|  |
| --- |
| **Chronic Kidney Disease Screening** |
| Predictive Modeling Using R Programming  Group 7 |
| By utilizing regression techniques through R programming, an easy-to-use survey tool is to be created that reflects a regression model. This model is optimized by utilizing available research, a CKD case study report, and data manipulation through R. |
|  |

|  |  |  |
| --- | --- | --- |
| Chronic Kidney Disease Screening  Predictive Modeling Using R Programming Task Offered by The University of Virginia through Darden Business Publishing, the case study has tasked our group with developing a “simple” screening tool to effectively suggest a CKD test. By utilizing 8819 adults 20 years and older. The case set is divided into two sets: A 6,000-case training set, and a validation sample contained 2,819 subjects. The validation sample has 34 variables which includes the CKD indication. This set is what we will utilize to build our model in order to effectively screen our 6,000 subject set. Problem The causes of CKD are known. Diabetes, hypertension, cardiovascular disease, age, and race are among the many factors that contribute to CKD. Although the causes of CKD may be evident, the optimal combination of factors and overall risk faced by a given patient aren’t quantified. This factor combination also does not have to be variables that *cause* CKD, but merely *indicate* it. These challenges require the use of a regression model to analyze subjects, and a basic survey tool to effectively and efficiently screen subjects. Although all subjects could hypothetically be tested, as a CKD test is relatively inexpensive, this survey tool should appeal to the every-day patient while encouraging a visit to a doctor. Solution We will utilize a combination of strategies in order to comprise subsets that indicate CKD in a subject. From the case study we know diabetes, hypertension, and age are proven causes and indicators. Using these as a base, we will create three sub-categories in relation to CKD; Basic Characteristics, Heart Health, and Health Conditions. By then analyzing correlation, we will be able to create our model based on these categories. Regression will be instrumental in finalizing our model and creating a simplistic screening tool. |  | Facts of Disease Diagnosis • • •  Two thirds of all cases of chronic kidney disease are caused by diabetes or high blood pressure.  People age 60 and over are at increased risk for developing CKD.  Certain ethnic groups such as First Nations and Pacific Islanders are at a higher risk of developing the disease.  Kidney Disease can be genetic. |

Missing Data

We used mice method to fill in missing data. MICE stands for multivariate imputation by chained equations. Mice used predictive mean matching to detect and fill in missing values. Regression techniques are used to analyze rows of data and fill in values that are closest to regression predicted values. These values are able to be predicted in continuous, binary, ordered, and un-ordered factor types. Mice method was only used on out sample which did not include CKD diagnosis. The in sample which included CKD diagnosis was left unaltered. Rows without any missing data were utilized to develop the regression model. This was determined the most accurate way to establish a model, rather than developing a model based on mice filled data. This also resulted in a model that included less error (AIC) during practice run throughs of both versions of data. We only used mice method on the sample that did not include CKD diagnosis because we needed to diagnose every subject fully and missing data would make the model increasingly inaccurate.

Creating Prime Groups

Based on case study documentation and external research, support for diabetes, hypertension, and old age are directly related to CKD. Because these are established factors, we choose to use these variables as building blocks to create our model. We utilize these variables to create 3 categories that explain CKD: **Characteristics**, **Heart Health**, and **Conditions**. All 33 variables were then grouped into these three categories.

**Conditions**

Diabetes

Dyslipidemia

Anemia

Stroke

**Heart Health**

SBP

DBP

Hypertension

PVD

CVD

CHF

**Characteristics**

Age

Female

Education

Income

Race Group

Care Source

Insured

Weight

Height

BMI

Obese

Waist

HDL

LDL

Cholesterol

Activity

Poor Vision

Smoker

Family Hypertension

Family Diabetes

Family CVD

Analyzing Correlation

We begin the variable selection with a pre-existing fact based bias. This bias is based on the currently known factors that already influence CKD. We will call these our prime variables as they will be our first priority. We also want to keep our categories in mind so we will hold our three groups in second priority. With these rules, we begin to filter out variables that are highly correlated with each other.

By analyzing all variables and their relationships, we can accomplish a model that includes independent variables. This is important because while doing a regression because correlated x variables can create misleading results in predicting for a y value. Too many variables can also prove to be difficult in achieving meaningful regression insights, which will be extremely important to achieve a simple survey tool.

Age-SBP: 0.551307834

Female-Height: -0.6706917203

Weight-BMI: 0.862178526

Weight-Obese: 0.663605806

Weight-Waist: 0.878096739

BMI-Obese: 0.7780751715

BMI-Waist: 0.8775057564

Obese-Waist: 0.676862352

LDL-Total.Chol: 0.930527688

Fam.Hypertension- Fam.CVD: 0.7832088603

Stoke-CVD: 0.649007670

Age- Hypertension: 0.4793500465

After creating the correlation matrix, these relationships proved to be the highest correlated. Most of which make intuitive sense or are backed by facts. Weight, BMI, Waist, and Obesity are all inter-correlated with each other because BMI and Obesity are measures of weight while waist size is a variable that increases when weight increases. Female and Height were highly negatively correlated, meaning that height was likely to decrease if the subject was a female. Total Cholesterol and LDL are highly related because Total Cholesterol is a measure that includes LDL. Family health conditions were related as well as current health conditions like Stroke and CVD. Age and Hypertension were also highly correlated. By placing priority on our prime groups, we can begin to use correlation to establish smaller and more insightful variable groups.

Final Variable Selection

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Age | Female | Activity | White | Hisp | Other | Hyper | CVD | PVD | Anemia | Diabetes |
| Age | 1.00 | -0.06 | -0.16 | 0.14 | -0.09 | -0.06 | 0.48 | 0.24 | 0.23 | -0.04 | 0.25 |
| Female | -0.06 | 1.00 | -0.14 | -0.02 | 0.01 | 0.02 | 0.00 | -0.08 | -0.01 | 0.10 | -0.03 |
| Activity | -0.16 | -0.14 | 1.00 | 0.00 | 0.06 | -0.02 | -0.09 | -0.08 | -0.07 | -0.01 | -0.10 |
| White | 0.14 | -0.02 | 0.00 | 1.00 | -0.65 | -0.19 | 0.00 | 0.06 | 0.03 | -0.09 | -0.10 |
| Hisp | -0.09 | 0.01 | 0.06 | -0.65 | 1.00 | -0.11 | -0.05 | -0.07 | -0.04 | 0.04 | 0.06 |
| Other | -0.06 | 0.02 | -0.02 | -0.19 | -0.11 | 1.00 | -0.03 | -0.01 | -0.03 | 0.01 | 0.01 |
| Hyper | 0.48 | 0.00 | -0.09 | 0.00 | -0.05 | -0.03 | 1.00 | 0.17 | 0.14 | -0.01 | 0.21 |
| CVD | 0.24 | -0.08 | -0.08 | 0.06 | -0.07 | -0.01 | 0.17 | 1.00 | 0.08 | 0.00 | 0.09 |
| PVD | 0.23 | -0.01 | -0.07 | 0.03 | -0.04 | -0.03 | 0.14 | 0.08 | 1.00 | 0.01 | 0.11 |
| Anemia | 0.00 | 0.10 | -0.01 | -0.09 | 0.04 | 0.01 | -0.01 | 0.00 | 0.01 | 1.00 | 0.00 |
| Diabetes | 0.25 | -0.03 | -0.10 | -0.10 | 0.06 | 0.01 | 0.21 | 0.09 | 0.11 | 0.00 | 1.00 |

**Characteristics**

Age- Proven over 60 has Increased risk of CKD.

Female- Case study data shows females are at slightly more risk.

Activity- Chosen instead of weight and chol. Encompasses total health and is independent.

RaceGroup- Proven that certain race groups are at higher risk.

**Heart Health**

Hypertension- Proven cause of CKD.

CVD- Heart condition

PVD- Heart condition

**Conditions**

Diabetes- Proven cause of CKD

Anemia- Independent condition.

Education, Marriage Status, Income, Insured, Weight Statistics, Cholesterol, Dyslipidemia, Poor Vision, Smoker, Family Statistics, and Stroke were all removed from the model. Correlation played a large part in removing weight statistics as well as Cholesterol. Because of what we know about CKD’s relationships with major health conditions, we placed more focus on conditions rather than health statistics. CKD is a sort of chain reaction in the body that is caused by already failing systems being affected by diabetes and hypertension. We know that poor cardiovascular health and bad diet can result in diabetes or hypertension, so weight statistics and cholesterol were removed to put more focus on already diagnosed conditions.

Family history of diseases were also removed from the model. The main thought process here was that since family history of CKD was not included, we did not choose to put weight on family history of conditions that were already asked for. By removing these variables, we could focus more on the present and future of their health relating to CKD, rather than analyzing their genetic past. The creation of prime variables and groups based on facts resulted in an optimal combination of factors that we could now do regression on. These groups will be important to us as they will be the basis of how we interpret our model as well as construct our survey.

Regression

model=glm(CKD ~ Age+Hypertension+PVD+Diabetes+CVD+Anemia+Activity, family="binomial",data=data\_in)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -7.745008 0.484453 -15.987 < 2e-16 \*\*\*

Age 0.080423 0.006105 13.173 < 2e-16 \*\*\*

Female 0.157592 0.146588 1.075 0.282347

Racegrphispa -0.811539 0.266240 -3.048 0.002303 \*\*

Racegrpother -0.076060 0.562647 -0.135 0.892468

Racegrpwhite 0.110372 0.205424 0.537 0.591067

Activity -0.263058 0.110798 -2.374 0.017586 \*

PVD 0.442028 0.225122 1.963 0.049588 \*

CVD 0.798840 0.189616 4.213 2.52e-05 \*\*\*

Hypertension 0.671116 0.176595 3.800 0.000145 \*\*\*

Anemia 1.100570 0.492743 2.234 0.025512 \*

Diabetes 0.713571 0.171275 4.166 3.10e-05 \*\*\*

---

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: 1357.1 on 4124 degrees of freedom

AIC: 1381.1

phatnew=predict(model, newdata = data\_in, type = "response")

phatnew1

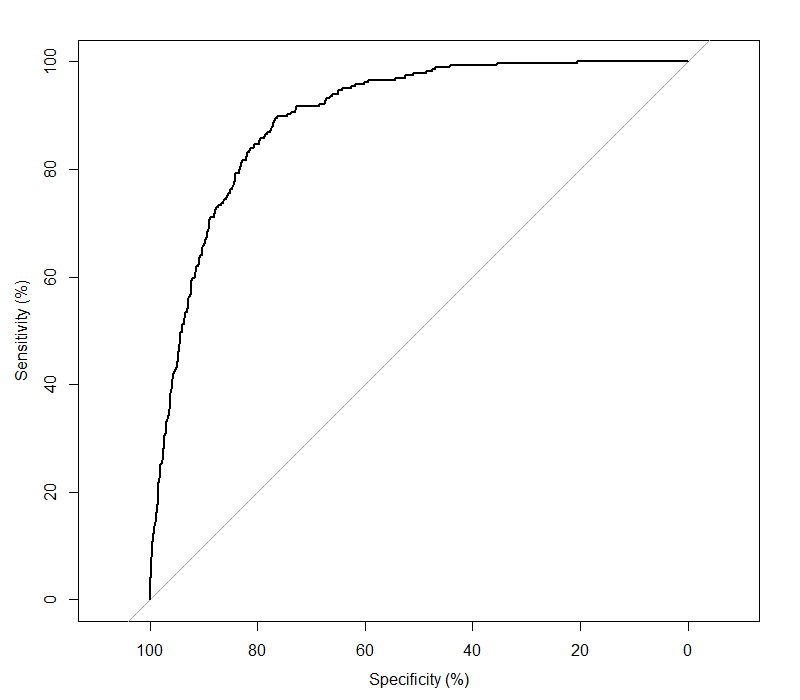
classify=ifelse(phatnew1>.08,1,0)

classify

recall precision accuracy tpr fpr fmeasure tp tn fp fn

0.81955 0.23904 0.82060 0.81955 0.17933 0.37012 218.00000 3176.00000 694.00000 48.00000

By running the regression on the 9 variables, we can see the significance of every variable relating to the prediction of Y. Female showed to be insignificant relative to the model, but was kept in the model due to case study interpretation. Some race groups were insignificant but were kept because of Hispanic significance.



After adjusting the threshold to .08, the accuracy and Total Positive Rate became similar while the false positive rate was relatively low. This regression is determined to be adequate in developing the survey. The total positive rate isn’t high where bias isn’t placed on providing positive CKD diagnosis, while accuracy and false positive rate are at ideal levels. At 81.955% TPR and 17.933 % FPR, the cut point is approximately in the position on the ROC curve that is closest to the optimal point. By optimizing true positive rate and false positive rate, we can achieve model with the ideal threshold to predict CKD diagnosis. Now that the model and threshold are optimized, we can begin classifying CKD patients in the out\_sample and start transforming our regression model into a working simple survey. In order to have a survey that accurately reflects our regression model, we must adapt the regression equation and scale our data in order to maintain model and survey compatibility.

Adapting Regression Model

Based on our logistic regression model, we can find the probabilities of having CKD with the equation: where p means the probabilities of having CKD.

ln(p/(1-p)) = -7.745008+0.080423\*Age+0.157592\*Female-0.811539\*Racegrphispa-0.07606\*Racegrpother+0.110372\*Racegrpwhite-0.263058\*Activity+0.442028\*PVD

+0.79884\*CVD+0.671116\*Hypertension+1.10057\*Anemia+0.713571\*Diabetes

From the value of coefficients, we can make our conclusions:

* As the increasing of age, the probabilities of having CKD is increasing,
* if the gender is female, the probabilities of having CKD is increasing,
* if the race is Hispanic, the probabilities of having CKD is decreasing,
* if the race is other, the probabilities of having CKD is decreasing,
* if the race is white, the probabilities of having CKD is increasing,
* as the increasing of activities, the probabilities of having CKD is decreasing,
* if the patient has PVD, the probabilities of having CKD is increasing,
* if the patient has CVD, the probabilities of having CKD is increasing,
* if the patient has hypertension, the probabilities of having CKD is increasing,
* if the patient has anemia, the probabilities of having CKD is increasing,
* if the patient has diabetes, the probabilities of having CKD is increasing.

Let us set t = ln(p/(1-p)),

P(CKD)=1/(1+e^-t)

Since we set our threshold as 8%, then when p = 8%, we can find t = -2.44,

Multiply 100 to the equation above and round it, we get:

**t=-775+8\*age+16\*Female-81\*Hispan-8\*Other+11\*White-26\*Activity+44\*PVD+80\*CVD+67\*Hypertension+110\*Anemia+71\*Diabetes.**

**t>-244, then say Yes**

Next, we did normalization of t, first step is to eliminate intercept value to get m = t+775,

When m>-244+775=531, then say Yes

Finally, to make the survey easier for patients to take, we set some intervals for age and scale the value ofm, when **m >= 52,** then say **Yes**

With our regression model scaled and normalized, we will provide the following point values for the variables:

**Age**  PTS

Younger than 35 25

Between 35 and 50 years of age 34

Between 51 and 66 years of age 46

Over 66 54

**Gender**

Female 2

**Race groups**

Hispanic -8

White 1

Other -1

**Activity**

Always sit -3

Stand or walk a lot -6

Lift light loads or climb stairs often -9

Heavy work and heavy loads -12

**Disease History**

I have a history of PVD 4

I have a history of CVD 8

I have a history of hypertension 7

I have anemia 10

I have a history of diabetes 7

Conclusion

After comparing the normalized and scaled survey results to the regression model prediction, we were able to achieve 95% accuracy relative to original regression model prediction rates. The survey is expected to screen patients for CKD approximately 82% of the time (True Positive Rate) while falsely screening subjects positively for CKD only 17% of the time (False Positive Rate).

Activity and activity were significant and were determined to be a good base indicator of a subjects health. For race groups and sex, females are at a slightly higher risk while Hispanics are at a lower risk for CKD. Use these 4 base characteristics, matrix tables were created to determine a patients starting point for CKD. From there, health conditions containing diabetes, Anemia, CVD, PVD and Hypertension were assigned points relative to their regression results. By using Age, Activity, Race, and Sex as a risk patients base risk, combinations of health conditions were then responsible for giving a screening result.

Survey

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Under 35 | 35-50 | 51-66 | Over 66 |
| I always sit | 22 | 31 | 43 | 51 |
| I stand and regularly walk | 19 | 28 | 40 | 48 |
| I lift light objects and climb stairs | 16 | 25 | 37 | 45 |
| I lift heavy things and work hard | 13 | 22 | 34 | 42 |

1. Find your age group and typical activity level. The box that corresponds to both characteristics is your starting point. Write your number to the right

Total #1

\_\_\_\_\_\_\_\_\_\_\_

1. Find your sex and race group. Locate the box that satisfies both of your characteristics. Write your number to the right.

Total #2

\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |  |
| --- | --- | --- | --- |
|  | White | Hispanic | Other |
| Female | +3 | -6 | +1 |
| Male | +1 | -8 | -1 |

1. If you have been diagnosed with any given condition, add the corresponding points and total them to the right.

Total #3

\_\_\_\_\_\_\_\_\_\_\_

|  |  |
| --- | --- |
| CVD | +8 |
| PVD | +4 |
| Hypertension | +7 |
| Anemia | +10 |
| Diabetes | +7 |

Add or subtract all of the columns on the right side and place the total in the box below.

Final Total

\_\_\_\_\_\_\_\_\_\_\_

If your final total is greater than or equal to **52 points**, you are at risk for Chronic Kidney Disease. A doctor visit to test for CKD is recommended.