**Forced Expiratory Volume (FEV) Data**

**Benjamin Porter, Mingming Yue, Neil Cashin**

**STAT 482/582**

**Spring 2021**

**Data**

**The data we have is Forced Expiratory Volume (FEV), and the source is a sample of 654 youths in an age range from 3 to 19 in the area of East Boston collected during the 1970s. The data are interesting for us because we never thought the smoker age range from 3 to 19 is possible before we gain this data set.**

**Our data set has 6 variables:**

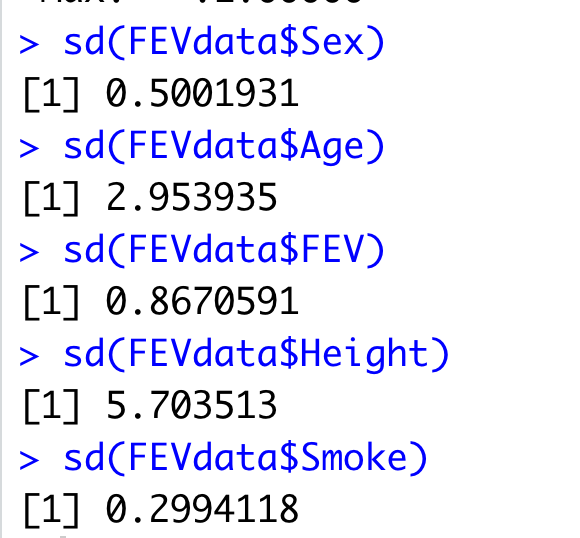
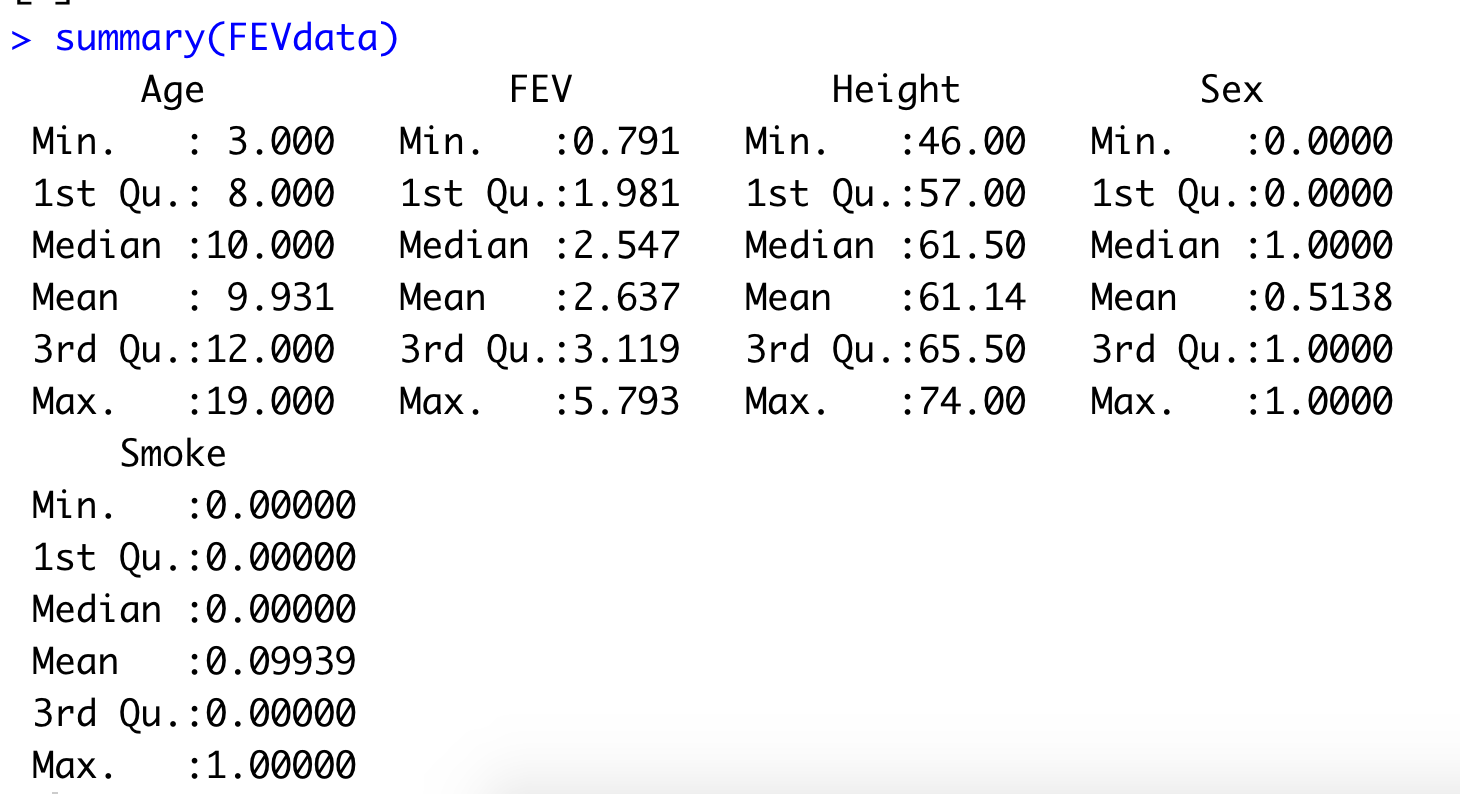
* **Age, measured in years**
* **The FEV, measured in liters**
* **Height, measured in inches**
* **The sex of the participants, where a male is measured by a 1 and a female is measured by 0**
* **The indication of smoking, where 1 = an individual smoker and 0 = non-smoker**

