BIST 5092 Phase III Project

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(stringr)
library(ggplot2)
```

import ADPA and ADSL datasets, select measurements at visit 6

```
adpa <- read.csv("../data_phase 3/ADPA.csv")
pasi <- adpa %>% filter(PARAMCD == "PASI")
pasi6 <- adpa %>% filter(PARAMCD == "PASI", AVISIT == "VISIT 6")
adsl <- read.csv("../data_phase 3/ADSL.csv")
adae <- read.csv("../data_phase 3/ADAE.csv")
```

perform logistic regression to assess the treatment effect on PASI75 adjusted by sex with the full dataset

```
dat_full <- merge(adsl, pasi6, by = "SUBJID")
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat_full, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat_full)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0185  -0.4241   0.5286   0.8783   2.2944
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.5574     0.2248 -11.377   <2e-16 ***
## SEXM                 0.1940     0.1239   1.566    0.117
## as.factor(TRTPN)2    3.2724     0.2438  13.421   <2e-16 ***
## as.factor(TRTPN)3    3.1169     0.2257  13.811   <2e-16 ***
## as.factor(TRTPN)4    4.2609     0.2395  17.791   <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: 2396.1  on 1830  degrees of freedom
## Residual deviance: 1804.7  on 1826  degrees of freedom
## AIC: 1814.7
##
## Number of Fisher Scoring iterations: 5
```

MCAR

create functions to generate dataset with certain percent missing

```
generate_missing <- function(data, percent){
  id <- sample(1:nrow(data), round(nrow(data)*percent), replace = FALSE)
  data_par1 <- data[id,]
  data_par1$AVAL <- NA
  data_par2 <- data[-id,]
  data_final <- rbind(data_par1, data_par2)
  data_final$PCHGCA1N[which(is.na(data_final$AVAL))] <- NA

  return(data_final)
}
```

1. generate dataset with 10% missing and perform logistic regression

```
set.seed(1234)
pasi6_final1 <- generate_missing(pasi6, 0.1)
dat1 <- merge(adsl, pasi6_final1, by = "SUBJID")
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat1, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0105  -0.4370   0.5332   0.8755   2.2534
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.4568     0.2306 -10.653   <2e-16 ***
## SEXM                0.1560     0.1309   1.191    0.234
## as.factor(TRTPN)2   3.1982     0.2515  12.715   <2e-16 ***
## as.factor(TRTPN)3   3.0620     0.2323  13.182   <2e-16 ***
## as.factor(TRTPN)4   4.1797     0.2464  16.965   <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: 2149.9  on 1647  degrees of freedom
## Residual deviance: 1628.0  on 1643  degrees of freedom
##   (183 observations deleted due to missingness)
## AIC: 1638
##
## Number of Fisher Scoring iterations: 4
```

2. Impute the missingness and reanalyze the data

```
dat1_im <- dat1
dat1_im$PCHGCA1N[which(is.na(dat1_im$AVAL))] <- 0
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat1_im, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat1_im)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7759  -1.3036   0.6803   0.9433   2.3003
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.5722     0.2261 -11.375   <2e-16 ***
## SEXM                0.1499     0.1157   1.295    0.195
## as.factor(TRTPN)2   3.0016     0.2439  12.304   <2e-16 ***
## as.factor(TRTPN)3   2.8640     0.2281  12.558   <2e-16 ***
## as.factor(TRTPN)4   3.7679     0.2344  16.078   <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: 2493.8  on 1830  degrees of freedom
## Residual deviance: 2020.7  on 1826  degrees of freedom
## AIC: 2030.7
##
## Number of Fisher Scoring iterations: 5
```

3. repeat 1 & 2 with 20% missingness

