Week 7: Regression

RG

2023-06-26

Note that this is a copy of NhanesAnalysis.Rmd

Load the data

There are 6063 observations, some are incomplete and have missing values for some covariates. There are 22 covariates, which have cryptic names and you need to use the meta-data to resolve them. The survey is very complex and typically any analysis requires a substantial amount of reading of the documentation. Here we will guide you past some of the hurdles.

We load up the data and the metadata. In the metadata we have a textual description of the phenotype, the short name, and the target. The target tells us which of the sampled individuals was eligible to answer the question.

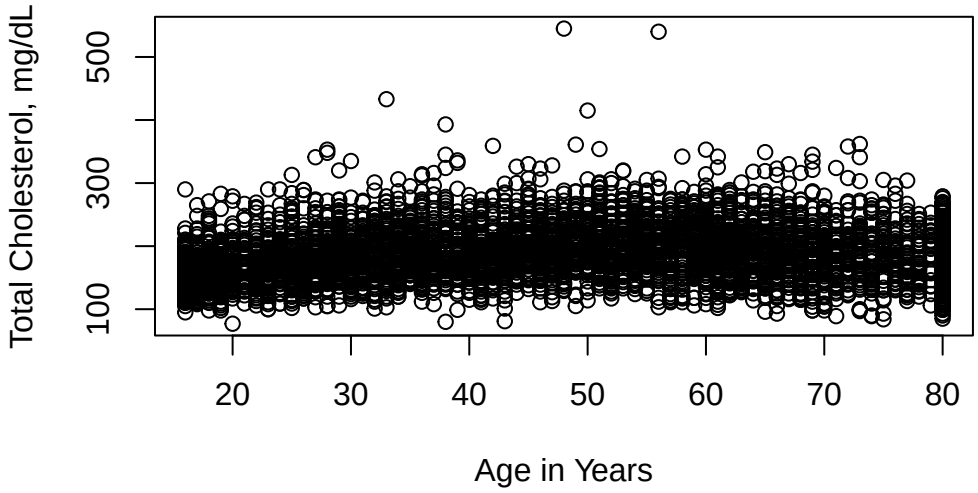
```
nhanesDataPath = ""
load("d4.rda")
load("metaD.rda")
DT::datatable(metaD)
```

```
PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installe
```

Show 10 entries Search:

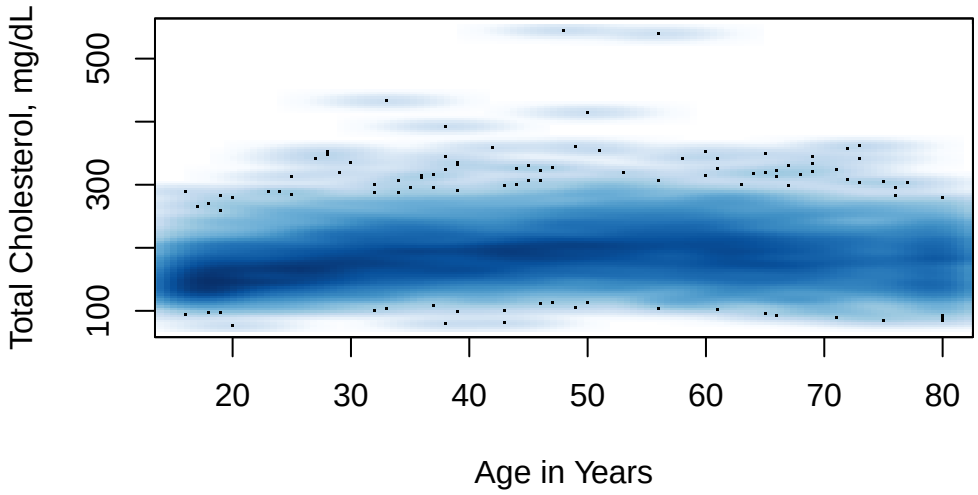
	VariableName	SASLabel	EnglishText	Target
1	RIDAGEYR	Age in years at screening	Age in years of the participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age.	Both males and females 0 YEARS - 150 YEARS
2	RIAGENDR	Gender	Gender of the participant.	Both males and females 0 YEARS - 150 YEARS
3	RIDRETH1	Race/Hispanic origin	Recode of reported race and Hispanic origin information	Both males and females 0 YEARS - 150 YEARS
4	DMDEDUC2	Education level - Adults 20+	What is the highest grade or level of school {you have/SP has} completed or the highest degree {you have/s/he has} received?	Both males and females 20 YEARS - 150 YEARS
5	WTINT2YR	Full sample 2 year interview weight	Full sample 2 year interview weight.	Both males and females 0 YEARS - 150 YEARS
6	SDMVPSU	Masked variance pseudo-PSU	Masked variance unit pseudo-PSU variable for variance estimation	Both males and females 0 YEARS - 150 YEARS
7	SDMVSTRA	Masked variance pseudo-stratum	Masked variance unit pseudo-stratum variable for variance estimation	Both males and females 0 YEARS - 150 YEARS
8	INDFMPIR	Ratio of family income to poverty	A ratio of family income to poverty guidelines.	Both males and females 0 YEARS - 150 YEARS
9	BPQ050A	Now taking prescribed medicine for HBP	HELP AVAILABLE (Are you/Is SP) now taking prescribed medicine	Both males and females 16 YEARS - 150 YEARS

We will look at the relationship between the variable LBXTC (which is Total Cholesterol in mg/dL measured by a blood draw) and the age of the participant in years.



And we can see that plotting, over-plotting is a substantial issue here. You might also notice what seems like a lot of data at age 80, this is because any age over 80 was truncated to 80 to prevent reidentification of survey participants. In a complete analysis, this should probably be adjusted for in some way, but we will ignore it for now.

We can try some other methods, such as `hexbin` plotting and `smoothScatter`



Now we can see a few outliers - with extremely high serum cholesterol. We get a sense that the trend is not exactly a straight line, but rather a parabola, lower for the young and the old and a bit higher in the middle.

We fit a linear model first.

```
lm1 = lm(d4$LBXTC ~ d4$RIDAGEYR)
summary(lm1)
```

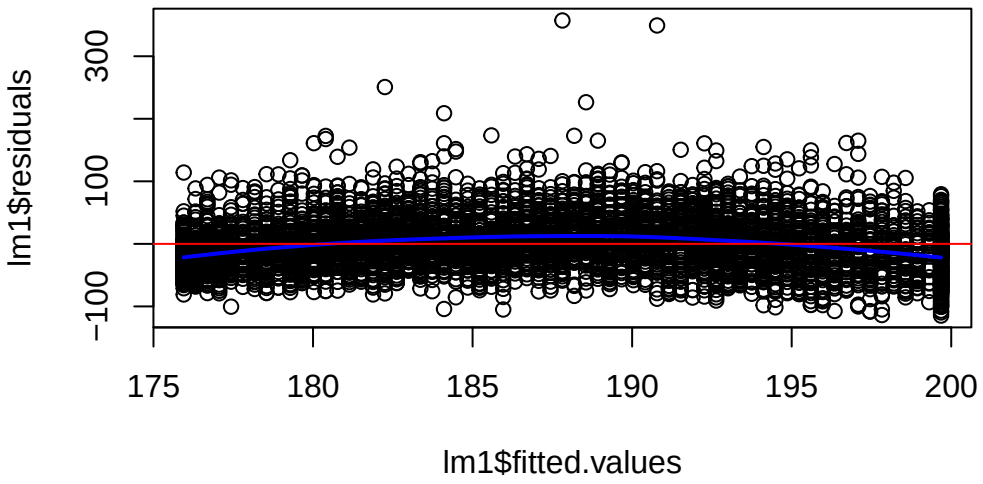
```
Call:
lm(formula = d4$LBXTC ~ d4$RIDAGEYR)

Residuals:
    Min      1Q  Median      3Q     Max
-114.68  -27.95   -2.91   23.34  357.19

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 170.0140     1.4340  118.56   <2e-16 ***
d4$RIDAGEYR   0.3708     0.0285   13.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 41.24 on 5689 degrees of freedom
  (372 observations deleted due to missingness)
Multiple R-squared:  0.02891,   Adjusted R-squared:  0.02874
F-statistic: 169.4 on 1 and 5689 DF,  p-value: < 2.2e-16
```

```r
plot(lm1$fitted.values, lm1$residuals)
##fit a loess curve
l2 = loess(lm1$residuals ~ lm1$fitted.values)
pl = predict(l2, newdata=sort(lm1$fitted.values))
lines(x=sort(lm1$fitted.values), y=pl, col="blue", lwd=2)
abline(h=0, col="red")
```



Notice that both terms in the model are very significant, but that the multiple R^2 is only around 2%. So age, in years, is not explaining very much of the variation. But because we have such a large data set, the parameter estimates are found to be significantly different from zero.

Spline Models

- when a linear model does not appear sufficient we can try other models.

- one choice is to use natural splines, which are very flexible
- they are based on B-splines with the previso that the model is linear outside the range of the data
- based on the initial analysis, we chose to use df=7, which gives five internal knots when fitting the splines
- you have almost 6,000 degrees of freedom here, so using up a few to get a more appropriate fit seems good.

```
library("splines")
lm2 = lm(d4$LBXTC ~ ns(d4$RIDAGEYR, df=7))
summary(lm2)
```

```
Call:
lm(formula = d4$LBXTC ~ ns(d4$RIDAGEYR, df = 7))

Residuals:
    Min      1Q  Median      3Q     Max
-113.43  -26.32   -2.88   22.47  343.31

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                154.799      2.178  71.071  < 2e-16 ***
ns(d4$RIDAGEYR, df = 7)1    39.956      3.379  11.826  < 2e-16 ***
ns(d4$RIDAGEYR, df = 7)2    32.705      4.074   8.028 1.20e-15 ***
ns(d4$RIDAGEYR, df = 7)3    55.583      3.637  15.283  < 2e-16 ***
ns(d4$RIDAGEYR, df = 7)4    42.275      3.725  11.347  < 2e-16 ***
ns(d4$RIDAGEYR, df = 7)5    30.111      3.352   8.984  < 2e-16 ***
ns(d4$RIDAGEYR, df = 7)6    41.098      5.758   7.137 1.07e-12 ***
ns(d4$RIDAGEYR, df = 7)7    15.992      2.478   6.453 1.19e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.62 on 5683 degrees of freedom
  (372 observations deleted due to missingness)
Multiple R-squared:  0.1044,    Adjusted R-squared:  0.1033
F-statistic: 94.65 on 7 and 5683 DF,  p-value: < 2.2e-16
```

- we can use standard tools for comparing models

```
anova(lm1, lm2)
```

```
Analysis of Variance Table

Model 1: d4$LBXTC ~ d4$RIDAGEYR
Model 2: d4$LBXTC ~ ns(d4$RIDAGEYR, df = 7)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   5689 9674185
2   5683 8922039  6    752146 79.848 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice also that the multiple R^2 went up to about 10%, a pretty substantial increase, suggesting that the curvilinear nature of the relationship is substantial.

The residual standard error also decreased by about 5%.

We have lost the simple explanation that comes from fitting a linear model. We cannot say that your serum cholesterol increases by a units per year, but that model was wrong, so it really shouldn't be used.

We can use the regression model we fit to make predictions for any one, and these are substantially more accurate.

Spline Models

- even though the regression summary prints out a different row for each spline term, they are not independent variables, and you need to either retain them all, or retain none of them

Sex

- next we might want to start to add other variables and explore the different relationships.
- let's consider sex, for now we will leave out age, and just try to understand what happens with sex
- first I will fit the model without an intercept

```
lm3 = lm(LBXTC ~ RIAGENDR-1, data=d4)
summary(lm3)
```

```
Call:
lm(formula = LBXTC ~ RIAGENDR - 1, data = d4)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-107.55  -29.55   -3.55   24.21  360.45

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
RIAGENDRMale   184.5534     0.7974   231.4   <2e-16 ***
RIAGENDRFemale 189.7914     0.7692   246.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.76 on 5689 degrees of freedom
  (372 observations deleted due to missingness)
Multiple R-squared:  0.9526,    Adjusted R-squared:  0.9526
F-statistic: 5.722e+04 on 2 and 5689 DF,  p-value: < 2.2e-16
```

- here we can see the mean for males is a bit higher than for females
- both are significant and notice how large the multiple R^2 value is
- this is not a very interesting test - we are asking if the mean is zero, which isn't even physically possible
- our model is

$$y_i = \beta_M \cdot 1_{M,i} + \beta_F \cdot 1_{F,i}$$

- where $1_{M,i}$ is 1 if the i^{th} case is male and zero otherwise, similarly for $1_{F,i}$
- instead we ask if the mean for males is different than that for females

$$y_i = \beta_0 + \beta_1 \cdot 1_{F,i}$$

- so that $E[Y|M] = \beta_0$ and $E[Y|F] = \beta_0 + \beta_1$
- β_1 estimates the difference in mean between male and female

```r
lm3 = lm(LBXTC ~ RIAGENDR, data=d4)
summary(lm3)
```

```
Call:
lm(formula = LBXTC ~ RIAGENDR, data = d4)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-107.55  -29.55   -3.55   24.21  360.45

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    184.5534     0.7974 231.434  < 2e-16 ***
RIAGENDRFemale   5.2380     1.1080   4.728 2.33e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.76 on 5689 degrees of freedom
  (372 observations deleted due to missingness)
Multiple R-squared:  0.003913,  Adjusted R-squared:  0.003738
F-statistic: 22.35 on 1 and 5689 DF,  p-value: 2.327e-06
```

- now we see an Intercept term (that will be the overall mean)
- and the estimate for females is represents how they differ, if it is zero then there is no difference in total cholesterol between men and women

Look at more variables

- now we will put together a set of features (variables) that we are interested in
- for simplicity we only keep partipants for which we have all the data

```
ivars = c("RIDAGEYR", "RIAGENDR", "RIDRETH1", "DMDEDUC2", "INDFMPIR", "LBDHDD", "LBXGH", "
```

```
d4sub = d4[,ivars]
compCases = apply(d4sub, 1, function(x) sum(is.na(x)))
cC = compCases==0
d4sub = d4sub[cC,]
dim(d4sub)
```

```
[1] 4592    9
```

##One quick transformation

- the variable `DMDEDUC2` is a bit too granular for our purposes
- we will modify it to be, less than high school, high school and more than high school

```
table(d4sub$DMDEDUC2)
```

```
                                Less than 9th grade
                                                523
9-11th grade (Includes 12th grade with no diploma)
                                                506
            High school graduate/GED or equivalent
                                               1007
                          Some college or AA degree
                                               1391
                          College graduate or above
                                               1164
                                         Don't Know
                                                  1
```

```
dd = d4sub$DMDEDUC2

dd[dd=="Don't Know"] = NA

eduS = ifelse(dd == "Less than 9th grade" | dd =="9-11th grade (Includes 12th grade with n

#stick this into our dataframe
#and drop the NA
d4sub$eduS = eduS
d4sub = d4sub[-which(is.na(eduS)), ]

table(eduS, dd, useNA = "always")
```

```
      dd
eduS   Less than 9th grade 9-11th grade (Includes 12th grade with no diploma)
  <HS                  523                                                506
  >HS                    0                                                  0
  HS                     0                                                  0
  <NA>                   0                                                  0
      dd
eduS   High school graduate/GED or equivalent Some college or AA degree
  <HS                                      0                          0
  >HS                                      0                       1391
  HS                                    1007                          0
  <NA>                                     0                          0
      dd
eduS   College graduate or above Don't Know <NA>
```

```
<HS                          0          0    0
>HS                       1164          0    0
HS                           0          0    0
<NA>                         0          0    1
```

Principle Components

- we can take the continuous variables and look at principle components
- plotting the first two PCs suggest that these directions are dominated by outliers and that a good step in our analysis might be to remove them

```
cvars = c("RIDAGEYR", "INDFMPIR", "LBDHDD", "LBXGH", "BMXBMI", "LBXTC")
contd4sub=d4sub[, cvars]
pcs = prcomp(contd4sub)

##based on pc plot we have at least 3 outliers that are dominating the first 2 pcs
contd4sub = contd4sub[-c(1077, 2876, 2933),]
d4sub = d4sub[-c(1077, 2876, 2933),]
pcs = prcomp(contd4sub)

pcvals=pcs$x

##which(abs(pcs$x[,1]) > 300)
##which(abs(pcs$x[,2]) > 100)
```

Random Forests

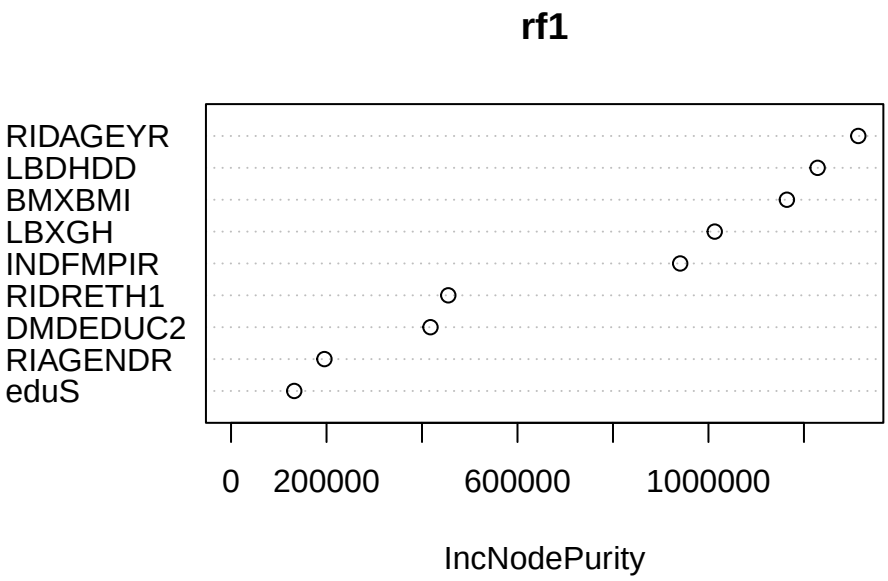
- Random Forests are a simple way to get a sense of how important different variables are in predicting a variable of interest.
- we will cover Random Forests in some detail in our AI/ML lecture for now we will just apply them

```
library("randomForest")
```

```
randomForest 4.7-1.1
```

```
Type rfNews() to see new features/changes/bug fixes.
```

```
rf1 = randomForest(LBXTC ~ ., proximity=TRUE, data=d4sub)
varImpPlot(rf1)
```



Back to Regression

```
lmF = lm(LBXTC ~ ., data=d4sub)
summary(lmF)
```

```
Call:
lm(formula = LBXTC ~ ., data = d4sub)

Residuals:
     Min       1Q   Median       3Q      Max
-121.697  -27.614   -3.266   22.511  230.947

Coefficients: (2 not defined because of singularities)
                                                    Estimate Std. Error
(Intercept)                                       139.83835    5.27657
RIDAGEYR                                            0.11313    0.03677
```

```
RIAGENDRFemale                                                1.21679    1.27112
RIDRETH1Other Hispanic                                       -2.40216    2.19839
RIDRETH1Non-Hispanic White                                   -2.11062    1.90895
RIDRETH1Non-Hispanic Black                                   -8.59043    2.03855
RIDRETH1Other Race - Including Multi-Racial                   1.37260    2.22474
DMDEDUC29-11th grade (Includes 12th grade with no diploma)    1.94689    2.54980
DMDEDUC2High school graduate/GED or equivalent               2.07336    2.29093
DMDEDUC2Some college or AA degree                             1.93297    2.26393
DMDEDUC2College graduate or above                            -1.44617    2.46516
INDFMPIR                                                      1.14015    0.42983
LBDHDD                                                        0.43105    0.03888
LBXGH                                                         2.30193    0.54514
BMXBMI                                                        0.21757    0.09225
eduS>HS                                                           NA         NA
eduSHS                                                            NA         NA
                                                            t value Pr(>|t|)
(Intercept)                                                  26.502  < 2e-16 ***
RIDAGEYR                                                      3.076  0.00211 **
RIAGENDRFemale                                                0.957  0.33849
RIDRETH1Other Hispanic                                       -1.093  0.27459
RIDRETH1Non-Hispanic White                                   -1.106  0.26894
RIDRETH1Non-Hispanic Black                                   -4.214 2.56e-05 ***
RIDRETH1Other Race - Including Multi-Racial                   0.617  0.53728
DMDEDUC29-11th grade (Includes 12th grade with no diploma)    0.764  0.44518
DMDEDUC2High school graduate/GED or equivalent               0.905  0.36550
DMDEDUC2Some college or AA degree                             0.854  0.39325
DMDEDUC2College graduate or above                            -0.587  0.55747
INDFMPIR                                                      2.653  0.00802 **
LBDHDD                                                       11.088  < 2e-16 ***
LBXGH                                                         4.223 2.46e-05 ***
BMXBMI                                                        2.359  0.01839 *
eduS>HS                                                          NA       NA
eduSHS                                                           NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.05 on 4573 degrees of freedom
Multiple R-squared:  0.04359,   Adjusted R-squared:  0.04067
F-statistic: 14.89 on 14 and 4573 DF,  p-value: < 2.2e-16
```

We see that being Non-hispanic black seems to have a pretty big effect, so we might want to just include that, and group all other ethnicities together Education level seems to have little to add, we can drop those.

It might be good to get a sense of the relationship with the poverty level variable.

```
Black = ifelse(d4sub$RIDRETH1 == "Non-Hispanic Black", "B", "nonB")
ivars = c("RIDAGEYR", "INDFMPIR", "LBDHDD", "LBXGH", "BMXBMI", "LBXTC")

d5sub = cbind(d4sub[,ivars], Black)
lmFx = lm(LBXTC ~ . , data=d5sub)
summary(lmFx)
```

```
Call:
lm(formula = LBXTC ~ ., data = d5sub)

Residuals:
     Min       1Q   Median       3Q      Max
-121.851  -27.794   -3.232   22.628  230.533

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 133.22251    5.13242  25.957  < 2e-16 ***
RIDAGEYR      0.10074    0.03558   2.831  0.00465 **
INDFMPIR      0.81062    0.37531   2.160  0.03083 *
LBDHDD        0.43813    0.03643  12.026  < 2e-16 ***
LBXGH         2.38969    0.54074   4.419 1.01e-05 ***
BMXBMI        0.22594    0.08860   2.550  0.01080 *
BlacknonB     7.32259    1.50452   4.867 1.17e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.05 on 4581 degrees of freedom
Multiple R-squared:  0.0415,    Adjusted R-squared:  0.04025
F-statistic: 33.06 on 6 and 4581 DF,  p-value: < 2.2e-16
```

Exercise: You can compare the two models using the `anova` function. Are the two models nested?

```
anova(lmFx, lmF)
```

Exercise: Plot RIDAGEYR versus INDFMPIR. What do you see in the plot? Does anything cause you concern about fitting a linear model?

Missing Indicator Approach

The missing indicator approach may be useful for data where there is a limit of detection, so that values below some lower bound α_L or above some upper bound α_U are set to α_L or to α_U respectively. A similar approach can be used for the Windsorization used for RIDAGEYR variable, where values over 80 are set to 80. We are not proposing using this for other types of missingness, although there is a righ literature and users may want to explore the broader use of this method. However, it is important to realize that often bias or increased variance may obtain, often dependent on untestable assumptions.

The reason that we believe this approach is appropriate for the cases listed is that we actually did measure the variable, and many over variables on those individuals, but we cannot report the exact value for any individual. Hence, one can interpret the indicator as being some form of average over all individuals affected by the reporting limits. It is probably worth noting that these limitations have implications for predication methods. While one in principle could estimate some outcome for an 88 year old person, the data won't support that prediction. Instead one should ignore the linear predictor for age and use the indicator variable. Chiou *et al* (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6812630/) provide some rationale for logistic regression and the references therein point to results for linear regression. Groenwold *et al* (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3414599/) provide advice on more general use of the method to address missing data.

Next let's try to fix out model to deal with the repeated values in these two variables. Now an important consideration is to try and assess just how to interpret them. For RIDAGEYR the documentation states that anyone over age 80 in the database has their age represented as 80. This is not censoring. The measurements (eg BMI, cholesterol etc.) were all made on a person of some age larger than 80. We just Windsorized their ages, and so these individuals really are not the same as the others, where we are getting accurate age values.

(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3414599/)

So basically what we want to do is to create a *dummy variable* for those whose age is reported as 80.

```
Age80 = d5sub$RIDAGEYR == 80
d6sub=cbind(d5sub, Age80)

lmFx2  = lm(LBXTC ~ . + Age80 , data=d6sub)
 summary(lmFx2)
```

```
Call:
lm(formula = LBXTC ~ . + Age80, data = d6sub)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-124.153  -27.990   -2.893   22.471  233.538

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 131.37007    5.10387  25.739  < 2e-16 ***
RIDAGEYR      0.23620    0.03930   6.011 1.99e-09 ***
INDFMPIR      0.63847    0.37346   1.710 0.087409 .
LBDHDD        0.43743    0.03619  12.087  < 2e-16 ***
LBXGH         2.08628    0.53854   3.874 0.000109 ***
BMXBMI        0.17431    0.08826   1.975 0.048333 *
BlacknonB     7.61583    1.49503   5.094 3.64e-07 ***
Age80TRUE   -22.52123    2.85514  -7.888 3.81e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.79 on 4580 degrees of freedom
Multiple R-squared:  0.05435,   Adjusted R-squared:  0.05291
F-statistic: 37.61 on 7 and 4580 DF,  p-value: < 2.2e-16
```

Exercise: What changes do you notice in the model fit? Use the anova function to compare `lmFx3` to `lmFx2`. How do you interpret the output?

Exercise: Try to fit missing indicator variables for both of the repeated values in the INDFMPIR variable. Then interpret the output.

```
Pov5 = d6sub$INDFMPIR == 5
Pov0 = d6sub$INDFMPIR == 0
d7sub = cbind(d6sub, Pov5, Pov0)

lmFx2  = lm(LBXTC ~ . + Age80 + Pov0 + Pov5, data=d7sub)
 summary(lmFx2)
```

```
Call:
lm(formula = LBXTC ~ . + Age80 + Pov0 + Pov5, data = d7sub)

Residuals:
     Min       1Q   Median       3Q      Max
-124.211  -28.113   -2.924   22.452  233.584
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 131.71138    5.14836  25.583  < 2e-16 ***
RIDAGEYR      0.23508    0.03937   5.972 2.53e-09 ***
INDFMPIR      0.48519    0.53208   0.912 0.361880
LBDHDD        0.43724    0.03621  12.076  < 2e-16 ***
LBXGH         2.09263    0.53888   3.883 0.000105 ***
BMXBMI        0.17328    0.08834   1.962 0.049867 *
BlacknonB     7.60308    1.49565   5.083 3.85e-07 ***
Age80TRUE   -22.49633    2.85926  -7.868 4.47e-15 ***
Pov5TRUE      0.77824    2.26671   0.343 0.731364
Pov0TRUE     -2.87497    6.20157  -0.464 0.642966
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.8 on 4578 degrees of freedom
Multiple R-squared:  0.05441,   Adjusted R-squared:  0.05256
F-statistic: 29.27 on 9 and 4578 DF,  p-value: < 2.2e-16
```

It seems that some of the apparent effect of INDFMPIR seems to be related to the fact that we are not fitting RIDAGEYR properly.

```r
lmFx3 = lm(LBXTC ~ ns(RIDAGEYR, df=7)+ INDFMPIR+ LBDHDD+LBXGH + BMXBMI + Black + Age80 +
summary(lmFx3)
```

```
Call:
lm(formula = LBXTC ~ ns(RIDAGEYR, df = 7) + INDFMPIR + LBDHDD +
    LBXGH + BMXBMI + Black + Age80 + Pov0 + Pov5, data = d7sub)

Residuals:
     Min       1Q   Median       3Q      Max
-107.624  -26.969   -3.422   22.446  235.007

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           123.61064    5.67125  21.796  < 2e-16 ***
ns(RIDAGEYR, df = 7)1  29.51675    3.83297   7.701 1.65e-14 ***
ns(RIDAGEYR, df = 7)2  27.22554    4.93196   5.520 3.57e-08 ***
ns(RIDAGEYR, df = 7)3  43.76063    4.41993   9.901  < 2e-16 ***
ns(RIDAGEYR, df = 7)4  27.43607    4.63596   5.918 3.49e-09 ***
ns(RIDAGEYR, df = 7)5  14.40261    4.39131   3.280  0.00105 **
```

```
ns(RIDAGEYR, df = 7)6  30.58254    7.98665   3.829  0.00013 ***
ns(RIDAGEYR, df = 7)7   8.03699    5.13361   1.566  0.11752
INDFMPIR               -0.01926    0.52258  -0.037  0.97061
LBDHDD                  0.45186    0.03551  12.727  < 2e-16 ***
LBXGH                   1.68200    0.53017   3.173  0.00152 **
BMXBMI                  0.07783    0.08683   0.896  0.37014
BlacknonB               8.06440    1.46712   5.497 4.08e-08 ***
Age80TRUE              -7.17874    5.48512  -1.309  0.19068
Pov0TRUE               -6.50137    6.08383  -1.069  0.28529
Pov5TRUE               -0.06740    2.22084  -0.030  0.97579
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.96 on 4572 degrees of freedom
Multiple R-squared:  0.09487,   Adjusted R-squared:  0.0919
F-statistic: 31.95 on 15 and 4572 DF,  p-value: < 2.2e-16
```

Exercise::

In the code below, we drop the terms that were not statistically significant in the model and then compare this smaller model to the larger one, above. Interpret the results.

```
lmFx4 = lm(LBXTC ~ ns(RIDAGEYR, df=7) + LBDHDD+LBXGH+Black, data=d7sub)
summary(lmFx4)
```

```
Call:
lm(formula = LBXTC ~ ns(RIDAGEYR, df = 7) + LBDHDD + LBXGH +
    Black, data = d7sub)

Residuals:
     Min       1Q   Median       3Q      Max
-105.937  -26.949   -3.407   22.193  234.664

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           125.94067    4.93936  25.497  < 2e-16 ***
ns(RIDAGEYR, df = 7)1  29.92780    3.81772   7.839 5.60e-15 ***
ns(RIDAGEYR, df = 7)2  27.19534    4.90144   5.548 3.04e-08 ***
ns(RIDAGEYR, df = 7)3  44.60181    4.35736  10.236  < 2e-16 ***
ns(RIDAGEYR, df = 7)4  26.40388    4.50739   5.858 5.02e-09 ***
ns(RIDAGEYR, df = 7)5  17.52411    3.80778   4.602 4.29e-06 ***
```

```
ns(RIDAGEYR, df = 7)6  28.18635    7.63846   3.690 0.000227 ***
ns(RIDAGEYR, df = 7)7   2.33978    2.68986   0.870 0.384428
LBDHDD                  0.44206    0.03402  12.993  < 2e-16 ***
LBXGH                   1.73292    0.52267   3.315 0.000922 ***
BlacknonB               7.88204    1.45294   5.425 6.10e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.96 on 4577 degrees of freedom
Multiple R-squared:  0.09412,   Adjusted R-squared:  0.09214
F-statistic: 47.56 on 10 and 4577 DF,  p-value: < 2.2e-16
```

```
anova(lmFx4, lmFx3)
```

```
Analysis of Variance Table

Model 1: LBXTC ~ ns(RIDAGEYR, df = 7) + LBDHDD + LBXGH + Black
Model 2: LBXTC ~ ns(RIDAGEYR, df = 7) + INDFMPIR + LBDHDD + LBXGH + BMXBMI +
    Black + Age80 + Pov0 + Pov5
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1   4577 6946291
2   4572 6940552  5    5739.5 0.7562 0.5814
```