assigment 2

omid

2024-02-06

```r
# Load necessary libraries
library(ggplot2)

#set the theme
theme_update(
  # Set plot background color to white
  plot.background = element_rect(fill = "#FFFFFF", color = "#FFFFFF"),
  # Set panel background color to white with no border
  panel.background = element_rect(fill = "#FFFFFF", color = NA),
  # Customize x-axis text appearance
  axis.text.x = element_text(angle=0, face = "plain",
                             color = "black", size = 10, hjust=1, vjust = 1,
                             margin = margin(r = 0)),
  # Customize y-axis text appearance
  axis.text.y = element_text(face = "plain", color = "black", size = 11,
                             hjust = 1, margin = margin(r = 5)),
  # Customize x-axis line appearance
  axis.line.x = element_line(color="black", size = .5),
  # Customize y-axis line appearance
  axis.line.y = element_line(color="black", size = .5),
  axis.ticks = element_blank(), # Remove axis ticks

  # Customize plot title appearance
  plot.title = element_text(face = "bold", color = "black", size = 15,
                            hjust = 0, margin = margin(b = 5, l=0)),
  # Customize plot subtitle appearance
  plot.subtitle = element_text(face = "plain", color = "#737c7d",
                               size = 10, hjust = 0,
                               margin = margin(b = 20, l=10)),
  # Customize legend text appearance
  legend.text=element_text(face = "plain", color = "black", size = 12),
  # Customize legend title appearance
  legend.title = element_text(face = "bold", color = "black", size = 12)
)
```

Problem 1: Covariance and Correlation

For this problem, use the Excel file called “`Uninsured.xslx`” in the “`Datasets`” folder on GitHub. This data set contains information on 20 municipalities in Massachusetts. For each municipality, the fraction of people without health insurance (frac uninsured) and the fraction of people declaring bankruptcy (frac bankrupt) are reported.

1-A. What is the covariance between these two variables? Make a nice-looking scatterplot of

the variables' relationship. How does the covariance you reported jibe with the graph? Why do you think this is?

- install the **`tidyverse`** package

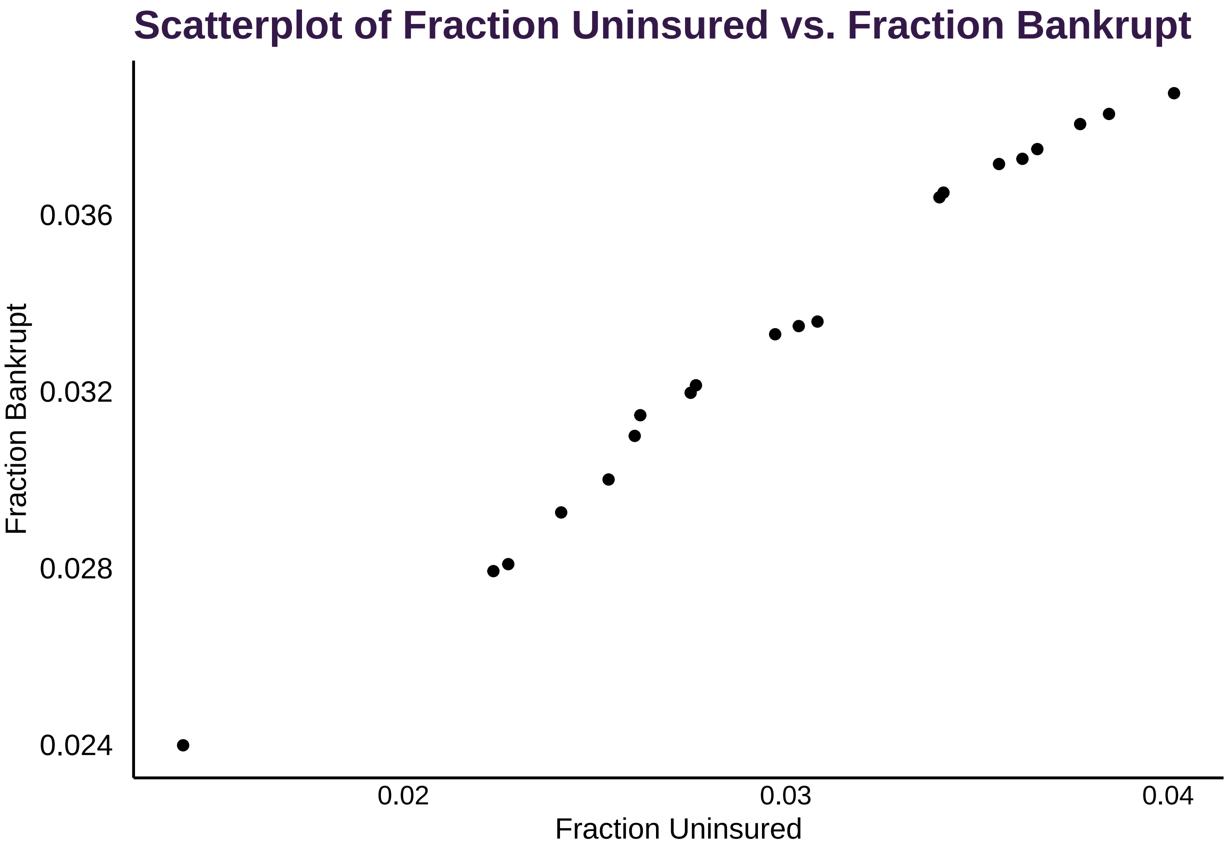
```r
# Check if tidyverse is already installed
if (!requireNamespace("tidyverse", quietly = TRUE)) {
  # If not installed, install tidyverse
  install.packages("tidyverse")
}

# Load tidyverse
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# Load necessary libraries
library(readxl)

# Read the Excel file
data <- read_excel("Uninsured.xlsx")

# Calculate covariance
covariance <- cov(data$frac_uninsured, data$frac_bankrupt)

# Print covariance
print(covariance)
```

```
## [1] 2.738676e-05
```

```r
# Load necessary libraries
library(ggplot2)

# Create a scatterplot
ggplot(data, aes(x = frac_uninsured, y = frac_bankrupt)) +
  geom_point() +
  labs(x = "Fraction Uninsured", y = "Fraction Bankrupt") +
  ggtitle("Scatterplot of Fraction Uninsured vs. Fraction Bankrupt") +
  theme(plot.title = element_text(color = "#341948"))
```



The covariance between the variables "`frac_uninsured`" and "`frac_bankrupt`" is approximately 2.738676×10^{-5}, and we observe that the scatterplot resembles a line with a slope similar to the diagonal, it suggests a weak positive linear relationship between the two variables. Since the covariance is positive, it indicates that as one variable increases, the other tends to increase as well. However, the covariance value being **close to zero suggests that the relationship is weak**. This interpretation aligns with both the numerical value of the covariance and the visual observation from the scatterplot.

The weak positive linear relationship may stem from economic disparities, high healthcare costs, policy variations, demographic differences, and regional factors. Another possibility for the weak positive linear relationship between the variables could be attributed to limitations in data gathering techniques. Variations in data collection methods, sample sizes, and measurement errors could introduce noise and weaken the observed relationship between the variables. Therefore, ensuring robust data collection techniques and minimizing biases in data sampling could enhance the accuracy of the relationship analysis. `However, in the mext question (1-B), where the variables were scaled to represent the absolute number of uninsured individuals and bankruptcies, the covariance significantly increased. This change in covariance indicates a stronger linear relationship between the variables when measured in terms of absolute numbers.`

1-B. Create new variables for both bankruptcy and (un-)insurance that is measured in people (rather than percentages). Use the population variable to do so. Does this change the linear relationship? What is the new covariance? What does this teach you about covariance and data viz?

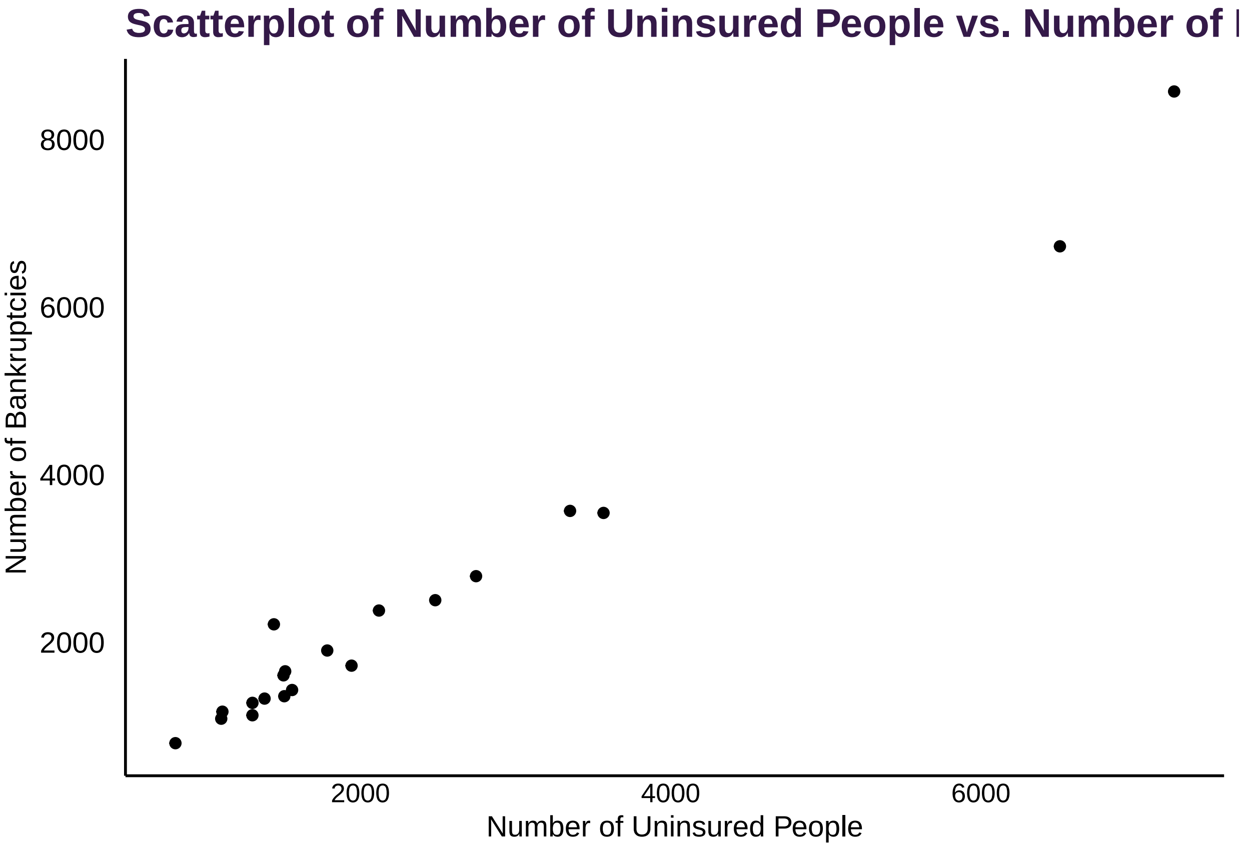
```r
# Create new variables for bankruptcy and uninsured individuals measured in people
uninsured_data <- data %>%
  mutate(bankruptcy_people = frac_bankrupt * Population,
         uninsured_people = frac_uninsured * Population)

# Calculate the covariance between the new variables
```

```r
covariance_normalized_with_population <- cov(uninsured_data$uninsured_people, uninsured_data$bankruptcy_

# Print the new covariance
print(covariance_normalized_with_population)
```

```
## [1] 3358976
```

The new covariance between the number of uninsured people and the number of bankruptcies is $3,358,976$. This represents the change in covariance when the variables are measured in terms of people rather than fractions. Comparing this new covariance with the previous one, which was 2.738676×10^{-5}, highlights the importance of considering the scale of variables when interpreting covariance. The significant increase in covariance after converting the variables to absolute numbers suggests a stronger linear relationship between the number of uninsured individuals and the number of bankruptcies. This change teaches us that the scale of variables can greatly influence their covariance and, consequently, our understanding of their relationship.

```r
# Load necessary libraries
library(ggplot2)

# Create a scatterplot
ggplot(uninsured_data, aes(x = uninsured_people, y = bankruptcy_people)) +
  geom_point() +
  labs(x = "Number of Uninsured People", y = "Number of Bankruptcies") +
  ggtitle("Scatterplot of Number of Uninsured People vs. Number of Bankruptcies") +
  theme(plot.title = element_text(color = "#341948"))
```



Creating new variables based on population allows us to express the fractions of uninsured individuals and people declaring bankruptcy in terms of actual counts of people rather than percentages. This change in scale can provide additional insights into the relationship between these variables. When we express the variables in terms of people rather than percentages, it enables us to directly compare the absolute numbers of uninsured individuals and bankruptcies across different municipalities. This can be particularly useful when analyzing the impact of these factors on a larger scale or when considering absolute numbers in policy or resource allocation decisions. Additionally, expressing the variables in terms of people may help in avoiding potential distortions that could arise from differences in population sizes across municipalities when comparing percentages alone. For example, a small increase in the percentage of uninsured individuals in a municipality with a large population could represent a larger absolute number of uninsured individuals compared to the same percentage increase in a municipality with a smaller population. Overall, creating new variables based on population provides a clearer and more interpretable representation of the data, allowing for a better understanding of the relationship between uninsured rates and bankruptcy rates across municipalities.

In other words, covariance values are sensitive to the scale of the data. and this makes them difficults to interpret.The solution, correlation, addresses this issue by standardizing the covariance values to a scale between -1 and 1. Correlation coefficients provide a standardized measure of the linear relationship between two variables, regardless of their scales.

1-C) As discussed in class, the correlation is a `unitless` measure that resolves some of the problems discussed above. What is the correlation between the two original variables? Does this correlation change when you use the new variables (based on people, not percentages) instead? Why (or why not)?

```r
# Calculate correlation between original variables
correlation_original <- cor(data$frac_uninsured, data$frac_bankrupt)

# Print correlation between original variables
print(correlation_original)
```

```
## [1] 0.9943376
```

```r
# Calculate correlation between new variables (based on people, not percentages)
correlation_new <- cor(uninsured_data$uninsured_people, uninsured_data$bankruptcy_people)

# Print correlation between new variables
print(correlation_new)
```

```
## [1] 0.9899327
```

The correlation coefficient ranges from `-1` to `1`. A correlation of `1` indicates a perfect positive linear relationship, `-1` indicates a perfect negative linear relationship, and `0` indicates no linear relationship. Since correlation is unitless, it is not affected by the scale of the variables.

The correlation coefficient between the original variables (`frac_uninsured` and `frac_bankrupt`) is approximately 0.994, indicating a very strong positive linear relationship between the fractions of uninsured individuals and bankruptcies. The correlation coefficient between the new variables based on absolute numbers of uninsured individuals and bankruptcies is approximately 0.990, also indicating a very strong positive linear relationship between the numbers of uninsured individuals and bankruptcies.

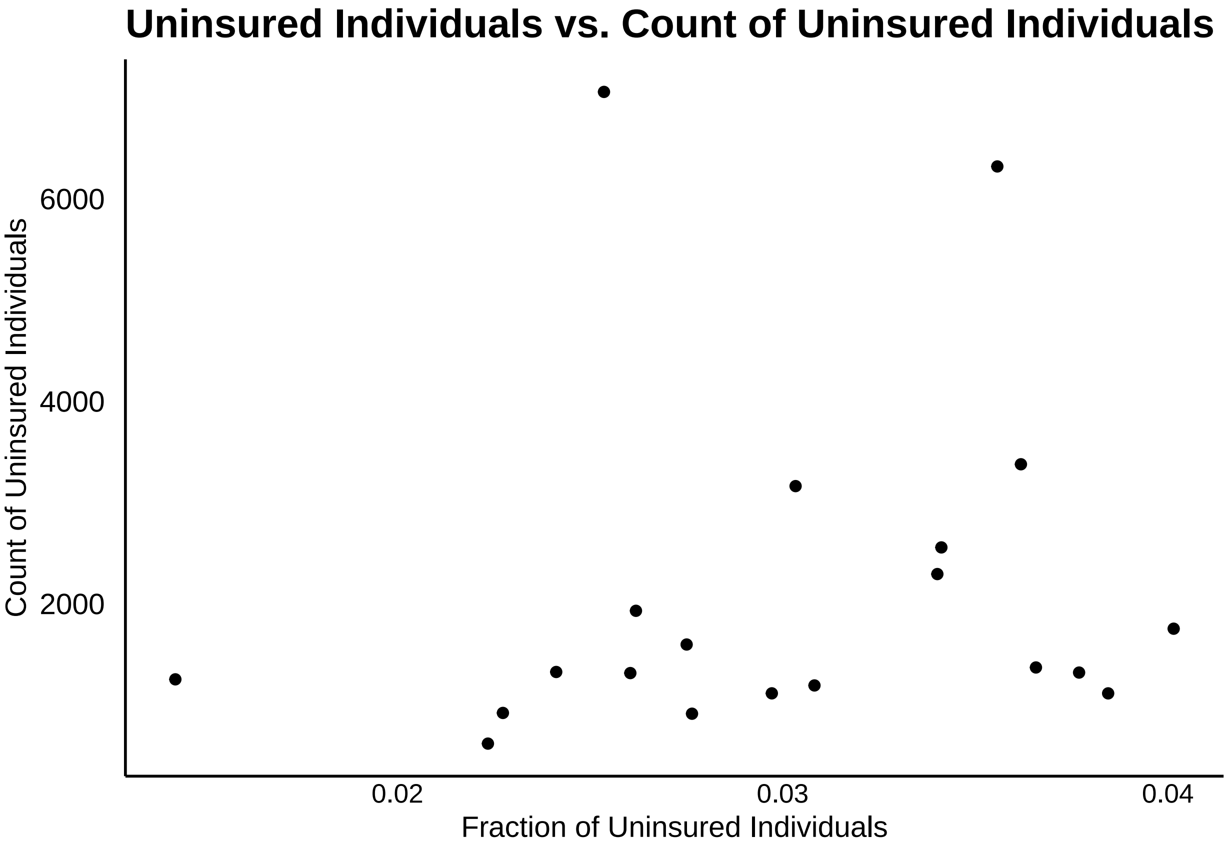
Despite the change in scale from percentages to absolute numbers, the correlation remains very high and stable. This suggests that the relationship between the variables is consistent regardless of the scaling, `confirming the robustness of the correlation coefficient as a unitless measure`.

1-D) What is the correlation between `frac uninsured` and your count of uninsured? What explains this? Why doesn't this cause a problem in calculating the correlation between your new variables? (A scatter plot—or multiple plots—may be useful here.)

```r
# Calculate correlation between frac_uninsured and uninsured_people
correlation_frac_uninsured_with_uninsured_people <- cor(uninsured_data$frac_uninsured, uninsured_data$un
# Print correlation between frac_uninsured and uninsured_people
print(correlation_frac_uninsured_with_uninsured_people)
```

```
## [1] 0.1558009
```

```
# Create scatterplot
ggplot(uninsured_data, aes(x = frac_uninsured, y = uninsured_people)) +
  geom_point() +
  labs(x = "Fraction of Uninsured Individuals", y = "Count of Uninsured Individuals") +
  ggtitle("Uninsured Individuals vs. Count of Uninsured Individuals")
```



The correlation coefficient between the original variable “frac_uninsured” (fraction of people without health insurance) and the new variable “uninsured_people” (count of uninsured individuals) is approximately 0.156. This correlation suggests a weak positive linear relationship between the fraction of uninsured individuals and the count of uninsured individuals across the municipalities. The weak correlation might be expected, as the fraction of uninsured individuals is a percentage that could vary independently of the population size. In contrast, the count of uninsured individuals takes into account both the fraction and the population size, potentially attenuating the strength of the linear relationship.

Despite this lower correlation between the fraction and count variables, it doesn’t cause a problem in calculating the correlation between the new variables based on absolute numbers. This is because the correlation coefficient is a measure of linear association that remains robust even when the relationship between variables is not perfect. Additionally, the normalization process in correlation calculation helps account for differences in scale and ensures that the correlation is not affected by changes in units or scaling of the variables.

1-E) However, even looking at the correlation can be misleading. Create a data set of 1000**observations and 2 variables:**X **that ranges continuously over the interval** $[0, 5]$ **and** Y **given by** $Y = -X * (X - 5)$**. Create a scatter plot of this relationship. What economic variables may**

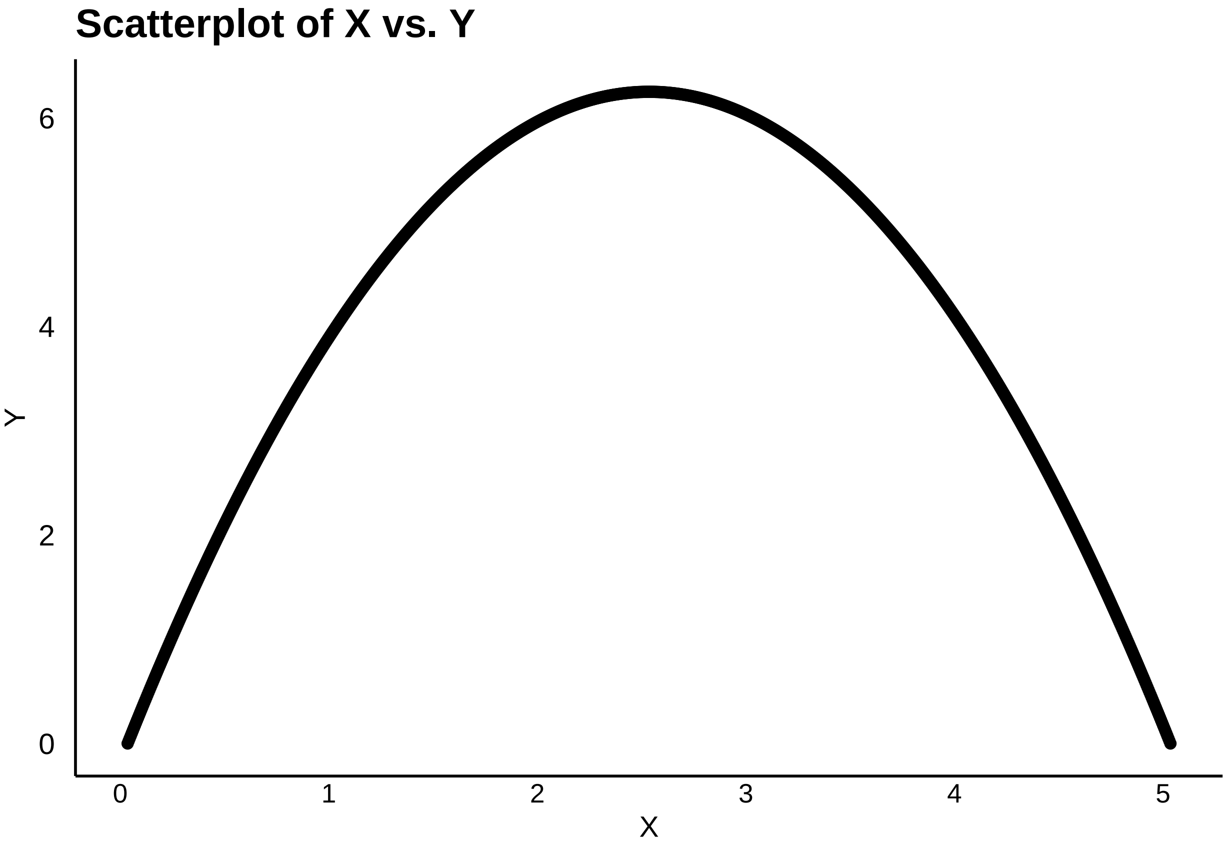
have this relationship? What is the correlation between X and Y, and what drives this result? When should I be careful of looking only at the correlation coefficient ρ?

- **To create X, I recommend looking at the *seq*() command.**

```
# Generate values for X ranging from 0 to 5
X <- seq(0, 5, length.out = 1000)

# Calculate corresponding values for Y using the relationship Y = -X * (X - 5)
Y <- -X * (X - 5)

# Create a dataframe with X and Y
data <- data.frame(X = X, Y = Y)

# Calculate correlation between X and Y
correlation <- cor(data$X, data$Y)
print(correlation)
```

```
## [1] -1.626454e-16
```

```
# Create a scatter plot
ggplot(data, aes(x = X, y = Y)) +
  geom_point() +
  labs(x = "X", y = "Y") +
  ggtitle("Scatterplot of X vs. Y")
```



The correlation coefficient between X and Y is approximately $-1.626454 \times 10^{-16}$. This value is extremely close to zero, indicating almost no linear relationship between X and Y.

The relationship between X and Y, as defined by $Y = -X \times (X - 5)$, creates a downward-opening parabolic curve. Economic variables that may exhibit a similar relationship could include costs and revenues in certain business models, where increasing costs lead to decreasing revenues due to competitive pressures or diminishing returns.

The correlation coefficient being close to zero highlights an important point: while correlation coefficients are valuable measures of **linear** association, they may not capture non-linear relationships. In this case, even though there's a clear functional relationship between X and Y, it's not captured by the correlation coefficient because it's not a linear relationship.

Therefore, it's crucial to be cautious when interpreting correlation coefficients and not rely solely on them, especially when the relationship between variables may be non-linear. Visual inspection of the data, as demonstrated by the scatter plot, can provide additional insights into the nature of the relationship.

Problem 2: Confidence Intervals in R. Suppose we sample n data points from a distribution $N(\theta, 36)$, where the value of the central mean θ is unknown.

- **a) Suppose that $n = 100$ and the sample mean is estimated to be $\bar{x} = 25$.** What is the 90% confidence interval for θ? The 95? The 99?
- **b) Repeat the exercise for $n = 1,000$.** Why are the confidence intervals always narrower?

```r
calculate_confidence_intervals <- function(n) {
  sample_mean <- 25   # Sample mean
  standard_deviation <- 6   # Standard deviation (sqrt(36))

  # Confidence levels
  confidence_levels <- c(0.90, 0.95, 0.99)

  # Calculate standard error of the mean
  standard_error <- standard_deviation / sqrt(n)

  # Calculate the margin of error for each confidence level
  margin_of_error <- qnorm(1 - (1 - confidence_levels) / 2) * standard_error

  # Calculate lower and upper bounds of the confidence intervals
  lower_bounds <- round(sample_mean - margin_of_error, 2)
  upper_bounds <- round(sample_mean + margin_of_error, 2)

  # Combine the results into a data frame
  confidence_intervals <- data.frame(
    "Confidence Level" = confidence_levels,
    "Lower Bound" = lower_bounds,
    "Upper Bound" = upper_bounds
  )

  print(confidence_intervals)
}

# Given values
n <- 100      # Sample size
calculate_confidence_intervals(n)
```

```
##   Confidence.Level Lower.Bound Upper.Bound
## 1             0.90       24.01       25.99
## 2             0.95       23.82       26.18
## 3             0.99       23.45       26.55
```

Repeat the exercise for $n = 1,000$

```
n <- 1000      # Sample size
calculate_confidence_intervals(n)
```

```
##   Confidence.Level Lower.Bound Upper.Bound
## 1             0.90       24.69       25.31
## 2             0.95       24.63       25.37
## 3             0.99       24.51       25.49
```

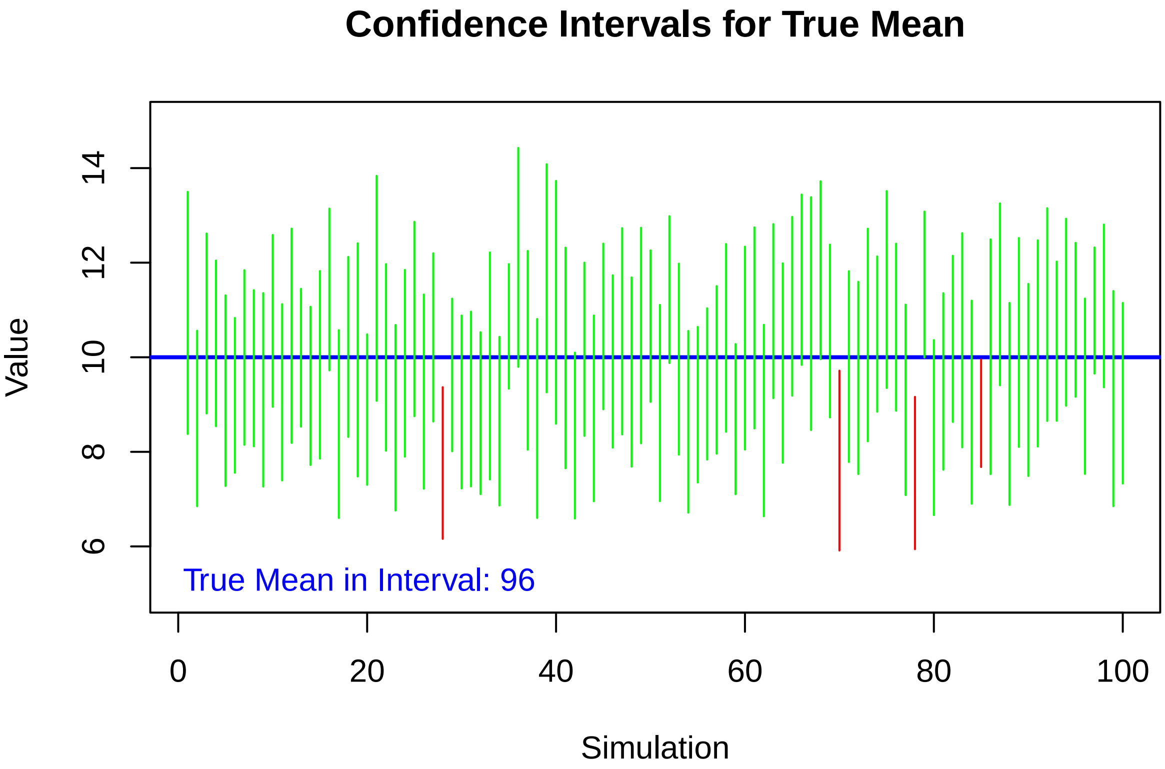
Smaller margins of error result in narrower confidence intervals. This means that as the sample size increases, our estimate of the population mean becomes more precise, leading to a more confident and narrower range of values for the true population mean. In summary, larger sample sizes provide more information and reduce the uncertainty in our estimates, resulting in narrower confidence intervals.

Problem 3: Confidence Intervals and Probabilities. People may mistake a confidence interval to mean there is a certain probability that the true parameter value lies in the interval. Let's explore more why this is incorrect.

A) Simulation. Suppose the true mean is $\mu = 10$. Write a loop that performs and stores this output 100 times:

- **Sample 25 observations from** $N(\mu, 25)$ **distribution**.
- **Compute the 95% confidence interval given your sample.**

```
# Set seed for reproducibility
set.seed(42)

# True population mean
true_mean <- 10

# Number of simulations
num_simulations <- 100

# Sample size
sample_size <- 25

# Confidence level
confidence_level <- 0.95

# Initialize a matrix to store the lower and upper bounds of confidence intervals
interval_bounds <- matrix(NA, nrow = num_simulations, ncol = 2)

# Perform the simulation and store the interval bounds
for (i in 1:num_simulations) {
  # Generate a random sample from N(true_mean, 25)
  sample_data <- rnorm(sample_size, mean = true_mean, sd = sqrt(25))

  # Calculate the sample mean and standard error
  sample_mean <- mean(sample_data)
  standard_error <- sd(sample_data) / sqrt(sample_size)

  # Calculate the margin of error
  margin_of_error <- qnorm(1 - (1 - confidence_level) / 2) * standard_error

  # Calculate the confidence interval
  lower_bound <- sample_mean - margin_of_error
  upper_bound <- sample_mean + margin_of_error
  # Store the lower and upper bounds
  interval_bounds[i, ] <- c(lower_bound, upper_bound)
}

# Plot the confidence intervals
plot(1:num_simulations, rep(true_mean, num_simulations), type = "n",
     xlab = "Simulation", ylab = "Value", main = "Confidence Intervals for True Mean",
     ylim = c(true_mean - 5, true_mean + 5))
abline(h = true_mean, col = "blue", lwd = 2)  # Add a vertical line for the true mean

# Plot confidence intervals as ranges
#
# Initialize counter for number of times true_mean is in the interval
num_true_mean_in_interval <- 0
for (i in 1:num_simulations) {
  if (true_mean >= interval_bounds[i, 1] & true_mean <= interval_bounds[i, 2]) {
    lines(rep(i, 2), interval_bounds[i, ], col = "green")
    num_true_mean_in_interval <- num_true_mean_in_interval + 1  # Increment counter
  } else {
    lines(rep(i, 2), interval_bounds[i, ], col = "red")
  }
}

# Add text to display the number of times true_mean is in the interval
text(0.5, 5.3, paste("True Mean in Interval:", num_true_mean_in_interval),
     col = "blue", adj = 0)
```



How many times is μ in your interval? 96

When we repeatedly conduct the simulation a large number of times, the proportion of times the true population mean falls within the calculated confidence interval is expected to converge to 95%. However, due to randomness inherent in the simulation process, the actual result may vary from the expected value, but it should generally be close to 95%.

```r
# Set seed for reproducibility
set.seed(42)

# True population mean
true_mean <- 10

# Number of simulations
num_simulations <- 100000

# Sample size
sample_size <- 25

# Confidence level
confidence_level <- 0.95

# Initialize a counter to store the number of times true_mean is in the interval
num_true_mean_in_interval <- 0

# Perform the simulation and count the number of times true_mean is in the interval
for (i in 1:num_simulations) {
  # Generate a random sample from N(true_mean, 25)
  sample_data <- rnorm(sample_size, mean = true_mean, sd = sqrt(25))

  # Calculate the sample mean and standard error
  sample_mean <- mean(sample_data)
  standard_error <- sd(sample_data) / sqrt(sample_size)

  # Calculate the margin of error
  margin_of_error <- qnorm(1 - (1 - confidence_level) / 2) * standard_error

  # Calculate the confidence interval
  lower_bound <- sample_mean - margin_of_error
  upper_bound <- sample_mean + margin_of_error

  # Check if true_mean is in the interval
  if (true_mean >= lower_bound & true_mean <= upper_bound) {
    num_true_mean_in_interval <- num_true_mean_in_interval + 1
  }
}

# Calculate the percentage of times true_mean is in the interval
percentage_true_mean_in_interval <- (num_true_mean_in_interval / num_simulations) * 100

# Display the number of times true_mean is in the interval
num_true_mean_in_interval
```

```
## [1] 93912
```

```r
# Display the percentage of times true_mean is in the interval
percentage_true_mean_in_interval
```

```
## [1] 93.912
```

B) An updated simulation. Suppose instead we have the following information about possible means and their (prior) probabilities:

μ	**0**	**1**	**2**
$p(\mu)$	**0.94**	**0.04**	**0.02**

Suppose that I pick μ at random from this distribution, and sample $N(\mu, 25)$ 64 times. I tell you that the sample mean $\bar{x} = 1.6$. (Note: this means that the confidence interval is [0.375,2.825], can you verify this?

To verify that the confidence interval for the sample mean $\bar{x} = 1.6$ is indeed $[0.375, 2.825]$, we can use the formula for a confidence interval:

$$\text{Confidence Interval} = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

Given that $n = 64$ (the sample size) and the standard deviation $\sigma = \sqrt{25} = 5$ (given that the data is sampled from $N(\mu, 25)$), we can calculate the standard error $\frac{s}{\sqrt{n}} = \frac{5}{\sqrt{64}} = \frac{5}{8}$.

Using a z-score of 1.96 for a 95% confidence interval, we get:

$$\text{Confidence Interval} = 1.6 \pm 1.96 \times \frac{5}{8} = [0.375, 2.825]$$

```r
# Sample mean
x_bar <- 1.6

# Standard deviation
sigma <- sqrt(25)

# Confidence level (95%)
conf_level <- 0.95

# Calculate the z-value for the specified confidence level
z_star <- qnorm((1 + confidence_level) / 2)

# Margin of error
margin_error <- z_star * (sigma / sqrt(64))

# Confidence interval
conf_interval <- c(x_bar - margin_error, x_bar + margin_error)

# Print confidence interval
cat("Confidence Interval:", format(conf_interval, digits = 3), "\n")
```

```
## Confidence Interval: 0.375 2.825
```

Now, let's fill in the table to calculate the posterior probabilities:

```r
# Given data
sample_mean <- 1.6
sample_size <- 64
confidence_interval <- c(0.375, 2.825)

# Possible means and their prior probabilities
means <- c(0, 1, 2)
```

```r
prior_prob <- c(0.94, 0.04, 0.02)

# Calculate the likelihood function for each hypothesis
likelihood <- dnorm(sample_mean, means, sqrt(25/64))

# Calculate the prior probabilities for each hypothesis
prior_prob_table <- data.frame(Hypothesis = means, Prior = prior_prob)

# Calculate the numerators (prior multiplied by likelihood)
numerators <- prior_prob * likelihood

# Calculate the posterior probabilities
posterior_prob <- numerators / sum(numerators)

# Create a table to display the results
results_table <- data.frame(Hypothesis = means, Prior = prior_prob,
                            Likelihood = likelihood, Numerator = numerators,
                            Posterior = posterior_prob)
results_table <- rbind(results_table, c("Total", sum(prior_prob), NA,
                                        sum(numerators), sum(posterior_prob)))

# Display the results
print(results_table)
```

```
##   Hypothesis Prior         Likelihood          Numerator         Posterior
## 1          0  0.94 0.0240953861238039 0.0226496629563757 0.460762702130413
## 2          1  0.04  0.402630945756987 0.0161052378302795 0.327629285937947
## 3          2  0.02  0.520099622534531 0.0104019924506906  0.21160801193164
## 4      Total     1               <NA> 0.0491568932373458                 1
```

The prior and posterior probabilities are not close to 95%.

C) Suppose we have data that a new cancer treatment increases patient survival by an average of 15 months longer than the conventional treatment, where the 95% confidence interval is [10,20]. Say whether each of the following statements is true or false. False means that the statement does not follow logically from the confidence-interval result.

1. **The true increase in survival is in the range [10,20] with 95% probability.**
   - **False:** This is a common misconception about confidence intervals. The 95% confidence interval means that if we were to repeatedly take samples and calculate confidence intervals from each, 95% of those intervals would capture the true population mean. It does not guarantee that any particular interval contains the true value with 95% certainty.
2. **The treatment increases survival time at all with at least 95% probability.**
   - **False:** The confidence interval only captures the average increase in survival. It doesn't tell us anything about whether the treatment is always effective for every patient. There could be cases where the treatment has no effect or even decreases survival.
3. **[10,20] is an estimate of the true average increase in survival.**
   - **True:** The confidence interval [10,20] provides an interval estimate for the true **average** increase in survival. It suggests that, based on the sample data, we are 95% confident that the true increase in survival falls within the interval [10,20].
4. **The null hypothesis that the treatment does not affect survival time is likely incorrect.**
   - **False:** The confidence interval gives a range of plausible values, but it doesn't directly test the null hypothesis. Hypothesis testing is a separate procedure, and the decision to reject the null hypothesis should be based on the significance level and p-value from the hypothesis test.
5. **After 100 experiments, in approximately 95 of them, the 95% confidence intervals would**

contain the true value of survival increases.

- **True:** This statement aligns with the correct interpretation of a 95% confidence interval. If we were to conduct many experiments and calculate 95% confidence intervals for each, we would expect about 95% of those intervals to contain the true parameter.

6. **We reject the null hypothesis of no improvement in survival at 5% significance.**
   - **False:** The decision to reject or not reject the null hypothesis is based on the p-value from a hypothesis test, not directly on the confidence interval.