**Research Questions**

**We want to know the relationship between smoke and FEV.**

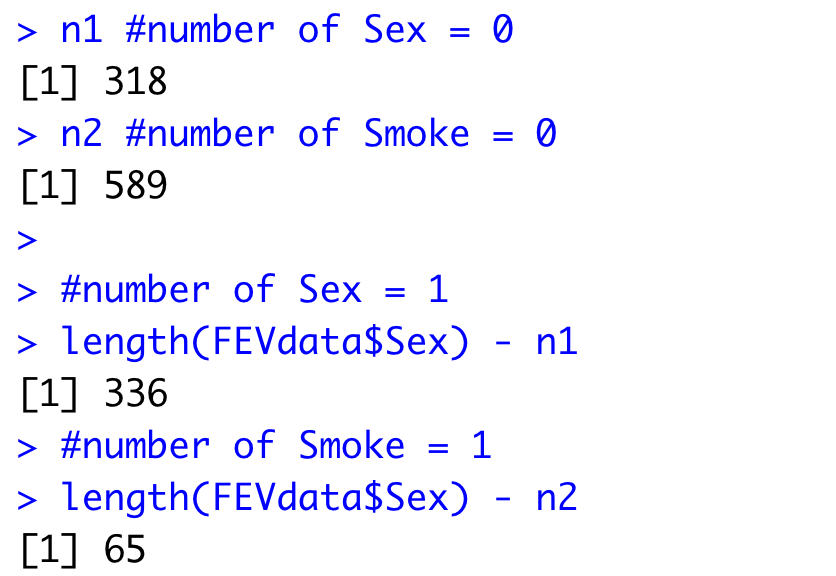
**Does the absence of either smoking or exposure to second-hand smoke affect their measured Forced EV value?**

**Descriptive Statistics**

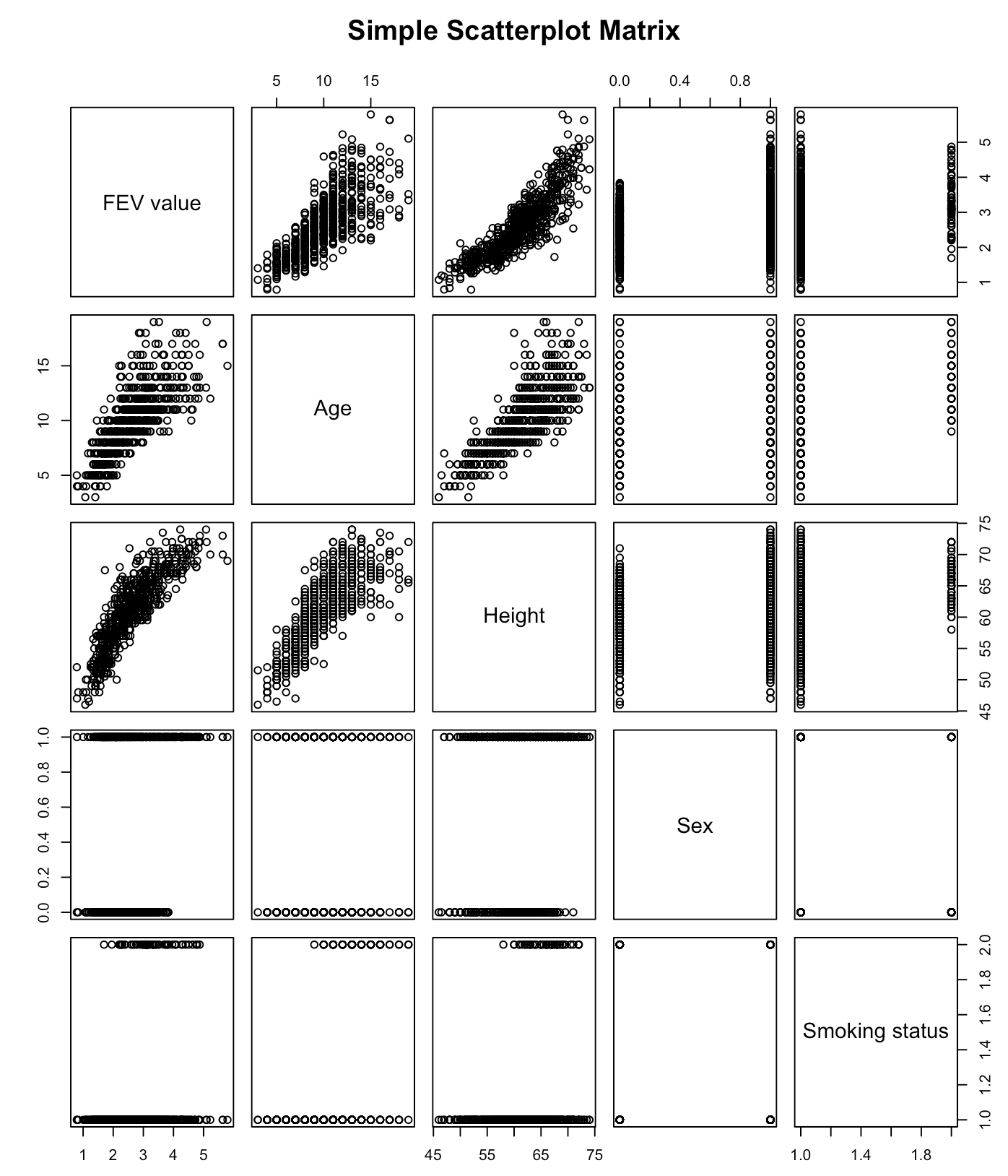
****

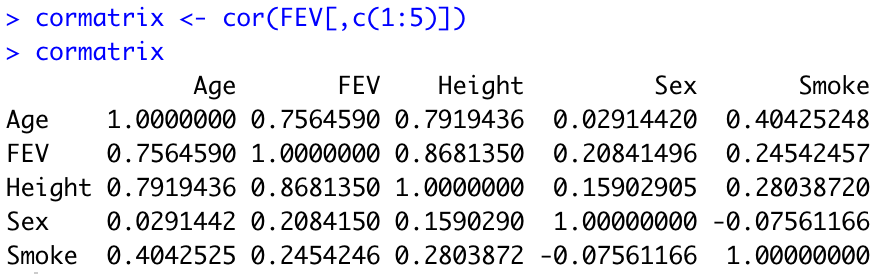
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable Name | Mean | Std. Dev. | Min | Q1 | Median | Q3 | Max |
| Age | 9.93 | 2.95 | 3 | 8 | 10 | 12 | 19 |
| Height | 61.14 | 5.7 | 46 | 57 | 61.5 | 65.5 | 74 |
| FEV | 2.64 | 0.87 | 0.79 | 1.98 | 2.55 | 3.12 | 5.79 |
| Sex | 0.51 | 0.50 |  |  |  |  |  |
| Smoke | 0.10 | 0.30 |  |  |  |  |  |

|  |  |  |
| --- | --- | --- |
| Variable Name | Count | Proportion |
| Sex | 318  336 | Sex = 0 (female) = .486  Sex = 1 (male) = .514 |
| Smoke | 589  65 | Smoke = 0 (non-smoker) = .90  Smoke = 1 (smoker) = .10 |

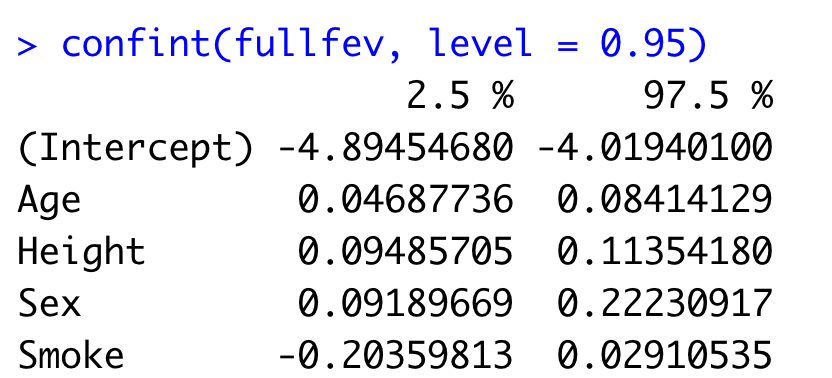
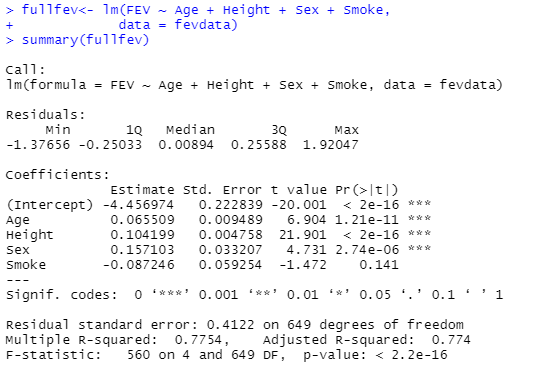


**Scatterplot and Correlation Matrix**





**Summary of Full Regression Model**

****

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable Name | Estimate | Std. Error | Test Statistic | p-value | Lower CI | Upper CI |
| Intercept | -4.456974 | 0.222839 | -20.001 | < 2e-16 | -4.894547 | -4.019401 |
| Age | 0.065509 | 0.009489 | 6.904 | 1.21e-11 | 0.04687736 | 0.08414129 |
| Height | 0.104199 | 0.004758 | 21.901 | < 2e-16 | 0.09485705 | 0.1135418 |
| Sex | 0.157103 | 0.033207 | 4.731 | 2.74e-06 | 0.09189669 | 0.2223092 |
| Smoke | -0.087246 | 0.059254 | -1.472 | 0.141 | -0.2035981 | 0.02910535 |

Explanation of Overall Model Significance

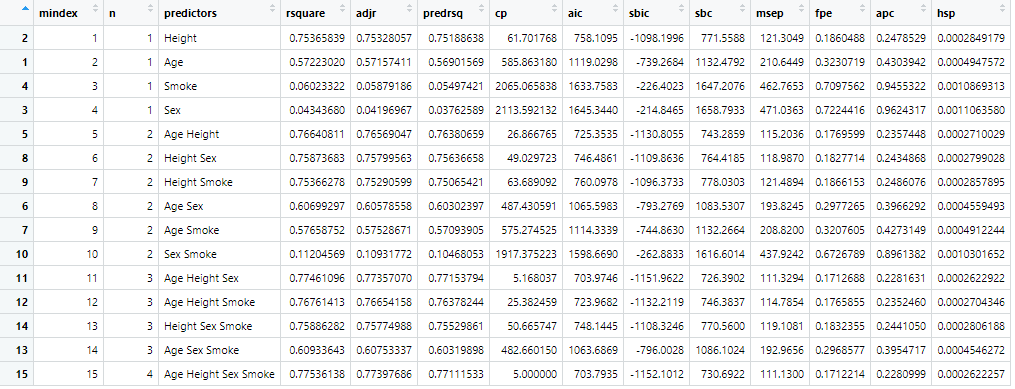
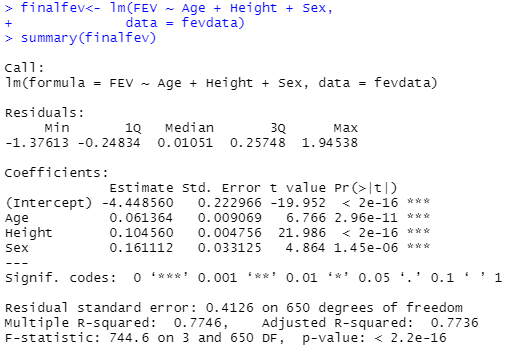
**Hypothesis test:**

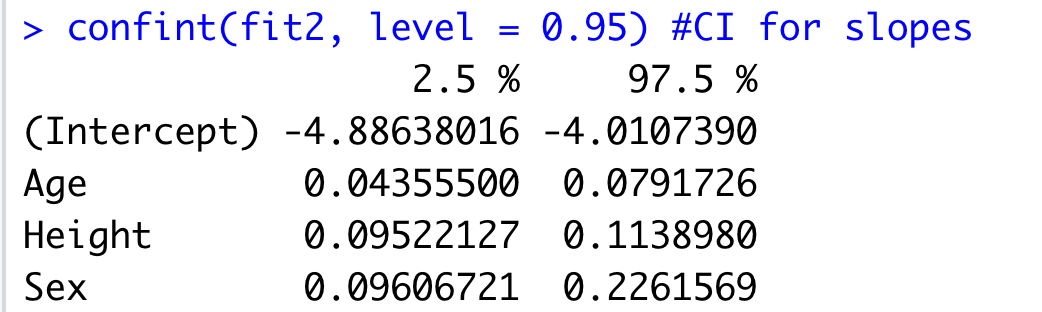
* **Null: H0: Beta1 = Beta2 = Beta3 = Beta4 = 0**
* **Alt: HA: At least one Betaj not equal to 0 for j = 1,2,3,4**
* **Test statistic: F = 560**
* **P-value: p < 0.0001**
* **Conclusion: There is extremely strong evidence that the overall linear model with Age, Height, Sex and Smoking status helps to explain the variability in the FEV in this population of youths.**

Explanation of Model Usefulness

**R^2 = 0.7754. The model with Age, Height, Sex and Smoking Status helps explain 77.54% of the variability in the FEV.**

**Summary of Final Regression Model**





|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable Name | Estimate | Std. Error | Test Statistic | p-value | Lower CI | Upper CI |
| Intercept | -4.448560 | 0.222966 | -19.952 | < 2e-16 | -4.89 | -4.01 |
| Age | 0.061364 | 0.009069 | 6.766 | 2.96e-11 | 0.04 | 0.08 |
| Height | 0.104560 | 0.004756 | 21.986 | < 2e-16 | 0.10 | 0.11 |
| Sex | 0.161112 | 0.033125 | 4.864 | 1.45e-06 | 0.10 | 0.23 |

Explanation of method for obtaining model

**We compared all possible models and we saw that the full model and the model that does not include the smoking status variable seemed to fit the data better and answer our research question. Therefore, we have not included the smoking status variable in this model.**

Explanation of Overall Model Significance-

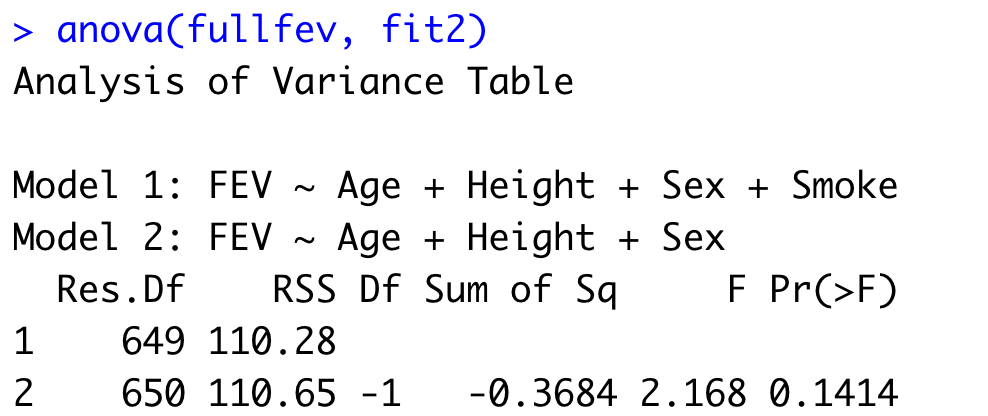
**Hypothesis Test:**

* **Null: H0: Beta1 = Beta2 = Beta3 = 0**
* **Alt: HA: At least 1 Betaj is =/= 0 for j = 1,2,3**
* **Test statistic- F = 744.6**
* **P-value- p < 0.0001**
* **Conclusion- There is extremely strong evidence that this model with Age, Height and Sex helps to explain the variability in the FEV in this population of youths.**

Explanation of Model Usefulness

**R^2 = 77.46%. The model with Age, Height and Sex helps explain 77.46% of the variability in the FEV.**

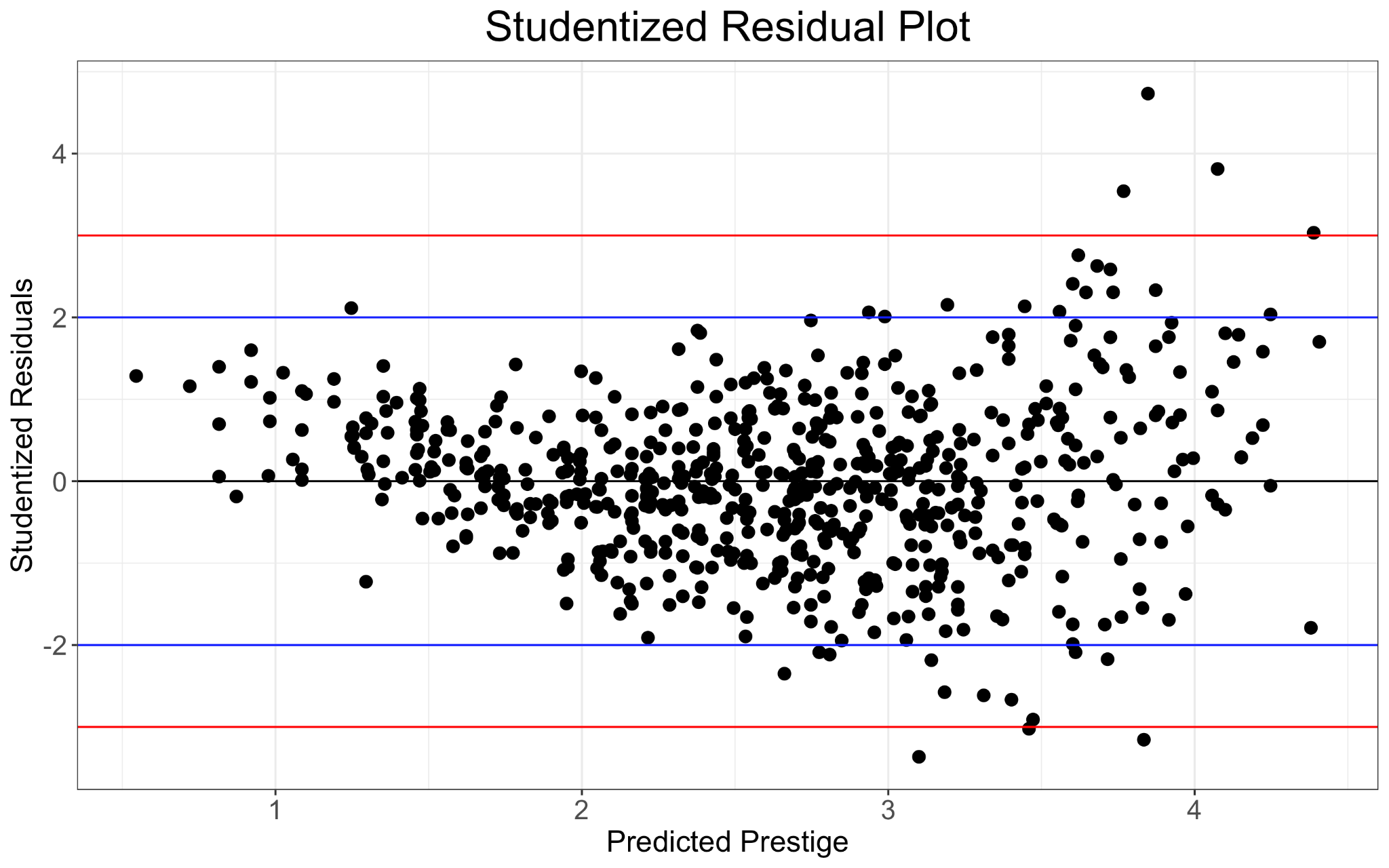
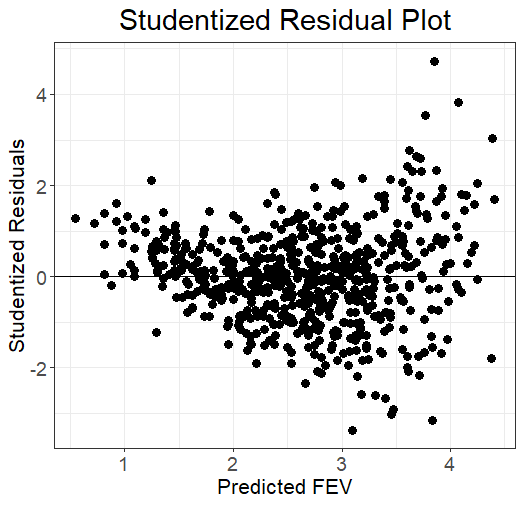
**Assumptions Check**

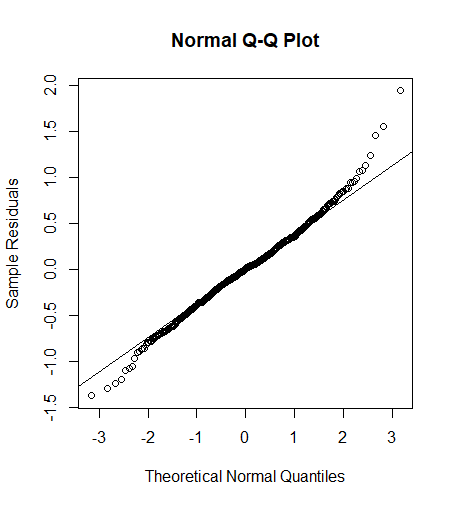


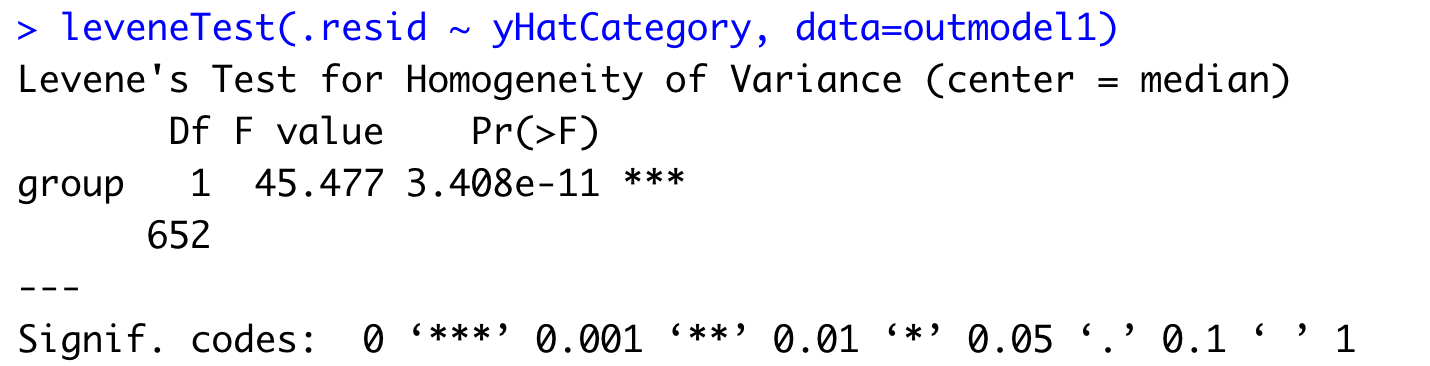
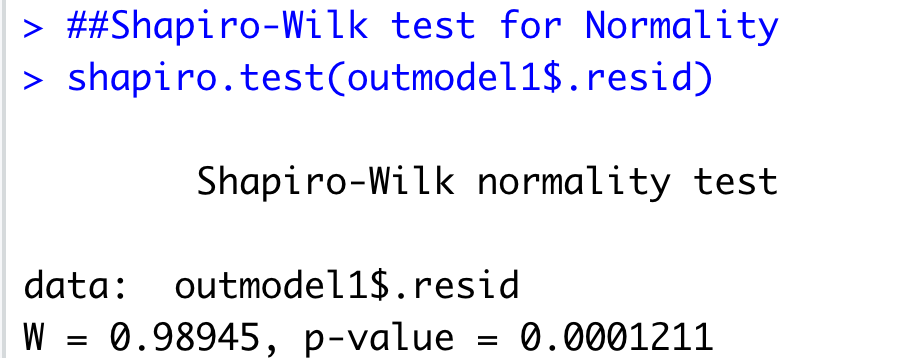
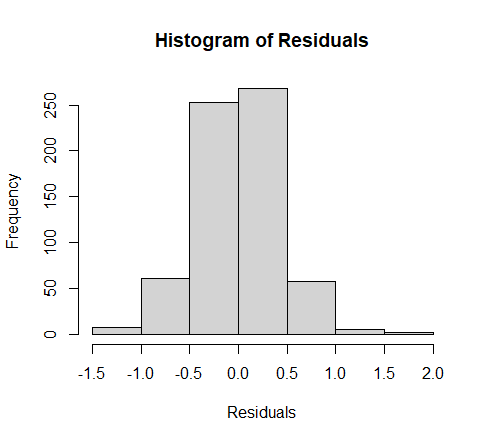
**Independence- our data was collected in a specific timeframe (1970-1979) and was drawn as a random sample from a larger population in a certain area (East Boston). This assumption is met.**

**Linearity- Based on the slight curved pattern in the residual plot and the fanning out of the residuals, we think that non-linearity exists in our data. This assumption is not met.**

**Normality- The histogram shows a normal distribution and the QQ plot shows a strong adherence to the regression line. The Shapiro-Wilks test is within acceptable levels as well. This assumption is met.**

****

****

****

**Summary of Findings**

**Based on our analysis, we could not find a significant link between the smoking status of a youth and its effect on their FEV. The correlation between the 2 variables is rather weak (0.245426) and hypothesis testing shows that little evidence exists for the significance of the smoking status. In fact, our data shows that demographics like height, age and sex have a larger effect on a youth’s FEV over the smoking status. When including and excluding the smoking status in our analysis of our linear models, we received near identical results in both models, indicating a lack of significance on each model.**