Predicting the Risk of Osteoporotic Fracture

Aigerim Kulzhabayeva
Department of Mathematics and Statistics York University

Introduction

This report outlines the analysis of The National Health and Nutrition Examination Survey (NHANES) data, collected between January 2007 and December 2008. NHANES is an ongoing program created to assess the health and nutritional status of adults and children in the United States. NHANES is run by National Center for Health Statistics (NCHS) which itself is part of the Centers for Disease Control and Prevention. The data set is obtained from a survey of about 5000 individuals, which are sampled annually from various counties across the US. For logistical reasons a sample of 15 counties are visited each year. The survey is a multistage cluster sampling design and consists of interviews and physical examinations. The interview portion includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel.

In this analysis we will focus on a subset of the main data set, selected with the goal to find the risk factors associated with osteoporotic fracture in the US population. Osteoporosis is a medical condition that causes bones to become thin and porous, decreasing bone strength and leading to increased risk of bone fracture. The most common sites of osteoporotic fracture are the wrist, spine and hip. The cause for osteoporosis has not been identified, however several factors aside from BMD factors have been linked to higher chance of developing osteoporosis. The objectives of this analysis are:

1. Identify predictors of osteoporotic fracture: fractures of the spine, wrist, or a hip.
2. To determine which bone mineral density (BMD) measures are the best predictors of osteoporotic fracture after controlling for various non-BMD risk factors.

We conduct the analysis using GLM without survey information as well as GLM which includes the information for the desing of the survey for comparison. We run both models first on the dataset with only complete observations and next on the data set will all missing values imputed using the missForest package. As a result of the analysis we have found that the most important variables to predict the risk of osteoporosis are:

1. Measure on the Ward's triangle
2. Self reported weight in the last 10 years
3. Having more than 12 alcoholic beverages in the last year. (Being more than occasional drinker)

4. Being of non-Hispanic white Americans ethnicity
5. Having arthritis
6. Use of steroids
7. Gender, to a lesser degree improves the overall model.

Additional interesting observation is that BMI seems to be insignificant variable in the models, however persons height and weight separately seem to be impacting the response. This seems logical, since the taller or heavier the person is, the more pressure is exerted on the bones. Where as BMI is just a ratio of weight to height, which can be low for a very tall and heavy individuals. This is a useful observation to the examiner who may be tempted to only collect BMI information on the individual.

The following are the variables in the data set.

TABLE 1. *Description of the Data set*

Variable	*Description*	*Coding: ref = refused, DK = I don't know*
SEQN	ID assigned to each person	num
RIAGENDR	Gender	1=male; 2=female; .
RIDAGEYR	Age (in years)	num , max=80; .
RIDRETH1	Race/ethnicity	1=Mexican American; 2=Other Hispanic; 3=Non-Hispanic white; 4=Non-Hispanic black; 5=other race; .
DXX0FBMD	Total femur BMD (g/cm2)	num;
DXXNKBMD	Femoral neck BMD (g/cm2)	num;
DXXTRBMD	Trochanter BMD (g/cm2)	num;
DXXINBMD	Intertrochanter BMD (g/cm2)	num;
DXXWDBMD	Ward's triangle BMD (g/cm2)	num;
DXXL1BMD	L1 BMD (g/cm2)	num;
DXXL2BMD	L2 BMD (g/cm2)	num;
DXXL3BMD	L3 BMD (g/cm2)	num;
DXXL4BMD	L4 BMD (g/cm2)	num;
DXXOSBMD	Total spine BMD (g/cm2)	num;
OSQ170	Did mother ever fracture a hip	1=yes; 2=no; 7=ref; 9=DK; .
OSQ200	Did father ever fracture a hip	1=yes; 2=no; 7=ref; 9=DK; .
DIQ010	Doctor told you have diabetes	1=yes; 2=no; 3=borderline; 7=ref; 9=DK;
DID040	Age when first told you have diabetes	num from 1 to 80; 666=less than 1 year; 777=ref; 999=DK;
DIQ220	When was diabetes diagnosed	0 = 3 months ago; 2 = 3-6 months ago; 3 = 6-9 months ago; 4 = 9-12 months ago, 5 = >12 months ago;7=ref;9=DK;
MCQ160A	Doctor ever said you have arthritis	1=yes; 2=no; 77777=ref; 99999=DK;
MCQ180A	Age when told you had arthritis	num 1-80; 77777=ref; 99999=DK;
MCQ190	Type of arthritis	1=rheumatoid arthritis; 2=osteoarthritis; 3=other arthritis; 7=ref; 9=DK; .
MCQ160C	Doctor said you had heart disease	1=yes; 2=no; 7=ref; 9=DK;
MCQ180C	Age when told you had heart disease	num 1-80; 77777=ref; 99999=DK;
MCQ160L	Told you had a liver condition?	1=yes; 2=no; 77777=ref; 99999=DK;

