Project1\_Zhang

#### James Cutler

#### Project 1

# Problem Description

#### It is known that smoking impairs lung function in adults. There is a statistically significant association between FEV and smoking in adults, with adult smokers having a lower mean FEV. It would be of interest to know whether this association exists in children, since there is less data on children in regards to this question.

# Objective

#### We want to find if there is an association between FEV and smoking status, in order to see if there is evidence that smoking impairs lung function in children, just like in adults.

# Available Data

#### We have data on the FEV, sex, height, age, and smoking status of 654 children, aged 3 to 19, including 65 smokers. There are 318 girls in the sample and 336 boys.

# Analysis Methods

#### We will run multiple linear regression on FEV and the following: smoking status, age, height, and sex, to see what associations are statistically significant.

#### We would run diagnostics with residual plots and QQ plots, but there is not enough space.

# Results (with interpretation)

## Descriptive statistics:

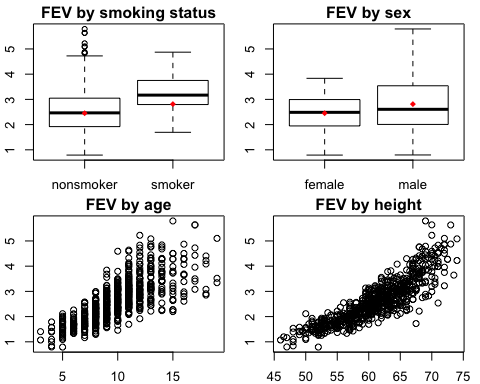
## subjid age fev height   
## Min. : 201 Min. : 3.000 Min. :0.791 Min. :46.00   
## 1st Qu.:15811 1st Qu.: 8.000 1st Qu.:1.981 1st Qu.:57.00   
## Median :36071 Median :10.000 Median :2.547 Median :61.50   
## Mean :37170 Mean : 9.931 Mean :2.637 Mean :61.14   
## 3rd Qu.:53638 3rd Qu.:12.000 3rd Qu.:3.119 3rd Qu.:65.50   
## Max. :90001 Max. :19.000 Max. :5.793 Max. :74.00   
## sex smoke   
## female:318 nonsmoker:589   
## male :336 smoker : 65   
##   
##   
##   
##

## # A tibble: 4 x 4  
## # Groups: sex [2]  
## sex smoke n prop  
## <fct> <fct> <int> <dbl>  
## 1 female nonsmoker 279 0.877   
## 2 female smoker 39 0.123   
## 3 male nonsmoker 310 0.923   
## 4 male smoker 26 0.0774

## # A tibble: 4 x 4  
## # Groups: smoke [2]  
## smoke sex n prop  
## <fct> <fct> <int> <dbl>  
## 1 nonsmoker female 279 0.474  
## 2 nonsmoker male 310 0.526  
## 3 smoker female 39 0.6   
## 4 smoker male 26 0.4

### There are more female smokers than male (39 to 26). There are more male nonsmokers than female (310 to 279).

## Descriptive plots:



### There appears to be an association between smoking status and FEV, based on the corresponding boxplot, above. There is also an apparent association between age/height and FEV, however. Sex and FEV might have a slight association as well.

# Hypothesis tests:

## Multiple regression (full and reduced models):

### Full:

##   
## Call:  
## lm(formula = fev ~ smoke \* sex \* age \* height, data = smoke)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.42585 -0.22543 0.00276 0.22556 1.62815   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.614576 0.844699 -3.095 0.002052 \*\*   
## smokesmoker 28.325085 13.332803 2.124 0.034015 \*   
## sexmale 2.982065 1.102206 2.706 0.007001 \*\*   
## age -0.024674 0.109691 -0.225 0.822099   
## height 0.073483 0.014592 5.036 6.2e-07 \*\*\*  
## smokesmoker:sexmale -34.154716 15.503374 -2.203 0.027948 \*   
## smokesmoker:age -1.929389 1.008290 -1.914 0.056128 .   
## sexmale:age -0.469953 0.135921 -3.458 0.000581 \*\*\*  
## smokesmoker:height -0.423226 0.205652 -2.058 0.039999 \*   
## sexmale:height -0.047472 0.018928 -2.508 0.012386 \*   
## age:height 0.001487 0.001759 0.846 0.398063   
## smokesmoker:sexmale:age 2.071779 1.219240 1.699 0.089762 .   
## smokesmoker:sexmale:height 0.520902 0.236961 2.198 0.028289 \*   
## smokesmoker:age:height 0.028604 0.015565 1.838 0.066568 .   
## sexmale:age:height 0.007616 0.002140 3.558 0.000401 \*\*\*  
## smokesmoker:sexmale:age:height -0.031718 0.018560 -1.709 0.087953 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3827 on 638 degrees of freedom  
## Multiple R-squared: 0.8096, Adjusted R-squared: 0.8051   
## F-statistic: 180.9 on 15 and 638 DF, p-value: < 2.2e-16

### Reduced:

##   
## Call:  
## lm(formula = fev ~ smoke + sex + age + height, data = smoke)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.37656 -0.25033 0.00894 0.25588 1.92047   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.456974 0.222839 -20.001 < 2e-16 \*\*\*  
## smokesmoker -0.087246 0.059254 -1.472 0.141   
## sexmale 0.157103 0.033207 4.731 2.74e-06 \*\*\*  
## age 0.065509 0.009489 6.904 1.21e-11 \*\*\*  
## height 0.104199 0.004758 21.901 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4122 on 649 degrees of freedom  
## Multiple R-squared: 0.7754, Adjusted R-squared: 0.774   
## F-statistic: 560 on 4 and 649 DF, p-value: < 2.2e-16

## 2.5 % 97.5 %  
## (Intercept) -4.89454680 -4.01940100  
## smokesmoker -0.20359813 0.02910535  
## sexmale 0.09189669 0.22230917  
## age 0.04687736 0.08414129  
## height 0.09485705 0.11354180

### The full model indicates that there is are interactions that are highly significant:

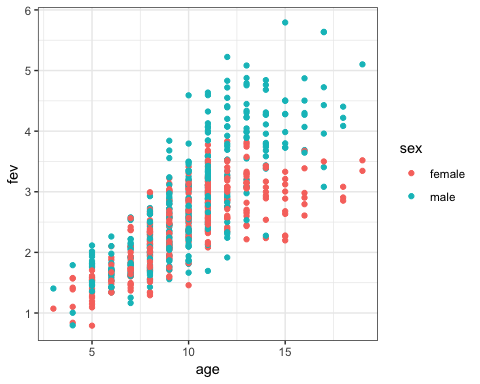
#### - sex and age

#### - sex, age and height

### The reduced model indicates that sex, age, and height all have a highly significant association with FEV, but that smoking status does not.

### Given the interaction between sex and age, it would be interesting to see that visually:

### Interaction between age and sex on FEV:



### Indeed, it looks like the older males have a higher FEV than the older females.

# Conclusions

#### We can conclude that there initially appears to be an association between smoking status and FEV based on the descriptive statistics. In the full MLR model the association between FEV and smoking status does appear to be significant, but in the reduced model the association is not significant. I would like to better understand why.

# Limitations

#### One limitation is that we don’t know how long the smokers had the habit of smoking, and how much they smoked (one pack a day, two?).