BMED 6700
Biostatistics

Chronic Kidney Disease Prediction

Nirav N Shah Sini Nagpal Parisa Yousefi Zowj

Abstract

The chronic kidney disease (CKD) describes the heterogeneous disorders that affect the functionality and the structure of kidney. The progression of CKD may lead to complications such as high blood pressure, anemia, weak bones, poor nutritional health and damage and if not treated, it may eventually lead to kidney failure which requires dialysis or kidney transplant. Therefore, it is vital to detect it and prevent its progression in the early stages. The existing markers for CKD diagnosis can only identify high-risk patients and do not improve the understanding of the pathogenesis and progression of disease. In this study, the objective is to analyze and determine the possible predictive factors that are reliable and easily measured in laboratory environment for the early recognition and prevention of the progression of the disease. In this report, we try to use GLM along with various model selection techniques to build a prediction model for CKD detection. We validate the model and test its predictive power via the generalized linear models.

Contents

1 Introduction

In the past two decades, there has been a rapid increase in the growth of big data and this data can be analyzed to predict effective treatments. Intensive Care Units in hospitals are now known to use disease prediction models to evaluate their treatment options. Keeping this in mind, we have built an effective prediction model for Chronic Kidney Disease.

The chronic kidney disease (CKD) describes the heterogeneous disorders that affect the functionality and the structure of the kidney. Chronic Kidney Disease describes the gradual loss of kidney functions which includes filtering waste and excess fluids within the body to urine. In extreme stages, the kidney is unable to perform the function efficiently and hence toxic waste along with electrolytes can increase in the blood causing the pressure within the blood to increase, thus affecting the cardiovascular status of the body.

Figure 1 shows the conceptual model of CKD that was developed by the National Kidney Foundation's Kidney Disease Quality Outcome Initiative (NKF-KDOQI) in 2002 and revised and adopted by the an international consensus in 2005. The model specifies the development, progression and complications of CKD. The progression of CKD may lead to complications such as high blood pressure, anemia, weak bones, poor nutritional health and damage and if not threated, it may eventually lead to kidney failure which requires dialysis or kidney transplant [2].

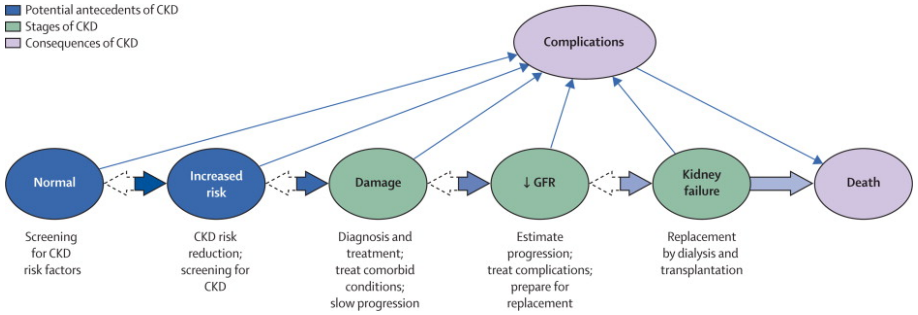

Figure 1: Conceptual Model for Chronic Kidney Disease [3]

In developed countries, CKD is associated with old age, diabetes, hypertension, obesity cardiovascular disease and with diabetic glomerulosclerosis and hypertensive nephroscerosis, whereas in developing countries, the common causes of CKD are glomerular and tubulointerstitial diseases which result from infections and exposure to drugs and toxins. In any case, it is difficult to diagnose the disease in early stages due to its asymptomatic nature and due to several comorbidities are associated with CKD [1].

The prevalence of CKD in U.S.A is about 14 % with about 661,000 Americans having kidney failure and 400,000+ on dialysis. End Stage Renal Disease (ESRD), which CKD is a part of is prevalent about 2.7 times more in African Americans and 1.4 times greater in Native Americans. This makes it a very important disease to study.

For each stage of CKD, there are different treatments available such as slowing the progression of the disease, preventing the progression, treating the complications of decreased glomerular filtration rate, reducing the cardiovascular disease risk factors and treating the cardio vascular disease.

Two key markers for the diagnosis of CKD are albuminuria and glomerular filtration rate (GFR). Albuminuria is a pathological condition in which the protein, albumin, is abnormally present in urine. Persistent albuminuria suggests some level of kidney damage. Although the albumin screening is a good predictor for CKD, the measurements of albuminuria cannot detect patients with kidney disease with 100 % guarantee [2].

In conjunction with albuminuria, GFR is an important factor to measure the severity of CKD. Basically, GFR measures the total filtration rate of functioning nephrons in the kidney. The best way to measure GFR is using plasma or urinary clearance of an exogenous filtration marker. Due to the complexity, it is not a routinely performed procedure. Instead, GFR is estimated from serum creatinine in combination of age, sex, ethnic origin and body size. Figure 2 shows the prognosis of CKD made by GFR and albuminuria, which suggest a categorization of CKD into five stages.

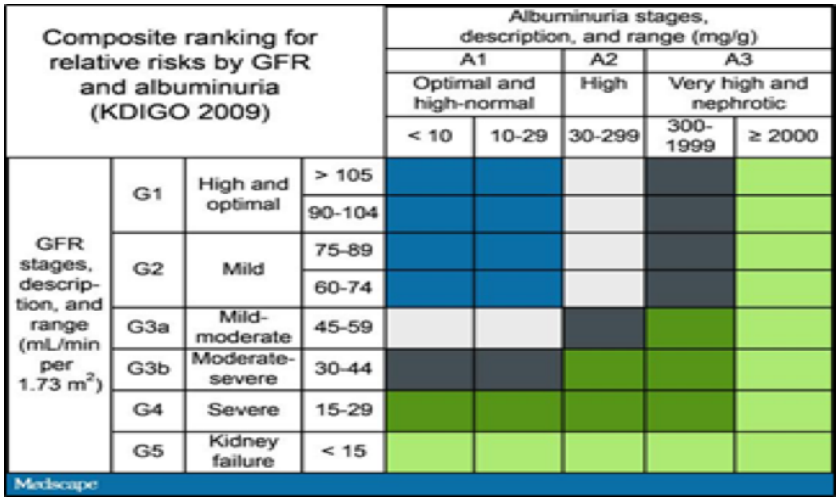

Figure 2: Prognosis of CKD by GFR and Albuminuria [4]

The five stages of CKD are:

Stage 1: Kidney damage with normal kidney function (estimated GFR $\geq$90 mL/min per 1.73 m^2) and persistent ($\geq$3 months) proteinuria

Stage 2: Kidney damage with mild loss of kidney function (estimated GFR 60-89 mL/min per 1.73 m^2) and persistent ($\geq$3 months) proteinuria

Stage 3: Mild-to-severe loss of kidney function (estimated GFR 30-59 mL/min per 1.73 m^2)

Stage 4: Severe loss of kidney function (estimated GFR 15-29 mL/min per 1.73 m^2)

Stage 5: Kidney failure requiring dialysis or transplant for survival. Also known as ESRD (estimated GFR$<$15 mL/min per 1.73 m^2)

The common approach to treat CKD is by performing dialysis which simulates the function of kidneys by filtering the waste and water from the blood. However, for long-term and old-age, this is a difficult task to perform. Kidneys perform other functions such as release of hormones (erythropoietin, renin and calcitriol) which are difficult to simulate thus affecting the production of blood and maintenance of calcium within the body. Thus, the earlier diagnosis of CKD is very critical for the health of the patient which can be pursued by using our prediction models discussed. Although both markers are good predictors of CKD, they mainly enable to identify high-risk patients. These markers do not improve the understanding of the pathogenesis and progression of CKD. Therefore, CKD calls the attention of public health approach for identifying new markers for prevention and early detection of the disease.

In this study, the objective is to determine and analyses the possible predictive factors that are reliable and easily measured in laboratory environment for the early recognition and prevention of the progression of the disease.

2 Data Acquisition and Description

The data set was obtained from U.C. Irvine Machine Learning Repository [5]. Dr. P. Soundarapandian M.D.,D.M who is a senior consultant nephrologist at Apollo Hospitals, India is the provider of the data and L. Jerlin Rubini, a research scholar at Alagappa University created the data set. The data was donated on July 3th 2015 and there has not been any publication based on this data set.
The data set consists of 24 variables (11 numeric, 13 nominal) which are related to medical measurements of 400 patients and one target (response) variable that represents whether a patient has a chronic kidney disease or not. Additionally, the age of patients is also present in the data set. Table ..... in Appendix summarizes the name of predictors, their abbreviations used throughout the project and the levels/unit of nominal/numerical variables.

3 Exploratory Data Analysis

As an initial step of statistical modeling, an explanatory data analysis is carried out to get familiar with the data set. The important aspect of any data is to look at the missing data. One drawback of this data set is that there are many missing values in it. Among 400 patients, the data set is complete for only 158 patients.
Figure 3 shows the counts of missing values for each of those 24 variables. For some of the variables there is more than 30% missing data present. As this number was high, we omitted the

observations which had one or the other data present. This reduced the data set size from 400 to 158 patients.Out of 158, 115 patients are diagnosed with CKD and the remaining 43 patients are not.

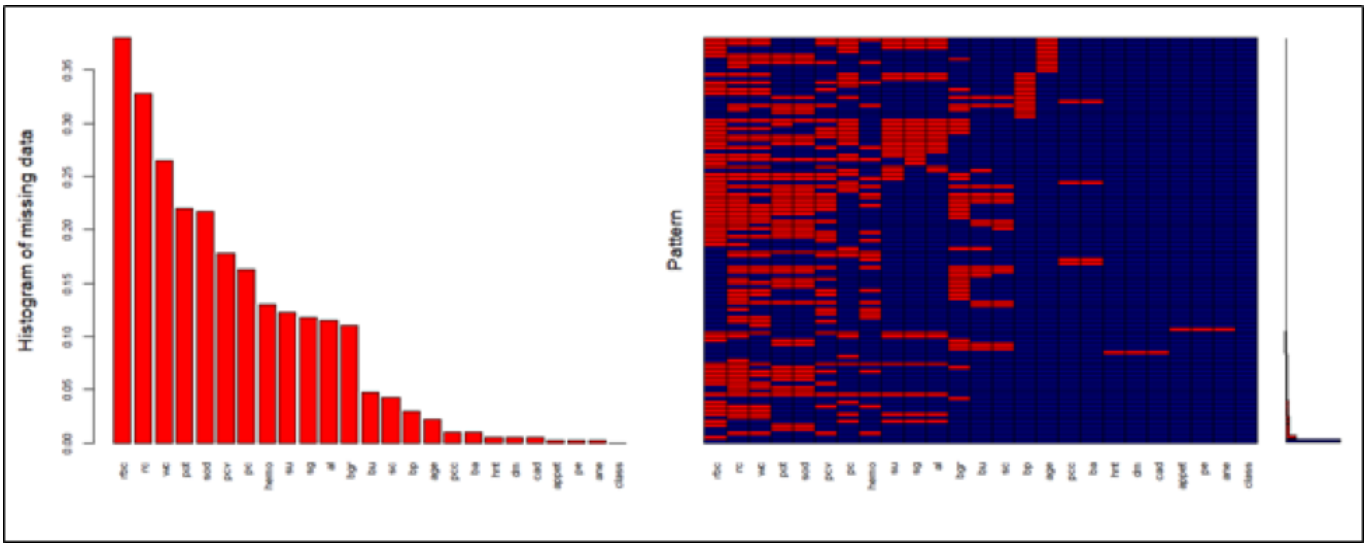

Figure 3: Numerical Predictors and CKD Diagnosis

Another way to deal with these missing values is to impute those.

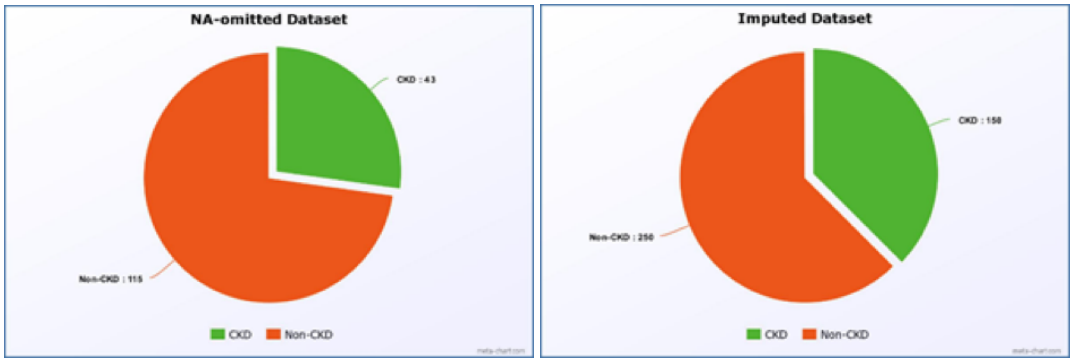

Figure 3: Numerical Predictors and CKD Diagnosis

Figure 3 displays the relationship between the response variable and the 11 numerical predictors. To begin with, there are couple of outliers that show themselves for different predictors such as for potassium and red blood cell. However, since the response is binary and the logistic regression does not allow to do model diagnostics, the outliers are kept in the data set. One important reason keeping the outliers is that they might reveal important relationships for the diagnosis of CKD. The boxplots for numerical predictors, other than blood pressure and potassium suggest that these predictors have different distributions for patients with CKD and patient without CKD.

For blood glucose randim, blood urea, sodium, hemoglobin, packed cell volume and red blood cell count, the interquartile range do not even overlap for the current dataset. This is a clear indication that some of these predictors might be statistically significant in predicting CKD.

To investigate the dependency between those numerical predictors, the correlation matrix is used given in Figure 4. The scale from blue to red represents the perfect positive correlation to perfect negative correlation.

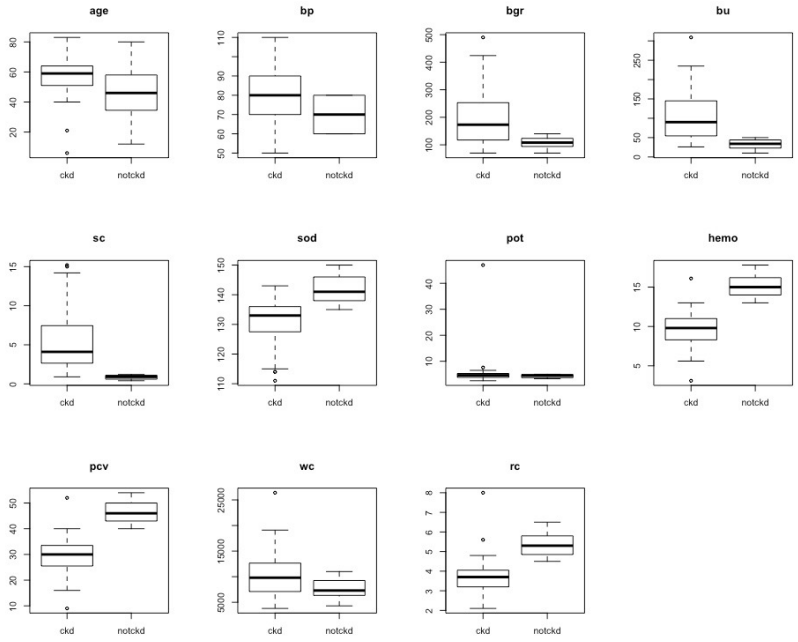

Figure 3: Numerical Predictors and CKD Diagnosis

The take-away from this figure is that blood urea and the serum creatinine are highly positive correlated while these are highly negative correlated with hemoglobin, packed cell volume and red blood cell count. This multicollinearity may cause convergence issue for the logistic regression.

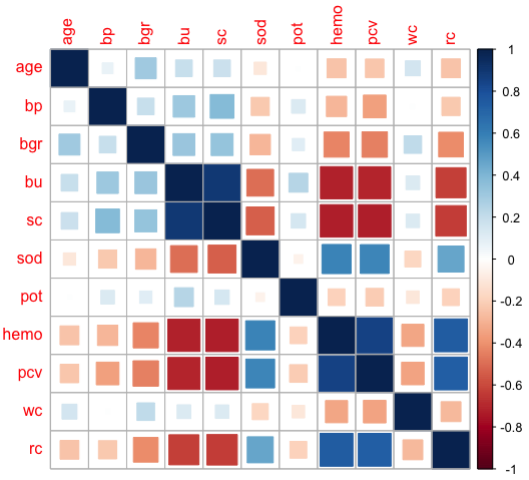

Figure 4: Correlation Matrix for Numerical Predictors

Finally, the relationship between the nominal predictors and the response is visualized using mosaic package in R. Figure 5 displays the relationship between sugar and response and Figure 8 shows the relationship for the remaining nominal predictors. The patients who are not diagnosed with CKD have only one level of sugar, while patients diagnosed with CKD have different levels of sugar. Expect for specific gravity, this perfect separation is observed for the remaining nominal predictors in the dataset.

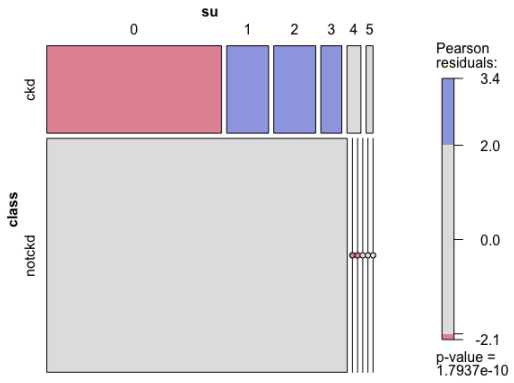

Figure 5: Relationship between Sugar and CKD Diagnosis

4 Statistical Modeling

4.1 Assessment of Current Diagnosis Technique in Literature

As mentioned before, Albuminuria (Al) is a pathological condition in which the protein, albumin, is abnormally present in urine. Persistent Albuminuria suggests some level of kidney damage. Another important factor to measure the severity of CKD is GFR which is estimated by Serum Ceratitine (sc) in combination of age, sex, ethnic origin and body size. To replicate the findings for the current practice, a logistic regression was fit with predictors in Model 1 as just Albuminuria and in Model 2 as serum ceratitine and age where age is the on additional information present in dataset. A low correlation of 0.19 between the two predictors is observed and according to the model, serum creatitine is statistically significant (p-value ¡0.001) but not age, i.e. for one unit increase in serum creatinine the log odds of being CKD increases by 8.81 taking into age into consideration. Hence, the model for the current dataset confirms the current diagnosis technique in literature. Serum creatitine is a good marker for the diagnosis CKD. Table 1 is summarizes the regression output.

	Estimate	**Std. Error**	**z value**	**p-value**	**significance**
(Intercept)	a	a	a	a	a
Albuminuria					

Table 1: Regression Output of The Model for the Current Diagnosis

	Estimate	**Std. Error**	**z value**	**p-value**	**significance**
(Intercept)	a	a	a	a	a
Age					
Serum Ceratitine					

Table 2: Regression Output of The Model for the Current Diagnosis

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	**significance**
(Intercept)	-12.663990	3.393768	-3.732	0.000190	***
age	-0.003689	0.036328	-0.102	0.919107	
sc	8.808505	2.576618	3.419	0.000629	***

Table 1: Regression Output of The Model for the Current Diagnosis

[1]

[1]Significance Codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

4.2 Model with Numerical Predictors

When fitting a logistic regression model considering all nominal and numerical predictors, the algorithm does not converge due to perfect separation where the response separates a combination of categorical predictor variables completely. Excluding the categorical variables, the algorithm for the logistic regression model still experience the convergence problem due to high correlation among predictors. An initial heuristic approach is to select subset of six predictors that have low correlation among them. Figure 6 shows the correlation between those selected predictors, age, blood pressure, blood glucose random, sodium, potassium and white blood cell count.

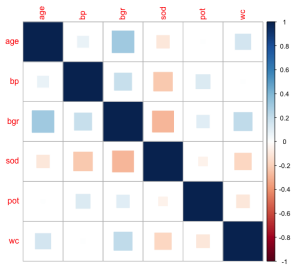

Figure 6: Correlation Matrix for Six Predictors

When fitting a logistic regression with these predictors, blood glucose random, sodium, white blood cell count and potassium are found as significant predictors at significance level of 0.05 as shown in Table 2. The AIC of the model comes out to be 43.53.

	Estimate	Std. Error	z value	Pr(>|z|)	significance
(Intercept)	51.3925918	26.5339154	1.937	0.05276	.
age	0.0843070	0.0505696	1.667	0.09548	.
bp	0.0604326	0.0516018	1.171	0.24155	
bgr	0.0382040	0.0147591	2.589	0.00964	**
sod	-0.5927399	0.2080570	-2.849	0.00439	**
pot	1.7926161	0.7331282	2.445	0.01448	*
wc	0.0007702	0.0002809	2.742	0.00611	**

Table 2: Regression Output of The Model with Six Predictors

- For one unit increase in sodium, the log odds of being CKD versus not decreased by 0.593.
- For one unit increase in white blood cell count, the logs of being CKD versus not increased by 0.001.
- For one unit increase in blood glucose random, the logs of being CKD versus not increased by 0.038.

- For one unit increase in potassium, the logs of being CKD versus not increased by 1.793.

Among all the significant predictors, potassium seems to be the more important in increasing the odds of being diagnosed with CKD while sodium is the only significant predictor whose increase decreases the odds of being diagnosed with CKD.

4.3 Model Selection

As there is no 'best' model and the model with six predictors in Section 4.2 is chosen by randomly selecting one of the predictors among highly correlated ones, a more methodological way to assess the explanatory and predictive power of predictors is to conduct model selection. Three different model selection techniques are compared. These are stepwise regression, lasso and elastic net.

4.3.1 'Both' Stepwise Regression

Stepwise regression is run in both (forward and backward) directions that toggles between one step of forward selection and one step of backward selection. A step is only performed if it lowers AIC, otherwise it is skipped. The algorithm stops if two consecutive steps are skipped. Although 'both' stepwise regression is not greedy, it does not guarantee finding the model with the lowest AIC. It is possible that the 'best' model might require exchanging sets of multiple variables but stepwise can only move one step at a time. In that sense, 'both' stepwise regression results in a local minimum, rather than the global minimum. Another drawback of stepwise regression is that it searches a large space of possible models. Hence it is prone to overfitting the data. In other words, stepwise regression will often fit much better in sample than it does on new out-of-sample data.

Since the algorithm for model with all numerical predictors does not converge, the 'both' stepwise regression is performed for the model with six predictors. This technique suggests a model with 5 predictors except the blood pressure. Including the age, all predictors are significant at significance level of 0.05. For one year increase in age, the log odds of being diagnosed as CKD increased by 0.093. Although age is statistically significant, it does not have a large influence on the odds of begin CKD. Confusion matrix given in Table 3 shows that the model has a accuracy of 93% (153/158) for predicting CKD. It should be noted that the accuracy is a high and might be caused by overfitting as explained above.

		Predict		Total
		Not CKD=0	CKD=1	
Actual	Not CKD=0	113	2	115
	CKD=1	3	40	43
Total		116	42	158

Table 3: Confusion Matrix

4.3.2 Lasso

In order to enhance the predictive accuracy, the lasso (least absolute shrinkage and selection) that performs both variable selection and regularization simultaneously is applied as an alternative technique considering all 11 numerical predictors. The lasso estimator $\hat{\beta}(\lambda)$ is the value of β that solves:

$$\min_{\beta \in \mathbb{R}^p} (\sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_1) \tag{1}$$

where $\lambda > 0$ and $\|\beta\|_1 = \sum_{j=1}^{p} \beta_j$ is the L_1 norm of the vector β. L_1 is a penalty that captures sparsity. This is a convex optimization problem which has a unique value for $\hat{\beta}(\lambda)$ that depends on λ. The selected model for a given λ is,

$$S(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0\} \tag{2}$$

where the constant λ can be chosen by cross-validation. R package *lars* is used to produce the lasso estimator. If a non-zero coefficient ever hits zeros, it is removed from the active set of predictors and the joint direction is recomputed.

Figure 7 shows the Lasso outputs in R. The optimal model design is selected with respect to the number of predictors in the model that minimizes the *Mallow* Cp criteria. As seen from the graph on the left, the smallest Cp value (10.14) is reached with 10 predictors (df=11). However, a more parsimonious model with 7 variables (df=8) has a Cp value of 14.81 which deviates slightly from the criteria $Cp \approx p$, where p is the number of predictors in the selected models.

If chosen, the model with 10 predictors has all the numerical variables except potassium. On the other, the model with 7 predictors includes hemoglobin, packed cell volume, blood glucose random, sodium, red blood cell count, serum creantinine and white blood cell count. As mentioned before, due to the high correlation of the predictors, the algorithm for the models Lasso suggests does not converge.

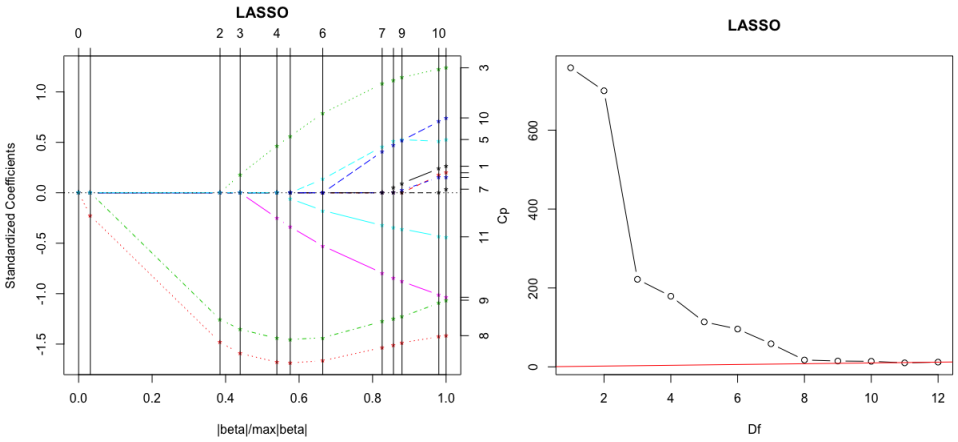

Figure 7: Model Selection-Lasso

4.3.3 Elastic Net

Another model selection technique to deal with highly correlated predictors is the elastic net regularized method. Recall that, if there is a group of highly correlated variables, Lasso tends to select one variable in the group and ignore the others. To overcome this limitation of Lasso, the elastic net adds a quadratic part ($\|\beta\|^2$) to the Lasso penalty ($\|\beta\|_1$), which - when used alone- is the ridge regression. The estimates from the elastic net method are defined by

$$\hat{\beta} = \underset{\beta}{argmin}\, f(y, X; \beta) + \lambda_2 ||\beta||^2 + \lambda_1 ||\beta||_1, \tag{3}$$

or

$$\hat{\beta} = \underset{\beta}{min}\, f(y, X; \beta) \quad s.t. \quad J(\beta) \leq t, \tag{4}$$

where $J(\beta) = \alpha||\beta||^2 + (1-\alpha)||\beta||_1$ with $\alpha = \lambda_2/(\lambda_2 + \lambda_1)$. $f(y, X; \beta) = ||\mathbf{y} - \mathbf{X}\beta||^2$ for normal distribution and is negative log-likelihood of binomial distribution in case of logistic regression.

Since the response variable is binary (0=Not CKD and 1=CKD), binomial distribution with 100 different λ values with $\alpha = 0.5$ for the elastic net application in R was selected. The optimal tuning parameter was obtained by performing 10-fold cross validation. This method was performed with only numerical predictors and also with both numerical and categorical predictors separately.

Model with Numerical Variables

Table 4 shows the prediction confusion matrix based on the optimal model, Table 5 gives the coefficient matrix and Table 6 is the model output with the selected predictors. The model has 98 % accuracy in prediction. The values of the coefficients for our 11 numerical predictors indicate that potassium with coefficient 0 does not play an important role in predicting the CKD. Excluding the predictors with coefficient close to zero, a model with predictors, hemoglobin, red blood cell count

and sodium is fitted. Among these three predictors, only hemoglobin is statistically significant at significance level of 0.01.

		Predict		Total
		Not CKD=0	CKD=1	
Actual	Not CKD=0	115	0	115
	CKD=1	3	40	43
Total		118	40	158

Table 4: Confusion Matrix

	age	bp	bgr	bu	sc	sod
Coefficient	0.0051	0.0026	0.0096	0.0076	0.1408	−0.0669
	pot	hemo	pcv	wc	rc	
Coefficient	0	−0.2666	−0.0756	0.0001	−0.1990	

Table 5: Coefficient Matrix

	Estimate	Std. Error	z value	Pr(>|z|)	significance
(Intercept)	52.7176	18.6216	2.831	0.00464	**
hemo	-1.8357	0.5898	-3.113	0.00185	**
rc	-0.4904	0.5114	-0.959	0.33761	
sod	-0.1979	0.1375	-1.439	0.15012	

Table 6: Model Output

Model with Numerical and Categorical Variables

Table 7 shows the prediction confusion matrix based on the optimal model with numerical and categorical variables and Table 8 gives the coefficient matrix. Again, due to perfect separation, the algorithm for the logistic regression model suggested by elastic net does not converge.

		Predict		Total
		Not CKD=0	CKD=1	
Actual	Not CKD=0	115	0	115
	CKD=1	0	43	43
Total		115	43	158

Table 7: Confusion Matrix

	age	bp	bgr	bu	sc	sod
Coefficient	0	0	0.0018	0.0002	0.0303	-0.0371
	pot	hemo	pcv	wc	rc	
Coefficient	0.0000	-0.1360	-0.0383	0.0001	-0.1321	
	sg (1.01)	sg (1.015)	sg (1.02)	sg (1.025)	al (1)	al (2)
Coefficient	0.8777	0.8635	-0.0180	0	0	0.9200
	al (3)	al (4)	al (5)	su (1)	su (2)	su (3)
Coefficient	0.2490	1.2375	0	0	0	0
	su (4)	su (5)	rbc (normal)	pc (normal)	pcc (present)	ba (present)
Coefficient	0	0	-0.9292	-0.7839	0	0
	hnt (yes)	dm (yes)	cad (yes)	appet (poor)	pe (yes)	ane (yes)
Coefficient	1.5092	0.6301	0	0	0	0

Table 8: Coefficient Matrix

5 Discussion

In this study, a thorough data analysis and statistical modeling are performed to determine which variables have predictive power on the diagnosis of CKD. Three model selection techniques were tested. Table 9 summarizes the models considered, the significant predictors in these models, AIC values and finally the p-value of Hosmer-Lemeshow goodness of fit test which is performed by splitting the observations into 10 groups according to their predicted probabilities. Model 1 is the model that assess the current practice for CKD diagnosis, Model 2 represents the model with 6 predictors, Model 3 is the model suggested by stepwise regression and finally Model 4 is the one found by elastic net method. Among the four models, Model 3 has the smallest AIC value and largest goodness of fit, while Model 1 has a p-value of 0.055 which is enough to reject the null hypothesis. In other words, Model 1 is not a good fit for the current dataset. To sum up,

testing predictors *age*, *blood glucose random*, *sodium*, *potassium* and *white blood cell count* might be an alternative way to to diagnose CKD in early stages.

Model Name	Model	Significant Predictors	AIC	p-value of GoF
Model 1	y~age+sc	sc	35.16	0.055
Model 2	y ~ age+bp+bgr+sod+pot+wc	bgr , sod , pot , wc	43.53	0.7425
Model 3	y ~ ag+bgr+sod+pot+wc	age , bgr , sod , pot , wc	34.03	0.7998
Model 4	y ~ hemo+rc+sod	hemo	36.45	0

Table 9: Model Comparison

6 Limitations and Future Study Directions

There are several limitations to be noted regarding this study. To begin with, there is an inbalance in the size of patient populations with diagnosis of CKD and no CKD (43/115). The original data set consists of a large patient population of 400 patients with missing values in predictors. Since the scope of this project is not imputing the missing values, a subset of 158 patients out of 400 patient is selected without any missing values. However, the imputation of missing values would result in a larger patient population that the current one. Replicating the results found with the current data set on the larger data set would imply the accuracy/fit of the selected models.

Another limitation is the absence of patient demographics such as gender and ethic group in the data set. As stated in literature, the diagnosis of CKD by estimating the serum ceratitine is done in combination of the patients demographic information such as gender and race. Having these predictors and their interactions with other predictors in the statistical modeling may reveal interesting associations with the diagnosis of CKD.

References

[1] Levey, Andrew S and Coresh, Josef. *Chronic Kidney Disease*.The Lancet. 379(9811): 165-180, 2012.

[2] (2016,April 15). Retrieved from https://www.kidney.org/kidneydisease/aboutckd.

[3] National Kidney Foundation. K/DOQI Clinical Practice Guidelines for Chronic Kidney Disease: Evaluation, classification, and stratification.Am J Kidney Dis, 39 (suppl 1),S1–S266,2002.

[4] Matshushita K, van de Velde M, Astor BC, et al, for the Chronic Kidney Disease Prognosis Consortium. Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. The Lancet. 375: 2073–81, 2010.

[5] (2016,March 1).Retrieved from http://archive.ics.uci.edu/ml/datasets/ChronicKidneyDisease.

A Appendix

Attribute Name	Abbreviation	Unit	Category
Age	age	years	Numerical
Blood Pressure	bp	mm/Hg	Numerical
Specific Gravity	sg	(1.005,1.010,1.015,1.020,1.025)	Nominal
Albumin	al	(0,1,2,3,4,5)	Nominal
Sugar	su	(0,1,2,3,4,5)	Nominal
Red Blood Cells	rbc	(normal,abnormal)	Nominal
Pus Cell	pc	(normal,abnormal)	Nominal
Pus Cell Clumps	pcc	(present,notpresent)	Nominal
Bacteria	ba	(present,notpresent)	Nominal
Blood Glucose Random	bgr	mgs/dl	Numerical
Blood Urea	bu	mgs/dl	Numerical
Serum Creatinine	sc	mgs/dl	Numerical
Sodium	sod	mEq/L	Numerical
Potassium	pot	mEq/L	Numerical
Hemoglobin	hemo	gms	Numerical
Packed Cell Volume hemo	pcv	-	Numerical
White Blood Cell Count	wc	cells/cumm	Numerical
Red Blood Cell Count	rc	millions/cmm	Numerical
Hypertension	hnt	(yes/no)	Nominal
Diabetes Mellitus	dm	(yes/no)	Nominal
Coronary Artery Disease	cad	(yes/no)	Nominal
Appetite	appet	(yes/no)	Nominal
Pedal Edema	pe	(yes/no)	Nominal
Anemia	ane	(yes/no)	Nominal
Class	class	(ckd,notckd)	Nominal

Table 10: Data Description

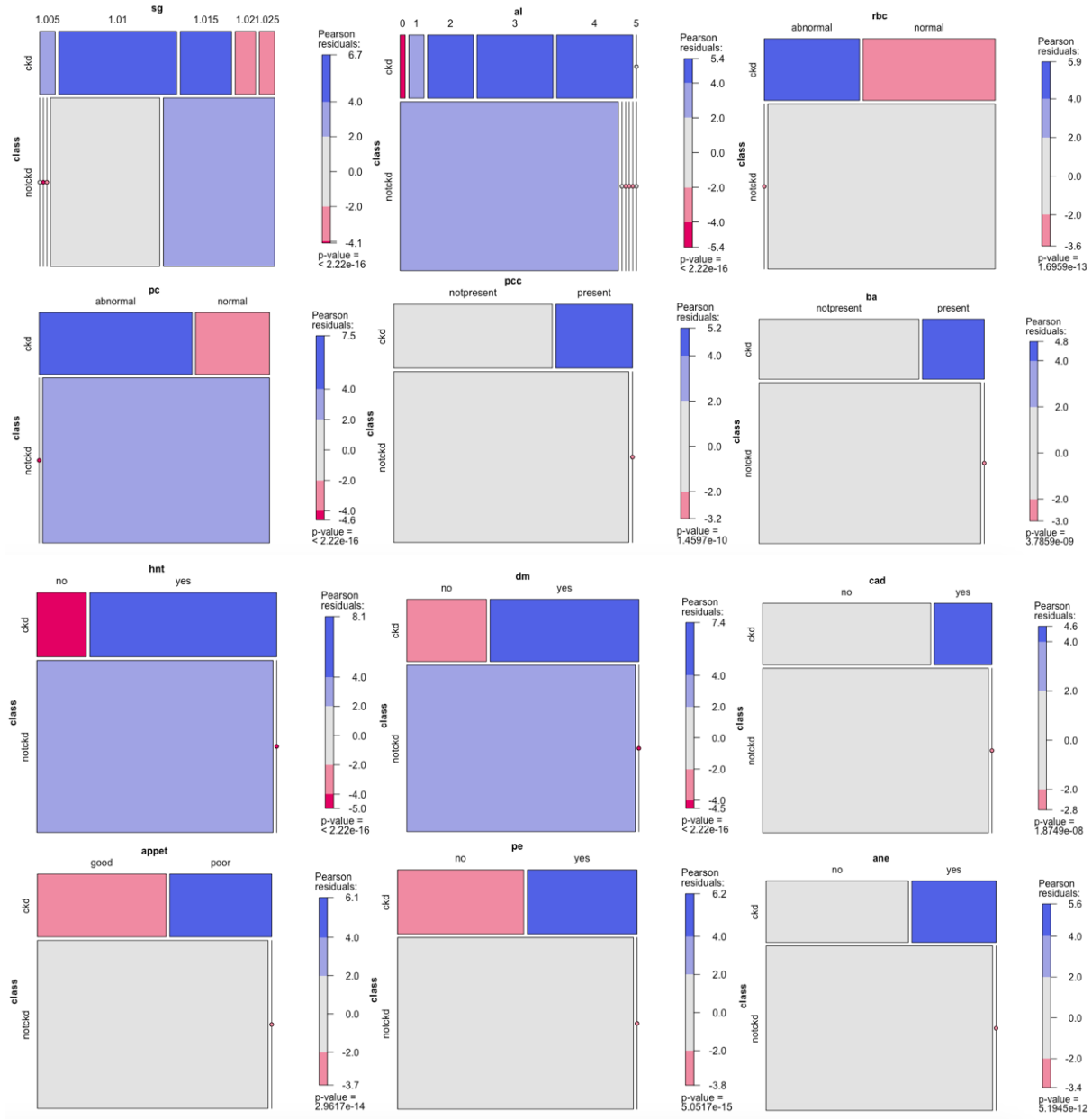

Figure 8: Nominal Predictors and Response Variable