C1M4_peer_reviewed

January 10, 2022

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.

[57]:

```r
# This cell loads the necesary libraries for this assignment
library(tidyverse)
library(ggplot2)
```

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).

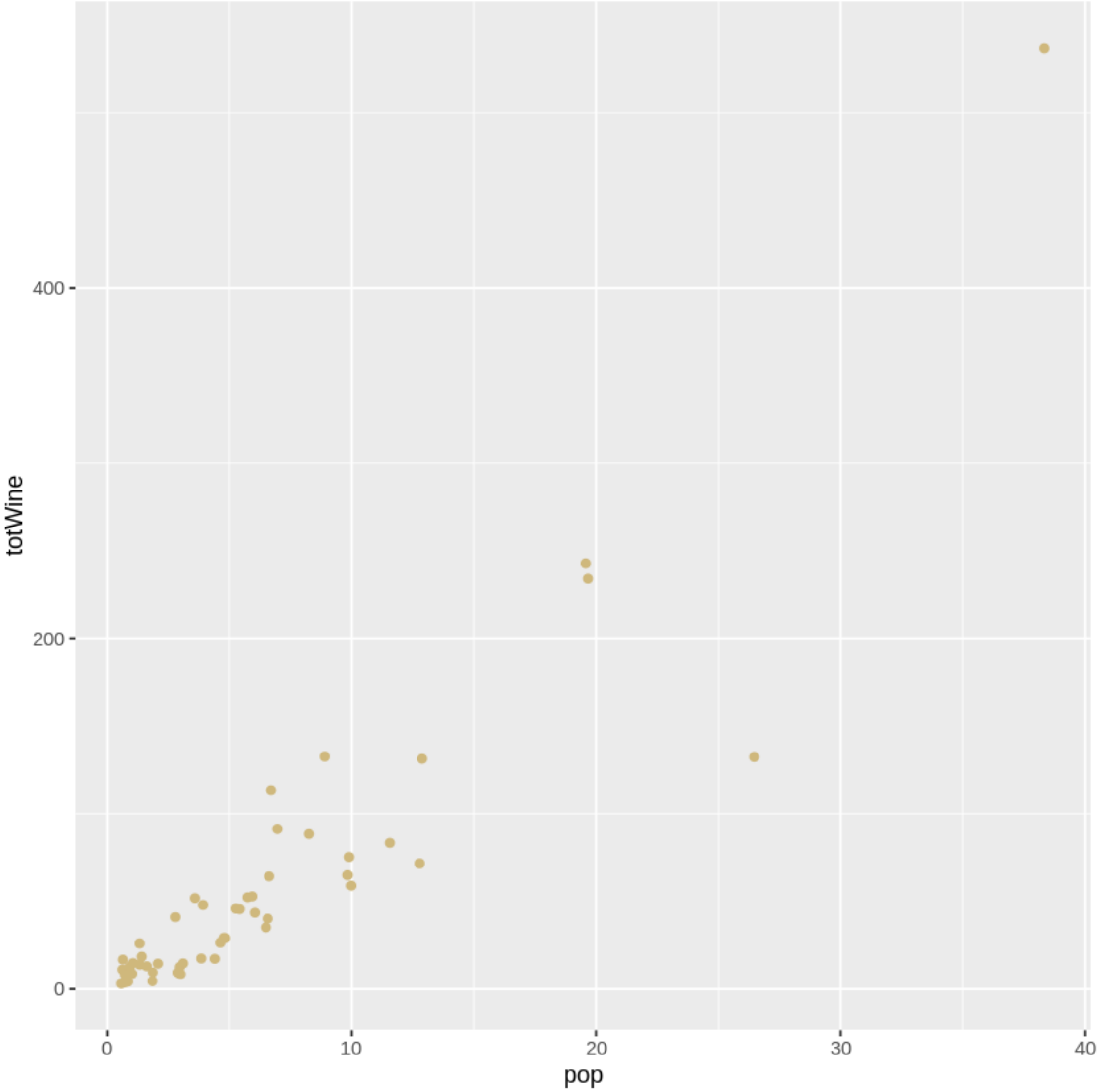
1. (a) Initial Inspections Load in the data and create a scatterplot with `population` on the x-axis and `totWine` on the y-axis. For fun, set the color of the point to be `#CFB87C`.

[58]:

```r
# Load the data
wine.data = read.csv("wine_state_2013.csv")
head(wine.data)

# Your Code Here
wine_scat = ggplot(wine.data, aes(x=pop, y=totWine)) +
    geom_point(color="#CFB87C")

wine_scat
```

A data.frame: 6 × 4

	State	pcWine	pop	totWine
	<fct>	<dbl>	<dbl>	<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



[59]:

```
dim(wine.data)
```

1. 51 2. 4

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.

