Homework 4

Denis Ostroushko

Problem 1

(a)

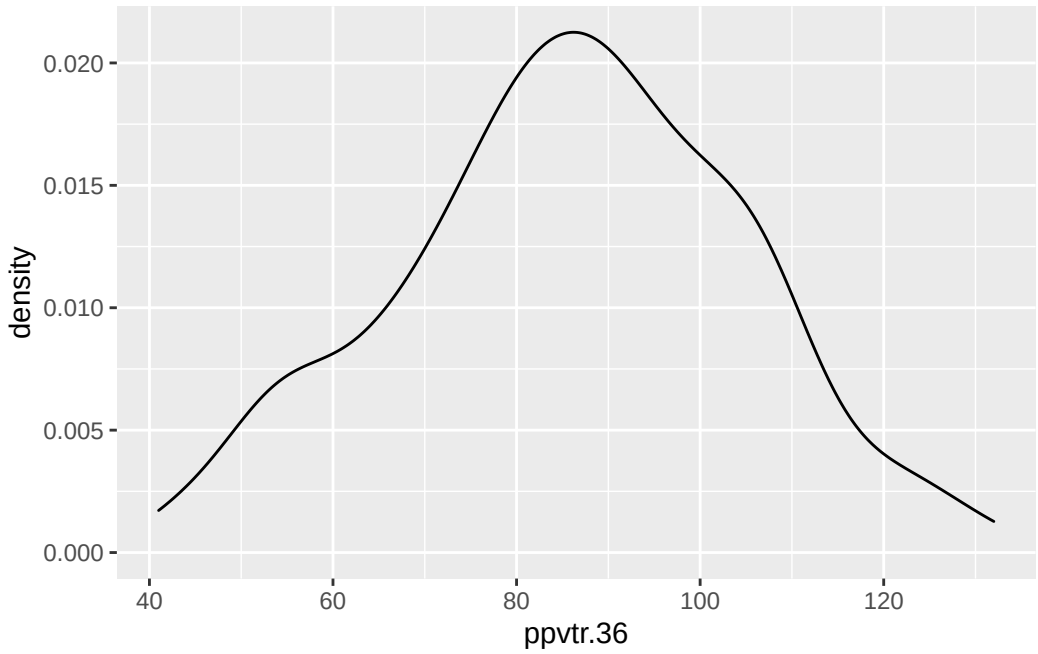
Chosing models

Before diving into the multiple imputation I decided to study the distribution of observed variables. While we covered that the distribution of observed vales does not give accurate information about the true distribution of a random variable. However, knowing the shape of observed distribution might be able to suggest a proper model for the imputaton.

ppvtr.36 distribution

We begin by studying the shape of a response variable in this study. We can use a Bayesian Normal Linear model for the inputaion task here.

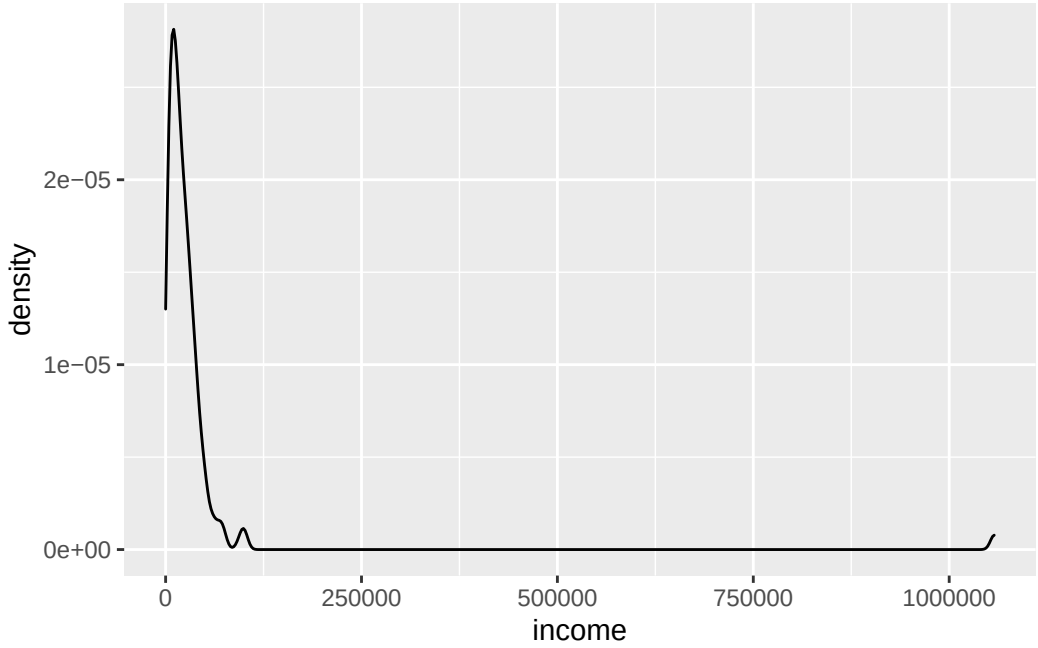
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  41.00   74.00   87.00   85.94   99.00  132.00      75
```



income

The distribution of income is heavily skewed. Perhaps, it would make more sense to use a log-transformation of the variable to achieve more accurate predictions, however, this will affect the estimates we eventually produce. For the purpose of being consistent with assignment expectations, use income on the original scale. Again, we will use a Bayesian model.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0    8590   17906   32041   31228 1057448      82
```



Variables `b.marr` and `momed` will be imputed using a logistic regression and proportional odds models respectively.

Multiple imputation

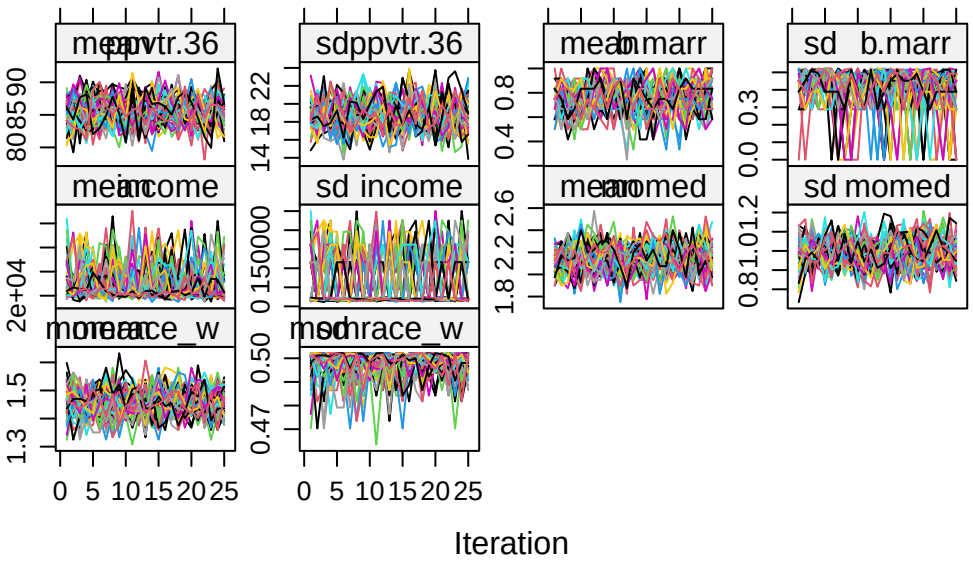
Code below provides my approach to multiple imputation problem. I use 50 data sets for pooling of results and require 25 iterations before a given data set results are considered converged.

```
# list of imputation models
# https://www.rdocumentation.org/packages/mice/versions/3.15.0/topics/mice

imp_1 <- mice(nlsyV,
              method = c("norm", "logreg", "logreg", "pmm","pmm", "polr", "logreg"),
              maxit = 25,
              m = 50,
              printFlag = F)
```

(b)

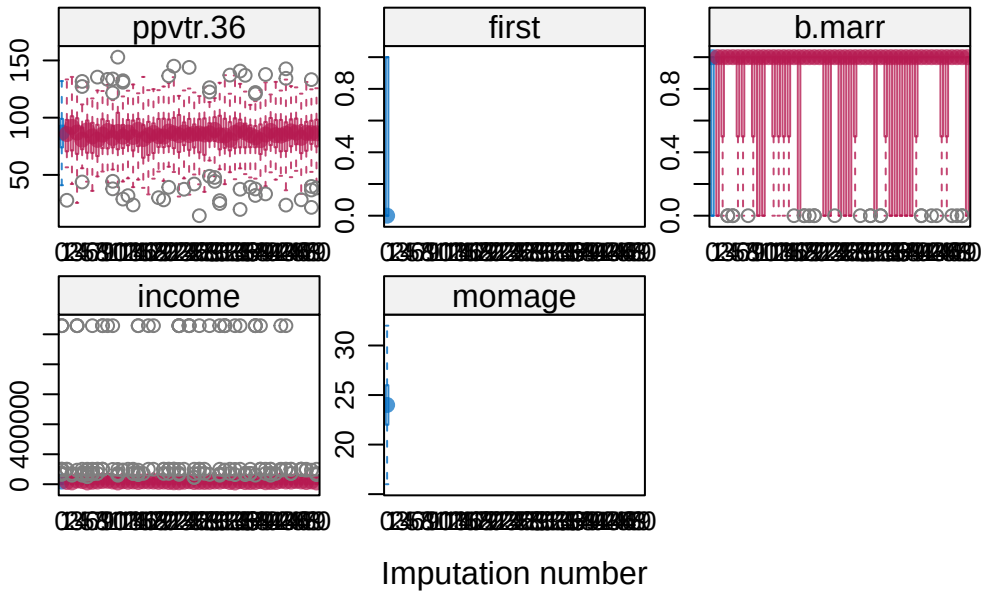
We begin evaluation of our imputations by looking over at the mean and standard deviation of each imputed variable over the course of all 50 imputations.



We are looking for evidence or absence of these things:

1. Trends in mean or standard deviation. It is clear here that there is no concrete upward or downward trend in the summary statistics. This is an indication that a set of imputations is appropriate
2. Constancy of variance in summary statistics. While there are some spikes, or outliers in imputations, especially in the income variable, overall, I see no obvious problems with the imputations.
3. The two point above make me feel convinced that the imputation process converged.

Additionally we will look at the diagnostic of distribution of imputed variables for each of 50 iterations.



Again, there are clear outliers with the income variable, but the we saw there this distribution is prone to very extreme values, so we are expecting this behavior. Overall, there is no reason to believe that the imputation process has some obvious flaws.

We can use these imputations for the inference on the data.

(c)

Code below provides pooled results for the regression model:

```
options(scipen = 999)

fit_mi <- with(imp_1, lm(ppvtr.36 ~ first + b.marr + income + momage + momed + momrace_w))

sum <- summary(pool(fit_mi))

sum[,2:length(sum)] <- round(sum[,2:length(sum)], 2)

sum
```

```
          term estimate std.error statistic     df p.value
1  (Intercept)    71.00      7.82      9.08 203.50    0.00
2        first     3.90      1.90      2.06 241.23    0.04
3      b.marr     3.30      2.37      1.39 134.33    0.17
4      income     0.00      0.00      0.61 284.00    0.54
```

```
5          momage    -0.10      0.33     -0.31 210.83    0.76
6          momed2     5.09      2.21      2.30 216.37    0.02
7          momed3     9.92      2.54      3.90 251.81    0.00
8          momed4    15.39      4.29      3.59 204.32    0.00
9 momrace_wwhite    15.36      2.17      7.07 126.58    0.00
```

`first` is an indicator that a child is a first born. Estimated effect is 4.08, meaning that first born children on average score 4.08 points higher on the Peabody Picture Vocabulary Test, when compared with second and later born children, after adjusting for other variables.

P-value is less than 0.05, indicating that the effect is statistically significant.

Confidence interval for the effect is (0.09, 8.07). We used a 97.5th percentile of a T-distribution with 176.96 degrees of freedom to construct this confidence interval.

Problem 2

In the simulation code provided, it appears that the true relationship between Y and predctors X_i foir i = 1,2,3,4,5 is given by this equation:

$$Y = 2 + X_1 + X_2 + X_3 + 2 * X_4 + 2 * X_5 + \epsilon$$

where ϵ is a random error, normally distributed with mean 0 and variance σ^2.

(a)

Based on the description of the model above, we will use $\beta_1 = 1$ as the true value, and calculate bias using this value.

To calculate average bias for each method by first calculating bias for each of 100 iterations and then averaging over the difference. I am including code below and some values of estimates and biases before making final conclusions.

```r
estiamtes %>%
  mutate_all(~. - 1) %>%
  summarise(across(everything(), mean)) %>%
  t() %>%
  round(.,4) -> r

colnames(r) <- "Average Bias"

r
```

```
         Average Bias
Method.a      -0.0094
Method.b      -0.0882
Method.c      -0.2154
Method.d       0.0021
Method.e      -0.0077
```

It appears that method D, a Single imputation with error, has the lowest bias. Although, I would stay that methods A, D, E have the lowest bias.

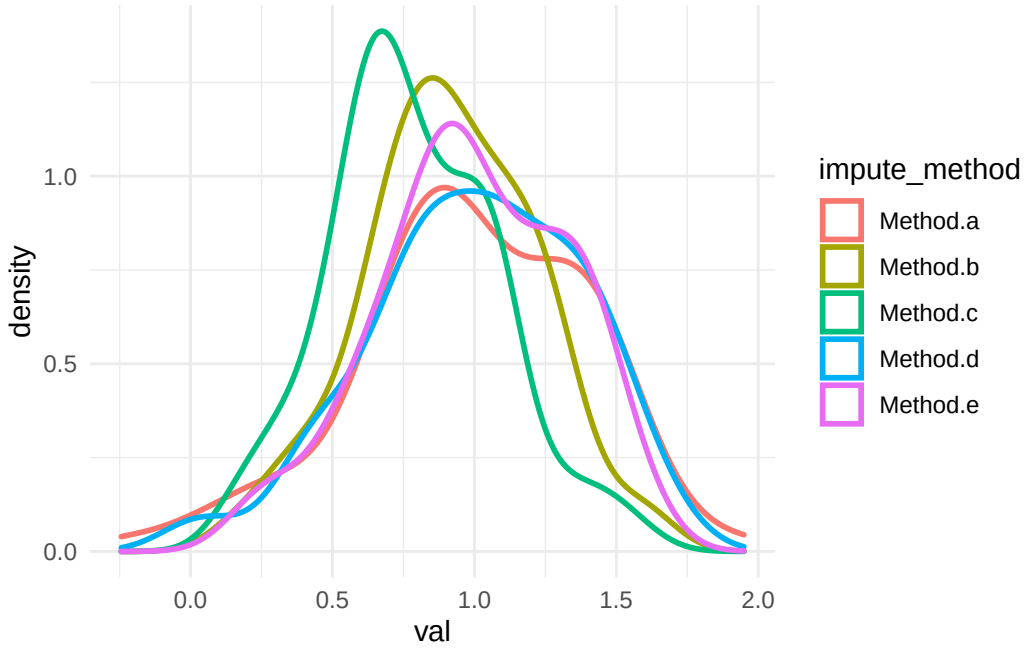
We covered that for complete case analysis the bias is close to zero if data missingness is independent of Y or other data. However, this is not the case here, and what we observed is due to chance.

Single and Multiple imputation with error are both fairly unbiased.

Imputing predicted value is a biased estimate because even if the model specification is correct, we do not include any error structure into the data, so we produce estimates that are highly biased.

(b)

Before looking at the standard error of estimates, we can visually examine the distributions.



I would say that all these look approximately normal. All distributions will have *similar* standard errors, however, method C, will probably have the smallest variance, even though we know method C is a flawed method.

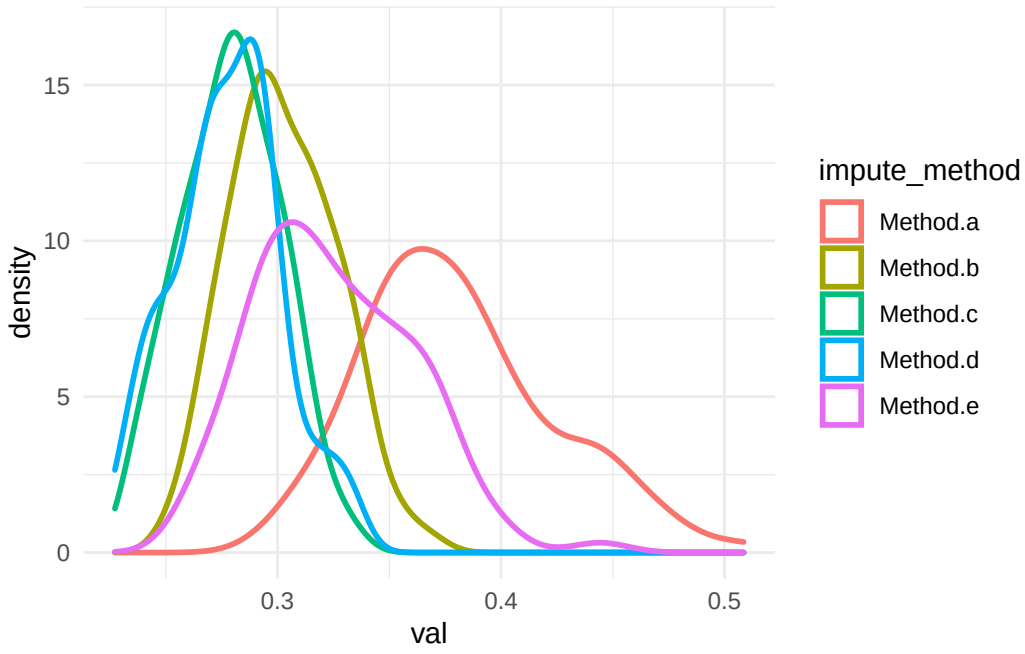
```
           [,1]
Method.a 0.4068
Method.b 0.3039
Method.c 0.2946
Method.d 0.3712
Method.e 0.3302
```

```
         Standard Deviation of Betas
Method.a                     -0.0094
Method.b                     -0.0882
Method.c                     -0.2154
Method.d                      0.0021
Method.e                     -0.0077
```

It appears that method C does have the smallest variance in the sampling distribution of method C. However, among *closely unbiased* estimates, method E, where we impute multiple times, has the smallest standard error. This is quite reassuring because we know that multiple imputation is the best way to handle missing data.

(c)

Before looking at the standard error of estimates, we can visually examine the distributions.



	Average Standard Error
Method.a	0.3809
Method.b	0.3025
Method.c	0.2792
Method.d	0.2774
Method.e	0.3248

Results obtained here are similar to those in part (b). We would like to see that the average SE and the SE of obtained Betas from multiple iterations agree, i.e. the two values are pretty close to each other. Otherwise, if we see a larger discrepancy, it would indicate that something is not right with the sampling distribution.

All methods except for D have approximately matching values. A single imputation with error probably have these mismatching results due to a single error. When we draw just one error, sometimes the value of such error can be too extreme, which then creates an extreme values that skew regression estimates.

Multiple imputation and many draws masks the effect of a single outlier.

(d)

Average squared errors are given below:

```r
estiamtes %>%
  mutate_all(~(.-1)^2) %>%
  summarise(across(everything(), mean))  %>%
  t() %>%
  round(.,4) -> r

colnames(r) <- "Average Squared Error"

r
```

```
         Average Squared Error
Method.a                0.1639
Method.b                0.0992
Method.c                0.1323
Method.d                0.1364
Method.e                0.1080
```

Again, among the methods that produce unbiased estimates, method E has the smallest standard error for the regression coefficient.

(e)

My overall conclusion is that we should always use multiple imputation process with some error structure. While is produces results that are similar to other methods, it seems intuitive to me that using more data and then averaging over results with some randomness should produce more reliable and robust results, at the cost of computing power.

Method with a simple prediction imputation was the most problematic. Average bias of the estimate was quite high, which is due to the fact that we did not have any error structure embedded in the model.

Complete case analysis showing the same performance as the multiple and single imputation methods is quite strange to me, I think this is simply a coincidence, I did not see a reason as to why this would be the case in Jared's code.

In conclusion, we saw that:

1. Using predicted values for imputation produces biased regression estimates, which affects the validity of inference
2. Using single imputation with error is not optimal because of the effect of a single outlier drawn can affect the results
3. Imputing mean value produces artificially low standard error for the estimated beta

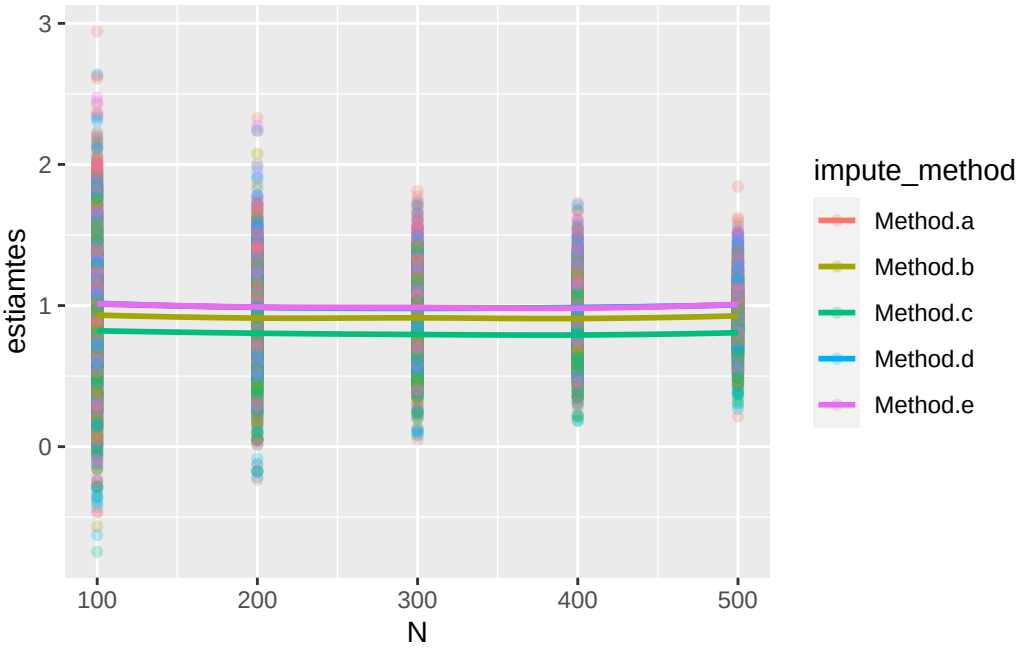
(f)

For the extra credit problem, I modified Jared's code to add more looping into the simulation code. I considered Sample size of 100, 200, … , 500, as well as modifying the intercept for probability of missingness indicator. I allowed the intercept to range from 0.25 to 2, in 0.25 increments.

Sample Size Effect

First, I plot Beta estimates against the sample size, and color these Beta's by their respective imputation method. While it appears that the variance of sampling distributions decreases as N increases, no method of imputation seems to be affected by the sample size in terms of the average estimated Beta.

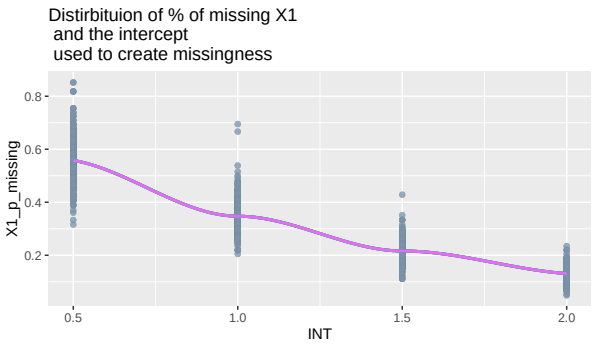
We can see that method C, which had large bias, retains Beta that is consistently lower than the true value over all considered sample sizes N.

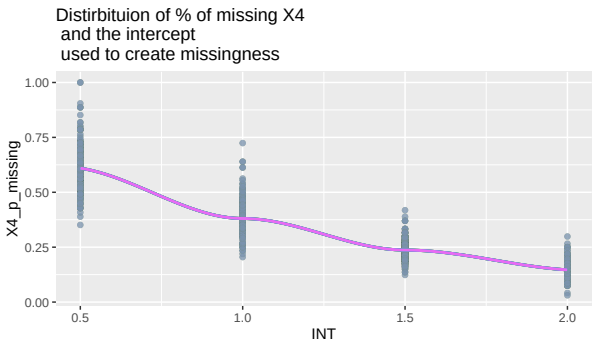


Missingness Amount Effect

Next we consider the plot of intercept against estimated Betas, colored by the respective imputation method again.

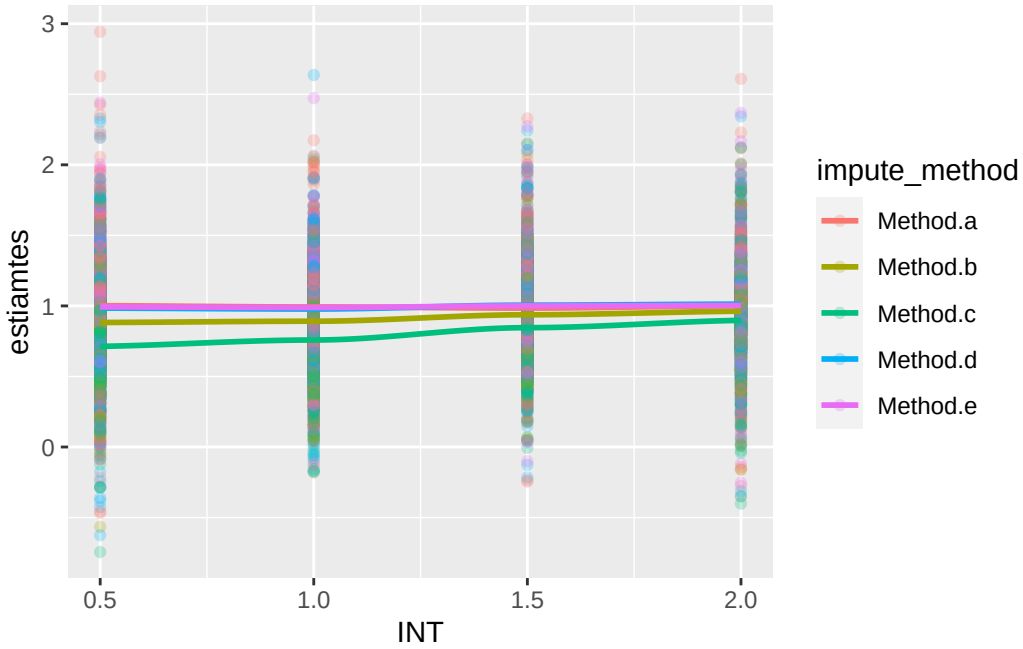
To help interpret these results, here are the two plots that show how increase in the intercept relates to the % of X_1 and X_4 are missing.





Clearly, as the intercept increases, we reduce the % of missing values for the two covariates that we impute

Now, we can look at the plot of intercepts considered and estimated betas.



It appears that the methods most affected by the missing values is method C, which has been the most problematic so far. As the proportion of missing values in X1 and X4 decreases, average bias of beta decreases as well. Method B seems to have similar behaviors, although not as strong.

Again, complete case analysis exhibits similar behavior to single and multiple imputation methods, which I did not expect.

It is, again, reassuring to see that imputation methods with error structures have very robust behavior regardless of variation in sample size and missingness levels.