C1M4_peer_reviewed

May 29, 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).

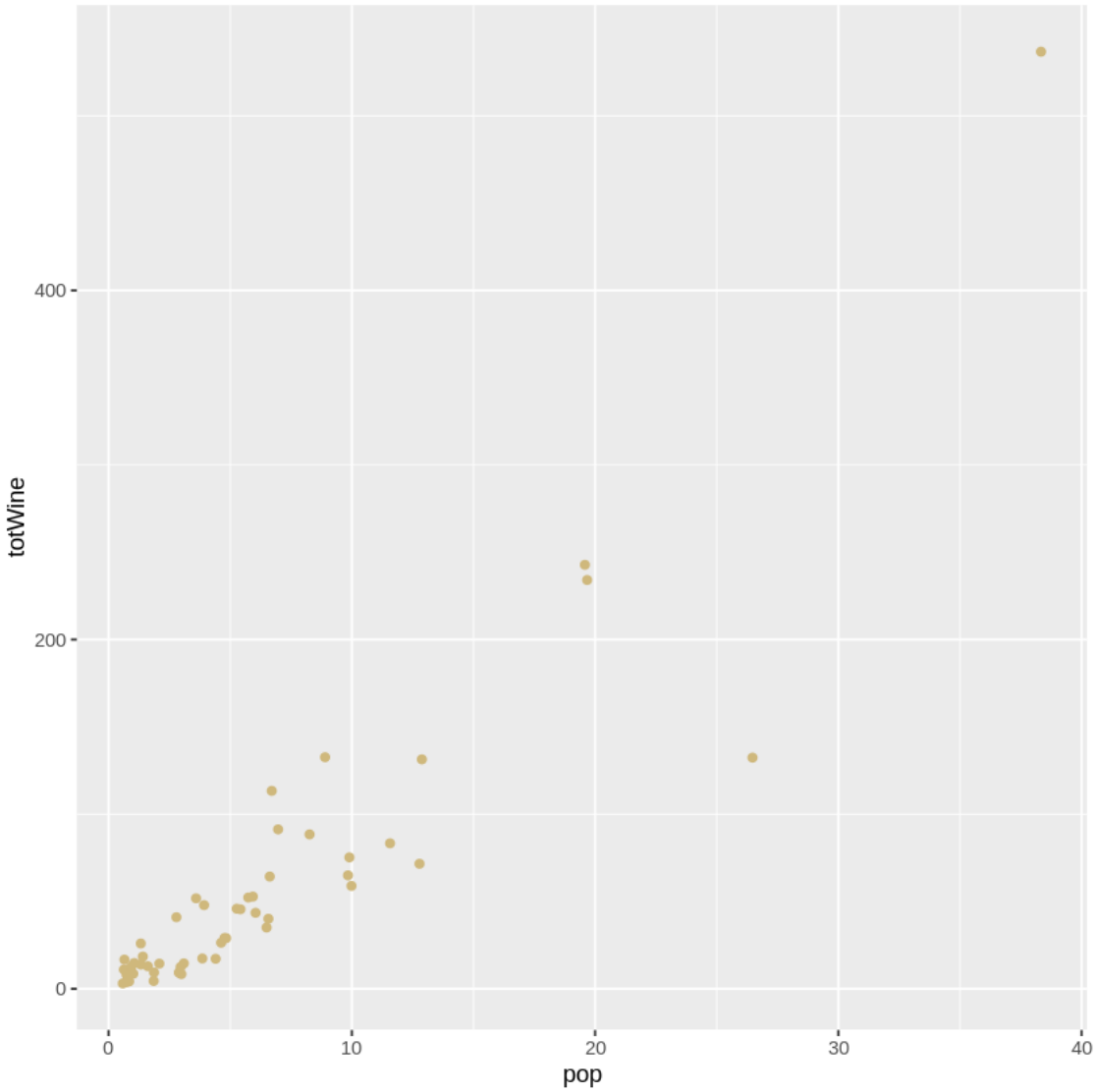
1. (a) Initial Inspections Load in the data and create a scatterplot with `population` on the x-axis and `totWine` on the y-axis. For fun, set the color of the point to be `#CFB87C`.