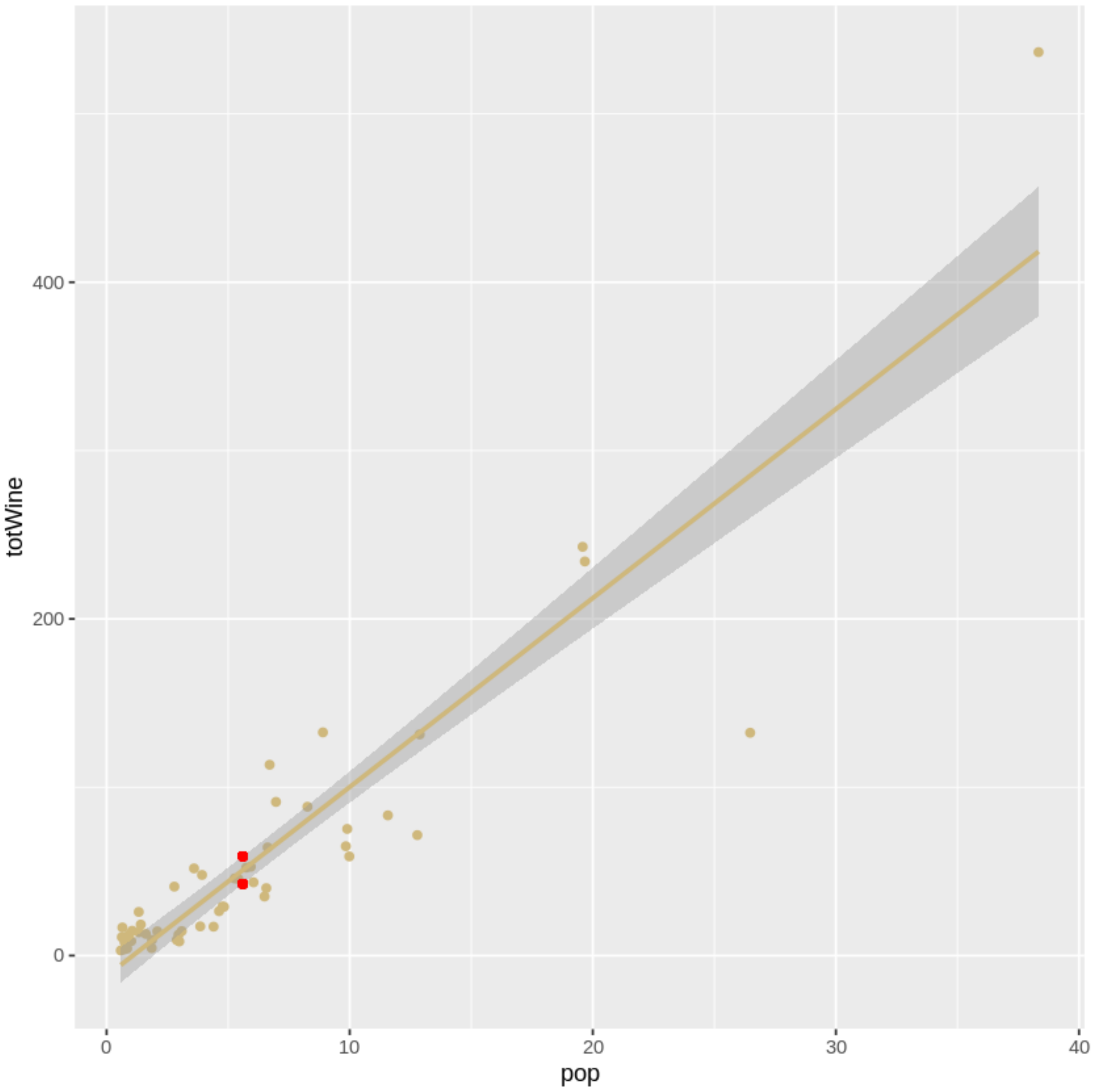
[77]:

```
# Your Code Here
wine_lm = lm(totWine~pop, data= wine.data)

x1 = data.frame(pop=5.6)
wine_conf= predict(wine_lm, new=x1, interval='confidence', level=0.90)
wine_conf

wine_scat +
    geom_smooth(method=lm, col="#CFB87C", level=0.9) +
    geom_point(aes(x=x1$pop, y=wine_conf[1,2]), color="red") +
    geom_point(aes(x=x1$pop, y=wine_conf[1,3]), color="red")
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	50.7431	42.54412	58.94208

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



After creating many samples and fitting a new regression line, each time, with the same popluation values, 90% of the regression lines would fall within our lower and upper bounds of the confidence interval of (42.54412 and 58.94208).

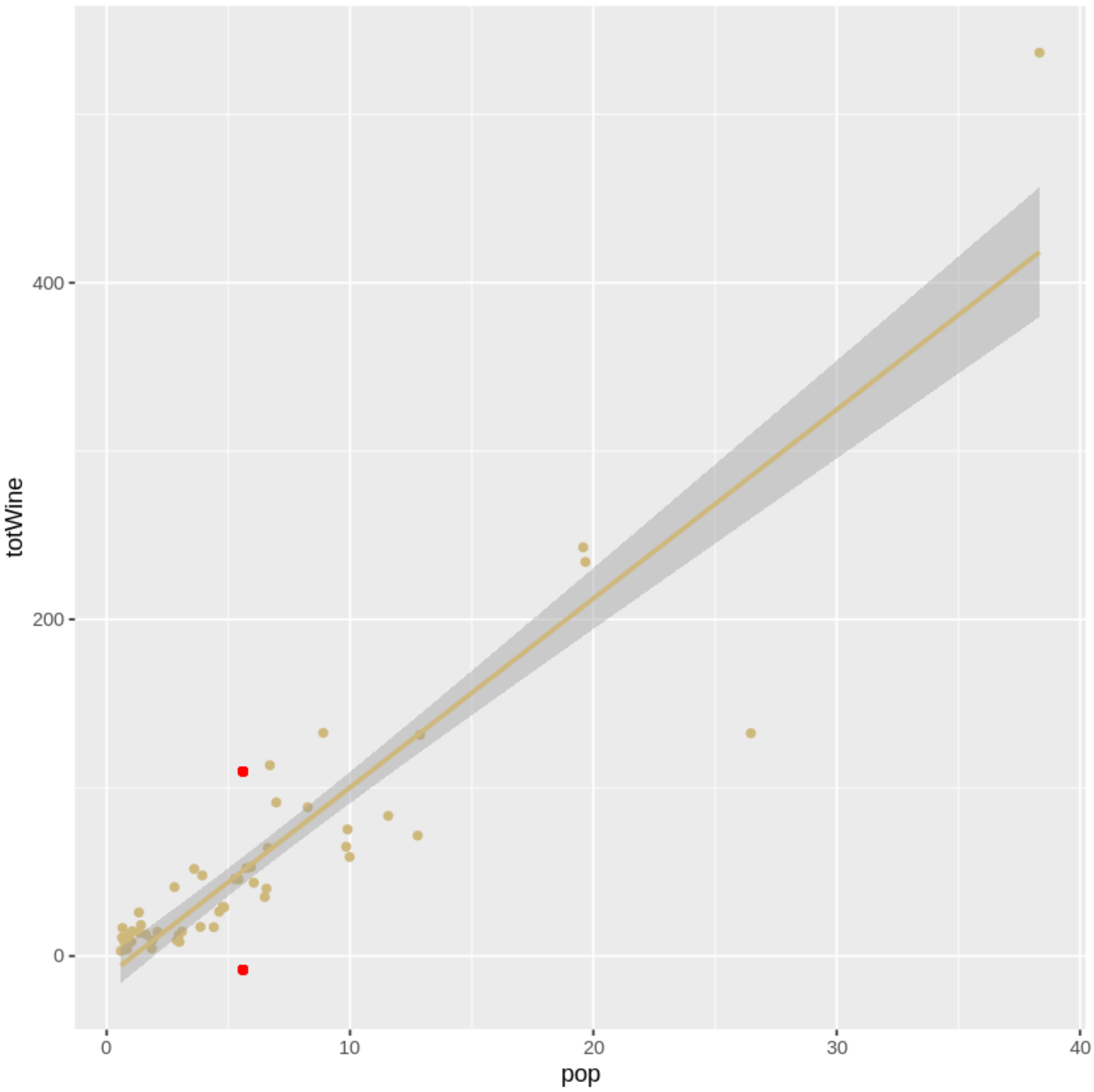
1. (c) Prediction Intervals Using the same `pop` point-value as in **1.b**, plot the prediction interval end points. In words, explain what this interval means for that data point.

[78]:
```
# Your Code Here
wine_pred= predict(wine_lm, new=x1, interval='prediction', level=0.90)
wine_pred

wine_scat +
    geom_smooth(method=lm, col="#CFB87C", level=0.9) +
    geom_point(aes(x=x1$pop, y=wine_pred[1,2]), color="red") +
    geom_point(aes(x=x1$pop, y=wine_pred[1,3]), color="red")
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	50.7431	-8.166836	109.653

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



The prediction interval, 90% of the time, will contain the true value of total wine consumption for 5.6 million people.

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!

[83]:

```
# Your Code Here
#1
res = resid(wine_lm)
sum(res)

#3
sum(wine.data['pop'] * res)

#4
y_hat = fitted(wine_lm)
sum(y_hat * res)

#5
x_mean = data.frame(pop=mean(wine.data$pop))
x_mean
predict(wine_lm, newdata=x_mean) - mean(wine.data$totWine)
```

```
-2.00672811700997e-14

-1.11632925126059e-12

-7.65254526413628e-12

                            pop
A data.frame: 1 × 1      <dbl>
                       6.200096

1: 7.105427357601e-15
```

For 2, least squares regression minimizes teh the sum of squares of the residuals as the best model. 2 exemplifies that state. This is very similar to minimizing the variance.

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? Our wine drinking data cam from an observational study. We cannot infer causation between population and the amount of wine drank from these data because there may be underlying factors that we haven't controlled for. No experiment was conducted.

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?

