DATA 621 - BLOG 5 - MULITPLE LINEAR REGRESSION DEMONSTRATION

Enid Roman

2023-12-12

STEPS FOR MULTIPLE LINEAR REGRESSION

Multiple linear regression uses two or more independent variables

The heart dataset contains observations on the percentage of people biking to work each day, the percentage of people smoking, and the percentage of people with heart disease in an imaginary sample of 500 towns.

Note: R Code is in Appendix section.

LOAD DATA

SUMMARY OF DATA

```
##        X             biking          smoking        heart.disease    
##  Min.   :  1.0   Min.   : 1.119   Min.   : 0.5259   Min.   : 0.5519  
##  1st Qu.:125.2   1st Qu.:20.205   1st Qu.: 8.2798   1st Qu.: 6.5137  
##  Median :249.5   Median :35.824   Median :15.8146   Median :10.3853  
##  Mean   :249.5   Mean   :37.788   Mean   :15.4350   Mean   :10.1745  
##  3rd Qu.:373.8   3rd Qu.:57.853   3rd Qu.:22.5689   3rd Qu.:13.7240  
##  Max.   :498.0   Max.   :74.907   Max.   :29.9467   Max.   :20.4535  
```

Executing this code yields a numerical summary of the data for the quantitative independent variables (smoking and biking) and the dependent variable (heart disease). The summary includes key statistics such as minimum, median, mean, and maximum values.

MEETING ASSUMPTION

In R, we can assess if our data meet the four key assumptions for mutiple linear regression.

INDEPENDENCE OF OBSERVATION (no autocorrelation)

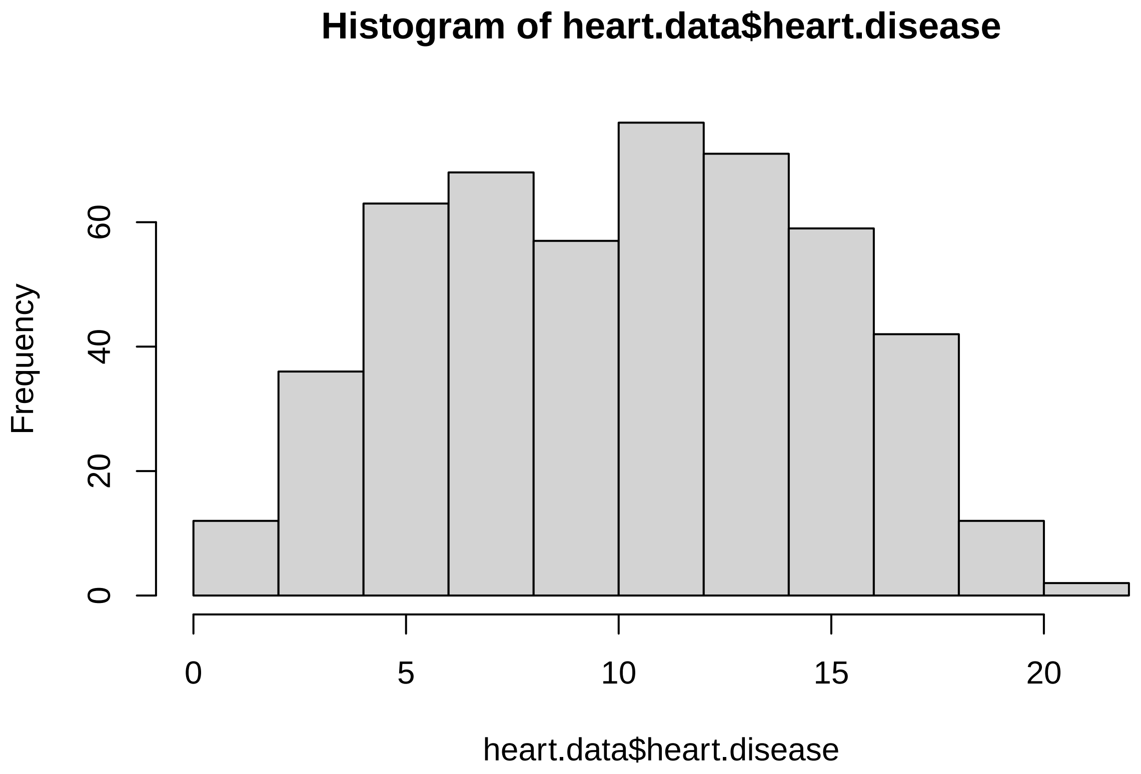
Use the cor() function to assess the correlation between your independent variables and ensure that they are not excessively correlated.

```
## [1] 0.01513618
```

When running this code, the output is 0.015, indicating a small correlation (1.5%) between biking and smoking. This suggests that both variables can be included in the model without significant multicollinearity issues.

