C1M4_peer_reviewed

June 7, 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]:

```r
# 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

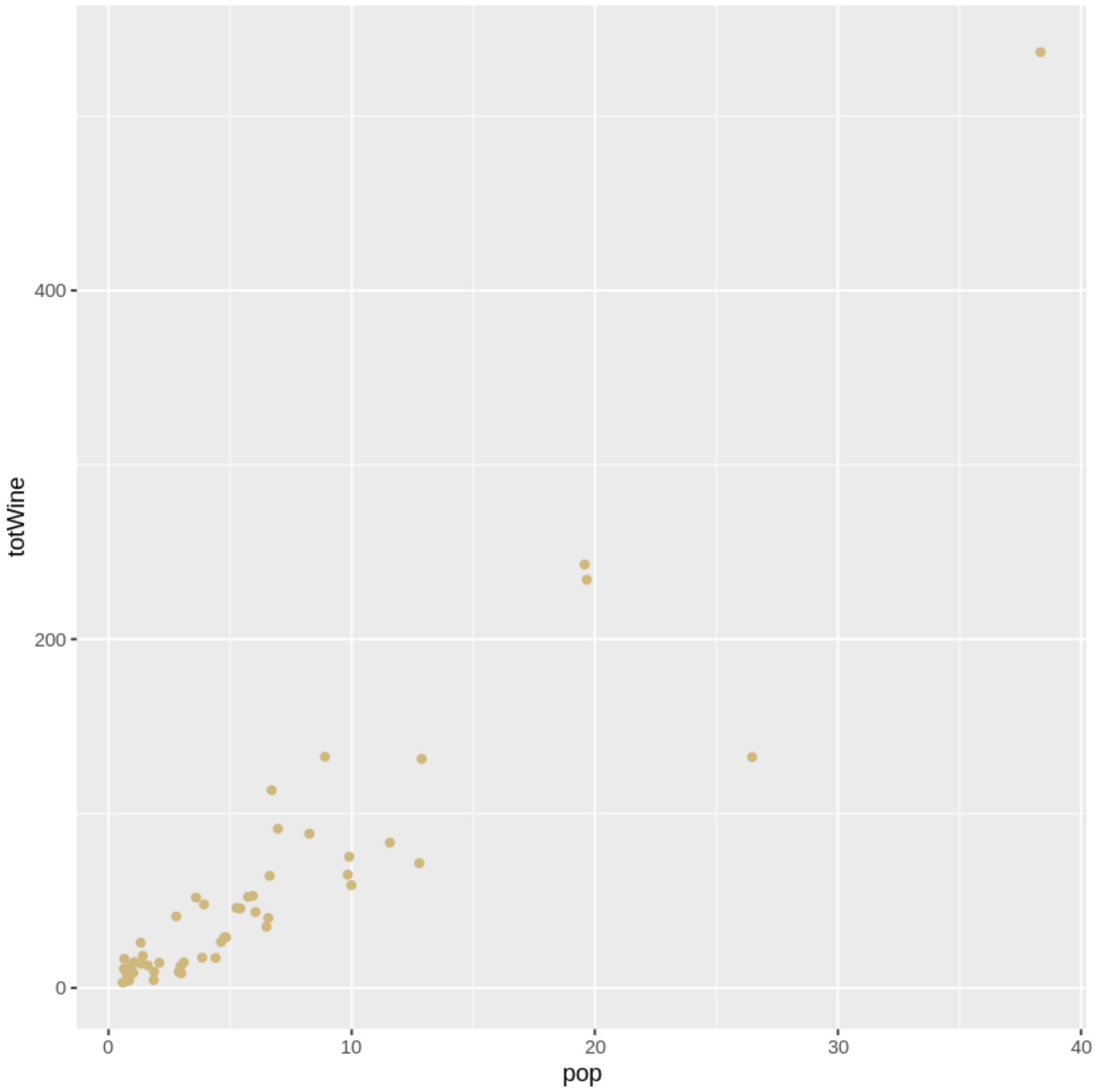
For this problem, we're going to practice creating and interpreting Confidence (Mean) Intervals and Prediction Intervals. To do so, we're going to use data in U.S. State Wine Consumption (millions of liters) and Population (millions).

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]:
```r
# Load the data
wine.data = read.csv("wine_state_2013.csv")
head(wine.data)
# Your Code Here
ggplot(wine.data, aes(x=pop,y=totWine)) + geom_point(col="#CFB87C")
```

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



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

mod <- lm(data=wine.data,totWine~pop)
ggplot(wine.data, aes(x=pop,y=totWine)) + geom_point(col="#CFB87C") +␣
↪stat_smooth(method = "lm", col="#CFB87C",level=0.9)
```

```r
wine.data[1,]
predict(mod, wine.data[1,], interval="confidence")
```

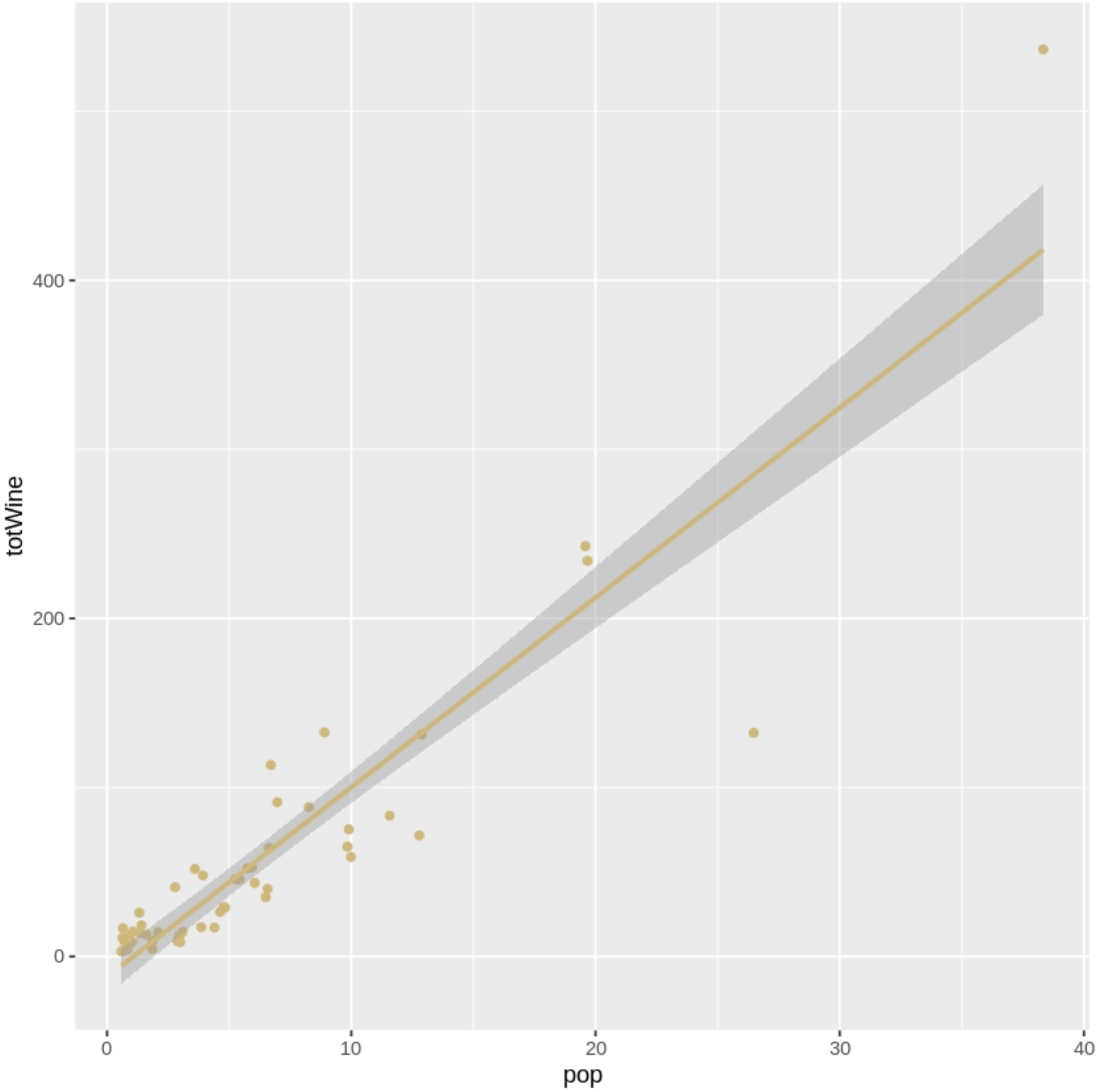
```
`geom_smooth()` using formula 'y ~ x'
```

A data.frame: 1 × 4

	State	pcWine	pop	totWine
	<fct>	<dbl>	<dbl>	<dbl>
1	Alabama	6	4.829479	28.97687

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	42.09342	32.11451	52.07234



If we were to repeatedly redo the sampling process from the same underlying population and computed the predicted value at pop=4.829479, then 95% of means of the corresponding predictions would be between 32.11451 and 52.07234.

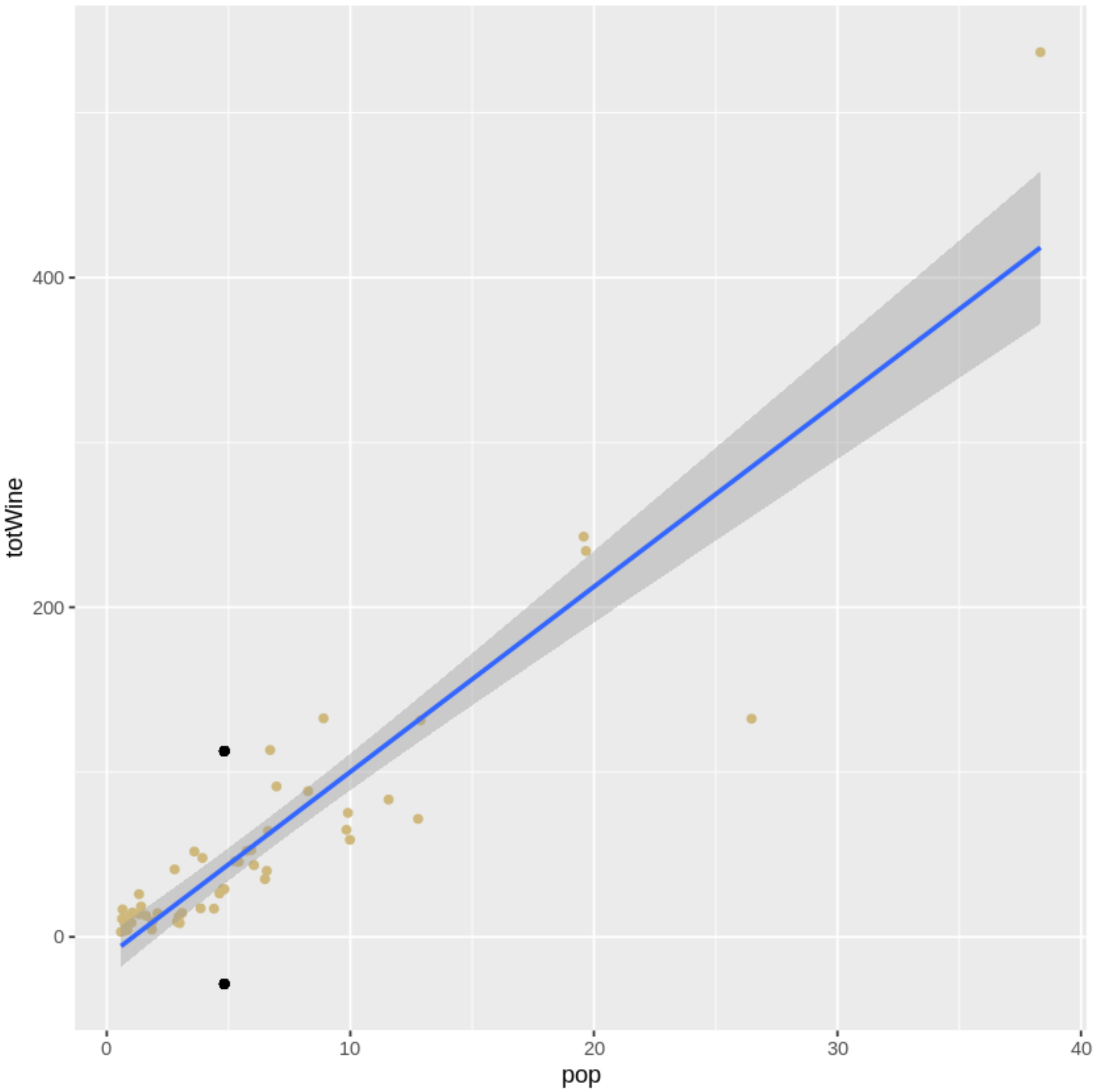
1. (c) Prediction Intervals Using the same `pop` point-value as in **1.b**, plot the prediction interval end points. In words, explain what this interval means for that data point.

```
[4]: # Your Code Here
pred_interval <- predict(mod, wine.data[1,], interval="prediction")
ggplot(wine.data, aes(x=pop,y=totWine)) + geom_point(col="#CFB87C") +␣
 ↪stat_smooth(method = "lm") +
geom_point(aes(x = wine.data[1,]$pop, y = pred_interval[2])) +
geom_point(aes(x = wine.data[1,]$pop, y = pred_interval[3]))
pred_interval
```

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

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	42.09342	-28.5394	112.7262



If we repeat the same sampling process a large number of times, then, approximately 95% of the predictions at pop=4.829479 will be between -28.5394 and 112.7262

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!

[31]:

```
# Your Code Here
# 1.
sum(wine.data$totWine - mod$fitted)

# 3.
sum(wine.data$pop*resid(mod))

# 4.
sum(mod$fitted*resid(mod))

# 5. the residual is zero for x = x_bar
mean(wine.data$totWine) - predict(mod,data.frame(pop = mean(wine.data$pop)))
```

1.68753899743024e-14

-1.11632925126059e-12

-7.65254526413628e-12

1: -7.105427357601e-15

2. the fit is the least squares line, i.e., when determining the best fitting line we are minimizing the sum of the squared residuals, therefore, the sum of the squared residuals is as small as it can be.

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 data come from an observational study. Although causation cannot be inferred in the strict sense, since this is not a randomized control trial, there is an obvious causal explanation in that the more people there are, the more wine will be drank in total.

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?

[7]:

```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))
```

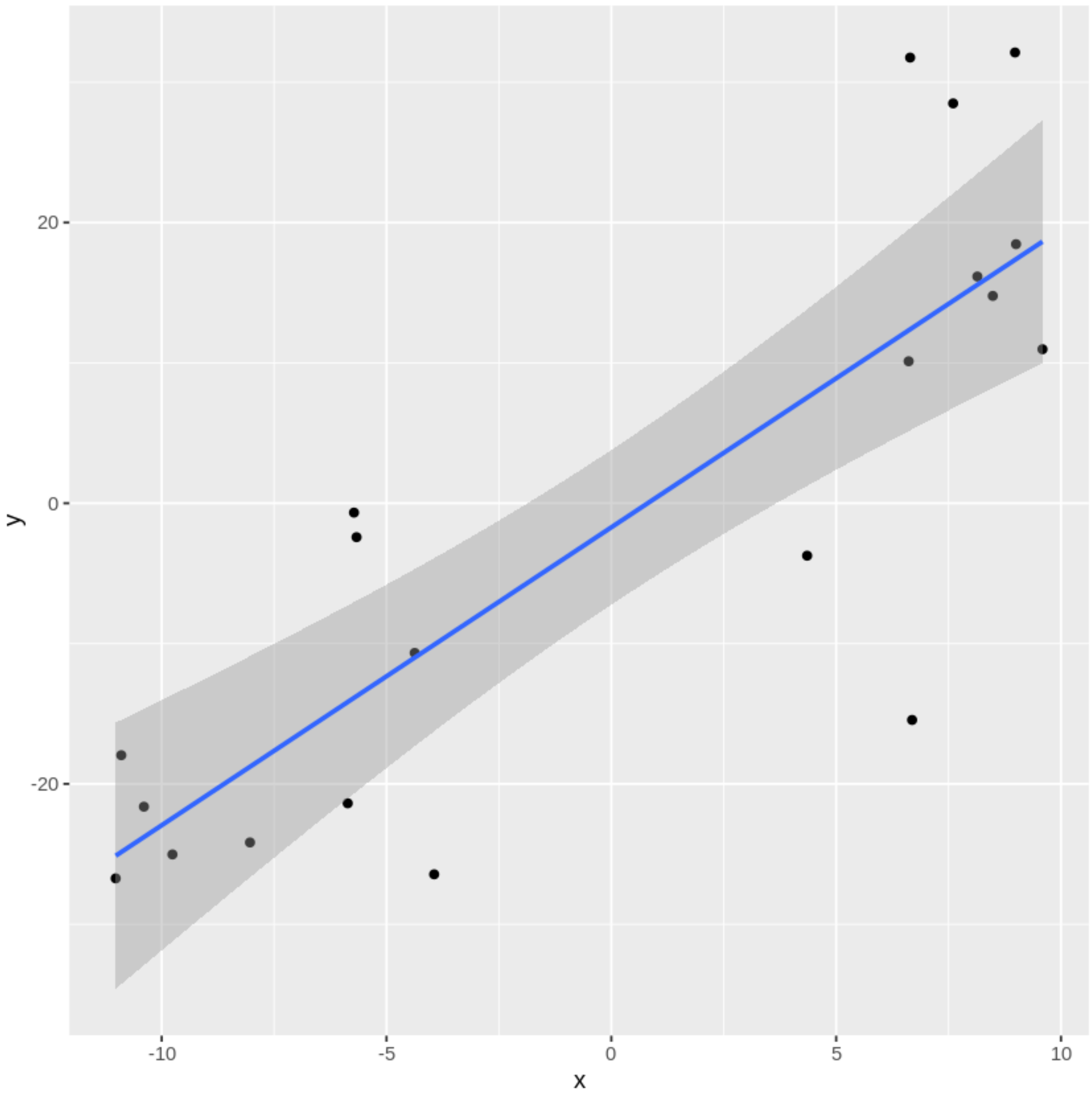
```
                           x            y
                           <dbl>        <dbl>
                        1  -5.474091    -11.1908617
                        2  -8.652467    -11.5309770
A data.frame: 6 × 2     3  -5.340401    -7.3474393
                        4  -5.455141    -0.8683876
                        5  -7.170717    -12.9125020
                        6  -11.079900   -15.1237204
```

[20]:

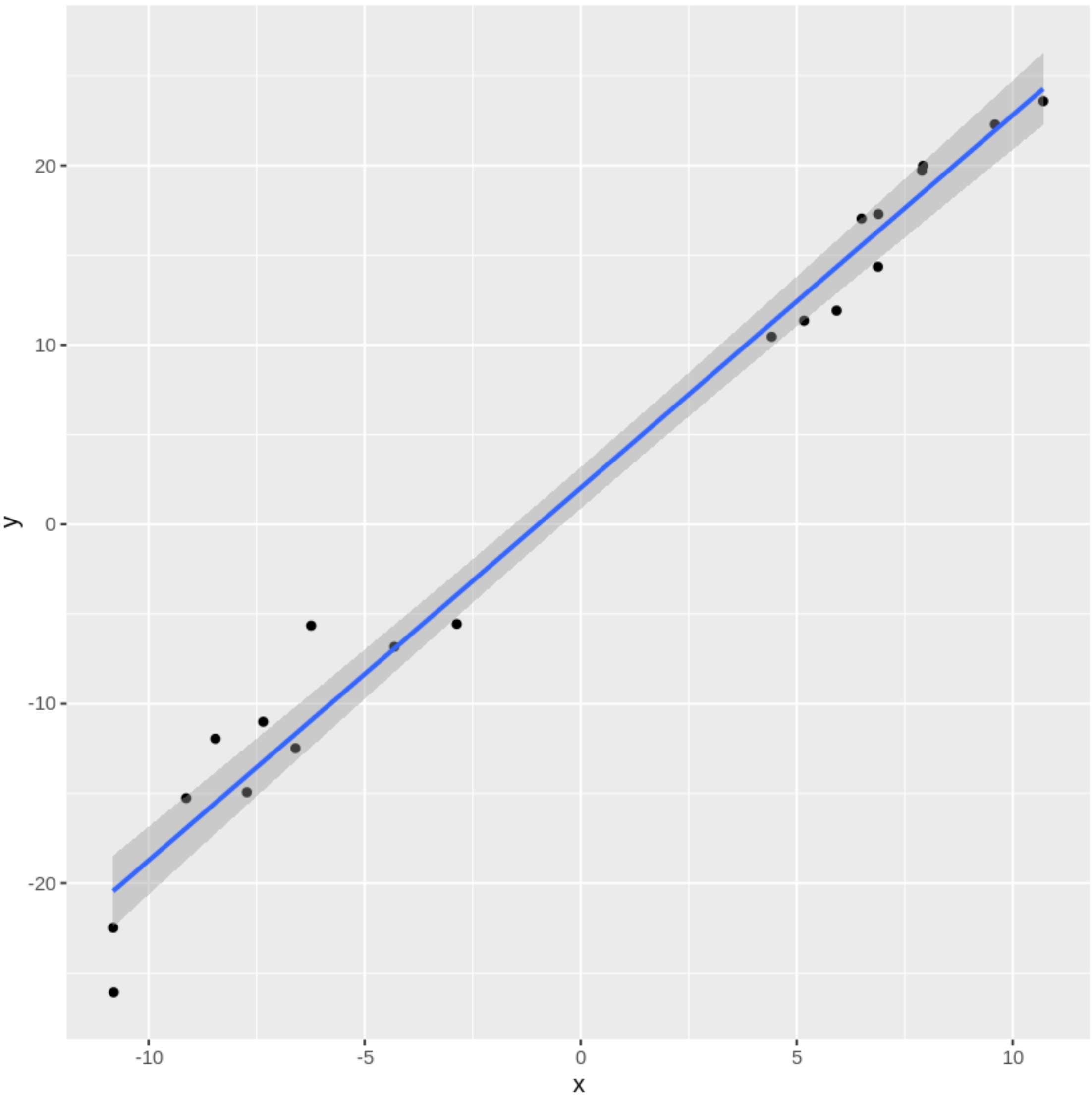
```r
# Your Code Here
df1 <- gen_data(-8, 8, 10, 10)
lm_mod1 <- lm(data = df1, y ~ x)
ggplot(data=df1,aes(x = x, y = y)) + geom_point() + stat_smooth(method="lm")

df2 <- gen_data(-8, 8, 2, 2)
lm_mod2 <- lm(data = df2, y ~ x)
ggplot(data=df2,aes(x = x, y = y)) + geom_point() + stat_smooth(method="lm")

df3 <- gen_data(-4, 4, 10, 10)
lm_mod3 <- lm(data = df3, y ~ x)
ggplot(data=df3,aes(x = x, y = y)) + geom_point() + stat_smooth(method="lm")

cat("confidence interval 1")
predict(lm_mod1, data.frame(x = 0), interval="confidence")
cat("confidence interval 2")
```

```
predict(lm_mod2, data.frame(x = 0), interval="confidence")
cat("confidence interval 3")
predict(lm_mod3, data.frame(x = 0), interval="confidence")


cat("prediction interval 1")
predict(lm_mod1, data.frame(x = 0), interval="prediction")
cat("prediction interval 2")
predict(lm_mod2, data.frame(x = 0), interval="prediction")
cat("prediction interval 3")
predict(lm_mod3, data.frame(x = 0), interval="prediction")
```

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

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



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



```
confidence interval 1
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	-1.717624	-7.230382	3.795133

```
confidence interval 2
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	2.037496	0.8819243	3.193067

```
confidence interval 3
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	0.5478181	-5.271303	6.366939

```
prediction interval 1
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	-1.717624	-26.98017	23.54492

```
prediction interval 2
```

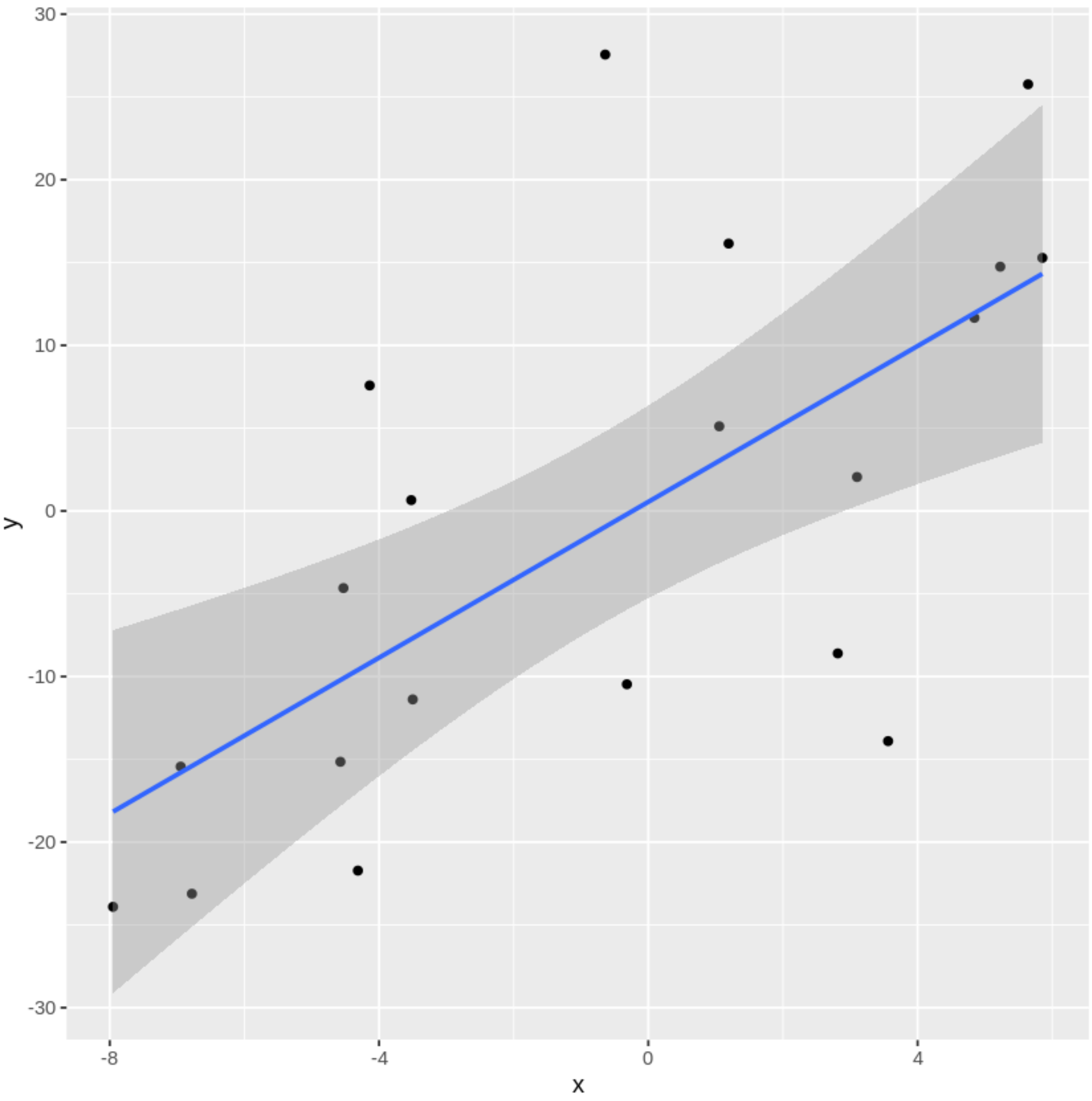
A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	2.037496	-3.257374	7.332366

```
prediction interval 3
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	0.5478181	-25.81649	26.91213



the larger the variance, the larger the confidence interval and the prediction interval, with the prediction interval being considerably wider than the confidence interval. The confidence interval is narrower near the mean of x.

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?

[25]:
```
# Your Code Here
# the value of x_k corresponding to the shortest CI width is the mean value of␣
 ↪x.
x_bar <- mean(df1$x)
```

the value of x_k corresponding to the shortest CI width is the mean value of x. There should be more certainty for estimates of x near the mean than for values of x nearer the extreme ends of the data.

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

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

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

yes, the width of the prediciton interval is narrower near the mean value of x, since there is more certainty in the estimate near the mean.

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 one can only observe one of two potential outcomes, i.e., one cannot observe what would have happened had the treatment not been given (and vice-versa).

2. the use of close substitutes is an attempt to observe the counterfactual, i.e., the close substitute is a statistical unit of analysis that has not had the treatment, but that is comparable to the treated unit, and can, therefore be taken as an approximation to the conterfactual outcome.

3. In a deterministic theory A is said to be the cause of B is B happens when A happens, while in the probabilistic theory, A is the cause of B is A happening increases the probability of B also happening.

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?

Controversy emerged from seemingly conflicting from Knox et al. on the one hand, and Gaebler et al on the other, on what type of causal claims may be made from analysis of police records and the effect of race in policing arrest or stopping decisions. In summary, Knox et al. claim that police records are already the product of a selection mechanism whereby black individuals are much more likely to be arrested when compared to white individuals. Moreover, since no records exits on individuals of both races that were merely observed by the police but not arrested, it is not possible to know the extent to which discrimination is actually taking place. Therefore, the authors claim, the impact of race on police cannot be completely determined, and any attempt to solve the issu by merely looking at what has actually been written in the record files will provide a necessarily incomplete, and potentially very biased, picture of reality. In turn Gaebler et al argue that, while it is true that the police records do not tell the whole story (since the very arrest decision that led to the existence of the record is likely to be an effect of discrimination), one can always make causal claims from the data that is available. A striking note is that the authors of the two papers do not disagree that there is discrimination. Where they disagree is if the available data provides sufficient evidence to make any causal claim regarding police racial discrimination. In a way, the debate arises from a pervasive theme in causal inference: missing data. The fundamental problem of causal inference lies in the fact that, by definition, one cannot observe the counterfactual: what would have happened had the individuals stopped by the police been white (black) instead of black (white). In a sense, this question cannot be answered in a meaningful way since no comparison group seems available, and therefore, it seems true that no overarching causal claim regarding the effect of race on police actions can be made. And it seems certainly true that selection mechanisms distort the data available for analysis, and make generalization of results tentative, at best (i.e., how to generalize the results to the broader population of individuals if the individuals on record were selected by very specific reasons, that are likely not present in the same way in the enlarged population?). Gaebler et al seem to take a safer approach, and makes the more modest attempt at making causal claims only from the point of arrest forward, i.e., Gaebler et al does not claim to be able to estimate the overall effect of race on policing decisions. In turn, in order to estimate the causal effects of race on policing decisions, even if only from the stopping decision forward, a strong assumption must be made by Gaebler et al in that no unobserved confounders are present in the system, i.e., confounders that would influence both the decision to arrest the person and the subsequent decision of the judicial system for that individual. In turn, Knox et al claim that, since the known data do not tell the whole story, it is best to be silent on the matter.