MASTER OF STATISTICS AND DATA SCIENCE

2021-22

PROJECT LEARNING FROM DATA

SAP (version 2)

Group 1

Author	*Student ID*
Bhanupriya Dixit	2157909
Adina Asim	2159804
Aaryan Kaushik	2159244

Version	Date	Revision
1.0	24 December 2021	Original
2.0	02 January 2022	Following content was revised: • Sample size calculation is removed as not required. • Software to be used updated (it was not included in the previous version) • Brief detail about model assumptions has been added. • Analysis plan table has been revised and more relevant information regarding variables has been added.

Table of Contents

1. Introduction:

The purpose of this study is to see how smoking cigarettes affects lung function in children aged 3 to 19. The FEV (Forced Expiratory Volume), which is the volume of air an individual can exhale in the first second of a strong breath, is used to assess pulmonary function (measured in liters). The higher the FEV, the greater the lung's pulmonary function.

1.1 Aims and objectives:

The main objective of this study is to determine the relationship between the level of pulmonary function and individual characteristics.

- The primary research question is to assess the effect of smoking cigarettes on the FEV. Smoking information is available as a binary indicator (smoker / non-smoker).

- The secondary research question is to assess the effect of parental smoking on the FEV of the children (i.e., the effect of passive smoking). Parental smoking is available as a binary indicator (none of the parent's smoke / at least one of the parent's smokes).

2. Study Methods:

In this section design of the proposed study and sample size calculations are presented.

Study Design:

To evaluate the impact of smoking on the brain, long-term observational research was established for Children's lung function. Families in the East Boston area with young children (USA) were requested to take part in this research. The data that must be analyzed in this case is a cross-sectional subset of data comprising the findings of each country's most recent survey family. Only one child's data is accessible per family (654 children). We have sufficient subjects for analysis and to see the effect on FEV of smoking and parental at a one -tailed significance level of < 0.05[i] .

3. Types of Analysis:

In this section, types of data analysis are reported, which will be further carried out in the detailed analysis.

3.1 Descriptive Data Analysis:

Descriptive analysis is a sort of data analysis that helps to explain, show, or summarize data points in a constructive way so that patterns can develop that satisfies all the data's conditions. It is one of the most crucial steps in the statistical data analysis process.

3.2 Exploratory Data Analysis:

Exploratory data analysis will be done for initially to provide an overview of the data set. Boxplot and scatter plot were used to study the potential relationship between two variables, and to evaluate the shape of the data distribution, and to discover possible outliers in the data.

3.3 Inferential Testing:

Inferential statistics aid in the development of hypotheses about a condition or event. It varies from descriptive statistics in that it allows extrapolations to be used to draw conclusions. Inferential statistics offer a quantitative tool for determining whether the null hypothesis H_0 should be accepted or rejected. Because H0 can only be true or false, an inferential test can only have two valid outcomes: accurate rejection of H0 when it is untrue and proper retention of H_0 when it is true[ii].

4. Model Selection:

We will use a forward selection strategy with cross-validation (CV) as a selection criterion to choose which variables are essential to answer the research question. With this, we can build a prediction model [iii].

5. Variable of Interest:

There are 654 observations on children aged 3 to 19 in this data set. The variable of interest is forced Expiratory Volume (FEV1), which is the amount of air you can force from your lungs in one second. Age, height, and BMI are continuous predictors. Gender, Smoke, Parent smoke, T1D, and lung disease are categorical predictors. We will convert the colorblind into categorical form (No=0, Yes=1).

6. Statistical Principles:

Statistical tests will be one-sided at a significance level of 0.05. Also, the results will be reported as point estimates with 95% confidence intervals (CIs).

Furthermore, if in this model we are comparing more than 2 groups, and then this can lead to multiplicity issue. Multiplicity refers to the inflation of the type I error rate. The type I error is nothing but the rate of rejecting the null hypothesis when it is true [iv].We'll apply the Tukey-Kramer approach to avoid Type I errors because it reduces them[v].

7. Summarize and Describe Data:

Quantitative variables are to be summarized by using descriptive statistics, i.e., Number of subjects (N), number of subjects with non-missing values mean and standard deviation (SD), Median, 25th Percentile - 75th Percentile (Q1-Q3), Minimum and maximum, Nominal/ordinal variables will be summarized by counts and percentages.

8. Descriptive statistics:

All Analysis will be presented using descriptive statistics and the data is expected to be normally distributed by the mean and standard deviation (SD). The subsections below will describe analysis in addition to descriptive statistics.

8.1 Primary outcome:

To examine the primary research question, at first a simple linear regression model[vi] will be used to find unadjusted results of Smoke on FEV, where smoke is the independent variable and considered as a key variable in this analysis. After analyzing this key variable, a

multiple linear regression[vii] will be conducted to assess whether or not independent variables predict the dependent variable (criterion). In order to assess this effect in the light of variables we will first do model selection to find out the significant regressors for our model. The model assumption which we will adopt for main effects is forward selection strategy with CV and for interaction effects a stepwise forward selection will be adapted.

8.2 Secondary outcome:

In order to assess the effect of parental smoking on the FEV of the children, the same approach as described above will be adapted. Primarily the effect of parent smoke (independent variable) on FEV of child smoke (dependent variable) will be assessed by using linear regression model and after that, main effects of variables selected by forward selection strategy with CV will be taken into consideration. Later on, the selected interaction effects will also be added to the model to assess the outcome variable from every aspect.

9. Modeling:

In this section how data analysis will be done is presented.

9.1 Main analysis

In order to analyze the data we will use the selected regressors and construct a linear regression model to predict the effect of smoke on FEV. Upon initial findings we will analyze the model and assess all model assumptions. If any of the model assumptions i.e. linearity, normality, homoscedasticity, leverages or multicollinearity will violate we will try to refit the model by adding other regressors in our model. After ascertaining the best model we will find out the summary of that model and interpret the results extracted from the analysis.

9.2 Statistical Hypothesis:

This study is designed to test the null hypothesis that there is no effect of smoking cigarettes on the FEV of children. The alternative hypothesis is that there is a significant effect of smoking on FEV of children.

9.3 Model Assumptions:

The assumptions of multiple regression linearity[viii], homoscedasticity[ix], leverages[x] and absence of multicollinearity will be assessed. Linearity and homoscedasticity will be assessed by examination of a scatter plot and QQ-Plots and if systematic deviations will be seen in plots the outliers will be reported. The absence of multicollinearity assumes that predictor variables are not too related and will be assessed using Variance Inflation Factor (VIF). When computing VIF of the desired independent variable the value needs to be smaller than 5.

10. Analysis Plan Table:

Objective	**Outcome/ variable**	**Variable Type**	**Predictors/comparisons**	**Analysis**
Determine the significant regressors (main effects and	FEV (Forced Expiratory Volume) (liters)	Continuous	Age, Gender, Smoke, Height, Lung disease, TID, SES, Colorblind, Sportdays and BMI	Forward selection strategy with CV (for main effects). stepwise forward

interaction terms) for the model.				selection (for interaction terms)
Assessment of model assumptions	FEV (Forced Expiratory Volume) (liters)	Continuous	Independent variables selected after model selection.	QQ Plot/ scatterplot/ Variance inflation factors (VIF), residual plots.
Determine the effect of smoking cigarettes on the FEV.	FEV (Forced Expiratory Volume) (liters)	Continuous	Independent variables selected after model selection.	Multiple Linear Regression Model/ Boxplots/Scatterplot.
Determine the effect of parental smoking on the FEV of the children	FEV (Forced Expiratory Volume) (liters)	Continuous	Independent variables selected after model selection.	Multiple Linear Regression Model/ Boxplots/Scatterplot.

11. Missing Values:

The lack of data diminishes statistical power, which refers to the likelihood that the test will reject the null hypothesis if it is wrong. Second, missing data can lead to parameter estimation bias. Third, it has the potential to reduce sample representativeness. In given data if missing value is present will be omitted.

12. Software:

R-4.1.1 statistical package will be used for analysis.

References:

[i] https://clinicaltrials.gov/ProvidedDocs/88/NCT03574688/Prot_SAP_000.pdf

[ii] Gail F. Dawson MD, MS, FAAEP, in Easy Interpretation of Biostatistics, 2008

[iii] Peter Kennedy, A Guide to Econometrics, 5th edition, p. 137

[iv] https://statswork.com/blog/multiplicity-problem-in-clinical-trials-and-some-statistical-approaches/

[vi] Godfrey, Katherine. "Simple linear regression in medical research" *New England Journal of Medicine* 313.26 (1985): 1629-1636.

[vii] Aiken, Leona S.,Stephen G. West, and Steven C. Pitts."Multiple linear regression".*Handbook of psychology* (2003): 481-507.

[viii] Hansen, Bruce. "Testing for linearity." *Journal of economic surveys* 13.5 (1999): 551-576.

[ix] Peter Kennedy, A Guide to Econometrics, 5th edition, p. 137

[x] https://othas.github.io/LIMO/