IHME Researcher Assessment Responses

Troy Edwards

2023-03-01

Question 1.1

```
smoking_data <- read.csv("./data.csv")
```

Question 1.2

```
printf <- function(...) {
  print(sprintf(...))
}

male_data <- smoking_data[smoking_data$sex == "male",]
female_data <- smoking_data[smoking_data$sex == "female",]

printf(
  "Smoking prevalence (male): mean = %.3f, sd = %.3f",
  mean(male_data$smoke),
  sd(male_data$smoke)
)
```

```
## [1] "Smoking prevalence (male): mean = 0.318, sd = 0.148"
```

```
printf(
  "Smoking prevalence (female): mean = %.3f, sd = %.3f",
  mean(female_data$smoke),
  sd(female_data$smoke)
)
```

```
## [1] "Smoking prevalence (female): mean = 0.109, sd = 0.099"
```

```
printf(
  "Overweight prevalence (male): mean = %.3f, sd = %.3f",
  mean(male_data$overweight),
  sd(male_data$overweight)
)
```

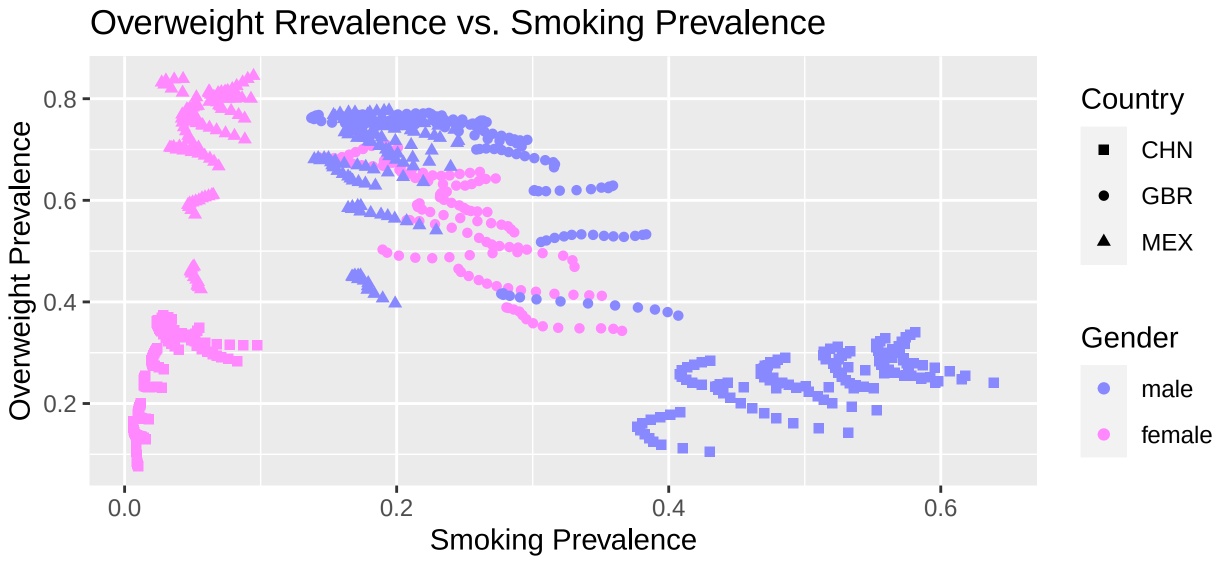
```
## [1] "Overweight prevalence (male): mean = 0.532, sd = 0.221"
```

```
printf(
  "Overweight prevalence (female): mean = %.3f, sd = %.3f",
  mean(female_data$overweight),
  sd(female_data$overweight)
)
```

```
## [1] "Overweight prevalence (female): mean = 0.517, sd = 0.218"
```

Question 1.3

```
ggplot(smoking_data, aes(smoke, overweight)) +
  geom_point(aes(color = smoking_data$sex, shape = smoking_data$location)) +
  scale_color_manual(
    name = "Gender",
    values = c("male" = "#8888ffff", "female" = "#ff88ffff")
  ) +
  scale_shape_manual(
    name = "Country",
    values = c("CHN" = 15, "GBR" = 16, "MEX" = 17)
  ) +
  labs(
    x = "Smoking Prevalence",
    y = "Overweight Prevalence",
    title = "Overweight Rrevalence vs. Smoking Prevalence"
  )
```



Question 1.4

```
printf("r = %.3f", cor(smoking_data$smoke, smoking_data$overweight))
```

```
## [1] "r = -0.305"
```

Question 1.5

```
summary(
  lm(
    overweight ~ smoke,
    data = smoking_data
  )
)
```

```
##
## Call:
## lm(formula = overweight ~ smoke, data = smoking_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53104 -0.15368  0.05585  0.19129  0.27166
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.61214    0.01231  49.745   <2e-16 ***
## smoke       -0.40945    0.04577  -8.946   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2092 on 778 degrees of freedom
## Multiple R-squared:  0.09328,    Adjusted R-squared:  0.09211
## F-statistic: 80.03 on 1 and 778 DF,  p-value: < 2.2e-16
```

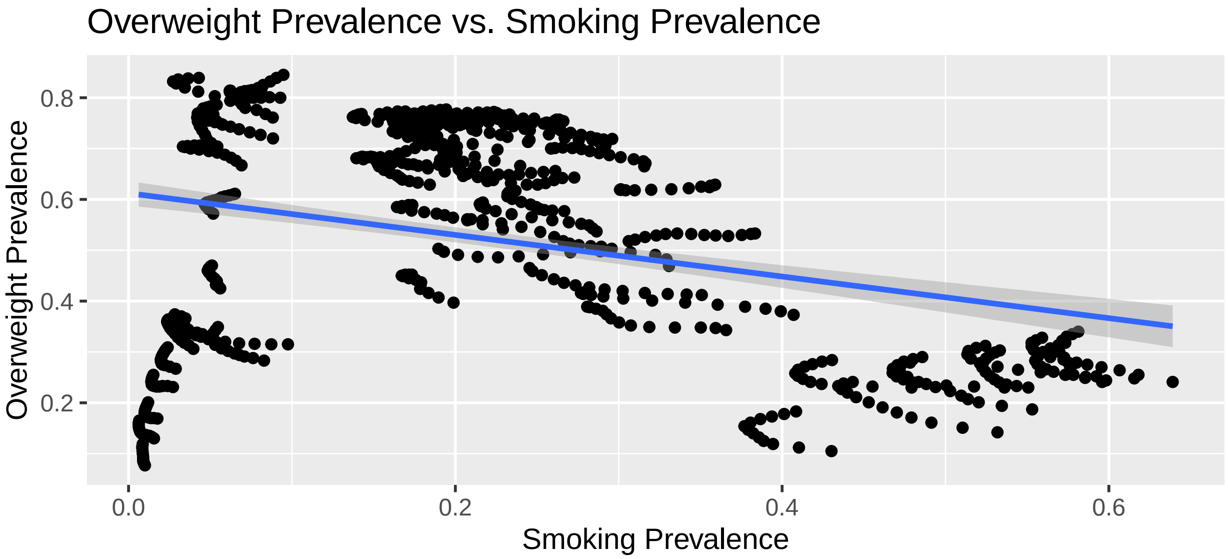
Question 1.6

The r^2 value of -0.305 indicates that there is a moderate negative linear correlation between smoking prevalence and overweight prevalence.

Question 1.7

$$\widehat{overweight} = 0.61214 - 0.40945(smoking)$$

```
ggplot(smoking_data, aes(smoke, overweight)) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(
    x = "Smoking Prevalence",
    y = "Overweight Prevalence",
    title = "Overweight Prevalence vs. Smoking Prevalence"
  )
```

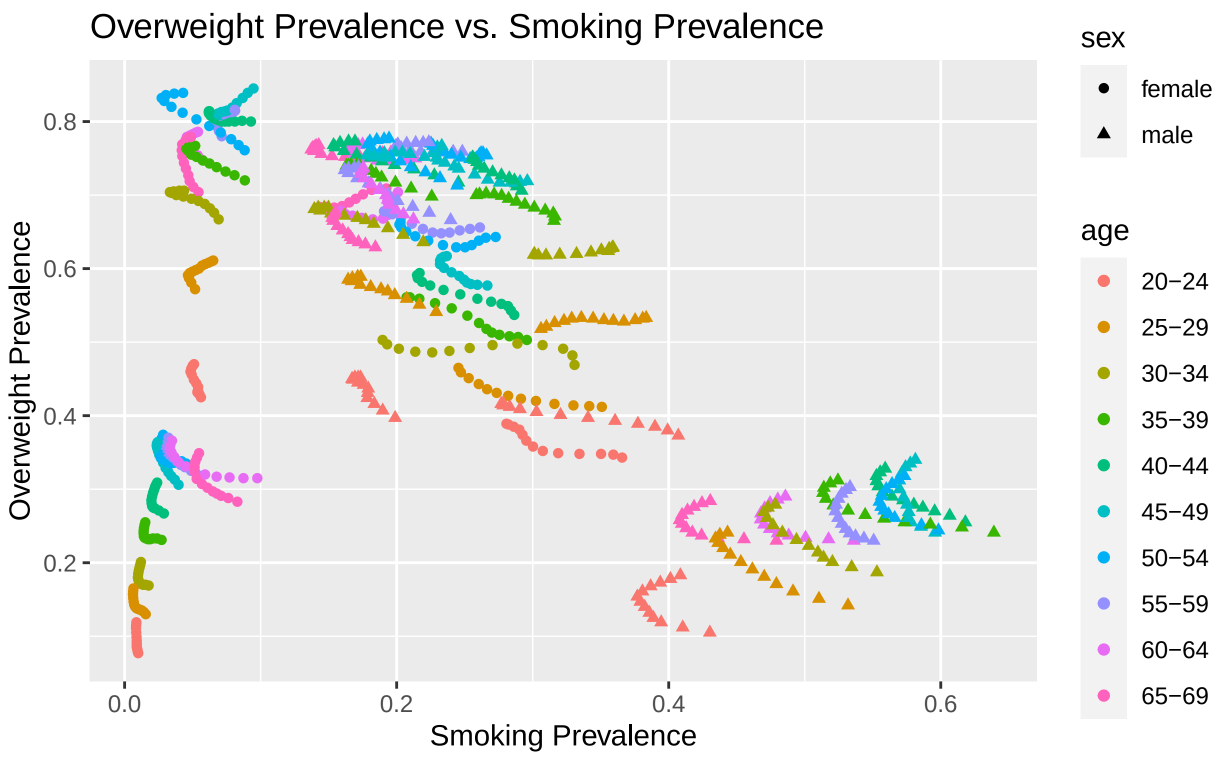


Question 1.8

No, we cannot conclude that reduction in smoking causes overweight because you cannot conclude causation from an observational study.

Question 1.9

```
ggplot(smoking_data, aes(smoke, overweight)) +
  geom_point(aes(color = age, shape = sex)) +
  labs(
    x = "Smoking Prevalence",
    y = "Overweight Prevalence",
    title = "Overweight Prevalence vs. Smoking Prevalence"
  )
```



One interesting thing is that each little "curve" represented by a group of points is comprised entirely of 1 age group, 1 gender, and 1 country.

Question 2

1. In general, smoking prevalence appears to have decreased in most countries.
2. For women, countries with less prevalence in 1980 appeared to have more extreme rates of change, both up and down.
3. Men tend to smoke much more than women.

Question 3

I would try to come up with a model to predict decrease in smoking prevalence.