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


wine_plot <- ggplot(data = wine.data, aes(x=pop, y=totWine)) +
                geom_point(color='#CFB87C')
wine_plot
```

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

wine_lm = lm(totWine ~ pop, data = wine.data)

wine_plot <- wine_plot + geom_smooth(method='lm', col='#CFB87C', level = 0.90)
wine_plot
```

```
single_point <- wine.data[5, ]
single_point

wine_conf <- predict(wine_lm, new = single_point, level = 0.90, interval =␣
↪"confidence")
wine_conf
```

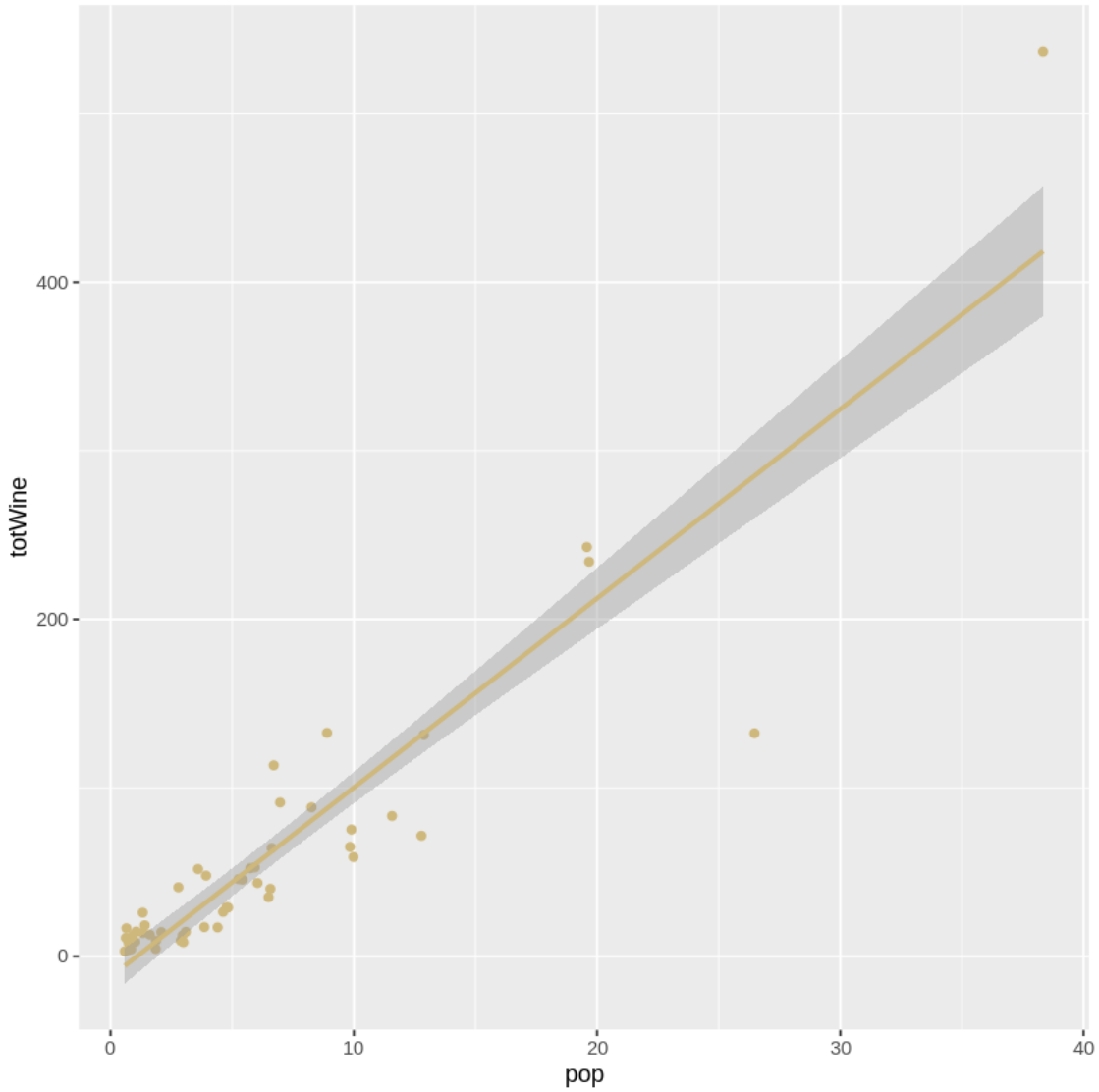
```
`geom_smooth()` using formula 'y ~ x'
```

A data.frame: 1 × 4

	State	pcWine	pop	totWine
	<fct>	<dbl>	<dbl>	<dbl>
5	California	14	38.3352	536.6928

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
5	418.2201	379.6794	456.7608



I chose California as my data point. the best point estimate was 418.2201 million liters of wine. The 90% confidence interval (379.6794,456.7608) means a 90% certainty that a new US state with a population of 38.3352 million would expect to consume somewhere between 379.6794 million and 456.7608 million liters of wine on average over many resamplings.

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.

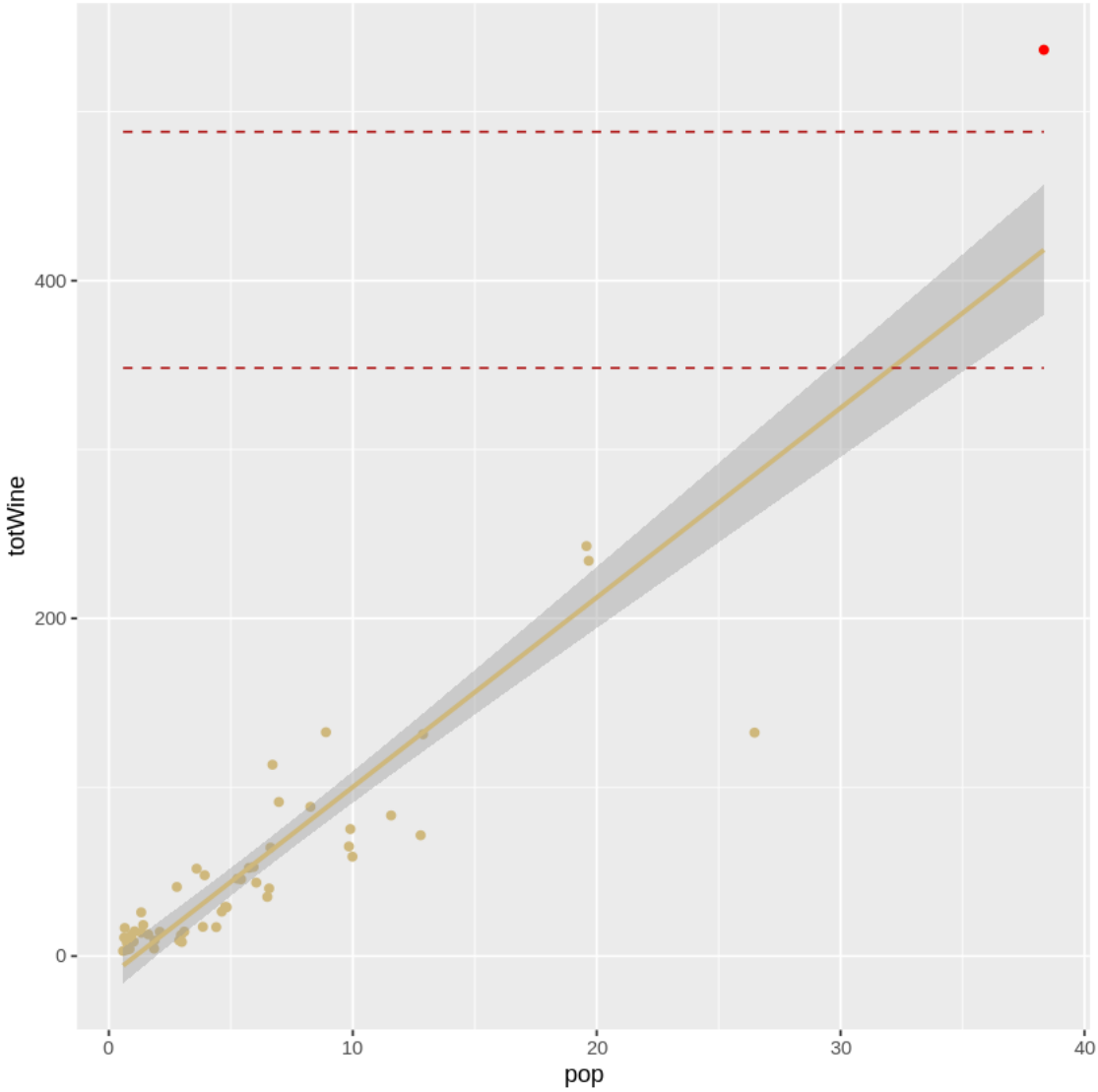
[4]:
```
# Your Code Here
wine_conf <- predict(wine_lm, new = single_point, level = 0.90, interval =␣
↪"prediction")
wine_conf
```

```
pred_plot <- ggplot(data = wine.data, aes(x=pop, y=totWine)) +
                geom_point(color='#CFB87C') +
                geom_point(data = wine.data %>% filter(State == "California"),␣
 ↪color = "red") +
                geom_smooth(method='lm', col='#CFB87C', level = 0.90) +
                geom_line(aes(y=wine_conf[ , 2]), color = "firebrick", linetype␣
 ↪= "dashed")+
                geom_line(aes(y=wine_conf[ , 3]), color = "firebrick", linetype␣
 ↪= "dashed")
pred_plot
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
5	418.2201	348.302	488.1383

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



For California (shown in red upper right), the point estimate is 418.2201 million liters of wine consumed, which is very very high. The 90% prediction interval for the data point is (348.302, 488.1383) indicated by the red dashed lines. This means that there is a 90% certainty that a new US state with a population of 38.3352 million would expect to consume somewhere between 348.302 million and 488.1383 million liters of wine for a single sample of that state.

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!

[5]:
```
# Your Code Here

e = resid(wine_lm)
fit = fitted(wine_lm)

Q_1 <- sum(e)
cat(round(Q_1,8), "\n")

Q_3 <- sum(e * wine.data$pop)
cat(round(Q_3,8), "\n")

Q_4 <- sum(e * fit)
cat(round(Q_4,8), "\n")


#part_5
left <- wine_lm$coefficients[1]
right <- mean(wine.data$totWine) - wine_lm$coefficients[2] * mean(wine.data$pop)
cat(round(right, 8) == round(left,8))
```

```
0
```

```
0
0
TRUE
```

The squared sum of residuals $\sum \hat{\varepsilon}_i^2$ should logically be minimized because the line of best fit should try to be as close to the data points as possible. If there existed a line that was closer to the data points, then that line would 'fit' the data better and be the line of best fit instead.

The simple linear regression line must pass through the point $(\bar{x}, \bar{y})$ because the OLS parameter estimate for the intercept of the line of best fit is $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$.

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?

The wine drinking data is an observational study of how much each state naturally drank and what the pre-existing population of that state was. There was no way for the researchers to control the population or to force entire states to drink more or less wine, so it cannot be an experiment. Causality cannot be inferred from observational studies, only correlation.

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]:
```r
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

[8]:
```
# Your Code Here

sample <- gen_data(-8, 50, 5, 20)  # change mean and variance here

lm_gen_data <- lm(y ~ x, data = sample)
summary(lm_gen_data)

temp <- predict(lm_gen_data, interval = "prediction", level = 0.95)
full_df <- cbind(sample, temp)

gen_plot <- ggplot(data = full_df, aes(x=x, y=y)) +
                geom_point(color='blue') +
                geom_smooth(method='lm', col='#CFB87C', level = 0.95) +
                geom_line(aes(y=lwr), color = "firebrick", linetype = "dashed")+
                geom_line(aes(y=upr), color = "firebrick", linetype = "dashed")
gen_plot
```

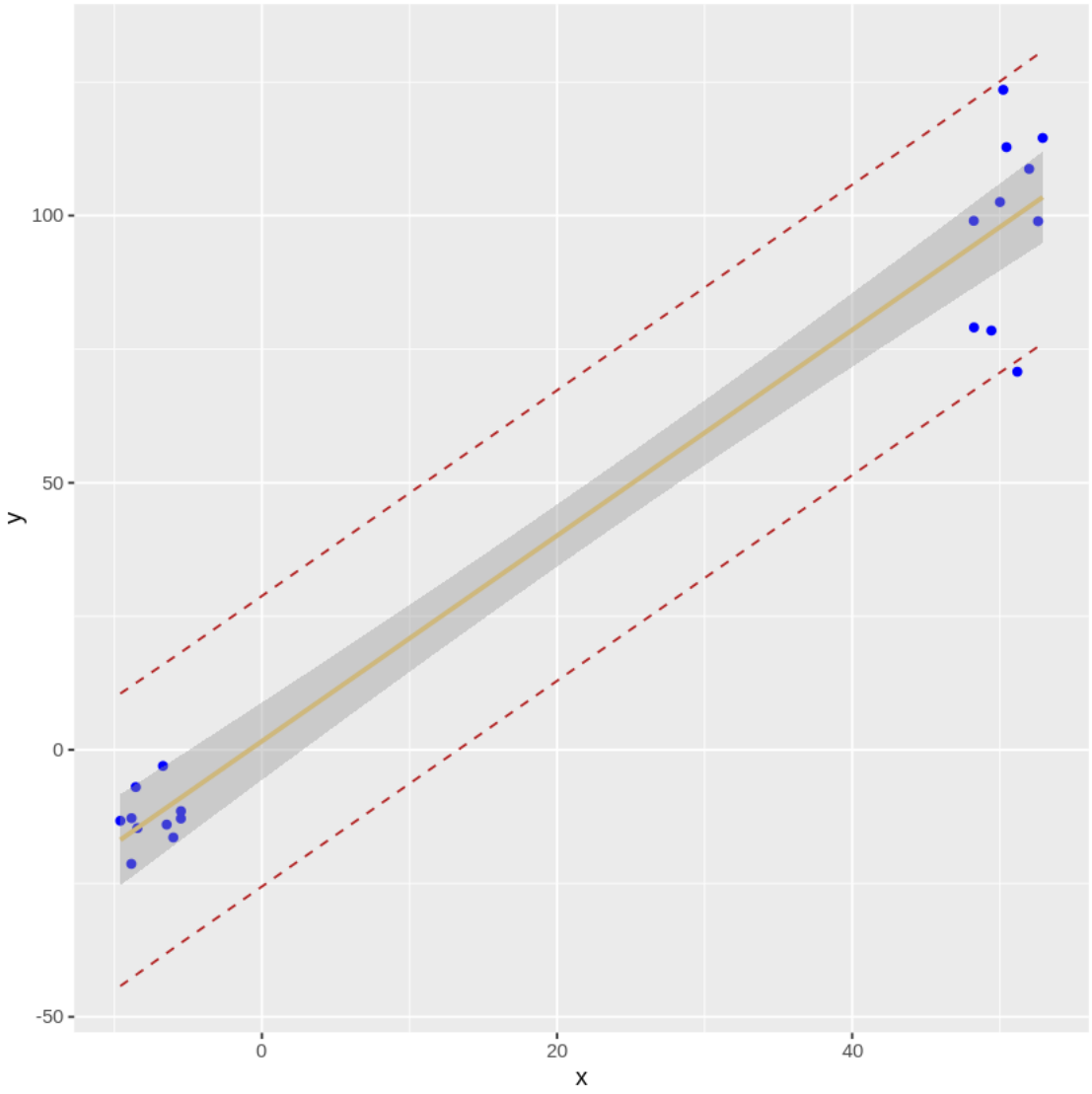
```
Call:
lm(formula = y ~ x, data = sample)

Residuals:
    Min      1Q  Median      3Q     Max
-29.357  -4.423   1.287   7.267  25.250

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.61857    3.44780   0.469    0.644
x            1.92449    0.09537  20.178 8.26e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.38 on 18 degrees of freedom
Multiple R-squared:  0.9577,Adjusted R-squared:  0.9553
F-statistic: 407.2 on 1 and 18 DF,  p-value: 8.263e-14
```

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



Changing the variance changes the width of the the CI and the PI because they both rely on the value of the standard error.

Increasing or decreasing the variance will increase or decrease the width of the interval.

Changing the mean will have no effect on the CI or PI.

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

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

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 $MSE \times \Big(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})}\Big)$ 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?

When x_k is equal to $\bar{x}$, the entire second term in the parentheses becomes 0, which minimizes the term and minimizes the entire interval. This matches up with the simulated data. It is good When data points are exactly the mean because that is the place where you would expect points to be on average.

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.

Yes, the width of the prediction interval is dependent on x_k in the formula. Since population would be x_k in this scenario, the population value for each state would change the width of the prediction.

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 there is only a singular outcome that occurs and therefore we cannot know the impact of other potential outcomes. The counterfactual theory of causality states that "event A causes event B if in the absence of event A, event B would be less likely to occur or not occur at all." The fundamental problem of causal inference means that we cannot easily find out what would have happened if an event that

occurred had instead not occurred, because we can only see one outcome in our true reality. We are not able to collect real data on counterfactual scenarios.

2. Since we cannot gather direct evidence on counterfactual scenarios, the best we can do is find another situation which is as similar as possible but with a one key difference, like a single predictor or event. This allows us to compare situation A to situation A', where A' is a close substitute for A. With two identical backgrounds in every other sense, we can investigate the effect of the one difference across multiple different outcomes.

3. The deterministic theory of causality mandates that if event A causes event B, then the occurrence of event A will always cause event B to occur. There is no uncertainty, events either happen or they don't. If outcomes appear to be stochastic, that is simply an indication that there are more variables at play that have not been taken into account for the model. The probabilistic theory of causality merely says that a cause only has to impact the chances of outcome, but need not necessarily manifest it. For example, increasing the amount of radiation one's body is exposed to will increase the odds that one gets melanoma, but does not guarantee it.

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?

There seem to be two main sides in the debate over applying causal inference to racial data related to criminal justice and its various constructs that exert influence over the public, such as the police force. One group seems to be making the claim that it might not be an issue to perform standard regression on data in a certain point in the justice system becuse there are multiple sources of bias and they could even out to being approximately unbiased. Lily Hu mentions in her blog post that one potential possible scenario could be that one source of bias is that the police are more likely to stop Black people for less severe crimes. This source of bias could be countered by another source of bias, such as encounters between the police and a Black civilian are more likely to become confrontational. The decreased average crime severity combined with the increased average threat level could theoretically mellow each other out to effectively have a non-biased situation. The other main party in the debate says that it is statistically unsound to estimate regression parameters for causal inference when there are sources of upstream bias. For example, using arrest data from administrative police records could easily have bias from the choice of which civilians to stop impacting the set of available potential civilians to arrest. This could distort the causal affects estimated in the arrest records. One overarching issue with settling the debate of whether or not certain methodology violates the conditions of causal inference is that data can never be gathered on the counterfactual scenario, the so-called "fundamental problem of causal inference". For example, if we believe there is bias in an upstream point in the system, there is no way to gather data for the same exact situation in which the bias is removed. This also makes it difficult to falsify claims about bias sources cancelling each other out. For a civilian that was never stopped by the police but was simply being "observed", there would be no way to observe the scenario in which the civilian had actually been stopped and gather data on the outcome. Hu mentions that because causal inference requires a certain set of assumptions on the part of the researcher, the ability to claim that the causal estimates are valid depends heavily on the researcher's own views. Namely, Hu writes "whether policing is racially discriminatory depends on one's prior views about which differences across racial groups are the 'relevant' differences that do and do not 'justify' differential police treatment."

[ ]: