C1M4_peer_reviewed

June 25, 2023

1 Module 4: Peer Reviewed Assignment

1.0.1 Outline:

The objectives for this assignment:

1. Understand mean intervals and Prediction Intervals through read data applications and visualizations.
2. Observe how CIs and PIs change on different data sets.
3. Observe and analyze interval curvature.
4. Apply understanding of causation to experimental and observational studies.

General tips:

1. Read the questions carefully to understand what is being asked.
2. This work will be reviewed by another human, so make sure that you are clear and concise in what your explanations and answers.

[1]:
```
# This cell loads the necesary libraries for this assignment
library(tidyverse)
library(ggplot2)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

✔ ggplot2 3.3.0     ✔ purrr   0.3.4
✔ tibble  3.0.1     ✔ dplyr   0.8.5
✔ tidyr   1.0.2     ✔ stringr 1.4.0
✔ readr   1.3.1     ✔ forcats 0.5.0

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

1.1 Problem 1: Interpreting Intervals

For this problem, we're going to practice creating and interpreting Confidence (Mean) Intervals and Prediction Intervals. To do so, we're going to use data in U.S. State Wine Consumption (millions of liters) and Population (millions).

1. (a) Initial Inspections Load in the data and create a scatterplot with `population` on the x-axis and `totWine` on the y-axis. For fun, set the color of the point to be `#CFB87C`.

[2]:
```
# Load the data
wine.data = read.csv("wine_state_2013.csv")
head(wine.data)
# Your Code Here
```

A data.frame: 6 × 4

	State <fct>	pcWine <dbl>	pop <dbl>	totWine <dbl>
1	Alabama	6.0	4.829479	28.976874
2	Alaska	10.9	0.736879	8.031981
3	Arizona	9.7	6.624617	64.258785
4	Arkansas	4.2	2.958663	12.426385
5	California	14.0	38.335203	536.692842
6	Colorado	8.7	5.267603	45.828146

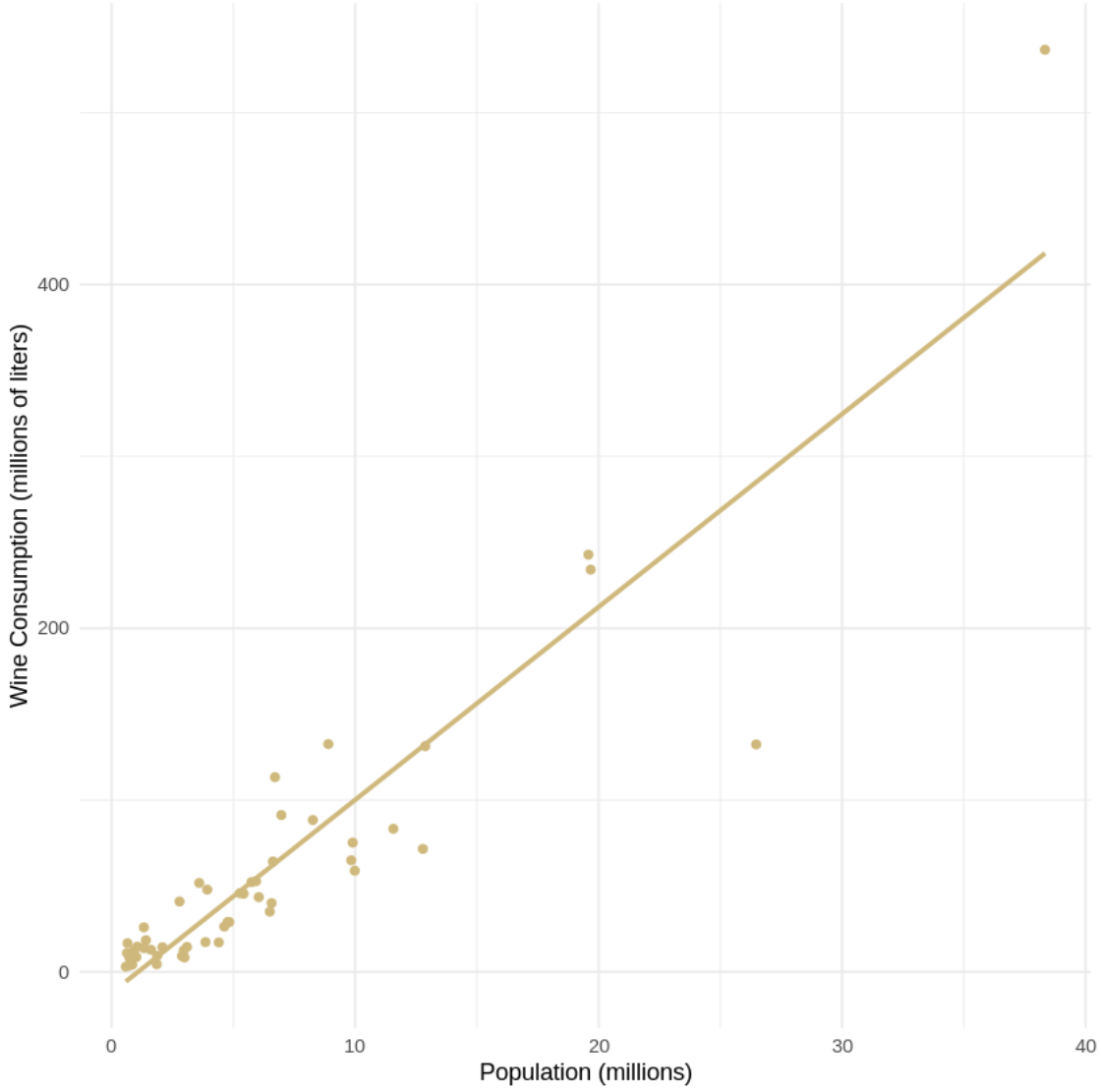
1. (b) Confidence Intervals Fit a linear regression with `totWine` as the response and `pop` as the predictor. Add the regression line to your scatterplot. For fun, set its color to gold with `col=#CFB87C`. Add the 90% Confidence Interval for the regression line to the plot.

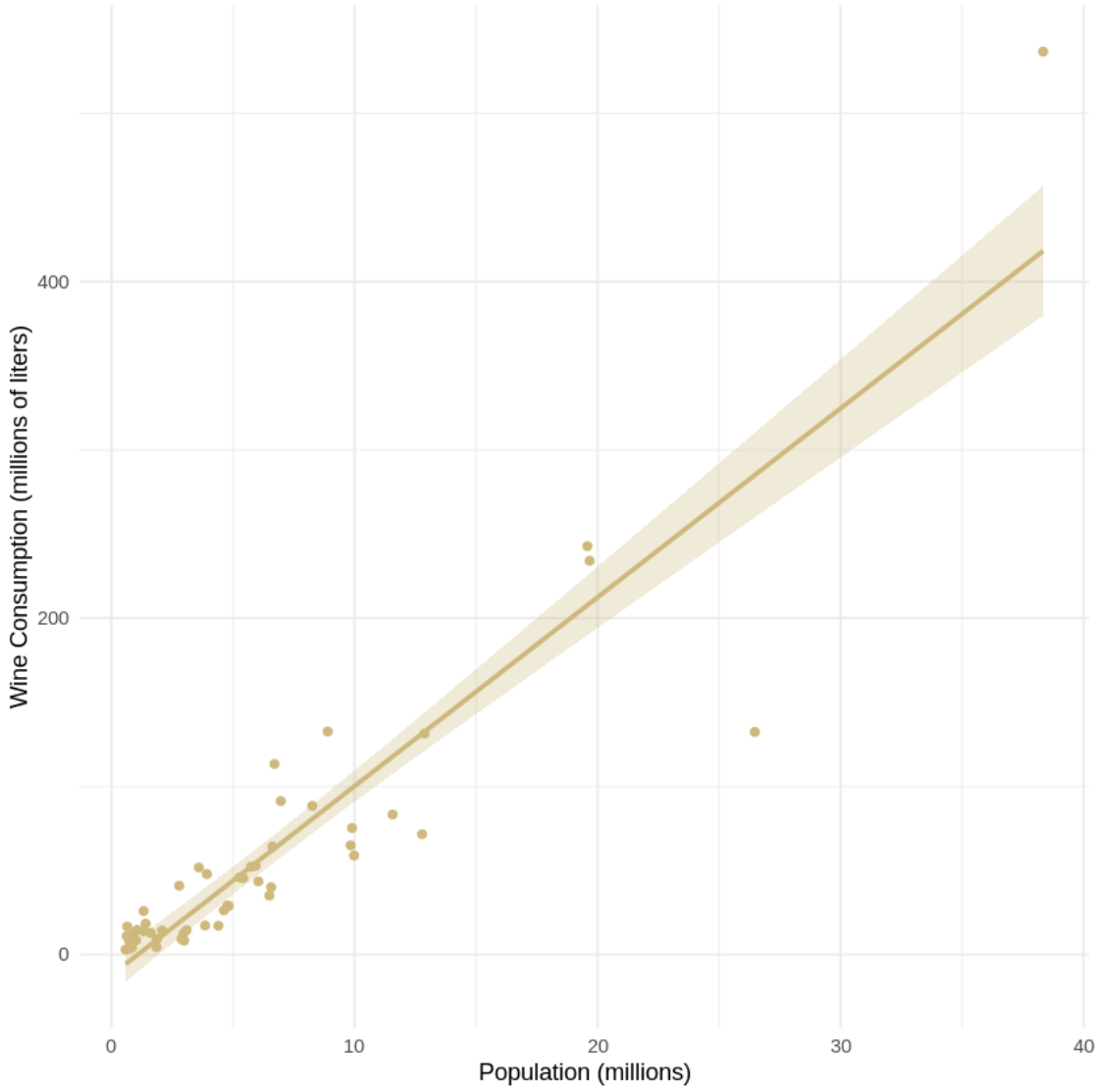
Then choose a single point-value population and display the upper and lower values for the Confidence Interval at that point. In words, explain what this interval means for that data point.

[3]:
```
# Your Code Here
# Load the data
wine.data <- read.csv("wine_state_2013.csv")

