Advanced Methods in Epidemiology

Practical Session on Missing Data Techniques

Prof. dr. Steven Abrams

Master of Epidemiology | Academic year: 2021-2022

Background

The incidence of human immunodeficiency virus (HIV) in Sub-Saharan Africa (SSA) remains high and disproportionate compared to the global average. More specifically, in 2017 a five-fold higher number of new HIV cases was reported in SSA compared to the global average (WHO, 2019). Two-thirds of the new HIV cases among youth occur in adolescent girls and young women (AGYW) aged between 15 and 24 years of age (UNAIDS, 2015, 2016). Four out of five AGYW living with HIV reside in SSA, and they acquire HIV infection on average five to seven years earlier than men (UNAIDS, 2014, 2015). The highest number of people living with HIV reside in South Africa (Shisana, 2014) and AGYW aged between 15- and 24-years old account for about 26% of all new HIV infections, despite that AGYW only make up 10% of the entire population (UNAIDS, 2017).

In general, AGYW are at a higher risk of acquiring HIV compared to young men and older women (Dellar et al., 2015; Harrison, 2014). Social and contextual factor including gender inequality and gender-based violence make them vulnerable for risk-taking behavior such as engagement in unprotected sexual intercourse or substance abuse. Consequently, these factors compound to their vulnerabilities and increase risk of HIV acquisition (Harrison et al., 2016; Arnett, 1992; Tapert et al., 2001). Despite the implementation of intervention programs for AGYW, the effects are suboptimal and the decrease in incidence in AGYW remains slow (UNAIDS, 2015).

An adequate level of risk perception is critical to adopt a protective behavior concerning HIV acquisition. The purpose of this study is to **identify factors associated with HIV risk perception among AGYW**. Here, we will focus on risk perception at baseline and which factors are associated therewith.

Data description

The **DREAMS** (Determined, Resilient, Empowered, AIDS-free, Mentored and Safe) initiative, funded by the US President's Emergency Plan for AIDS Relief (PEPFAR) (Saul et al., 2018), aimed at keeping adolescent girls and young women (AGYW) HIV-negative through pre-exposure prophylaxis (PrEP) and secondary distribution of HIV self-test kits to their male sexual partners. The study was conducted between September 18, 2017, and October 31, 2018, at primary healthcare facilities or community outreach programs in Northern Johannesburg, South Africa. Participants were eligible if they were between 16 and 24 years old, tested HIV-negative, had a sexual male partner who was HIV-negative or of unknown HIV status and which were in a heterosexual relationship for at least 3 months.

```
## Import the data
##-----------------
all_dat = read.table("Dreams_reduced.csv", header = T, sep = ";",dec=",")
head(all_dat, 5)
names(all_dat)
str(all_dat)
```

Primary endpoint

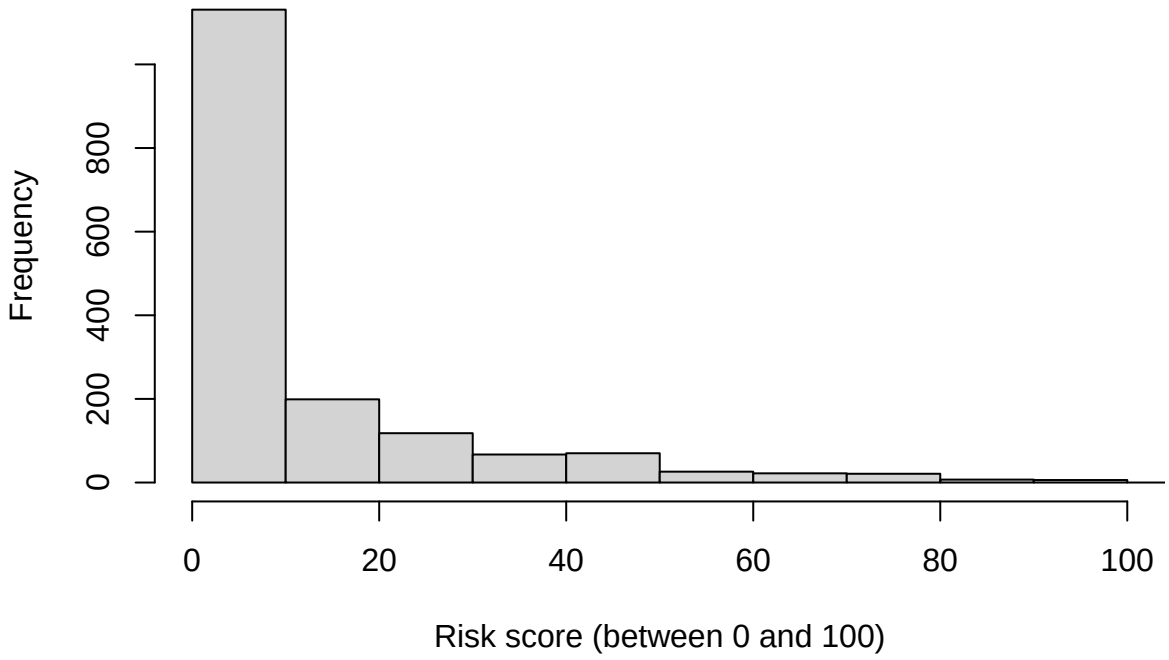
The primary endpoint of this analysis was self-assessed risk of acquiring HIV infection. At baseline, participants were asked "What do you think the chance is that you will get HIV in the next six months?", thereby expressing this on a scale ranging from 0 to 100, with 100 being the maximum value.

A first exploration of the primary endpoint (i.e., self-assessed risk of acquiring HIV infection) is shown below:

```
## Primary endpoint
##------------------
tab_primary = table(all_dat$risk_score)
table(all_dat$risk_score > 0)
```

```
##
## FALSE  TRUE
##   614  1057
```

```
h_primary = hist(all_dat$risk_score, xlim = c(0,100), breaks = 100, main = "",
                 xlab = "Risk score (between 0 and 100)")
```




```
## Some remarks:
## Although the perceived risk, in theory, takes values between 0 and 100, the
## majority of the respondents indicate to experience zero risk of acquiring
## HIV infection. Consequently, the marginal distribution of risk scores is
## characterized by so-called zero-inflation, meaning that, as compared to
## existing parametric distributions there is an over-representation of zeros.
## This is an important aspect to account for in subsequent (missing data)
## analyses.
##
## Next to zero-inflation, the observed response data is subject to digit
## preference, meaning that individuals tend to 'round' their risk score to
## specific (rounded) values. Although statistical techniques exist to cope with
## digit preference, including for example composite link models, we will not
## deal with this complexity in this tutorial.
```

Furthermore, next to individual-level characteristics other potential factors related to HIV risk perception of AGYW were recorded including socio-economic and demographic variables, overall and sexual health behavior, source of HIV information and relationship characteristics.

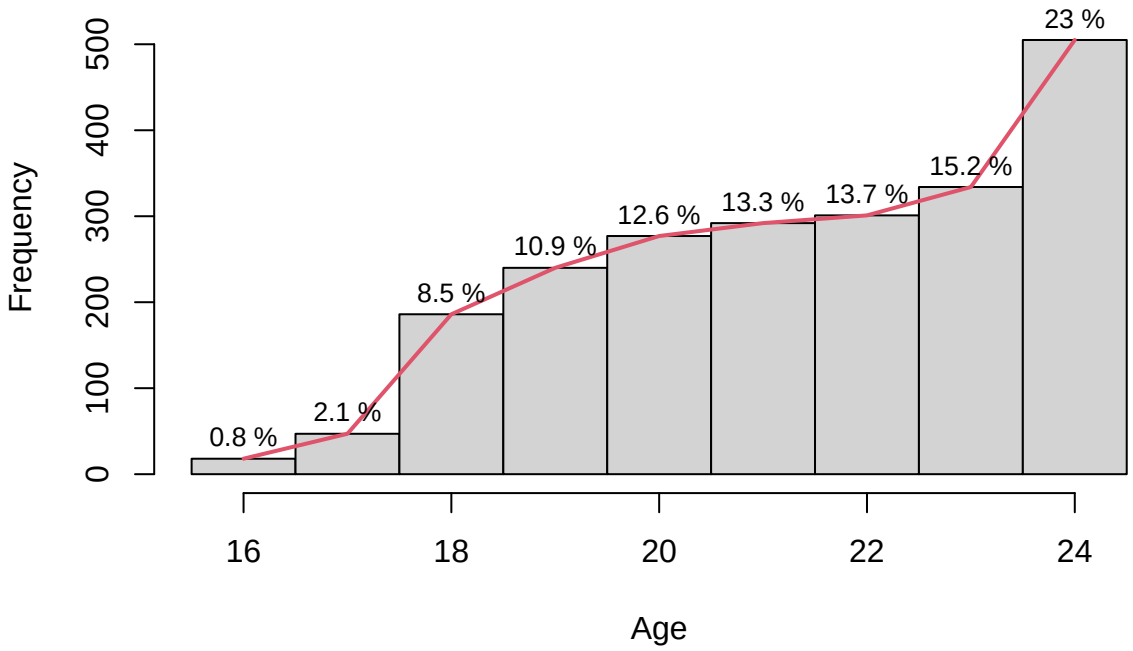
1. How many participants are included in the study? How many variables?

```
## Participants
##--------------
n <- nrow(all_dat)
print(n)
```

```
## [1] 2200
```

2. What is the age distribution of individuals in the DREAMS study?

```
## Age distribution
##--------------
tab_age = table(all_dat$age)
rel_freq = tab_age/n

xgrid = seq(min(all_dat$age)-0.5,max(all_dat$age)+0.5,1)
h = hist(all_dat$age, main = "", ylab = "Frequency", xlab = "Age",
         breaks = xgrid, ylim = c(0,550))
lines(as.numeric(names(tab_age)), tab_age, col = 2, lwd = 2)
text(xgrid[-length(xgrid)] + 0.5, tab_age + 25,
     paste(round(rel_freq*100,1),"%"), cex = 0.8)
```



Missing data

For some of the participants, information about the primary endpoint and/or additional covariate information is lacking. Quantify the extent of the missing data in the dataset:

1. For some of the variables, missing observations are coded as 999 instead of NA. First recode missing observations as NA. **NOTE**: for the primary endpoint, some observations are equal to 777 which requires recoding to NA as well.

```r
## Recoding for primary endpoint
##-------------------------------
all_dat$risk_score[all_dat$risk_score == 777 | all_dat$risk_score == 999] = NA

## Recoding for other variables
##-----------------------------
n_recodings = apply(all_dat, 2, function(x){sum(x == 999, na.rm = T)})
all_dat[which(all_dat == 999, arr.ind = T)] = NA

## Number of missing observations for the primary endpoint
##--------------------------------------------------------
sum(is.na(all_dat$risk_score))
```

```
## [1] 533
```

2. Explore the level of missing data in the dataset by calculating the frequency of missing observations for each variable.

```r
## Missing data
##--------------
n_missing = apply(all_dat, 2, FUN = function(x){sum(is.na(x))})
perc_missing = n_missing/nrow(all_dat)
perc_missing
```

```
##                     study_id                          age
##                0.0000000000               0.0000000000
##                     studying           educational_level
##                0.0000000000               0.0000000000
##           country_of_origin                mother_alive
##                0.0000000000               0.0000000000
##               mother_status                father_alive
##                0.0000000000               0.0000000000
##               father_status            residential_area
##                0.0000000000               0.0000000000
##           household_members             household_water
##                0.0000000000               0.0000000000
##         electricity_at_home        no_food_in_household
##                0.0000000000               0.0000000000
##             not_enough_food      whole_day_without_food
##                0.0000000000               0.0000000000
##              monthly_income               steady_income
##                0.0000000000               0.0000000000
##            government_grant   r1000_in_personal_accoount
##                0.0000000000               0.0000000000
##                  dependants           financial_support
##                0.0000000000               0.0000000000
##  control_on_household_money              away_from_home
##                0.0000000000               0.0004545455
##               have_children           how_many_children
##                0.0000000000               0.4559090909
##            kids_living_with          ever_been_pregnant
##                0.4513636364               0.0000000000
##          currently_pregnant   want_to_have_more_children
##                0.3263636364               0.1618181818
##        preventing_pregnancy     method_prevent_pregnancy
```

```
##                 0.1622727273                 0.3936363636
##                  alcohol_use                          sti
##                 0.0000000000                 0.0000000000
##         discuss_health_issue                     hiv_talk
##                 0.0000000000                 0.0000000000
##              hiv_talk_mostly          hiv_on_social_media
##                 0.1009090909                 0.0000000000
##            hiv_info_internet          ever_tested_for_hiv
##                 0.0000000000                 0.0000000000
##                last_hiv_test       first_sexual_encounter
##                 0.0445454545                 0.0000000000
##              sexual_partners        how_many_intercourses
##                 0.0000000000                 0.0000000000
##            sex_after_alcohol       sex_with_drunk_partner
##                 0.0000000000                 0.0000000000
##     sex_for_service_or_goods                sex_for_money
##                 0.0000000000                 0.0000000000
##            physical_violence physical_violence_by_partner
##                 0.0000000000                 0.0000000000
##              sexual_violence          relationship_status
##                 0.0000000000                 0.0000000000
##      current_sexual_partners                   risk_score
##                 0.0009090909                 0.2422727273
##      sex_partners_past_6_mos                  risk_cat_bl
##                 0.8268181818                 0.0000000000
##                     risk_cat                      agediff
##                 0.2404545455                 0.0000000000
##                     age5diff       total_relationship_yrs
##                 0.0000000000                 0.0000000000
```

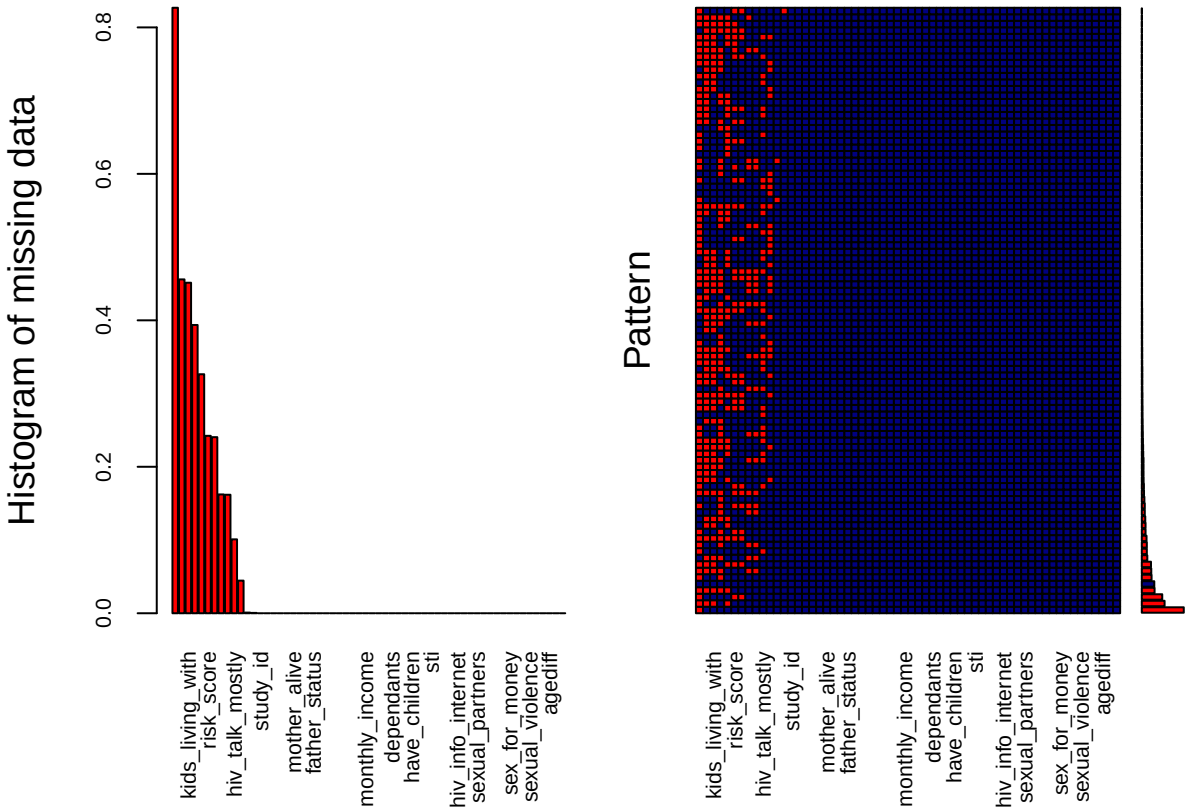
3. How many participants have at least one missing observation for the recorded variables?

```
## Individuals with missing observations
##---------------------------------------
n_subjects_missing_values = sum(apply(all_dat, 1, function(x){sum(is.na(x))}) > 0)
n_subjects_missing_values
```

```
## [1] 2071
```

4. Alternatively, provide a graphical exploration of the amount of missing data by variable using the function *aggr* in the R package VIM.

```
## Graphical exploration
##-----------------------
suppressPackageStartupMessages(library(VIM))
aggr_plot <- aggr(all_dat, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE,
labels=names(all_dat), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##                       Variable        Count
##       sex_partners_past_6_mos 0.8268181818
##             how_many_children 0.4559090909
##              kids_living_with 0.4513636364
##      method_prevent_pregnancy 0.3936363636
##            currently_pregnant 0.3263636364
##                    risk_score 0.2422727273
##                      risk_cat 0.2404545455
##          preventing_pregnancy 0.1622727273
##    want_to_have_more_children 0.1618181818
##               hiv_talk_mostly 0.1009090909
##                 last_hiv_test 0.0445454545
##       current_sexual_partners 0.0009090909
##                away_from_home 0.0004545455
##                      study_id 0.0000000000
##                           age 0.0000000000
##                      studying 0.0000000000
##             educational_level 0.0000000000
##             country_of_origin 0.0000000000
##                  mother_alive 0.0000000000
##                 mother_status 0.0000000000
##                  father_alive 0.0000000000
##                 father_status 0.0000000000
##              residential_area 0.0000000000
##             household_members 0.0000000000
##               household_water 0.0000000000
```

```
##           electricity_at_home 0.0000000000
##         no_food_in_household 0.0000000000
##              not_enough_food 0.0000000000
##       whole_day_without_food 0.0000000000
##                monthly_income 0.0000000000
##                 steady_income 0.0000000000
##             government_grant 0.0000000000
##   r1000_in_personal_accoount 0.0000000000
##                   dependants 0.0000000000
##            financial_support 0.0000000000
##   control_on_household_money 0.0000000000
##                have_children 0.0000000000
##           ever_been_pregnant 0.0000000000
##                  alcohol_use 0.0000000000
##                          sti 0.0000000000
##         discuss_health_issue 0.0000000000
##                     hiv_talk 0.0000000000
##          hiv_on_social_media 0.0000000000
##            hiv_info_internet 0.0000000000
##          ever_tested_for_hiv 0.0000000000
##       first_sexual_encounter 0.0000000000
##              sexual_partners 0.0000000000
##        how_many_intercourses 0.0000000000
##            sex_after_alcohol 0.0000000000
##       sex_with_drunk_partner 0.0000000000
##     sex_for_service_or_goods 0.0000000000
##                sex_for_money 0.0000000000
##            physical_violence 0.0000000000
## physical_violence_by_partner 0.0000000000
##              sexual_violence 0.0000000000
##          relationship_status 0.0000000000
##                  risk_cat_bl 0.0000000000
##                      agediff 0.0000000000
##                     age5diff 0.0000000000
##       total_relationship_yrs 0.0000000000
```

Missing data techniques and questions

Let us first focus on the primary endpoint. Study the marginal distribution of the observed risk perception scores for the participants in the DREAMS study.

1. Which missingness mechanisms do exist? Explain.

```
## Three missingness mechanisms exist: MCAR, MAR and MNAR (see definitions in
## course notes). Little (1988)* introduced a chi-square test statistic to
## assess whether data are missing completely at random (MCAR). Under the null
## hypothesis, the data are considered missing completely at random. However, a
## distinction between data being MAR or MNAR is impossible.
## * Little, R. J. A. (1988). A Test of Missing Completely at Random for
##   Multivariate Data with Missing Values. Journal of the American Statistical
##   Association 83 (404): 1198-1202. doi:10.1080/01621459.1988.10478722.

## Apply Little's MCAR test
##-------------------------
```

```r
#install.packages('naniar')
library(naniar)

mcar_test(all_dat)
```

```
## # A tibble: 1 x 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int>
## 1 19688262.  5093       0               93
```

```r
## Based on all available data, the null hypothesis of MCAR is rejected at a
## 5% significance level (p-value being very small). NOTE: one should be careful
## when using the MCAR test of Little, given that the asymptotics for the test
## statistic are mainly valid for quantitative data. For categorical data, more
## appropriate large-sample tests of the MCAR assumption can be based on the
## chi-squared statistics in Fuchs (1982)**. Next to that, other limitations
## are listed by Little (1988)*.
## ** Fuchs, C. (1982), "Maximum Likelihood Estimation and Model Selection in
## Contingency Tables With Missing Data," Journal of the American Statistical
## Association, 77, 270-278.

## As an alternative, one could model the probability of a missing observation
## in relation to the observed covariate information (i.e., predictors and
## auxiliary variables). In order to do so, we consider the risk perception
## score as outcome variable and define a variable R indicating whether Z has
## a missing observation or not.

## Define the variable W
##-----------------------
all_dat$r <- as.numeric(is.na(all_dat$risk_score))

## Check the missingness percentage (see above)
##---------------------------------------------
abs_missing <- table(all_dat$r)
rel_missing <- abs_missing/sum(abs_missing)
rel_missing
```

```
##
##         0         1
## 0.7577273 0.2422727
```

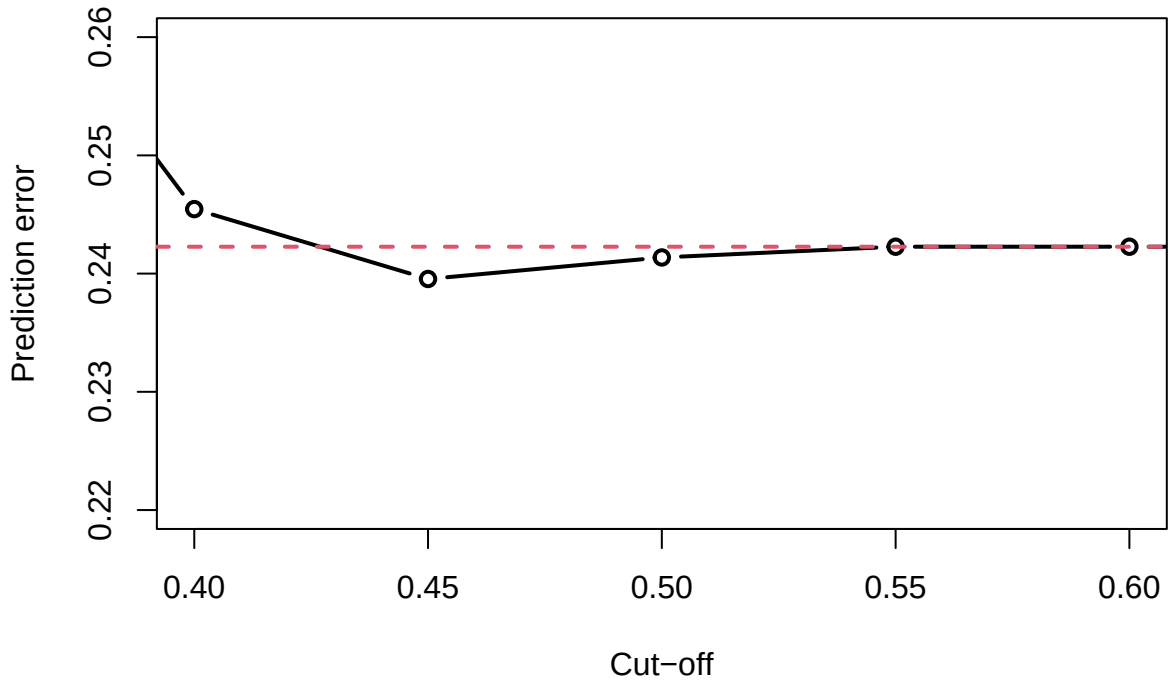
```r
## We now fit a logistic regression model to investigate whether the missingness
## probability can be explained based on additional information in the dataset.

## Logistic regression model (for missingness indicator)
##------------------------------------------------------
missing_glm = glm(r ~ age + factor(educational_level) +
                factor(residential_area) + monthly_income, data = all_dat,
                family = binomial(link = "logit"))
summary(missing_glm)
```

```
##
## Call:
## glm(formula = r ~ age + factor(educational_level) + factor(residential_area) +
##     monthly_income, family = binomial(link = "logit"), data = all_dat)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3842  -0.7833  -0.6503  -0.3637   2.3518
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       -3.257e-01  5.680e-01
## age                                               -7.961e-02  2.376e-02
## factor(educational_level)Completed primary school  9.446e-01  3.341e-01
## factor(educational_level)No formal schooling      -1.241e+01  3.576e+02
## factor(educational_level)Some high/secondary schooling  4.200e-01  1.151e-01
## factor(educational_level)Some primary school       8.968e-01  4.358e-01
## factor(educational_level)Tertiary education       -4.026e-01  1.658e-01
## factor(residential_area)Diepsloot                  7.941e-01  2.260e-01
## factor(residential_area)Fourways                  -1.213e-01  4.213e-01
## factor(residential_area)Kyasand/pipeline           8.361e-01  2.971e-01
## factor(residential_area)Msawawa                    1.212e+00  3.036e-01
## factor(residential_area)Other                      8.376e-01  2.480e-01
## monthly_income                                    -1.323e-06  9.468e-06
##                                                   z value Pr(>|z|)
## (Intercept)                                        -0.573 0.566359
## age                                                -3.351 0.000805 ***
## factor(educational_level)Completed primary school   2.828 0.004691 **
## factor(educational_level)No formal schooling       -0.035 0.972315
## factor(educational_level)Some high/secondary schooling   3.647 0.000265 ***
## factor(educational_level)Some primary school        2.058 0.039611 *
## factor(educational_level)Tertiary education        -2.429 0.015161 *
## factor(residential_area)Diepsloot                   3.513 0.000443 ***
## factor(residential_area)Fourways                   -0.288 0.773477
## factor(residential_area)Kyasand/pipeline            2.814 0.004888 **
## factor(residential_area)Msawawa                     3.992 6.56e-05 ***
## factor(residential_area)Other                       3.377 0.000733 ***
## monthly_income                                     -0.140 0.888902
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2436.2  on 2199  degrees of freedom
## Residual deviance: 2337.5  on 2187  degrees of freedom
## AIC: 2363.5
##
## Number of Fisher Scoring iterations: 12
```

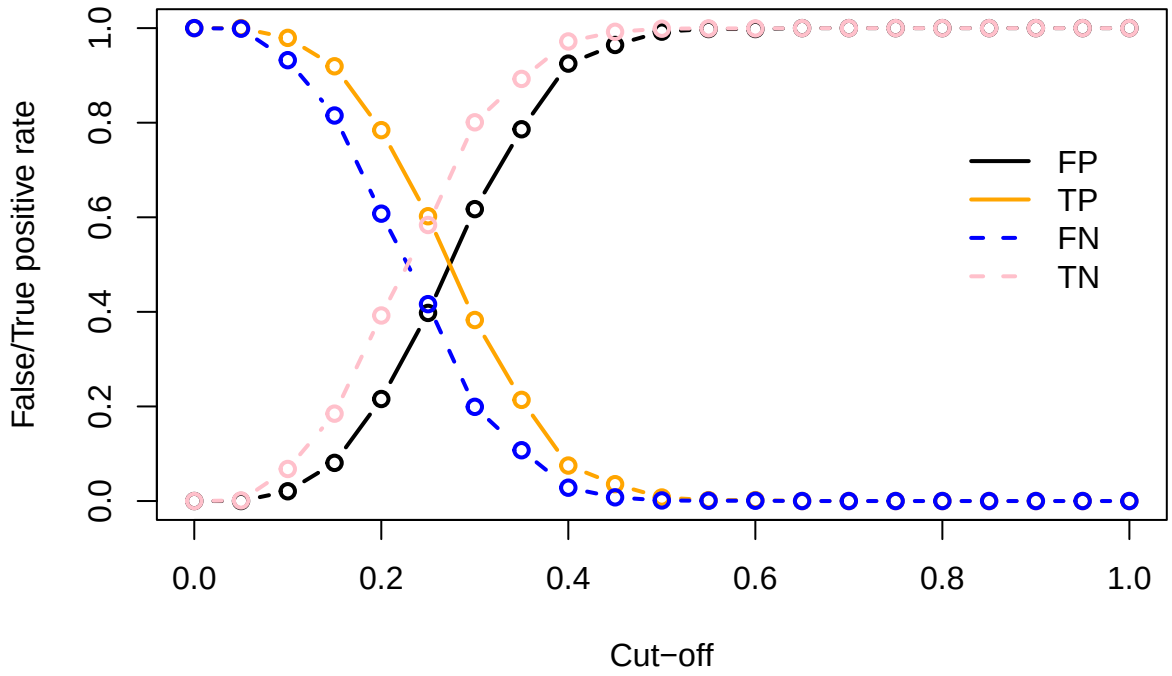
```
cutoffs = seq(0,1,0.05)
prediction_error = vector();
precision = vector();
recall = vector();
F1 = vector();
TotP = sum(all_dat$r == 1);
TotN = sum(all_dat$r == 0);
FP = vector(); TP = vector();
FN = vector(); TN = vector();
```

```r
for (j in 1:length(cutoffs)){
  r_prediction = as.numeric(predict(missing_glm, type = "response") > cutoffs[j])
  prediction_error[j] = sum(abs(all_dat$r - r_prediction))/length(all_dat$r);

  FP[j] = sum(abs(all_dat$r[all_dat$r == 1] - r_prediction[all_dat$r == 1]));
  TP[j] = TotP - FP[j];

  FN[j] = sum(abs(all_dat$r[all_dat$r == 0] - r_prediction[all_dat$r == 0]));
  TN[j] = TotN - FN[j];

  precision[j] = TP[j]/(TP[j] + FP[j]);
  recall[j] = TP[j]/(TP[j] + FN[j]);

  F1[j] = (2*precision[j]*recall[j])/(precision[j] + recall[j]);
}

plot(cutoffs, prediction_error, xlab = "Cut-off",  ylab = "Prediction error",
     lwd = 2, type = "b", xlim = c(0.4, 0.6), ylim = c(0.22, 0.26))
abline(h = rel_missing[2], col = 2, lwd = 2, lty = 2)
```



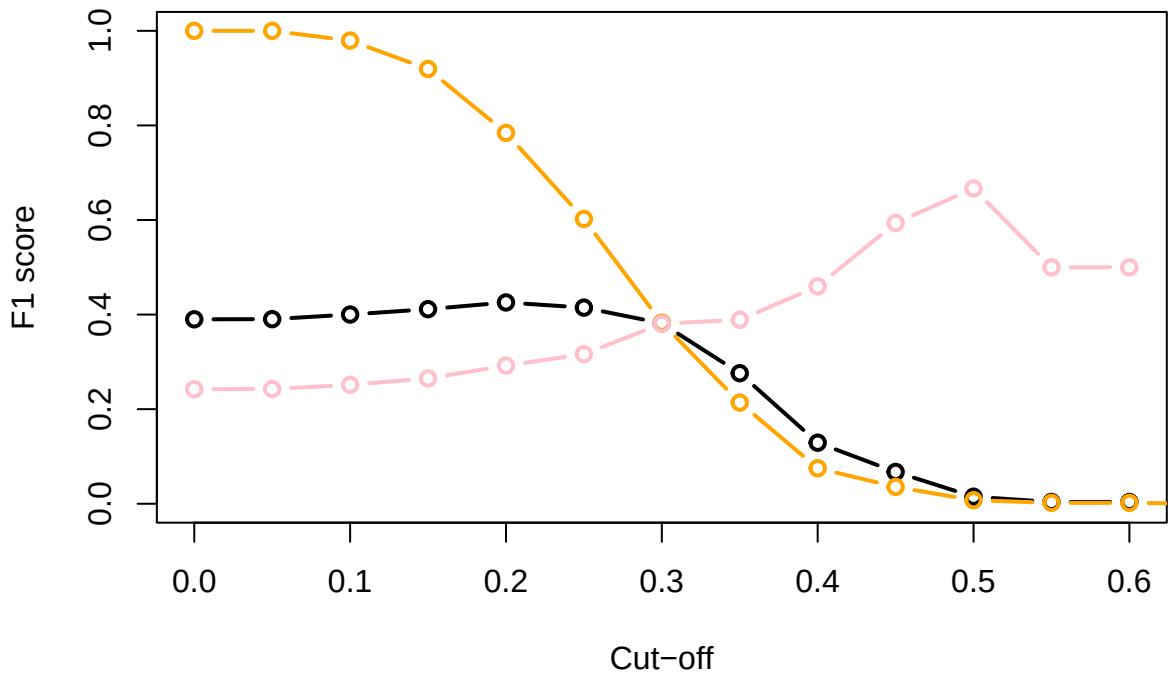

```r
## The prediction error of a model represents the number of times a model made a
## wrong prediction across the entire dataset. Equivalently, performance of a
## prediction model can be expressed in terms of model accuracy defined as the
## number of times a model correctly predicts the outcome value of interest, in
## this case the presence of missingness. Needless to say, accuracy (and
## prediction error) is only a useful metric in case of class-balanced, meaning
## that each outcome class has the same number of samples. In our case, there
## is class-imbalance given that only about 24.2% of the risk score values are
## missing. Class-imbalance implies that labelling all observations as non-
## missing leads to a prediction error of only 24.2% (red dashed line), hence,
## the overall performance of the prediction model should be evaluated in
## relation to this level. Alternatively, other metrics have been proposed in
```

```
## the literature to quantify the performance of a prediction model.

## 1. True and false positives: number of samples correctly or wrongly predicted
##    as positive
## 2. True and false negatives: number of samples correctly or wrongly predicted
##    as negative
plot(cutoffs, FP/TotP, xlab = "Cut-off",  ylab = "False/True positive rate",
     lwd = 2, type = "b")
lines(cutoffs, TP/TotP, lwd = 2, col = "orange", type = "b")
lines(cutoffs, FN/TotN, lwd = 2, col = "blue", type = "b", lty = 2)
lines(cutoffs, TN/TotN, lwd = 2, col = "pink", type = "b", lty = 2)
legend(0.8, 0.8, c("FP", "TP", "FN", "TN"),
       col = c("black","orange","blue","pink"),
       lty = c(1,1,2,2), lwd = 2, bty = "n")
```




```
## A metric used in case of class-imbalance is the F1 score, balancing precision
## and recall using a harmonic mean. Consequently, maximising the F1 score
## implies simultaneously maximising both precision and recall.
## Based on TP, FP, TN, FN, one can define precision and recall scores,
## mathematically defined as:
##      Precision = TP/(TP + FP)
##      Recall = TP/(TP + FN)
## Precision is defined as the proportion of true positives among all positive
## predictions; measures the extent of error caused by FPs
## Recall is defines as the proportion of positive predictions, among all true
## positive cases; measures the extent of error caused by FNs
## The use of recall is preferred when you value FN as being worse than FPs
## while precision is used in case you want to maximise TP.
## The F1 score provides a way to maximise both precision and recall and is
## defined as the harmonic mean of the precision and recall scores, ranging
## between 0 and 100% with higher F1 values representing a better performance of
## the prediction model (or classifier).
```

```
plot(cutoffs, F1, lwd = 2, type = "b", xlab = "Cut-off",
     ylab = "F1 score", xlim = c(0,0.6), ylim = c(0,1))
lines(cutoffs, precision, type = "b", lwd = 2, col = "orange")
lines(cutoffs, recall, type = "b", lwd = 2, col = "pink")
```




```
## Based on the F1 score, a cut-off of 0.2 produces a prediction ability with
## optimal balance between precision and recall. However, the prediction model
## has still a suboptimal performance.

## In order to improve prediction, one could either (1) consider an extended
## model in terms of the observed covariate information included in the model or
## (2) train a random forest classifier. However, this is not the topic of this
## practical session. In conclusion, there seems to be some predictive ability
## in terms of missingness based on the covariate information (which can be
## improved by adding additional covariates - not shown here). Consequently, we
## assume hereunder that data are missing at random (MAR) and that the afore-
## mentioned variables (i.e., age, educational level, residential area and
## monthly income) can be used to construct an appropriate imputation model.
```

2. Formulate an imputation model for the (pseudo-)continuous endpoint Z defined as the risk perception score (on a range of 0 – 100).

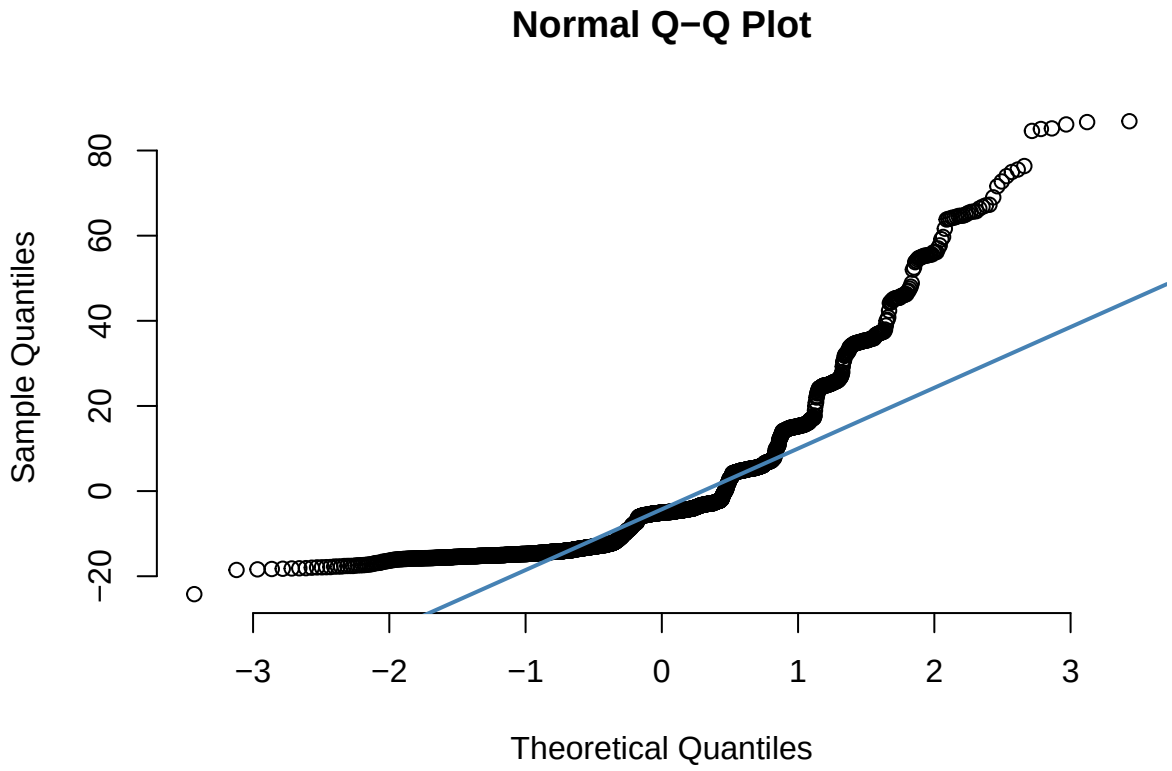
a. More specifically, fit a linear regression model for Z in terms of age, educational level, residential area and monthly income.

```
## Define imputation model for Z = risk perception score
##------------------------------------------------------

## A first naive approach is to consider directly a linear regression model to
## impute missing observations for the risk perception score (on the range
## 0-100)

## Linear regression model without transformation
##-----------------------------------------------
```

```r
lm_fit1 = glm(risk_score ~ age + factor(educational_level) +
                factor(residential_area) + monthly_income, data = all_dat)
summary(lm_fit1)
```

```
##
## Call:
## glm(formula = risk_score ~ age + factor(educational_level) +
##     factor(residential_area) + monthly_income, data = all_dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -24.207  -13.899   -5.062    5.352   86.876
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       1.180e+01  5.186e+00
## age                                               1.256e-01  2.246e-01
## factor(educational_level)Completed primary school -3.077e+00  3.887e+00
## factor(educational_level)No formal schooling       1.311e+01  1.355e+01
## factor(educational_level)Some high/secondary schooling  6.517e-01  1.098e+00
## factor(educational_level)Some primary school      -5.869e+00  5.341e+00
## factor(educational_level)Tertiary education        3.034e-01  1.260e+00
## factor(residential_area)Diepsloot                  7.432e-01  1.551e+00
## factor(residential_area)Fourways                   2.186e+00  2.652e+00
## factor(residential_area)Kyasand/pipeline           3.168e+00  2.517e+00
## factor(residential_area)Msawawa                   -1.109e+00  2.810e+00
## factor(residential_area)Other                     -1.207e+00  1.807e+00
## monthly_income                                    -1.327e-04  7.695e-05
##                                                   t value Pr(>|t|)
## (Intercept)                                         2.275   0.0230 *
## age                                                 0.559   0.5761
## factor(educational_level)Completed primary school  -0.792   0.4287
## factor(educational_level)No formal schooling        0.967   0.3335
## factor(educational_level)Some high/secondary schooling   0.594   0.5528
## factor(educational_level)Some primary school       -1.099   0.2720
## factor(educational_level)Tertiary education         0.241   0.8097
## factor(residential_area)Diepsloot                   0.479   0.6318
## factor(residential_area)Fourways                    0.824   0.4098
## factor(residential_area)Kyasand/pipeline            1.259   0.2082
## factor(residential_area)Msawawa                    -0.395   0.6932
## factor(residential_area)Other                      -0.668   0.5044
## monthly_income                                     -1.724   0.0849 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 359.8073)
##
##     Null deviance: 599739  on 1666  degrees of freedom
## Residual deviance: 595121  on 1654  degrees of freedom
##   (533 observations deleted due to missingness)
## AIC: 14557
##
## Number of Fisher Scoring iterations: 2
```

```r
## Model diagnostics (normality)
##------------------------------
qqnorm(residuals(lm_fit1), pch = 1, frame = FALSE)
qqline(residuals(lm_fit1), col = "steelblue", lwd = 2)
```



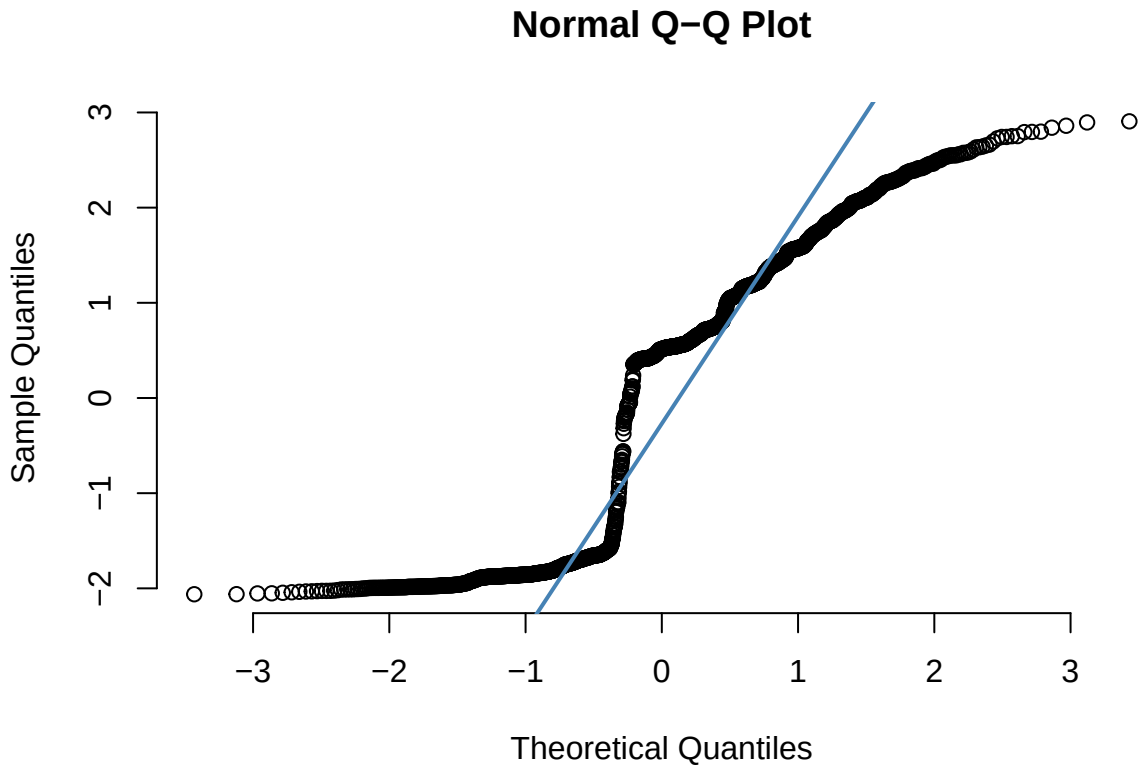
b. What about the excess number of zero observations? Would a transformation help

```r
## Linear regression model with transformation
##--------------------------------------------------
lm_fit2 = glm(log(risk_score + 1) ~ age + factor(educational_level) +
              factor(residential_area) + monthly_income, data = all_dat)
summary(lm_fit2)
```

```
##
## Call:
## glm(formula = log(risk_score + 1) ~ age + factor(educational_level) +
##     factor(residential_area) + monthly_income, data = all_dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0619  -1.7374   0.5138   1.1987   2.9066
##
## Coefficients:
##                                                        Estimate Std. Error
## (Intercept)                                           1.905e+00  4.146e-01
## age                                                  -5.389e-03  1.795e-02
## factor(educational_level)Completed primary school    -2.097e-01  3.107e-01
## factor(educational_level)No formal schooling          1.187e+00  1.083e+00
## factor(educational_level)Some high/secondary schooling 1.087e-01  8.775e-02
## factor(educational_level)Some primary school         -5.117e-01  4.270e-01
```

```
## factor(educational_level)Tertiary education           -1.688e-02  1.007e-01
## factor(residential_area)Diepsloot                      1.060e-01  1.240e-01
## factor(residential_area)Fourways                      -6.072e-02  2.120e-01
## factor(residential_area)Kyasand/pipeline               1.584e-01  2.012e-01
## factor(residential_area)Msawawa                       -1.554e-01  2.246e-01
## factor(residential_area)Other                         -9.006e-02  1.445e-01
## monthly_income                                        -7.594e-06  6.152e-06
##                                                       t value Pr(>|t|)
## (Intercept)                                             4.594 4.67e-06 ***
## age                                                    -0.300    0.764
## factor(educational_level)Completed primary school      -0.675    0.500
## factor(educational_level)No formal schooling            1.095    0.273
## factor(educational_level)Some high/secondary schooling  1.239    0.215
## factor(educational_level)Some primary school           -1.198    0.231
## factor(educational_level)Tertiary education            -0.168    0.867
## factor(residential_area)Diepsloot                       0.855    0.393
## factor(residential_area)Fourways                       -0.286    0.775
## factor(residential_area)Kyasand/pipeline                0.787    0.431
## factor(residential_area)Msawawa                        -0.692    0.489
## factor(residential_area)Other                          -0.623    0.533
## monthly_income                                         -1.234    0.217
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.299272)
##
##     Null deviance: 3839.4  on 1666  degrees of freedom
## Residual deviance: 3803.0  on 1654  degrees of freedom
##   (533 observations deleted due to missingness)
## AIC: 6133.6
##
## Number of Fisher Scoring iterations: 2
```

```r
## Model diagnostics (normality)
##------------------------------
qqnorm(residuals(lm_fit2), pch = 1, frame = FALSE)
qqline(residuals(lm_fit2), col = "steelblue", lwd = 2)
```




```
## Needless to say based on the fitted models, the underlying assumption of
## normality is violated irrespective of the use of a logarithmic transformation.
## One of the reasons is the zero-inflation of the distribution of the observed
## values for the primary endpoint. Hence, we consider now a different strategy
## to construct an imputation model for the data at hand. More specifically,
## we assume that the underlying distribution for the response variable,
## conditional on a subset of covariates that are available, is a two-component
## mixture distribution with either a risk score value equal to zero or, when
## larger than zero coming from a (log)normal distribution with mean mu and
## variance sigma2 (parameters estimated from the observed data; see below).
```

3. Consider the random variable Y defined as having a risk score equal to zero (no risk of HIV acquisition). Construct an imputation model for Y, under the assumption of missing at random (MAR), and depending on the covariates age, educational level, residential area, and monthly income. What type of model will you consider?

```
## We first formulate a logistic regression model for the random variable Y,
## conditional on covariates, i.e., modelling the probability of having a
## zero risk perception score

## Define imputation model for Y = risk perception score equal to zero
##-------------------------------------------------------------------------
all_dat$y1 = as.numeric(all_dat$risk_score == 0)

## Logistic regression model
##--------------------------
glm_fit = glm(y1 ~ age + factor(educational_level) + factor(residential_area) +
                monthly_income, data = all_dat,
              family = "binomial"(link = logit))
summary(glm_fit)
```

```
##
## Call:
## glm(formula = y1 ~ age + factor(educational_level) + factor(residential_area) +
##     monthly_income, family = binomial(link = logit), data = all_dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5366  -0.9535  -0.8869   1.3626   1.5567
##
## Coefficients:
##                                                        Estimate Std. Error
## (Intercept)                                          -7.472e-01  5.709e-01
## age                                                   1.365e-02  2.476e-02
## factor(educational_level)Completed primary school     2.075e-01  4.148e-01
## factor(educational_level)No formal schooling         -1.331e+01  3.785e+02
## factor(educational_level)Some high/secondary schooling -1.680e-01  1.213e-01
## factor(educational_level)Some primary school          5.915e-01  5.684e-01
## factor(educational_level)Tertiary education          -4.293e-03  1.369e-01
## factor(residential_area)Diepsloot                    -1.484e-01  1.686e-01
## factor(residential_area)Fourways                      2.231e-01  2.840e-01
## factor(residential_area)Kyasand/pipeline             -1.873e-01  2.794e-01
## factor(residential_area)Msawawa                       2.117e-01  3.030e-01
## factor(residential_area)Other                         9.322e-02  1.950e-01
## monthly_income                                        4.870e-06  8.305e-06
##                                                      z value Pr(>|z|)
## (Intercept)                                           -1.309    0.191
## age                                                    0.551    0.581
## factor(educational_level)Completed primary school      0.500    0.617
## factor(educational_level)No formal schooling          -0.035    0.972
## factor(educational_level)Some high/secondary schooling -1.386    0.166
## factor(educational_level)Some primary school           1.041    0.298
## factor(educational_level)Tertiary education           -0.031    0.975
## factor(residential_area)Diepsloot                     -0.880    0.379
## factor(residential_area)Fourways                       0.785    0.432
## factor(residential_area)Kyasand/pipeline              -0.670    0.503
## factor(residential_area)Msawawa                        0.699    0.485
## factor(residential_area)Other                          0.478    0.633
## monthly_income                                         0.586    0.558
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2194.0  on 1666  degrees of freedom
## Residual deviance: 2178.8  on 1654  degrees of freedom
##   (533 observations deleted due to missingness)
## AIC: 2204.8
##
## Number of Fisher Scoring iterations: 12
```

4. Based on the fitted imputation model (for Y), predict the missing y-values, and store the results in a new data frame in R. Prediction of the probability of a zero risk can be performed using the predict-function in R and prediction of y (being equal to zero or one) should be based on an estimated probability of having a zero-risk perception exceeding 0.5.

```
## In this part, we first define a new dataset to store the imputed values
```

```
## Define new imputed dataset
##----------------------------
imp_dat = all_dat

## Single imputation
##------------------
imp_dat$y1_imp = as.numeric(predict(glm_fit, newdata = all_dat,
                                    type = "response") > 0.5)
imp_dat$y1_imp2 = imp_dat$y1_imp

## RESTORE previous observed values (after assessment of prediction error)
##-------------------------------------------------------------------------
imp_dat$y1_imp[is.na(imp_dat$y1) == F] = imp_dat$y1[is.na(imp_dat$y1) == F]
```

5. How would you evaluate the performance of this imputation (or prediction) model?

```
## Prediction error (among observed values)
##-----------------------------------------
table(imp_dat$y1, imp_dat$y1_imp2)
```

```
##
##        0    1
##   0 1049    4
##   1  604   10
```

```
sum(abs(imp_dat$y1 - imp_dat$y1_imp2) == 1, na.rm = T)/sum(!is.na(imp_dat$y1))
```

```
## [1] 0.3647271
```

```
## The prediction error is relatively high (for the cut-off value of 0.5)
## meaning that predicting whether a AGYW has a zero perceived risk solely based
## on the covariates age, educational level residential area and monthly income
## implies a large number of misclassifications. This is especially true given
## that approximately 37% of participants have a perceived risk score of zero,
## hence, classifying everyone as having a non-zero risk score would lead to
## a prediction error of 37%. Consequently, the gain in fitting the logistic
## model is very small. In general, the performance can be improved by using all
## available covariate information in the dataset (cfr. the use of the mice
## package for imputing the variable y) or by considering tree-based methods for
## classification. For now, however, we will work with these predictions as a
## first step in remediating the missingness in the risk score variable Z.
```

6. The procedure in step 4. describes a single imputation process. How would you extend the approach to account for the uncertainty related to the imputation step? More specifically, how could you repeatedly and multiple times impute the missing observations for the random variable Y? Create 10 different imputation sets based on such an approach.

```
## Multiple imputation
##--------------------
M = 10

set.seed(2504)
for (imp_id in 1:M){
    imp_set_name = paste0("imp_dataset",imp_id)
    ynew = rbinom(n, size = 1, prob = predict(glm_fit, newdata = all_dat,
                                              type = "response"))
```

```
    ynew[is.na(all_dat$y1) == F] = all_dat$y1[is.na(all_dat$y1) == F]
    assign(imp_set_name, cbind(all_dat, ynew))
}

## This routine enables the construction of M datasets with complete information
## on the variable Y, through the use of the distributional assumption with
## regard to Y. More specifically, we generate random values from a binomial
## distribution (with repetitions equal to one - Bernoulli distribution) and
## probability of success equal to the estimated probability based on the
## logistic regression model.
```

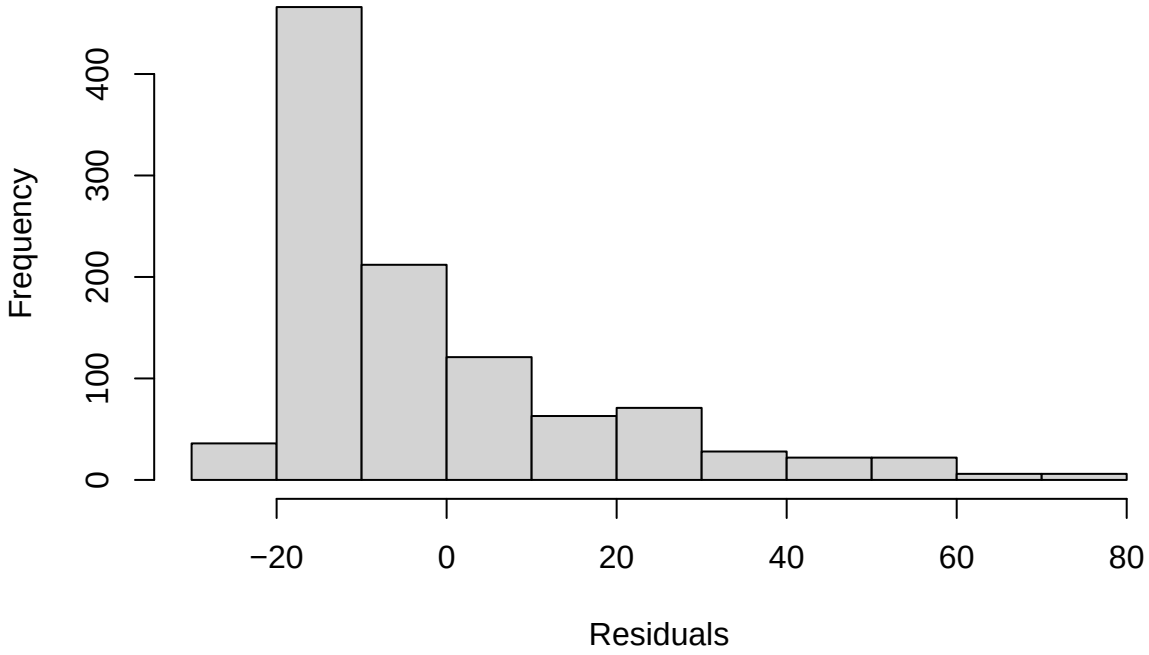
7. If we impute Y = 1, the resulting imputed value for Z is equal to zero. However, how can we impute the missing values for our primary endpoint Z, conditional on Y being equal to 0 (i.e., implying that the risk perception score differs from zero). Formulate an imputation model for Z depending on age, educational level, residential area, and monthly income. What type of model will you consider?

```
## Define data with non-zero risk scores
##--------------------------------------
reduced_dat = all_dat[all_dat$risk_score != 0, ]

## Linear regression models
##-------------------------
lm_fit = lm(risk_score ~ age + factor(educational_level) +
              factor(residential_area) + monthly_income, data = reduced_dat)
summary(lm_fit)
```

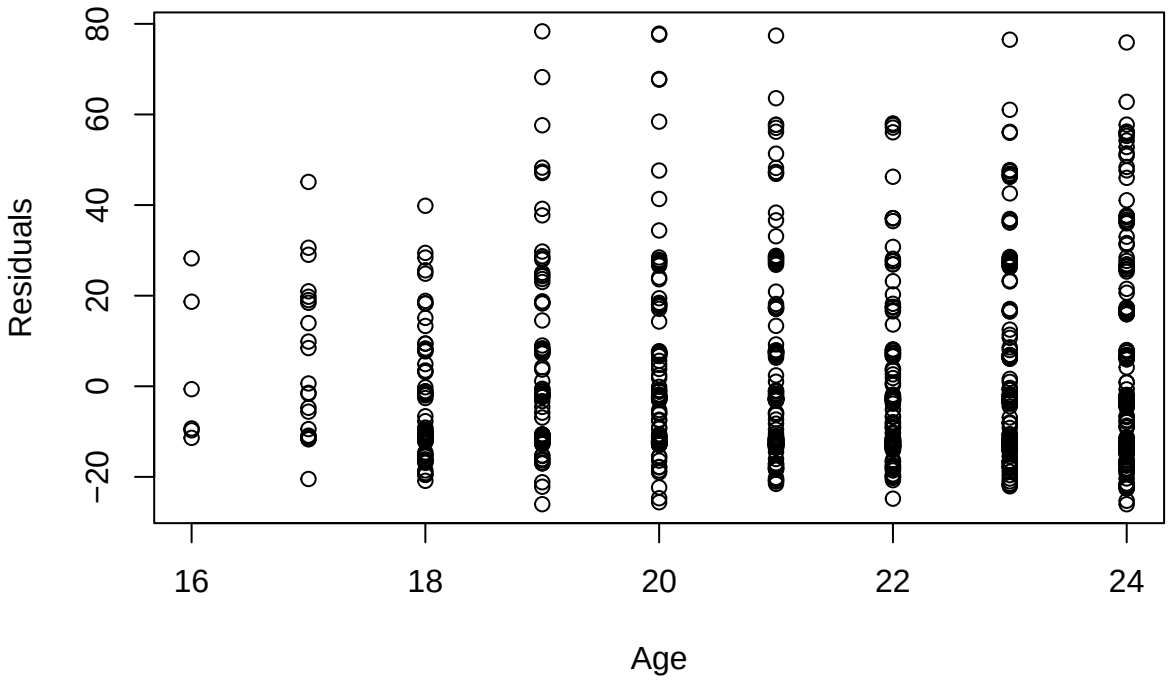
```
##
## Call:
## lm(formula = risk_score ~ age + factor(educational_level) + factor(residential_area) +
##     monthly_income, data = reduced_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.036 -12.966  -7.700   7.324  78.346
##
## Coefficients:
##                                                        Estimate Std. Error
## (Intercept)                                           1.635e+01  6.827e+00
## age                                                   3.276e-01  2.927e-01
## factor(educational_level)Completed primary school    -2.871e+00  5.300e+00
## factor(educational_level)No formal schooling          2.095e+00  1.399e+01
## factor(educational_level)Some high/secondary schooling -2.804e-01  1.404e+00
## factor(educational_level)Some primary school         -4.965e+00  8.139e+00
## factor(educational_level)Tertiary education           5.741e-01  1.648e+00
## factor(residential_area)Diepsloot                     6.840e-02  2.021e+00
## factor(residential_area)Fourways                      6.043e+00  3.596e+00
## factor(residential_area)Kyasand/pipeline              3.546e+00  3.200e+00
## factor(residential_area)Msawawa                       3.548e-01  3.729e+00
## factor(residential_area)Other                        -1.156e+00  2.392e+00
## monthly_income                                       -1.480e-04  9.555e-05
##                                                      t value Pr(>|t|)
## (Intercept)                                            2.395   0.0168 *
## age                                                    1.119   0.2634
## factor(educational_level)Completed primary school     -0.542   0.5882
```

```
## factor(educational_level)No formal schooling             0.150   0.8810
## factor(educational_level)Some high/secondary schooling  -0.200   0.8418
## factor(educational_level)Some primary school            -0.610   0.5420
## factor(educational_level)Tertiary education              0.348   0.7277
## factor(residential_area)Diepsloot                        0.034   0.9730
## factor(residential_area)Fourways                         1.681   0.0932 .
## factor(residential_area)Kyasand/pipeline                 1.108   0.2681
## factor(residential_area)Msawawa                          0.095   0.9242
## factor(residential_area)Other                           -0.483   0.6291
## monthly_income                                          -1.549   0.1217
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.45 on 1040 degrees of freedom
##   (533 observations deleted due to missingness)
## Multiple R-squared:  0.009575,   Adjusted R-squared:  -0.001853
## F-statistic: 0.8379 on 12 and 1040 DF,  p-value: 0.6112
```

```
hist(residuals(lm_fit), main = "", xlab = "Residuals")
```




```
plot(reduced_dat$age[!is.na(reduced_dat$age)], residuals(lm_fit), xlab = "Age",
     ylab = "Residuals")
```

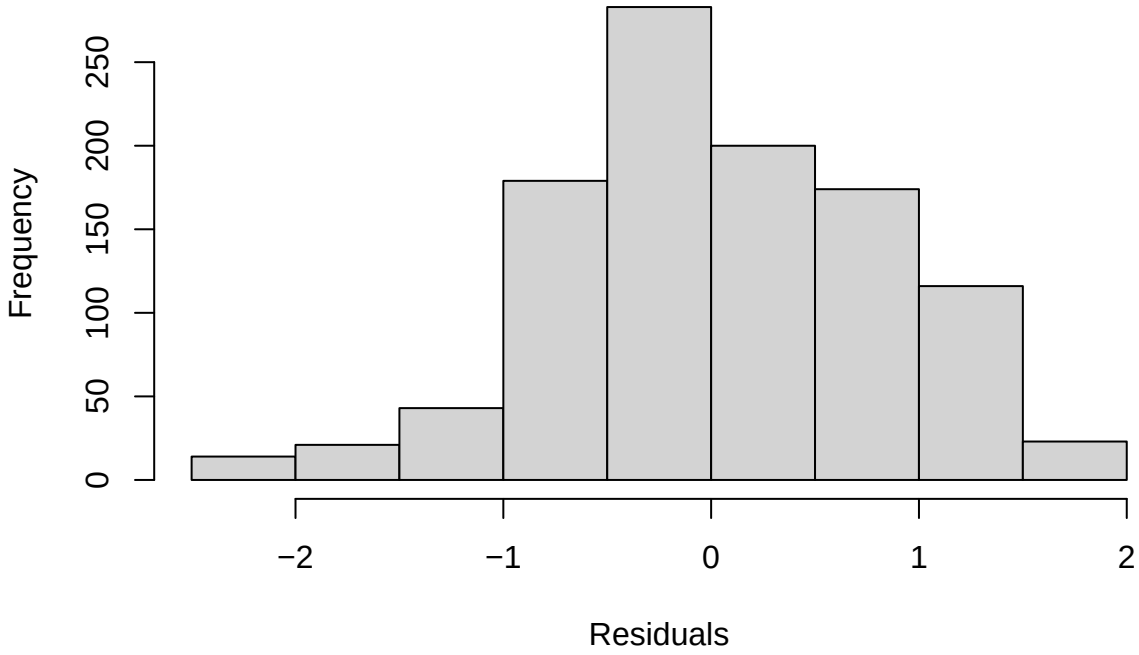



```
lm_fit2 = lm(log(risk_score + 1) ~ age + factor(educational_level) +
               factor(residential_area) + monthly_income, data = reduced_dat)
summary(lm_fit2)
```

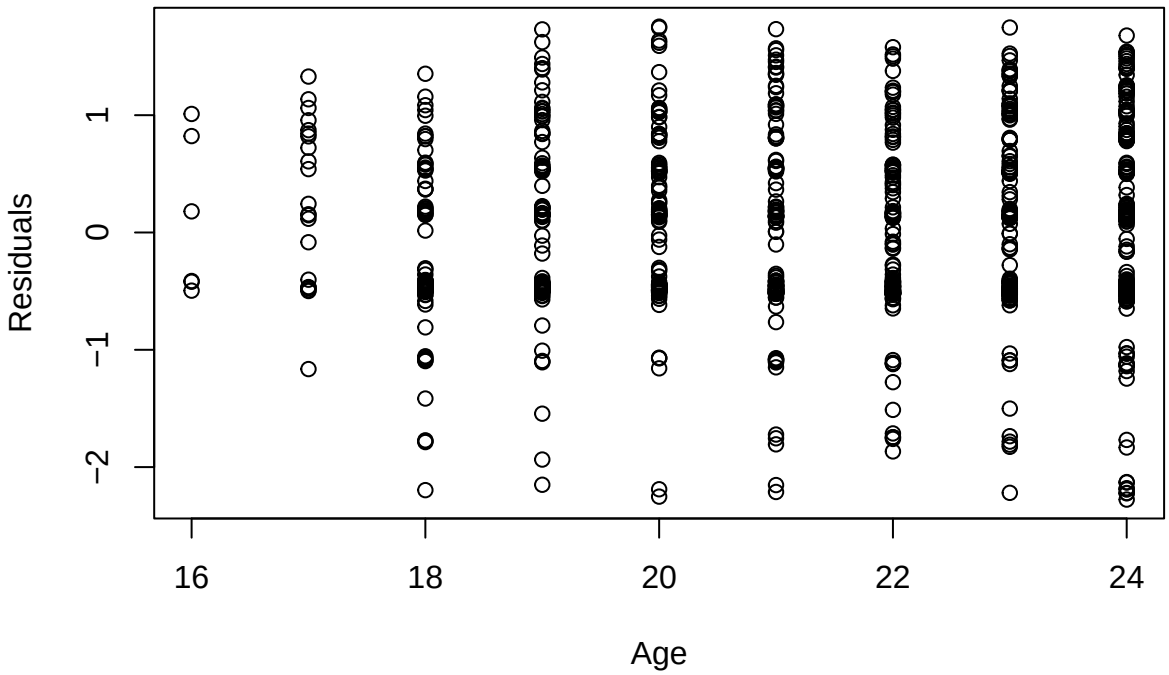
```
##
## Call:
## lm(formula = log(risk_score + 1) ~ age + factor(educational_level) +
##     factor(residential_area) + monthly_income, data = reduced_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2767 -0.4994 -0.1111  0.5400  1.7545
##
## Coefficients:
##                                                      Estimate Std. Error
## (Intercept)                                         2.783e+00  2.687e-01
## age                                                 6.044e-03  1.152e-02
## factor(educational_level)Completed primary school  -1.047e-01  2.086e-01
## factor(educational_level)No formal schooling       -9.645e-02  5.505e-01
## factor(educational_level)Some high/secondary schooling -1.270e-03  5.527e-02
## factor(educational_level)Some primary school       -1.751e-01  3.203e-01
## factor(educational_level)Tertiary education        -2.825e-02  6.486e-02
## factor(residential_area)Diepsloot                   1.517e-02  7.956e-02
## factor(residential_area)Fourways                    1.598e-01  1.415e-01
## factor(residential_area)Kyasand/pipeline            6.064e-02  1.260e-01
## factor(residential_area)Msawawa                    -9.736e-03  1.468e-01
## factor(residential_area)Other                      -4.310e-02  9.412e-02
## monthly_income                                     -5.884e-06  3.761e-06
##                                                     t value Pr(>|t|)
## (Intercept)                                          10.357   <2e-16 ***
## age                                                   0.525    0.600
## factor(educational_level)Completed primary school    -0.502    0.616
```

```
## factor(educational_level)No formal schooling           -0.175    0.861
## factor(educational_level)Some high/secondary schooling -0.023    0.982
## factor(educational_level)Some primary school           -0.547    0.585
## factor(educational_level)Tertiary education            -0.436    0.663
## factor(residential_area)Diepsloot                       0.191    0.849
## factor(residential_area)Fourways                        1.129    0.259
## factor(residential_area)Kyasand/pipeline                0.481    0.630
## factor(residential_area)Msawawa                        -0.066    0.947
## factor(residential_area)Other                          -0.458    0.647
## monthly_income                                         -1.565    0.118
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7655 on 1040 degrees of freedom
##   (533 observations deleted due to missingness)
## Multiple R-squared:  0.005956,   Adjusted R-squared:  -0.005514
## F-statistic: 0.5193 on 12 and 1040 DF,  p-value: 0.9034
```

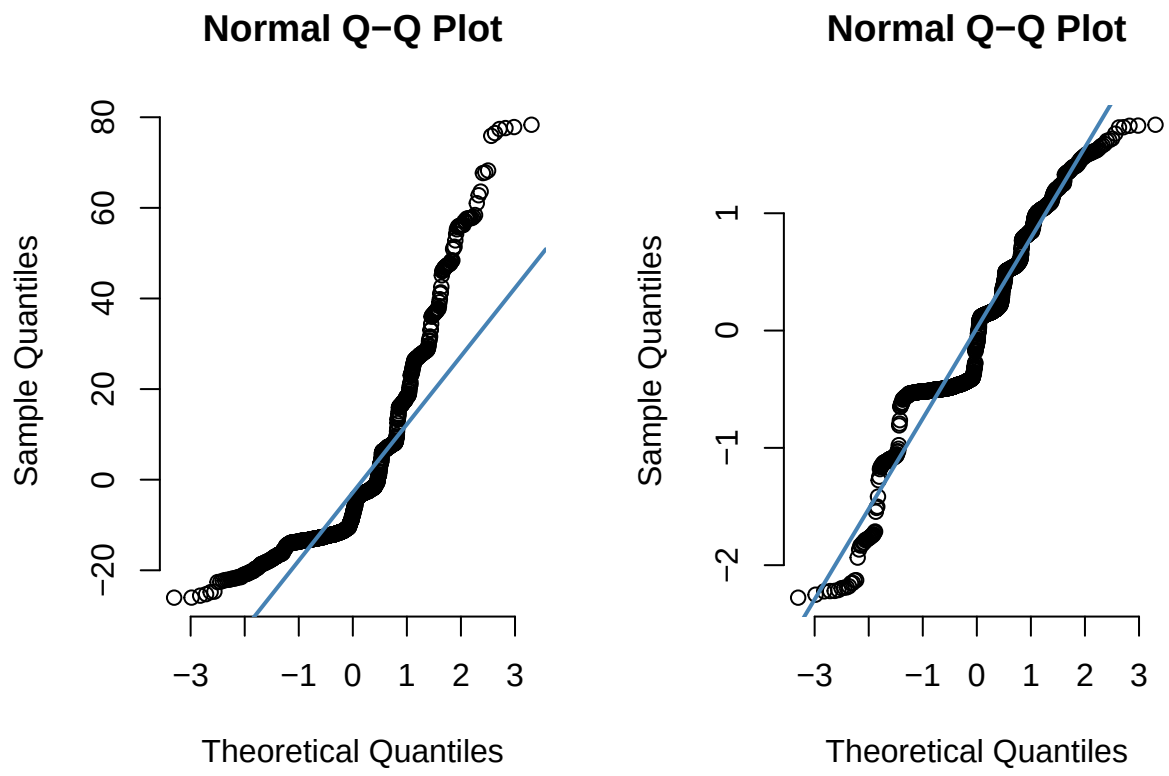
```
hist(residuals(lm_fit2), main = "", xlab = "Residuals")
```




```
plot(reduced_dat$age[!is.na(reduced_dat$age)], residuals(lm_fit2), xlab = "Age",
     ylab = "Residuals")
```




```
## Model diagnostics (normality)
##------------------------------
par(mfrow = c(1,2))
qqnorm(residuals(lm_fit), pch = 1, frame = FALSE)
qqline(residuals(lm_fit), col = "steelblue", lwd = 2)

qqnorm(residuals(lm_fit2), pch = 1, frame = FALSE)
qqline(residuals(lm_fit2), col = "steelblue", lwd = 2)
```



```
## Although the model based on the transformed outcome performs better in terms
## of underlying normality assumption, the presence of digit preference still
## induces deviations from normality as well as problems with the assumption of
## homoscedasticity (constancy of variance of the error terms). Moreover, the
## explanatory power of the 'final model' (lm_fit2) is very low indicating that
## the predictive ability of the set of covariates is low. The problem regarding
## digit preference will not be discussed further in this tutorial, while the
## issue about the predictive power of the imputation model is partly remediated
## by considering the full conditional specification (FCS) approach in the mice
## package. More specifically, the FCS approach will rely on the information
## about all covariates in the construction of the imputation model for the
## primary endpoint.
```

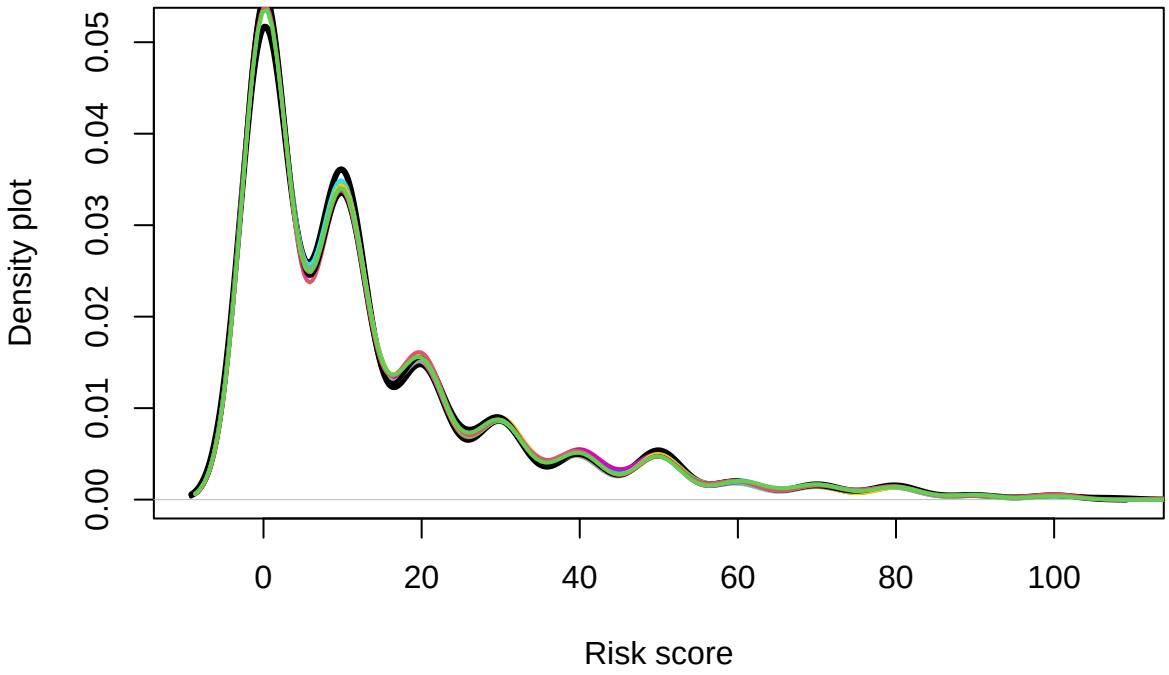
8. Based on the fitted imputation model (for Z, conditional on Y = 0), predict the missing z-values, and store the results in a new data frame in R. Prediction can be done using the predict-function in R.

```
## Imputing values for Z (based on model 2 - with transformation)
##-----------------------------------------------------------------
imp_dat$znew_si = exp(predict(lm_fit2, newdata = all_dat) + rnorm(n, mean = 0,
                                                    sd = sigma(lm_fit2)))-1
imp_dat$znew_si[is.na(all_dat$risk_score) == F] = all_dat$risk_score[is.na(all_dat$risk_score) == F
imp_dat$znew_si[imp_dat$y1_imp == 1] = 0
```

9. The procedure in step 8. (i.e., also referred to as regression imputation) describes a single imputation process. How would you extend the approach to perform multiple imputation?

a. Consider random generation of error terms (random noise);

