250
13     20     3 education_ordinal         education_ordinal ON default     1250
14     21     3 education_ordinal             education_ordinal ON age     1250
15     22     3 education_ordinal         education_ordinal ON housing     1250
16     23     3 education_ordinal            education_ordinal ON loan     1250
17     24     3 education_ordinal        education_ordinal ON previous     1250
18     25     3 education_ordinal     education_ordinal difflogthresh2     1250
19     10     2     jobUnemployed         jobUnemployed ON (Intercept)     1250
20     11     2     jobUnemployed   jobUnemployed ON education_ordinal     1250
21     12     2     jobUnemployed             jobUnemployed ON balance     1250
22     13     2     jobUnemployed             jobUnemployed ON default     1250
23     14     2     jobUnemployed                 jobUnemployed ON age     1250
24     15     2     jobUnemployed             jobUnemployed ON housing     1250
25     16     2     jobUnemployed                jobUnemployed ON loan     1250
26     17     2     jobUnemployed            jobUnemployed ON previous     1250
27      1     1   contactCellular       contactCellular ON (Intercept)     1250
28      2     1   contactCellular     contactCellular ON jobUnemployed     1250
29      3     1   contactCellular contactCellular ON education_ordinal     1250
30      4     1   contactCellular           contactCellular ON balance     1250
31      5     1   contactCellular           contactCellular ON default     1250
32      6     1   contactCellular               contactCellular ON age     1250
33      7     1   contactCellular           contactCellular ON housing     1250
34      8     1   contactCellular              contactCellular ON loan     1250
35      9     1   contactCellular          contactCellular ON previous     1250
   effsize accrate Rhat
1     11.3    0.39 1.04
2     38.0    0.42 1.03
3    446.5    0.54 1.00
4    100.2    0.57 1.00
5    400.1    0.49 1.00
6    439.1    0.48 1.00
7     20.8    0.46 1.03
```

```
8    167.3    0.40 1.00
9    485.4    0.52 1.01
10   514.0    0.47 1.01
11    19.8    0.50 1.06
12   366.9    0.54 1.02
13   558.0    0.41 1.00
14    26.1    0.49 1.07
15   214.3    0.46 1.00
16   410.0    0.51 1.00
17   470.1    0.46 1.00
18   189.2    0.42 1.00
19    13.9    0.55 1.22
20   107.8    0.53 1.07
21   405.9    0.47 1.01
22   536.0    0.48 1.00
23    18.0    0.52 1.14
24   208.7    0.40 1.02
25   476.3    0.45 1.00
26   596.5    0.43 1.01
27    17.9    0.51 1.08
28   330.5    0.56 1.00
29    90.4    0.54 1.01
30   569.1    0.39 1.01
31   301.5    0.48 1.02
32    23.5    0.47 1.12
33   160.1    0.50 1.01
34   280.6    0.51 1.02
35   435.3    0.29 1.01

Iterations = 501:2999
Thinning interval = 2
Number of chains = 1
Sample size per chain = 1250

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                                          Mean        SD  Naive SE Time-series SE
contactCellular ON (Intercept)        3.742e+00 8.273e-02 2.340e-03      1.957e-02
```

```
contactCellular ON jobUnemployed     -8.280e-02 1.102e-01 3.117e-03      6.062e-03
contactCellular ON education_ordinal  4.350e-01 3.254e-02 9.204e-04      3.422e-03
contactCellular ON balance           -1.688e-05 4.796e-06 1.356e-07      2.010e-07
contactCellular ON default            5.451e-01 2.017e-01 5.705e-03      1.162e-02
contactCellular ON age               -4.527e-02 1.445e-03 4.087e-05      2.982e-04
contactCellular ON housing            1.668e-01 3.991e-02 1.129e-03      3.154e-03
contactCellular ON loan               1.037e-01 5.959e-02 1.686e-03      3.557e-03
contactCellular ON previous          -2.304e-03 7.352e-03 2.079e-04      3.524e-04
jobUnemployed ON (Intercept)         -2.381e+00 1.303e-01 3.685e-03      3.498e-02
jobUnemployed ON education_ordinal   -3.548e-01 4.120e-02 1.165e-03      3.968e-03
jobUnemployed ON balance              1.102e-05 7.972e-06 2.255e-07      3.957e-07
jobUnemployed ON default              3.427e-01 1.935e-01 5.473e-03      8.358e-03
jobUnemployed ON age                 -8.498e-03 2.587e-03 7.317e-05      6.096e-04
jobUnemployed ON housing             -6.364e-01 5.895e-02 1.667e-03      4.080e-03
jobUnemployed ON loan                -7.688e-01 1.010e-01 2.857e-03      4.628e-03
jobUnemployed ON previous            -3.136e-02 1.755e-02 4.964e-04      7.186e-04
education_ordinal ON (Intercept)      2.021e+00 2.313e-02 6.541e-04      5.196e-03
education_ordinal ON balance          3.194e-05 1.908e-06 5.398e-08      9.963e-08
education_ordinal ON default         -7.591e-02 4.095e-02 1.158e-03      1.733e-03
education_ordinal ON age             -2.162e-02 4.748e-04 1.343e-05      9.292e-05
education_ordinal ON housing         -2.614e-01 1.143e-02 3.232e-04      7.805e-04
education_ordinal ON loan            -5.707e-02 1.534e-02 4.340e-04      7.577e-04
education_ordinal ON previous         1.783e-02 2.879e-03 8.144e-05      1.328e-04
education_ordinal difflogthresh2      4.428e-01 5.252e-03 1.485e-04      3.818e-04
y ON (Intercept)                     -2.303e+00 8.882e-02 2.512e-03      2.640e-02
y ON contactCellular                  2.240e-01 5.387e-02 1.524e-03      8.737e-03
y ON jobUnemployed                    2.368e-01 8.147e-02 2.304e-03      3.856e-03
y ON education_ordinal                2.573e-01 2.168e-02 6.132e-04      2.165e-03
y ON balance                          2.288e-05 3.958e-06 1.120e-07      1.979e-07
y ON default                         -5.058e-01 1.429e-01 4.041e-03      6.818e-03
y ON age                              2.895e-03 1.499e-03 4.239e-05      3.285e-04
y ON housing                         -8.556e-01 3.277e-02 9.269e-04      2.533e-03
y ON loan                            -6.121e-01 5.186e-02 1.467e-03      2.354e-03
y ON previous                         1.250e-01 5.967e-03 1.688e-04      2.632e-04

2. Quantiles for each variable:

                                          2.5%        25%        50%        75%
contactCellular ON (Intercept)       3.592e+00  3.687e+00  3.735e+00  3.788e+00
```

```
contactCellular ON jobUnemployed     -2.992e-01 -1.631e-01 -8.299e-02 -9.874e-03
contactCellular ON education_ordinal  3.661e-01  4.129e-01  4.380e-01  4.570e-01
contactCellular ON balance           -2.583e-05 -2.024e-05 -1.706e-05 -1.361e-05
contactCellular ON default            1.582e-01  4.086e-01  5.528e-01  6.875e-01
contactCellular ON age               -4.826e-02 -4.621e-02 -4.526e-02 -4.423e-02
contactCellular ON housing            8.214e-02  1.403e-01  1.678e-01  1.918e-01
contactCellular ON loan              -1.254e-02  6.239e-02  1.033e-01  1.437e-01
contactCellular ON previous          -1.535e-02 -7.433e-03 -2.521e-03  1.909e-03
jobUnemployed ON (Intercept)         -2.617e+00 -2.475e+00 -2.380e+00 -2.298e+00
jobUnemployed ON education_ordinal   -4.299e-01 -3.825e-01 -3.550e-01 -3.294e-01
jobUnemployed ON balance             -5.239e-06  5.917e-06  1.093e-05  1.693e-05
jobUnemployed ON default             -4.901e-02  2.109e-01  3.474e-01  4.823e-01
jobUnemployed ON age                 -1.393e-02 -1.011e-02 -8.462e-03 -6.664e-03
jobUnemployed ON housing             -7.507e-01 -6.747e-01 -6.340e-01 -5.964e-01
jobUnemployed ON loan                -9.634e-01 -8.289e-01 -7.712e-01 -7.005e-01
jobUnemployed ON previous            -6.771e-02 -4.241e-02 -3.102e-02 -1.866e-02
education_ordinal ON (Intercept)      1.983e+00  2.005e+00  2.017e+00  2.037e+00
education_ordinal ON balance          2.834e-05  3.072e-05  3.194e-05  3.323e-05
education_ordinal ON default         -1.560e-01 -1.005e-01 -7.685e-02 -4.889e-02
education_ordinal ON age             -2.267e-02 -2.192e-02 -2.159e-02 -2.129e-02
education_ordinal ON housing         -2.825e-01 -2.696e-01 -2.612e-01 -2.532e-01
education_ordinal ON loan            -8.811e-02 -6.732e-02 -5.677e-02 -4.759e-02
education_ordinal ON previous         1.215e-02  1.594e-02  1.787e-02  1.971e-02
education_ordinal difflogthresh2      4.324e-01  4.395e-01  4.426e-01  4.464e-01
y ON (Intercept)                     -2.468e+00 -2.368e+00 -2.303e+00 -2.240e+00
y ON contactCellular                  1.074e-01  1.900e-01  2.290e-01  2.593e-01
y ON jobUnemployed                    8.078e-02  1.825e-01  2.385e-01  2.897e-01
y ON education_ordinal                2.146e-01  2.426e-01  2.577e-01  2.719e-01
y ON balance                          1.519e-05  2.035e-05  2.275e-05  2.553e-05
y ON default                         -7.822e-01 -5.977e-01 -5.079e-01 -4.175e-01
y ON age                              2.085e-04  1.796e-03  2.777e-03  4.032e-03
y ON housing                         -9.202e-01 -8.779e-01 -8.544e-01 -8.323e-01
y ON loan                            -7.223e-01 -6.459e-01 -6.112e-01 -5.775e-01
y ON previous                         1.127e-01  1.207e-01  1.253e-01  1.291e-01

                                         97.5%
contactCellular ON (Intercept)        3.936e+00
contactCellular ON jobUnemployed      1.388e-01
contactCellular ON education_ordinal  4.935e-01
contactCellular ON balance           -7.082e-06
```

```
contactCellular ON default             9.315e-01
contactCellular ON age                -4.252e-02
contactCellular ON housing             2.427e-01
contactCellular ON loan                2.201e-01
contactCellular ON previous            1.261e-02
jobUnemployed ON (Intercept)          -2.111e+00
jobUnemployed ON education_ordinal    -2.698e-01
jobUnemployed ON balance               2.563e-05
jobUnemployed ON default               7.120e-01
jobUnemployed ON age                  -3.682e-03
jobUnemployed ON housing              -5.267e-01
jobUnemployed ON loan                 -5.741e-01
jobUnemployed ON previous              4.439e-04
education_ordinal ON (Intercept)       2.074e+00
education_ordinal ON balance           3.580e-05
education_ordinal ON default           7.303e-03
education_ordinal ON age              -2.075e-02
education_ordinal ON housing          -2.389e-01
education_ordinal ON loan             -2.589e-02
education_ordinal ON previous          2.340e-02
education_ordinal difflogthresh2       4.531e-01
y ON (Intercept)                      -2.139e+00
y ON contactCellular                   3.256e-01
y ON jobUnemployed                     4.038e-01
y ON education_ordinal                 2.986e-01
y ON balance                           3.105e-05
y ON default                          -2.276e-01
y ON age                               5.741e-03
y ON housing                          -7.945e-01
y ON loan                             -5.121e-01
y ON previous                          1.365e-01
```

7.3.6 MI Analysis Output

First, the MI code produced this table.

	results	se	(lower	upper)	missInfo
	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
(Intercept)	-2.043403e+00	9.238874e-02	-2.224726e+00	-1.862080e+00	7 %
age	2.912813e-03	1.426887e-03	1.161046e-04	5.709522e-03	1 %
jobUnemployed	2.399357e-01	7.977967e-02	8.356811e-02	3.963032e-01	1 %
education_ordinal.L	3.716716e-01	3.673828e-02	2.996094e-01	4.437337e-01	5 %
education_ordinal.Q	-1.040895e-02	2.784736e-02	-6.508814e-02	4.427024e-02	8 %
defaultTRUE	-4.843047e-01	1.463555e-01	-7.711563e-01	-1.974531e-01	0 %
balance	2.283557e-05	3.943117e-06	1.510719e-05	3.056394e-05	0 %
housingTRUE	-8.567157e-01	3.157153e-02	-9.185949e-01	-7.948366e-01	0 %
loanTRUE	-6.144875e-01	5.058156e-02	-7.136255e-01	-5.153494e-01	0 %
contactCellular	2.196515e-01	5.918104e-02	1.032640e-01	3.360390e-01	11 %
previous	1.246892e-01	6.213500e-03	1.125109e-01	1.368674e-01	0 %

Then, the MI code also produced this console output, containing the confidence intervals.

