MASTER OF STATISTICS AND DATA SCIENCE

2021-22

PROJECT LEARNING FROM DATA

Final Report

Group 1

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

Table of Contents

1. Introduction

A common measurement of lung function is the forced expiratory volume (FEV), which measures how much air you can blow out of your lungs in one second. A higher FEV is usually associated with better respiratory function. It is well known that prolonged smoking diminishes FEV in adults, and those adults with diminished FEV also tend to have decreased pulmonary function as measured by other clinical variables, such as blood oxygen and carbon dioxide levels.

A common clinical technique to measure FEV is through spiromery [1]. The result through this technique heavily depends on technicality of implementation as well as personal attributes of a patient. Personal attributes that will help in determining an accurate FEV score will be patient's Gender, Age Height, and indication of being smoker or non-smoker. FEV values greater than 80% of the predicted average value are considered as normal. These results depend heavily on the technicality of implementation as well as personal attributes of the patient. Personal attributes that will help determine an accurate FEV score will be patients' age, height, gender and indication of being smoker or non-smoker.

The data used during this simulation, is inspired by the data of tiger et.al [2]. The dataset consists of 15 variables some of which are directly measured, and some are qualitative in nature.

The data set is composed of a sample population consisting of 654 children, male and female aged between 3 to 19 years old from the East Boston area in the late 1970's. This dataset consists of 15 variables of measurement of children including Age (years), height (inches), gender (male/female), smoke (Yes/No), FEV (liters), BMI (kg/m^2), Socio-economic status (Low, middle and high), Smoker status of the parents (Yes/No), sport, school results (poor, average and good), T1D (Yes, No), colorblind (Yes, No), sport days, and lung disease (Yes/No).

An investigation of the relationship between child a child's FEV and their current smoking status will be sought as well as other comparisons between predictor variables. This is an observational study i.e., the child made an indication if they, themselves were smokers or not. It is important to note that the younger the child the lower their FEV lung capacity will be due to the stature of their body alone. Therefore, in a normal case the older the child the higher the lung capacity as their body grows.

Aim of this study is to assess the effect of children smoke on FEV and the effect of parental smoke on child's FEV. In order to predict this effect a linear regression model has been used. Upon initial findings the model assumptions have been tested and best fit model for the given dataset has been tried to find out by the authors in order to conduct further analysis. Additionally, an analysis has been conducted with the help of a multiple linear regression model to describe the effect of each predictor variable and FEV.

Organization of this analysis is as follows: in section 2 exploratory data analysis has b^2een conducted in order to get familiar with the predictors and outcome variable in the data set. In section 3 the effect of smoke on FEV has been assess by using a simple linear regression

model and after that, we begin with a preliminary model and have seen that it violates model assumptions. So, a final prediction model has been proposed to carry out further analysis. In section 4 the same procedure has been followed to assess the effect of parental smoke on child's FEV. In section 5 we end up with a discussion of our findings.

2. Data Exploration

Exploratory Data Analysis (EDA) is a way of evaluating datasets to summarize their essential properties, generally using visual methods, in data mining. Before beginning the modeling work, EDA is used to see what the data can tell us. It's not easy to deduce essential data qualities from a column of numbers or an entire spreadsheet. Deriving insights from raw data can be tiresome, uninteresting, and overpowering. In this case, exploratory data analysis approaches have been developed as a help.

There are two ways to categorize exploratory data analysis. The first distinction is that each method is either non-graphical or graphical. Second, each strategy is univariate or multivariate (usually just bivariate). Initially, exploratory data analysis was performed to get a sense of the data collection. The boxplot and scatter plot were used to investigate the possible association between two variables and assess the form of the data distribution and identify any outliers.

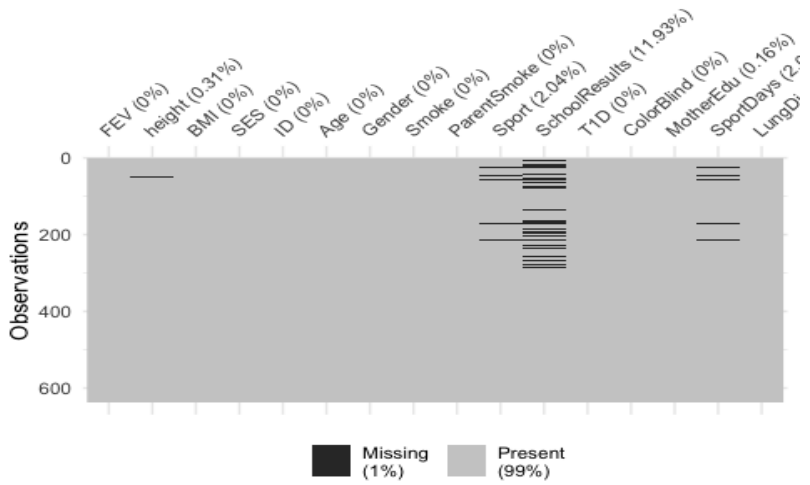
2.1 Data Description

In this study, 654 subjects were recruited and measured their forced expiratory volume (FEV, liters) and assessed how age, gender, height, smoking habit, lung disease and sports activities are related to FEV.

Software R-4.1.1 statistical package was used for statistical analysis

2.2 Missing Values

Missing data can significantly impact quantitative research, with some of the most common consequences being biased parameter estimations, loss of information, decreased statistical power, increased standard errors, and diminished generalizability of findings. As a rule, if your analytic dataset has 10% or fewer missing data for the main outcome variable for a single component, you can usually proceed with your analysis without further evaluation or correction. If more than 10% of the data for a variable is missing, you may need to investigate respondents and non-respondents further concerning the primary outcome variable. We used the missing value map and missing data matrix to analyze the missingness pattern. The absence of missing values in variables with more than 10% missing values has not been seen as a significant influence.



SchoolResults=76, MotherEdu=1, and SportDays=13 are all missing data in the graph. So, in the original data,99% of the data was present. As the absence of missing values in variables with more than 10% missing values has not been seen as a significant influence.

2.3 Data Cleaning

Data cleansing is the approach of going through all the data and removing or updating any information that is missing, wrong, incorrectly structured, duplicated, or irrelevant (source). Data cleansing usually entails cleaning up data that has been gathered in one location. Though data cleaning can include deleting information, but it is mainly concerned with updating, correcting, and combining data to make your system as efficient as possible.

For this analysis data were cleaned before analysis as column Height, row 19, which contains height 150m (height cannot be 150m), and row 29, which contains lung disease 2(not usable) because lung disease is a binary indicator in 0 1.A multivariate

3. Assess the effect of cigarette smoking of children on lung function as measured by forced expiratory volume in liters (FEV)

3.1 Methods

Regression methods provide tools for explaining how an outcome variable varies for different values of explanatory variables and predict the value of FEV for subjects. Such methods can also be used to predict future outcomes for new individuals based on their set of explanatory variable values. The main objective of this study is to determine the effect of smoking on FEV.

3.2 Descriptive Statistics

Descriptive statistics provides the summarized population data by describing what is observed in the sample numerically. Descriptive statistics for continuous variables include the minimum, 25%, median, mean, 75%, max, and standard deviation.

Table 1:Description

Variables	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation
FEV	0.791	1.95	2.53	2.61	3.09	5.638	0.844
height	1.17	1.45	1.56	1.55	1.65	1.88	0.144
BMI	14.5	18.6	20.1	20.12	21.5	27.0	2.11
Age	3	8	10	9.815	12	19	2.86
Sport	0	1	2	2.17	3	15	2.35
SportDays	0	1	2	1.45	2	5	1.03

3.3 Exploratory Data Analysis

This section is a summary of the exploratory data analysis for the response variable as well as predictor variables. From **Figure 1** below the average forced expiratory volume in liters for non-smokers was lower as compared to smoker subjects.

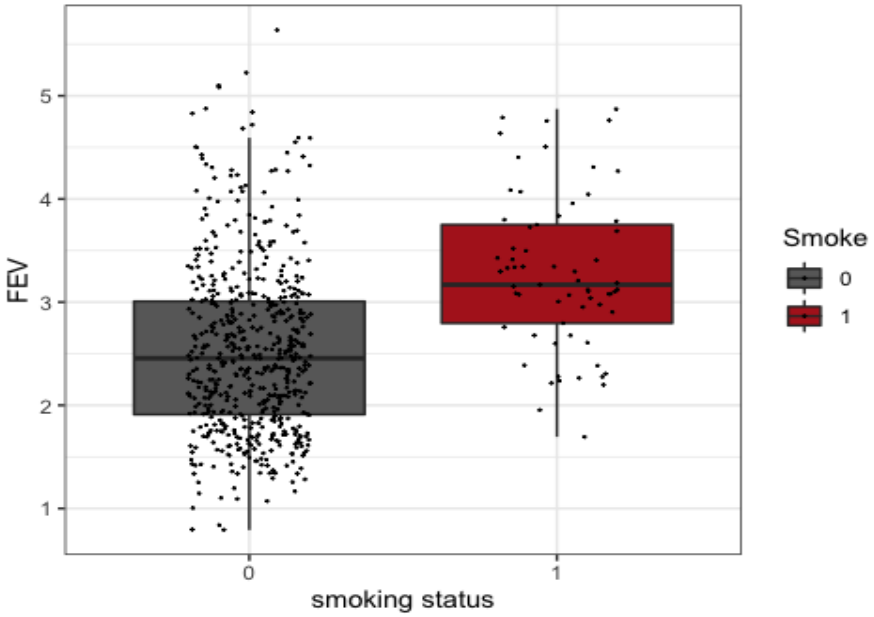


Figure 1 FEV versus Smoking

In order to visualize the association between predictors and outcome variable, scatter plot is presented below. From scatter plot, we visualize that the FEV variable is highly correlated with age and height.