```
## with 20% missingness
set.seed(1234)
pasi6_final2 <- generate_missing(pasi6, 0.2)
dat2 <- merge(adsl, pasi6_final2, by = "SUBJID")
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat2, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat2)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max  
## -2.0328  -0.4278   0.5205   0.8583   2.2780  
## 
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)    
## (Intercept)          -2.5170     0.2510 -10.026   <2e-16 ***
## SEXM                  0.1718     0.1402   1.225    0.221    
## as.factor(TRTPN)2     3.2763     0.2740  11.956   <2e-16 ***
## as.factor(TRTPN)3     3.1540     0.2528  12.475   <2e-16 ***
## as.factor(TRTPN)4     4.2759     0.2684  15.932   <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: 1898.6  on 1464  degrees of freedom
## Residual deviance: 1425.3  on 1460  degrees of freedom
##   (366 observations deleted due to missingness)
## AIC: 1435.3
## 
## Number of Fisher Scoring iterations: 5
```

```
## with 20% missingness imputed
dat2_im <- dat2
dat2_im$PCHGCA1N[which(is.na(dat2_im$AVAL))] <- 0
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat2_im, family = binomial))
```

```
## 
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial, 
##     data = dat2_im)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.5847  -1.2246   0.8189   1.0687   2.3790  
## 
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)    
## (Intercept)          -2.7691     0.2429 -11.399   <2e-16 ***
## SEXM                  0.1508     0.1115   1.352    0.176    
## as.factor(TRTPN)2     2.8794     0.2588  11.126   <2e-16 ***
## as.factor(TRTPN)3     2.8523     0.2451  11.635   <2e-16 ***
## as.factor(TRTPN)4     3.5386     0.2477  14.285   <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: 2535.6  on 1830  degrees of freedom
## Residual deviance: 2140.8  on 1826  degrees of freedom
## AIC: 2150.8
## 
## Number of Fisher Scoring iterations: 5
```

repeat 1 & 2 with 30% missingness

```
## with 30% missingness
set.seed(1234)
pasi6_final3 <- generate_missing(pasi6, 0.3)
dat3 <- merge(adsl, pasi6_final3, by = "SUBJID")
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat3, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0680  -0.4346   0.5008   0.8330   2.2785
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.5182     0.2668  -9.438   <2e-16 ***
## SEXM                 0.2061     0.1511   1.364    0.172
## as.factor(TRTPN)2    3.1821     0.2894  10.996   <2e-16 ***
## as.factor(TRTPN)3    3.1922     0.2684  11.893   <2e-16 ***
## as.factor(TRTPN)4    4.3250     0.2849  15.182   <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: 1651.6  on 1281  degrees of freedom
## Residual deviance: 1225.1  on 1277  degrees of freedom
##   (549 observations deleted due to missingness)
## AIC: 1235.1
##
## Number of Fisher Scoring iterations: 5
```

```
## with 30% missingness imputed
dat3_im <- dat3
dat3_im$PCHGCA1N[which(is.na(dat3_im$AVAL))] <- 0
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat3_im, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat3_im)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4471  -1.1401  -0.3355   0.9689   2.4106
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.8492     0.2535 -11.241   <2e-16 ***
## SEXM                 0.1022     0.1101   0.929    0.353
## as.factor(TRTPN)2    2.6584     0.2691   9.880   <2e-16 ***
## as.factor(TRTPN)3   2.6690     0.2561  10.422   <2e-16 ***
## as.factor(TRTPN)4   3.3617     0.2572  13.072   <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: 2525.8  on 1830  degrees of freedom
## Residual deviance: 2193.2  on 1826  degrees of freedom
## AIC: 2203.2
##
## Number of Fisher Scoring iterations: 5
```

MAR

```
# select subject id, visit number, percent change from baseline at each visit
# reshape the dataset into wide format
pasi_inter <- pasi %>%
  select(SUBJID, AVISIT, PCHG) %>%
  spread(AVISIT, PCHG)

# generate missing data conditioned on previous visits
# in miss4, miss5 and miss6, 0 is missing, 1 is not missing
set.seed(1234)
for (i in 1:nrow(pasi_inter)){
  if (pasi_inter$`VISIT 3`[i] < 10){
    pasi_inter$miss4[i] <- rbinom(1, 1, 0.7)
  } else{ pasi_inter$miss4[i] <- rbinom(1, 1, 0.95)}
}

for (i in 1:nrow(pasi_inter)){
  if (pasi_inter$miss4[i] == 0){
    pasi_inter$miss5[i] = 0
  } else if (pasi_inter$`VISIT 4`[i] < 10){
    pasi_inter$miss5[i] <- rbinom(1, 1, 0.7)
  } else{ pasi_inter$miss5[i] <- rbinom(1, 1, 0.95)}
}

for (i in 1:nrow(pasi_inter)){
  if (pasi_inter$miss4[i] == 0){
    pasi_inter$miss6[i] = 0
  } else if (pasi_inter$`VISIT 5`[i] < 10){
    pasi_inter$miss6[i] <- rbinom(1, 1, 0.7)
  } else{ pasi_inter$miss6[i] <- rbinom(1, 1, 0.95)}
}

# check percent of missing data at visit 6
1 - sum(pasi_inter$miss6)/nrow(pasi_inter) # about 17% missingness
```

```
## [1] 0.1693064
```

```
# generate missingness in the pasi6 data according to the missigness pattern obtained in the previous s
pasi6_mar <- pasi6
pasi6_mar$PCHGCA1N[which(pasi_inter$miss6 == 0)] <- NA # PCHGCA1N contains visit 6 results with missing
```

```
pasi6_mar$im6 <- pasi6_mar$PCHGCA1N
pasi6_mar$im6[which(is.na(pasi6_mar$PCHGCA1N))] <- 0 # impute all missing data as non-responders

dat_mar <- merge(adsl, pasi6_mar, by = "SUBJID")
```

```
# logistic regression on PASI75 with missing values excluded
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat_mar, family = binomial))
```

```
## 
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial, 
##     data = dat_mar)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.0431  -0.5322   0.5146   0.8693   2.1126  
## 
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         -2.1180     0.2494  -8.491   <2e-16 ***
## SEXM                 0.2350     0.1330   1.767   0.0773 .  
## as.factor(TRTPN)2    2.8004     0.2693  10.400   <2e-16 ***
## as.factor(TRTPN)3    2.6614     0.2509  10.608   <2e-16 ***
## as.factor(TRTPN)4    3.8378     0.2645  14.511   <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: 1893.6  on 1520  degrees of freedom
## Residual deviance: 1553.0  on 1516  degrees of freedom
##   (310 observations deleted due to missingness)
## AIC: 1563
## 
## Number of Fisher Scoring iterations: 4
```

```
# logistic regression on PASI75 with missing data imputed as 0
summary(glm(im6 ~ SEX + as.factor(TRTPN), data = dat_mar, family = binomial))
```