TABLE 2. *Description of the Data set*

Variable	*Description*	
MCQ170L	Do you still have a liver condition	1=yes; 2=no; 77777=ref; 99999=DK;
MCQ180L	Age when told you had a liver condition	num 1 - 80; 77777=ref; 99999=DK;
OSQ130	Ever taken prednisone or cortisone every	1=yes; 2=no; 7=ref; 9=DK; day for a month or longer
SMQ020	Smoked at least 100 cigarettes in life	
SMQ040	Do you now smoke?	
ALQ101	Had > 12 alcoholic drinks last year	1=yes; 2=no; 7=ref; 9=DK
ALQ130	Average of alcohol drinks/day last year	num , max=95; 777=ref; 999=DK;
ALQ140Q	Number of days having 5+ drinks last year	num , max=365; 777=ref; 999=DK;
DBQ197	Past 30 day milk consumption	0=never; 1=<once a week; 2= >once a week, but < once a day; 3 = once a day; 4 = varied; 7 = ref; 9 = DK;
DBQ229	Regular milk use 5 times per week	1= milk drinker for most or all of my life, including childhood; 2= never been a regular milk drinker; 3= varied, sometimes a regular milk drinker; 7=ref; 9=DK;
OSQ010A	Broken or fractured hip	1=yes; 2=no; 7=ref; 9=DK
OSQ010B	Broken or fractured wrist	1=yes; 2=no; 7=ref; 9=DK
OSQ010C	Broken or fractured spine	1=yes; 2=no; 7=ref; 9=DK
OSQ020A	Number of times broken/fractured hip	num, max=5; 7777=ref; 9999=DK;
OSQ020B	Number of times broken/fractured wrist	num , max=7; 7777=ref; 9999=DK;
OSQ020C	Number of times broken/fractured spine	num , max=5; 7777=ref; 9999=DK;
OSQ40AA	Under/over 50 fractured hip first time	1=under 50; 2=>50; 7=ref; 9=DK;
OSQ40BA	Under/over 50 fractured wrist first time	1=under 50; 2=>50; 7=ref; 9=DK;
OSQ40CA	Under/over 50 fractured spine first time	1=under 50; 2=>50; 7=ref; 9=DK;
OSQ070	Ever treated for osteoporosis	1=yes; 2=no; 7=ref; 9=DK;
BMXBMI	Body mass index	num;
WHD020	Current self-report weight (pounds)	num; 7777=ref; 9999=DK; .
WHD110	Self-reported weight 10 years ago	num; 7777=ref; 9999=DK; .
WHD140	Greatest self-reported weight;	num; 7777=ref; 9999=DK; .
WTMEC2YR	Full 2 year sample weight for individuals with a medical examination	num
SDMVPSU	Masked variance pseudo PSU variable for variance estimation	num
SDMVSTRA	Masked variance unit pseudo-stratum variable for variance estimation	num

Notes: Data from https://ssc.ca/en/case-study/risk-factors-osteoporotic-fracture-national-health-and-nutrition-examination-survey

Survey Design

When we are dealing with a survey data, information on survey design plays a significant part in removing the bias and adjustment of the variances for proper inference. The NHANES samples are not generated using a simple random samples,

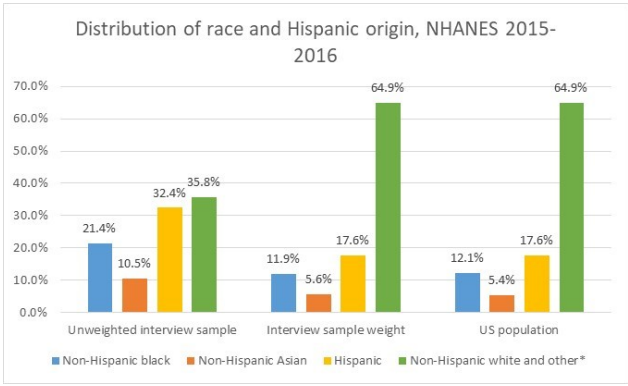
but rather use a complex, multistage, probability sampling design to sample the civilian noninstitutionalized U.S. population. The stages of sample selection areas follows:
The NHANES samples are not simple random samples

1. First, select primary Sampling Units (PSUs). PSUs are counties or small groups of counties which are selected with probability proportional to a measure of size (PPS). Usually 15 PSUs are selceted in any given year;
2. Then PSUs are segmented into clusters of households. Sample segments are selected with PPS;
3. Households within segments are randomly selected; and
4. One or more participants within households are chosen at random.

The survey consists of an interview and a health examination administered by a highly trained medical personnel. Interviews are usually conducted in interviewee's house and includes demographic, socioeconomic, dietary, and health-related questions. Health measurements are performed in specially-designed and equipped mobile centers, called MACs, which travel to various locations throughout the US. The medical examination consists of medical, dental, and physiological measurements, as well as laboratory tests. The sample for the survey is selected to represent the U.S. population of all ages. To produce reliable statistics for underrepresented sub-populations, NHANES over-samples persons 60 and older, African Americans, and Hispanics.

Weighting

Since the aim of the analysis on the collected data is to generalize to US population, which means that the proportion of individuals from different sub-samples, for example people from different racial backgrounds, has to be proportional to the US population. However, this sometimes results in having inadequate amounts of data on minority sub-population of interest. Solution to this problem is to use sample weighting. Simply put a sample weight is a measure of the number of people in the population represented by that sample person. Weighting helps to remove bias form the population estimates. For example, lets look at the graph from the NHANES website:



If we were to estimate mean blood pressure in the US population and people of Hispanic origin would on average have higher blood pressure, the estimation of the population mean would be biased. Thus, the sampling design needs to be accounted for during the analysis. Sample weights and can be found using the formula below:

$$\text{Weight of the sampled person} = \frac{1}{\text{Probability of selection}} \tag{1}$$

where probability of selection for the NHANES sample is given by:

$$\begin{aligned}\text{Probabiity of selection} = {} & P(\text{PSU is selected}) * P(\text{cluster of the PSU selected}) \\ & * P(\text{household selected}) * P(\text{individual is selected})\end{aligned} \tag{2}$$

Further, these estimates need to be adjusted for non-response and post stratification adjustments.

Adjustment for nonresponse to NHANES subsample components

An important mechanism of missingness which we will encounter in the following section is "missing by design", which pertains to the missing data which was dictated by the design of the survey. In particular, NHANES respondents are asked to participate various survey sub-components. For example, participant who are scheduled to have their medical tests to be conducted in the morning may be asked to do a blood sugar test which requires fasting. This means that blood sugar test is conducted only on a sub-population which will need to be reflected in the sample weights which will account for additional probability of selection into sub-sample. This will also introduce a missing data for the respondents who did not have such a test and will have missing values in the test variable.

Variance Estimation

As we saw in the previous subsection, using statistical weights for the sampled individuals will affect the bias of the parameter estimates. Another thing we have to account for when analyzing survey data is while incorporating complex sample design, for example clustering and stratification, will inevitably affect variance estimates. This because within clusters individuals tend to be more similar to one another, thus if the data analyzed as if it was sampled using simple random sample, the variance will be too low and biased because the correlation among sample individuals within a cluster will not be accounted for.

