Zhang_Project1_Cutler

James Cutler

Project 1

```
## Warning: package 'qwraps2' was built under R version 3.5.2
```

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 will run diagnostics with residual plots and QQ plots and report the results here without showing the figures.

Results (with interpretation)

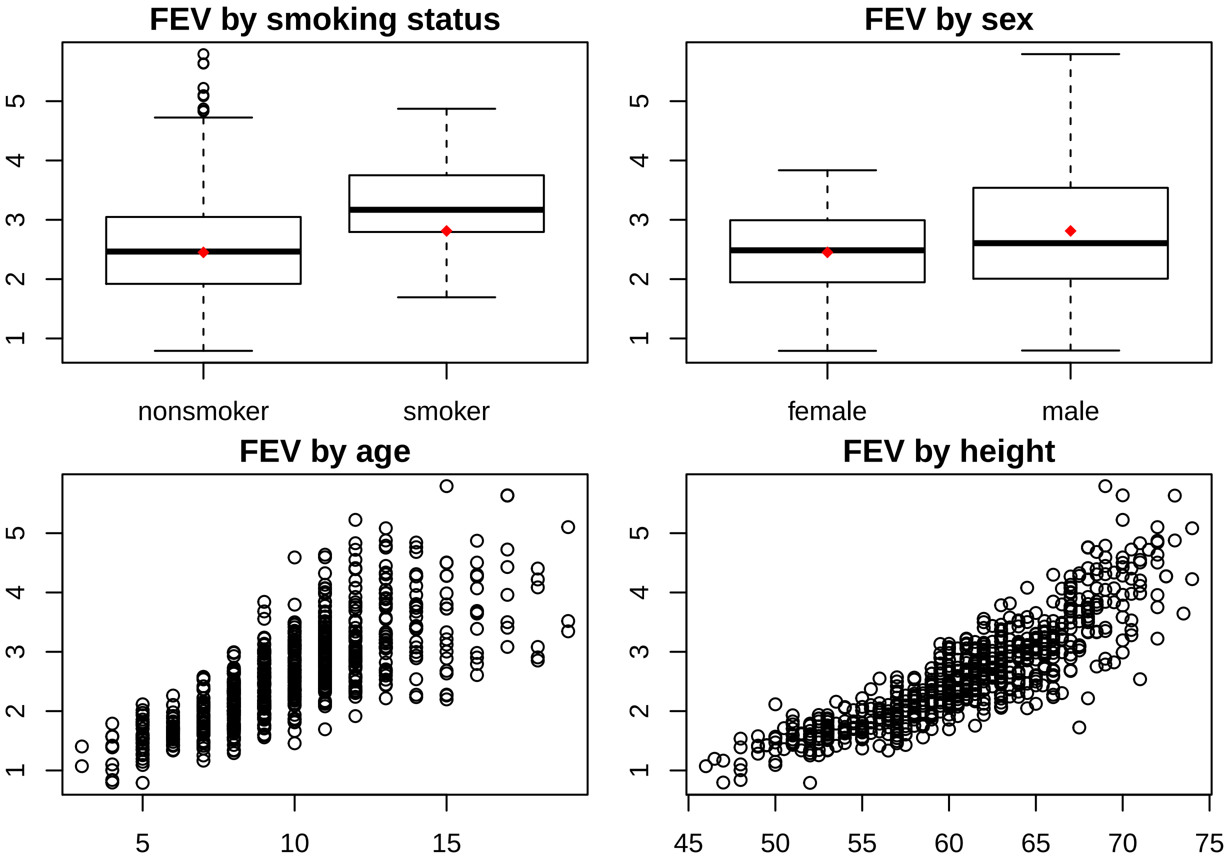
Descriptive statistics:

	smoke (N = 654)
Age	
min	3
max	19
mean (SD)	9.93 ± 2.95

	smoke (N = 654)
FEV	
min	0.791
max	5.793
mean (SD)	2.64 ± 0.87
Height	
min	46
max	74
mean (SD)	61.14 ± 5.70
Sex	
female	318 (48.62%)
male	336 (51.38%)
Smoke	
smokers	65 (9.94%)
non-smokers	589 (90.06%)

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 model:

Table 2: Full model with all interactions

	Dependent variable:
	fev
sexmale	0.140^{***} p = 0.0001
height	0.100^{***} p = 0.000
age	0.075^{***} p = 0.000
smokesmoker	−1.842 p = 0.147
sexmale:smokesmoker	0.146 p = 0.256
height:smokesmoker	0.033^{*} p = 0.099
age:smokesmoker	−0.037 p = 0.138
Constant	-4.262^{***} p = 0.000
Observations	654
R^2	0.778
Adjusted R^2	0.776
Residual Std. Error	0.410 (df = 646)
F Statistic	324.062^{***} (df = 7; 646)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: Full model with sex and smoking interaction removed

	Dependent variable:
	fev
sexmale	0.151^{***} p = 0.00001
height	0.099^{***} p = 0.000
age	0.075^{***} p = 0.000
smokesmoker	-2.553^{**} p = 0.021
height:smokesmoker	0.045^{***} p = 0.010
age:smokesmoker	-0.036 p = 0.146
Constant	-4.252^{***} p = 0.000
Observations	654
R^2	0.778
Adjusted R^2	0.776
Residual Std. Error	0.411 (df = 647)
F Statistic	377.685^{***} (df = 6; 647)
Note:	*p<0.1; **p<0.05; ***p<0.01

The full model (with all two way interactions that involve smoking status) does not reveal any significant interactions. We can now remove the least significant interaction from the above model–the interaction between smoking and sex. The new model looks like this:

We can see, from this new model, that the interaction between height and smoking status is significant (p=.00098). We can also see that every predictor variable by itself has a significant association with FEV.

The residual plots (not printed here) for this second model (with sex*smoking removed) show more or less constant variance of the residuals, while the residuals vs the fitted plot shows a slightly not-so-random pattern, with somewhat of a megaphone pointing upwards (so V-shaped). This could indicate a potential violation of our assumptions, or my concern might be overblown.

Conclusions

FEV appears to be negatively affected by smoking in the final model choice, decreasing by 2.5 liters per second across smoking status, from non-smoking to smoking. FEV increases .15 liters per second with male gender, .099 liters per second with each inch increase in height, and .075 liters per second with each year increase in age. FEV also appears to increase .045 liters per second for each inch in height across smoking status (from non-smoker to smoker). This could make sense given the negative correlation with smoking by itself, and the higher positive correlation with height by itself.

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?).