NORMALITY

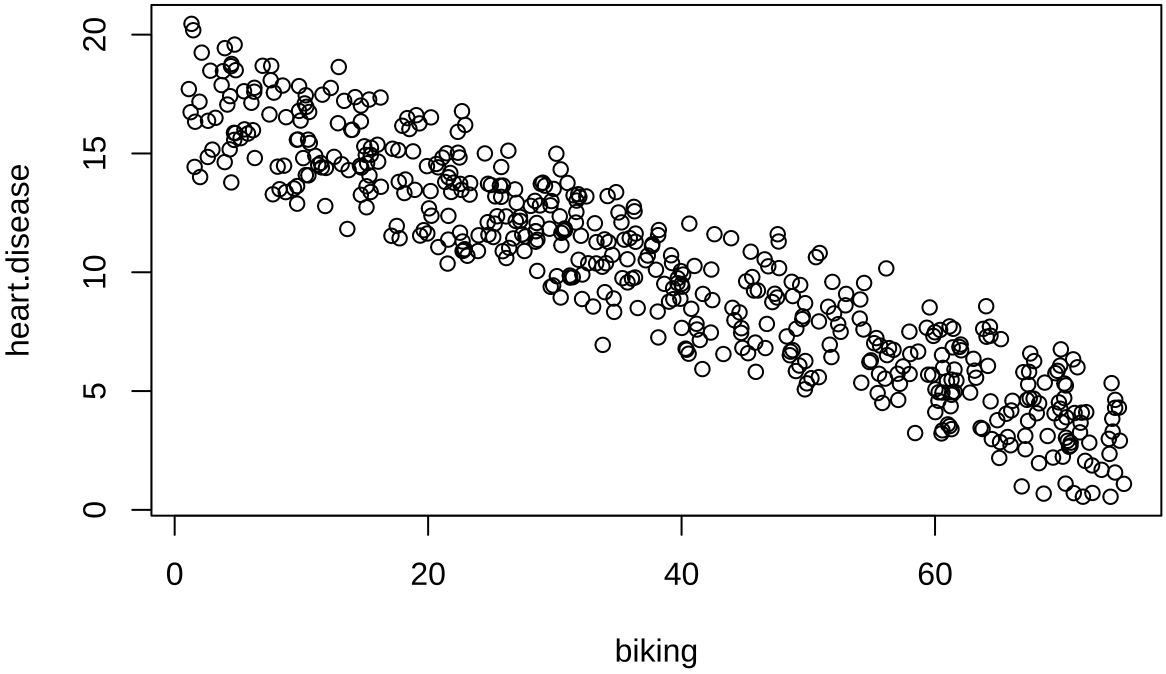
To check whether the dependent variable follows a normal distribution, use the hist() function.