To avoid this problem usually more clusters are selected, with less individuals within clusters. However, NHANES can only sample 15 PSUs during each year because of operational limitations such as the cost associated with of moving the survey MECs between PSUs.

To account for this, NHANES created two additional variables: *Masked pseudo-stratum* variable and *masked pseudo PSU* variable for variance estimation. These

variables should be used for all statistical tests and the construction of confidence limits during the analysis. Both variables are masked for confidentiality reasons, however when used, then allow to produce variance estimates that closely approximate the variances that would have been estimated using the true design variables.

Analyzing Subgroups

Unlike with most of the non-survey data, when we are interested in analyzing a sub-population from the data set we can not just remove the observations which do not fit our criteria. Instead, we have to use the special statements provided in the software to perform subgroup analyses. This is because statistical software requires all observations with a non-zero value of the weight variable, as well as an indicator variable indicating which records are in your sub-population of interest.

Data Preparation

We take initial look at the data using the xqplot() function in spida2 package and we find the following are prominent features of the data:

1. Since this is the survey data we find the three variables "WTMEC2YR", "masked.psu", "masked.strat" needed to model the survey design.
2. There are a lot of variables with category "don't know" and "refused to answer" which need to be dealt with.
3. There are a lot of variables with a missing data, some over 90% missing. We will look at the missing data in the next section.
4. Variables pertaining to fractures are severely unbalanced, since they are in the category of response variables we may encounter problems when we are modelling the response.
5. The data set contains a lot of nested variables, meaning that data structure is hierarchical. For example, there are questions about medical conditions such as diabetes with consequent variables pertaining only to individuals who have diabetes asking for additional detail on when condition was diagnosed etc.

   In the next subsection we will be addressing each of the above hurdles individually.

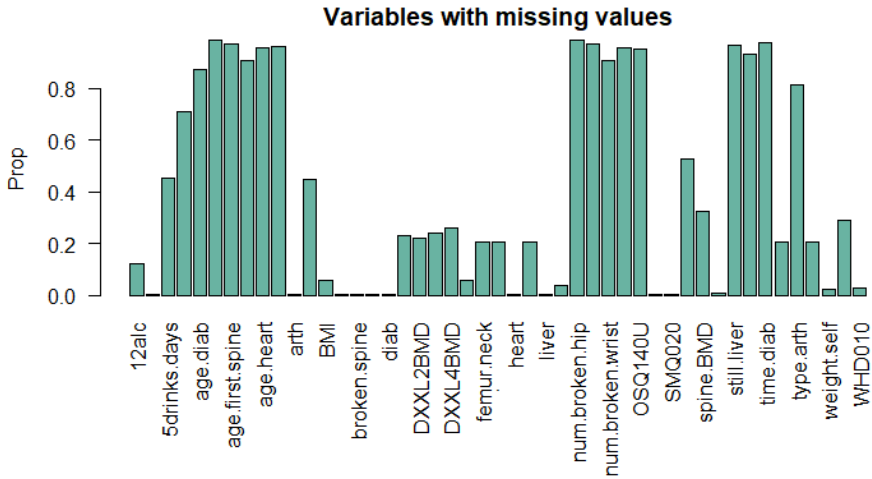
"I don't knows" and "Refuse to answer" categories

In this section we look at all people who replies "I don't know" or "refused" to answer the question. It was initially surprising that even for some of the variables which pertain to the fracture of writs or a spine, some individuals have answered "I don't know". Initially the idea was to manually impute there values as "no", since it would be hard to assume that a person would not notice a broken spine. However, since US has private health care system some patients, even though, in pain may decide to not go to a doctor for an x-ray to confirm a broken bone due to financial reasons. After looking at the observations and observing no clear pattern, I decided that since we

will be doing imputation on the data set I will mark all such categories as missing so that we can get the most probable estimate for those values.

Missing values

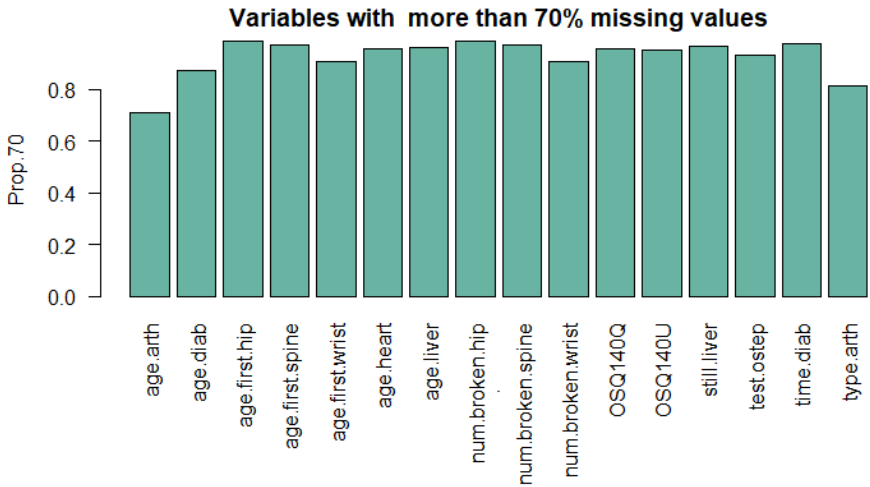
This data set has quite a few missing values. Overall 36% of data is missing, with some variable with more than 90% missing. We will try to understand the reason behind missingness and remedy it by looking at individual variables. First we take look at all of the variables with missing values depicted in Figure below.



There are too many variables with missing values so we will split them into 3 groups: variables with more than 70% missing values, variables with more than 30% missing, and the remaining variables. We will analyze the groups separately to identify the possible reasons behind missingness.

More than 70% missing

The following variables have more than 70% of missing values.



These variables can be grouped by the missingness mechanism or by the way they will be manually imputed.

1. First age of occurrence of a medical condition: categorical "age.first.hip", "age.first.wrist", "age.first.diab" - these three variables are factors which have "NA" whenever no fracture occured. Note that this variable is just a more granular variable of the "broke" variable, where the category "broken" is split between age >50 or<50, thus it carries similar information. This kind of missing mechanism is called "missing by design", since the missingness relies on answer in another variable and these type of varaibles are called nested varaibles. Nested varaibles are common in surveys where the first question asks about for example if the person smokes, and in follwoing questions askes about particulars of the habit (i.e. How many cigarettes a day? etc.). In this variable we will just add a third category for people who didn't have fracture. This mechanism of filling in is called "imputation based on logical rules" or "deductive imputation", since missing-data mechanism is known. This simple and often useful approach to imputation by adding an extra category for the variable indicating missingness is often used for unordered categorical predictors.

2. First age of occurrence of a medical condition: numerical "age.heart", "age.liver", "age.diab","age.arth", "time.diab", - are numerical variables, and we don't know which age cut off point to use for these, but we will follow the cutoff for age variables above and binary into >50 or <50, once we have binary values we will create a category "No heart", "No liver", "No arthritis" and "No diabetes" respectively.

3. Number of fractures "num.broken.wrist", "num.broken.hip", "num.broken.spine"

- for these values it is reasonable to put value zero if there was no fracture, again, this is more refined variable of "broken" variables.

4. Steroid and Prednisone use "OSQ140Q", "OSQ140U" - Both variables pertain to the use of steroids, "OSQ140Q" is a number of days using steroids daily, which can be imputed as zero if no steroids have been used. "OSQ140U" is a factor, for which we will create an additional category whenever steroids have not been used.

5. Liver condition "still.liver" - will be "No" if there was never a liver disease.

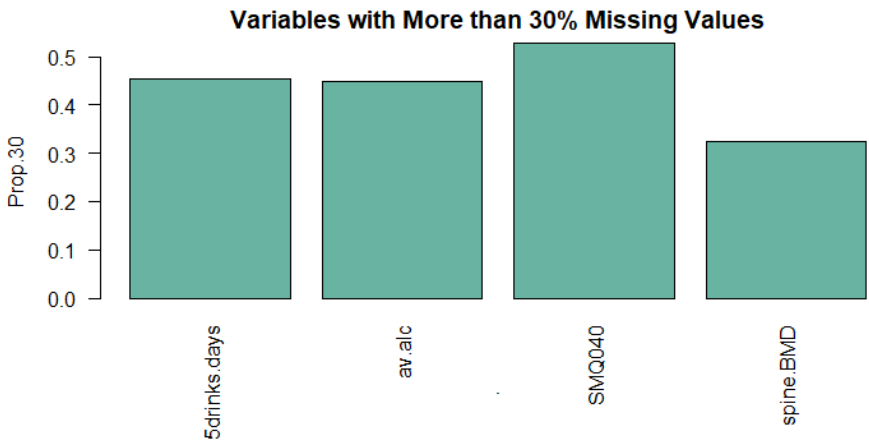
6. Osteoporosis test "test.ostep" - this variable is a factor encoding whether person was ever treated for osteoporosis. Since this variable has over 90% missing values which can not be derived form other variables we will remove it from the data set.

7. Type of arthritis "type.arth" - is also a nested varaible, which depend on whether the person has arthritis. Since the varaible is a factor we will simply create a category "0", which will correspond to the level when person does not have an arthritis.

After manually imputing the missing values for the variables above, next we will look at the variables with more than 30% values missing.

More than 30% missing

We look at the graph below to see which variables have more than 30% of missing values.



There are are now only 4 variables left, meaning that the imputations from previous subsection were successful. We will deal with remaining variables as follows

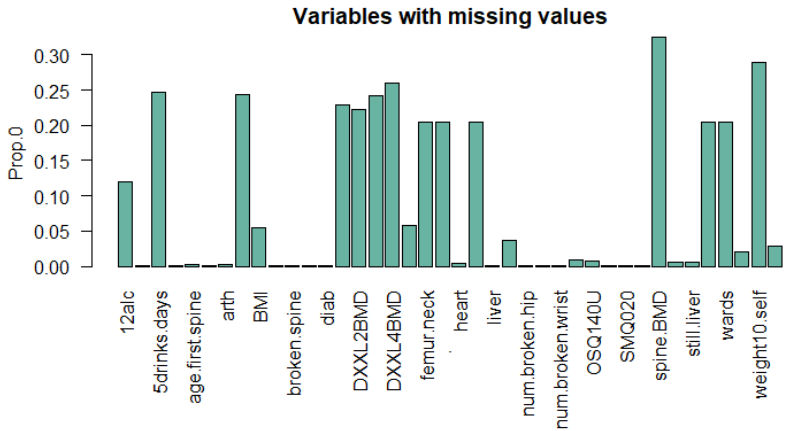
1. Alcohol consumption variables "5drinks.days", "av.alc" - these variable ask in the last year how many days have you had more than 5 drinks, and how many alcoholic beverages you have on average. These variables are highly unreliable because they involve sensitive information. However, we could fill in some "NA" values from the variable - average day consumption (av.alc), which asks how many alcoholic beverages in a year (alc12). We will can safely fill in values of "0" in av.alc and 5drink.days when person has consumed less that 12 drinks last year, this person is not a drinker, thus on average they have 0 drinks per day, and it is highly unlikely they will be consuming more than 5 drinks per day. Even if they do then they had at most two such days in a year which is insignificant in terms of health impact.

2. Smoking variable "SMQ040" - variable asks whether the person is currently a smoker. People were directed to this question if they previously answered "yes" to smoking at least 100 cigarettes a day in variable "SMQ020". This means we can safely impute value "No" for all NA where the person smoked less than 100 cigarettes in their lifetime.

3. Spine BMD "spine.BMD" is one of the BMD variables and we will be dealing with those together.

More than 0% missing

After addressing all the variables above, we again look at variables with any missing values.



The remaining groups of variables with missing values are:

1. Alcohol variables For the remaining values in the alcohol variables we can only use imputation.

2. Bone mineral density (BMD) variables Bone mineral density (BMD) variables - all variables seem to have very similar number of missing values. According to the NHANES website, only certain subsets of individuals had certain medical tests conducted on them. To deal with missing values here we can either two-phase estimation for missing data, in which case we adjust for the weights or do a simple imputation. We will elaborate on both methods in the next section.

3. Weight variables Weight variables are perfect candidate for imputation since there are 4 variables: self reported weight, self reported greatest weight, self reported weight 10 years ago and measured BMI.

Response variable

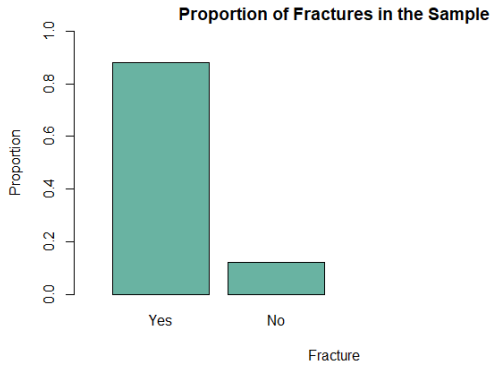
In the data set we don't have a single response variable.Instead we have a group of variables which can considered as a response. These variables are:

1. Broken.hip
2. Broken.wrist
3. Broken spine

4. Num.broken.hip
5. Num.broken.wrist
6. Num.broken.spine

7. Age.first.hip
8. Age.first.wrist
9. Age.first.spine

Since the main goal of the analysis is to identify risk factors associate with osteoporotic fracture we will create a binary variable called "fracture" by merging all types of fracture. Fracture variable will have two level: fracture occurred (yes), fracture hasn't occurred (no). We make a simplifying assumption that all fractures occurred are osteoporotic fractures. The reason is that osteoporosis may occur in children and teenagers and it may be argued that fractures occur commonly in people with weak bones and very rarely in people with strong bones. We look at the fracture variable in the figure below



The response variable is quite unbalanced which will need to be accommodated during modelling.

Modelling

Recall from section on Data Preparation, that one of the features of the data are nested variables, so prior to modelling we would like to describe our approach to dealing with such variables. After that, we will move onto modelling the data using GLM and GLM with a survey design. We will first use only complete cases in the data, and then compare it to the data set with imputed values.

Nested Variables in modelling

As we saw previously, some of the variables in the data set are nested variables. If this data was not a survey data we could analyze it using hierarchical models which allow to have such leveling in the data. However, currently R does not have a package that would allow multilevel survey modelling, thus our choice will be to use the survey package and try to adjust for those variables in the model.

There is very good answer on StackExchange in regards to how to accommodate for such variables in the model:

One situation that occasionally arises in regression analysis involving nested variables is the case where the nested variable has a sufficient amount of detail that it fully determines the initial explanatory variable. An example of this occurs in this question, where the analyst has an indicator variable DrugA for whether or not a drug has been taken, and a nested variable DrugA Concentration for the concentration of the drug. In this example, the latter variable allows a concentration value of zero, which is equivalent to the drug not being taken, and so DrugA is equivalent to DrugA Concentration != 0.

In these types of cases it is possible to remove the initial explanatory variable from the model altogether, and simply use the nested variable on its own. This is legitimate

in this case, because the values in the nested variable determine the value of the initial explanatory variable. For this reason we will remove some of the indicator variables for the imputed data set. In particular variables:

1. Steroid
2. Diabetes
3. Arthritis
4. Heart
5. Liver

However, during the modelling stage it was discovered that none of these nested variables were significant.

Logistic regression without survey design

In order to look at the difference between fitting a regular logistic regression on the data vs. logistic regression with a survey design information we will fit both models. In this section we start with logistic regression on fractures with no survey information. Due to large number of nested variables there is quite a bit of correlation in the data so we were adding various variables in the model and keeping track of correlations using Variance Inflation Factor (VIF). First we will look at the model with just BMD variables.

```
Call:
glm(formula = fracture ~ femur.tot + femur.neck + trochanter +
    intertrochanter + wards + spine.BMD + DXXL1BMD + DXXL2BMD +
    DXXL3BMD + DXXL4BMD, family = binomial(link = "logit"), data = dd)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9465  -0.5316  -0.4637  -0.3949   2.4930

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.9135     0.4134  -2.210   0.0271 *
femur.tot         4.4315     4.6576   0.951   0.3414
femur.neck        0.1918     1.2712   0.151   0.8801
trochanter       -1.1441     1.8026  -0.635   0.5256
intertrochanter  -1.8586     2.4958  -0.745   0.4565
wards            -2.4464     0.6001  -4.077 4.57e-05 ***
spine.BMD       -12.9822    11.1169  -1.168   0.2429
DXXL1BMD          3.2004     2.5358   1.262   0.2069
DXXL2BMD          2.5027     2.8907   0.866   0.3866
DXXL3BMD          3.4532     3.0807   1.121   0.2623
DXXL4BMD          2.9030     3.2166   0.902   0.3668
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2717.1  on 3837  degrees of freedom
Residual deviance: 2665.3  on 3827  degrees of freedom
  (2097 observations deleted due to missingness)
AIC: 2687.3

Number of Fisher Scoring iterations: 5
```

From all of the variables then only significant seems to the measurement on Wards triangle, however it only reasonable to assume that some of the BMD variables are correlated, thus we take a look at the VIF.

```
  femur.tot     femur.neck     trochanter intertrochanter          wards      spine.BMD       DXXL1BMD
 203.828051      13.102498      21.290396      81.303208       4.759849    1054.695736      61.262194
   DXXL2BMD       DXXL3BMD       DXXL4BMD
  79.299370      89.928195      96.718447
```

Some VIFs are extremely high, after several round of adding and removing BMD variable's we arrive at the following model.

```
Call:
glm(formula = fracture ~ intertrochanter + wards + spine.BMD,
    family = binomial(link = "logit"), data = dd)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8735  -0.5317  -0.4673  -0.3978   2.5163

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.9364     0.3818  -2.453   0.0142 *
intertrochanter  0.9669     0.4430   2.182   0.0291 *
wards           -2.0553     0.4027  -5.103 3.34e-07 ***
spine.BMD       -0.8396     0.4530  -1.854   0.0638 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2717.1  on 3837  degrees of freedom
Residual deviance: 2669.7  on 3834  degrees of freedom
  (2097 observations deleted due to missingness)
AIC: 2677.7

Number of Fisher Scoring iterations: 5
```

with the VIF for the three variables

```
intertrochanter           wards       spine.BMD
       2.561782        2.137763        1.745538
```

From this we conclude that wards, spine.BMD and intertrochanter variables are best associated with fractures. Now we are moving on to adding non-BMD variables to the model in order to identify the most significant variables given important BMD variables identified earlier.

```
Call:
glm(formula = fracture ~ wards + spine.BMD + gender + race +
    SMQ040 + alc.12 + steroid + weight10.self + WHD010 + age.arth,
    family = binomial(link = "logit"), data = dd)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3473  -0.5620  -0.4365  -0.3158   2.6910

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.34847    1.59960  -1.468 0.142061
wards         -1.54869    0.48566  -3.189 0.001428 **
spine.BMD     -1.28434    0.53983  -2.379 0.017351 *
gender2        0.02648    0.18046   0.147 0.883352
race2         -0.44166    0.31007  -1.424 0.154334
race3          0.43542    0.20885   2.085 0.037085 *
race4         -0.43370    0.26939  -1.610 0.107412
race5          0.52707    0.34869   1.512 0.130646
SMQ0402       -0.06969    0.37363  -0.187 0.852028
SMQ0403       -0.37548    0.15117  -2.484 0.012997 *
alc.12.L      -0.16556    0.10869  -1.523 0.127697
steroid2      -0.40266    0.22913  -1.757 0.078860 .
weight10.self  0.00412    0.00204   2.020 0.043360 *
WHD010         0.04308    0.02456   1.754 0.079400 .
age.arth2     -0.76803    0.21845  -3.516 0.000438 ***
age.arth0     -0.58984    0.15588  -3.784 0.000154 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1872.2  on 2463  degrees of freedom
Residual deviance: 1738.2  on 2448  degrees of freedom
  (3471 observations deleted due to missingness)
AIC: 1770.2

Number of Fisher Scoring iterations: 5
```

To see any of the variables are highly correlated we will look at the VIF below. It seems like all the VIFs are appropriate.

```
                    GVIF Df GVIF^(1/(2*Df))
wards           1.559236  1        1.248694
spine.BMD       1.708995  1        1.307285
gender          2.073238  1        1.439874
race            1.260777  4        1.029390
SMQ040          1.088919  2        1.021525
alc.12          1.113581  1        1.055263
steroid         1.042564  1        1.021060
weight10.self   1.638244  1        1.279939
WHD010          2.443083  1        1.563036
age.arth        1.152789  2        1.036185
```

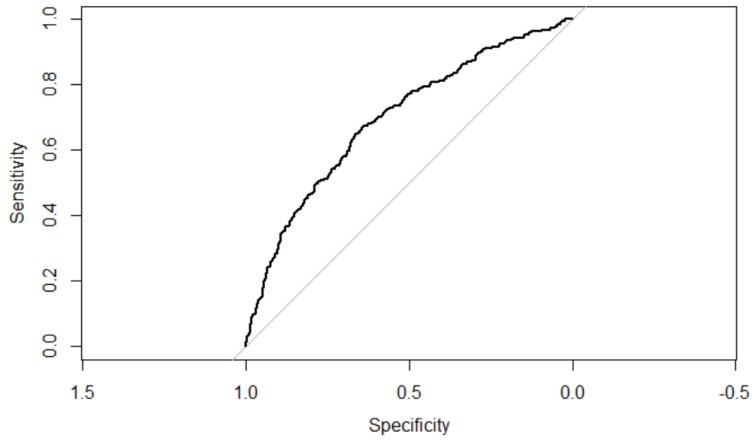
Next we take a look at the deviance table and see how sequentially added variables change the deviance.

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: fracture

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         2463     1872.2
wards         1   30.778      2462     1841.5 2.893e-08 ***
spine.BMD     1    1.196      2461     1840.3 0.2740594
gender        1    6.971      2460     1833.3 0.0082850 **
race          4   47.090      2456     1786.2 1.460e-09 ***
SMQ040        2   10.441      2454     1775.8 0.0054041 **
alc.12        1    2.803      2453     1773.0 0.0940722 .
steroid       1    5.197      2452     1767.8 0.0226280 *
weight10.self 1    9.676      2451     1758.1 0.0018666 **
WHD010        1    3.027      2450     1755.0 0.0818993 .
age.arth      2   16.863      2448     1738.2 0.0002179 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This was the best model that we were able to fit to the data. Recall that this data has unbalanced response variable, thus it is not appropriate to look only at the accuracy of the model, which in our case is 87%, it is also important to see whether the model classifies all of the observations into the bigger class, thus obtaining high accuracy. The table of prediction is given below:

```
pred    0    1
   0 2106  284
   1   46   28
```

Model does not perform exceptionally well. The AUC for this model is 0.6951 with the ROC curve given below:



Logistic Regression with survey design

To fit logistic regression with survey design information we will be using the `survey` package. To specify the survey design we have to run `svydesign` function where we will input observation weights as well as information about PSUs and stratas. The following is the summary of the survey design

```
Stratified 1 - level Cluster Sampling design (with replacement)
With (32) clusters.
svydesign(data = dd, ids = ~masked.psu, strata = ~masked.strat,
    weights = ~WTMEC2YR, nest = TRUE)
Probabilities:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
5.188e-06 2.112e-05 3.874e-05       Inf 6.343e-05       Inf
Stratum Sizes:
            59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74
obs        466 508 493 453 418 376 363 390 508 412 343 290 289 317 161 148
design.PSU   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
actual.PSU   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
Data variables:
 [1] "gender"           "age"              "race"             "WTMEC2YR"         "masked.psu"       "masked.strat"
 [7] "broken.hip"       "broken.wrist"     "broken.spine"     "num.broken.hip"   "num.broken.wrist" "num.broken.spine"
[13] "age.first.hip"    "age.first.wrist"  "age.first.spine"  "steroid"          "OSQ140Q"          "OSQ140U"
[19] "mother.hip"       "father.hip"       "SMQ020"           "SMQ040"           "femur.tot"        "femur.neck"
[25] "trochanter"       "intertrochanter"  "wards"            "spine.BMD"        "DXXL1BMD"         "DXXL2BMD"
[31] "DXXL3BMD"         "DXXL4BMD"         "alc.12"           "av.alc"           "5drinks.days"     "diab"
[37] "age.diab"         "time.diab"        "arth"             "age.arth"         "type.arth"        "heart"
[43] "age.heart"        "liver"            "still.liver"      "age.liver"        "BMI"              "30day.milk"
[49] "reg.milk.use"     "WHD010"           "weight.self"      "weight10.self"    "fracture"         "frac.age"
[55] "bmd.miss"
```

As we have done previously, we will start with BMD variables only. The following is the model with intertrochanter, ward and spine.BMD variables. Note that coefficients have changes and spine.BMD variables is even more significant.

```
Call:
svyglm(formula = fracture ~ intertrochanter + wards + spine.BMD,
    design = design.full, family = "quasibinomial")

Survey design:
svydesign(data = dd, ids = ~masked.psu, strata = ~masked.strat,
    weights = ~WTMEC2YR, nest = TRUE)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.7934     0.2786  -2.848  0.01372 *
intertrochanter   1.5257     0.5887   2.592  0.02236 *
wards            -1.9901     0.3457  -5.756 6.64e-05 ***
spine.BMD        -1.4942     0.4889  -3.056  0.00919 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.115869)

Number of Fisher Scoring iterations: 4
```

Initially we ran the model with non-BMD variables used in the glm model without the survey design, however none of the variables seem to be significant. Thus we have conducted model selection anew and the best logistic model with survey design seems to be the following

```
Call:
svyglm(formula = fracture ~ wards + weight10.self + alc.12 +
    race + age.arth + steroid + gender, design = design.full,
    family = "quasibinomial", data = dd)

Survey design:
svydesign(data = dd, ids = ~masked.psu, strata = ~masked.strat,
    weights = ~WTMEC2YR, nest = TRUE)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.389607   0.465649  -0.837  0.44091
wards         -1.806643   0.328361  -5.502  0.00271 **
weight10.self  0.001857   0.001924   0.965  0.37878
alc.12.L      -0.178196   0.100535  -1.772  0.13651
race2         -0.326348   0.175910  -1.855  0.12273
race3          0.329875   0.128618   2.565  0.05035 .
race4         -0.381926   0.239270  -1.596  0.17133
race5          0.579404   0.339344   1.707  0.14845
age.arth2     -0.547024   0.199971  -2.736  0.04101 *
age.arth0     -0.594008   0.203865  -2.914  0.03326 *
steroid2      -0.324756   0.223782  -1.451  0.20643
gender2       -0.332906   0.157950  -2.108  0.08888 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.022036)

Number of Fisher Scoring iterations: 5
```

A recent feature of the `survey` package is `anova`() test, which has not been available for the survey data previously.

```
Anova table:  (Rao-Scott LRT)
svyglm(formula = fracture ~ wards, design = design.full, family = "quasibinomial",
    data = dd)
                  stats       DEff         df ddf         p
wards          748.5935    0.90389    1.00000  15 1.753e-14 ***
weight10.self 1112.5033    1.16884    1.00000  14 3.238e-14 ***
alc.12          95.5102    1.17456    1.00000  13 6.560e-07 ***
race            20.7518    0.88713    4.00000   9  0.023891 *
age.arth        23.1516    1.99320    2.00000   7  0.035986 *
steroid         20.8865    1.38947    1.00000   6  0.008788 **
gender           7.4601    1.67077    1.00000   5  0.091674 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

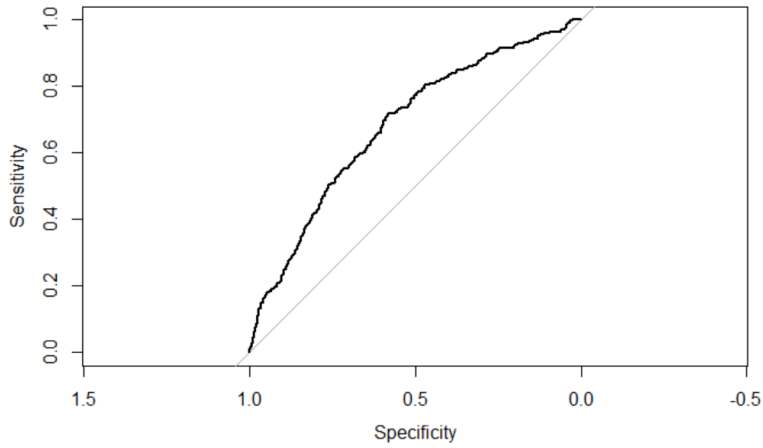
From this we conclude that the most important variables to predict the risk of osteoporosis are:

1. Measure on Wards triangle
2. Self reported weight in the last 10 years
3. Whether the person consumed more than 12 alcoholic beverages in the last year
4. Being of non-Hispanic white Americans ethnicity
5. Having arthritis
6. Use of steroids
7. Gender, to a lesser degree improves the overall model.

However, just like with the regular logistic model this model is not the greatest even with 87% accuracy rate, because of unbalanced response variable. Lets take a look at the predicted values

```
pred.3    0    1
     0 2139  306
     1   13    6
```

The AUC for the model is 0.6806 with the ROC curve depicted below



Imputation

In the previous sections we have omitted all the missing values from the model. In this section we will fit the logistic regression with survey design on the dataset imputed using random forest in `missForest` package. To do the imputation we will be using package missForest. Some of the advantages and disadvantages of `missForest` imputation are given below.

Advantages:

1. Can be applied to both numeric categorical variables
2. No pre-processing required
3. Beside assumption of MAR/MCAR missingness, no additional assumptions are required.
4. Robust to noisy data
5. Non-parametric
6. Allows for non-linear and interaction between variables
7. Gives an OOB error estimate for its predictions. MSE for numeric and PFC for categorical.
8. Works well on high dimensional data

Disadvantages:

1. Lack of interpretability inherent to random forests, however, since we need just prediction this disadvantage is does not apply.

To impute the data we use the following function

```
dd.impute<-missForest(dd, maxiter = 10, variablewise = TRUE)
ddimp<-dd.impute$ximp
```

As mentioned previously, missForest provided Out Of Bag (OOB) estimates of errors for the imputed values. OOB errors are calculated by fitting random forest on part of the data set and using the rest as a remaining data set, which is very similar to the cross validation on different models.

Where PFC is proportion falsely classified, measured on categorical variables and MSE is mean squared error measure on continuous variables. Note that some variables were imputed with minimal errors, but some have really large errors. In particular variables:

1. "5drinks.days"
2. "weight.self"
3. "weight10.self"

We will not use these variables in the modelling stage. However, the variable weight10.self was a significant variable in the previous models, to aid this we will substitute it with the variable WHD010, which is the height of the person. Surprisingly both weight and height are significant, however whenever BMI is in the model it is not significant.

```
  PFC   MSE   PFC   MSE   PFC   MSE   PFC   PFC   PFC   PFC   PFC   PFC
0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.001 0.016 0.003
  PFC   PFC   PFC   PFC   MSE   PFC   PFC   PFC   PFC   PFC   MSE   MSE
0.003 0.015 0.006 0.000 1.172 0.018 0.052 0.021 0.223 0.169 0.001 0.003
  MSE   MSE   MSE   MSE   MSE   MSE   MSE   MSE   PFC   MSE   MSE   PFC
0.002 0.003 0.006 0.000 0.003 0.002 0.002 0.003 0.000 5.808 133.803 0.019
  PFC   PFC   PFC   PFC   PFC   PFC   PFC   PFC   PFC   PFC   MSE   PFC
0.000 0.000 0.004 0.000 0.000 0.001 0.000 0.002 0.016 0.000 7.879 0.520
  PFC   MSE   MSE   MSE
0.444 5.707 219.530 637.417
```

Next we fit the model on the imputed data set with height variable instead of weight variable, we also add some additional variables since the model performance dropped after using the imputed data set and removing the weight variable.

```
Call:
svyglm(formula = fracture ~ wards + spine.BMD + WHD010 + alc.12 +
    race + age.arth + steroid + gender + father.hip + mother.hip,
    design = design.full, family = "quasibinomial", data = dd.rforest)

Survey design:
svydesign(data = dd, ids = ~masked.psu, strata = ~masked.strat,
    weights = ~WTMEC2YR, nest = TRUE)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.40516    1.63423  -0.860   0.4805
wards       -0.68160    0.36513  -1.867   0.2029
spine.BMD   -1.51022    0.39777  -3.797   0.0629 .
WHD010       0.02108    0.02300   0.917   0.4561
alc.12.L    -0.13228    0.10988  -1.204   0.3518
race2       -0.27022    0.22392  -1.207   0.3509
race3        0.69633    0.15202   4.580   0.0445 *
race4       -0.17956    0.24409  -0.736   0.5385
race5        0.61562    0.32417   1.899   0.1980
age.arth2   -0.83134    0.29868  -2.783   0.1085
age.arth0   -0.66393    0.19914  -3.334   0.0794 .
steroid2    -0.05804    0.33669  -0.172   0.8790
gender2     -0.45966    0.18074  -2.543   0.1260
father.hip2  0.43178    0.45436   0.950   0.4423
mother.hip2 -0.14190    0.20562  -0.690   0.5615
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.147257)

Number of Fisher Scoring iterations: 5
```

We run anova() to see how the deviance is affected with each additional variable.

```
Anova table:  (Rao-Scott LRT)
svyglm(formula = fracture ~ wards, design = design.full, family = "quasibinomial",
    data = dd.rforest)
             stats      DEff        df ddf         p
wards      748.5935   0.90389   1.00000  15 1.753e-14 ***
spine.BMD  722.2932   0.54952   1.00000  14 3.477e-15 ***
WHD010      75.9618   1.54898   1.00000  13 1.027e-05 ***
alc.12     101.8221   1.18206   1.00000  12 8.879e-07 ***
race        33.4867   1.14598   4.00000   8  0.016376 *
age.arth    27.3535   1.57591   2.00000   6  0.018010 *
steroid     14.5034   1.86875   1.00000   5  0.040696 *
gender       7.7609   1.77067   1.00000   4  0.108297
father.hip 187.6556   1.65111   1.00000   3  0.001965 **
mother.hip  20.6975   0.78961   1.00000   2  0.039141 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The AUC for this model is 0.6882 which is on par with the rest of the models.

Conclusion

Merging all the results in the analysis we conclude that important variables increase the risk of fracture:

1. Measure on Wards triangle
2. Spine BMD
3. Height
4. Whether the person consumed more than 12 alcoholic beverages in the last year

5. Being of non-Hispanic white Americans ethnicity
6. Having arthritis
7. Use of steroids