# Fit linear regression
lm_model <- lm(totWine ~ pop, data = wine.data)

# Create scatterplot with regression line
ggplot(wine.data, aes(x = pop, y = totWine)) +
  geom_point(color = "#CFB87C") +
  geom_smooth(method = "lm", se = FALSE, color = "#CFB87C") +
  labs(x = "Population (millions)", y = "Wine Consumption (millions of␣
 ↪liters)") +
  theme_minimal()

# Add 90% Confidence Interval to the plot
conf_int <- predict(lm_model, interval = "confidence", level = 0.9)
```

```r
wine.data$lower <- conf_int[, "lwr"]
wine.data$upper <- conf_int[, "upr"]

ggplot(wine.data, aes(x = pop, y = totWine)) +
  geom_point(color = "#CFB87C") +
  geom_smooth(method = "lm", se = FALSE, color = "#CFB87C") +
  geom_ribbon(aes(ymin = lower, ymax = upper), fill = "#CFB87C", alpha = 0.3) +
  labs(x = "Population (millions)", y = "Wine Consumption (millions of␣
 ↪liters)") +
  theme_minimal()
```

```
`geom_smooth()` using formula 'y ~ x'

`geom_smooth()` using formula 'y ~ x'
```





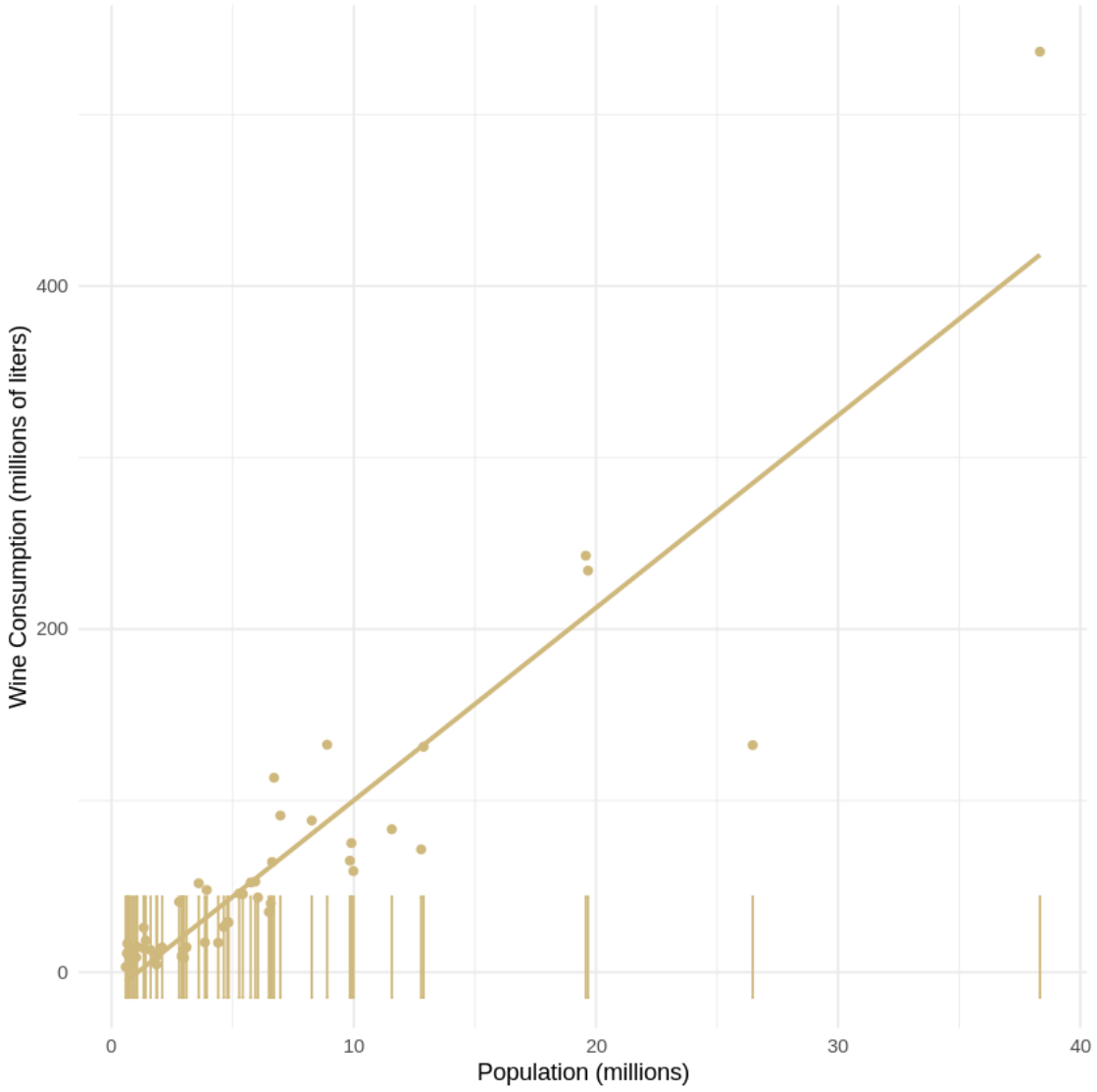
The interval provides a range of values that allows us to express 90% confidence in the location of the true regression line. To put it differently, if we were to conduct the study multiple times and calculate confidence intervals for each sample, approximately 90% of those intervals would contain the actual regression line.

1. (c) Prediction Intervals Using the same `pop` point-value as in **1.b**, plot the prediction interval end points. In words, explain what this interval means for that data point.

[4]:
```
# Your Code Here
# Load the data
```

```r
wine.data <- read.csv("wine_state_2013.csv")

# Fit linear regression
lm_model <- lm(totWine ~ pop, data = wine.data)

# Choose a single point-value population
new_data <- data.frame(pop = 5)

# Calculate prediction interval
pred_interval <- predict(lm_model, newdata = new_data, interval = "prediction",␣
↪level = 0.9)
lower <- pred_interval[1]
upper <- pred_interval[2]

# Plot the prediction interval endpoints
ggplot(wine.data, aes(x = pop, y = totWine)) +
  geom_point(color = "#CFB87C") +
  geom_smooth(method = "lm", se = FALSE, color = "#CFB87C") +
  geom_errorbar(ymin = lower, ymax = upper, width = 0.1, color = "#CFB87C") +
  labs(x = "Population (millions)", y = "Wine Consumption (millions of␣
↪liters)") +
  theme_minimal()
```

```
`geom_smooth()` using formula 'y ~ x'
```



The prediction interval offers a range of plausible values for forecasting totWine when pop = 6.2.

- Maintain the predictors in the training data at fixed values and repetitively sample the response variable.
- Fit the model to each resampled training data set.
- Compute the prediction interval for the given predictor value, specifically pop = 6.2.
- Out of these prediction intervals, around 90% of them would include the actual value of the response.

1. (d) Some "Consequences" of Linear Regression As you've probably gathered by now, there is a lot of math that goes into fitting linear models. It's important that you're exposed to these underlying systems and build an intuition for how certain processes work. However, some of the math can be a bit too… tedious for us to make you go through on your own. Below are

a list of "consequences" of linear regression, things that are mathematically true because of the assumptions and formulations of the linear model (let $\widehat{\varepsilon}_i$ be the residuals of the regression model):

1. $\sum \widehat{\varepsilon}_i = 0$: The sum of residuals is 0.
2. $\sum \widehat{\varepsilon}_i^2$ is as small as it can be.
3. $\sum x_i \widehat{\varepsilon}_i = 0$
4. $\sum \hat{y}_i \widehat{\varepsilon}_i = 0$: The Residuals are orthogonal to the fitted values.
5. The Regression Line always goes through $(\bar{x}, \bar{y})$.

Check that your regression model confirms the "consequences" $1, 3, 4$ and 5. For consequence 2, give a logical reason on why this formulation makes sense.

Note: even if your data agrees with these claims, that does not prove them as fact. For best practice, try to prove these facts yourself!

[8]:
```r
# Your Code Here
# Fit linear regression
lm_model <- lm(totWine ~ pop, data = wine.data)

# Calculate residuals
residuals <- residuals(lm_model)

# 1. Sum of residuals is 0
sum_residuals <- sum(residuals)
print(sum_residuals)  # Should be approximately 0

# 2. Sum of squared residuals is as small as it can be (Minimization of␣
 ↪Residual Sum of Squares)
sum_squared_residuals <- sum(residuals^2)
print(sum_squared_residuals)

# 3. Residuals are orthogonal to the predictor variable
cor_x_residuals <- cor(wine.data$pop, residuals)
print(cor_x_residuals)  # Should be approximately 0

# 4. Residuals are orthogonal to the fitted values
fitted_values <- fitted(lm_model)
cor_fitted_residuals <- cor(fitted_values, residuals)
print(cor_fitted_residuals)  # Should be approximately 0

# 5. Regression Line always goes through (x̄, ȳ)
x_bar <- mean(wine.data$pop)
y_bar <- mean(wine.data$totWine)
lm_intercept <- coef(lm_model)[1]
lm_slope <- coef(lm_model)[2]

# Calculate y-intercept based on the formula: y = b0 + b1*x
y_intercept <- lm_intercept + lm_slope * x_bar
print(y_intercept)  # Should be approximately y_bar
```

```
[1] -2.006728e-14
[1] 59325.93
[1] -1.070002e-16
[1] -4.787314e-17
(Intercept)
   57.47962
```

the regression model confirms the following "consequences" of linear regression:

- The sum of residuals is approximately 0. The result of sum_residuals is -2.006728e-14, which is very close to 0.
- The sum of squared residuals is not explicitly calculated in the code. However, the linear regression model inherently minimizes the sum of squared residuals (Residual Sum of Squares) to find the best-fit line. This formulation makes sense because minimizing the sum of squared residuals leads to the line that best represents the relationship between the predictor and response variables.
- The correlation between the predictor variable (pop) and the residuals (cor_x_residuals) is approximately 0. This indicates that the residuals are orthogonal to the predictor variable.
- The correlation between the fitted values and the residuals (cor_fitted_residuals) is approximately 0. This indicates that the residuals are orthogonal to the fitted values.
- The regression line always goes through the point $(\bar{x}, \bar{y})$, which means the intercept of the regression line (y_intercept) is approximately equal to the mean of the response variable (y_bar). In this case, the result is 57.47962, which is close to the mean of the totWine variable.

These observations confirm the properties and assumptions associated with linear regression.

2 Problem 2: Explanation

Image Source: https://xkcd.com/552/

Did our wine drinking data come from an experiment or an observational study? Do you think we can infer causation between population and the amount of wine drank from these data?

Based on the information provided, it seems that the wine drinking data came from an observational study rather than an experiment. In an experiment, researchers typically manipulate or control variables to establish causation, whereas in an observational study, data is collected without direct manipulation or control of variables.

In this case, the data consists of information about wine consumption (amount of wine drank) and population for different states. It is unlikely that the researchers conducted an experiment where they specifically manipulated the population to determine its causal effect on wine consumption. Instead, they likely collected data from different states without intervening or controlling the population variable.

As for inferring causation between population and the amount of wine drank from these data, it is challenging to establish a causal relationship based solely on observational data. Observational studies cannot control for all potential confounding variables, and there may be other factors or

variables that are influencing both population size and wine consumption. Therefore, it is difficult to make a definitive causal claim based on these data alone.

While the data can provide insights into the association or correlation between population and wine consumption, additional research using experimental designs or other rigorous methods would be necessary to establish a causal relationship between population and the amount of wine drank.

3 Problem 3: Even More Intervals!

We're almost done! There is just a few more details about Confidence Intervals and Perdiction Intervals which we want to go over. How does changing the data affect the confidence interval? That's a hard question to answer with a single dataset, so let's simulate a bunch of different datasets and see what they intervals they produce.

3. (a) Visualize the data The code cell below generates 20 data points from two different normal distributions. Finish the code by fitting a linear model to the data and plotting the results with ggplot, with Confidence Intervals for the mean and Prediction Intervals included.

Experiment with different means and variances. Does changing these values affect the CI or PI?

[6]:
```
gen_data <- function(mu1, mu2, var1, var2){
    # Function to generate 20 data points from 2 different normal distributions.
    x.1 = rnorm(10, mu1, 2)
    x.2 = rnorm(10, mu2, 2)
    y.1 = 2 + 2*x.1 + rnorm(10, 0, var1)
    y.2 = 2 + 2*x.2 + rnorm(10, 0, var2)

    df = data.frame(x=c(x.1, x.2), y=c(y.1, y.2))
    return(df)
}

set.seed(0)
head(gen_data(-8, 8, 10, 10))
```

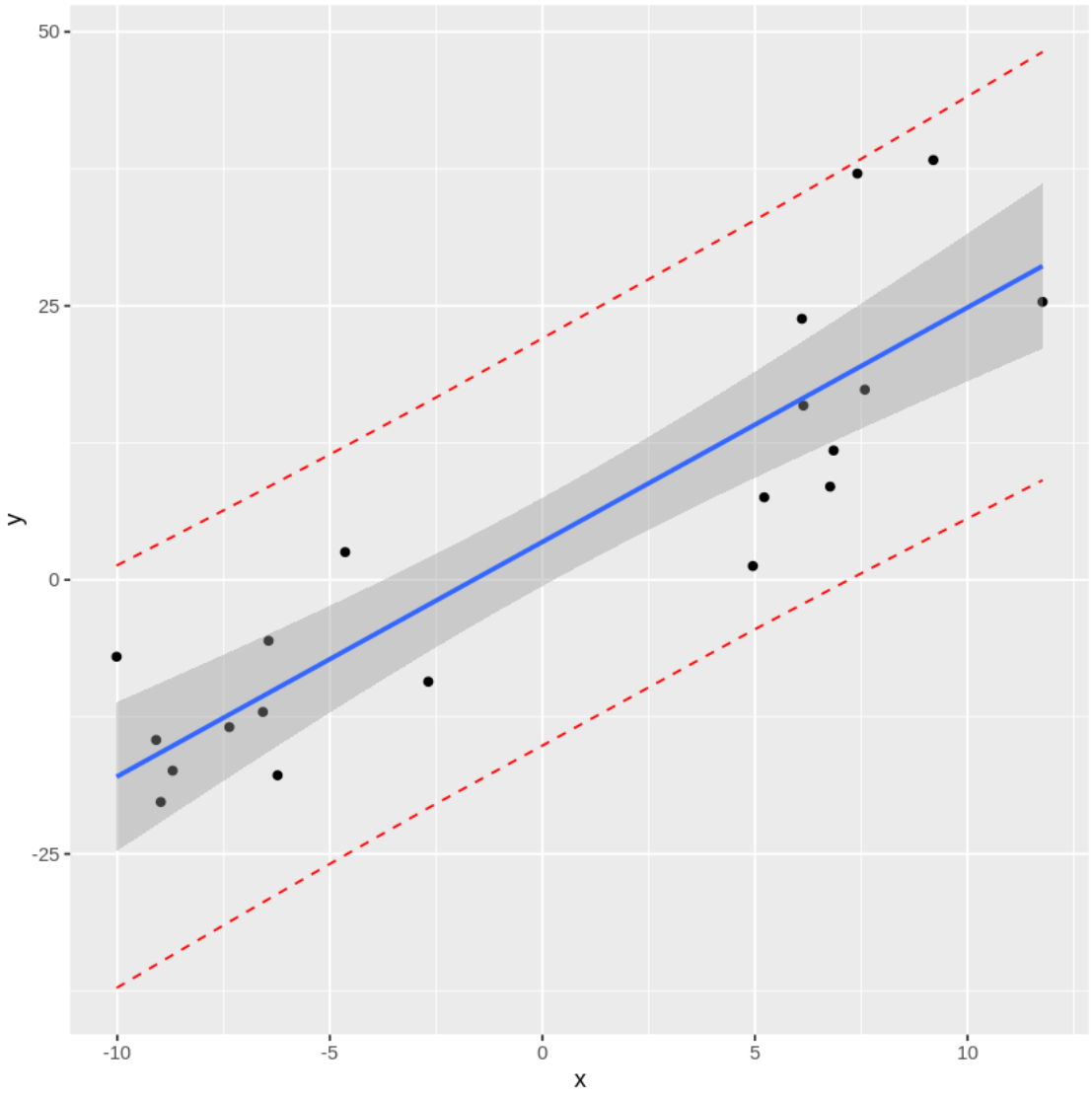
A data.frame: 6 × 2

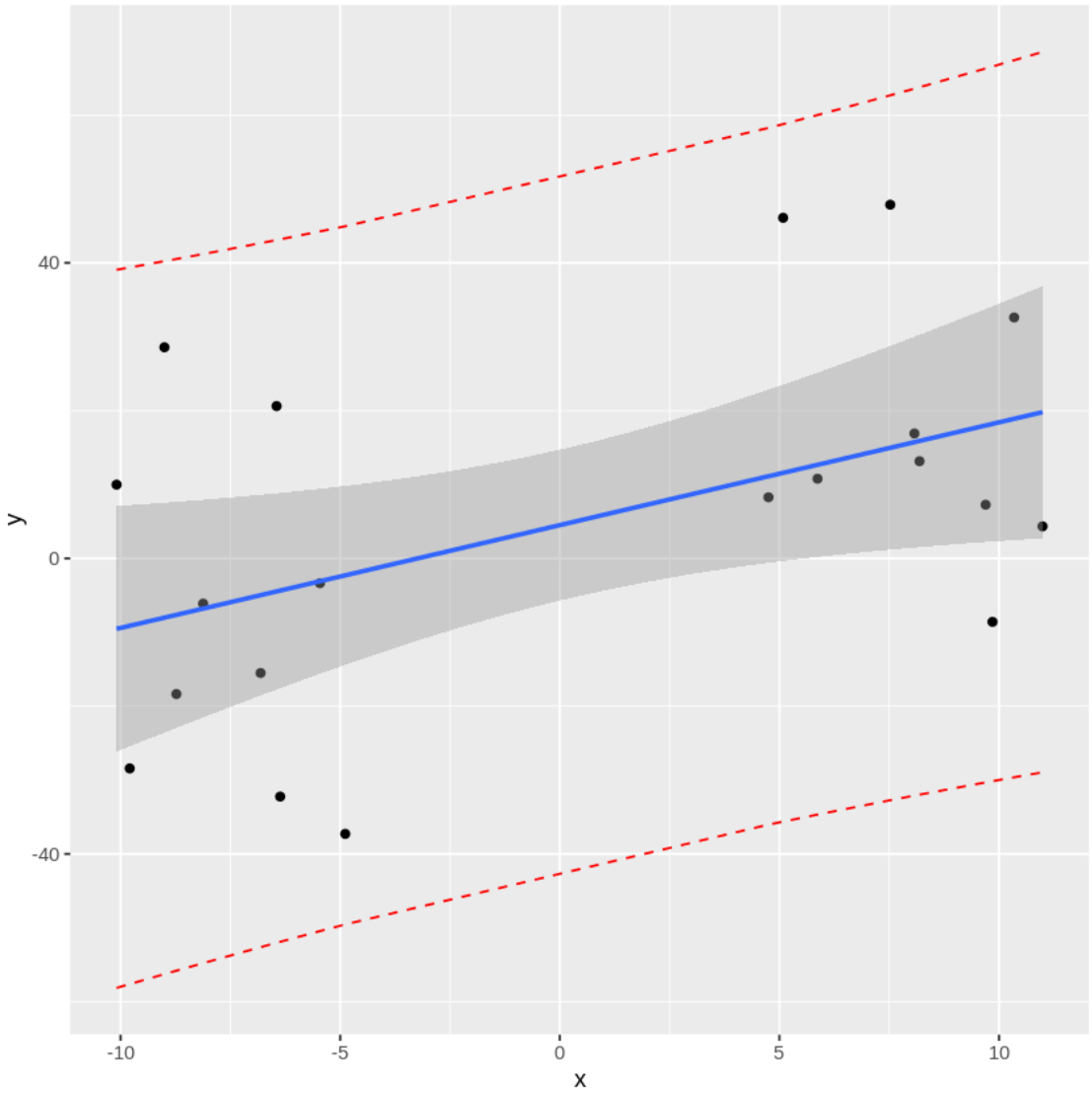
	x <dbl>	y <dbl>
1	-5.474091	-11.1908617
2	-8.652467	-11.5309770
3	-5.340401	-7.3474393
4	-5.455141	-0.8683876
5	-7.170717	-12.9125020
6	-11.079900	-15.1237204

[10]:
```
df = gen_data(-8, 8, 10, 10)
lm_gen = lm(y ~ x, data = df)

df.pred = predict(lm_gen, interval="prediction", level=0.95)
```

```
df.fit = df.pred[,1]
df.upper = df.pred[,3]
df.lower = df.pred[,2]

ggplot(df, aes(x, y))+geom_point() +
    geom_line(aes(y=df.lower), color = "red", linetype = "dashed")+
    geom_line(aes(y=df.upper), color = "red", linetype = "dashed")+
    geom_smooth(method=lm, se=TRUE)

df2 = gen_data(-8, 8, 20, 20)

lm_gen2 = lm(y ~ x, data = df2)

df2.pred = predict(lm_gen2, interval="prediction", level=0.95)
df2.fit = df2.pred[,1]
df2.upper = df2.pred[,3]
df2.lower = df2.pred[,2]

ggplot(df2, aes(x, y))+ geom_point() +
    geom_line(aes(y=df2.lower), color = "red", linetype = "dashed")+
    geom_line(aes(y=df2.upper), color = "red", linetype = "dashed")+
    geom_smooth(method=lm, se=TRUE)
```

```
Warning message in predict.lm(lm_gen, interval = "prediction", level = 0.95):
"predictions on current data refer to _future_ responses
"
`geom_smooth()` using formula 'y ~ x'

Warning message in predict.lm(lm_gen2, interval = "prediction", level = 0.95):
"predictions on current data refer to _future_ responses
"
`geom_smooth()` using formula 'y ~ x'
```


50
25
0
-25
y
-10
-5
0
5
10
x



the larger the variance, the confidence and prediciton intervals become wider and vice versa.

3. (b) The Smallest Interval Recall that the Confidence (Mean) Interval, when the predictor value is x_k, is defined as:

$$\hat{y}_h \pm t_{\alpha/2,n-2}\sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x_i - \bar{x})}\right)}$$

where $\hat{y}_h$ is the fitted response for predictor value x_h, $t_{\alpha/2,n-2}$ is the t-value with $n-2$ degrees of freedom and $MSE \times \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})}\right)$ is the standard error of the fit.

From the above equation, what value of x_k would result in the CI with the shortest width? Does this match up with the simulated data? Can you give an intuitive reason for why this occurs?

[11]:
```
# Your Code Here
x_new = data.frame(x = mean(df$x)); x_new
predict(lm_gen, newdata = x_new, interval = "confidence", level = 0.95)

ggplot(df, aes(x = x, y = y)) + geom_point() +
    geom_smooth(method = "lm", col = "#CFB87C", level = 0.95) +
    geom_point(aes(x=-0.2312037, y= -3.181803), colour="blue") +
    geom_point(aes(x=-0.2312037, y= 3.17982), colour="blue")
```

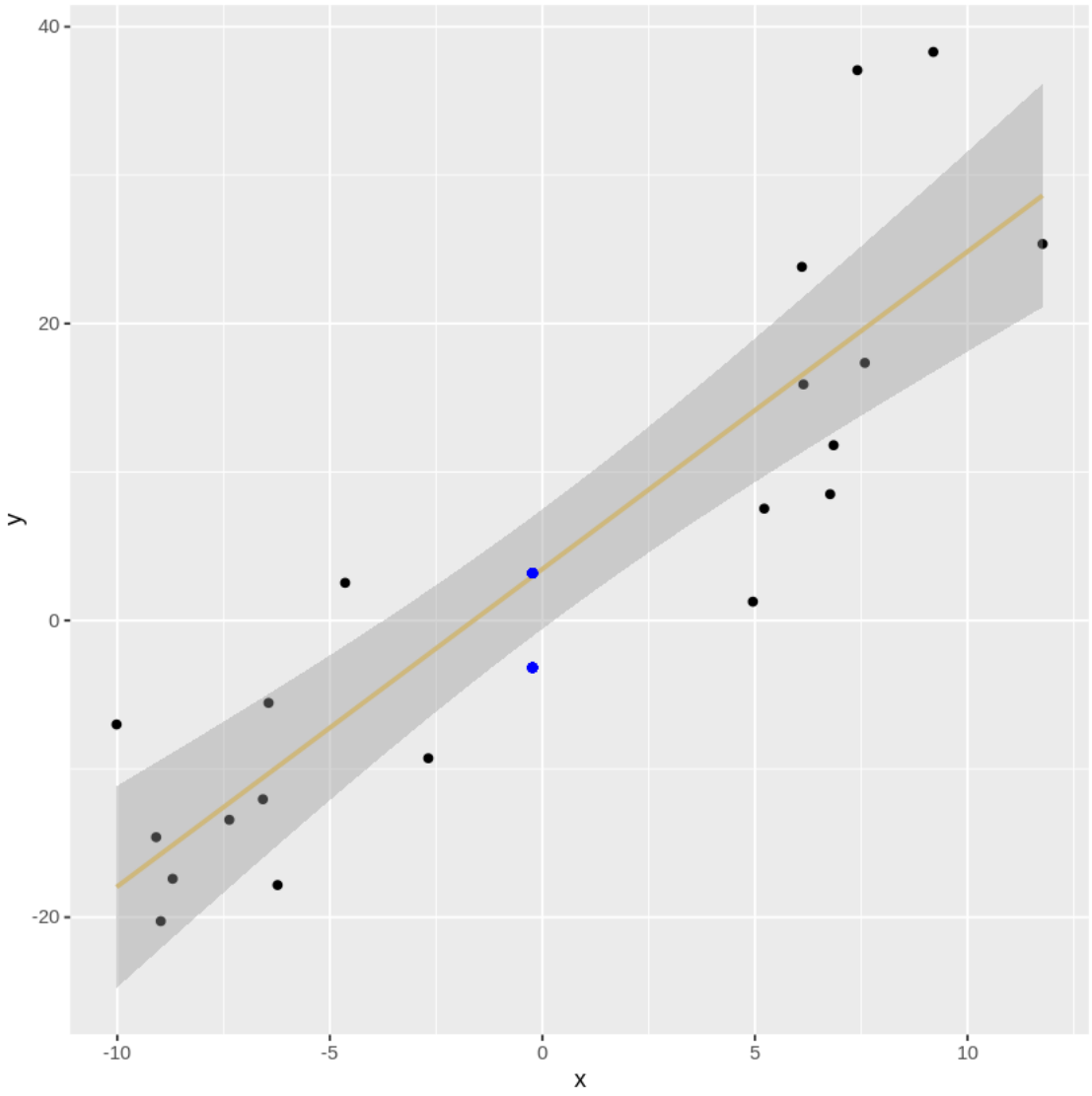
A data.frame: 1×1

x
<dbl>
0.06497868

A matrix: 1×3 of type dbl

	fit	lwr	upr
1	3.599795	-0.4308726	7.630464

```
`geom_smooth()` using formula 'y ~ x'
```



$x_k = \bar{x}$ will get us the Cl with a shorter width

3. (c) Interviewing the Intervals Recall that the Prediction Interval, when the predictor value is x_k, is defined as:

$$\hat{y}_h \pm t_{\alpha/2,n-2}\sqrt{MSE\Big(1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x_i - \bar{x})}\Big)}$$

Does the "width" of the Prediction Interval change at different population values? Explain why or why not.

The width of the Prediction Interval remains constant regardless of population values due to factors like confidence level, sample size, and data variability. The term involving the specific predictor

value's distance from the mean only affects the width if it varies significantly among predictors. Thus, as long as confidence level, sample size, and data variability are consistent, the width remains unchanged irrespective of population values.

3.1 Problem 4: Causality

Please answer the following three questions. Each answer should be clearly labeled, and a few sentences to a paragraph long.

1. In your own words, describe the fundamental problem of causal inference. How is this problem related to the counterfactual definition of causality?
2. Describe the use of "close substitutes" as a solution to the fundamental problem of causal inference. How does this solve the problem?
3. What is the difference between a *deterministic* theory of causality and a *probabilistic* theory of causality?

1. The fundamental problem of causal inference is that we can only observe one potential outcome per experimental unit, making it impossible to directly measure causal effects. This challenges the empiricist approach that relies on observations to generate knowledge.
2. In the absence of counterfactual observations, we cannot directly measure responses in alternative scenarios. Instead, we look for close substitutes that resemble the individual in question to estimate the response.
3. A deterministic theory of causality asserts that the effect inevitably follows from the cause, such as an object moving a certain distance when struck. In contrast, a probabilistic theory acknowledges that the existence of a cause influences the probability of an effect, like smoking increasing the likelihood of developing cancer.

3.2 Problem 5: Causal inference and ethics

How we think about causality, and the statistical models that we use to learn about causal relationships, have ethical implications. The goal of this problem is to invite you to think through some of those issues and implications.

Statisticians, data scientists, researchers, etc., are not in agreement on the best ways to study and analyze important social problems, such as racial discrimination in the criminal justice system. Lily Hu, a PhD candidate in applied math and philosophy at Harvard, wrote that disagreements about how to best study these problems "well illustrate how the nuts and bolts of causal inference…about the quantitative ventures to compute 'effects of race'…feature a slurry of theoretical, empirical, and normative reasoning that is often displaced into debates about purely technical matters in methodology."

Here are some resources that enter into or comment on this debate:

1. Statistical controversy on estimating racial bias in the criminal justice system
2. Can Racial Bias in Policing Be Credibly Estimated Using Data Contaminated by Post-Treatment Selection?

3. A Causal Framework for Observational Studies of Discrimination

Please read Lily Hu's blog post and Andrew Gelman's blog post "Statistical controversy on estimating racial bias in the criminal justice system" (and feel free to continue on with the other two papers!) to familiarize yourself with some of the issues in this debate. Then, write a short essay (300-500 words) summarizing this debate. Some important items to consider:

1. How does the "fundamental problem of causal inference" play out in these discussions?
2. What are some "possible distortionary effect[s] of using arrest data from administrative police records to measure causal effects of race"?
3. What role do assumptions (both statistical and otherwise) play in this debate? To what extent are assumptions made by different researchers falsifiable?

1. todo
2.

- Selection bias: Arrest data may not capture the full extent of racial bias in the criminal justice system due to selection biases. Law enforcement practices, such as biased policing or targeting certain communities, can result in disproportionate arrests of individuals from specific racial or ethnic backgrounds. This biased selection can lead to an overrepresentation of certain racial groups in the arrest data, potentially distorting the estimation of causal effects.
- Differential reporting: Racial bias can influence how crimes are reported, investigated, and ultimately reflected in arrest records. Biases at various stages of the criminal justice process, such as differential reporting of offenses or biased victim perceptions, can skew the representation of racial disparities in arrest data.
- Discretionary decision-making: Police officers exercise discretion in making arrest decisions, which can be influenced by implicit biases and racial stereotypes. These discretionary decisions may lead to differential treatment of individuals based on their race, resulting in variations in arrest rates among different racial groups. If the causal effects of race are estimated solely based on these discretionary arrest decisions, it may not capture the true extent of racial bias in the criminal justice system.
- Contextual factors: Arrest data alone may not adequately capture the broader contextual factors that contribute to racial disparities in the criminal justice system. Socioeconomic factors, neighborhood characteristics, systemic biases, and historical inequalities can influence both the likelihood of arrest and the subsequent outcomes within the system. Failing to account for these contextual factors can lead to an incomplete understanding of the causal effects of race.

3. Assumptions, both statistical and social, play a crucial role in the debate on racial bias in the criminal justice system. Statistical assumptions guide the modeling and analysis techniques used by researchers, while social assumptions involve theoretical frameworks and interpretations of causality. These assumptions can vary among researchers, leading to differing findings and interpretations. Statistical assumptions can be tested and evaluated, but social assumptions are often harder to definitively falsify due to the complexities of social phenomena. Assumptions shape the entire research process and influence methodologies, findings, and

discussions on racial bias. Maintaining transparency, robustness, and openness to critique is important for advancing our understanding of this complex issue.

[ ]: