Assessment 2

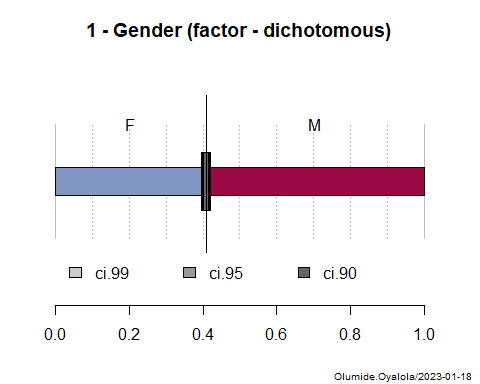
Halimah

2023-01-18

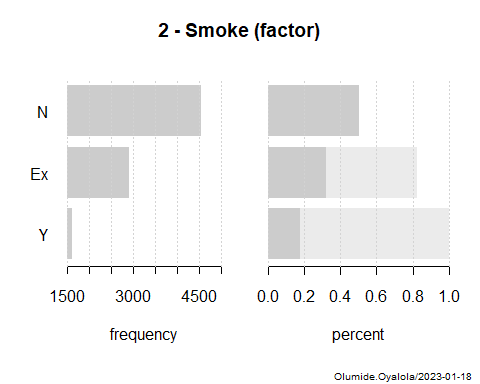
## Question 2

### (A) Data Summary

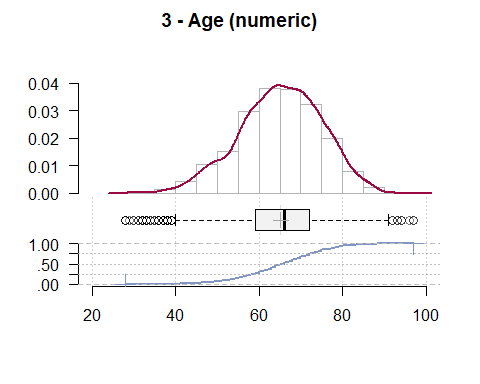
------------------------------------------------------------------------------   
Describe . (data.frame):  
  
data frame: 11742 obs. of 10 variables  
 5851 complete cases (49.8%)  
  
 Nr ColName Class NAs Levels   
 1 Gender factor . (2): 1-F, 2-M   
 2 Smoke factor 2700 (23.0%) (3): 1-Ex, 2-N, 3-Y  
 3 Age numeric 1917 (16.3%)   
 4 TimeFromDiagnosis numeric 15 (0.1%)   
 5 HbA1c numeric .   
 6 BMI numeric 3936 (33.5%)   
 7 SBP numeric 3222 (27.4%)   
 8 Cholesterol numeric 3328 (28.3%)   
 9 Triglycerides numeric 650 (5.5%)   
 10 MetforminTreatment numeric .   
  
  
------------------------------------------------------------------------------   
1 - Gender (factor - dichotomous)  
  
 length n NAs unique  
 11'742 11'742 0 2  
 100.0% 0.0%   
  
 freq perc lci.95 uci.95'  
F 4'793 40.8% 39.9% 41.7%  
M 6'949 59.2% 58.3% 60.1%  
  
' 95%-CI (Wilson)



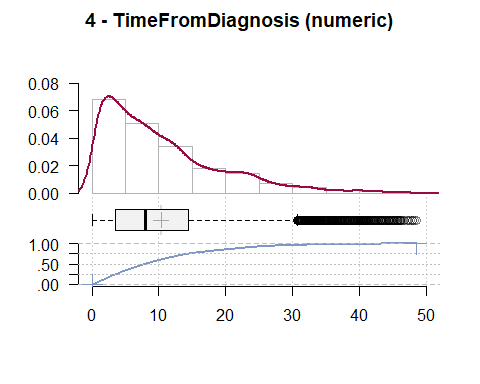
------------------------------------------------------------------------------   
2 - Smoke (factor)  
  
 length n NAs unique levels dupes  
 11'742 9'042 2'700 3 3 y  
 77.0% 23.0%   
  
 level freq perc cumfreq cumperc  
1 N 4'540 50.2% 4'540 50.2%  
2 Ex 2'904 32.1% 7'444 82.3%  
3 Y 1'598 17.7% 9'042 100.0%



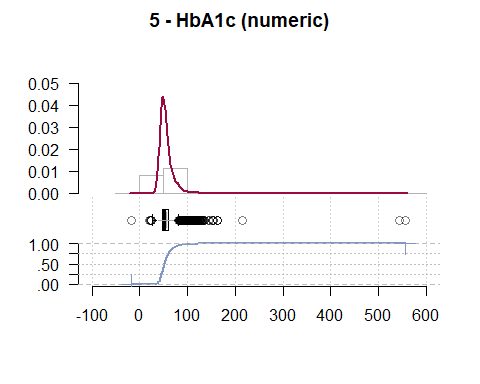
------------------------------------------------------------------------------   
3 - Age (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 11'742 9'825 1'917 69 0 65.18 64.98  
 83.7% 16.3% 0.0% 65.37  
   
 .05 .10 .25 median .75 .90 .95  
 48.00 52.00 59.00 66.00 72.00 78.00 81.00  
   
 range sd vcoef mad IQR skew kurt  
 69.00 9.92 0.15 10.38 13.00 -0.26 -0.08  
   
lowest : 28.0 (4), 29.0, 30.0 (2), 31.0 (2), 32.0 (4)  
highest: 92.0 (2), 93.0 (2), 94.0 (2), 96.0, 97.0 (2)  
  
' 95%-CI (classic)



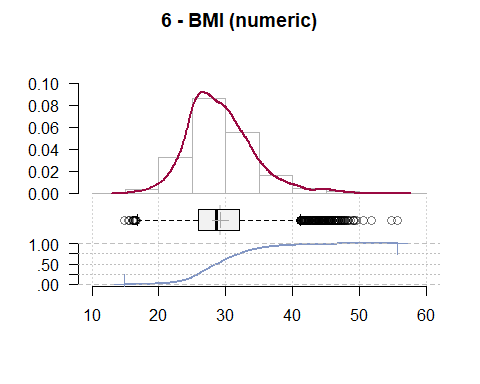
------------------------------------------------------------------------------   
4 - TimeFromDiagnosis (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 11'742 11'727 15 2'932 0 10.34 10.18  
 99.9% 0.1% 0.0% 10.50  
   
 .05 .10 .25 median .75 .90 .95  
 0.89 1.48 3.57 7.95 14.41 23.04 27.85  
   
 range sd vcoef mad IQR skew kurt  
 48.44 8.76 0.85 7.41 10.84 1.29 1.55  
   
lowest : 0.02 (2), 0.04, 0.05 (4), 0.06 (2), 0.07 (2)  
highest: 47.67, 47.86, 47.96, 48.21, 48.46  
  
' 95%-CI (classic)



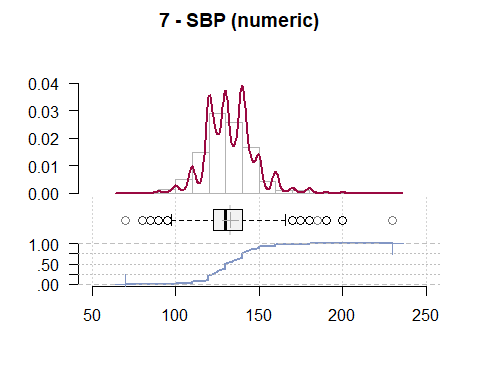
------------------------------------------------------------------------------   
5 - HbA1c (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 11'742 11'742 0 998 0 55.16 54.89  
 100.0% 0.0% 0.0% 55.43  
   
 .05 .10 .25 median .75 .90 .95  
 39.89 42.08 46.45 51.96 60.66 71.58 80.33  
   
 range sd vcoef mad IQR skew kurt  
 572.68 14.89 0.27 9.79 14.21 7.70 212.93  
   
lowest : -16.94, 20.22, 21.31, 22.4, 23.5 (2)  
highest: 154.65 (2), 163.39 (4), 214.0, 544.81, 555.74  
  
' 95%-CI (classic)



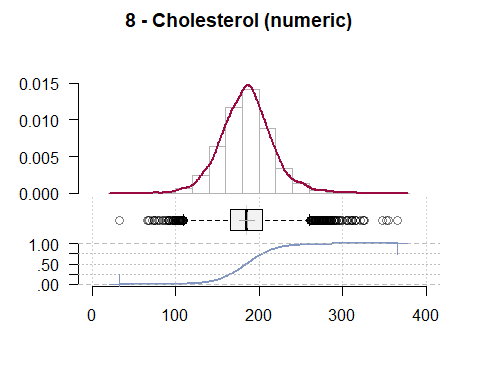
------------------------------------------------------------------------------   
6 - BMI (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 11'742 7'806 3'936 1'747 0 29.20 29.09  
 66.5% 33.5% 0.0% 29.31  
   
 .05 .10 .25 median .75 .90 .95  
 22.27 23.77 25.88 28.65 31.99 35.45 38.19  
   
 range sd vcoef mad IQR skew kurt  
 40.75 4.93 0.17 4.42 6.11 0.76 1.28  
   
lowest : 14.91, 15.43, 15.61, 16.03, 16.16  
highest: 49.52, 50.67, 51.83, 54.82, 55.66  
  
' 95%-CI (classic)



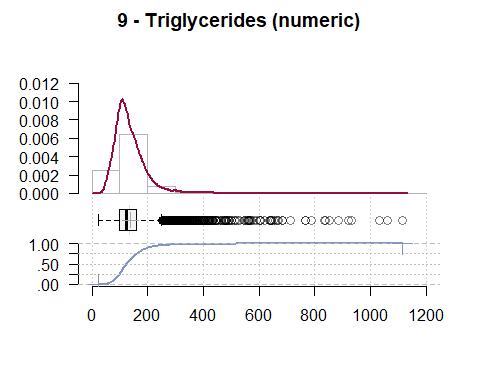
------------------------------------------------------------------------------   
7 - SBP (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 11'742 8'520 3'222 3'384 0 132.82 132.52  
 72.6% 27.4% 0.0% 133.13  
   
 .05 .10 .25 median .75 .90 .95  
 110.00 120.00 122.38 130.00 140.00 150.00 160.00  
   
 range sd vcoef mad IQR skew kurt  
 160.00 14.36 0.11 14.83 17.62 0.51 1.69  
   
lowest : 70.0, 80.0 (2), 85.0 (3), 90.0 (26), 95.0 (12)  
highest: 180.0 (75), 185.0, 190.0 (10), 200.0 (11), 230.0  
  
heap(?): remarkable frequency (12.6%) for the mode(s) (= 130)  
  
' 95%-CI (classic)



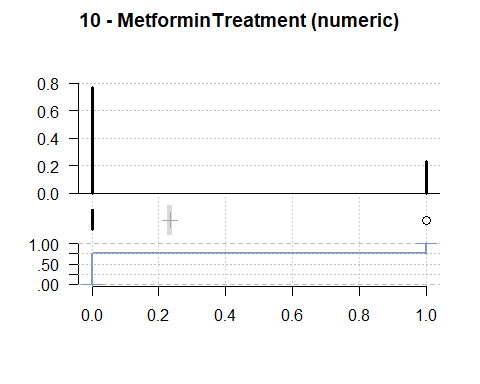
------------------------------------------------------------------------------   
8 - Cholesterol (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 11'742 8'414 3'328 3'780 0 185.43 184.76  
 71.7% 28.3% 0.0% 186.09  
   
 .05 .10 .25 median .75 .90 .95  
 136.00 148.00 166.00 185.00 203.95 222.00 237.00  
   
 range sd vcoef mad IQR skew kurt  
 332.00 31.06 0.17 28.17 37.95 0.28 1.38  
   
lowest : 33.0, 66.0, 68.0, 69.0, 72.0  
highest: 326.0, 347.0, 352.0, 355.0, 365.0  
  
' 95%-CI (classic)



------------------------------------------------------------------------------   
9 - Triglycerides (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 11'742 11'092 650 5'240 0 137.88 136.65  
 94.5% 5.5% 0.0% 139.11  
   
 .05 .10 .25 median .75 .90 .95  
 68.00 80.00 100.00 124.96 160.43 202.00 238.00  
   
 range sd vcoef mad IQR skew kurt  
 1'093.00 66.30 0.48 42.95 60.43 3.80 30.22  
   
lowest : 22.0, 25.0, 35.0 (2), 36.0 (2), 38.0  
highest: 919.0, 930.0, 1'032.0, 1'059.0, 1'115.0  
  
' 95%-CI (classic)



------------------------------------------------------------------------------   
10 - MetforminTreatment (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 11'742 11'742 0 2 8'995 0.23 0.23  
 100.0% 0.0% 76.6% 0.24  
   
 .05 .10 .25 median .75 .90 .95  
 0.00 0.00 0.00 0.00 0.00 1.00 1.00  
   
 range sd vcoef mad IQR skew kurt  
 1.00 0.42 1.81 0.00 0.00 1.26 -0.42  
   
  
 value freq perc cumfreq cumperc  
1 0 8'995 76.6% 8'995 76.6%  
2 1 2'747 23.4% 11'742 100.0%  
  
' 95%-CI (classic)



* The dataset contains 11742 observations with 11 variables. Only 5851(49.8%) complete cases were found in the dataset.
* The Gender variable has no missing value. On the other hand, the BMI variable has the most missing values with 3936 (33.5%) missing values.

### (B) Missing Values Imputation

iter imp variable  
 1 1 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 1 2 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 1 3 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 1 4 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 1 5 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 2 1 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 2 2 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 2 3 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 2 4 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 2 5 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 3 1 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 3 2 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 3 3 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 3 4 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 3 5 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 4 1 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 4 2 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 4 3 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 4 4 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 4 5 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 5 1 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 5 2 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 5 3 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 5 4 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides  
 5 5 Smoke Age TimeFromDiagnosis BMI SBP Cholesterol Triglycerides

## Question 3

### (A) Fitting a model

**Hypothesis**

* Null hypothesis:
* Alternative hypothesis:

===============================================  
 Dependent variable:   
 ---------------------------  
 HbA1c   
-----------------------------------------------  
MetforminTreatment 3.200\*\*\*   
 (0.320)   
   
Constant 54.000\*\*\*   
 (0.160)   
   
-----------------------------------------------  
Observations 11,742   
R2 0.008   
Adjusted R2 0.008   
Residual Std. Error 15.000 (df = 11740)   
F Statistic 99.000\*\*\* (df = 1; 11740)   
===============================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

**Hypothesis Testing**

Linear hypothesis test  
  
Hypothesis:  
MetforminTreatment = 0  
  
Model 1: restricted model  
Model 2: HbA1c ~ MetforminTreatment  
  
 Res.Df RSS Df Sum of Sq F Pr(>F)   
1 11741 2601427   
2 11740 2579593 1 21835 99.4 <0.0000000000000002 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### (B) Model Discussion

* The linear hypothesis test result suggests that the estimate of MetforminTreatment is significantly different from zero. This implies that MetforminTreatment has a causal effect on HbA1c.
* For the missing value imputation, it is assume that the missing data are missing at Random (MAR), which mean that the probability that a value is missing depends only on observed value and can be predicted using them.