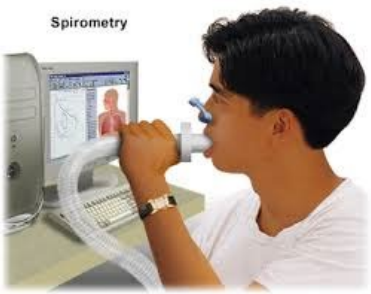
Exploration 4.4: FEV and Smoking

Forced expiratory volume (FEV) is a measure of the strength of a person's lungs—the maximum volume of air (measured in liters) a person can blow out in the first second. Larger values indicate healthier lungs. Researchers in the 1970s were interested in how smoking impacted FEV in adolescents. In particular, does smoking have a particularly strong impact on adolescents whose lung capacity is still developing and maturing?



Step 1: Ask a research question. The researchers wanted to explore whether smokers tend to have lower FEV values than non-smokers.

1. In addition to smoking status, what other variables could be associated with FEV values?

Step 2: Design a study and collect data. The researchers couldn't very well randomly assign individuals to be smokers and non-smokers so they conducted an observational study. Data were collected on 654 adolescent youth in the East Boston area during the middle- to late-1970s. Variables measured included smoking status, age, height, and gender. The data can be found in the file ***FEV***.

Step 3: Explore the data.

Copy the data into the Multiple Variables applet (or use the Select data pull-down menu to select FEV and press Use Data) Drag *FEV* into the **Response** box and *Smoker* into the **Subset By** box. Check the **Show descriptive** box.

2. What do you learn about the difference in average FEV between smokers and non-smokers? Which group tends to have larger FEV values? Is this what you would have predicted? How much variation in FEV is explained by smoking status?

Move the *Smoker* variable into the **Explanatory** variable box and check the **Show Equation** box.

3. Provide an interpretation of the slope coefficient (using effect coding) of the smoking variable in context. (*Hint*: What does this slope coefficient tell you about the difference in group means?)

Scroll down past the pie chart and use the pull-down menu to change from Effect coding to **Indicator coding**.

4. How does the equation change with indicator coding?

Remove the *Smoker* variable and drag the *Age* variable to the Explanatory box. Keep the Show Equation box checked.

5. Describe the association between FEV and age. Does it behave as you would expect? Provide an interpretation of the slope coefficient of age, in context.

Because we expect lung capacity to increase with age, age is potentially a confounding variable in this study. (What else needs to be true for age to be a confounding variable?) So, what we would really like to know about is the association between smoking and FEV after adjusting for age.

Drag the Smoker variable to the Explanatory box, placing it below the Age variable. Your graph of FEV vs. Smoker should now be color coded by age, darker colors (e.g., red, black) indicating younger individuals. Note: Smoker = 1 and Non-smoker = 0 in this graph with indicator coding..

6. What does this graph tell us about the association between age and smoking status? In particular, are younger people more or less likely to be smokers? How can you tell? (You can also check the Show 2-variable graphs box.)

7. If we were to adjust the FEV values for age, lowering the FEV of the older individuals and raising the FEV of the younger individuals to "even the playing field," how will that change the mean FEV of the non-smokers (Smoker = 0)? How will that change the mean FEV of the smokers (Smoker = 1)?

Check the **Adjust values** box to check your answer. (The original line is grey and the purple line shows the age-adjusted regression line predicting FEV from smoking status.)

8. What are the values of the adjusted and unadjusted slopes between smoking status and FEV? Why are they different? Using the displayed prediction equation, what are the age-adjusted means for smokers and non-smokers (*Hint*: Remember you have indicator coding selected).

Check the **Statistical model** box (with Indicator coding).

9. Use the Statistical model output to write out the single prediction equation for this two-variable model. Based on the pie chart or ANOVA table, how much total variation in FEV is explained by this two-variable model? What is the R^2 for this two-variable model?

[Type here]

From this model we learn that if we compare two individuals of the same age, we predict smokers to have a lower FEV than non-smokers. But, does this really tell the whole story? This model makes an assumption—it assumes that the impact of smoking on FEV is the same for older people and younger people. This means that we are assuming there is no interaction between age and smoking status on FEV. How can we check this assumption?

Definition: A ***statistical interaction*** occurs between two variables when the effect (or association) of one explanatory variable on the response variable changes based on the value of the other variable. An interaction between a quantitative variable and a categorical variable occurs when the slopes of the regression lines are different for different categories. Equivalently, this means that the distance between the lines changes as the quantitative variable changes.

Drag the Smoker variable to the **Subset By** box. The applet will now fit two separate lines, one for the smokers and one for the non-smokers.

10. What is the equation of the line for non-smokers? For smokers?

    Non-smokers:

    Smokers:

11. What characteristic of these lines suggests that is an interaction between age and smoking status?

12. For non-smokers, what is the predicted FEV at age 8? What is the predicted FEV for smokers at age 8? What about at age 12? Age 16? Fill in the table below, including the difference in the predicted FEV for non-smokers and smokers of each age.

	Predicted FEV		
	Age 8	**Age 12**	**Age 16**
Non-smokers			
Smokers			
Difference = non-smokers – smokers			

13. What does the "difference in the differences" indicate about how the relationship between FEV and age differs between smokers and non-smokers? Why?

14. Write a sentence summarizing the nature of the statistical interaction between age and smoking on FEV so that a non-statistician would understand. For example, how does age modify the association between FEV and whether someone is a smoker?

[Type here]

To see where these separate equations come from, imagine that we multiply the value of the quantitative explanatory variable and the value of the indicator variable for the binary categorical variable for each row in the data table. In this study, that means for each study participant, we multiply the smoking indicator variable (1 = smoker, 0 = non-smoker) by the age variable. This creates an interaction term, which we typically denote as smoker × age. An example of this multiplication is shown in **Figure 1**.

FIGURE 1 Age × Smoker_Ind column.

	FEV	Height	Age	Gender	Smoker	Smoker_Ind	Age*Smoker_Ind
1	4.506	71	15	Male	yes	1	15
2	2.884	69	11	Male	no	0	0
3	2.328	64	10	Male	no	0	0
4	1.708	57	9	Female	no	0	0

Most software packages create the interaction term for you if you specify you want to include the interaction in the model.

Key Idea: To model the interaction, we create and include a new variable that is the product of the quantitative variable and the categorical variable.

Return to the applet and obtain the output for the **Statistical model** with the interaction using **Indicator coding**.

15. Using indicator coding, write out the full statistical model equation to predict FEV:

predicted FEV = _____ + _____ *age* + _____ *smoker (1 = yes)* + _____ *age x smoker (1 = yes)*

16. Provide an interpretation of the intercept coefficient in this model. (*Hint*: Which category, smokers or non-smokers, is the reference category? How does the equation simplify for those in the reference category?)

17. Using the full statistical model equation, verify the two separate equations for smokers and non-smokers shown in the applet. (*Hint*: Because non-smokers are the reference group, what is the equation when the Smoker variable is set to 0? What is the equation when the Smoker variable is set to 1? Keep in mind that the last term is the product of age and the smoking indicator variable.)

18. According to the full prediction equation, among the smokers, how much of an increase in FEV is associated with a one-year age increase?

[Type here]

19. According to the prediction equation, among the non-smokers, how much of an increase in FEV is associated with a one-year age increase? How does this compare to the prediction for smokers?

20. Notice that the coefficient of the interaction term is negative. How does this change the line for smokers compared to non-smokers?

One other important note: Because smoking and age were associated, when we fit the two variable model with no interaction the association between smoking and FEV changed when we adjusted for age. The association between smoking and age means there is variation in FEV that cannot be uniquely explained by either the smoking variable or the age variable. This variation is called covariation. When we add an interaction between smoking and age there is even more variation in FEV that we cannot uniquely assign to either of the explanatory variables or to the interaction term. As you might expect, this is not an issue in well-designed experiments where there is no association between factors or blocks.

Step 4: Draw inferences beyond the data.

21. Is the interaction between age and smoking status statistically significant? (Be sure to state hypotheses and cite both a standardized statistic and a p-value.)

Step 5: Formulate conclusions.

22. Report a rough 95% confidence interval for the interaction coefficient in the population. To what population are you willing to generalize these results? Can you draw a cause and effect conclusion about smoking and the rate of growth of lung capacity from this study?

Step 6: Look back and ahead.

Check the **Show residuals** box.

23. Do you consider the validity conditions met for this study? Justify your answer.

24. What suggestions do you have for a follow-up study?