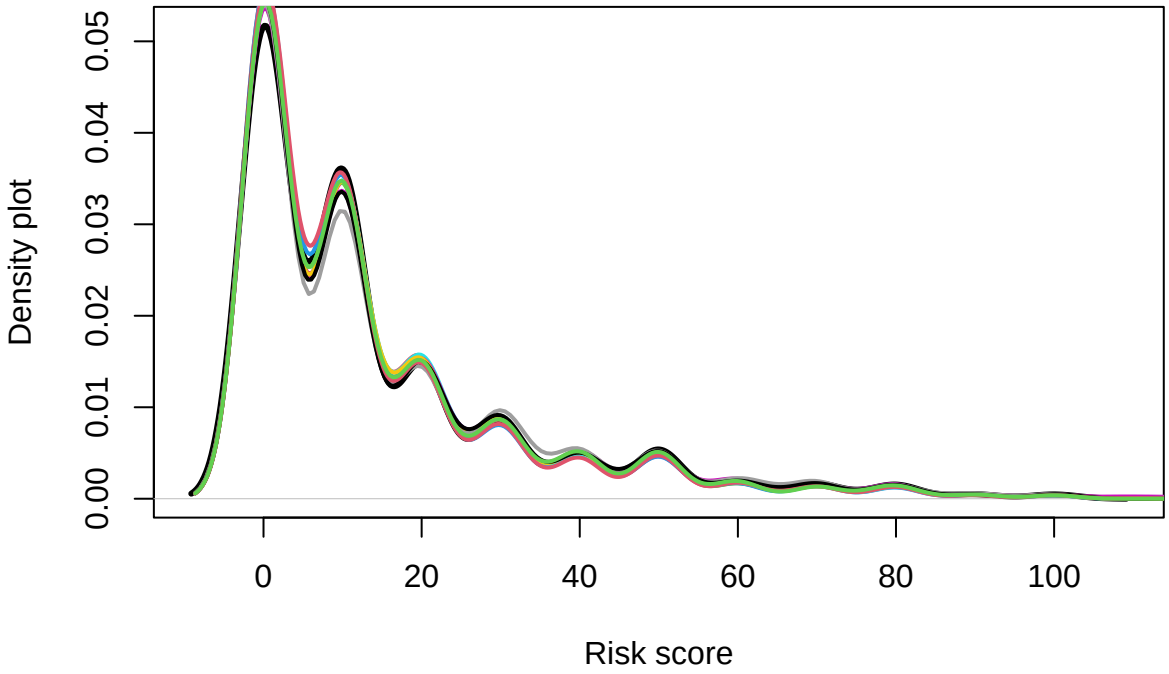
```
## Multiple imputation
##--------------------
M = 10

d_imp = plot(density(all_dat$risk_score, na.rm = T), main = "",
             xlab = "Risk score", ylab = "Density plot", lwd = 3, col = 1)

for (imp_id in 1:M){
  imp_set_old = paste0("imp_dataset",imp_id)
    imp_set_name = paste0("imp_dataset_full",imp_id)

    ## Without transformation
  ##------------------------
    #znew = predict(lm_fit, newdata = all_dat) + rnorm(n, mean = 0,
    #                                                  sd = sigma(lm_fit))

    ## With transformation
  ##---------------------
    znew = exp(predict(lm_fit2, newdata = all_dat) + rnorm(n, mean = 0,
                                                    sd = sigma(lm_fit2)))-1
    znew[is.na(all_dat$risk_score) == F] =
      all_dat$risk_score[is.na(all_dat$risk_score) == F]
    znew[get(imp_set_old)$ynew == 1] = 0

    lines(density(znew), col = imp_id + 1, lwd = 2)
    assign(imp_set_name, cbind(get(imp_set_old), znew))
}
```



b. Random generation of the asymptotic distribution of the estimators of the model parameters in combination with random noise.

```
## Multiple imputation
##---------------------
M = 10

d_imp = plot(density(all_dat$risk_score, na.rm = T), main = "",
             xlab = "Risk score", ylab = "Density plot", lwd = 3, col = 1)

for (imp_id in 1:M){
  set.seed(imp_id)
  imp_set_old = paste0("imp_dataset",imp_id)
  imp_set_name = paste0("imp_dataset_full2",imp_id)

  ## Without transformation
  ##-----------------------
  #sm_values = summary(lm_fit)$coefficients
  #new_coef = rnorm(nrow(sm_values), mean = sm_values[,1], sd = sm_values[,2])
  #new_pred = apply(model.matrix.lm(lm_fit, data = all_dat,
  #                                 na.action = "na.pass")%*%new_coef, 1, sum) +
  #  rnorm(n, mean = 0, sd = sigma(lm_fit))
  #znew = new_pred

  ## With transformation
  ##---------------------
  sm_values = summary(lm_fit2)$coefficients
  new_coef = rnorm(nrow(sm_values), mean = sm_values[,1], sd = sm_values[,2])
  new_pred = apply(model.matrix.lm(lm_fit2, data = all_dat,
                                   na.action = "na.pass")%*%new_coef, 1, sum) +
    rnorm(n, mean = 0, sd = sigma(lm_fit2))

  znew = exp(new_pred)-1
```

```
  znew[is.na(all_dat$risk_score) == F] =
    all_dat$risk_score[is.na(all_dat$risk_score) == F]
    znew[get(imp_set_old)$ynew == 1] = 0

    lines(density(znew), col = imp_id + 1, lwd = 2)
  assign(imp_set_name, cbind(get(imp_set_old), znew))
}
```




```
### The difference between option a.) and option b.) is that in the latter the
### uncertainty in terms of the imputation model, especially with regard to the
### estimation of the model parameters, is fully accounted for in the imputation
### process. Under option a.) only stochastic variation as a result of
### measurement error is accommodated, while uncertainty with respect to the
### estimates of the model parameters is ignored. Therefore, option b.) is
### preferred, which is also translated to MICE, at least in a Bayesian context.
```

10. Construct M = 10 imputation sets and estimate whether there exists a significant difference in average baseline risk perception score among AGYW living in Cosmocity versus those living in Fourways. How would you combine the results of the different analyses?

```
## Here we demonstrate the use of Rubin's rules for the logistic model modelling
## the probability of having a zero perceived risk in relation to the area in
## which AGYW are living (i.e., Cosmocity vs. Fourways). An analysis focusing
## on the average baseline risk perception score, conditional on the score being
## larger than zero, can be performed in a similar way. Moreover, we rely on the
## imputed datasets according to the regression imputation approach including
## variability with regard to the model estimates (cfr. Question 9b.)

## MI analyses
##-------------
coef_imp = vector()
se_imp = vector()
```

```r
for (imp_id in 1:M){

  imp_set_name = paste0("imp_dataset_full2",imp_id)

  glm_fit_imp = glm(ynew ~ age + factor(educational_level) +
                      factor(residential_area) + monthly_income,
                    data = get(imp_set_name), family = binomial(link = "logit"))
  sm_glm_imp = summary(glm_fit_imp)$coefficients
  cov_id = which(rownames(sm_glm_imp) == "factor(residential_area)Fourways")
  coef_imp[imp_id] = sm_glm_imp[cov_id,1]
  se_imp[imp_id] = sm_glm_imp[cov_id,2]
}

mean_estimate = mean(coef_imp); mean_estimate;
```

```
## [1] 0.2489494
```

```r
between_var = var(coef_imp); between_var;
```

```
## [1] 0.008874793
```

```r
within_var = mean(se_imp**2); within_var;
```

```
## [1] 0.07140182
```

```r
total_var = within_var + (1 + 1/M)*between_var; total_var;
```

```
## [1] 0.08116409
```

```r
### Odds ratio estimation
###-----------------------
OR = exp(mean_estimate); OR;
```

```
## [1] 1.282677
```

```r
### Hypothesis problem H0: beta = 0; H1: beta != 0
###-----------------------------------------------
wald_test_statistic = mean_estimate/sqrt(total_var);
wald_test_statistic;
```

```
## [1] 0.8738344
```

```r
n = nrow(imp_dataset_full21)
k = nrow(sm_glm_imp)
lambda = (between_var + (between_var/M))/total_var
df_old = (M - 1)/(lambda^2)
df_obs = (((n - k) + 1)/((n - k) + 3)) * ((n - k)*(1 - lambda))
df_adj = (df_old*df_obs)/(df_old + df_obs)

## Inference is based on the univariate Wald test (Rubin, 1987; Van Buuren,
## 2018); Marshall et al., 2009) with two distinct ways to calculate the
## degrees of freedom for the Wald statistic following a t-distribution under
## the null hypothesis.

### Calculation of the two-sided p-value
###-------------------------------------
p_old = 2*pt(wald_test_statistic, df = df_old, lower.tail = FALSE)
p_adj = 2*pt(wald_test_statistic, df = df_adj, lower.tail = FALSE)
```

```
p_old; p_adj;
```

```
## [1] 0.3825456
```

```
## [1] 0.3826547
```

```
### 95% CI for the odds ratio
###---------------------------
logOR = mean_estimate;
alpha = 0.05
df = df_adj
cv = qt(1-alpha/2, df = df)
ll_logOR = logOR - cv*sqrt(total_var)
ul_logOR = logOR + cv*sqrt(total_var)
ll_OR = exp(ll_logOR)
ul_OR = exp(ul_logOR)

print(c(OR, ll_OR, ul_OR));
```

```
## [1] 1.2826771 0.7328073 2.2451477
```

```
## Based on the unadjusted (old) and adjusted p-value (< 0.05), we do not reject
## the null hypothesis (beta = 0) at 5% s.l. implying that there does not exist
## a significant difference in probability of reporting zero perceived risk
## between Fourways and Cosmocity AGYW (reference category). Alternatively,
## the pooled odds ratio (after multiple imputation) is found to be
## NOT significantly different from one given that the 95% CI for the pooled OR
## contains value one.

## NOTE: the use of Rubin's rules is based on the assumption of asymptotic
## normality of the quantity of interest. If this assumption is violated one
## can not rely on the Wald test statistic (and corresponding t-distribution).

### In order to ensure that the assumption of asymptotic normality holds for the
### odds ratio, which by definition takes values between 0 and infinity, the
### application of Rubin's rule is done at the log-transformed scale, i.e., for
### the log(OR). Given that this is the respective parameter in a logistic
### regression model, estimates of the standard error at the transformed scale
### are readily available.

### Now we calculate the pooled difference in mean perceived risk score between
### Cosmocity and Fourways, conditional on having a non-zero perceived risk
### score. More specifically, among AGYW which have a perceived risk that is
### higher than zero, is there a significant difference in average risk score
### between the two areas of interest. In order to do so, we use Rubin's rules
### directly, relying on asymptotic normality of the estimated conditional
### average risk score. Therefore, we should consider a subset of the full
### data, including only those observations with a perceived risk score larger
### than zero.

## MI analyses
##-------------
coef_imp = vector()
se_imp = vector()
```

```r
for (imp_id in 1:M){

  imp_set_name = paste0("imp_dataset_full2",imp_id)
  dat_used = get(imp_set_name)

  glm_fit_imp = glm(log(znew + 1) ~ age + factor(educational_level) +
                      factor(residential_area) + monthly_income,
                    data = dat_used[dat_used$znew > 0, ],
                    family = gaussian(link = "identity"))
  sm_glm_imp = summary(glm_fit_imp)$coefficients
  cov_id = which(rownames(sm_glm_imp) == "factor(residential_area)Fourways")
  coef_imp[imp_id] = sm_glm_imp[cov_id,1]
  se_imp[imp_id] = sm_glm_imp[cov_id,2]
}

mean_estimate = mean(coef_imp); mean_estimate;
```

```
## [1] 0.1417174
```

```r
between_var = var(coef_imp); between_var;
```

```
## [1] 0.001626971
```

```r
within_var = mean(se_imp**2); within_var;
```

```
## [1] 0.01816553
```

```r
total_var = within_var + (1 + 1/M)*between_var; total_var;
```

```
## [1] 0.0199552
```

```r
### Difference of mean transformed risk score
###--------------------------------------------
mean_estimate;
```

```
## [1] 0.1417174
```

```r
### Hypothesis problem H0: beta = 0; H1: beta != 0
###-------------------------------------------------
wald_test_statistic = mean_estimate/sqrt(total_var);
wald_test_statistic;
```

```
## [1] 1.003217
```

```r
n = nrow(dat_used[dat_used$znew > 0, ])
k = nrow(sm_glm_imp)
lambda = (between_var + (between_var/M))/total_var
df_old = (M - 1)/(lambda^2)
df_obs = (((n - k) + 1)/((n - k) + 3)) * ((n - k)*(1 - lambda))
df_adj = (df_old*df_obs)/(df_old + df_obs)

## Inference is based on the univariate Wald test (Rubin, 1987; Van Buuren,
## 2018); Marshall et al., 2009) with two distinct ways to calculate the
## degrees of freedom for the Wald statistic following a t-distribution under
## the null hypothesis.

### Calculation of the two-sided p-value
###-------------------------------------
```

```r
p_old = 2*pt(wald_test_statistic, df = df_old, lower.tail = FALSE)
p_adj = 2*pt(wald_test_statistic, df = df_adj, lower.tail = FALSE)
p_old; p_adj;
```

```
## [1] 0.3159728
```

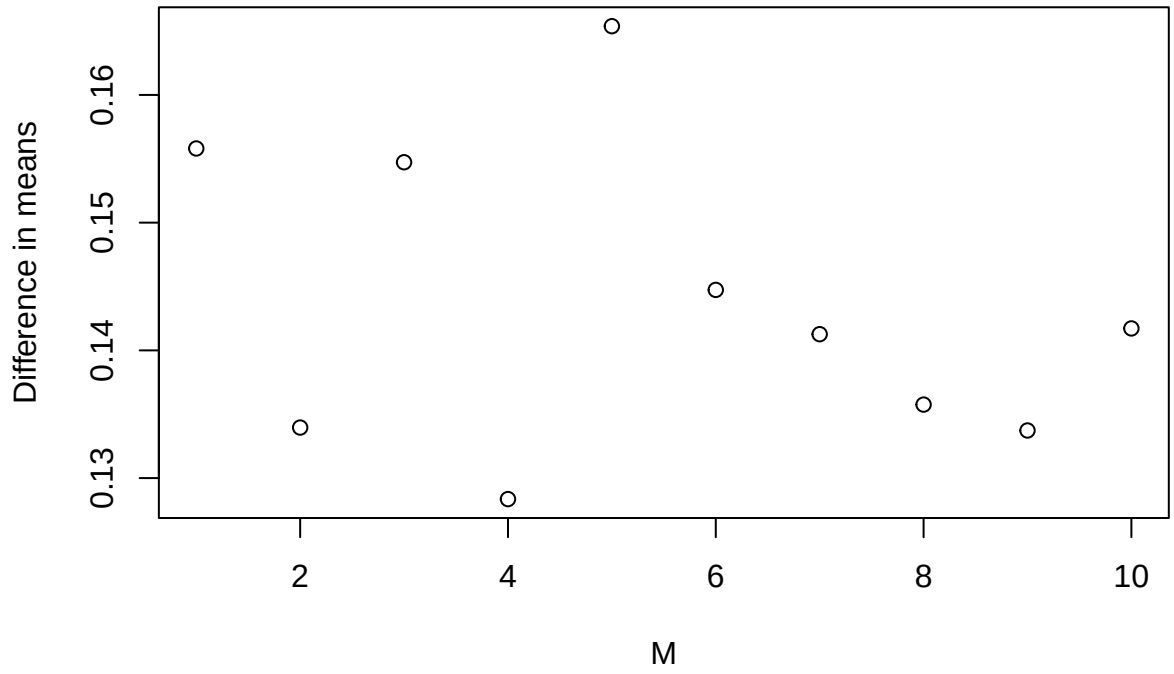
```
## [1] 0.3161652
```

```r
## Based on the unadjusted (old) and adjusted p-value (< 0.05), we do not reject
## the null hypothesis (beta = 0) at 5% s.l. implying that there does not exist
## a significant difference in (transformed) perceived risk between Fourways and
## Cosmocity AGYW (reference category), while correcting for the other variables
## in the final regression model.
```

11. Check whether the number of imputations is sufficient. More specifically, create a graph in which you assess the convergence of the pooled difference in average baseline risk perception score among AGYW living in Cosmocity versus those living in Fourways depending on the number of imputations.

```r
mean_estimate_vec = vector()
total_var_vec = vector()

for (k in 1:10){
  set.seed(k)
  select_imp = sample(1:10, size = k, replace = F)
  mean_estimate_vec[k] = mean(coef_imp[select_imp]);
  between_var = var(coef_imp[select_imp]);
  within_var = mean(se_imp[select_imp]**2);
  total_var_vec[k] = within_var + (1 + 1/M)*between_var;
}

plot(1:10, mean_estimate_vec, xlab = "M", ylab = "Difference in means")

## Increasing the number of imputations leads to a convergence in terms of the
## quantity you are investigating. Based on this plot, however, I would perform
## slightly more imputations to make sure that the estimated difference in means
## has converged to a stable value.

## Repeat the imputation process adding 10 more imputations
##-----------------------------------------------------------
for (imp_id in 11:(2*M)){
  set.seed(imp_id)

  imp_set_name = paste0("imp_dataset",imp_id)
    ynew = rbinom(n, size = 1, prob = predict(glm_fit, newdata = all_dat,
                                              type = "response"))
    ynew[is.na(all_dat$y1) == F] = all_dat$y1[is.na(all_dat$y1) == F]
    assign(imp_set_name, cbind(all_dat, ynew))

  imp_set_old = paste0("imp_dataset",imp_id)
  imp_set_name = paste0("imp_dataset_full2",imp_id)

  ## With transformation
  ##--------------------
  sm_values = summary(lm_fit2)$coefficients
  new_coef = rnorm(nrow(sm_values), mean = sm_values[,1], sd = sm_values[,2])
```

```
  new_pred = apply(model.matrix.lm(lm_fit2, data = all_dat,
                                   na.action = "na.pass")%*%new_coef, 1, sum) +
    rnorm(n, mean = 0, sd = sigma(lm_fit2))

  znew = exp(new_pred)-1
  znew[is.na(all_dat$risk_score) == F] =
    all_dat$risk_score[is.na(all_dat$risk_score) == F]
    znew[get(imp_set_old)$ynew == 1] = 0

    lines(density(znew), col = imp_id + 1, lwd = 2)
  assign(imp_set_name, cbind(get(imp_set_old), znew))
}
```




```
for (imp_id in 11:(2*M)){

  imp_set_name = paste0("imp_dataset_full2",imp_id)
  dat_used = get(imp_set_name)

  glm_fit_imp = glm(log(znew + 1) ~ age + factor(educational_level) +
                      factor(residential_area) + monthly_income,
                    data = dat_used[dat_used$znew > 0, ],
                    family = gaussian(link = "identity"))
  sm_glm_imp = summary(glm_fit_imp)$coefficients
  cov_id = which(rownames(sm_glm_imp) == "factor(residential_area)Fourways")
  coef_imp[imp_id] = sm_glm_imp[cov_id,1]
  se_imp[imp_id] = sm_glm_imp[cov_id,2]
}

mean_estimate = mean(coef_imp); mean_estimate;
```

```
## [1] 0.1561229
```

```
between_var = var(coef_imp); between_var;
```

```
## [1] 0.002110888
```

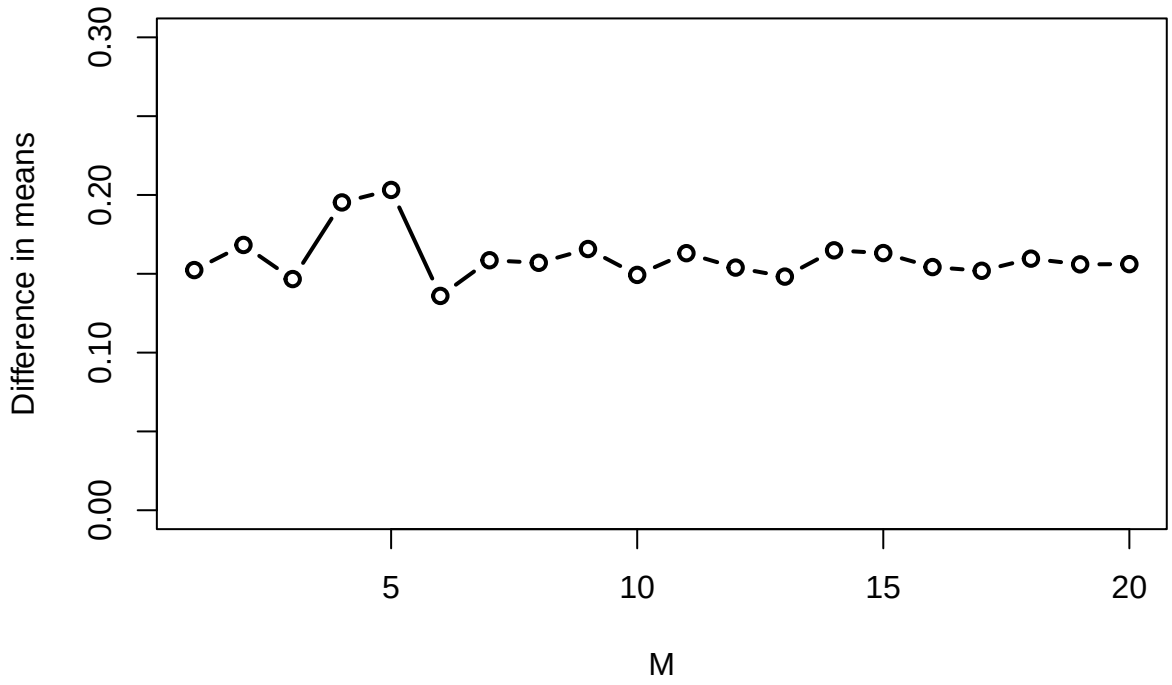
```
within_var = mean(se_imp**2); within_var;
```

```
## [1] 0.01806747
```

```
total_var = within_var + (1 + 1/M)*between_var; total_var;
```

```
## [1] 0.02038944
```

```
mean_estimate_vec = vector()
total_var_vec = vector()

for (k in 1:(2*M)){
  set.seed(k)
  select_imp = sample(1:20, size = k, replace = F)
  mean_estimate_vec[k] = mean(coef_imp[select_imp]);
  between_var = var(coef_imp[select_imp]);
  within_var = mean(se_imp[select_imp]**2);
  total_var_vec[k] = within_var + (1 + 1/M)*between_var;
}

plot(1:20, mean_estimate_vec, xlab = "M", ylab = "Difference in means",
     ylim = c(0,0.3), type = "b", lwd = 2)
```




```
## From the graph it is clear that from M = 10 imputations onwards the estimated
## difference in means is not changing dramatically, and especially near 20
## imputations convergence seems to be achieved.

## NOTE: data points are not the same on both graphs (i.e., M = 10 to 20) given
## that different random selections of imputation-specific estimates are
## combined in the pooling process.
```

12. The MICE package in R provides an automated way of performing Multiple Imputation with Chained Equations (Van Buuren, 2011). The MICE approach is also referred to as Full Conditional Specification, meaning that available information on all variables, except for the one to be imputed, will be used in the imputation model to impute missing values. Use the MICE package to perform multiple imputation for the DREAMS dataset at hand.

```
## Based on the previous exercise, we conclude that an imputation strategy is
## already cumbersome for a single variable (here risk_score). The MICE package
## enables an automated way of performing multiple imputation. Note, however,
## that the use of the MICE package (and corresponding function) to perform
## multiple imputation is by no means a way of avoiding to check the underlying
## imputation model(s) with regard to their fit to the observed data.

## MICE is based on Full Conditional Specification meaning that the default
## setting is to use all available information (read for all variables) to
## impute missing observations in a sequential manner. More specifically, the
## algorithm starts with imputing a specific variable based on the available
## information regarding all other variables in the dataset and subsequently
## performs imputation on the 'next' variable until a complete dataset is
## retained. For specific details concerning the MICE algorithm, please have a
## look at the MICE documentation or the book by Van Buuren (Van Buuren, 2011).

## Hereunder, an easy example concerning the use of the MICE package is given.
## Note that we start here from a subset of the entire DREAMS dataset, which is
## strictly speaking not necessary, but for illustration purposes we consider
## this more convenient. As mentioned previously, the imputation model for the
## perceived risk score still suffers from the fact that the data are subject to
## digit preference and the lack of predictive power for the subset of covariates
## considered in the imputation model. Under FCS, we are able to improve the
## prediction model, at least under the assumption of data being missing at
## random.

## Define the dataset to be used
##-------------------------------
mice_dat = subset(all_dat, select = c("age", "educational_level",
                                      "residential_area", "household_members",
                                      "household_water", "electricity_at_home",
                                      "monthly_income","alcohol_use","sti",
                                      "discuss_health_issue","hiv_talk",
                                      "ever_tested_for_hiv","sexual_partners",
                                      "physical_violence","preventing_pregnancy",
                                      "risk_score"))

## Although the imputation is done directly at the level of the risk score, an
## alternative approach could be to impute Y first (as defined above), and
## consequently impute Z using the mice procedure. By default, mice uses
## predictive mean matching for numeric data, logistic regression imputation
## for binary data, polytomous regression imputation for unordered categorical
## data and proportional odds ordinal logistic regression for ordered
## categorical data. Consequently, for the (pseudo-)continuous risk score, the
## pmm method provides imputations in line with the observed data within a
## neighborhood of specific (missing) risk score observations, thereby
## accounting for the likelihood of imputating zero values.
```

```r
## MICE package
##--------------
suppressPackageStartupMessages(library(mice))

## Perform multiple imputation using the mice function
##----------------------------------------------------
imp = mice(mice_dat, m = 10, maxit = 10, printFlag = FALSE, seed = 123)

## Here we use the default method for imputation depending on the type of
## variable in the dataset. By default, the method uses (1) pmm, predictive mean
## matching for numeric data, (2) logreg, logistic regression imputation for
## binary data, (3) polyreg, polytomous regression imputation for unordered
## categorical data (factor > 2 levels) and (4) polr, proportional odds model
## for (ordered, > 2 levels).

## NOTE: the default method is not necessarily the best option, please verify
## the underlying assumptions of each of these models in the context of the
## different variables that require imputation.

## Analyse the imputed dataset
##----------------------------
m1 <- with(data = imp, expr = lm(log(risk_score + 1) ~ age +
                                   factor(educational_level) +
                                   factor(residential_area) +
                                   factor(household_water) +
                                   monthly_income +
                                   factor(alcohol_use) +
                                   sti +
                                   hiv_talk +
                                   ever_tested_for_hiv +
                                   sexual_partners +
                                   physical_violence +
                                   preventing_pregnancy))
pooled_analysis <- pool(m1)

## Difference in mean perceived risk score for AGYW living in Fourways versus
## Cosmocity (based on the univariate Wald test discussed above)
##-------------------------------------------------------------------------
pooled_analysis$pooled$estimate[9]
```

```
## [1] -0.1540071
```

```r
pvalue = 2*pt(pooled_analysis$pooled$t[9], df = pooled_analysis$pooled$df[9],
              lower.tail = FALSE)
pvalue
```

```
## [1] 0.9402899
```

```r
## Hence, based on the Wald test, there does not exist a significant difference
## in mean perceived risk score for AGYW living in Fourways versus those living
## in Cosmocity. NOTE: We rely here on the assumption of normality of the
## outcome variable which was pointed out to be questionable, even when
## transforming the risk score, due to digit preference and zero-inflation. As
## an alternative one could opt to either model the data using a two component
## mixture approach thereby conditioning on the risk score being equal to zero
```

```
## or not (see description above) or use more advanced statistical methods such
## as composite link models to remediate the digit preference and zero-inflation
## (to some extent).
```

13. Explain the difference between MICE and Multivariate Normal Imputation? Which method is to be preferred?
14. What is the difference between predictive mean matching (pmm) and regression imputation considered in the previous steps? What are the advantages and disadvantages of both approaches?