```
## 
## Call:
## glm(formula = im6 ~ SEX + as.factor(TRTPN), family = binomial, 
##     data = dat_mar)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.8053  -1.2527   0.6605   0.9640   2.3722  
## 
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         -2.7518     0.2393  -11.50   <2e-16 ***
## SEXM                 0.2029     0.1159    1.75   0.0801 .  
## as.factor(TRTPN)2    3.0741     0.2554   12.04   <2e-16 ***
## as.factor(TRTPN)3    2.9271     0.2405   12.17   <2e-16 ***
## as.factor(TRTPN)4    3.9604     0.2472   16.02   <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: 2502.7  on 1830  degrees of freedom
## Residual deviance: 2004.2  on 1826  degrees of freedom
## AIC: 2014.2
##
## Number of Fisher Scoring iterations: 5
```

MNAR

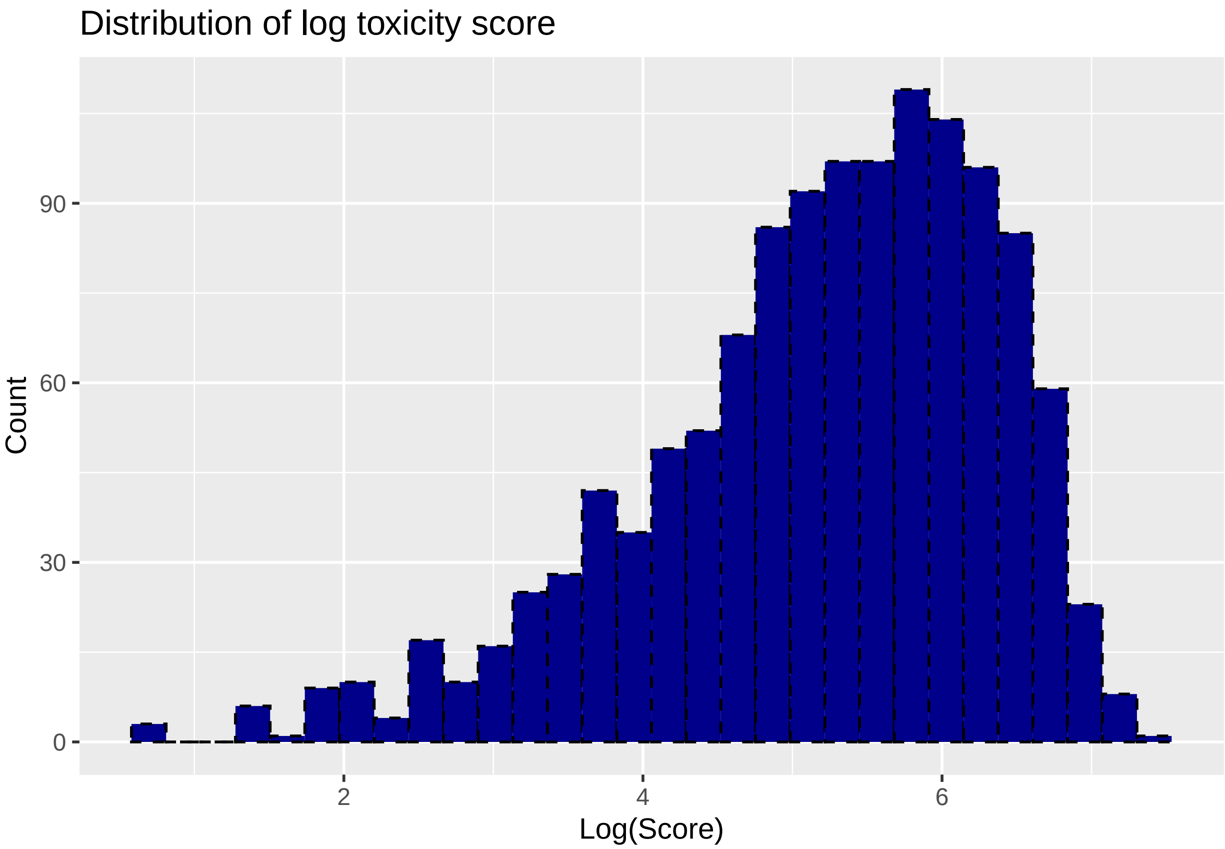
```
# sum the duration of each adverse event to generate total days of adverse event the patient had during
adae$duration <- adae$AEENDY - adae$AESTDY
#adduration <- adae %>% group_by(USUBJID) %>% summarise(duration = sum(duration))

# create severity score for each patient with AE using AESER, AESEV, and AEREL
# AESER_rate = 1 for Y, AESEV_rate : 1 for mild, 2 for moderate, 3 for severe, AEREL_rate = 1 for Y
adae$AESER_rate <- ifelse(adae$AESER == "Y", 1, 0)
adae$AESEV_rate <- ifelse(adae$AESEV == "MILD", 1, ifelse(adae$AESEV == "MODERATE", 2, 3))
adae$AEREL_rate <- ifelse(adae$AEREL == "Y", 1, 0)

# severity score score per event = AEREL_rate * 5 + AESER_rate * 3 + AESEV_rate * 2
adae$serscore <- adae$AEREL_rate * 5 + adae$AESER_rate * 3 + adae$AESEV_rate * 2

# toxic score = severity score * duration
adae$toxscore <- adae$serscore * adae$duration

# compute toxicity scale for each subject by sum the toxic score from each adverse event
toxscale <- adae %>% group_by(USUBJID) %>% summarise(toxscore = sum(toxscore))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
# check the distribution of the toxic scale
# hist(toxscale$toxscore)
toxscale$logscore <- log(toxscale$toxscore)
toxscale %>% ggplot(aes(x = logscore)) +
  geom_histogram(color="black", fill="darkblue", linetype = "dashed") +
  labs(title = "Distribution of log toxicity score", x = "Log(Score)", y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

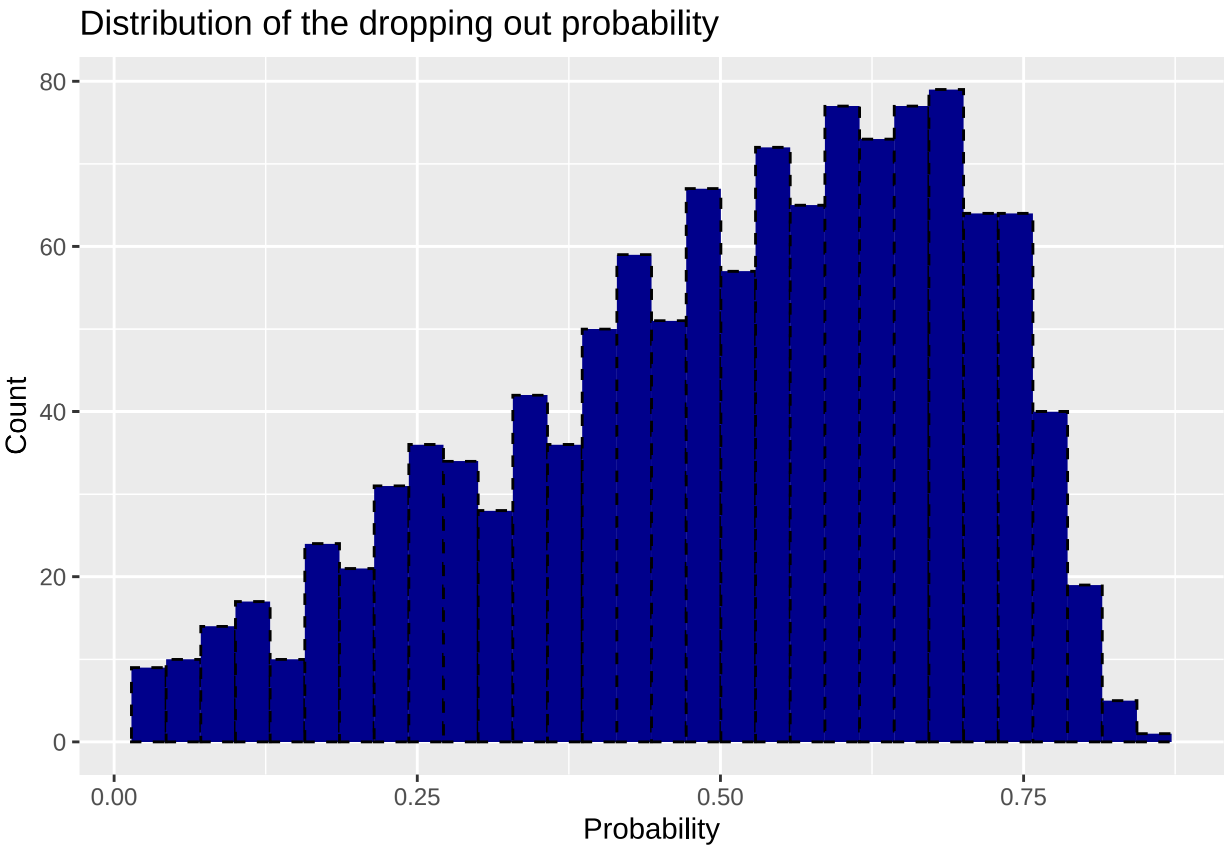


```
# assume the probability of dropping out is positively related to the toxicity scale and has an uniform
# the probability then can be simply determined by dividing the toxicity score with a common denominato
toxscale <- toxscale[which(toxscale$logscore != -Inf),]
mu <- mean(toxscale$logscore)
sigma <- sd(toxscale$logscore) + 1
toxscale$problog2 <- pnorm(toxscale$logscore, mu, sigma)

# impute the missingness of the subjects with adverse events with their probability of dropping out
aemissing <- toxscale
set.seed(1234)
for (i in 1:nrow(aemissing)){aemissing$misslog2[i] <- rbinom(1, 1, 1 - aemissing$problog2[i])}
```

```
## Warning: Unknown or uninitialised column: `misslog2`.
```

```
# check the probability distribution
toxscale %>% ggplot(aes(x = problog2)) +
  geom_histogram(color="black", fill="darkblue", linetype = "dashed") +
  labs(title = "Distribution of the dropping out probability", x = "Probability", y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
aemissing$SUBJID <- as.numeric(substr(aemissing$USUBJID, nchar(aemissing$USUBJID) - 4 + 1, nchar(aemiss

# check the percent of missingness
# table(aemissing$misslog2) # 33% missing values
# 605 / nrow(pasi6)

# join the aemissing table to pasi6 and set visit6 results to be missing based on results from the prev
pasi6_mnar <- left_join(pasi6, aemissing, by = "SUBJID")
pasi6_mnar$PCHGCA1N[which(pasi6_mnar$misslog2 == 0)] <- NA
pasi6_mnar$im6 <- pasi6_mnar$PCHGCA1N
pasi6_mnar$im6[which(is.na(pasi6_mnar$PCHGCA1N))] <- 0
dat_mnar <- merge(pasi6_mnar, adsl, by = "SUBJID")

# perform logistic regression on pasi75 with incomplete dataset and imputed dataset
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat_mnar, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat_mnar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9875  -0.4321   0.5466   0.8402   2.2899
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.5463     0.2541 -10.020   <2e-16 ***
## SEXM                0.2220     0.1495   1.485    0.138    
## as.factor(TRTPN)2   3.1769     0.2817  11.280   <2e-16 ***
## as.factor(TRTPN)3   3.1840     0.2559  12.442   <2e-16 ***
## as.factor(TRTPN)4   4.1499     0.2742  15.136   <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: 1649.5  on 1242  degrees of freedom
## Residual deviance: 1224.4  on 1238  degrees of freedom
##   (588 observations deleted due to missingness)
## AIC: 1234.4
## 
## Number of Fisher Scoring iterations: 5
```

```
summary(glm(im6 ~ SEX + as.factor(TRTPN), data = dat_mnar, family = binomial))
```

```
## 
## Call:
## glm(formula = im6 ~ SEX + as.factor(TRTPN), family = binomial, 
##     data = dat_mnar)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.2519  -1.1519  -0.3694   1.1280   2.3543  
## 
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)        -2.70675    0.24169 -11.199   <2e-16 ***
## SEXM                0.05606    0.10849   0.517    0.605    
## as.factor(TRTPN)2   2.38617    0.25901   9.213   <2e-16 ***
## as.factor(TRTPN)3   2.59042    0.24502  10.572   <2e-16 ***
## as.factor(TRTPN)4   2.82420    0.24511  11.522   <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: 2493.1  on 1830  degrees of freedom
## Residual deviance: 2247.2  on 1826  degrees of freedom
## AIC: 2257.2
## 
## Number of Fisher Scoring iterations: 5
```