CKD Screening Case

Logistic Regression

By Hoang Nha Nguyen

Agenda

1. Case Objective
2. Data Exploration
3. Imputation - “Mice” Package
4. Backward vs Forward Selection Technique

Case Objective

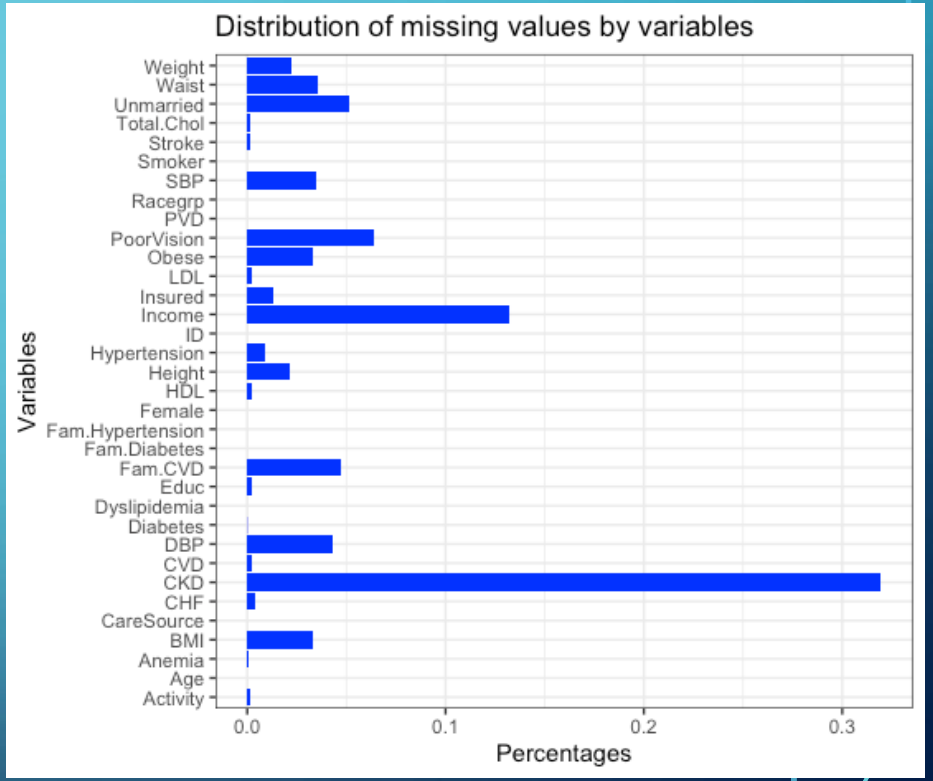
- To build a predictive model that could be turned into a quick screening tool that identifies patients who are at higher risk of developing CKD.
- Two main causes of CKD are diabetes and high blood pressure/ hypertension.
- Heart disease is major cause of death for people with CKD.

Data Exploration

- The data set consists of selected information from 8,819 adults 20 years of age or older taken from the 1999 to 2000 and 2001 to 2002 surveys.
- The dataset is divided into 2 sets:
  - 6,000-case training set
  - 2,819-case hold-out sample for testing. -> CKD has to be predicted
- A test for CKD was given to everyone in the study population.
- Target Variable is CKD, a binary variable indicating whether or not the subject had CKD.
- The 33 independent variables include age, weight, income, cholesterol level, systolic/diastolic blood pressure, family history of diabetes, cardiovascular diseases, etc.

Missing Values

- Our dataset consists of 8819 responses against 33 attributes (8819 x 33) 291027 individual responses are to be recorded.
- Only 283285 are recorded and 7742 records are missing (which is about 2.6 % of the data set)
- Issue: Class Imbalance problem -> Only **464 out of 6000** have CKD
- There are 24 variables with missing values.



Imputation

- Use MICE package with Predictive Mean Matching method to fill in the missing data.
- Combine multiple implication methods to predict missing values based on known values
- Skip variables with no missing values -> predictors
- Specify imputation methods for each variable:
  - **'logreg'** - logistic regression imputation (binary data, factor with 2 levels)
  - **'polyreg'** - polytomous regression imputation for unordered categorical data (factor > 2 levels)
  - **'norm'** - continuous variables

Before Imputation

```
> sapply(casedata, function(x) sum(is.na(x))) # Check number of missing values before imputation
```

ID	Age	Female	Racegrp	Educ	Unmarried
0	0	0	0	20	452
Income	**CareSource**	**Insured**	**Weight**	**Height**	**BMI**
1166	0	113	194	191	290
Obese	**Waist**	**SBP**	**DBP**	**HDL**	**LDL**
290	314	308	380	17	18
Total.Chol	**Dyslipidemia**	**PVD**	**Activity**	**PoorVision**	**Smoker**
16	0	0	10	567	0
Hypertension	**Fam.Hypertension**	**Diabetes**	**Fam.Diabetes**	**Stroke**	**CVD**
80	0	2	0	11	23
Fam.CVD	**CHF**	**Anemia**	**CKD**		
419	36	6	2819		

After Imputation

```
> sapply(finaldata, function(x) sum(is.na(x))) # Check the number of missing values after imputation
```

Weight	Height	BMI	SBP	DBP	Waist
95	191	290	27	35	112
HDL	**LDL**	**Total.Chol**	**Racegrp**	**CareSource**	**Age**
1	1	1	0	0	0
Income	**Educ**	**Unmarried**	**Insured**	**Obese**	**PoorVision**
20	5	9	1	191	90
Hypertension	**Diabetes**	**Stroke**	**CVD**	**Fam.CVD**	**CHF**
0	1	1	2	22	5
Anemia	**Smoker**	**PVD**	**Female**	**Fam.Hypertension**	**Fam.Diabetes**
0	0	0	0	0	0
Dyslipidemia	**Activity**	**CKD**			
0	0	2819			

Divide datasets

```
hold_out_sample=which(is.na(data$CKD)==1)
data_without=data[hold_out_sample,]   ## the ones without a disease status
data_with=data[-hold_out_sample,]   ## the ones with a disease status
summary(data_with)
```

```
> dim(data_in)
[1] 4136   33
> sapply(data_in, function(x) sum(is.na(x)))
              Age            Female          Racegrp             Educ        Unmarried            Income
                0                 0                 0                0                0                 0
       CareSource          Insured            Weight           Height              BMI             Obese
                0                 0                 0                0                0                 0
            Waist               SBP               DBP              HDL              LDL        Total.Chol
                0                 0                 0                0                0                 0
     Dyslipidemia               PVD          Activity       PoorVision           Smoker      Hypertension
                0                 0                 0                0                0                 0
Fam.Hypertension          Diabetes     Fam.Diabetes           Stroke              CVD           Fam.CVD
                0                 0                 0                0                0                 0
              CHF           Anemia               CKD
                0                 0                 0
```

LRM (CKD ~ AGE)

- The coefficient of age is positive indicating that an increase in age will lead to an increase in the probability of someone having CKD

```
> model=glm(CKD~Age,family="binomial",data=data_in)
> summary(model)

Call:
glm(formula = CKD ~ Age, family = "binomial", data = data_in)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0570  -0.3451  -0.1511  -0.0805   3.4148

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.547378   0.389045  -21.97   <2e-16 ***
Age          0.097145   0.005498   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1974.3  on 4135  degrees of freedom
Residual deviance: 1455.2  on 4134  degrees of freedom
AIC: 1459.2

Number of Fisher Scoring iterations: 7
```

```
## Run the Logistic Regression with one variable
model=glm(CKD~Age,family="binomial",data=data_with)
summary(model)
```

Base Model

```
Call:
glm(formula = CKD ~ ., family = "binomial", data = data_with)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6986  -0.2845  -0.1210  -0.0582   3.4183

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.543e+01  3.248e+02  -0.047  0.96212
Age                9.622e-02  8.548e-03  11.257  < 2e-16 ***
Female             6.933e-01  2.401e-01   2.888  0.00388 **
Racegrphispa      -4.857e-01  2.907e-01  -1.671  0.09476 .
Racegrpother       3.201e-01  5.777e-01   0.554  0.57949
Racegrpwhite       2.330e-01  2.216e-01   1.052  0.29298
Educ              -2.128e-01  1.688e-01  -1.261  0.20738
Unmarried          2.668e-01  1.686e-01   1.583  0.11353
Income             6.053e-02  1.808e-01   0.335  0.73781
CareSourceclinic   7.156e+00  3.247e+02   0.022  0.98242
CareSourceDrHMO    7.089e+00  3.247e+02   0.022  0.98259
CareSourcenoplace  6.939e+00  3.247e+02   0.021  0.98295
CareSourceother    7.422e+00  3.247e+02   0.023  0.98177
Insured            2.644e-01  3.790e-01   0.697  0.48552
Weight             4.507e-02  3.934e-02   1.146  0.25193
Height             8.386e-03  3.866e-02   0.217  0.82829
BMI               -6.697e-02  1.099e-01  -0.609  0.54235
Obese              2.868e-01  2.453e-01   1.169  0.24244
Waist             -3.168e-02  1.415e-02  -2.239  0.02517 *
SBP               -5.643e-03  4.182e-03  -1.349  0.17723
DBP               -7.674e-04  6.379e-03  -0.120  0.90425
HDL               -1.875e-02  5.833e-03  -3.215  0.00131 **
LDL                3.435e-03  1.920e-03   1.789  0.07360 .
Total.Chol                NA         NA      NA       NA
Dyslipidemia      -2.768e-01  2.524e-01  -1.097  0.27282
PVD                4.243e-01  2.333e-01   1.819  0.06898 .
PoorVision         4.858e-02  2.327e-01   0.209  0.83464
Smoker            -3.544e-02  1.557e-01  -0.228  0.81998
Hypertension       8.166e-01  2.004e-01   4.075 4.61e-05 ***
Fam.Hypertension  -1.484e-01  3.086e-01  -0.481  0.63069
Diabetes           6.495e-01  1.888e-01   3.441  0.00058 ***
Fam.Diabetes      -1.453e-01  1.645e-01  -0.883  0.37721
Stroke             1.546e-01  3.516e-01   0.440  0.66002
CVD                6.613e-01  2.646e-01   2.499  0.01245 *
Fam.CVD            1.447e-01  2.761e-01   0.524  0.60022
CHF                1.094e-01  3.034e-01   0.361  0.71841
Anemia             1.178e+00  5.192e-01   2.269  0.02329 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1974.3  on 4135  degrees of freedom
Residual deviance: 1313.0  on 4099  degrees of freedom
AIC: 1387

Number of Fisher Scoring iterations: 11
```

Backward v.s Forward Selection

- **Forward selection**, which starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.
- **Backward selection** , which starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.

Evaluation Metrics

- Deviance is a measure of goodness of fit of a model. Higher numbers always indicate bad fit. The null deviance shows how well the dependent variable is predicted by a model that includes only the intercept (grand mean), while residual includes independent variables.
- AIC (Akaike Information Criterion) and P-value

  AIC - estimation of prediction error -> lower means better

Model 2 - Forward Selection

```
> summary(model2)

Call:
glm(formula = CKD ~ Age + Female + Racegrp + Educ + Unmarried +
    Income + CareSource + Insured + Weight + Height + BMI + Obese +
    Waist + SBP + DBP + HDL + LDL + Total.Chol + Dyslipidemia +
    PVD + Activity + PoorVision + Smoker + Hypertension + Fam.Hypertension +
    Diabetes + Fam.Diabetes + Stroke + CVD + Fam.CVD + CHF +
    Anemia, family = "binomial", data = data_with)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6986  -0.2845  -0.1210  -0.0582   3.4183

Coefficients: (1 not defined because of singularities)
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.543e+01  3.248e+02  -0.047  0.96212
Age                9.622e-02  8.548e-03  11.257  < 2e-16 ***
Female             6.933e-01  2.401e-01   2.888  0.00388 **
Racegrphispa      -4.857e-01  2.907e-01  -1.671  0.09476 .
Racegrpother       3.201e-01  5.777e-01   0.554  0.57949
Racegrpwhite       2.330e-01  2.216e-01   1.052  0.29298
Educ              -2.128e-01  1.688e-01  -1.261  0.20738
Unmarried          2.668e-01  1.686e-01   1.583  0.11353
Income             6.053e-02  1.808e-01   0.335  0.73781
CareSourceclinic   7.156e+00  3.247e+02   0.022  0.98242
CareSourceDrHMO    7.089e+00  3.247e+02   0.022  0.98259
CareSourcenoplace  6.939e+00  3.247e+02   0.021  0.98295
CareSourceother    7.422e+00  3.247e+02   0.023  0.98177
Insured            2.644e-01  3.790e-01   0.697  0.48552
Weight             4.507e-02  3.934e-02   1.146  0.25193
Height             8.386e-03  3.866e-02   0.217  0.82829
BMI               -6.697e-02  1.099e-01  -0.609  0.54235
Obese              2.868e-01  2.453e-01   1.169  0.24244
Waist             -3.168e-02  1.415e-02  -2.239  0.02517 *
SBP               -5.643e-03  4.182e-03  -1.349  0.17723
DBP               -7.674e-04  6.379e-03  -0.120  0.90425
HDL               -1.875e-02  5.833e-03  -3.215  0.00131 **
LDL                3.435e-03  1.920e-03   1.789  0.07360 .
Total.Chol                NA         NA      NA       NA
Dyslipidemia      -2.768e-01  2.524e-01  -1.097  0.27282
PVD                4.243e-01  2.333e-01   1.819  0.06898 .
Activity          -2.399e-01  1.148e-01  -2.090  0.03662 *
PoorVision         4.858e-02  2.327e-01   0.209  0.83464
Smoker            -3.544e-02  1.557e-01  -0.228  0.81998
Hypertension       8.166e-01  2.004e-01   4.075 4.61e-05 ***
Fam.Hypertension  -1.484e-01  3.086e-01  -0.481  0.63069
Diabetes           6.495e-01  1.888e-01   3.441  0.00058 ***
Fam.Diabetes      -1.453e-01  1.645e-01  -0.883  0.37721
Stroke             1.546e-01  3.516e-01   0.440  0.66002
CVD                6.613e-01  2.646e-01   2.499  0.01245 *
Fam.CVD            1.447e-01  2.761e-01   0.524  0.60022
CHF                1.094e-01  3.034e-01   0.361  0.71841
Anemia             1.178e+00  5.192e-01   2.269  0.02329 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1974.3  on 4135  degrees of freedom
Residual deviance: 1313.0  on 4099  degrees of freedom
AIC: 1387

Number of Fisher Scoring iterations: 11
```

Model 3 - Backward Elimination

```
> summary(model3)

Call:
glm(formula = CKD ~ Age + Female + Racegrp + Unmarried + Weight +
    BMI + Waist + SBP + HDL + LDL + PVD + Activity + Hypertension +
    Diabetes + CVD + Anemia, family = "binomial", data = data_with)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7818  -0.2880  -0.1238  -0.0597   3.3790

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.339337   1.013424  -7.242 4.42e-13 ***
Age          0.097217   0.007530  12.911  < 2e-16 ***
Female       0.703990   0.232362   3.030 0.002448 **
Racegrphispa -0.503167   0.285432  -1.763 0.077929 .
Racegrpother  0.231132   0.572526   0.404 0.686428
Racegrpwhite  0.206131   0.213694   0.965 0.334743
Unmarried     0.271115   0.162362   1.670 0.094956 .
Weight        0.051595   0.012619   4.089 4.34e-05 ***
BMI          -0.071436   0.038069  -1.876 0.060587 .
Waist        -0.029776   0.013922  -2.139 0.032459 *
SBP          -0.005660   0.003872  -1.462 0.143767
HDL          -0.017954   0.005685  -3.158 0.001586 **
LDL           0.002901   0.001818   1.596 0.110439
PVD           0.450718   0.231282   1.949 0.051321 .
Activity     -0.249975   0.113097  -2.210 0.027086 *
Hypertension  0.818823   0.197299   4.150 3.32e-05 ***
Diabetes      0.622495   0.178919   3.479 0.000503 ***
CVD           0.779438   0.192209   4.055 5.01e-05 ***
Anemia        1.226600   0.520013   2.359 0.018335 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1974.3  on 4135  degrees of freedom
Residual deviance: 1320.5  on 4117  degrees of freedom
AIC: 1358.5

Number of Fisher Scoring iterations: 7
```

Model Comparison

Interpret: The final result aboves gives us the most important variables. Model 3 has the lower AIC (1358.5) and hence the best model.

```
> formula(model3)
CKD ~ Age + Female + Racegrp + Unmarried + Weight + BMI + Waist +
    SBP + HDL + LDL + PVD + Activity + Hypertension + Diabetes +
    CVD + Anemia
> formula(model2)
CKD ~ Age + Female + Racegrp + Educ + Unmarried + Income + CareSource +
    Insured + Weight + Height + BMI + Obese + Waist + SBP + DBP +
    HDL + LDL + Total.Chol + Dyslipidemia + PVD + Activity +
    PoorVision + Smoker + Hypertension + Fam.Hypertension + Diabetes +
    Fam.Diabetes + Stroke + CVD + Fam.CVD + CHF + Anemia
```

Prediction on test sample

```
> pred<-predict(model3, newdata=data_out,type="response")
> pred
        6001         6002         6003         6004         6005         6006         6007         6008
0.1483879986 0.0027414058           NA 0.1100339520 0.0005173282 0.0049633258 0.0615737247 0.0011983188
        6009         6010         6011         6012         6013         6014         6015         6016
0.0005871798           NA 0.0098783136 0.0020459303 0.0169761466 0.0448158467 0.8297321657 0.0242184008
        6017         6018         6019         6020         6021         6022         6023         6024
0.0009472520 0.0022792852 0.5282086324 0.0173951979 0.1991658288 0.0021437212 0.0077980124 0.0005910961
        6025         6026         6027         6028         6029         6030         6031         6032
0.2364244474 0.1393929673 0.0024466244 0.0005465623 0.0010211575           NA 0.2231157923 0.1756738428
        6033         6034         6035         6036         6037         6038         6039         6040
0.0108566863 0.4409265717 0.1545567484 0.1921065612 0.0294354010 0.1827858216 0.0005357277 0.0032217203
        6041         6042         6043         6044         6045         6046         6047         6048
0.0580190593 0.2087532905 0.0003099863 0.0006154434 0.0003003420 0.0044617165 0.3879784744 0.0331697864
        6049         6050         6051         6052         6053         6054         6055         6056
0.0740686537 0.3485669181 0.0067772034 0.0621192617 0.0139313293 0.0008501255 0.0016098810 0.0472704579
```

```
> class<-as.factor(ifelse(pred>=0.5,"YES","NO"))
> class
6001 6002 6003 6004 6005 6006 6007 6008 6009 6010 6011 6012 6013 6014 6015 6016 6017 6018 6019 6020 6021 6022
  NO   NO <NA>   NO   NO   NO   NO   NO   NO <NA>   NO   NO   NO   NO  YES   NO   NO   NO  YES   NO   NO   NO
6023 6024 6025 6026 6027 6028 6029 6030 6031 6032 6033 6034 6035 6036 6037 6038 6039 6040 6041 6042 6043 6044
  NO   NO   NO   NO   NO   NO <NA>   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO
6045 6046 6047 6048 6049 6050 6051 6052 6053 6054 6055 6056 6057 6058 6059 6060 6061 6062 6063 6064 6065 6066
  NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO
6067 6068 6069 6070 6071 6072 6073 6074 6075 6076 6077 6078 6079 6080 6081 6082 6083 6084 6085 6086 6087 6088
  NO   NO   NO   NO   NO   NO   NO <NA>   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO <NA>   NO   NO
6089 6090 6091 6092 6093 6094 6095 6096 6097 6098 6099 6100 6101 6102 6103 6104 6105 6106 6107 6108 6109 6110
  NO   NO   NO <NA>   NO <NA> <NA>   NO   NO   NO   NO   NO   NO   NO <NA>   NO   NO   NO   NO   NO   NO <NA>
6111 6112 6113 6114 6115 6116 6117 6118 6119 6120 6121 6122 6123 6124 6125 6126 6127 6128 6129 6130 6131 6132
<NA>   NO   NO <NA>   NO <NA>   NO   NO <NA> <NA>   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO   NO <NA>
6133 6134 6135 6136 6137 6138 6139 6140 6141 6142 6143 6144 6145 6146 6147 6148 6149 6150 6151 6152 6153 6154
  NO <NA>   NO   NO   NO   NO   NO   NO <NA>   NO   NO   NO   NO   NO   NO <NA>   NO   NO   NO   NO   NO  YES
```