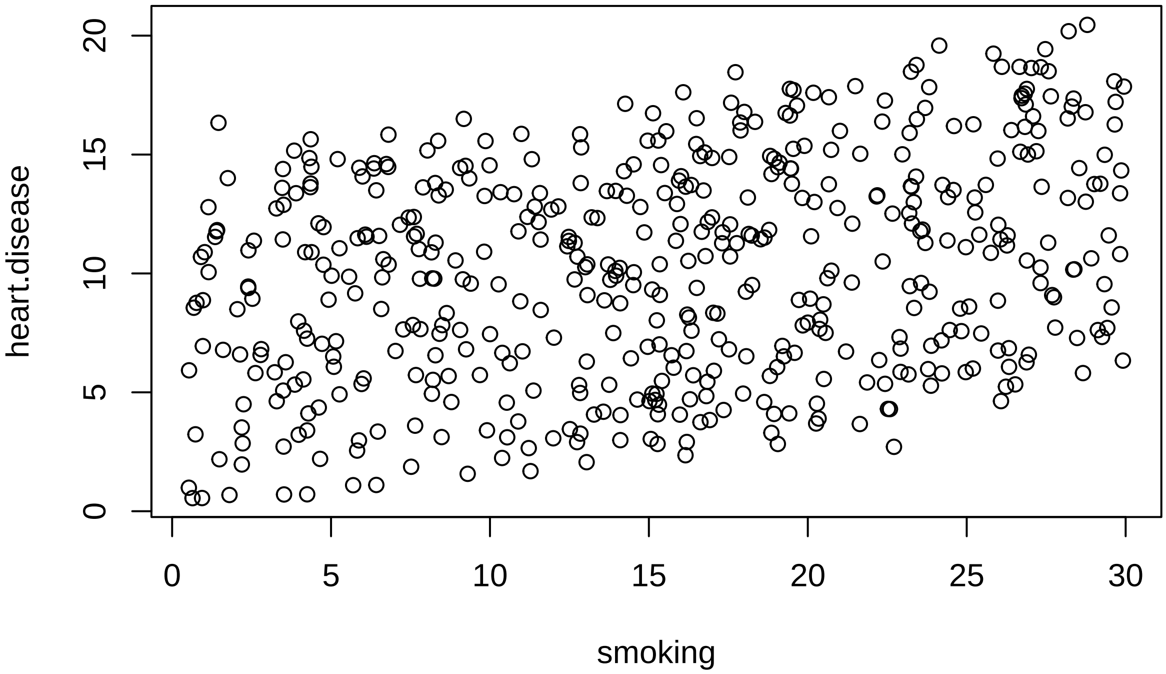


The roughly bell-shaped distribution of observations allows us to proceed with linear regression analysis.

LINEARITY

We can verify this by examining two scatterplots: one for biking and heart disease, and another for smoking and heart disease.





While the relationship between smoking and heart disease is somewhat less evident, it still appears to be linear, allowing us to proceed with linear regression.

HOMOSCEDASTICITY

We will check this after we make the model.

ANALYSIS

MUTIPLE REGRESSION: BIKING, SMOKING, AND HEART DISEASE

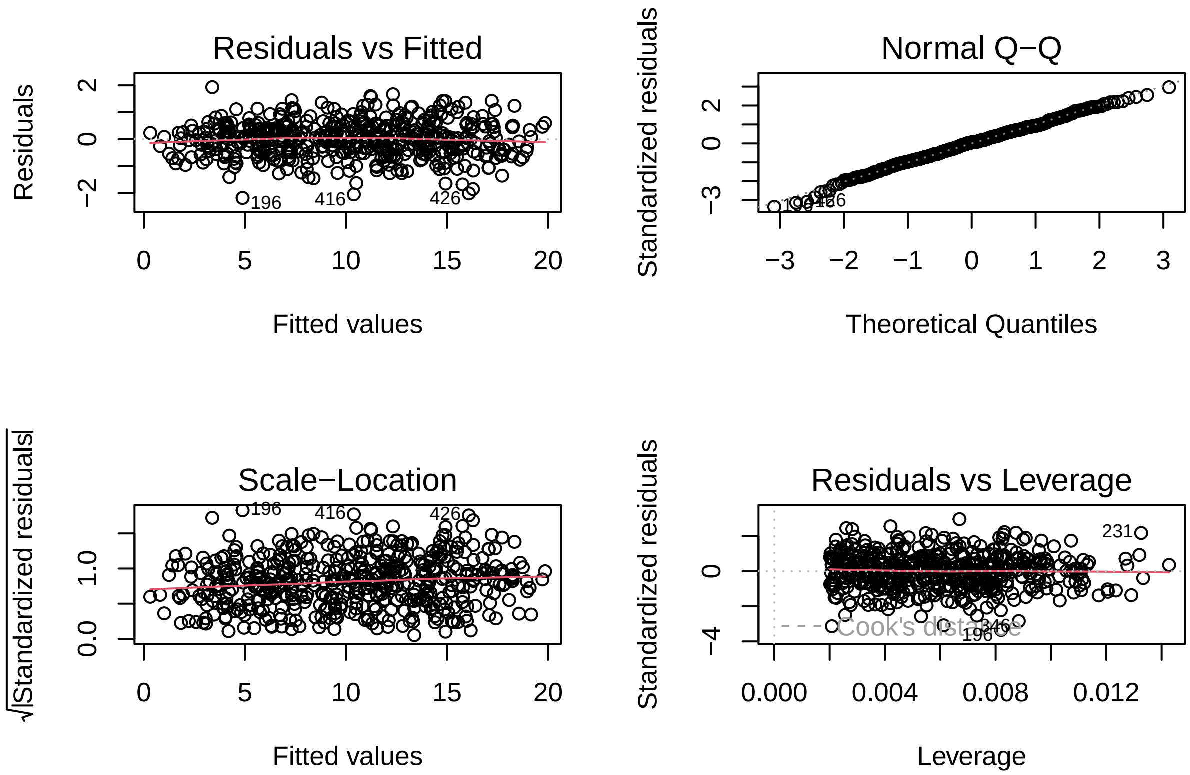
To test the linear relationship between biking to work, smoking, and heart disease in our hypothetical survey of 500 towns, we fitted a linear model with heart disease as the dependent variable and biking and smoking as the independent variables. The rates of biking to work ranged between 1% and 75%, rates of smoking between 0.5% and 30%, and rates of heart disease between 0.5% and 20.5%.

```
##
## Call:
## lm(formula = heart.disease ~ biking + smoking, data = heart.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1789 -0.4463  0.0362  0.4422  1.9331
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.984658   0.080137  186.99   <2e-16 ***
## biking      -0.200133   0.001366 -146.53   <2e-16 ***
## smoking      0.178334   0.003539   50.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.654 on 495 degrees of freedom
## Multiple R-squared:  0.9796, Adjusted R-squared:  0.9795
## F-statistic: 1.19e+04 on 2 and 495 DF,  p-value: < 2.2e-16
```

The estimated effects of biking on heart disease and smoking on heart disease are -0.2 and 0.178, respectively. This indicates that for every 1% increase in biking to work, there is a correlated 0.2% decrease in the incidence of heart disease. Conversely, for every 1% increase in smoking, there is a 0.178% increase in the rate of heart disease. The standard errors for these regression coefficients are very small, and the t statistics are very large (-147 and 50.4, respectively). The p values suggest that there is almost zero probability that these effects are due to chance.

HOMOSCEDASTICITY

We should check that our model is a good fit for the data and that there is not large variation in the model error by running this code.



The residuals show no bias, indicating that our model fits the assumption of homoscedasticity.

GRAPH VISUALIZATION

Visualizing multiple regression can be challenging due to having two predictors. While one option is to plot a plane, these can be hard to interpret and are not commonly published.

As an alternative method, we will visualize the relationship between biking and heart disease at different levels of smoking. In this example, smoking will be treated as a factor with three levels, solely for the purpose of displaying the relationships in our data.

There are 7 steps to follow.

CREATE A NEW DATAFRAME WITH THE INFORMATION NEEDED TO PLOT THE MODEL

The expand.grid() function is utilized to generate a dataframe with specified parameters. This involves creating a sequence spanning from the lowest to the highest observed biking data values and selecting the minimum, mean, and maximum values of smoking to establish three levels of smoking for predicting rates of heart disease.

Running this code does not produce any new output in the console, but a new data frame should appear in the Environment tab for viewing.

PREDICT THE VALUES OF HEART DISEASE BASED ON YOUR LINEAR MODEL

Next, we save our predicted ‘y’ values as a new column in the dataset we just created.

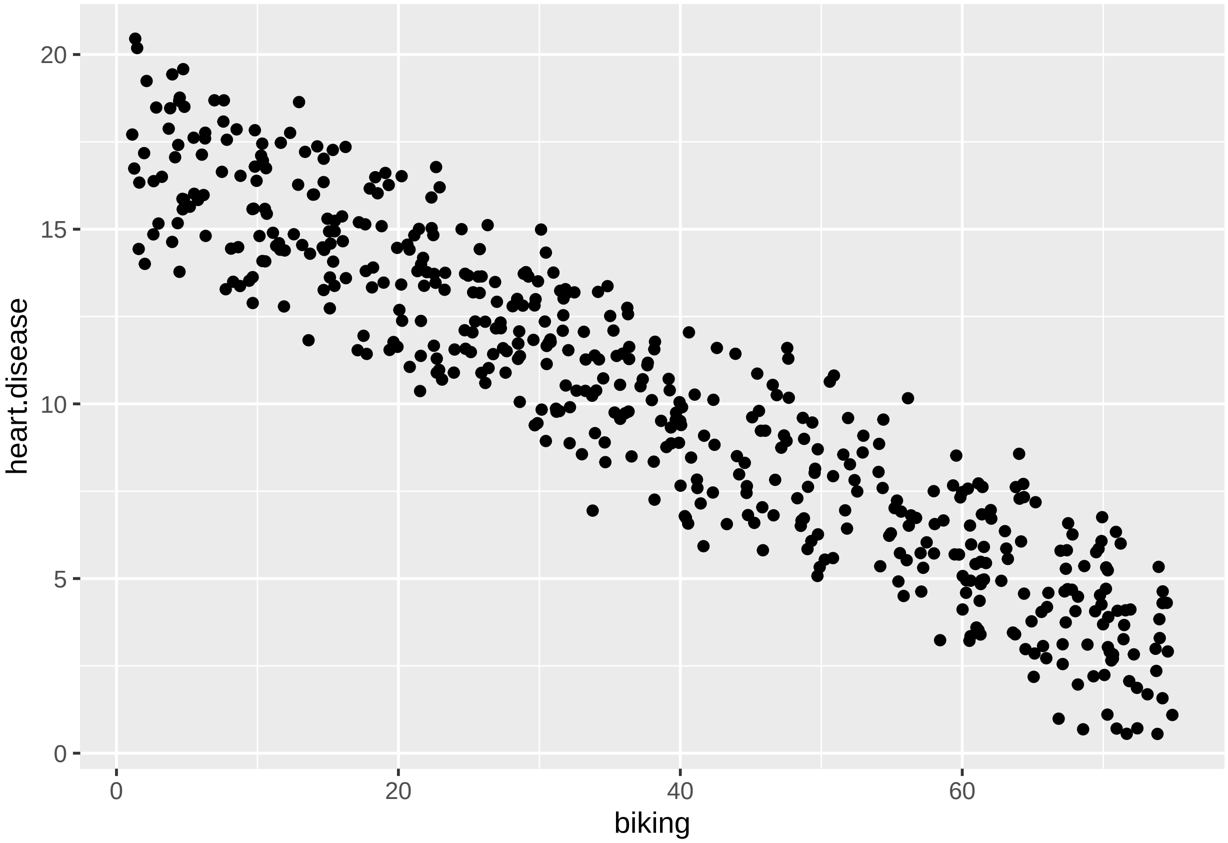
ROUND THE SMOKING NUMBERS TO TWO DECIMALS

This will make the legend easier to read later on.

CHANGE THE SMOKING VARIABLE INTO A FACTOR

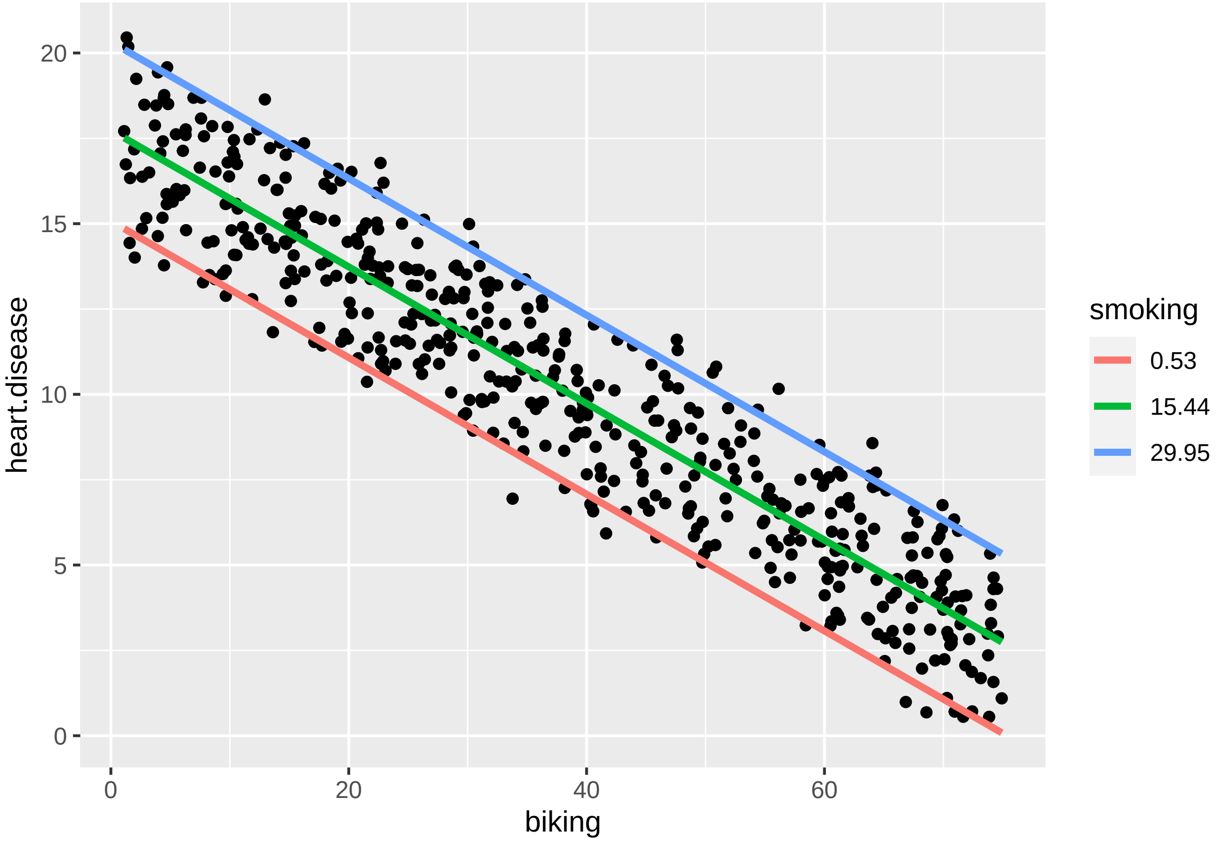
This allows us to plot the interaction between biking and heart disease at each of the three levels of smoking we chose.

PLOT THE ORIGINAL DATA

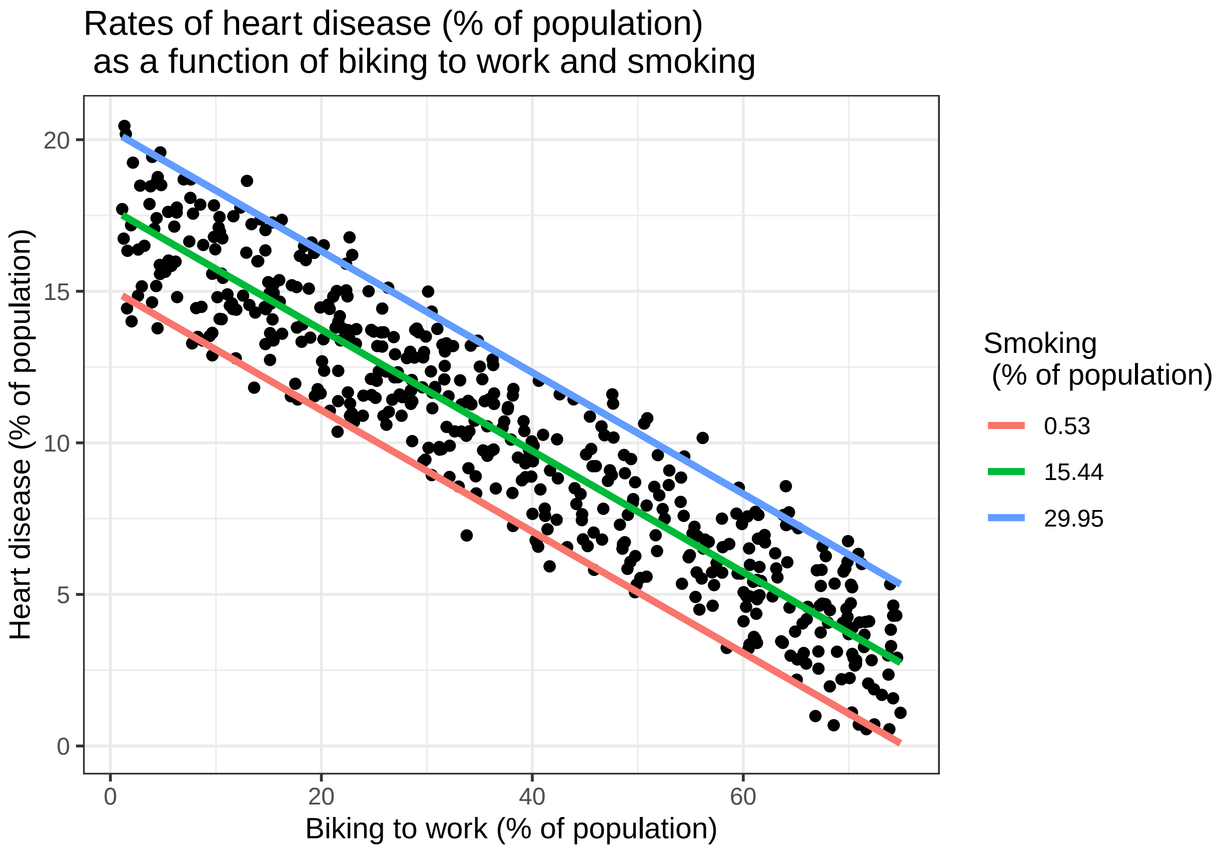


ADD THE REGRESSION LINES

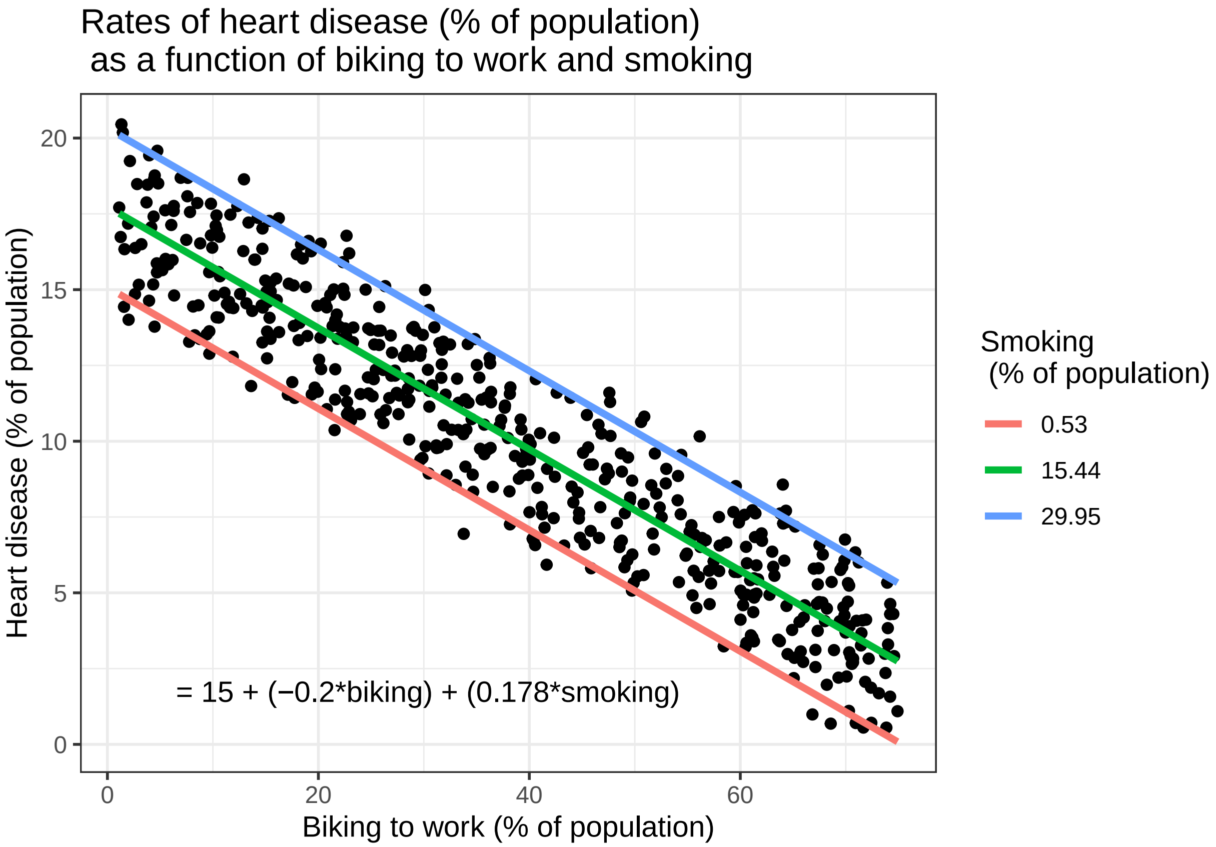
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



MAKE THE GRAPH READY FOR PUBLICATION



Because this graph has two regression coefficients, the stat_regline_equation() function won't work here. But if we want to add our regression model to the graph, we can do so like this:



This is the finished graph that you can include in your papers!

REPORT THE RESULTS

In our survey of 500 towns, we found significant relationships between the frequency of biking to work and the frequency of heart disease and the frequency of smoking and frequency of heart disease ($p < 0$ and $p < 0.001$, respectively). Specifically, we found a 0.2% decrease (± 0.0014) in the frequency of heart disease for every 1% increase in biking, and a 0.178% increase (± 0.0035) in the frequency of heart disease for every 1% increase in smoking.

APPENDIX

R-CODES

```r
knitr::opts_chunk$set(echo = FALSE)
# load libraries
suppressWarnings({
  # Code that generates specific warnings
  # Other code
  library(ggplot2)
  library(dplyr)
  library(broom)
  library(ggpubr)
})

suppressMessages({
  library(ggplot2)
  library(dplyr)
  library(broom)
  library(ggpubr)
})
heart.data <- read.csv("https://raw.githubusercontent.com/enidroman/DATA-621-Business-Analytics-and-Dat
summary(heart.data)
cor(heart.data$biking, heart.data$smoking)
hist(heart.data$heart.disease)
plot(heart.disease ~ biking, data=heart.data)
plot(heart.disease ~ smoking, data=heart.data)
heart.disease.lm<-lm(heart.disease ~ biking + smoking, data = heart.data)

summary(heart.disease.lm)
par(mfrow=c(2,2))
plot(heart.disease.lm)
par(mfrow=c(1,1))
plotting.data<-expand.grid(
  biking = seq(min(heart.data$biking), max(heart.data$biking), length.out=30),
    smoking=c(min(heart.data$smoking), mean(heart.data$smoking), max(heart.data$smoking)))
plotting.data$predicted.y <- predict.lm(heart.disease.lm, newdata=plotting.data)
plotting.data$smoking <- round(plotting.data$smoking, digits = 2)
plotting.data$smoking <- as.factor(plotting.data$smoking)
heart.plot <- ggplot(heart.data, aes(x=biking, y=heart.disease)) +
  geom_point()

heart.plot
heart.plot <- heart.plot +
  geom_line(data=plotting.data, aes(x=biking, y=predicted.y, color=smoking), size=1.25)

heart.plot
heart.plot <-
heart.plot +
  theme_bw() +
  labs(title = "Rates of heart disease (% of population) \n as a function of biking to work and smoking
      x = "Biking to work (% of population)",
      y = "Heart disease (% of population)",
      color = "Smoking \n (% of population)")

heart.plot
heart.plot + annotate(geom="text", x=30, y=1.75, label=" = 15 + (-0.2*biking) + (0.178*smoking)")
```