Reducing the Influence of Confounders on Predictive Models

Ricardo Brito Alves[1]

[1]Department of Electrical Engineering – Pontifícia Universidade Católica de Minas Gerais – Belo Horizonte – MG – Brazil

rbalves@sga.pucminas.br

Abstract. *The analysis of Big Data has become so important with the progressive increase of the information stored in digital media. Extracting more value from diversified and unstructured data is really challenging. With the help of predictive models, it is possible to find new patterns and trends that could be innovation bases. Predictive models need to have a relevant reliability rate to aid us in decision-making processes. In this context, this article discusses the influence of confounding variables on predictive models and proposes techniques for identifying and minimizing their effect. Through a database with information collected in a hospital, it was possible to construct a predictive model, to identify possible confounding variables, to apply a technique to minimize its influences and to evaluate the accuracy of the model through machine learning techniques. The result was an efficient prediction model.*
Keywords: Big Data, Predictive Model, Confounders, Multicollinearity, Machine Learning.

1. Introduction

The objective of this study was to apply techniques that identify and minimize the influence of the confounding variables so that the predictive model obtain maximum efficiency in the prediction process. Confusion can arise when in an unbalanced sample an interfering variable distorts the association between an exposure variable and a response variable, changing the strength or even the direction. Build a predictive model is not simply to write an equation but rather to perform more consistent analyzes on the data and their relationships. As result in our study, through the use of Pearson and Spearman Correlation, VIF - Variance Inflation Factor, linear regression by the outcome it was possible to minimize the effect of a potential confounding variable and through classification techniques such as SVM and Logistic Regression, we verified the efficiency of the model after reducing the influence of the confounding factor.

The Data mining has been used to exploit large amounts of data in search of consistent patterns, such as association rules or time sequences. Once systematic relationships between variables have been detected, it is possible to obtain subsets of data [Han et al. 2011, Maurizio 2011]. Predicting or inferring about causality are two of the great scientific motivations of verifying the statistical association between variables. As part of this process, we have the predictive analysis that has been applied in many areas and is able to use data, algorithms, and machine learning techniques in an attempt to predict future situations. The increasing information digitization by society, ease of storage and the possibility of processing large volumes of data has made this analysis more accessible to those who seek to know it [Waller and Fawcett 2013].

Predictive modelling uses statistics to predict outcomes and in most often the event one wants to predict is in the future, but predictive modelling can be applied to any type of unknown event, regardless of when it occurred [Leatherman et al. 2018].

A predictive model is nothing more than a mathematical function that, applied to a set of data, can identify hidden patterns and can enable the prediction of scenarios that interest us with a relative margin of accuracy. Predictive models can either be used directly to estimate a response (output) given a defined set of characteristics (input), or indirectly to drive the choice of decision rules [Steyerberg 2009].

In many cases, the model is chosen on the basis of detection theory to try to guess the probability of an outcome given a set amount of input data. Models can use one or more classifiers. Most of the regression models can be used for prediction purposes. Broadly speaking, there are two classes of predictive models: parametric and non-parametric. A third class, semi-parametric models, includes features of both. Parametric models make ”specific assumptions with regard to one or more of the population parameters that characterize the underlying distributions”, while non-parametric regressions make fewer assumptions than their parametric counterparts [Sheskin 2011].

As part of predictive modelling, it is necessary to analyze variables and their correlation. A confounding variable or confounding factor is a variable that influences both a dependent variable and an independent variable, causing a spurious association. It is a situation where the effects of two variables are difficult to separate from each other [Austin 2011]. To be considered as confusion, the variable must be associated with the outcome and be associated with exposure, and not be part of the causal chain linking exposure to outcome. Variables with a bias of confusion are potential effect modifiers, that is when the effect of an exposure on an outcome varies according to the level of a third variable. Therefore, when between two exposures there is a potentiated of one to another, we can say that the effect has been changed. One factor alone has one effect, but in the presence of another, its effect is increased. The union of two factors results in a different risk than simply the effect of one plus the effect of the another. Unlike most biases, confusion bias can be controlled after data collection, provided they have been collected in a way that allows for such control. When the effect is suspected, a stratified analysis of the confounding factor should be done, showing that the risks between the different strata are similar to each other but different from the gross risk. By reducing the effects of bias, the possibility of misclassification is reduced.
In the scope of statistics, for a sample, we can have two types of errors, random or by chance, that we can say are associated with the precision of the measurements and systematic or bias, which are associated with the validity of the measurements. Bias is a random distortion as a result of some sampling process. Also known as ”bias deviation”, it consists of the difference between the average value of a statistical estimator and the value that it intends to estimate. Often, to correct the deviation, the estimator is changed. Systematic errors or bias can be classified as selection bias, information bias, and confounding.

- Selection bias: the measure of association estimated in the study is distorted because of the way in which individuals are selected to make up the study population.
- Information Bias: The measure of association estimated in the study is distorted due to errors in the way information on exposure and/or disease has been obtained.

- Confusion or confounding: part of the observed association arises from the existence of one or more variables, called confounding.

This article was structured in introduction, related works, contextualization, materials and methods, results and conclusion. In the contextualization, it was trying to clarify more the subjects approached. In materials and methods, there is a more detailed description of the methods used, divided into three stages: building a predictive model and reduction of the characteristics to be worked, identification of potential confounding variables and the application of a technique to correct the effect of confounding variables using linear regression. The results have been discussed obeying the same three stages.

2. Related Works

[Li et al. 2011] present us in this article a Support Vector Machine classifier that can correct the prediction for observed confounding factors. This is achieved by minimizing the statistical dependence between the classifier and the confounding factors.
[Li and Zhang 2015] analyzes the existence of confounders such as population structure in genome-wide association study makes it difficult to apply machine learning methods directly to solve biological problems.
[Low et al. 2016] show us a observational studies from EHR in real time, particularly in emergencies, rapid confounder control methods that can handle numerous variables and adjust for biases are imperative. This study compares the performance of 18 automatic confounder control methods.
[Schnitzer et al. 2016] investigates the appropriateness of the integration of flexible propensity score modeling (nonparametric or machine learning approaches) in semiparametric models for the estimation of a causal quantity, such as the mean outcome under treatment.
[Berk et al. 2018] provide an alternative methods draw on work in econometrics and statistics from several decades ago, updated with the most recent thinking to provide a way to properly work with misspecified models. They show how asymptotically, unbiased regression estimates can be obtained along with valid standard errors.

3. Materials and Methods

Data from patients were collected at a hospital in India. We attempted to minimize the effects of confounding factors, by the multivariate analysis method where the predictor variables are analyzed simultaneously so that the effect of each variable is adjusted for the effect of the others. Thus, it tried to identify the direct effect of each variable in the prediction of the outcome, an effect that is independent of other variables. This is called as an independent association. Confusion is as an intermediate variable with a false association between two other variables. Adjusting for this confounding variable, we will know about the relationship between predictor variable and outcome. This concept of a need for adjustment for confusion by multivariate analysis applies to the construction of a predictive model. The multivariate analysis makes the adjustment for the predictive variables, determining an independent association, a necessary condition for the variable to be part of a predictive model or considered one of the causes of the outcome.

The database was extracted from UCI - Machine Learning Repository [http://mlr.cs.umass.edu/ml/ 2017], which aimed to classify two groups, carrier and non-

carrier to chronic kidney disease (CKD). CKD is a progressive pathology, with a high mortality rate. Discovering the disease before reaching the chronic phase and progressing to dialysis is fundamental to guarantee quality of life and to increase survival [Jena and Kamila 2015, Sinha and Sinha 2015, Kumar 2016]. According to the UCI - Machine Learning Repository data set extracted over a 2-month period in India, there are 400 patient records with 24 attributes to predict CKD or not CKD (age, blood pressure, relative density, albumin, sugar, red blood cells, pus cell, pus cell clusters, bacteria, random blood glucose, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia).

As a statistical analysis tool, was used public domain software R [Team 2014]. R software was used mainly in multivariate analyses, which estimated the association of each independent variable with the dependent variable, after adjusting for the effects of all other variables.

As mentioned before, this study was divided into three stages: building a predictive model and selection of variables more adherent to the predictive model, identification of potential confounding variables and the application of a technique to correct the effect of confounding variables. However, before all, one of the important steps for model development was to eliminate records with undefined values, leaving 158 sets of data for use in this work without missing values.

Among the 25 variables in the database, the variable to be predicted is the class variable. The others are explanatory or independent variables: age, blood.pressure, specific.gravity, albumin, sugar, red.blood.cells, pus.cell, pus.cell.clumps, bacteria, blood.glucose.random, blood.urea, serum.creatinine, sodium, potassium, hemoglobin, packed.cell.volume, white.blood.cell.count,, red.blood.cell.count, hypertension, diabetes.mellitus, coronary.artery.disease, appetite, pedal.edema, anemia.

3.1. Stage 1

The first stage was about the predictive model construction. Linear regression was used in order to build an equation to estimate the conditional (expected value) of a variable y, given the values of some other variables. In a set of 25 variables, the ”class” dichotomic variable was considered as the expected value that has CKD and not CKD information. The others were considered as independent or explicative variables. Thus, the model takes on the standard form (1) which describes a line with slope β_k and y-intercept β_0 involving the error term ϵ_i:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i \quad (1)$$

To optimize the predictive model it was necessary to analyze all independent variables in order to select the ones with the most relevant characteristics of the predictive model. It was an important step in the application of machine learning methods. Data sets are often described with many variables and some of these variables may be irrelevant to classification, and their use is a disadvantage for constructing models.Firstly, analyses of the set of variables were performed using the Pearson correlation coefficient and its respective p-value and the Spearman correlation coefficient in the search for the set of

characteristics more adherent to the predictive model. The correlation coefficient measures the degree to which two variables tend to change together. The coefficient describes the strength and direction of the relationship. Pearson's correlation evaluates the linear relationship between two continuous variables while Spearman's correlation evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together but not necessarily at a constant rate.

Then a new analysis was done using CFS - Correlation Feature Selection. This analysis could provide a numerical estimate of the importance of resource. The random forest classification algorithm could be run without setting parameters and provides a numerical estimate of the importance of the resource. It is a joint method in which classification is performed by voting on multiple nonindependent weak classifiers.

3.2. Stage 2

The second stage had the objective of identifying potentially confounding variable following the steps below.
- Identify variables with strong correlations, which tend to be redundant in the sense of adding little to the model and being potential confounding variables.
- Adjust relative risk analysis of more than 10 % relative to the gross relative risk, which suggests that it is a confounding variable. By the use of the linear regression, it is possible to analyze the influence of the confounding variables. The first linear regression model considers all variables, including those of confusion. The second linear regression model does not consider the confounding variables. If the estimates of the two models vary by more than 10 %, it is suggested that the variable is confusing [Austin 2011].
- Identify multicollinearity that is defined as the presence of a high degree of correlation between the independent variables [García et al. 2015]. Presence of multicollinearity means that there is a presence of collinearity between the variables, so any plane along the data dispersion axis will be unstable and results in the same sum of squares of the error. Variance Inflation Factor (VIF) was use to detect multicollinearity, assuming that the variables are centered and standardized, we have $R = (X^T X)^{-1}$ in which the diagonal elements of this matrix are called VIF and represent the increment of the variance due to the presence of multicollinearity [Montgomery et al. 2012]. The VIF can be calculated using the equation (2).

$$VIF_j = \frac{1}{1 - R_j^2} \ \ and \ \ j = 1, 2, ..., p \tag{2}$$

Where p is the number of predictor variables; R^2 is the multiple correlation coefficients resulting from the regression of Xj on the other $p - 1$ regressors.

3.3. Stage 3

The third stage was the application of a technique to correct the effect of the confounding variable by the use of linear regression by the outcome. The linear regression model was used to correct the effect of the confounding variable.

NSUB is the number of subjects, NVAR is the number of variables and NCOV is the number of covariables. For each variable y in Y and nCovariables we have:

- y = observed value for the dependent variable Y i-nth level of the independent variable nCOV.
- b = regression constant, which represents the intercept of the line with the y-axis.
- a = regression coefficient, which is the variation of Y as a function of the variation of a unit of variable a.
- C = coefficient of regression covariates.

Y is a NSUBxNVAR matrix C is a NSUBxNCOV matrix

For each variable y in Y, with NSUB samples:

Using the linear regression, we look for the coefficients of b, a and C (3).

$$y = b + a[1]\ C[1] + \cdots + a[nCOV]\ C[nCOV] \tag{3}$$

Correction of the variable y to each line using the equation (4):

$$y = y + mean(y) - b - a[1]\ C[1] - \cdots - a[nCOV]\ C[nCOV] \tag{4}$$

To check results at the corrected database, it was necessary to apply classification techniques to compare the accuracy of the predictive model. For this were used two technics, SVM and Logistic Regression. In technics, the corrected database was used, removing the predicted variable. New classifications were made that could be compared with the original classification. In the logistic regression, the database was separated on a training base and a test base, being the test base used for classification. Results of logistic regression was checked by confusion matrix.

In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.
Computing the SVM classifier amounts to minimizing an expression of the form (5):

$$[\frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i(\vec{w}.\vec{x}_i - b))] + \lambda \left\|\vec{w}\right\|^2 \tag{5}$$

Where the parameter λ determines the tradeoff between increasing the margin-size and ensuring that the $\vec{x}_i$ lie on the correct side of the margin. Thus, for sufficiently small values of λ , the second term in the loss function will become negligible, hence, it will behave similar to to the hard-margin SVM, if the input data are linearly classifiable, but will still learn if a classification rule is viable or not.

The logistic model or logit model is a statistical model that is usually taken to apply to a binary dependent variable. In regression analysis, logistic regression or logit regression is estimating the parameters of a logistic model. More formally, a logistic model is one where the log-odds of the probability of an event is a linear combination of independent or predictor variables. The two possible dependent variable values are often labelled as ”0” and ”1”, which represent outcomes. The binary logistic regression model

can be generalized to more than two levels of the dependent variable: categorical outputs with more than two values are modelled by multinomial logistic regression, and if the multiple categories are ordered, by ordinal logistic regression.

The logistic regression can be understood simply as finding the β parameters (6):

$$\begin{cases} 1 & \beta_0 + \quad \beta_1 x + \quad \varepsilon \quad > 0 \\ 0 & else \end{cases} \tag{6}$$

Where A(x,t) and d (x,t) =: D(x,t) are constant rank singular, possibly rectangular matrix functions which are in some sense well matched.

4. Results

As mentioned, the discussion of the results were divided according to the three stages of the study.

4.1. Stage 1

Firstly, analyses of the set of variables were performed using the Pearson correlation coefficient and its respective p-value and the Spearman correlation coefficient in the search for the set of characteristics more adherent to the predictive model. The correlation coefficient measures the degree to which two variables tend to change together.

The correlation between variables was evaluated using the Pearson correlation coefficient and the Spearman correlation coefficient, which measure the degree of correlation and the direction of this correlation. 17 of the 24 independent variables had significant correlations and significance with p-value < 0.05 as we can see in table 1.

Staying with many relevant variables, it was important to apply another method, the Correlation Feature Selection (CFS), which provides a numerical estimate of the importance of the resource. In this case, the result indicated the variables in order of importance for the predictive model. 23 attributes was confirmed important: albumin, anemia, appetite, bacteria, blood.glucose.random and 18 more. It is possible to verify all attributes by importance using the CFS method, which is represented in Figure 1.

Table 1. Pearson and Spearman Correlations

Class	Pearson Correlation	p-Value	Spearman Correlation
specific.gravity	0.790101503	2.20E-16	0.697503862
albumin	-0.925816188	2.20E-16	-0.97014838
sugar	-0.510615424	7.18E-09	-0.58507828
red.blood.cells	0.586390502	5.72E-13	0.586390502
pus.cell	0.775387573	2.20E-16	0.775387573
pus.cell.clumps	-0.50991462	7.75E-09	-0.50991462
blood.glucose.random	-0.591217296	2.88E-13	-0.44888816
blood.urea	-0.677610639	2.20E-16	-0.64139796
serum.creatinine	-0.820232963	2.20E-16	-0.7432218
sodium	0.640901871	2.20E-16	0.61444839
packed.cell.volume	-0.82798329	2.20E-16	0.740012593
red.blood.cell.count	0.736366841	2.20E-16	0.701921008
hypertension	-0.856334238	2.20E-16	-0.85633424
diabetes.mellitus	-0.758965495	2.20E-16	-0.75896549
appetite	-0.604622205	2.20E-16	-0.6046222
pedal.edema	-0.622572806	2.20E-16	-0.62257281
anemia	-0.548947121	8.16E-11	-0.54894712

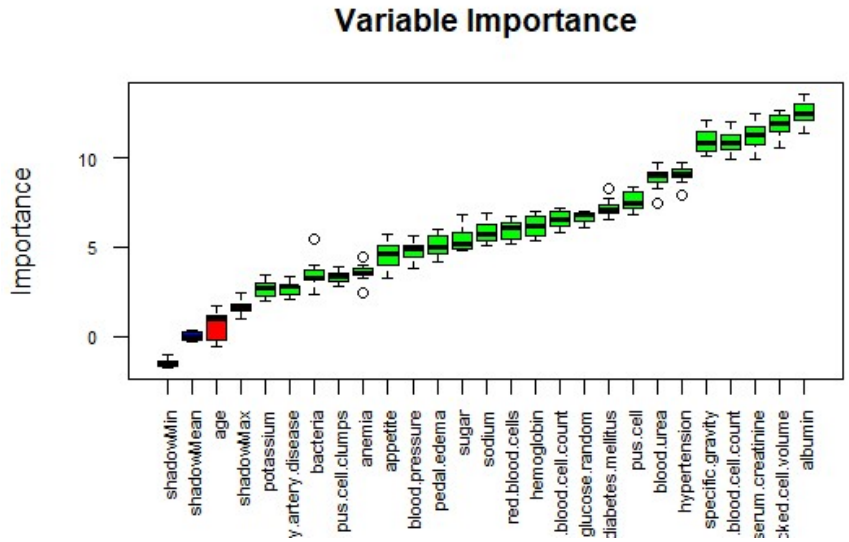

Figure 1. Correlation Feature Selection

Below in table 2 we have the list of the variables in order of significance and relevance. For the purpose of the study, was selected the variables considered most relevant to compose the predictive model.

Table 2. Order of Significance and Relevance

Significance	Relevance	Variable
1	-	blood.pressure
2	4	specific.gravity
3	1	albumin
4	15	sugar
5	14	red.blood.cells
6	8	pus.cell
7	-	pus.cell.clumps
8	-	bacteria
9	10	blood.glucose.random
10	7	blood.urea
11	3	serum.creatinine
12	13	sodium
13	-	potassium
14	11	hemoglobin
15	2	packed.cell.volume
16	12	white.blood.cell.count
17	5	red.blood.cell.count
18	6	hypertension
19	9	diabetes.mellitus
20	-	coronary.artery.disease
21	-	appetite
22	-	pedal.edema
23	-	anemia

4.2. Stage 2

The variables with strong correlations tend to be redundant, we first evaluate the albumin variable. According to the table 3, it can be seen albumin related to the dependent variable and other independent variables.

Table 3. Correlation Results

	albumin
class	-0.92582
specific.gravity	-0.71233
pus.cell	-0.75296
blood.urea	0.66194
serum.creatinine	0.802923
packed.cell.volume	-0.77553
hypertension	0,796876

The adjusted relative risk analysis showed that linear regression models considering and disregarding the variable albumin had large variations in the regression coefficients. Below we have the regression coefficients for the variables blood.pressure and specific.gravity.

- Blood.pressure variating from -3.575e-04 to -6.333e-04.
- Specific.gravity variating from 5.902e-02 to 7.697e-02.

To detect multicolinearity, VIF was used. A maximum VIF above 10 indicates that multicollinearity may be influencing least squares estimates and below, it is possible to see albumin variable with VIF = 10.506670. The highest VIF results is represented in table 4.

Table 4. VIF Results

Variable	**VIF**
albumin	10.506670
hypertension	7.423139
packed.cell.volume	5.221808
diabetes.mellitus	4.779383
pus.cell	4.612907
serum.creatinine	4.322035

In Figure 2, the presence of collinearity between the variables can be detected, so any plane along the data dispersion axis will be unstable.

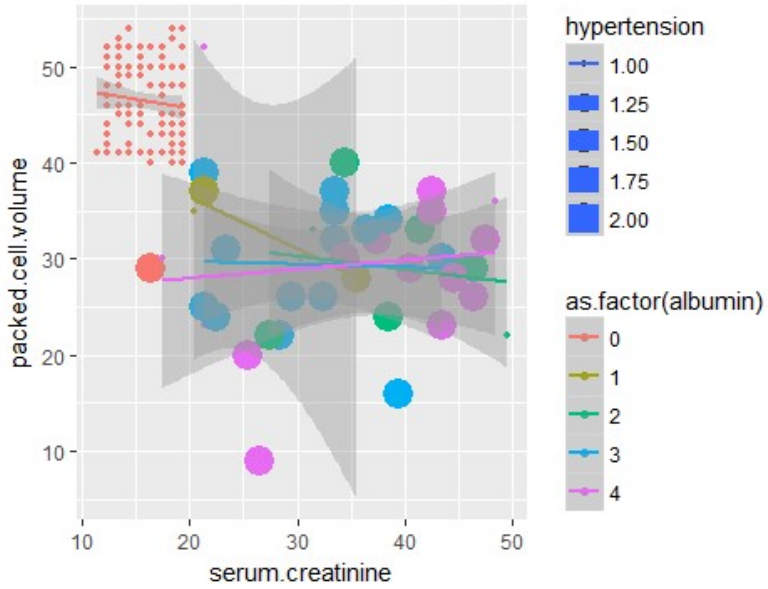

Figure 2. Comparing packed.cell.volume x serum.creatinine x hypertension x albumin

4.3. Stage 3

A technique of correction of the effect of the confounding variable was applied successful using the linear regression by the outcome to correct the effect of the confounding variable. With a set of corrected data, classifications techniques were applied, obtaining the results of the study.

Both technics SVM and logistic regression reached similar results. The accuracy of the predictive model was proved.

Applying the SVM technic in the original database, the classification in the presence of the confounder variable had 2 items incorrectly classified. In the absence of the confounder variable, 3 items were incorrectly classified. Both misclassifications cases were about CKD.

Applying the SVM technic in the corrected database, the result was 43 registers correctly classified as CKD and 115 registers correctly classified as NOT CKD in a total of 158 registers. None of the registers were incorrectly classified as we can see in table 5.

Table 5. SVM Results

	CKD	NOT CKD
CKD	43	0
NOT CKD	0	115

Applying Logistic Regression, the result was 40 registers correctly classified as CKD and 28 registers correctly classified as NOT CKD in a total of 68 registers. None of the registers were incorrectly classified. Using the original database, we had misclassifications cases, as we can see in table 6.

Table 6. Logistic Regression Results

	CKD	NOT CKD
CKD	40	0
NOT CKD	0	28

5. Conclusion

The development of the present study made it possible to analyze techniques in order to identify and reduce the influence of confounders in predictive models by the use of a reduced set of data. In addition, the study also shows us that to build a predictive model is not simply to write an equation but rather to perform more consistent analyzes on the data and their relationships. In our study, through the use of Pearson and Spearman Correlation, VIF - Variance Inflation Factor, linear regression by the outcome it was possible to minimize the effect of a potential confounding variable and through classification techniques such as SVM and Logistic Regression, we verified the efficiency of the model after reducing the influence of the confounding factor. These techniques were used in three different databases of different sizes and the results were similar. For the other two databases tested, the McNemar test was used, which showed strong evidence of a statistically significant association. In general, it was possible to perceive that the accuracy of the model was maintained. Given the importance of the subject, it is necessary to deepen in other techniques of analysis of the relationship between variables. This analysis allows us to know more about the data, their relationships and the results we want to predict, so the resources described here are relevant in the discussion of the creation of predictive models. As a contribution, the techniques described in this study can be applied and different databases with a classification of unbalanced classes and for future work, it would be interesting to extend the studies in others in techniques that identify the relevance and the relation of the variables within the model.

References

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

Berk, R., Brown, L., Buja, A., George, E., and Zhao, L. (2018). Working with misspecified regression models. *Journal of Quantitative Criminology*, 34(3):633.

García, C., García, J., López Martín, M., and Salmerón, R. (2015). Collinearity: Revisiting the variance inflation factor in ridge regression. *Journal of Applied Statistics*, 42(3):648–661.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

http://mlr.cs.umass.edu/ml/ (2017). Uci machine learning repository.

Jena, L. and Kamila, N. K. (2015). Distributed data mining classification algorithms for prediction of chronic-kidney-disease. *International Journal of Emerging Research in Management &Technology*, 4(11):110–118.

Kumar, M. (2016). Prediction of chronic kidney disease using random forest machine learning algorithm. *International Journal of Computer Science and Mobile Computing*, 5(2):24–33.

Leatherman, E. R., Santner, T. J., and Dean, A. M. (2018). Computer experiment designs for accurate prediction. *Statistics and Computing*, 28(4):739.

Li, L., Rakitsch, B., and Borgwardt, K. (2011). ccsvm: correcting support vector machines for confounding factors in biological data classification. *Bioinformatics (Oxford, England)*, 27:i342–i348.

Li, L. and Zhang, S. (2015). Orthogonal projection correction for confounders in biological data classification. *International journal of data mining and bioinformatics*, 13:181–196.

Low, Y. S., Gallego, B., and Shah, N. H. (2016). Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records. *Journal of comparative effectiveness research*, 5:179–192.

Maurizio, M. (2011). Data mining concepts and techniques. *domenica*.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to linear regression analysis*, volume 821. John Wiley & Sons.

Schnitzer, M. E., Lok, J. J., and Gruber, S. (2016). Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. *The international journal of biostatistics*, 12:97–115.

Sheskin, D. J. (2011). Parametric versus nonparametric tests. *International Encyclopedia of Statistical Science*.

Sinha, P. and Sinha, P. (2015). Comparative study of chronic kidney disease prediction using knn and svm. *International Journal of Engineering Research and Technology*, 4(12):608–12.

Steyerberg, E. (2009). Lessons from case studies. *Clinical Prediction Models*.

Team, R. C. (2014). R: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria. 2013.

Waller, M. A. and Fawcett, S. E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2):77–84.