[63]:

```
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))
```

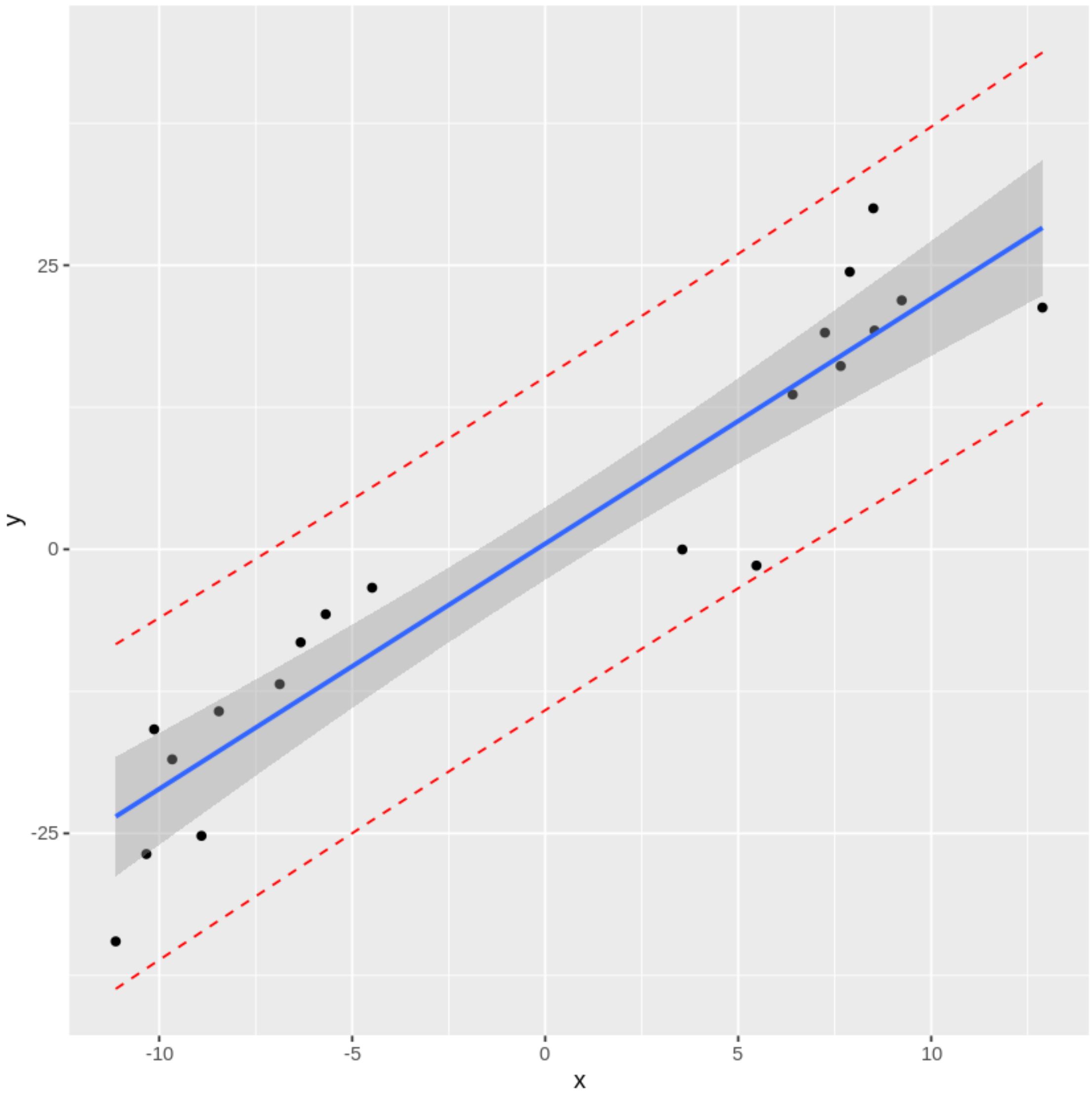
A data.frame: 6 × 2

	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

[64]:

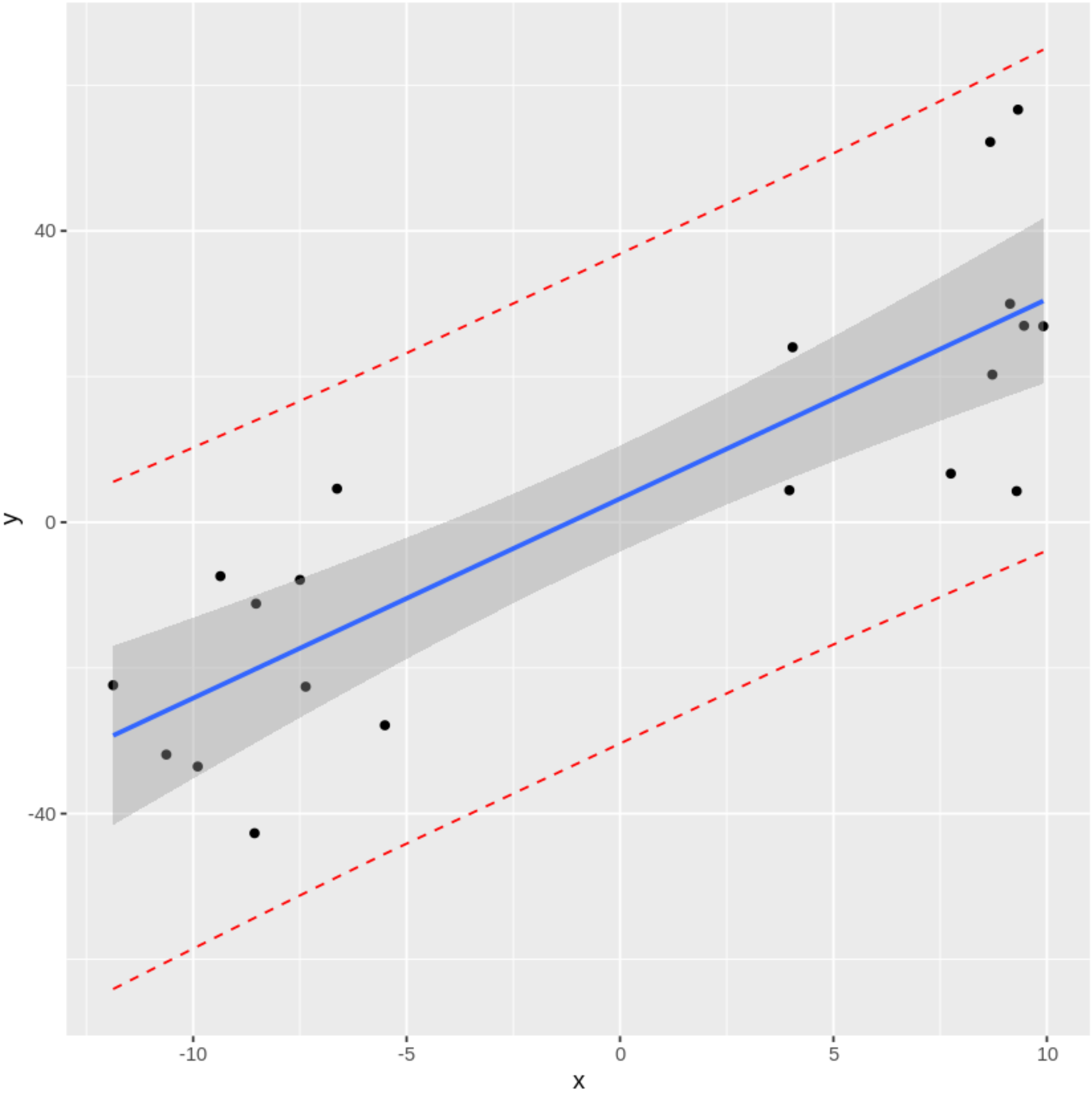
```
# Your Code Here
# Means -8, 8
# Vars 10, 10
viz.data1 = gen_data(-8,8, 10, 10)
viz.lm1 = lm(y~x, data=viz.data1)

viz.predict1 = predict(viz.lm1, new = viz.data1['x'], interval='prediction')
viz.data.fit = viz.predict1[ ,'fit']
viz.data.lower = viz.predict1[ ,'lwr']
viz.data.upper = viz.predict1[ ,'upr']

final.viz1 = cbind(viz.data1, viz.predict1)
ggplot(final.viz1, aes(x,y)) +
    geom_point() +
    geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
    geom_line(aes(y=upr), color = "red", linetype = "dashed")+
    geom_smooth(method=lm, se=TRUE)

set.seed(12049)
```

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



```
[65]: # Means -8, 8
      # Vars 20, 20
      viz.data2 = gen_data(-8,8, 20, 20)
      viz.lm2 = lm(y~x, data=viz.data2)

      viz.predict2 = predict(viz.lm2, new = viz.data2['x'], interval='prediction')
      viz.data.fit2 = viz.predict2[ ,'fit']
      viz.data.lower2 = viz.predict2[ ,'lwr']
      viz.data.upper2 = viz.predict2[ ,'upr']

      final.viz2 = cbind(viz.data2, viz.predict2)
      ggplot(final.viz2, aes(x,y)) +
          geom_point() +
          geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
          geom_line(aes(y=upr), color = "red", linetype = "dashed")+
          geom_smooth(method=lm, se=TRUE)

      set.seed(12050)
```

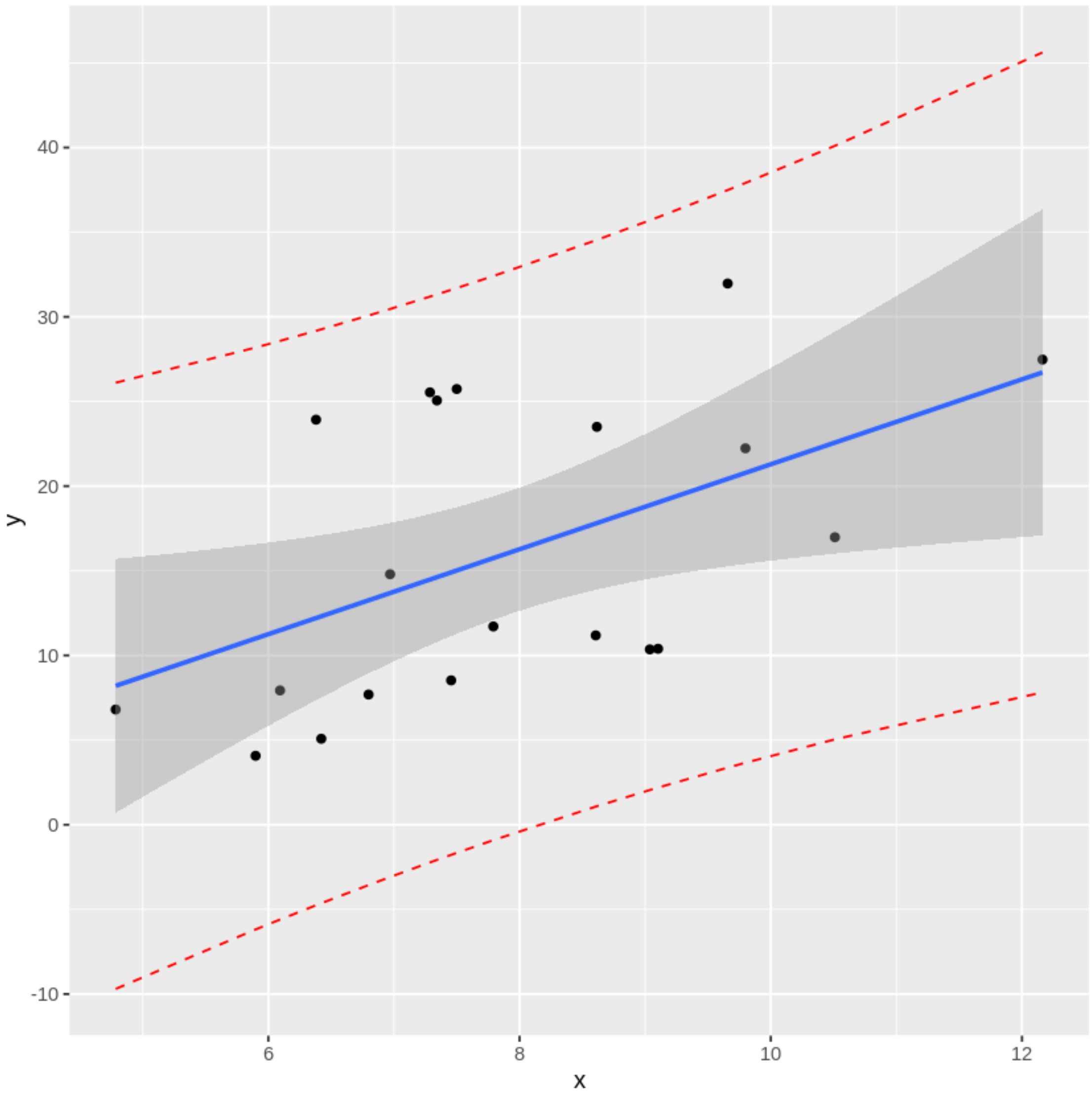
```
`geom_smooth()` using formula 'y ~ x'
```



[66]:

```
# Means 8, 8
# Vars 10, 10
viz.data3 = gen_data(8,8, 10, 10)
viz.lm3 = lm(y~x, data=viz.data3)
viz.predict3 = predict(viz.lm3, new = viz.data3['x'], interval='prediction')
viz.data.fit3 = viz.predict3[ ,'fit']
viz.data.lower3 = viz.predict3[ ,'lwr']
viz.data.upper3 = viz.predict3[ ,'upr']

final.viz3 = cbind(viz.data3, viz.predict3)
ggplot(final.viz3, aes(x,y)) +
    geom_point() +
    geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
    geom_line(aes(y=upr), color = "red", linetype = "dashed")+
    geom_smooth(method=lm, se=TRUE)

set.seed(12051)
```

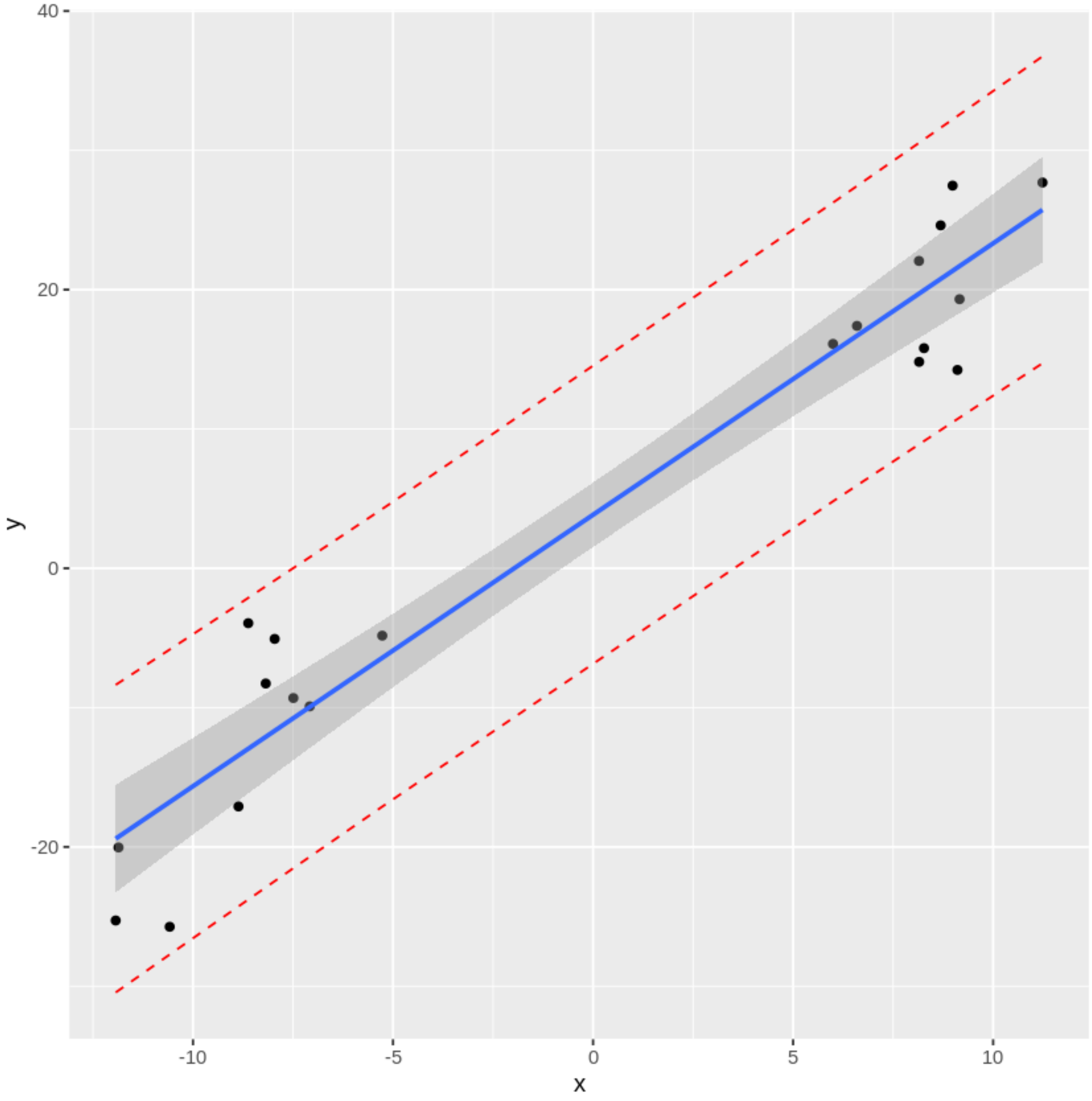
```
`geom_smooth()` using formula 'y ~ x'
```



[67]:

```
# Means -8, 8
# Vars 4, 4
viz.data4 = gen_data(-8,8, 4, 4)
viz.lm4 = lm(y~x, data=viz.data4)

viz.predict4 = predict(viz.lm4, new = viz.data4['x'], interval='prediction')
viz.data.fit4 = viz.predict4[ ,'fit']
viz.data.lower4 = viz.predict4[ ,'lwr']
viz.data.upper4 = viz.predict4[ ,'upr']

final.viz4 = cbind(viz.data4, viz.predict4)
ggplot(final.viz4, aes(x,y)) +
    geom_point() +
    geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
    geom_line(aes(y=upr), color = "red", linetype = "dashed")+
    geom_smooth(method=lm, se=TRUE)

set.seed(12052)
```

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



An increase in the variance widens both the prediction interval and the confidence interval. However, decreasing the variances decreases the prediction interval and the confidence interval.

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?

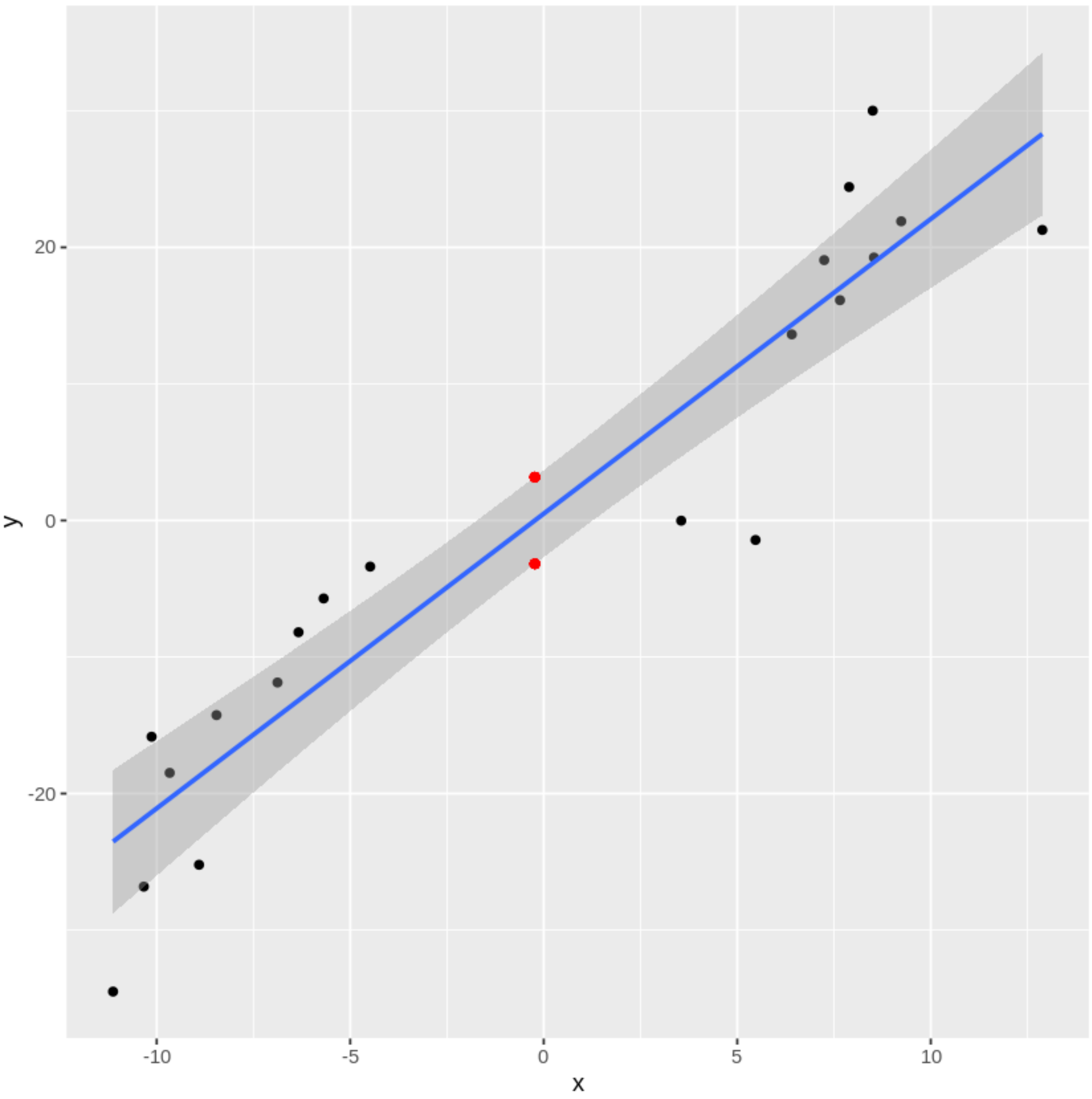
[96]:

```r
# Your Code Here
x_mean2 = data.frame(x=mean(viz.data1$x))
x_mean2
x_mean2_pred = predict(viz.lm1, newdata=x_mean2, interval="confidence")
x_mean2_pred

ggplot(viz.data1, aes(x=x, y=y)) +
    geom_point() +
    geom_smooth(method="lm") +
    geom_point(aes(x=x_mean2[1,1], y=x_mean2_pred[1,2]), color="red") +
    geom_point(aes(x=x_mean2[1,1], y=x_mean2_pred[1,3]), color="red")
```

```
                        x
A data.frame: 1 × 1   <dbl>
                     -0.2312037
```

```
                           fit            lwr        upr
A matrix: 1 × 3 of type dbl
                       1 | -0.0009917047  -3.181803  3.17982
```

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



The value would need to be equal to the mean of the x-values. Yes, it matches with the simulated data as shown above. This would occur because the regression line calculates the predicted Y by using the means of all the X's. The confidence interval is hte mean of predicted values. Therefore, the two values would be the same.

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 smaller the population, the wider the prediction interval. On the other hand, the larger the population, the shorter the prediction interval. Because n represents the population value, anbd it is in the demoniator of the equation, the larger it is, the smaller the standard error. Intuitively, a larger population/sample size would help us gather more information about the data.

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 an issue because we can observe many values or outcomes. However, we cannot directly measure causal effects for all outcomes for each observation. This is related to counterfactual definition of causality because we can't state causality to non-observed outcome. We don't know what would happen to that same unit if we tested it with the non observed outcome.
2. The 'close substitutes" use is a great solution to the problem of causal inference. Using close subsititutes we can evaluate different outcomes on observations that are relatively the same. It would be similar to observing an observation on the unit, reversing everthing we did, and then evaluating with the other observation. At the end of it, we would have information for both possible outcomes.
3. Deterministic theory of causality states that if A causes B then A will always be followed by B. On the other hand, B may not always follow A in probabilistic theory of causality. A probabilisticly causes B if it's occurence increases the probability of B.

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?

Lily Hu and Andrew Gelman both argue the validity, morality and ethics of both causal and statistical inference. Hu argues although law verdicts an social decisions are heavily based on data, neither the proper statistical asssumptions nor 'normative senstivity' are considered when evaluating; we cannot use statistical inference for the source of "normative innovation". Instead, it must be the beginning picture which we evaluate with other measures. Similarly, Gelman states the results are misleading when not carefully making decisions about the model.

First, the fundamental problem of causual inference is a large problem when discussing racial discrimination. Recently, there has been a large surge in the United States police reform due to racial bias. an argument states that black americans face harsher treatment with police than other races: in particular, white. The counter argument is the people arrested deserve what they are arrested for; there is no racial bias and the situation was handled poorly byt the person arrested. In this case, the fundamental problem of causual inference relies on the question of if the opposite race was in a particular situation, would it have the same result?

Second, both authors question the validity of using arrest data from police records to measure the causal effects of race. When developing experiments to measuere causal inference, we must carefully control for as many confounding variables as possible. However, when using arrest data from police records, we are unsure of what variables were controlled for. How was the data collected? Does the data accurately protray the people the policy observe without arresting? Does polce behavior change on varying races? None of these factors are measured within the arrest data. However, all of these factors can make an impact on whether the person was arrested or not.

Also, Hu argues one's personal beliefs creates too many differences between different people to develop consistent and valid models. Some people may justify some differences about race while others may find it problematic. This inconsisntency can create differing results among causality inference. This can be in the interpretation in the results or the predictors used to develop the inference. For example, Gelman argues that police and judges, when discussing racial bias, fail to control for neighborhoods. However, controlling for neighborhoods would be important as different neighorboods face different problems.

Third, as stated previously, one’s personal beliefs can result in differing models. The beliefs can result in different normative assumptions. However, Hu argues that statistical assumptions are better. Many believe in the assumption that ‘more data is better’; thus, they can use the Gaussian distribution. However, other statistical assumption are met. Some of the observations used are not indetically and independently distributed. Some arrests involve multiple people in the same case. This, again, can create racial bias. Are people of the same race more likely to be together?. Also, many fail to considere the correlation between predictors. Is whether a policy encounter violent independent of the race, location, etc.? These ignored assumptions can results in misleading statistical assumptions that can cause problems within the legal system.

In conclusion, regarding racial discrimaination, data is good to use to get an idea of what is occurring. However, to gain further insight, we will need to develop better methods.

[ ]: