CarAcceleration

Kevin Cummiskey

November 12, 2019

Smoking and Lung Function

Today, we are going to investigate the relationship betweeen smoking and lung function in teenagers. In the early 1980s, researchers recruited teenagers in South Boston to participate in a study on the health effects of smoking. The data set `teens` contains the age (years), height (inches), gender, forced expiratory volume - FEV (liters), and whether or not the subject smoked for 654 subjects in the study. FEV is the volume of air a person can exhale in a period of time and is a measure of lung function.

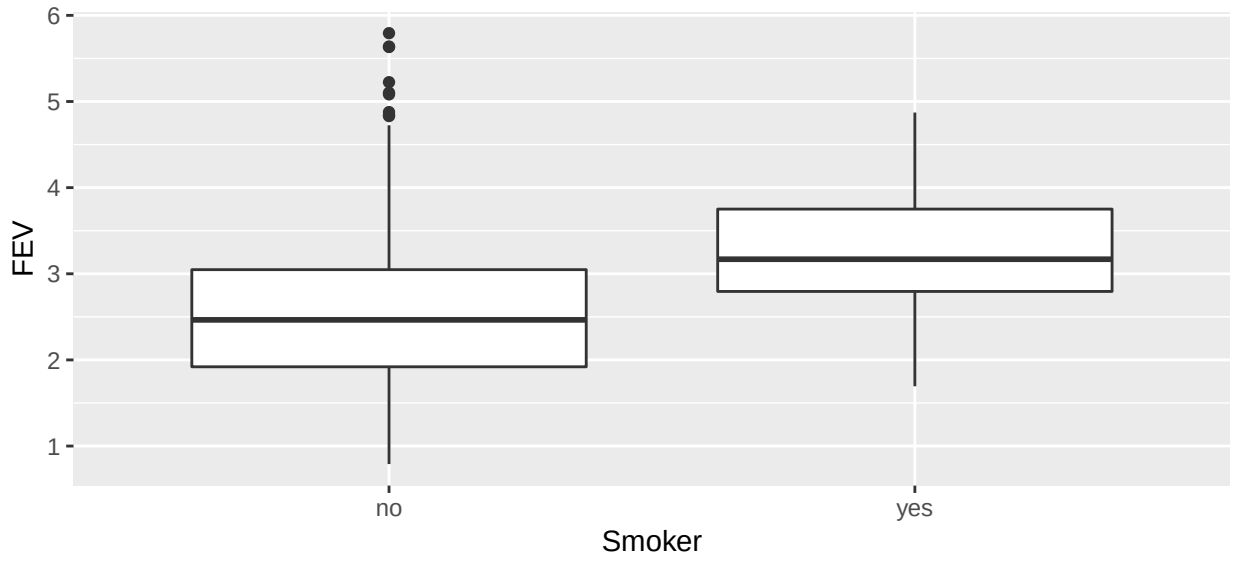
```
library(tidyverse)
teens = read.table(file = "http://www.isi-stats.com/isi2/data/FEV.txt",
                   header = T)
teens$Smoker = factor(teens$Smoker) #convert quantitative variable to categorical
head(teens)
```

```
##   Age   FEV Height Gender Smoker
## 1  15 4.506     71   Male    yes
## 2  11 2.884     69   Male     no
## 3  10 2.328     64   Male     no
## 4   9 1.708     57 Female     no
## 5  14 3.381     63   Male     no
## 6  11 2.170     58 Female     no
```

What would we expect the relationship to be between smoking and FEV?

Let's see what the data says.

```
teens %>% ggplot(aes(x = Smoker, y = FEV)) + geom_boxplot()
```



What does the data say?

Here's a model:

$$FEV_i = \beta_0 + \beta_1 Smoker_i + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

where FEV_i is the forced expiratory volume of subject i and $Smoker_i$ is whether or not patient i is a smoker. How do we interpret β_0, β_1?

Let's fit the model. (Based on the boxplot above, what are reasonable estimates for β_1 and β_0?)

```
model_smoker = lm(FEV ~ Smoker, data = teens)
summary(model_smoker)
```

```
##
## Call:
## lm(formula = FEV ~ Smoker, data = teens)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7751 -0.6339 -0.1021  0.4804  3.2269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.56614    0.03466  74.037  < 2e-16 ***
## Smokeryes    0.71072    0.10994   6.464 1.99e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8412 on 652 degrees of freedom
## Multiple R-squared:  0.06023,    Adjusted R-squared:  0.05879
## F-statistic: 41.79 on 1 and 652 DF,  p-value: 1.993e-10
```

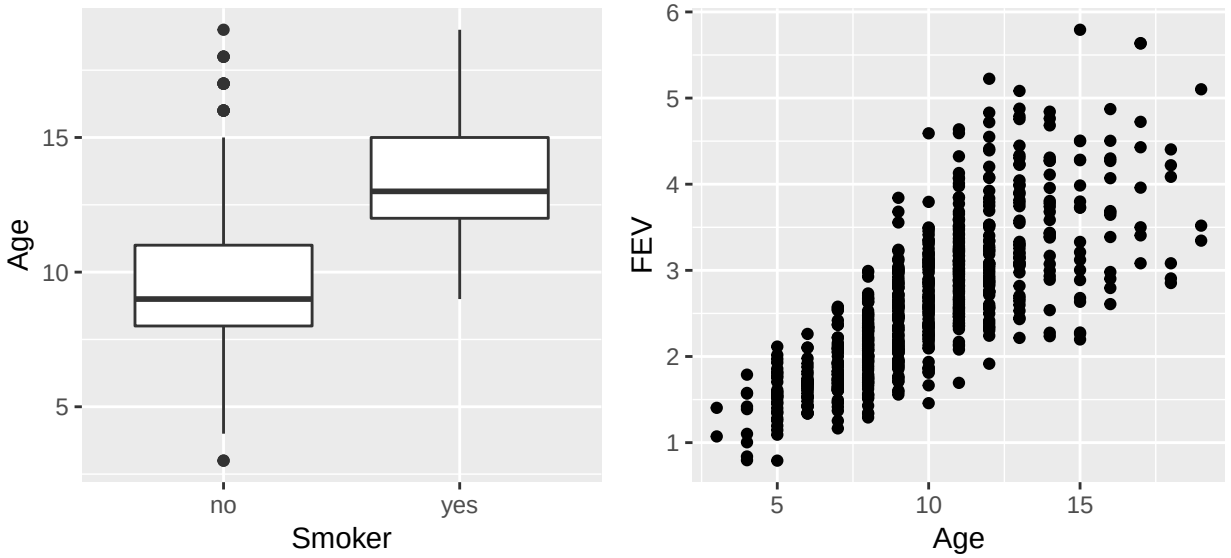
What would you conclude from this model? Comment on the size, direction, and strength of the effect of smoking on FEV. In addition, comment on the ability of the model to predict a subject's FEV.

Based on this analysis, would you conclude smoking is good for lung function? Explain.

In observational studies, if the goal is to assess the effect of one variable (smoking) on another variable (FEV), then we need to consider other variables associated with these two variables. List two other variables that may be associated with smoking and FEV.

Let's look at age and its relationship to smoking and FEV.

```
library(gridExtra)
p1 = teens %>% ggplot(aes(x = Smoker, y = Age)) + geom_boxplot()
p2 = teens %>% ggplot(aes(x = Age, y = FEV)) + geom_point()
grid.arrange(p1,p2, ncol = 2)
```



Is age associated with smoking and FEV? How do these associations explain the beneficial effect of smoking we observed above?

Consider the following model:

$$FEV_i = \alpha_0 + \alpha_1 Smoker_i + \alpha_2 Age_i + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

How do we interpret α_0, α_1, and α_2?

How would we expect α_1 to compare to β_1?

Let's fit the model.

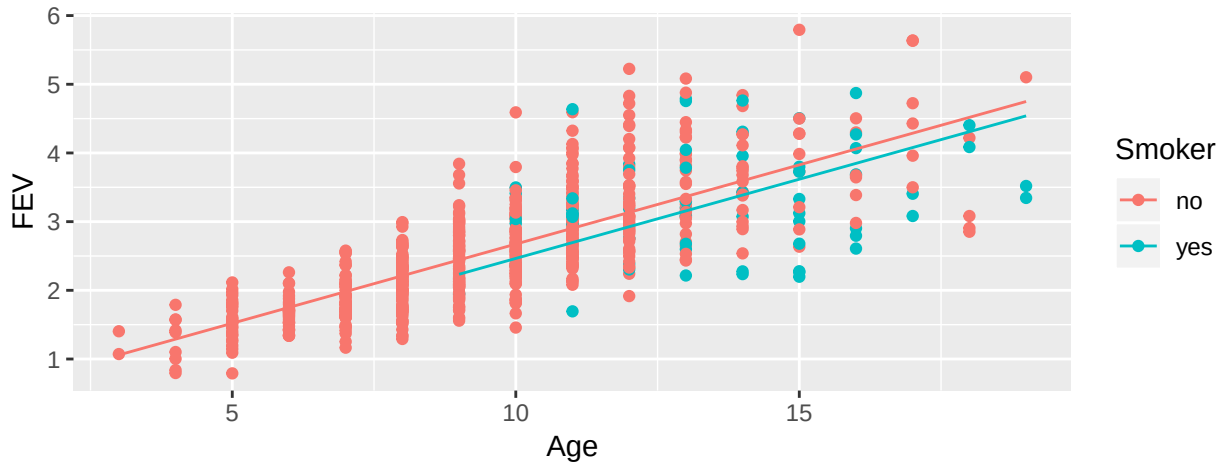
```r
model_smoker_age = lm(FEV ~ Smoker + Age, data = teens)
summary(model_smoker_age)
```

```
##
## Call:
## lm(formula = FEV ~ Smoker + Age, data = teens)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6653 -0.3564 -0.0508  0.3494  2.0894
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.367373   0.081436   4.511 7.65e-06 ***
## Smokeryes   -0.208995   0.080745  -2.588  0.00986 **
## Age          0.230605   0.008184  28.176  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5651 on 651 degrees of freedom
## Multiple R-squared:  0.5766, Adjusted R-squared:  0.5753
## F-statistic: 443.3 on 2 and 651 DF,  p-value: < 2.2e-16
```

What would you conclude from this model? Comment on the size, direction, and strength of the effect of smoking on FEV. In addition, comment on the ability of the model to predict a subject's FEV.

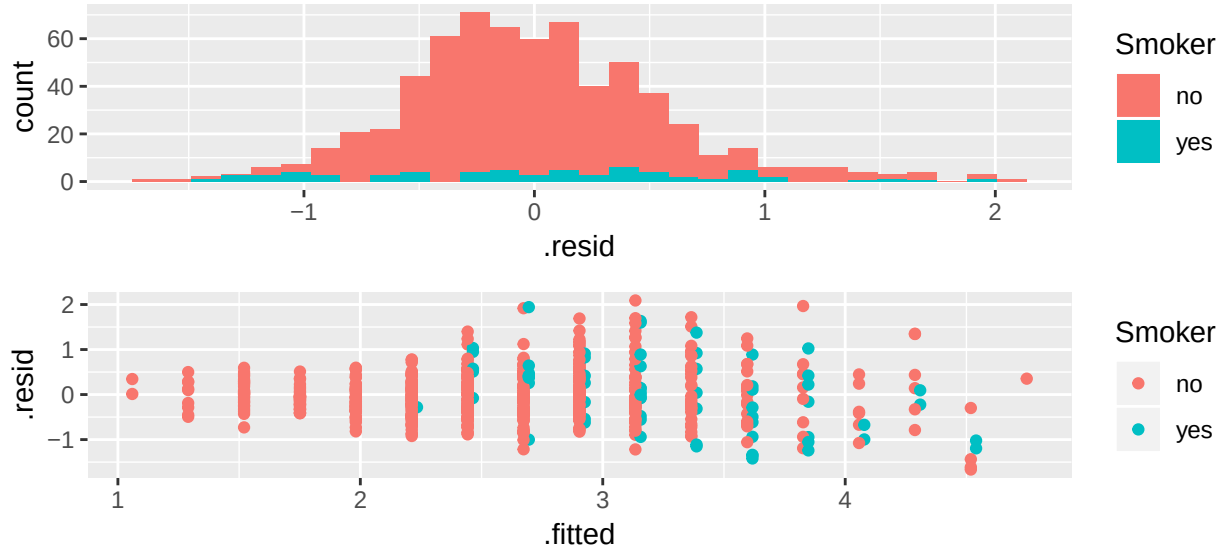
What assumption does this model make about the relationship between smoking and FEV at each age? See the plot below.

```r
model_smoker_age %>%
  fortify() %>%
  ggplot(aes(x = Age, y = FEV, color = Smoker)) +
  geom_point() +
  geom_line(aes(y = .fitted, color = Smoker))
```



Lastly, let's check some of our model assumptions. Discuss what the plots below tell us about each assumption.

```
p3 = model_smoker_age %>%
  fortify() %>%
  ggplot(aes(x = .resid, fill = Smoker)) +
  geom_histogram()
p4 = model_smoker_age %>%
  fortify() %>%
  ggplot(aes(x = .fitted, y = .resid, color = Smoker)) +
  geom_point()
grid.arrange(p3,p4)
```



Write a paragraph summarizing the results of this analysis for a nonstatiscian friend.