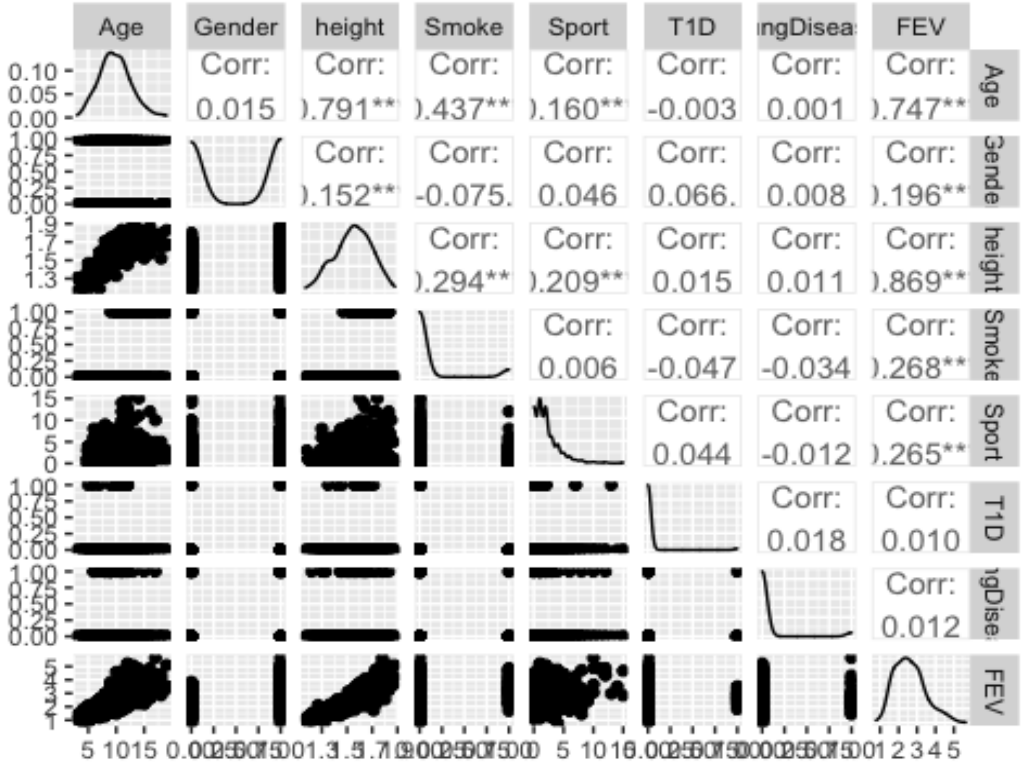


Figure 2: Correlation between regressors

From scatter plot, we visualize that the FEV variable is highly correlated with age and height. So, let's explore these variables.

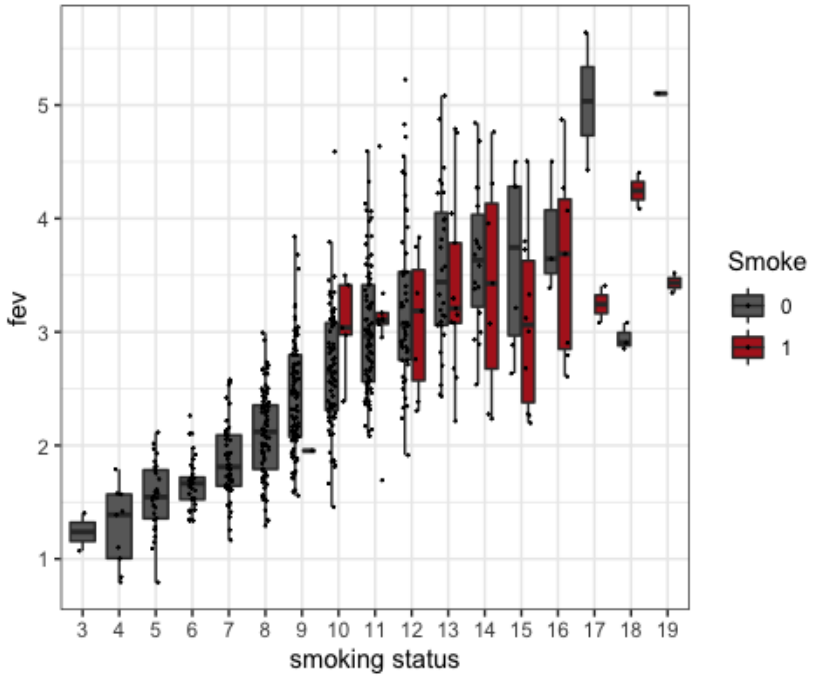


Figure 3: Comparison of FEV for smokers and nonsmokers, accounting for age.

Figure 3 indicated that the effect smokers on FEV as compared to non-smoker subjects by adjusting age of the individuals. So, age may be considered as a cofounder for the relationship between smoking and FEV.

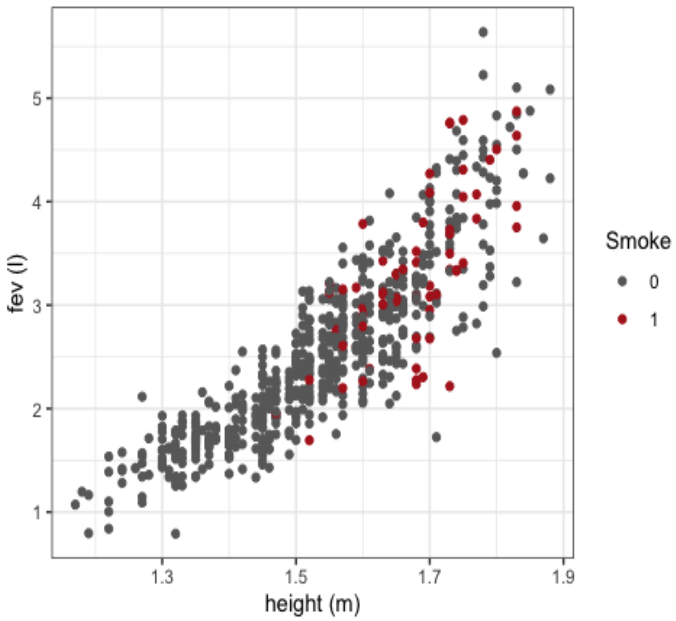


Figure 4: Comparison of FEV for smokers and nonsmokers, accounting for height.

Figure 4: Comparison of FEV for smokers and nonsmokers, accounting for height. suggests that, generally, the difference between the smokers' and nonsmokers' FEV values appears nominal

when accounting for height. Further, the average FEV value as a function of height appears curved, possibly quadratic.

3.4 Unadjusted Analysis

In the unadjusted analysis [5], smoking is positively associated with lung function. Statistical analysis is shown in **Table 2**

Table 2: Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.53045	0.03402	74.372	<2e-16
Smoke	0.74641	0.10651	7.008	6.2e-12

- **Residual standard error: 0.8137 on 635 degrees of freedom**
- **Multiple R-squared: 0.07178**
- **Adjusted R-squared: 0.07032**
- **F-statistic: 49.11 on 1 and 635 DF**
- **P-value: 6.204e-12**

In summary, the mean difference in FEV in smokers and non-smokers is 0.71. This is surprisingly given what we know about smoking. How can a positive relation between SMOKE and FEV exist given what we know about the physiological effects of smoking? The answer lies in understanding the confounding effects of AGE on SMOKE and FEV. In this child and adolescent population, nonsmokers are younger than nonsmokers (mean ages: 9.5 years vs. 13.5 years, respectively). It is therefore not surprising that AGE confounds the relation between SMOKE and FEV.

3.5 Multivariate Linear Regression Models

Confounding [3] is a bias due to non-comparability of groups. For the current dataset smokers are older than nonsmokers, and age is a strong predictor of the outcome. Therefore, the relation between FEV and SMOKE is confounded by AGE. Fortunately, we can use multiple regression model [6] to adjust for this problem. In effect, multiple regression will balancing the effects of age at all levels as we compare smokers and non-smokers. The model can help us to understand the effect of a change in the regressor to the outcome variable. Beside the main effects, potential interactions between regressors were included in the model, and their effect on the outcome was studied. The significance of those possible interactions is checked and if they turn out to be significant, they remain in the model.

The number of variables we have chosen in answering the research question is rather large. Thus, we will make use of a model selection procedure to choose which variables are important to answer the research question. The stepwise selection procedure was chosen, making use of forward selection strategy with CV [4] in each step of the model building procedure after performing the stepwise procedure (See Appendix), we end up with the following variables as predictor: Age , Smoke, height, gender, Sport, T1D and Lung Disease. Although, standard error of variable lung Disease is higher but we have included this

variable in our analysis because of its association with FEV. For interaction terms a stepwise selection procedure has been adopted.

We added two-way interaction of regressors Age: Smoke, Smoke: Gender, Height: Age and LungDisease: Sport at the 5% level of significance (See Appendix). Due to high correlation between variables age and height we have decided not to include two way interactions of these variables. Before doing the analysis let's explore the interaction plots with outcome variable.

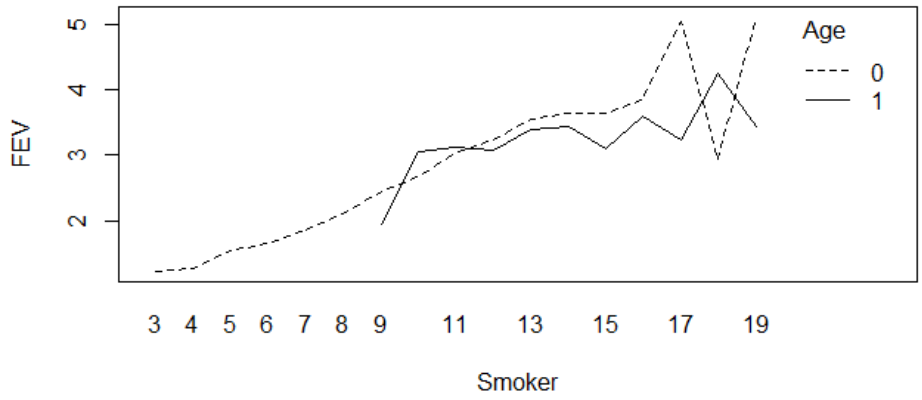


Figure 5: Interaction effect of smoke and age on FEV

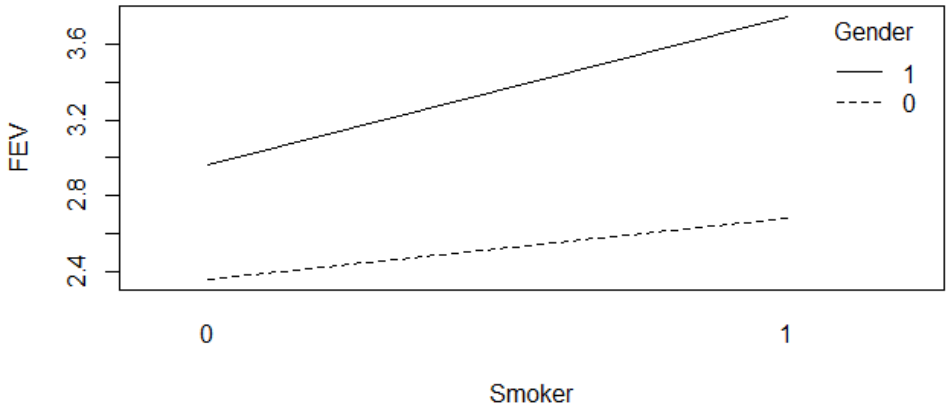


Figure 6: Interaction effect of smoke and Gender on FEV

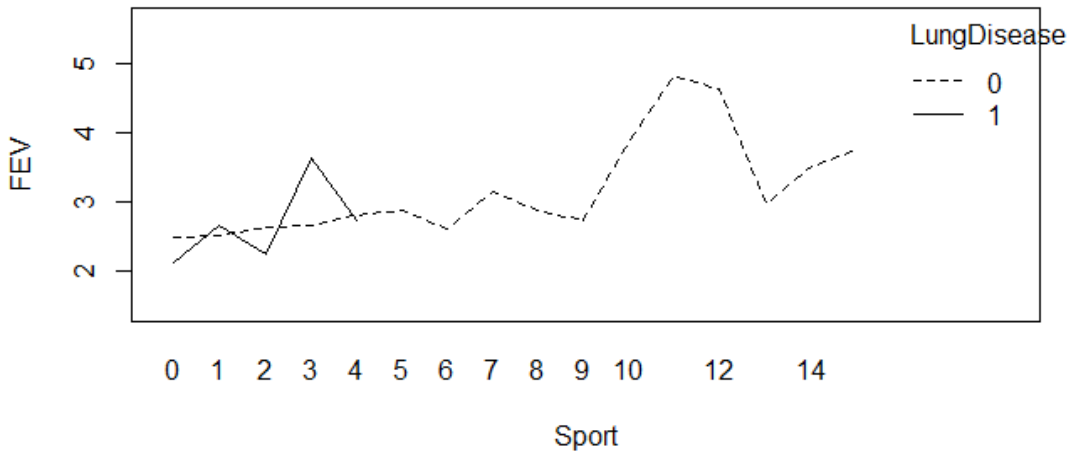


Figure 7: Two-way Interaction of sport and LungDisease with outcome variable

3.6 Adjusted Analysis (Non-Additive Model)

To begin, we will look at the model which includes all the main and two-way interaction effects, to assess if they are or not significant. We need to look at the model's assumptions first.

```
model<-lm(FEV~(Smoke+height+Age+Gender+Sport+T1D+LungDisease+
                 Smoke*Age+Smoke*Gender+LungDisease*Sport), data=data_fev)
```

3.7 Model Assumptions

When designing a multiple linear regression model, there are numerous assumptions and diagnostics to consider:

- Each observation should be self-contained.
- The error terms are distributed normally.
- The error term is anticipated to have a value of zero.
- The variances of the error terms are constant (homoscedasticity).
- The output variable and the regressors' regression function is linear.
- There is no significant multicollinearity among the regressors.
- Detecting outliers and their impact on the fit.

3.7.1 Normality Assumption

Normality tests [7] are used to determine if a data set is well-modeled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed**. Figure 8** shows the plot deviated from the reference line at tail

and head. The normality plot shows heavy tails on both ends. There is extreme observation at the top and not clearly indicated whether normality assumption met or not.

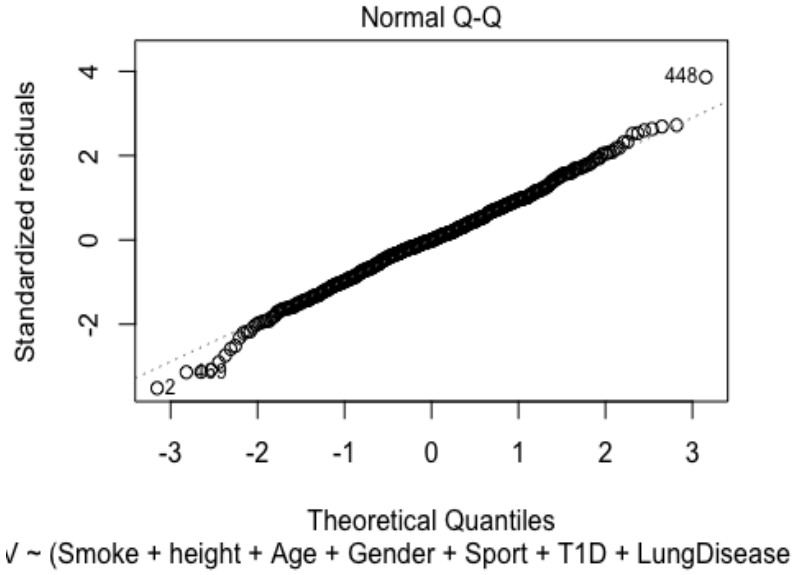


Figure 8: QQ Plot

3.7.2 Linearity Assumption

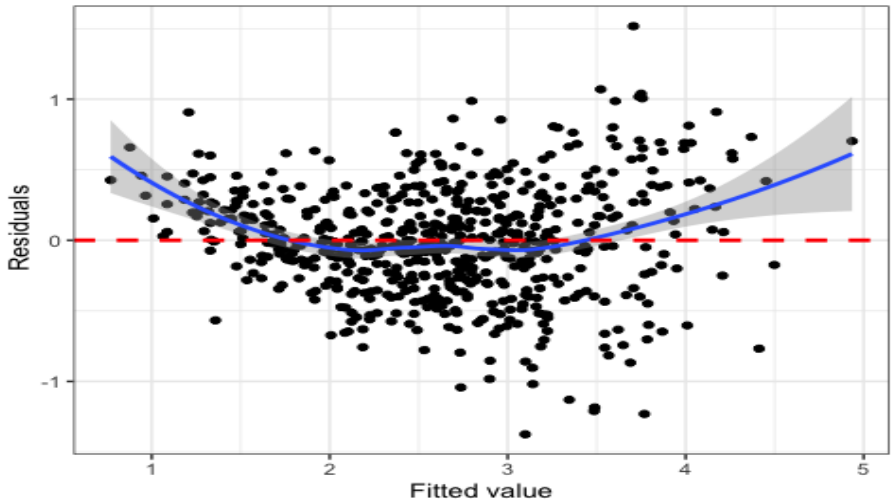


Figure 9: Residuals

In **Figure 9**, there is a linearity and homogeneity issues. The plot shows that residuals vary as the fitted values increase [8]. So, we should first try to improve the model fit, and then assess again the constant-variance assumption. The approach that we have adopted to make model fit is transformation of the regressor. As we have seen in exploratory data analysis that the nature of variable 'height' seems to be quadratic and adding quadratic

effect of height into the model might be a good option. Note that even nonlinear transformations still result in linear regression models (the model is still linear in the β-parameters). So let's refit the model:

```
model1<-lm(FEV~(Smoke+height+I(height^2)+Age+Gender+Sport+
                  T1D+LungDisease+Smoke*Age+Smoke*Gender+LungDisease*Sport),
data=data_fev)
```

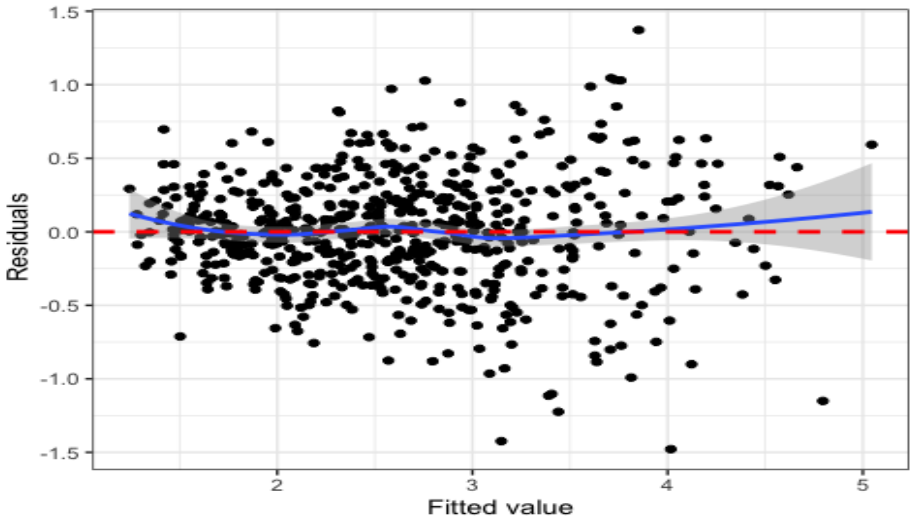


Figure 10: Residuals

From the above plot the fitted blue line seems a linear. The points are also systematically distributed and variance seems like a constant.

3.7.3 Multicollinearity

Linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. Variance Inflation Factor (VIF) – the variance inflation factor of the linear regression is defined as VIF = 1/T. With VIF > 5 there is an indication that multicollinearity may be present; with VIF > 10 there is certainly multicollinearity among the variables [9]. The values of the VIFs for the interaction's terms and their regressors will be high. They can be ignored; the high values are normal because the interaction terms and their regressors are correlated. In what follows we will present only the VIF values for the main effects in the model. If value of VIF is higher than 5 then it is a cause of concern. Here we can see that all regressors have a VIF of less than 3, which is not worrisome.

Table 3: VIF values of regressors

SMOKE	HEIGHT	AGE	GENDER	SPPORT	T1D	LUNG DISEASE
1.267000	2.675327	2.907060	1.061453	1.051617	1.008318	1.001933

3.7.4 Leverages

An outlier is a point that falls far from the other data points. It's a point that's extreme in some way. If the parameter estimates change considerably when a value is removed from the calculations, the value is influential.

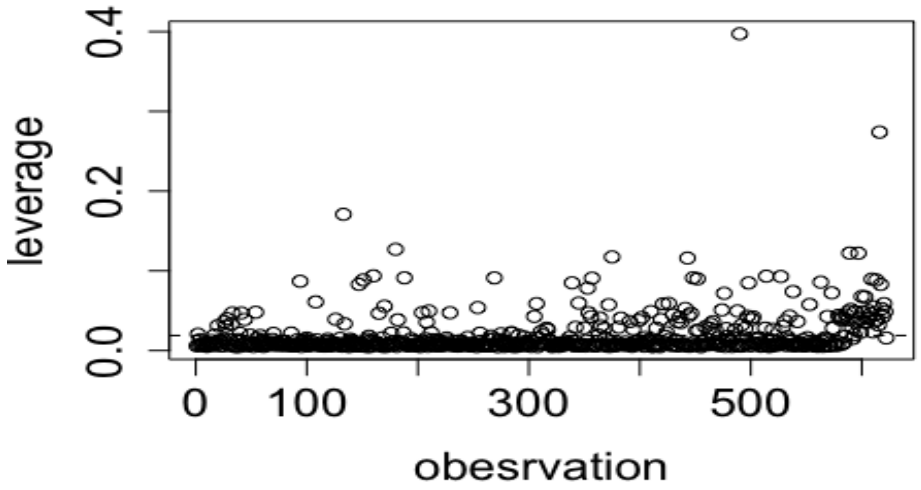


Figure 11: Detection of outliers:

Table 4: Bonferroni test

obs	rstudent	*unadjusted p-value*	*Bonferroni p*
469	-3.958774	8.4227e-05	0.052389

From above **figure**, one observation was extremely close to Cook's distance line [10]. That Bonferroni test [11] is showed that observation 469 is detected as an outlier with p-value 8.422e-05 at 5% level of significance respectively. Observation has a negative residual which may be underestimated the FEV data of the individuals. Removing this FEV observation will have a significant impact on the values of the intercept and slope in our regression model. But since sample size is large so we can ignore this term. So, there is no indication of strong influential outliers. Hence the following model is our best fitted model

3.8 Transformed model

Now we can look at the summary of the regression model, which will give us an idea on which regressors have a significant effect on the FEV. Recall our research question we are supposed to access the effect of smoking on FEV. The hypotheses to be tested are as follows: $H_0: \beta_i = 0$ i.e. there is no significant effect of smoking on FEV. The alternative hypothesis is $H_0: \beta_i \neq 0$ i.e. there is a significant effect of smoking on FEV.

Table 5: Residuals

Min	1Q	Median	3Q	Max
-1.47841	-0.22193	0.00112	0.23201	1.37219

Table 6: Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)	Lower CI	Upper CI
(Intercept)	6.123768	1.666763	3.674	0.000260	2.850478491	9.39705765
Smoke	0.191957	0.302602	0.634	0.526086	-0.402309925	0.78622467
height	-9.673273	2.165057	-4.468	9.42e-06	13.925143160	-5.42140357
I(height^2)	4.431716	0.696898	6.359	3.98e-10	3.063104551	5.80032677
Age	0.067555	0.010486	6.442	2.39e-10	0.046961560	0.08814921
Gender	0.072321	0.034263	2.111	0.035200	0.005033050	0.13960923
Sport	0.022030	0.006990	3.152	0.001703	0.008302411	0.03575754
T1D	0.006914	0.112412	0.062	0.950976	-0.213847289	0.22767530
LungDisease	-0.152799	0.084380	1.811	0.070658	-0.318509690	0.01291195
Smoke:Age	-0.026836	0.022953	-1.169	0.242782	-0.071911668	0.01823966
Smoke:Gender	0.175769	0.104589	1.681	0.093360	-0.029630003	0.38116790
Sport:LungDisease	0.081042	0.024354	3.325	0.000937	0.033176147	0.12890702

- **Residual standard error: 0.3828 on 610 degrees of freedom**
- **(15 observations deleted due to missingness)**
- **Multiple R-squared: 0.7921**
- **Adjusted R-squared: 0.7883**
- **F-statistic: 211.2 on 11 and 610 DF**
- **P-value: <2.2e-16.**

3.9 Discussion

From summary, we infer that there is no evidence against null hypothesis (p=0.526) at 5% level of significance. Hence there is no significant effect of smoke on FEV. The result is strange as we know smoking has an adverse effect on lungs. However, if we see other regressors in the summary statistics, height has a significant negative effect on FEV which means the larger the height the lower the FEV which seems to be odd. As older children have higher FEV than younger ones. This is due to the nature of this variable. So, we consider the quadratic effect of height which shows that there is a significant positive effect of 4.432/cm2 (CI 3.063104551 to 5.80032677) on FEV. Similar in case of Age it has significant positive effect of 0.067/year (CI 0.046961560 to 0.08814921) on FEV which is true that each year FEV of child increases 6%. The variables Gender and sports also have a significant positive effect of 0.072321 (CI 0.005033050 to 0.13960923) and 0.022030 (CI 0.008302411 to 0.03575754) on FEV. The interaction effect of sport and lung disease also have a significant positive effect which is obvious, the more anyone is engaged in a physical activity the more he/she is physically strong and have higher FEV.

4. Assess the effect of parent cigarette smoking on lung function as measured by forced expiratory volume in liters (FEV)

For this research question, the same methodology explained how an outcome variable varies for different explanatory variables and predicts the value of FEV for subjects. The main objective of this study is to determine the effect of parents' smoking on the FEV of children. All measured are reported to see the impact of all variables of interest on the outcome.

4.1 Exploratory Data Analysis

Exploratory data analysis is used to provide an overview of the data set. Boxplot is used to study the potential relationship between two variables.

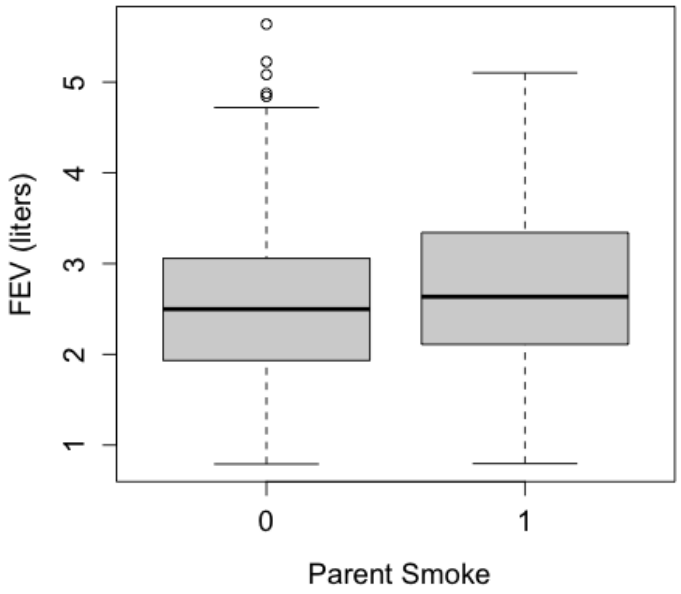


Figure 12: FEV vs. Parent Smoke

Figure 12 shows that the Parent who smokes their kids seems to have "Higher FEV". Although it was not expected but there could be some reason behind that. Let's explore some more analysis.

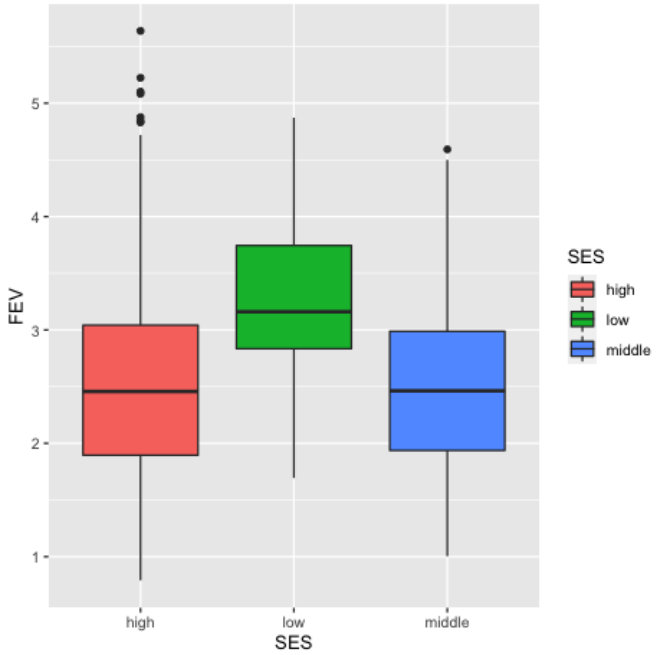


Figure 13: FEV vs. Parent Smoke

Figure 13 Shows that if the Socio-Economic Status of a kid is high, they will have "lower FEV" This seems quite interesting and meaningful that lower status has better FEV as they are less inclined to smoke due to the economic situation.

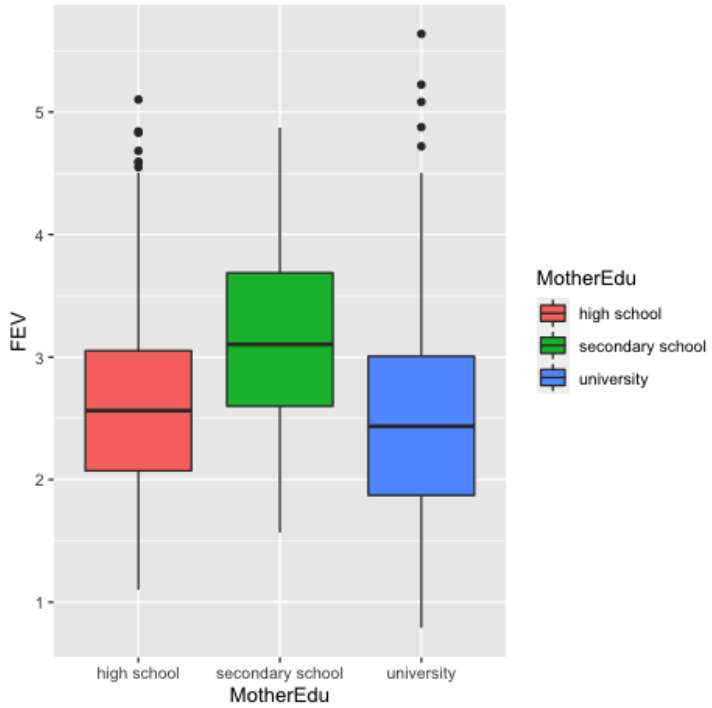


Figure 14: FEV vs. Mother Education

Figure 14 shows that Higher the Mother Educated "Lower the FEV" This is identical result from previous box plot i.e., lower mother education has better FEV as they are less inclined to smoke due to the economic situation. By these two results we can say that mother education and SES are related to each other.

4.2 Unadjusted Analysis

In the unadjusted analysis, Parent smoking is positively associated with lung function.

Table 7: Residuals

Min	1Q	Median	3Q	Max
-1.97760	-0.63983	-0.08283	0.50817	3.06817

Table 8: Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.56983	0.03681	69.817	<2e-16
ParentSmoke	0.20377	0.08663	2.352	0.019

- **Residual standard error: 0.841 on 635 degrees of freedom**
- **Multiple R-squared: 0.008638**
- **Adjusted R-squared: 0.007077**
- **F-statistic: 5.533 on 1 and 635 DF**
- **P-value: 0.01897**

4.3 Multivariate Linear Regression Models

For this study, a multiple linear regression was used to estimate the relationship between the outcome and the predictor variables. The model can help us to understand the effect of a change in the regressor to the outcome variable. In this model we are taking all variables of Primary Research question with SES and MotherEdu.

4.4 Adjusted Analysis (Non-Additive Model)

To begin, we will look at the model which includes all the main and two-way interaction effects, to assess if they are or not significant. We need to look at the model's assumptions first.

```
modelA<-lm(FEV~ParentSmoke+height+Age+Gender+Sport+T1D+LungDisease+SES+Mother
Edu+Smoke*Age+ Smoke*Gender+Sport*LungDisease,data=data_fev)
```

4.5 Model Assumptions:

4.5.1 Normality Assumption

For assessing the normality of error term ε, a normal probability plot, specifically a Quantile-Quantile Plot can be used using the residuals of the model.

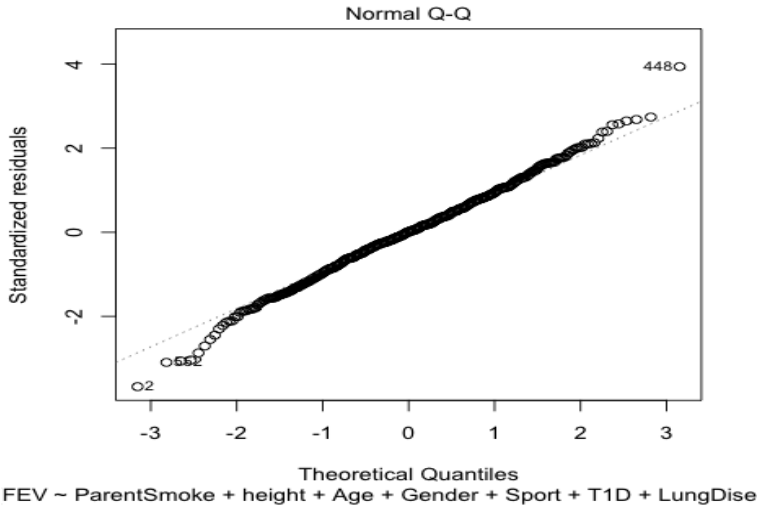


Figure 15: QQ Plot

The plot deviated from the reference line at tail and head. The normality plot shows heavy tails on both ends. There is extreme observation at the top and not clearly indicated whether normality assumption met or not.

4.5.2 Linearity Assumption

Residuals are the difference between "true value" and the "predicted value. “The residual plots are used to assess the linearity assumption, by plotting the residuals against each of the predictor variables, together with the horizontal line at zero to discover if there is still remaining any pattern in the data. If there are no systematic patterns in the residual plots, then the relation between the regressors and the output variable is considered to be linear.

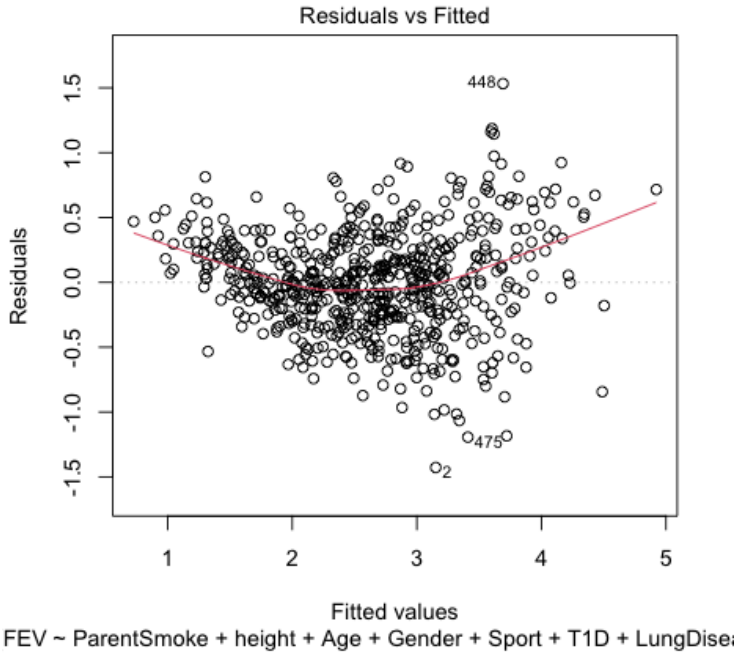


Figure 16: Residual Plot

In **Figure 16** shows that there is a linearity and homogeneity issues. The plot shows that residuals vary as the fitted values increase. So, we should first try to improve the model fit, and then assess again the constant-variance assumption. The approach that we have adopted to make model fit is transformation of the regressor. As we have seen in exploratory daa analysis that the nature of variable 'height' seems to be quadratic and adding quadratic effect of height into the model might be a good option. Note that even nonlinear transformations still result in linear regression models (the model is still linear in the β- parameters). So let's refit the model:

```
modelA1<-lm(FEV~(ParentSmoke+Smoke+height+I(height^2)+Age+Gender+Sport+
T1D+LungDisease+SES+MotherEdu+Smoke*Age+Smoke*Gender+LungDisease*Sport),
data=data_fev)
```

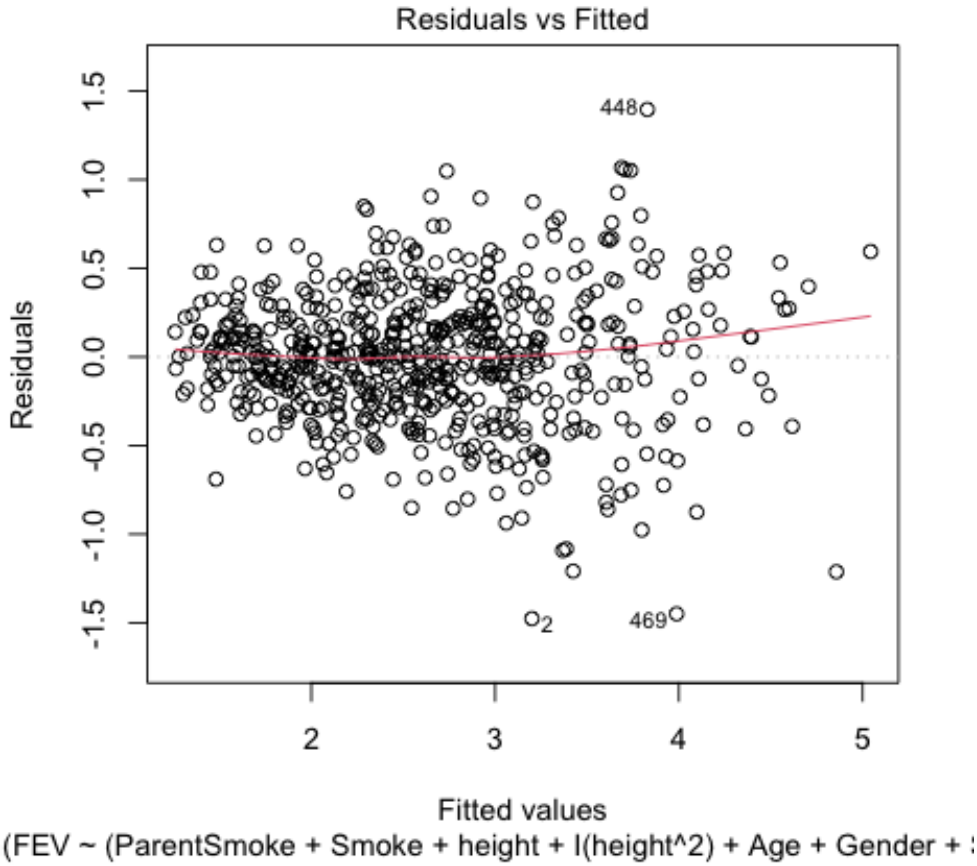


Figure 17: Residual Plot

From the above plot, the fitted Red line seems a linear. The points are also systematically distributed and variance is seems like a constant.

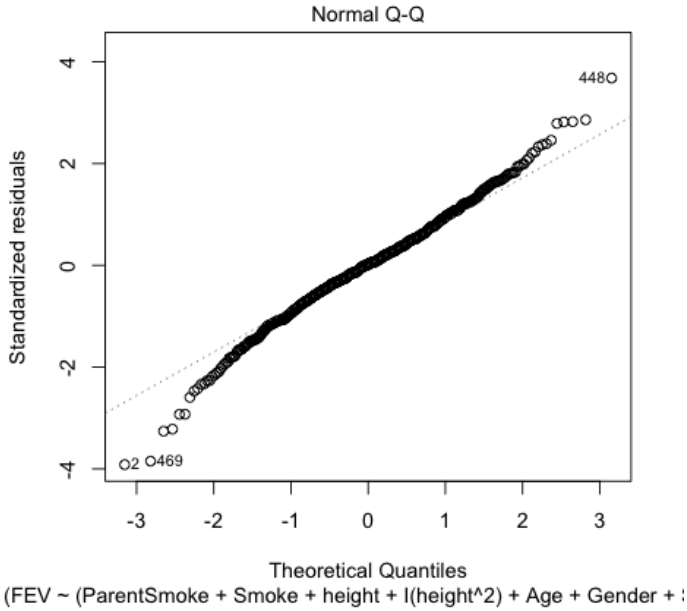


Figure 18: QQ Plot

From **Figure 18** there are some deviations at the tails, but they are not worrisome because of the large sample size. The deviated tails show presence of outliers in the data, but these outliers are influential and can be ignored.

4.5.3 Multicollinearity

Multicollinearity refers to a situation in which two or more explanatory variables are highly linearly related. VIF can't be used if there are variables with more than one degree of freedom[12]. Gvif is the square root of the VIF for individual predictors and thus can be used equivalently GVIF = $VIF^{[1/(2*df)]}$

Table 9: Generalized Variance Inflation Factor (GVIF)

	GVIF	DF	GVIF^(1/(2*DF)
ParentSmoke	1.145742	1	1.070393
Smoke	8.734328	1	2.955399
height	2.695599	1	1.641828
Age	2.941148	1	1.714978
Gender	1.07317	1	1.036203
Sport	1.061540	1	1.030311
T1D	1.009825	1	1.004900
LungDisease	1.006899	1	1.003443
SES	19.452069	2	2.100107
MotherEdu	10.923276	2	1.817976

Here we can see that all regressors have a GVIF^(1/(2*Df)) of less than 3, which is not worrisome.

4.5.4 Leverages

An outlier is a point that falls far from the other data points. It's a point that's extreme in some way. If the parameter estimates change considerably when a value is removed from the calculations, the value is influential.

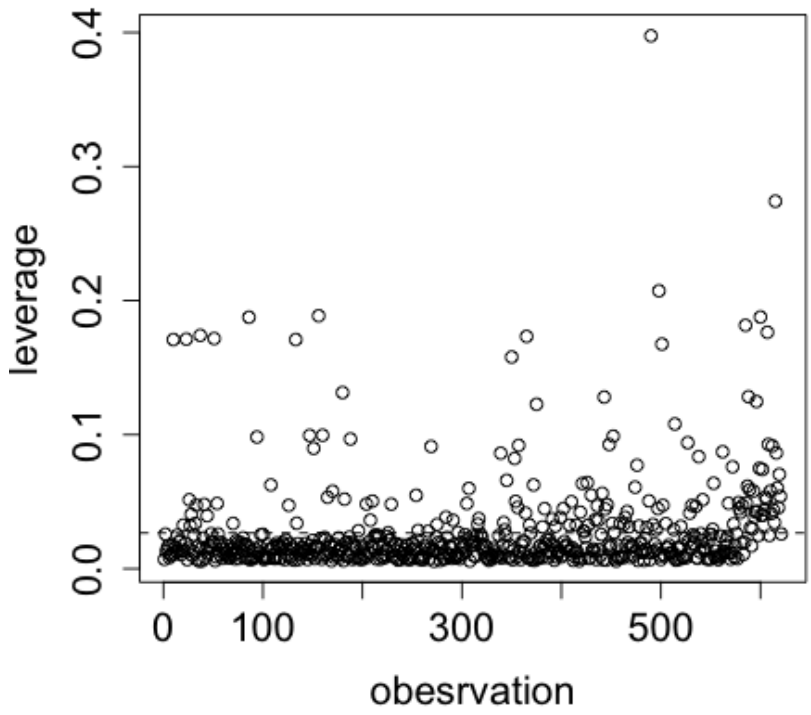


Figure 19:Leverages

Table 10: Bonferroni Test

obs	rstudent	*unadjusted p-value*	*Bonferroni p*
2	-3.962631	8.3002e-05	0.051544

From above **Figure 19** one observation was extremely close to Cook's distance line. That Bonferroni test is showed that observation 2 is detected as an outlier with p-value 8.3002e-05 at 5% level of significance respectively. observation has a negative residual which may be underestimated the FEV data of the individuals. removing this FEV observation will have a significant impact on the values of the intercept and slope in our regression model. But since sample size is large so we can ignore this term. So, there is no indication of strong influential outliers. Hence the following model is our best fitted model.

4.6 Transformed model

Now we can look at the summary of the regression model, which will give us an idea on which regressors have a significant effect on the FEV. Recall our research question we are supposed to access the effect of Parent Smoking on FEV. The hypotheses to be tested is as follows: $H_0: \beta_{ParentSmoke} = 0$ i.e. there is no significant effect of Parent Smoking on FEV. The alternative hypothesis is $H_1: \beta_{ParentSmoke} \neq 0$ i.e., there is a significant effect of Parent Smoking on FEV.

Table 11: Residuals

Min	1Q	Median	3Q	Max
-1.47653	-0.21550	0.00969	0.21752	1.39561

Table 12: Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)	Lower CI	Upper CI
(Intercept)	6.208916	1.667461	3.724	0.000215	2.934189256	9.483642064
ParentSmoke	-0.003560	0.042845	-0.083	0.933810	-0.087703700	0.080583893
Smoke	0.342495	0.327935	1.044	0.296719	-0.301536813	0.986527534
height	-9.712074	2.165559	-4.485	8.74e-06	-13.965014131	-5.459134433
I(height^2)	4.446091	0.696939	6.379	3.54e-10	3.077373196	5.814809425
Age	0.066520	0.010573	6.291	6.05e-10	0.045754565	0.087284479
Gender	0.075650	0.034347	2.203	0.028006	0.008195886	0.143104286
Sport	0.022098	0.007007	3.154	0.001692	0.008336917	0.035858552
T1D	0.002839	0.112283	0.025	0.979836	-0.217673969	0.223352055
LungDisease	-0.146861	0.084292	-0.094	0.081969	-0.312402173	0.018681134
SESlow	-0.212389	0.224129	0.275	0.343702	-0.652556357	0.227777616
SESmiddle	0.011998	0.043605	0.275	0.783296	-0.073638486	0.097634365
MotherEdusecondary school	-0.015260	0.163170	-0.094	0.925522	-0.335708104	0.305189051
MotherEduuniversity	-0.076013	0.037636	-2.020	0.043857	-0.149927189	-0.002099581
Smoke:Age	-0.025778	0.023034	-1.119	0.263517	-0.071013855	0.019457410
Smoke:Gender	0.169659	0.104941	1.617	0.106464	-0.036435170	0.375752650
Sport:LungDisease	0.083080	0.024354	3.411	0.000690	0.035250808	0.130909299

- **Residual standard error: 0.3821 on 604 degrees of freedom**
- **(16 observations deleted due to missingness)**
- **Multiple R-squared: 0.7947**
- **Adjusted R-squared: 0.7893**

- **F-statistic: 146.1 on 16 and 604 DF**
- **P-value: <2.2e-16**

4.7 Discussion

Multiple regression was carried out to investigate whether Parent Smoke could significantly affect the FEV on Children between 3 to 19 years. The results of the regression indicated that the model explains 79% variation in the response variable around its mean and show that there is no significant effect of Parent smoke on FEV.
From summary, we derive that there is no evidence against null hypothesis (p=0.93) at 5% level of significance. Hence there is no significant effect of parent smoke on FEV. Although the result is quite strange as some study shows that parent smoke has an adverse effect on lungs. Age shows that there is a significant positive effect of $\beta = 0.067$/year (SE=0.010) at 5% significance level with 95% CI (0.0457 to 0.0873) on FEV which is true that each year FEV of child increases 6%. The variables Gender and sports also have a significant positive effect on FEV. The interaction effect of sport and lung disease also have a significant positive effect $\beta = 0.083$ (SE=0.024354) at 5% level of significance, which is obvious, the more anyone is engaged in a physical activity the more he/she is physically strong and have higher FEV. For this given research question, it was also required to include the mother's education and social status as mother education can somehow impact social status. Mother Edu university shows a significant negative result, $\beta = -0.076013$ (SE=0.037) at 5% significance level with 95% CI (-0.149 to -0.0029),
From summary we conclude that there is no evidence against null hypothesis at 5% level of Figure 2: Correlation between regressors significance. Hence there is no significant effect of results Parent smoke on FEV. This shows that the Results are same as primary research.

5. Conclusion

The first objective of the study was to assess the effect of smoking on FEV. We found that the significant variable were mainly the height, age, and gender. From the research it is observed that there is no significant results were found on the outcome of given regressor at 5% significant level. Age and gender were seen cofounder as they have impact on outcome and exposure variable. The result does not have enough evidence to reject the null hypothesis and it make sense because given data is very specific. **Figure 3** shows that the child smoking starts after age 10 which shows smoking is impacting the lung strength in 11- to 19-year-old children. Because they are only children, some would not have been smoking for very long. We would expect smokers who are 40 years old or more to have a different FEV value. This is because they would have been smoking for longer and hence would have damaged and weakened their lungs even more. Model assumption was also done as it reflects the accuracy of model. Initially it does not meet some property of linearity model, but some transformations are applied for better conclusions. However, after looking from this data that might cause this insignificant result. Now it is understandable that child smoking starts after age 10 which shows smoking is impacting the lung strength in 11- to 19-year-old children. Because they are only children, However, if we see other regressors in the summary statistics, Age has a significant positive effect on

FEV which means the larger the age the lower the FEV which seems to be odd. As older children have higher FEV than younger ones. This is due to the nature of this variable,

The second objective of the study was to look if there is any impact of parent's smoking on FEV. The result here shows that there is no significant effect of parent's smoking on FEV. Social status (SES) and Mother Education were also seen variable of interest that could affect each other as mother education can have impact on social status of family. But result shows there is as such no evidence found. For this given research question, it was also required to include the mother's education and social status as mother education can somehow impact social status. The significant result indicate that higher mother education, higher status, and higher status leads to less FEV. By which we can conclude that higher status leads to less FEV in children. As a consequence of the SES result, we get an insignificant impact ($p>0.05$); thus, we can't make any predictions

This research, while it is not as straightforward as expected, could still be useful for health agencies and other stakeholders. The findings must also alarm them that there exists a potential possibility that if the children are not made to quit the smoking, then this would have detrimental effects on their FEV. Healthcare providers would appear to have a central and critical role to play in educating parents on this point, as they interact with parents at key times, such as during pregnancy, at birth, and at well-child visits, as well as on visits for illness. We believe that continued efforts to provide a smoke-free home for all children remain essential.

References

[1]. Kavitha, A., Sujatha, M., & Ramakrishnan, S. (2011). Evaluation of forced expiratory volume prediction in spirometric test using Principal Component Analysis. *International Journal of Biomedical Engineering and Technology*, *5*(2-3), 292-301.
[2]. Tager, I. B., Weiss, S. T., Rosner, B., & Speizer, F. E. (1979). Effect of parental cigarette smoking on the pulmonary function of children. *American Journal of Epidemiology*, *110*(1), 15-26.
[3].https://www.icpsr.umich.edu/web/pages/instructors/setups2012/exercises/notes/confounding-variable.html
[4]. https://othas.github.io/LIMO/
[5]. Godfrey, Katherine. "Simple linear regression in medical research" New England Journal of Medicine 313.26 (1985): 1629-1636.
[6]. Aiken, Leona S.,Stephen G. West, and Steven C. Pitts."Multiple linear regression".Handbook of psychology (2003): 481-507.
[7]. https://en.wikipedia.org/wiki/Normality_test
[8]. https://en.wikipedia.org/wiki/Homoscedasticity
[9]. https://www.investopedia.com/terms/v/variance-inflation factor.asp#:~:text=Variance%20inflation%20factor%20(VIF)%20is,only%20that%20single%20independent%20variable.
[10]. https://en.wikipedia.org/wiki/Cook%27s_distance
[11]. https://www.investopedia.com/terms/b/bonferroni-test.asp
[12]. https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/Collinearity

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
data_fev<-read.csv("/Users/aaryankaushik/Desktop/PLF/FullData_OC.csv",TRUE,se
p=";")
#convert variables ,(comma) to .(decimal)
new_FEV<-gsub(",",".",data_fev$FEV)
new_height<-gsub(",",".",data_fev$height)
new_BMI<-gsub(",",".",data_fev$BMI)
new_SES<-gsub("Middle","middle",data_fev$SES)
SES<-gsub("High","high",new_SES)

#then convert char to numeric
FEV<-as.numeric(new_FEV)
height<-as.numeric(new_height)
BMI<-as.numeric(new_BMI)

#removing old char variables from original data set
data_fev<-subset(data_fev,select = -c(FEV,height,BMI,SES))

#adding updated variables in data set
data_fev<-data.frame(FEV,height,BMI,SES,data_fev)

#removing row 19 which contain height 150m and row 29 which contain lungdisea
se 2
data_fev <- data_fev[-c(19,29), ]

#checking missing values
library(skimr)
skim(data_fev)
library(naniar)
vis_miss(data_fev)
#the dataset
head(data_fev)
#codding of factor variables
library(tidyverse)
library(dplyr)
fev <- data_fev %>%
  mutate(Gender = as.factor(Gender)) %>%
  mutate(Smoke = as.factor(Smoke)) %>%
  mutate(height)


#exploratory data analysis
library(ggplot2)

fev %>%
  ggplot(aes(x=Smoke,y=FEV,fill=Smoke)) +
  scale_fill_manual(values=c("dimgrey","firebrick")) +
```

```
  theme_bw()+
  geom_boxplot(outlier.shape=NA) +
  geom_jitter(width = 0.2, size=0.1)+
  ylab("FEV") +
  xlab("smoking status")
library(GGally)
data_fev %>% select(Age,Gender,height,Smoke, Sport, T1D, LungDisease, FEV) %>
%
ggpairs()
#BoxPlot of FEV vs Smoking stratified on age
fev%>%
  ggplot(aes(x=as.factor(Age),y=FEV,fill=Smoke)) +
  geom_boxplot(outlier.shape=NA) +
  geom_point(width = 0.2, size = 0.1, position = position_jitterdodge()) +
  theme_bw() +
  scale_fill_manual(values=c("dimgrey", "firebrick")) +
  ylab("fev") +
  xlab("smoking status")
#BoxPlot of FEV vs height
fev %>%
  ggplot(aes(x=height,y=FEV,color=Smoke)) +
  geom_point() +
  scale_color_manual(values=c("dimgrey", "firebrick")) +
  theme_bw() +
  ylab("fev (l)") +
  xlab("height (m)")
m<-lm(FEV~Smoke,data=data_fev)
summary(m)
#model selection step 1
m.age<-lm(FEV~Smoke+Age, data=data_fev)
mean((m.age$residuals/(1-influence(m.age)$h))^2)

m.gender<-lm(FEV~Smoke+Gender, data=data_fev)
mean((m.gender$residuals/(1-influence(m.gender)$h))^2)

m.height<-lm(FEV~Smoke+height, data=data_fev)
mean((m.height$residuals/(1-influence(m.height)$h))^2)

m.bmi<-lm(FEV~Smoke+BMI, data=data_fev)
mean((m.bmi$residuals/(1-influence(m.bmi)$h))^2)

m.sport<-lm(FEV~Smoke+Sport, data=data_fev)
mean((m.gender$residuals/(1-influence(m.gender)$h))^2)

m.result<-lm(FEV~Smoke+SchoolResults, data=data_fev)
mean((m.result$residuals/(1-influence(m.result)$h))^2)

m.td<-lm(FEV~Smoke+T1D, data=data_fev)
mean((m.td$residuals/(1-influence(m.td)$h))^2)
```

```r
m.colorblind<-lm(FEV~Smoke+ColorBlind, data=data_fev)
mean((m.colorblind$residuals/(1-influence(m.colorblind)$h))^2)

m.sportdays<-lm(FEV~Smoke+SportDays, data=data_fev)
mean((m.sportdays$residuals/(1-influence(m.sportdays)$h))^2)

m.lungdisease<-lm(FEV~Smoke+LungDisease, data=data_fev)
mean((m.lungdisease$residuals/(1-influence(m.lungdisease)$h))^2)

# smoke+height= 0.1766641
#model selection step 2
m.age<-lm(FEV~Smoke+height+Age, data=data_fev)
mean((m.age$residuals/(1-influence(m.age)$h))^2)

m.gender<-lm(FEV~Smoke+height+Gender, data=data_fev)
mean((m.gender$residuals/(1-influence(m.gender)$h))^2)

m.bmi<-lm(FEV~Smoke+height+BMI, data=data_fev)
mean((m.bmi$residuals/(1-influence(m.bmi)$h))^2)

m.sport<-lm(FEV~Smoke+height+Sport, data=data_fev)
mean((m.gender$residuals/(1-influence(m.gender)$h))^2)

m.result<-lm(FEV~Smoke+height+SchoolResults, data=data_fev)
mean((m.result$residuals/(1-influence(m.result)$h))^2)

m.td<-lm(FEV~Smoke+height+T1D, data=data_fev)
mean((m.td$residuals/(1-influence(m.td)$h))^2)

m.colorblind<-lm(FEV~Smoke+height+ColorBlind, data=data_fev)
mean((m.colorblind$residuals/(1-influence(m.colorblind)$h))^2)

m.sportdays<-lm(FEV~Smoke+height+SportDays, data=data_fev)
mean((m.sportdays$residuals/(1-influence(m.sportdays)$h))^2)

m.lungdisease<-lm(FEV~Smoke+height+LungDisease, data=data_fev)
mean((m.lungdisease$residuals/(1-influence(m.lungdisease)$h))^2)

# smoke+height+AGE=0.1702385
#model selection step 3

m.gender<-lm(FEV~Smoke+height+Age+Gender, data=data_fev)
mean((m.gender$residuals/(1-influence(m.gender)$h))^2)

m.bmi<-lm(FEV~Smoke+height+Age+BMI, data=data_fev)
mean((m.bmi$residuals/(1-influence(m.bmi)$h))^2)
```

```
m.sport<-lm(FEV~Smoke+height+Age+Sport, data=data_fev)
mean((m.gender$residuals/(1-influence(m.gender)$h))^2)

m.result<-lm(FEV~Smoke+height+Age+SchoolResults, data=data_fev)
mean((m.result$residuals/(1-influence(m.result)$h))^2)

m.td<-lm(FEV~Smoke+height+Age+T1D, data=data_fev)
mean((m.td$residuals/(1-influence(m.td)$h))^2)

m.colorblind<-lm(FEV~Smoke+height+Age+ColorBlind, data=data_fev)
mean((m.colorblind$residuals/(1-influence(m.colorblind)$h))^2)

m.sportdays<-lm(FEV~Smoke+height+Age+SportDays, data=data_fev)
mean((m.sportdays$residuals/(1-influence(m.sportdays)$h))^2)

m.lungdisease<-lm(FEV~Smoke+height+Age+LungDisease, data=data_fev)
mean((m.lungdisease$residuals/(1-influence(m.lungdisease)$h))^2)

#Smoke+height+Age+Gender+Sport=0.165599
#model selection step 4

m.bmi<-lm(FEV~Smoke+height+Age+Gender+Sport+BMI, data=data_fev)
mean((m.bmi$residuals/(1-influence(m.bmi)$h))^2)

m.result<-lm(FEV~Smoke+height+Age+Gender+Sport+SchoolResults, data=data_fev)
mean((m.result$residuals/(1-influence(m.result)$h))^2)

m.td<-lm(FEV~Smoke+height+Age+Gender+Sport+T1D, data=data_fev)
mean((m.td$residuals/(1-influence(m.td)$h))^2)

m.colorblind<-lm(FEV~Smoke+height+Age+Gender+Sport+ColorBlind, data=data_fev)
mean((m.colorblind$residuals/(1-influence(m.colorblind)$h))^2)

m.sportdays<-lm(FEV~Smoke+height+Age+Gender+Sport+SportDays, data=data_fev)
mean((m.sportdays$residuals/(1-influence(m.sportdays)$h))^2)

m.lungdisease<-lm(FEV~Smoke+height+Age+Gender+Sport+LungDisease, data=data_fe
v)
mean((m.lungdisease$residuals/(1-influence(m.lungdisease)$h))^2)

# Smoke+height+Age+Gender+Sport+T1D=0.1615807
#model selection step 5 BMI+SchoolResults+ColorBlind+SportDays

m.bmi<-lm(FEV~Smoke+height+Age+Gender+Sport+T1D+BMI, data=data_fev)
mean((m.bmi$residuals/(1-influence(m.bmi)$h))^2)

m.result<-lm(FEV~Smoke+height+Age+Gender+Sport+T1D+SchoolResults, data=data_f
ev)
```

```
mean((m.result$residuals/(1-influence(m.result)$h))^2)

m.colorblind<-lm(FEV~Smoke+height+Age+Gender+Sport+T1D+ColorBlind, data=data_
fev)
mean((m.colorblind$residuals/(1-influence(m.colorblind)$h))^2)

m.sportdays<-lm(FEV~Smoke+height+Age+Gender+Sport+T1D+SportDays, data=data_fe
v)
mean((m.sportdays$residuals/(1-influence(m.sportdays)$h))^2)

m.lungdisease<-lm(FEV~Smoke+height+Age+Gender+Sport+T1D+LungDisease, data=dat
a_fev)
mean((m.lungdisease$residuals/(1-influence(m.lungdisease)$h))^2)

model<-lm(FEV~(Smoke+height+Age+Gender+Sport+T1D+LungDisease+
                  Smoke*Age+Smoke*Gender+LungDisease*Sport), data=data_fev)
interaction.plot(data_fev$Age data_fev$Smoke, response = data_fev$FEV, xlab="Sport", ylab
= "FEV", trace.label = "Age")
interaction.plot(data_fev$Gender, data_fev$Smoke, response = data_fev$FEV, xlab="Sport",
ylab = "FEV", trace.label = "Gender")
interaction.plot(data_fev$Sport, data_fev$LungDisease, response = data_fev$FEV, xlab="Sp
ort", ylab = "FEV", trace.label = "LungDisease")


plot(model, which =2 )
res <-ggplot(model, aes(.fitted, .resid))+geom_point()+
  geom_smooth()+
  geom_hline(yintercept=0, col="red", linetype="dashed", size= 1)+
  labs(x="Fitted value",y=" Residuals")+
  theme_bw();res
model1<-lm(FEV~(Smoke+height+I(height^2)+Age+Gender+Sport+
                   T1D+LungDisease+Smoke*Age+Smoke*Gender+LungDisease*Sport),
data=data_fev)

res <-ggplot(model1, aes(.fitted, .resid))+geom_point()+
  geom_smooth()+
  geom_hline(yintercept=0, col="red", linetype="dashed", size= 1)+
  labs(x="Fitted value",y=" Residuals")+
  theme_bw();res
plot(model1, which =2 )
library(car)
vif(lm(FEV~(Smoke+height+Age+Gender+Sport+T1D+LungDisease), data=data_fev))
h<-influence(model1)$h
plot(h,xlab="obesrvation",ylab="leverage",cex.axis=1.5,cex.lab=1.5)
abline(h=sum(h)/nrow(data_fev[NA,]),lty=2)
# Bonferonni test
outlierTest(model1)
model1<-lm(FEV~(Smoke+height+I(height^2)+Age+Gender+Sport+T1D+LungDisease+
                   Smoke*Age+Smoke*Gender+LungDisease*Sport), data=data_fev)
```

```
summary(model1)
confint(model1)
#2nd research que begins code
fev1 <- data_fev %>%
  mutate(Gender = as.factor(Gender)) %>%
  mutate(ParentSmoke = as.factor(ParentSmoke)) %>%
  mutate(height)
fev1 %>%
  ggplot(aes(x=ParentSmoke,y=FEV,fill=ParentSmoke)) +
  scale_fill_manual(values=c("dimgrey","firebrick")) +
  theme_bw()+
  geom_boxplot(outlier.shape=NA) +
  geom_jitter(width = 0.2, size=0.1) +
  ylab("FEV") +
  xlab("Parentsmoke status")

data_fev %>% select(Age,Gender,height,ParentSmoke,SES,MotherEdu,Smoke, Sport,
T1D, LungDisease, FEV) %>%
ggpairs()

m<-lm(FEV~ParentSmoke,data=data_fev)
summary(m)
modelA<-lm(FEV~ParentSmoke+height+Age+Gender+Sport+T1D+LungDisease+SES+Mother
Edu+Smoke*Age+Sport*LungDisease,data=data_fev)
plot(modelA, which =2 )
interaction.plot(data_fev$Gender, data_fev$Smoke, response = data_fev$FEV, xlab="Smoker
", ylab = "FEV", trace.label = "Age")
interaction.plot(data_fev$Age, data_fev$Smoke, response = data_fev$FEV, xlab="Smoker", y
lab = "FEV", trace.label = "Age")


res <-ggplot(modelA, aes(.fitted, .resid))+geom_point()+
  geom_smooth()+
  geom_hline(yintercept=0, col="red", linetype="dashed", size= 1)+
  labs(x="Fitted value",y=" Residuals")+
  theme_bw();res
modelA1<-lm(FEV~(ParentSmoke+Smoke+height+I(height^2)+Age+Gender+Sport+
                   T1D+LungDisease+SES+MotherEdu+Smoke*Age+Smoke*Gender+LungDi
sease*Sport), data=data_fev)
res <-ggplot(modelA1, aes(.fitted, .resid))+geom_point()+
  geom_smooth()+
  geom_hline(yintercept=0, col="red", linetype="dashed", size= 1)+
  labs(x="Fitted value",y=" Residuals")+
  theme_bw();res
plot(modelA1, which =2 )
vif(lm(FEV~(ParentSmoke+Smoke+height+Age+Gender+Sport+T1D+LungDisease+SES+Mot
herEdu), data=data_fev))
h<-influence(modelA1)$h
plot(h,xlab="obesrvation",ylab="leverage",cex.axis=1.5,cex.lab=1.5)
```

```
abline(h=sum(h)/nrow(data_fev[NA,]),lty=2)
# Bonferonni test
outlierTest(modelA1)
summary(modelA1)
confint(modelA1)
```