Journal of Data Science,17(1). P. 55-80 ,2019
DOI:10.6339/JDS.201901_17(1).0003

AUTOMATING DATA ANALYSIS METHODS IN EPIDEMIOLOGY

George Choueiry[*,2], Pascale Salameh[*, 2, 3]

[*]*Department of Epidemiology & Biostatistics, School of Public Health, Lebanese University, Fanar, Lebanon*

[2]*School of Pharmacy, Lebanese University, Hadath, Lebanon*

[3]*School of medicine, Lebanese university, Hadath, Lebanon*

Abstract

Technological advances in software development effectively handled technical details that made life easier for data analysts, but also allowed for non-experts in statistics and computer science to analyze data. As a result, medical research suffers from statistical errors that could be otherwise prevented such as errors in choosing a hypothesis test and assumption checking of models. Our objective is to create an automated data analysis software package that can help practitioners run non-subjective, fast, accurate and easily interpretable analyses. We used machine learning to predict the normality of a distribution as an alternative to normality tests and graphical methods to avoid their downsides. We implemented methods for detecting outliers, imputing missing values, and choosing a threshold for cutting numerical variables to correct for non-linearity before running a linear regression. We showed that data analysis can be automated. Our normality prediction algorithm outperformed the Shapiro-Wilk test in small samples with Matthews correlation coefficient of 0.5 vs. 0.16. The biggest drawback was that we did not find alternatives for statistical tests to test linear regression assumptions which are problematic in large datasets. We also applied our work to a dataset about smoking in teenagers. Because of the open-source nature of our work, these algorithms can be used in future research and projects.

Keywords: automation; computer software; machine learning; normal distribution

[*]Corresponding author: George Choueiry
email: georgecymail@gmail.com

1. Introduction

Statistical errors are abundant in medical literature, and it can be proven that most claimed research findings are false (Ioannidis, 2005). In particular, conditions and assumptions of hypothesis tests are rarely checked and reported (Hanif & Ajmal, 2011). This can be done either with statistical testing, which does not work well especially with small sample sizes (Barker & Shaw, 2015), or with graphical methods which have their own problem of introducing subjectivity into the analysis, especially because researchers are already biased towards getting statistically significant results (Dwan, Gamble, Williamson, Kirkham, & Reporting Bias, 2013).

On the other hand, advancements in the field of statistical computing coupled with the power of open-source software has lead the statistical programming language R to grow in popularity in epidemiological research (haine, 2017). The community has built more than 12,000 packages for R that help solving a large variety of problems and provide data analysts with cutting edge technology *. However huge this growth was in the last years, it only impacted a minority of researchers who know how to code, as R is command driven and has a steep learning curve (Ozgur, Colliau, Rogers, Hughes, & Myer-Tyson, 2017) which is a serious disadvantage for non-programmers (Khan, 2013).

Recent developments with graphical user interface software improved the situation a lot since the data analyst was required to write code and deal with mathematical details to get things done. Packages like SPSS statistics opened the door for non-professional statisticians to work with data. However, this did not reduce the number of statistical errors in medical research (Ercan et al., 2007; Felson, Cupples, & Meenan, 1984) and it did not solve the subjectivity problem in data analysis.

* The comprehensive R archive network, www.cran.r-project.org

We hypothesise that by automating assumption checking and result interpretation of the most used statistical tests and models in epidemiology we can create a more objective analysis and reduce the rate of errors in the literature related to these subjects. Also, by creating a graphical user interface (GUI) for the R programming language, we can bring cutting edge technology to non-experts in programming and statistics. Therefore, our goal is to build an open-source web and desktop based GUI application that automates data analysis. Our secondary objective is to use this software to analyze the predictors of smoking among teenage students.

2. Methods

In this section we will describe how our software handles outliers, imputes missing data and automates the bivariate and multivariable analyses.

2.1 Outlier detection

The software detects and handles 2 types of outliers: numerical and categorical outliers. Numerical outliers are defined as values that fall more than 1.5 times the interquartile range above the third quartile or below the first quartile. The software offers the user the option to replace them with missing values. Categorical outliers are defined as variables that have at least one category that constitutes less than 10% of the total sample. The software detects and flags those variables so the user can choose to include them or not in subsequent analyses.

2.2 Missing values treatment

To impute missing values we implemented random forests, available through the R package missForest (Stekhoven, 2013).

2.3 Bivariate analysis

The software performs parametric and nonparametric hypothesis testing. Parametric tests include: Student t-test, adjusted t-test, Chi-squared, Pearson's correlation, one-way analysis of variance (ANOVA). Non-parametric tests include: Fisher's exact test, Mann-Whitney U test, Spearman's rank correlation, Kruskal-Wallis one-way analysis of variance.

In choosing between parametric and nonparametric alternatives, our software automatically examines corresponding conditions and reports how the decision was made. Many parametric tests require the variance to be approximately the same across compared groups. Therefore to confirm equal spread, our software computes the standard deviation of each group and no standard deviation should be 1.5 times larger than the other. When comparing several groups, such as with ANOVA, the ratio of the largest standard deviation to the smallest should be less than 1.5 (Falissard, 2011).

Another important assumption of parametric tests is that, for sample sizes smaller than 30 (Falissard, 2011), the studied variable must have a normal distribution. In our software, this is done by using a statistical model that predicts normality from features extracted from the histogram and normality tests. To train the model, we ran a simulation using 6000 samples with sample sizes ranging from 8 to 50, drawn at random from the following symmetric and asymmetric distributions: Normal(0,1), Uniform(0,1), Beta(2,2), Beta(6,2), Beta(2,1), Beta(3,2), t(5), t(7), t(10), Gamma(1,5), Gamma(4,5), $\chi 2(4)$, $\chi 2(20)$. These distributions were selected to cover various values of skewness and kurtosis. For every sample we computed:The adjusted sum of squared errors (SSE_{adj}): after drawing the histogram, we overlaid a normal distribution curve and the distance between the center of each bin and the curve is squared and added, then the sum is adjusted for the number of bins, as shown in equation 1

$$SSE_{adj} = \frac{1}{n} \sum_{i=1}^{n} \Delta_i \tag{1}$$

Where:
"n" is the number of bins of the histogram
"Δ_i" is the error of the ith bin

1. The histogram is split into 3 parts: the mean height of the bins is calculated for each part, then the vertical distance between the first part and the second is calculated, and between the second and the third to obtain the 2 distances: dist1 and dist2

2. The skewness and kurtosis of the distribution

3. The results of normality tests: Shapiro-Wilk (SW), Kolmogorov-Smirnov (KS), Anderson-Darling (AD) and Jarque-Bera (JB)

4. The dependent variable (target) was labeled manually by looking at the histogram and deciding whether the distribution is normal or not

We used as independent variables: "SSE_{adj}", "dist1", "dist2", "skew", "kurt", "SW", "KS", "AD" and "JB" and as dependent variable: "target". In order to avoid over-fitting, we split the sample into training and testing sets (70/30 split). Using the training set, we trained 2 models, a logistic regression and a random forests (Ho, 1995) whose parameters were tuned by cross-validation. The normality prediction thresholds for both models were set using the receiver operating characteristic (ROC) curve. We then used the testing set to compare the performance of the 2 models using raw accuracy, Matthews correlation coefficient and the area under the curve (AUC) of the ROC curve. Matthews correlation coefficient is a measure that takes into account true and false positives and negatives, works well with unbalanced size classes (Boughorbel, Jarray, & El-Anbari, 2017), it takes values between -1 (total disagreement between prediction and observation) and 1 (total agreement). Logistic regression and random forests were also compared to the Shapiro-Wilk test and a base model (that always predicts the majority class – non-normality) using raw accuracy and Matthews correlation coefficient. The model with the best performance was implemented in the software to predict normality for sample sizes smaller than 30.

2.4 Variable selection for multivariable models

We implemented several methods for automatic variable selection:

(1) Focused principal component analysis: is a graphical display that shows correlations of independent variables with the dependent and with each other (Falissard, 1999).

(2) Bivariate analysis: the user can then chose to include in the multivariable model independent variables that have a p-value < 0.2 (Bouyer, 2009).

(3) Forward stepwise regression

(4) Backward stepwise regression

(5) LASSO regression (Tibshirani, 1996).

2.5 Linear regression

Linear regression assumptions are automatically checked using statistical tests. If one of them is violated, the software tries using logarithmic transformation on the dependent variable and re-checks these conditions. The Shapiro-Wilk test (Shapiro & Wilk, 1965) is used to check for normality of the residuals since it is considered to have the best power for a given significance when compared to other normality tests (Mohd Razali & Yap, 2011). The Breusch-Pagan test (Breusch & Pagan, 1979) checks homoscedasticity of residuals and the Durbin-Watson test (Durbin & Watson, 1950, 1951) checks for serial correlation (Barker & Shaw, 2015).

An important assumption of linear regression is the linear relationship between every independent variable and the dependent variable. This assumption is hardly met with real life data. Most of the time we need to intervene to make sure this assumption is not violated. We implemented 2 strategies to make this correction:

(1) Transforming the independent variable: When running a linear regression, for each numerical independent variable, the software tries:
   (a) Fitting a regression line using the dependent and independent variable (without any transformation)
   (b) Fitting a regression line using the dependent and the logarithmic form of the independent variable.

(2) Then it compares the sum of squared errors of the 2 models. The independent variable form that has the smallest error – therefore a more linear relationship with the dependent variable, will be used in the final model. This enables the software to do automatic logarithmic transformation when needed. The software also helps the user by interpreting the final model, since the interpretation of a logarithmically transformed coefficient is not straightforward (Yang, 2012).

(3) Cutting the variable is a valid alternative if the relationship between dependent and independent variables has a breakpoint that represents an abrupt change and the sample is large enough. The software automatically chooses the threshold using a brute force algorithm – trying every possible data value as threshold and fitting 2 regression lines before and after the threshold, the computer then computes the sum of squared errors of the two lines for each threshold and chooses the threshold that yields the lowest error, thereby automating threshold selection. The user can then split the sample according to the threshold chosen and analyze data in each subset alone.

2.6 Logistic regression

The Hosmer-Lemeshow tests the model's goodness of fit. Nagelkerke's pseudo R-squared is also reported, and the results and model coefficients are automatically interpreted.

2.7 Tools used

This desktop and web application was coded using the shiny package (Chang et al., 2018) of Microsoft R Open version 3.4.4 in RStudio version 1.1.453, HTML and JavaScript. The complete list of R packages used can be found in the appendix.

The normality simulation and the predictive models were written in Microsoft R Open.

3. Results

3.1 Software description

Our software consists of a graphical user interface for R in which the user can work with data without the need of typing commands. The program can read data stored in Microsoft Excel, CSV (comma separated values), SPSS, STATA, SAS and JSON (JavaScript object notation) files. The complete menu structure is presented in Table 1.

Table 1: The program’s menu structure

Menu	Options
Home	Load and save a dataset
Change variable type	Change the type of a variable to: numeric, categorical or text
Cut variable	- Visualize the relationship between 2 variables - Automatically choose the best cutoff point, cut and save the 2 subsets in different files
Outliers treatment	- Detect, remove numeric outliers - Detect categorical outliers
Missing values treatment	- Visualize missing values patterns - Impute using random forests
Describe and analyze	- Table visualization of a dataset selection - Summary statistics for each variable: type, minimum, maximum, mean, median, category frequency (for categorical variables), number of missing values - Graph the dependent variable: bar plot or histogram with normal curve overlay - Heatmap of the correlation matrix with automatic detection of multicollinearity - Automatic bivariate analysis with explanation of tests assumptions - Automatic variable selection: focused principal component analysis, bivariate analysis with p-value ≤ 0.2, forward and backward stepwise selection, and LASSO regression - Linear and logistic regression with automatic assumption checking and interpretation of results

This menu can be divided into two large parts: a data pre-processing or preparation part and the data analysis part.

In the data preparation stage, users can change the variable type to specify if it should be considered as numeric, categorical or text information. They have the option to cut a numerical variable using an automatically chosen threshold as described in the methods section. The user can replace numeric outliers with missing values; the software also suggests some actions to handle categorical outliers. Visualizing missing patterns of data can help identify observations or variables that have a lot of missing values. We used the data from a student health behavior study (Abdo, Zeenny, & Salameh, 2016), which will be discussed in a future section, to see the pattern of missing data using the following variables: age (the student's age), total average score (the student's average grade score in school) and do you have a boy/girlfriend (If the student is in an intimate relationship or not).   Figure 1 shows missing values in black, one variable, the total average score, has only 1.2% missing values but these are restricted to the first and last few observations in the dataset, which lead us to investigate the reason. We can either delete these observations or choose to impute missing values using a random forest model.

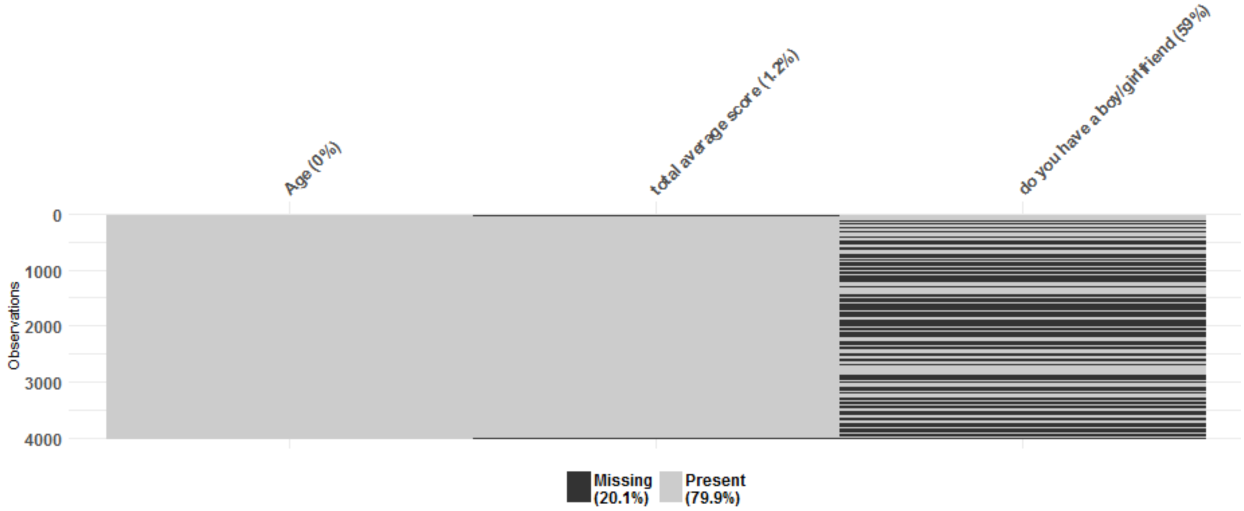


Figure 1: Visualizing missing pattern
Missing values are in black, non-missing are in grey

Next we discuss the analytical part of the software. Summary statistics show the minimum, maximum, mean, median and number of missing values for each numerical variable; and category frequency and number of missing values for each categorical variable. The correlation matrix is represented using a heatmap and automatic detection of collinearity.

A bivariate analysis examines the relationship between each independent variable and the dependent variable using statistical tests, the user also has the option to view the conditions and assumptions that led the software to choose a specific test. Variable selection algorithms implemented include: focused principle component analysis, bivariate analysis with p-value ≤ 0.2, forward and backward stepwise selection and LASSO regression.

When the user clicks on multivariable linear model, a linear regression is run (after verifying its assumptions) for continuous dependent variables and a logistic regression is run for binary dependent variables, the software tries logarithmic transformations as discussed in the methods section and helps the user also by interpreting the model's coefficients.

3.2 Accuracy results of the normality prediction algorithm

Using 70% of the simulation dataset (4200 samples), we trained a logistic regression model Nagelkerke's pseudo R2 was 0.201 andhe threshold for predicting normality was set to 0.15 using the area under the ROC curve. We also trained a random forests model, the threshold for predicting normality was set to 0.15 using the area under the ROC curve. On the remaining 30% of the dataset (1800 samples), we compared the performance of 4 normality predictors (Table 2). As a benchmark, we introduced a base model – a model that predicts non-normality for all samples. It's accuracy of 89.83% is the percentage of non-normal labelled samples in the test set. It has a Matthews correlation coefficient of 0 because it has a true positive rate of 0. Of all four predictors, random forests has the highest Matthews correlation coefficient of 0.501 and AUC under the ROC curve of 89.87%, as a result it was implemented in the software to predict normality of distributions with sample sizes less than 30.

Table 2: Comparing the test set performance of different normality predictors

Prediction method	Accuracy	Matthews correlation coefficient	AUC ROC[a]
Base model	89.83%	0	
Shapiro-Wilk test	35.72%	0.157	
Logistic regression	66.94%	0.280	78.39%
Random forest	87.5%	0.501	89.87%

a – Area under the receiver operating characteristic curve

4. Application – Predictors of smoking among Lebanese school adolescents

As means to test our newly developed software, we chose to work on a classical subject from the field of epidemiology, to find factors associated with smoking among adolescent school students.

4.1 Introduction

Aiming to limit risky health behaviors among these teenagers, it is important to identify high risk groups and implement effective health programs that target these individuals in early life when intervention is more beneficial. We will focus on cigarette and waterpipe smoking among teenage students. Tobacco smoking in the form of cigarettes and waterpipe is common among Lebanese students (Bejjani, El Bcheraoui, & Adib, 2012; El-Roueiheb et al., 2008). It has short-term respiratory and non-respiratory effects, causes addiction and leads to other form of drug use. One important long-term consequence is that most teenagers who smoke will continue to do so as adults (Elders, Perry, Eriksen, & Giovino, 1994).

Research has already shown many factors associated with higher risk of smoking among adolescents, such as: Age (Zhang, Wang, Zhao, & Vartiainen, 2000), male gender (El-Roueiheb et al., 2008), parents with lower socio-economic status (Hanson & Chen, 2007), family conflicts and psychosocial variables, including social and interpersonal factors, attitudinal and belief factors (Flay, Hu, & Richardson, 1998). Having a smoking family member (Fleming, Kim, Harachi, & Catalano, 2002) or friend (Zhang et al., 2000) is also associated with more tobacco use. A longitudinal study collected data from 15,705 adolescents, from 6 European countries, on their own smoking status and that of their parents and friends in general, found that the influence of the best friend and friends were comparable to that of parental smoking (de Vries, Engels, Kremers, Wetzels, & Mudde, 2003). Similarly, the mother's level of education, parent and best friend waterpipe smoking were positively associated with adolescent waterpipe smoking(Schröder, Chaaya, Saab, & Mahfoud, 2016). Using other drugs (Flay et al., 1998) such as alcohol, having lower grades (Fleming et al., 2002) and spending more time watching TV are associated with more smoking (Gidwani, Sobol, DeJong, Perrin, & Gortmaker, 2002). On the other hand, a review of 19 studies shows that not all studies found a positive relationship between body mass index and smoking among adolescents (Potter, Pederson, Chan, Aubut, & Koval, 2004), and measures of child attachment to parent and parent involvement with the child's school have a protective effect (Fleming et al., 2002).

Most adults who use and have trouble quitting tobacco started at a young age (Walker & Loprinzi, 2014). Whether out of curiosity or secondary to psychological issues, smoking can become a gateway to other substance abuse and major health problems. Therefore, efforts should be made to improve the quality of life of adolescent smokers in order to prevent short- and long-term consequences.

4.2 Methods

Our data comes from a cross-sectional study, carried out in 2014 on 4000 private school students, which asked about their health habits (smoking, alcoholism and eating habits) and various other things like standard of living and relationship with their parents. Details of the study design and data collection can be found in Abdo et al. (Abdo et al., 2016).

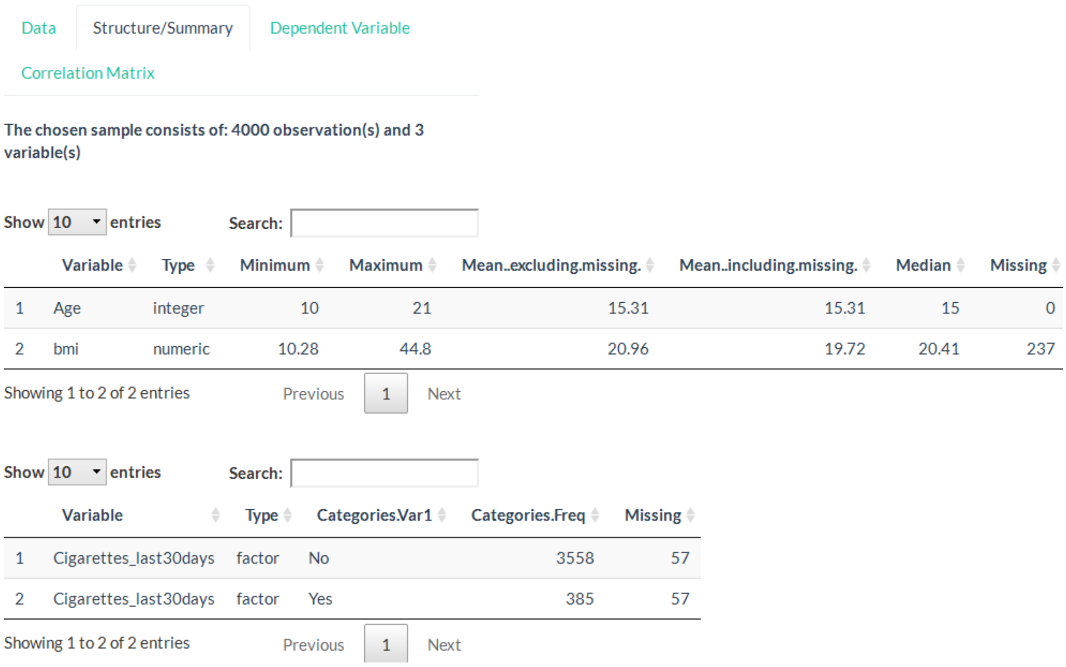
For the descriptive part, we used means with standard deviations to summarize continuous variables, and percentages for categorical variables.

To compare groups, student-t and chi-squared tests were used after checking appropriate conditions and assumptions. A p-value of less than 5% was considered statistically significant.

Multivariable logistic regressions were performed to assess factors associated with cigarette and waterpipe smoking by controlling for confounding variables. Variables included in the models were risk factors found in the literature review and others we found logically plausible to include.

4.3 Results

The sample consists of 4000 school students, mostly from Mount Lebanon (74.5%) and Beirut (24.5%), with roughly 51% females and 49% males. The average student age was 15.31 ±2.01 years, the smallest student was 10 years old and the oldest was 21 years old (Figure 2).



The chosen sample consists of: 4000 observation(s) and 3 variable(s)

	Variable	Type	Minimum	Maximum	Mean..excluding.missing.	Mean..including.missing.	Median	Missing
1	Age	integer	10	21	15.31	15.31	15	0
2	bmi	numeric	10.28	44.8	20.96	19.72	20.41	237

Showing 1 to 2 of 2 entries

	Variable	Type	Categories.Var1	Categories.Freq	Missing
1	Cigarettes_last30days	factor	No	3558	57
2	Cigarettes_last30days	factor	Yes	385	57

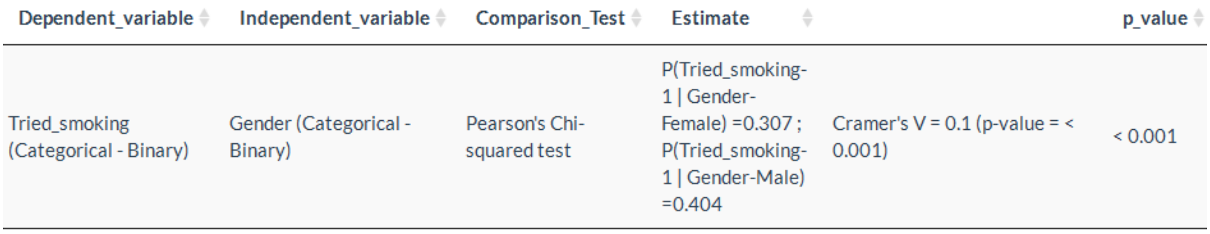
Showing 1 to 2 of 2 entries

Figure 2: Summary statistics

Shows how the software handles numerical and categorical variables independently and outputs accordingly

In our sample, 23.2% (n = 927) of students reported trying cigarette, pipe or cigar smoking. 30.7% (n = 1226) reported trying waterpipe at least once in their lifetime. In the past month, 4.5% of students smoked at least one cigarette daily, and 10.5% smoked at least one waterpipe per week. 1.8% reported smoking more than 10 cigarettes per day the last month and 1.9% smoked more than 4 waterpipes per week.

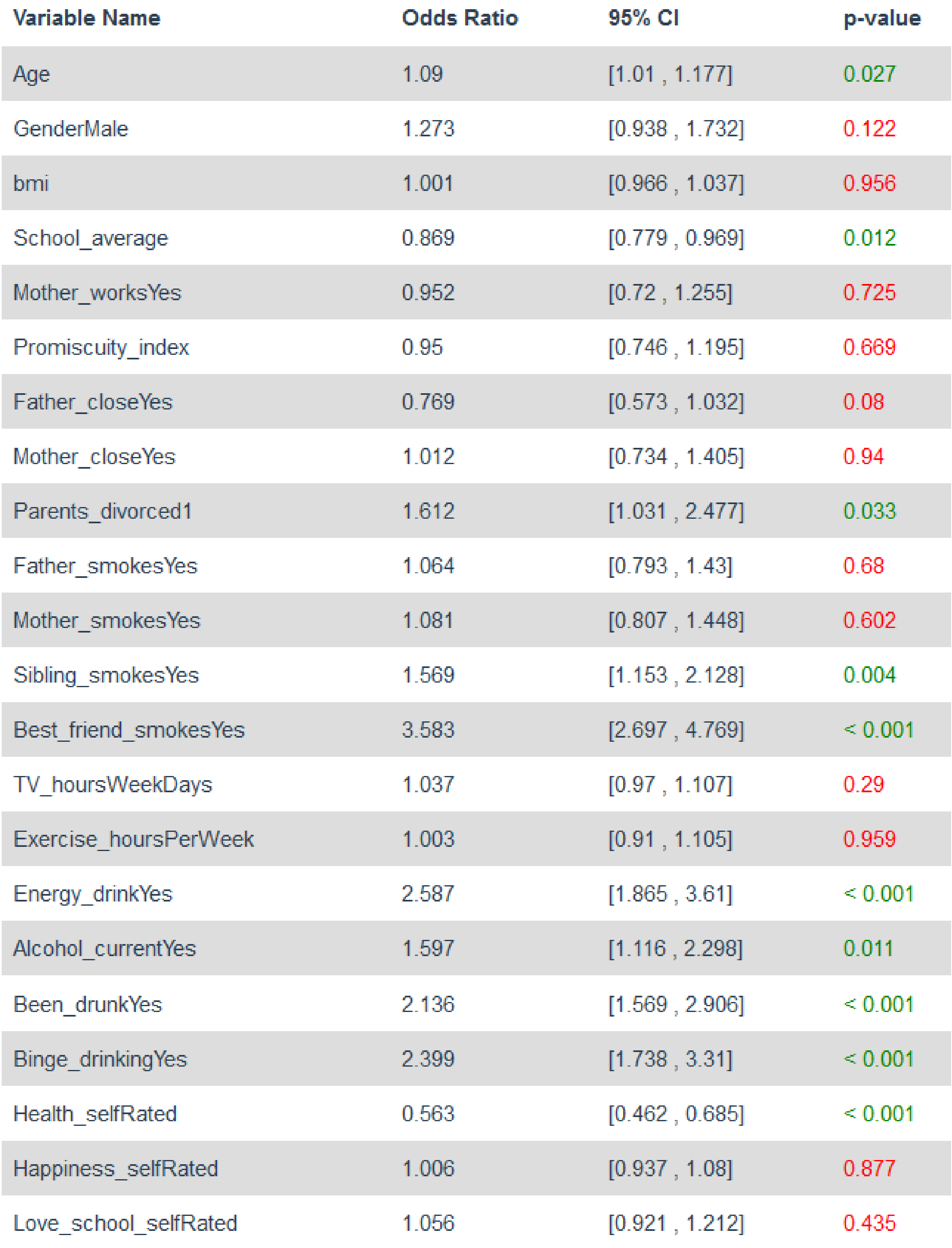
Boys' tendency of trying to smoke (40.4%) was significantly higher than girls' (30.7%), ($p < 0.001$) (Figure 3). They also tend to be heavier smokers, as they smoke on average 2 more cigarettes per week (95% CI [1.3-2.7]; $p < 0.001$).



Dependent_variable	Independent_variable	Comparison_Test	Estimate		p_value
Tried_smoking (Categorical - Binary)	Gender (Categorical - Binary)	Pearson's Chi-squared test	P(Tried_smoking-1 | Gender-Female) =0.307 ; P(Tried_smoking-1 | Gender-Male) =0.404	Cramer's V = 0.1 (p-value = < 0.001)	< 0.001

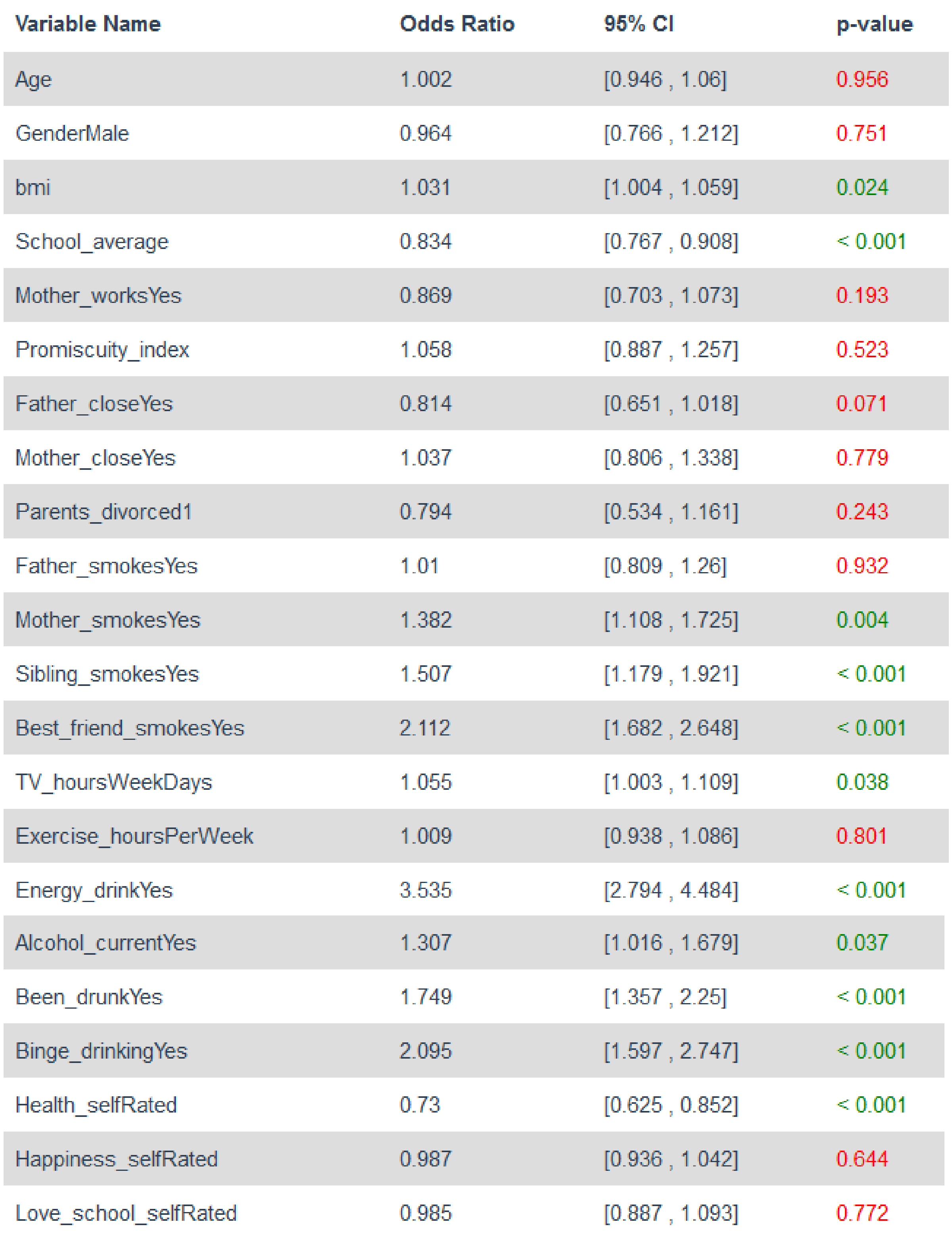
Figure 3: Gender differences in trying to smoke

Two logistic regression models were run; cigarette and waterpipe smoking in the past month were the dependent variables. Nagelkerke's pseudo R2 of the two models were 0.4 and 0.344, respectively. The adjusted odds ratios (aOR) for substance correlate with their 95% confidence intervals and p-values produced by our software are shown in figure 4 and 5.



Variable Name	Odds Ratio	95% CI	p-value
Age	1.09	[1.01 , 1.177]	0.027
GenderMale	1.273	[0.938 , 1.732]	0.122
bmi	1.001	[0.966 , 1.037]	0.956
School_average	0.869	[0.779 , 0.969]	0.012
Mother_worksYes	0.952	[0.72 , 1.255]	0.725
Promiscuity_index	0.95	[0.746 , 1.195]	0.669
Father_closeYes	0.769	[0.573 , 1.032]	0.08
Mother_closeYes	1.012	[0.734 , 1.405]	0.94
Parents_divorced1	1.612	[1.031 , 2.477]	0.033
Father_smokesYes	1.064	[0.793 , 1.43]	0.68
Mother_smokesYes	1.081	[0.807 , 1.448]	0.602
Sibling_smokesYes	1.569	[1.153 , 2.128]	0.004
Best_friend_smokesYes	3.583	[2.697 , 4.769]	< 0.001
TV_hoursWeekDays	1.037	[0.97 , 1.107]	0.29
Exercise_hoursPerWeek	1.003	[0.91 , 1.105]	0.959
Energy_drinkYes	2.587	[1.865 , 3.61]	< 0.001
Alcohol_currentYes	1.597	[1.116 , 2.298]	0.011
Been_drunkYes	2.136	[1.569 , 2.906]	< 0.001
Binge_drinkingYes	2.399	[1.738 , 3.31]	< 0.001
Health_selfRated	0.563	[0.462 , 0.685]	< 0.001
Happiness_selfRated	1.006	[0.937 , 1.08]	0.877
Love_school_selfRated	1.056	[0.921 , 1.212]	0.435

Figure 4: Logistic regression results of cigarette smoking



Variable Name	Odds Ratio	95% CI	p-value
Age	1.002	[0.946 , 1.06]	0.956
GenderMale	0.964	[0.766 , 1.212]	0.751
bmi	1.031	[1.004 , 1.059]	0.024
School_average	0.834	[0.767 , 0.908]	< 0.001
Mother_worksYes	0.869	[0.703 , 1.073]	0.193
Promiscuity_index	1.058	[0.887 , 1.257]	0.523
Father_closeYes	0.814	[0.651 , 1.018]	0.071
Mother_closeYes	1.037	[0.806 , 1.338]	0.779
Parents_divorced1	0.794	[0.534 , 1.161]	0.243
Father_smokesYes	1.01	[0.809 , 1.26]	0.932
Mother_smokesYes	1.382	[1.108 , 1.725]	0.004
Sibling_smokesYes	1.507	[1.179 , 1.921]	< 0.001
Best_friend_smokesYes	2.112	[1.682 , 2.648]	< 0.001
TV_hoursWeekDays	1.055	[1.003 , 1.109]	0.038
Exercise_hoursPerWeek	1.009	[0.938 , 1.086]	0.801
Energy_drinkYes	3.535	[2.794 , 4.484]	< 0.001
Alcohol_currentYes	1.307	[1.016 , 1.679]	0.037
Been_drunkYes	1.749	[1.357 , 2.25]	< 0.001
Binge_drinkingYes	2.095	[1.597 , 2.747]	< 0.001
Health_selfRated	0.73	[0.625 , 0.852]	< 0.001
Happiness_selfRated	0.987	[0.936 , 1.042]	0.644
Love_school_selfRated	0.985	[0.887 , 1.093]	0.772

Figure 5: Logistic regression results of waterpipe smoking

Compared to young students, older ones reported higher rates of cigarette smoking in the last month, with every year increasing the risk of 9%. Students with higher body mass index reported more waterpipe smoking. School grades had a protective effect on the two types of smoking (aOR of 0.87 and 0.83 for cigarette and waterpipe use, respectively).

Divorce between parents is associated with 61% more risk of cigarette smoking in adolescents. Communication between family members did not affect significantly the smoking status. However, having a sibling or a close friend who smoked significantly increased the risk of cigarette (aOR of 1.57 and 3.58) and waterpipe use (aOR of 1.5 and 2.1). The socioeconomic status represented here by the promiscuity index (number of individuals in the house divided by the number of rooms) did not affect smoking status.

Consuming energy drinks is associated with an increased risk of cigarette and waterpipe smoking with an aOR of 2.59 and 3.53 respectively. Alcohol use, binge drinking and been drunk at least in one occasion were all significant predictors of cigarette and waterpipe use. Students who smoke had an inferior evaluation of their own health, aOR for cigarette and waterpipe were 0.56 and 0.73, respectively. Watching TV or playing video games in weekdays is associated with 5.5% more waterpipe smoking.

4.5 Discussion

As an application to our work, we analyzed tobacco smoking behavior of Lebanese school adolescents. We found that the proportion of students who have ever tried cigarette smoking is close to that found by Zahlan et al. (Zahlan, Ghandour, Yassin, Afifi, & Martins, 2014), as for waterpipe our results were lower. Contrary to other studies (El-Roueiheb et al., 2008; Saade, Warren, Jones, Asma, & Mokdad, 2008; Zahlan et al., 2014), male gender was not associated with smoking in the multivariable analysis, an explanation could be that it is confounded with other substance use. More students currently smoke waterpipe than cigarettes, could be because they believe that waterpipe is safer and socially more acceptable than cigarettes (Smith et al., 2011). We showed that the likelihood of becoming a current cigarette smoker, as opposed to waterpipe, increases with age (Zeenny Rony, 2015; Zhang et al., 2000), family conflicts (Flay et al., 1998) such as parents' divorce. One explanation could be that cigarette smoking can be considered as an individual and a stress relief activity, contrary to waterpipe smoking which is a social and recreational one. Contrary to what we expected, attachment to parents and their smoking status (Fleming et al., 2002) did not influence smoking in children, except that mother's smoking status is associated to child's waterpipe smoking. However, sibling and friends were strong predictors of both cigarette and waterpipe smoking. A high school grades average (Fleming et al., 2002) and self-rated health score (Mazur & Woynarowska, 2004) have a protective effect on smoking. Our results show that smoking and alcohol use are highly linked, which matches several other studies (Alikaşifoğlu et al., 2004).

In general, our results were close to those previously reported in Lebanon and other countries. Few differences such as the relationship of parent-child and the parents smoking status were not good predictors of smoking in adolescents, apparently in this age range, kids are more influenced by their peers.

The study limitations are fully discussed in Abdo et al. (Abdo et al., 2016).

5. Discussion

We found that data analysis can be rendered faster and more objective with automation by using a combination of programming by specific instructions coupled with machine learning techniques.

Specific instructions were used, for instance, when choosing a statistical test. This is essentially following branches in a decision tree where we check conditions or assumptions and decide, based on the answer, which path we end up taking. Some of these conditions are straightforward, for instance checking if the variable is numeric or categorical, the sample size, etc. Others are more difficult to assess, such as the normality of a distribution or equal variance between 2 groups. We can use statistical tests to decide on normality and homoscedasticity but the problem with these tests is that with small sample sizes ($n < 30$) they do not have enough power to detect an effect (Barker & Shaw, 2015; Mohd Razali & Yap, 2011) and the opposite happens as the sample size gets large where they tend to detect the smallest effect size that would not affect the results much (Falissard, 2011). This is why we preferred using the rule of thumb definition of homoscedasticity between groups – where a standard deviation of one group is not larger than 1.5 times the standard deviation of the other (Falissard, 2011). Although this method cannot be regarded more than an approximation, it is practically accurate enough, independent of sample size and easily programmable. As for normality in hypothesis testing, it should be assessed only when we have a sample size smaller than 30, otherwise the central limit theorem ensures the normality of the sampling distribution (Falissard, 2011). In order to avoid using a low powered normality test with a small sample size, we can use a graphical method to assess normality, where the epidemiologist looks at the histogram or QQ plot to decide visually if the distribution is normal. This has the advantage of eliminating the dependency on low powered normality tests but introduces subjectivity into the analysis. To deal with this problem, we used a model to predict normality, which did better than the Shapiro-Wilk test, for sample sizes between 7 and 50, both in terms of raw accuracy (87.5% vs. 35.72%), and when considering true and false, positives and negatives (Matthews correlation coefficient was 0.50 vs. 0.16). It also has an advantage of non-subjectivity over the graphical method.

Imputing missing data by replacing numerical values with the mean and categorical values with the column mode does not make use of the underlying correlation structure of the data and thus performs poorly (P, J, & M, 2015). Other more accurate model-based methods can be used, but most of them focus only of numeric variables (Finch, 2010),or treat categorical and numeric data types separately thus ignoring possible relationships between these different variables. We implemented a random forest method, available from the package missForest, which is a non-parametric method that can handle categorical and numerical data types simultaneously and has many advantages over other methods as discussed in Stekhoven et al (Stekhoven, 2013).

One of the assumptions of linear regression is the linear relationship between dependent and numeric independent variables. One method of correcting non-linearity is cutting the independent variable in 2 groups, each having a linear relationship with the dependent variable. The problem is then reduced to choosing a cutoff point. Our brute force algorithm described in the methods section provides a mathematically more accurate alternative than the visual method or using the rule of thumb of choosing the median, because it chooses a threshold from all possible data points by finding the minimal sum of squared errors of the 2 regression lines.

5.1 Limitations

Statistical software with graphical user interface, especially with automatic checking of conditions and interpretation of results, lowers the entry bar for non-experts to analyze data. This can be both a good and a dangerous feature. Having a menu with point and click options encourages or permits blind and incorrect use of statistical methods (Nyirongo, Mukaka, & Kalilani-Phiri, 2008).

Three main assumptions should be assessed before running a linear regression, namely the residuals must be uncorrelated, have a normal distribution, and a constant variance (Berry, 1993). In our software, these assumptions were assessed using statistical tests. This has the downside of rejecting the null, in large samples, even for a small change that would not affect regression coefficients, p-values and confidence intervals much, which leads to trying logarithmic transformations of the dependent variable or using a generalized linear model instead of linear regression.

Granted that automation ensures a certain level of objectivity, however, forcing specific methods constricts data analysis in a way that experienced users can no longer get the results of a statistical test or a model unless the software judges appropriate by testing its assumptions in its specific implemented manner. This is both a feature and a drawback that we certainly considered but decided that it would be, in general, more advantageous than limiting.

5.2 Future and big picture

We can ameliorate the software by testing the assumptions of linear regression without resorting to statistical tests and handling bugs and errors caused by special cases such as running logistic regression on a binary dependent variable with one level being severely underrepresented.

Consider the following question: “does cranberry juice reduce the risk of urinary tract infection in immunosuppressed men older than 70? “. Automated data analysis combined with the natural language processing of a search engine can answer this type of specific queries whose answer is not addressed in any research paper, by searching for a dataset on the internet with appropriate variables and cases (and if needed combine datasets from different sources to increase the sample size) and running automatic analysis. We believe that this individualized, guided-by-demand human-computer interaction complements traditional research, specifically because medicine is always about answering specific case and personalized questions and never about averages in a population which is what the medical literature is all about. We believe that automated data analysis is an essential ingredient in the big picture solution that bridges this gap.

5.3 Software availability

The web application is available on

https://automated-data-analysis.shinyapps.io/automation_app/, the source code is available for download from the app itself, and anyone can use it to run the software locally and for free.

The desktop application can be recreated from the same source code using instructions from http://blog.analytixware.com/2014/03/packaging-your-shiny-app-as-windows.html. It is also available on demand for non-programmers, please contact white.softapp@gmail.com.

6. Conclusion

We look at this work as a contribution to automation of data analysis. In general, it should resolve the problem of errors in checking conditions and assumptions in statistics that are found in the medical literature, and give epidemiologists the opportunity not to be lost in statistical details, to see the big picture and focus on the question and consequences of the results. It also takes advantages that the R environment has in terms of cutting edge improvements and offers them to non-programmers through a graphical user interface.

Finally, we showed how easy it was for a machine learning model to outperform normality tests such as Shapiro-Wilk. On the other hand, our assumption checking in linear regression is still based on statistical tests. More research is needed to find new ways to automate certain parts of data analysis, either by improving methods used in our work or by automating others and adding them to the project.

7. Acknowledgement

We would like to thank the R community for the packages that made our work possible and everyone working on open-source projects and knowledge sharing.

References

[1] Abdo, R., Zeenny, R., & Salameh, P. (2016). Health Behaviors Among School-Aged Children: a Cross Sectional Study in Lebanese Private Schools. International Journal of Mental Health and Addiction, 14(6), 1003-1022. doi: 10.1007/s11469-016-9677-z

[2] Alikaşifoğlu, M., Erginöz, E., Ercan, O., Uysal, O., Albayrak-Kaymak, D., & Ilter, O. (2004). Alcohol drinking behaviors among Turkish high school students. The Turkish Journal of Pediatrics, 46(1), 44-53.

[3] Barker, L. E., & Shaw, K. M. (2015). Best (but of t-forgotten) practices: checking assumptions concerning regression residuals. The American Journal of Clinical Nutrition, 102(3), 533-539. doi: 10.3945/ajcn.115.113498

[4] Bejjani, N., El Bcheraoui, C., & Adib, S. M. (2012). The social context of tobacco products use among adolescents in Lebanon (MedSPAD-Lebanon). Journal of Epidemiology and Global Health, 2(1), 15-22. doi: 10.1016/j.jegh.2012.02.001

[5] Berry, W. D. (1993). Understanding regression assumptions: Newbury Park, Calif. : Sage Publications.

[6] Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLOS ONE, 12(6), e0177678. doi: 10.1371/journal.pone.0177678

[7] Bouyer, J. (2009). Épidémiologie: principes et méthodes quantitatives: Lavoisier.

[8] Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. Econometrica, 47(5), 1287-1294. doi: 10.2307/1911963

[9] Chang, W., Cheng, J., Allaire, J. J., Xie, Y., McPherson, J., Rstudio, . . . R, R. C. T. (2018). shiny: Web Application Framework for R (Version 1.1.0). Retrieved from https://CRAN.R-project.org/package=shiny

[10] de Vries, H., Engels, R., Kremers, S., Wetzels, J., & Mudde, A. (2003). Parents' and friends' smoking status as predictors of smoking onset: findings from six European countries. Health Education Research, 18(5), 627-636. doi: 10.1093/her/cyg032

[11] Durbin, J., & Watson, G. S. (1950). Testing for Serial Correlation in Least Squares Regression: I. Biometrika, 37(3/4), 409-428. doi: 10.2307/2332391

[12] Durbin, J., & Watson, G. S. (1951). Testing for Serial Correlation in Least Squares Regression. II. Biometrika, 38(1/2), 159-177. doi: 10.2307/2332325

[13] Dwan, K., Gamble, C., Williamson, P. R., Kirkham, J. J., & Reporting Bias, G. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. PLOS ONE, 8(7), e66844. doi: 10.1371/journal.pone.0066844

[14] El-Roueiheb, Z., Tamim, H., Kanj, M., Jabbour, S., Alayan, I., & Musharrafieh, U. (2008). Cigarette and waterpipe smoking among Lebanese adolescents, a cross-sectional study, 2003-2004. Nicotine & Tobacco Research: Official Journal of the Society for Research on Nicotine and Tobacco, 10(2), 309-314. doi: 10.1080/14622200701825775

[15] Elders, M. J., Perry, C. L., Eriksen, M. P., & Giovino, G. A. (1994). The report of the Surgeon General: preventing tobacco use among young people. American Journal of Public Health, 84(4), 543-547.

[16] Ercan, I., Yazıcı, B., Yang, Y., Özkaya, G., Cangur, S., Ediz, B., & Kan, I. (2007). Misusage of statistics in medical research. European Journal of General Medicine, 4(3), 128-134. doi: 10.29333/ejgm/82507

[17] Falissard, B. (1999). Focused Principal Component Analysis: Looking at a Correlation Matrix with a Particular Interest in a Given Variable. Journal of Computational and Graphical Statistics, 8(4), 906-912. doi: 10.1080/10618600.1999.10474855

[18] Falissard, B. (2011). Analysis of Questionnaire Data with R (1 edition ed.). Boca Raton, FL: Chapman and Hall/CRC.

[19] Felson, D. T., Cupples, L. A., & Meenan, R. F. (1984). Misuse of statistical methods in Arthritis and Rheumatism. 1982 versus 1967-68. Arthritis and Rheumatism, 27(9), 1018-1022.

[20] Finch, W. H. (2010). Imputation Methods for Missing Categorical Questionnaire. Journal of Data Science, 8(3), 361-378.

[21] Data: A Comparison of Approaches. Journal of Data Science, 8(3), 361-378.

[22] Flay, B. R., Hu, F. B., & Richardson, J. (1998). Psychosocial predictors of different stages of cigarette smoking among high school students. Preventive Medicine, 27(5 Pt 3), A9-18.

[23] Fleming, C. B., Kim, H., Harachi, T. W., & Catalano, R. F. (2002). Family processes for children in early elementary school as predictors of smoking initiation. The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine, 30(3), 184-189.

[24] Gidwani, P. P., Sobol, A., DeJong, W., Perrin, J. M., & Gortmaker, S. L. (2002). Television viewing and initiation of smoking among youth. Pediatrics, 110(3), 505-508.

[25] haine, D. (2017, 2017/11/15/). Popularity of statistical softwares in epidemiology. Retrieved from https://denishaine.ca/blog/popepi-rmd/files/571/popepi-rmd.html

[26] Hanif, A., & Ajmal, T. (2011). Statistical Errors in Medical Journals (A Critical Appraisal). Annals of King Edward Medical University, 17(2), 178-178. doi: 10.21649/akemu.v17i2.295

[27] Hanson, M. D., & Chen, E. (2007). Socioeconomic status and health behaviors in adolescence: a

review of the literature. Journal of Behavioral Medicine, 30(3), 263-285. doi: 10.1007/s10865-007-9098-3

[28] Ho, T. K. (1995, 1995/08//). Random decision forests. Paper presented at the Proceedings of 3rd International Conference on Document Analysis and Recognition.

[29] Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. PLOS Medicine, 2(8), e124. doi: 10.1371/journal.pmed.0020124

[30] Khan, A. M. (2013). R-software: A Newer Tool in Epidemiological Data Analysis. Indian Journal of Community Medicine : Official Publication of Indian Association of Preventive & Social Medicine, 38(1), 56-58. doi: 10.4103/0970-0218.106630

[31] Mazur, J., & Woynarowska, B. (2004). [Risk behaviors syndrome and subjective health and life satisfaction in youth aged 15 years]. Medycyna Wieku Rozwojowego, 8(3 Pt 1), 567-583.

[32] Mohd Razali, N., & Yap, B. (2011). Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. J. Stat. Model. Analytics, 2.

[33] Nyirongo, V. B., Mukaka, M. M., & Kalilani-Phiri, L. V. (2008). Statistical Pitfalls in Medical Research. Malawi Medical Journal, 20(1), 15-18.

[34] Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., & Myer-Tyson, E. B. (2017). MatLab vs. Python vs. R. Journal of Data Science, 15(3), 355-372.

[35] P, S., J, E., & M, G. (2015). A Comparison of Six Methods for Missing Data Imputation. Journal of Biometrics & Biostatistics, 6(1). doi: 10.4172/2155-6180.1000224

[36] Potter, B. K., Pederson, L. L., Chan, S. S. H., Aubut, J.-A. L., & Koval, J. J. (2004). Does a relationship exist between body weight, concerns about weight, and smoking among adolescents? An integration of the literature with an emphasis on gender. Nicotine & Tobacco Research: Official Journal of the Society for Research on Nicotine and Tobacco, 6(3), 397-425. doi: 10.1080/14622200410001696529

[37] Saade, G., Warren, C. W., Jones, N. R., Asma, S., & Mokdad, A. (2008). Linking Global Youth Tobacco Survey (GYTS) data to the WHO Framework Convention on Tobacco Control (FCTC): the case for Lebanon. Preventive Medicine, 47 Suppl 1, S15-19. doi: 10.1016/j.ypmed.2008.06.003

[38] Schröder, C., Chaaya, M., Saab, D., & Mahfoud, Z. (2016). The determinants of intention to smoke waterpipe among adolescents in Lebanon: a national household survey. Journal of Public Health (Oxford, England), 38(1), 84-91. doi: 10.1093/pubmed/fdv004

[39] Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52(3/4), 591-611. doi: 10.2307/2333709

[40] Smith, J. R., Novotny, T. E., Edland, S. D., Hofstetter, C. R., Lindsay, S. P., & Al-Delaimy, W. K. (2011). Determinants of Hookah Use among High School Students. Nicotine & Tobacco Research, 13(7), 565-572. doi: 10.1093/ntr/ntr041

[41] Stekhoven, D. J. (2013). missForest: Nonparametric Missing Value Imputation using Random Forest (Version 1.4). Retrieved from https://CRAN.R-project.org/package=missForest

[42] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267-288.

[43] Walker, J. F., & Loprinzi, P. D. (2014). Longitudinal examination of predictors of smoking cessation in a national sample of U.S. adolescent and young adult smokers. Nicotine & Tobacco Research: Official Journal of the Society for Research on Nicotine and Tobacco, 16(6), 820-827. doi: 10.1093/ntr/ntu005

[44] Yang, J. (2012, 2012/06//). [Interpreting Coefficients in Regression with Log-Transformed Variables].

[45] Zahlan, L., Ghandour, L., Yassin, N., Afifi, R., & Martins, S. S. (2014). Double trouble: Exploring the association between waterpipe tobacco smoking and the nonmedical use of psychoactive prescription drugs among adolescents. Drug and Alcohol Dependence, 145, 217-223. doi: 10.1016/j.drugalcdep.2014.10.020

[46] Zeenny Rony, S. P. (2015). Is Waterpipe Smoking a Gateway to Cigarette Smoking among Youth? Journal of Addictive Behaviors, Therapy and Rehabilitation, 04(02). doi: 10.4172/2324-9005.1000136

[47] Zhang, L., Wang, W., Zhao, Q., & Vartiainen, E. (2000). Psychosocial predictors of smoking among secondary school students in Henan, China. Health Education Research, 15(4), 415-422.

Appendix: packages used in the software

Package name	Description	Author(s)	Web reference
car	An R companion to applied regression	John Fox, Sanford Weisberg	http://socserv.socsci.mcmaster.ca/jfox/Books/Companion
caret	Classification and regression training	Max Kuhn, contribution from many other authors	https://CRAN.R-project.org/package=caret
caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	Jarek Tuszynski	https://CRAN.R-project.org/package=caTools
corrplot	Visualization of a correlation matrix	Taiyun Wei, Viliam Simko	https://github.com/taiyun/corrplot
d3heatmap	Interactive heat maps using htmlwidgets and D3.js	Joe Cheng, Tal Galili	https://CRAN.R-project.org/package=d3heatmap
DT	A wrapper of the JavaScript library ‘DataTables’	Yihui Xie	https://CRAN.R-project.org/package=DT
dplyr	A grammar of data manipulation	Hadley Wickham, Romain Francois, Lionel Henry, Kirill Muller	https://CRAN.R-project.org/package=dplyr
foreign	Read data stored by ‘Minitab’, ‘S’, ‘SAS’, ‘SPSS’, ‘Stata’, …	R core team	https://CRAN.R-project.org/package=foreign
glmnet	Regularization paths for generalized linear models via coordinate descent	Jerome Friedman, Trevor Hastie, Robert Tibshirani	http://www.jstatsoft.org/v33/i01/
gridExtra	Miscellaneous functions for ‘Grid’ Graphics	Baptiste Auguie	https://CRAN.R-project.org/package=gridExtra
e1071	Miscellaneous functions of the department of statistics, probability	David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel,	https://CRAN.R-project.org/package=e1071

	theory group, TU Wein	Friedrich Leisch	
jsonlite	A practical and consistent mapping between JSON data and R objects	Jeroen Ooms	https://arxiv.org/abs/1403.2805
lmtest	Diagnostic checking in regression relationships	Achim Zeileis, Torsten Hothorn	https://CRAN.R-project.org/doc/Rnews/
moments	Moments, cumulants, skewness, kurtosis and related tests	Lukasz Komsta, Frederick Novomestky	https://CRAN>R-project.org/package=moments
missForest	Nonparametric missing value imputation using random forest	Daniel Stekhoven	https://CRAN.R-project.org/package=missForest
naniar	Data structures, summaries and visualizations for missing data	Nicholas Tierney, Di Cook, Miles McBain, Colin Fay	https://CRAN.R-project.org/package=naniar
plotly	Create interactive web graphics via 'plotly.js'	Carson Sievert, et al.	https://CRAN.R-project.org/package=plotly
pROC	An open-source package for R and S+ to analyze and compare ROC curves	Xavier Robin, et al.	https://CRAN.R-project.org/package=pROC
psy	Various procedures used in psychometrics	Bruno Falissard	https://CRAN.R-project.org/package=psy
randomForest	Classification and regression by randomForest	Andy Liaw, Matthew Wiener	https://CRAN.R-project.org/doc/Rnews
rattle	Data mining with rattle and R: the art of excavating data for knowledge discovery	Graham Williams	https://www.amazon.com/gp/product/1441998896/ref=as_li_qf_sp_asin_tl?ie=UTF8&tag=togaware-20&linkCode=as2&camp=217145&creative=399373&creativeASIN=1441998896

rcompanion	Functions to support extension education program evaluation	Salvatore Mangiafico	https://CRAN.R-project.org/package=rcompanion
readxl	Read Excel files	Hadley Wickham, Jennifer Bryan	https://CRAN.R-project.org/package=readxl
reshape2	Reshaping data with the ‘reshape’ package	Hadley Wickham	http://www.jstatsoft.org/v21/i12/
ResourceSelection	Resource selection (probability) functions for use-availability	Subhash Lele, Jonah Keim, Peter Solymos	https://CRAN.R-project.org/package=ResourceSelection
rpart	Recursive partitioning and regression trees	Terry Therneau, Beth Atkinson	https://CRAN.R-project.org/package=rpart
RVAideMemoire	Testing and plotting procedures for biostatistics	Maxime Hervé	https://CRAN.R-project.org/package=RVAideMemoire
sas7bdat	SAS database reader	Matt Shotwell	https://CRAN.R-project.org/package=sas7bdat
scales	Scale functions for visualizations	Hadley Wickham	https://CRAN.R-project.org/package=scales
Shiny	Web application framework for R	Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, Jonathan McPherson	https://CRAN.R-project.org/package=shiny
shinythemes	Themes for ‘shiny’	Winston Chang	https://CRAN.R-project.org/package=shinythemes
shinyWidgets	Custom inputs widgets for ‘shiny’	Victor Perrier, Fanny Meyer	https://CRAN.R-project.org/package=shinyWidgets
stringr	Simple, consistent wrappers for common string operations	Hadley Wickham	https://CRAN.R-project.org/package=stringr