Imputation(add new variables)

Xinyi Lin

9/29/2019

## Import data

Import data “R1\_R18\_partcleaned\_20190913.dta”.

## Creat the variable

Notation:

Variable name: student\_r

Created from: occup1\_r, occup2\_r

Label: student\_r — yes-is a student, no-not a student, NA-occup1\_r and occup2\_r are NAs

## Analysis missing data

### Select variables and tidy data

Predictors including: ageyrs\_r, sex, student\_r, area, educate\_r, currmarr\_r, sexp1yr, SEScat, study\_id, pregnow\_r, numchild\_r, eversex\_r, currrltn.

First, check each variable.

## student\_r visit ageyrs\_r sex   
## no :21824 18 : 3596 Min. :13.00 Length:47699   
## yes :19669 17 : 3407 1st Qu.:16.00 Class :character   
## NA's: 6206 16 : 3187 Median :17.00 Mode :character   
## 14 : 3126 Mean :17.17   
## 6 : 2972 3rd Qu.:18.00   
## 3 : 2939 Max. :19.00   
## (Other):28472   
## area educate\_r currmarr\_r sexp1yr   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. : 0.00   
## 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.: 1.00   
## Median :0.0000 Median :1.0000 Median :0.0000 Median : 2.00   
## Mean :0.2955 Mean :0.9775 Mean :0.1954 Mean :42.69   
## 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:98.00   
## Max. :2.0000 Max. :1.0000 Max. :1.0000 Max. :99.00   
## NA's :8 NA's :4 NA's :1391   
## SEScat study\_id pregnow\_r numchild\_r   
## Min. :0.000 Length:47699 Min. :0.000 Min. :0.00   
## 1st Qu.:1.000 Class :character 1st Qu.:0.000 1st Qu.:0.00   
## Median :2.000 Mode :character Median :0.000 Median :1.00   
## Mean :1.644 Mean :0.126 Mean :0.78   
## 3rd Qu.:3.000 3rd Qu.:0.000 3rd Qu.:1.00   
## Max. :3.000 Max. :1.000 Max. :4.00   
## NA's :208 NA's :22586 NA's :43771   
## eversex\_r currrltn   
## Min. :0.0000 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:2.000   
## Median :1.0000 Median :2.000   
## Mean :0.5849 Mean :1.783   
## 3rd Qu.:1.0000 3rd Qu.:2.000   
## Max. :1.0000 Max. :9.000   
## NA's :347

For sexp1yr:

## 0 1 2 3 4 5 6 7 8 9 10 11   
## 3840 17949 3162 865 276 118 79 54 29 12 25 2   
## 12 13 15 16 17 18 20 21 22 24 30 93   
## 6 2 12 2 2 2 5 1 1 1 2 33   
## 97 98 99 NA's   
## 1 19826 1 1391

93 means a lot, 97 means don’t know, 98 means not applicable and 99 means not response. 97, 98, 99 can be marked as NA.

For pregnow\_r:

## # A tibble: 4 x 3  
## # Groups: sex [2]  
## sex pregnow\_r n  
## <chr> <dbl+lbl> <int>  
## 1 female 0 [no] 21952  
## 2 female 1 [yes] 3161  
## 3 female NA 662  
## 4 male NA 21924

For male, pregnow\_r can be marked as 98. And pregnow\_r should be catagorical variable.

For currrltn:

## 0 1 2 7 8 9   
## 54 10415 37202 12 5 11

7 means don’t know. *0, 8, 9 can be marked as missing value(not sure).*

As there are too much missing value in numchild\_r, and too much categories in study\_id it is not suitable to be a predictor.

## student\_r visit ageyrs\_r sex area   
## no :21824 18 : 3596 Min. :13.00 female:25775 0:36220   
## yes :19669 17 : 3407 1st Qu.:16.00 male :21924 1: 8861   
## NA's: 6206 16 : 3187 Median :17.00 2: 2618   
## 14 : 3126 Mean :17.17   
## 6 : 2972 3rd Qu.:18.00   
## 3 : 2939 Max. :19.00   
## (Other):28472   
## educate\_r currmarr\_r sexp1yr SEScat pregnow\_r   
## 0 : 1074 0 :38374 Min. : 0.000 0 : 9894 0 :21952   
## 1 :46617 1 : 9321 1st Qu.: 1.000 1 :10996 1 : 3161   
## NA's: 8 NA's: 4 Median : 1.000 2 :12704 98 :21924   
## Mean : 1.273 3 :13897 NA's: 662   
## 3rd Qu.: 1.000 NA's: 208   
## Max. :93.000   
## NA's :21219   
## eversex\_r currrltn   
## 0 :19657 1 :10415   
## 1 :27695 2 :37202   
## NA's: 347 7 : 12   
## NA's: 70   
##   
##   
##

### Permutation

Resample each variable separately to test their contribution to imputation. Average misclassification rates are shown below. We can find that eversex\_r(Have you ever had a sexual relationship?) and currrltn(Do you currently have a sexual relationship with someone to whom you are not officially married or in a consensual union?) contribute most to permutation process.

## ageyrs\_r sex area educate\_r currmarr\_r sexp1yr   
## 0.2446691 0.2472702 0.2458968 0.2506929 0.2579233 0.2632014   
## SEScat pregnow\_r eversex\_r currrltn   
## 0.2718768 0.2724077 0.4035861 0.4262156

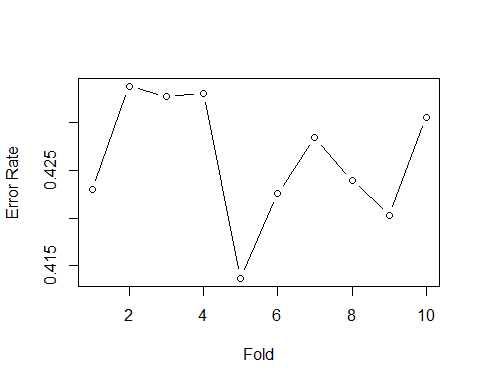
### Test data

First, choose observations with known student\_r as test dataset. There are 41493 observations in total.

Then, randomly split test dataset into ten subsets and set them as missing values to test error rates of each fold when using different methods.

For misclassification rate:

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.4137 0.4227 0.4262 0.4262 0.4322 0.4338



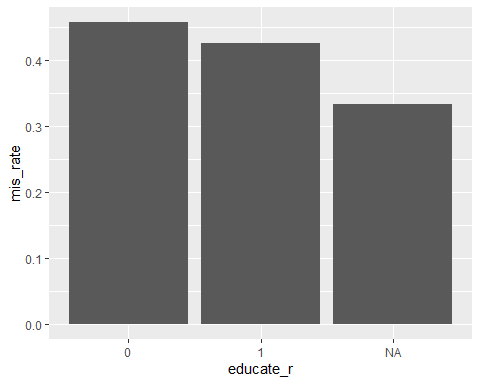
For the misclassification dataset:

There are 41493 observations are misclassified. Marginal distribution of misclassification rate in different gourps are significant among following variables: educate\_r， currmarr\_r, sexp1yr， pregnow\_r.

For educate\_r: **significant difference**

## Warning: Factor `educate\_r` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

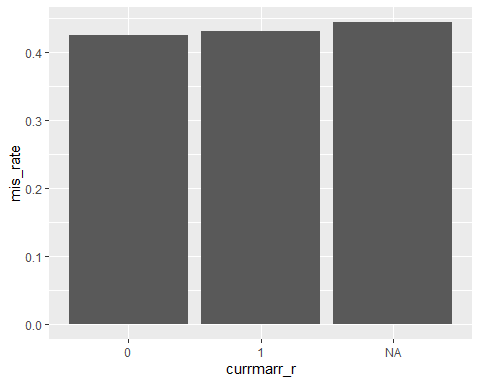
|  |  |  |
| --- | --- | --- |
| educate\_r | mis\_rate | sum |
| 0 | 0.4578454 | 854 |
| 1 | 0.4255900 | 40633 |
| NA | 0.3333333 | 6 |



For currmarr\_r: **significant difference**

## Warning: Factor `currmarr\_r` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

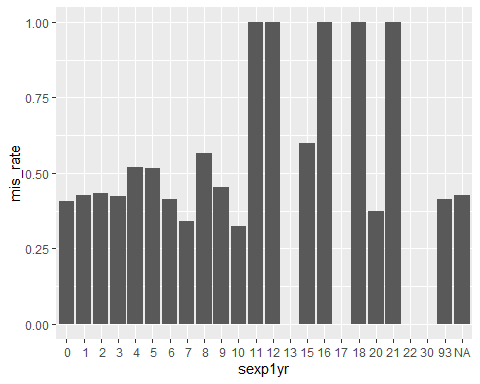
|  |  |  |
| --- | --- | --- |
| currmarr\_r | mis\_rate | sum |
| 0 | 0.4250194 | 33582 |
| 1 | 0.4314098 | 7902 |
| NA | 0.4444444 | 9 |



For sexp1yr: **significant difference**

## Warning: Factor `sexp1yr` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

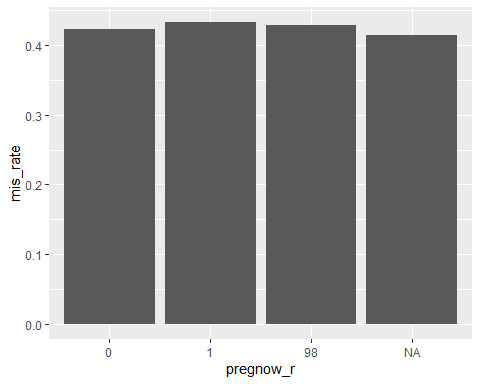
|  |  |  |
| --- | --- | --- |
| sexp1yr | mis\_rate | sum |
| 0 | 0.4074778 | 3156 |
| 1 | 0.4261734 | 15787 |
| 2 | 0.4337266 | 2882 |
| 3 | 0.4218959 | 749 |
| 4 | 0.5200000 | 275 |
| 5 | 0.5157895 | 95 |
| 6 | 0.4133333 | 75 |
| 7 | 0.3409091 | 44 |
| 8 | 0.5652174 | 23 |
| 9 | 0.4545455 | 11 |
| 10 | 0.3225806 | 31 |
| 11 | 1.0000000 | 2 |
| 12 | 1.0000000 | 1 |
| 13 | 0.0000000 | 2 |
| 15 | 0.6000000 | 10 |
| 16 | 1.0000000 | 1 |
| 17 | 0.0000000 | 2 |
| 18 | 1.0000000 | 2 |
| 20 | 0.3750000 | 8 |
| 21 | 1.0000000 | 1 |
| 22 | 0.0000000 | 1 |
| 30 | 0.0000000 | 1 |
| 93 | 0.4137931 | 29 |
| NA | 0.4267686 | 18305 |



For pregnow\_r: **significant difference**

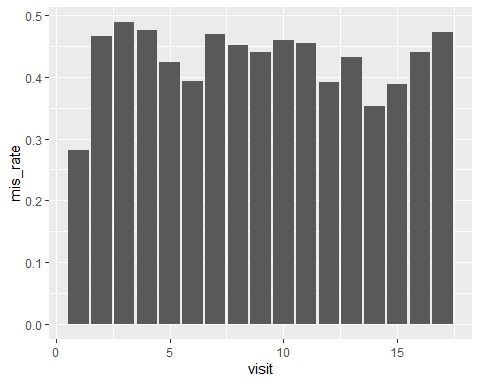
## Warning: Factor `pregnow\_r` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

|  |  |  |
| --- | --- | --- |
| pregnow\_r | mis\_rate | sum |
| 0 | 0.4229225 | 19422 |
| 1 | 0.4329897 | 2716 |
| 98 | 0.4290737 | 18773 |
| NA | 0.4140893 | 582 |



For visit:

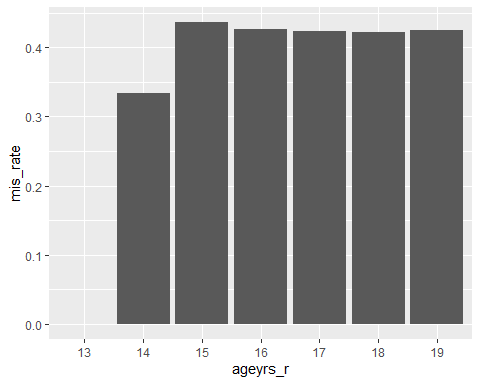
|  |  |  |
| --- | --- | --- |
| visit | mis\_rate | sum |
| 1 | 0.2816011 | 2848 |
| 2 | 0.4659989 | 1897 |
| 3 | 0.4888603 | 2334 |
| 4 | 0.4765343 | 2770 |
| 5 | 0.4238247 | 2829 |
| 6 | 0.3925144 | 3126 |
| 7 | 0.4689843 | 2934 |
| 8 | 0.4517965 | 2811 |
| 9 | 0.4403541 | 2146 |
| 10 | 0.4602336 | 3596 |
| 11 | 0.4548043 | 1405 |
| 12 | 0.3911765 | 1360 |
| 13 | 0.4314721 | 985 |
| 14 | 0.3526245 | 2972 |
| 15 | 0.3880831 | 2551 |
| 16 | 0.4394682 | 2858 |
| 17 | 0.4717528 | 2071 |



For ageyrs\_r:

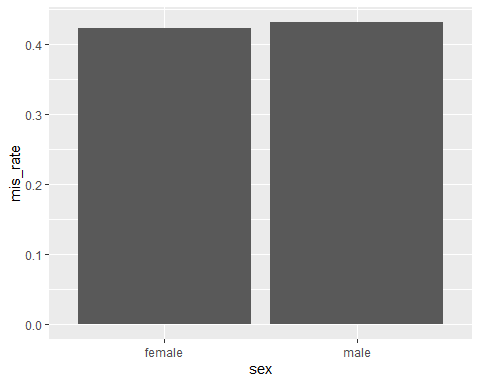
|  |  |  |
| --- | --- | --- |
| ageyrs\_r | mis\_rate | sum |
| 13 | NA | NA |
| 14 | 0.3333333 | 3 |
| 15 | 0.4365387 | 7477 |
| 16 | 0.4262731 | 8267 |
| 17 | 0.4235294 | 7735 |
| 18 | 0.4213382 | 9401 |
| 19 | 0.4250203 | 8609 |

## Warning: Removed 1 rows containing missing values (position\_stack).



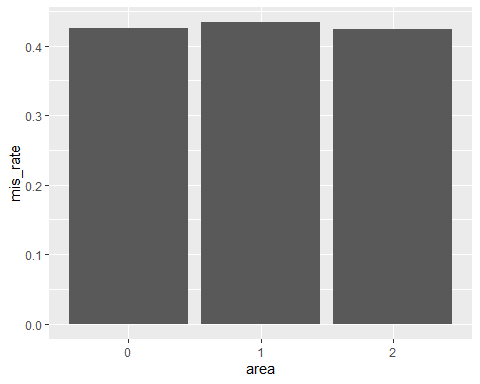
For sex:

|  |  |  |
| --- | --- | --- |
| sex | mis\_rate | sum |
| female | 0.4221810 | 22623 |
| male | 0.4311076 | 18870 |



For area:

|  |  |  |
| --- | --- | --- |
| area | mis\_rate | sum |
| 0 | 0.4245455 | 31297 |
| 1 | 0.4339382 | 7826 |
| 2 | 0.4232068 | 2370 |



For SEScat:

## Warning: Factor `SEScat` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

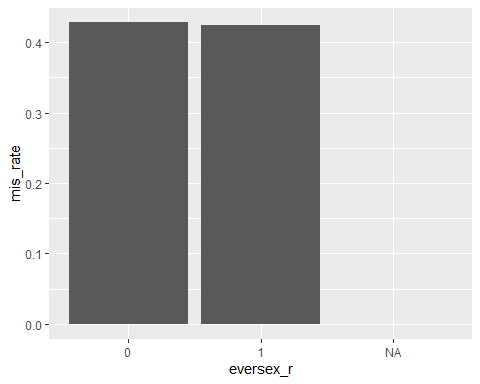
|  |  |  |
| --- | --- | --- |
| SEScat | mis\_rate | sum |
| 0 | 0.4245623 | 8053 |
| 1 | 0.4311106 | 9283 |
| 2 | 0.4296361 | 11405 |
| 3 | 0.4203301 | 12602 |
| NA | 0.4533333 | 150 |



For eversex:

## Warning: Factor `eversex\_r` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

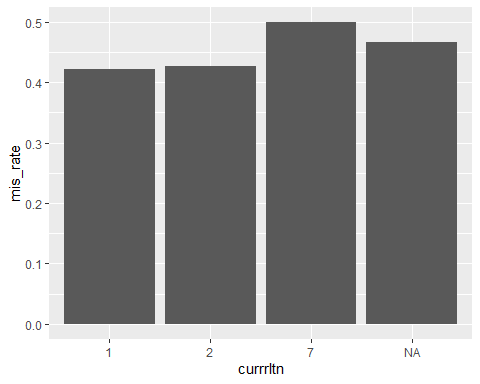
|  |  |  |
| --- | --- | --- |
| eversex\_r | mis\_rate | sum |
| 0 | 0.4288192 | 17877 |
| 1 | 0.4243246 | 23614 |
| NA | 0.0000000 | 2 |



For currrltn:

## Warning: Factor `currrltn` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

|  |  |  |
| --- | --- | --- |
| currrltn | mis\_rate | sum |
| 1 | 0.4226965 | 8997 |
| 2 | 0.4271722 | 32467 |
| 7 | 0.5000000 | 14 |
| NA | 0.4666667 | 15 |



knitr::opts\_chunk$set(echo = FALSE)  
library(tidyverse)  
library(missForest)  
library(ggplot2)  
library(caret)  
library(haven)  
new\_data <- read\_dta("R1\_R18\_partcleaned\_20190913.dta")  
student\_data = new\_data %>%  
 as.data.frame() %>%   
 mutate(student\_r = ifelse(occup1\_r == 8 | occup2\_r == 8, "yes", "no"),  
 student\_r = ifelse(is.na(student\_r) & occup1\_r == 20, "no", student\_r))  
#head(student\_data)  
target\_data = student\_data %>%   
 filter(ageyrs\_r <= 19) %>%   
 mutate(student\_r = as.factor(student\_r),  
 visit = as.factor(visit),  
 sex = ifelse(female==1, "female", "male")) %>%   
 select(student\_r, visit, ageyrs\_r, sex, ageyrs\_r, sex, student\_r, area, educate\_r, currmarr\_r, sexp1yr, SEScat, study\_id, pregnow\_r, numchild\_r, eversex\_r, currrltn)  
summary(target\_data)  
summary(as.factor(target\_data$sexp1yr))  
target\_data %>%   
 group\_by(sex, pregnow\_r) %>%   
 dplyr::summarize(n = n())  
summary(as.factor(target\_data$currrltn))  
impu\_data = target\_data %>%   
 mutate(sexp1yr = ifelse(sexp1yr %in% c(97, 98, 99), NA, sexp1yr)) %>%   
 mutate(pregnow\_r = ifelse(sex=="male", 98, pregnow\_r),  
 pregnow\_r = as.factor(pregnow\_r)) %>%   
 mutate(currrltn = ifelse(currrltn %in% c(0, 8,9), NA, currrltn),  
 currrltn = as.factor(currrltn)) %>%   
 select(-c(numchild\_r, study\_id)) %>%   
 mutate(visit = as.factor(visit),  
 area = as.factor(area),  
 educate\_r = as.factor(educate\_r),  
 currmarr\_r = as.factor(currmarr\_r),  
 SEScat = as.factor(SEScat),  
 eversex\_r = as.factor(eversex\_r),  
 sex = as.factor(sex))  
summary(impu\_data)  
test\_data = impu\_data[!is.na(impu\_data$student\_r),]  
set.seed(123)  
var\_mis <- matrix(data=NA,nrow=10,ncol=10) # 10 variables and 10 folds for each variables  
for (j in 3:12) {  
 flds <- createFolds(1:nrow(test\_data), k = 10, list = TRUE, returnTrain = FALSE)  
 target\_variable = test\_data[,j]  
 new\_variable = sample(target\_variable, nrow(test\_data), replace = T)  
 test\_data[,j] = new\_variable  
 test\_data = as.data.frame(as.matrix(test\_data))  
 for (n in 1:10) {  
 na\_data = test\_data  
 na\_data[flds[[n]], 1] = NA  
 # missForest  
 impu\_forest = missForest(na\_data)  
 impu\_forest\_df = impu\_forest$ximp  
 # calculate error rates  
 # yes = 2, no = 1  
 var\_mis[j-2,n] = sum(abs(as.numeric(impu\_forest\_df[flds[[n]], 1]) - as.numeric(test\_data[flds[[n]], 1])))/length(test\_data[flds[[n]], 1])  
 }  
}  
colnames(var\_mis) = c("fold1", "fold2", "fold3", "fold4", "fold5", "fold6", "fold7", "fold8", "fold9", "fold10")  
rownames(var\_mis) = c("ageyrs\_r","sex","area","educate\_r","currmarr\_r","sexp1yr","SEScat","pregnow\_r","eversex\_r","currrltn")  
apply(var\_mis,1,mean)  
flds <- createFolds(1:nrow(test\_data), k = 10, list = TRUE, returnTrain = FALSE)  
#flds  
# error\_rate = vector("list", 10)  
set.seed(123)  
test\_data = as.data.frame(as.matrix(test\_data))  
error\_forest = rep(NA, 10) # indicate misclassification rate  
error\_index = rep(0,nrow(test\_data)) # indcate misclassification rows  
for (n in 1:10){  
 na\_data = test\_data  
 na\_data[flds[[n]], 1] = NA  
 # missForest  
 impu\_forest = missForest(na\_data)  
 impu\_forest\_df = impu\_forest$ximp  
 # calculate error rates  
 # yes = 2, no = 1  
 error\_forest[n] = sum(abs(as.numeric(impu\_forest\_df[flds[[n]], 1]) - as.numeric(test\_data[flds[[n]], 1])))/length(test\_data[flds[[n]], 1])  
 error\_vector = abs(as.numeric(impu\_forest\_df[, 1]) - as.numeric(test\_data[, 1]))  
 error\_index = error\_index + error\_vector  
}  
summary(error\_forest)  
plot(1:10,error\_forest,xlab="Fold",ylab="Error Rate",type="b",lty=1)  
error\_data = test\_data %>%   
 mutate(miscla = error\_index,  
 miscla = as.factor(miscla))  
educate\_error = error\_data %>%   
 group\_by(miscla, educate\_r) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(educate\_r, mis\_rate, sum)   
educate\_error %>% knitr::kable()  
educate\_error %>%   
 ggplot(aes(educate\_r, mis\_rate)) +  
 geom\_bar(stat = "identity")  
currmarr\_error = error\_data %>%   
 group\_by(miscla, currmarr\_r) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(miscla\_1 = ifelse(is.na(miscla\_1), 0, miscla\_1)) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(currmarr\_r, mis\_rate, sum)   
currmarr\_error %>% knitr::kable()  
currmarr\_error %>%   
 ggplot(aes(currmarr\_r, mis\_rate)) +  
 geom\_bar(stat = "identity")  
sexp1yr\_error = error\_data %>%   
 group\_by(miscla, sexp1yr) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(miscla\_1 = ifelse(is.na(miscla\_1), 0, miscla\_1),  
 miscla\_0 = ifelse(is.na(miscla\_0), 0, miscla\_0)) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(sexp1yr, mis\_rate, sum)   
sexp1yr\_error %>% knitr::kable()  
sexp1yr\_error %>%   
 ggplot(aes(sexp1yr, mis\_rate)) +  
 geom\_bar(stat = "identity")  
pregnow\_error = error\_data %>%   
 group\_by(miscla, pregnow\_r) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(pregnow\_r, mis\_rate, sum)   
pregnow\_error %>% knitr::kable()  
pregnow\_error %>%   
 ggplot(aes(pregnow\_r, mis\_rate)) +  
 geom\_bar(stat = "identity")  
visit\_error = error\_data %>%   
 mutate(visit = as.numeric(visit)) %>%   
 group\_by(miscla, visit) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(visit, mis\_rate, sum)   
visit\_error %>% knitr::kable()  
visit\_error %>%   
 ggplot(aes(visit, mis\_rate)) +  
 geom\_bar(stat = "identity")  
age\_error = error\_data %>%   
 group\_by(miscla, ageyrs\_r) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(miscla\_1 = ifelse(is.na(miscla\_1), 0, miscla\_1)) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(ageyrs\_r, mis\_rate, sum)   
age\_error %>% knitr::kable()  
age\_error %>%   
 ggplot(aes(ageyrs\_r, mis\_rate)) +  
 geom\_bar(stat = "identity")  
sex\_error = error\_data %>%   
 group\_by(miscla, sex) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(sex, mis\_rate, sum)   
sex\_error %>% knitr::kable()  
sex\_error %>%   
 ggplot(aes(sex, mis\_rate)) +  
 geom\_bar(stat = "identity")  
area\_error = error\_data %>%   
 group\_by(miscla, area) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(area, mis\_rate, sum)   
area\_error %>% knitr::kable()  
area\_error %>%   
 ggplot(aes(area, mis\_rate)) +  
 geom\_bar(stat = "identity")  
SEScat\_error = error\_data %>%   
 group\_by(miscla, SEScat) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(SEScat, mis\_rate, sum)   
SEScat\_error %>% knitr::kable()  
SEScat\_error %>%   
 ggplot(aes(SEScat, mis\_rate)) +  
 geom\_bar(stat = "identity")  
eversex\_error = error\_data %>%   
 group\_by(miscla, eversex\_r) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(miscla\_1 = ifelse(is.na(miscla\_1), 0, miscla\_1)) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(eversex\_r, mis\_rate, sum)   
eversex\_error %>% knitr::kable()  
eversex\_error %>%   
 ggplot(aes(eversex\_r, mis\_rate)) +  
 geom\_bar(stat = "identity")  
currrltn\_error = error\_data %>%   
 group\_by(miscla, currrltn) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(currrltn, mis\_rate, sum)   
currrltn\_error %>% knitr::kable()  
currrltn\_error %>%   
 ggplot(aes(currrltn, mis\_rate)) +  
 geom\_bar(stat = "identity")