```
                           2.5 %        97.5 %
(Intercept)         -2.2244815835 -1.862324e+00
age                  0.0001161670  5.709460e-03
jobUnemployed        0.0835703822  3.963009e-01
education_ordinal.L  0.2996658479  4.436773e-01
education_ordinal.Q -0.0649887708  4.417087e-02
defaultTRUE         -0.7711562519 -1.974531e-01
balance              0.0000151072  3.056393e-05
housingTRUE         -0.9185947776 -7.948366e-01
loanTRUE            -0.7136255268 -5.153494e-01
contactCellular      0.1036587546  3.356442e-01
previous             0.1125109409  1.368674e-01
```

Afterwards, the MI diagnostics code produced the following output on the console.

```
Found 1 observations with probability 0 or 1.

--- For Comparison: Averages for the whole dataset ---
    balance          previous
 Min.   : -8019   Min.   :  0.0000
 1st Qu.:    72   1st Qu.:  0.0000
 Median :   448   Median :  0.0000
 Mean   :  1362   Mean   :  0.5803
```

```
 3rd Qu.:  1428   3rd Qu.:  0.0000
 Max.   :102127   Max.   :275.0000
```

The observation with probability 0 or 1 was displayed as below.

	y	balance	previous	housing	loan
	<*lgl*>	<*int*>	<*int*>	<*lgl*>	<*lgl*>
29183	FALSE	543	275	TRUE	FALSE

The diagnostics code also produced the trace plot, which can be found in Section 7.3.3. The code also produced boxplots of the distributions of `balance` and `previous` by outcome from the first imputed dataset, as the aforementioned observation had very high values for these two variables. The boxplots shows that, while the observation did not have a particularly high or low value for `balance`, it did have an extremely high value for `previous`.

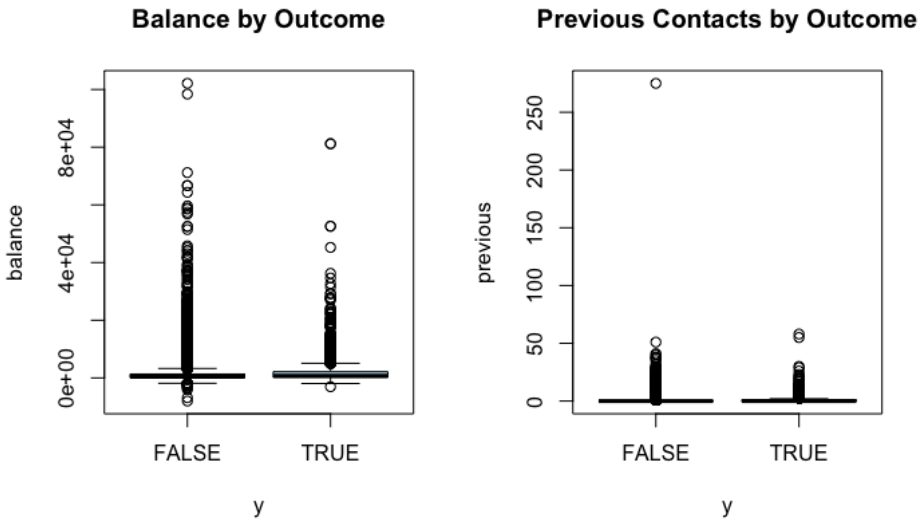

Figure 8: Boxplots of Distributions of Balance and Previous Contacts by Outcome