Module 4: Peer Reviewed Assignment

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.

In [2]:

```
# 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()
```

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).

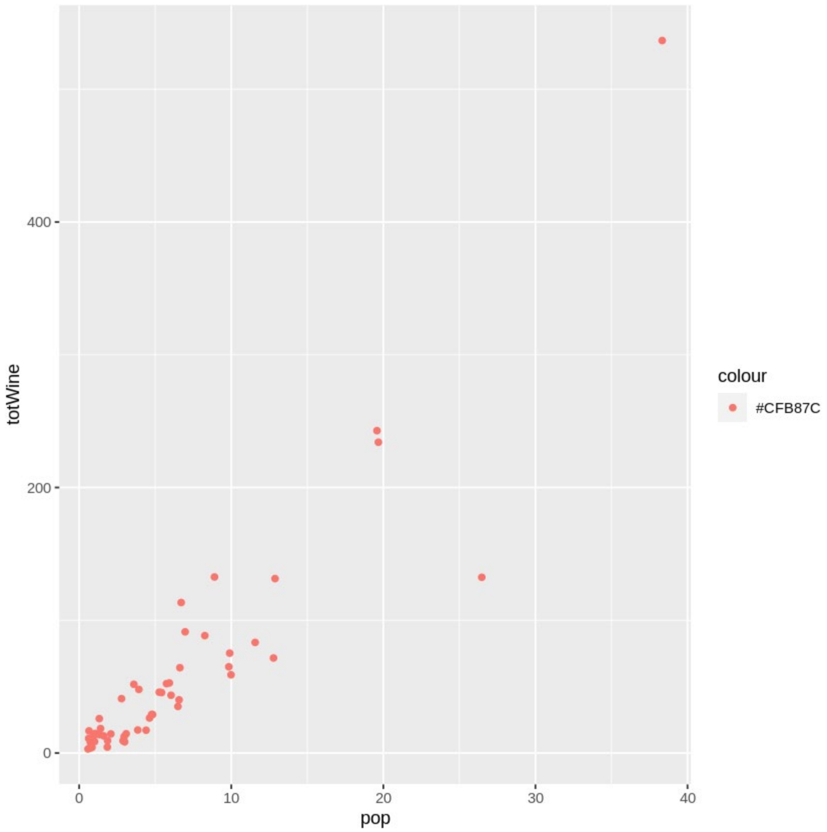
1. (a) Initial Inspections

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`.

```
In [8]: # Load the data
        wine.data = read.csv("wine_state_2013.csv")
        head(wine.data)
        # Your Code Here
        ggplot(data = wine.data) + geom_point(mapping = aes(x = pop, y = totWine,
        colour = "#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.

In [35]:

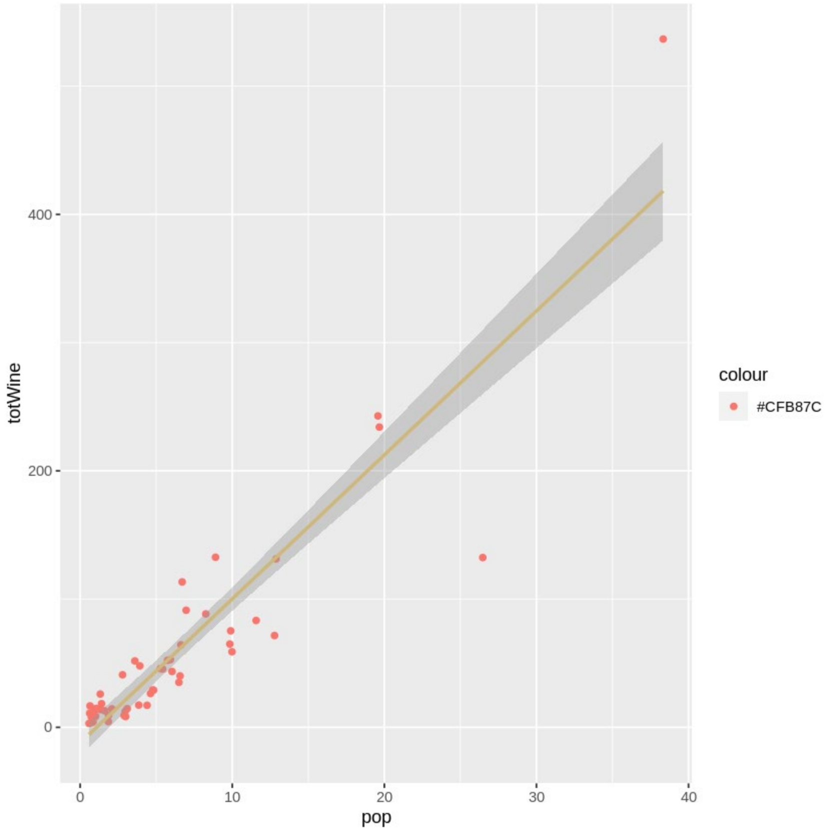
```r
# Your Code Here
lmod1 = lm(totWine ~ pop, data = wine.data)
summary(lmod1)

ggplot(wine.data, aes(x = pop, y = totWine, color = '#CFB87C')) + geom_point() + geom_smooth(method='lm', col='#CFB87C', se = TRUE, level=0.90)

# point-value population upper and lower values for the confidence interval.
predict(lmod1, newdata = data.frame(pop = 20), interval = "confidence", level = 0.90)
```

```
Call:
lm(formula = totWine ~ pop, data = wine.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-152.696  -12.697    0.488   14.900  118.473 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12.1211     6.5213  -1.859   0.0691 .  
pop          11.2257     0.6991  16.057   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.8 on 49 degrees of freedom
Multiple R-squared:  0.8403,	Adjusted R-squared:  0.837 
F-statistic: 257.8 on 1 and 49 DF,  p-value: < 2.2e-16
```

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

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	212.3938	194.2735	230.5142



At population = 20, the upper and lower values for the confidence interval are 194.2735 and 230.5142, respectively. This means that the uncertainty regarding the expected value of y-value (totWine) at the given x-value (population) has an upper bound of 230.5142 and a lower bound of 194.2735, with the true expected y-value being within the two bounds 90% of the time.

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.

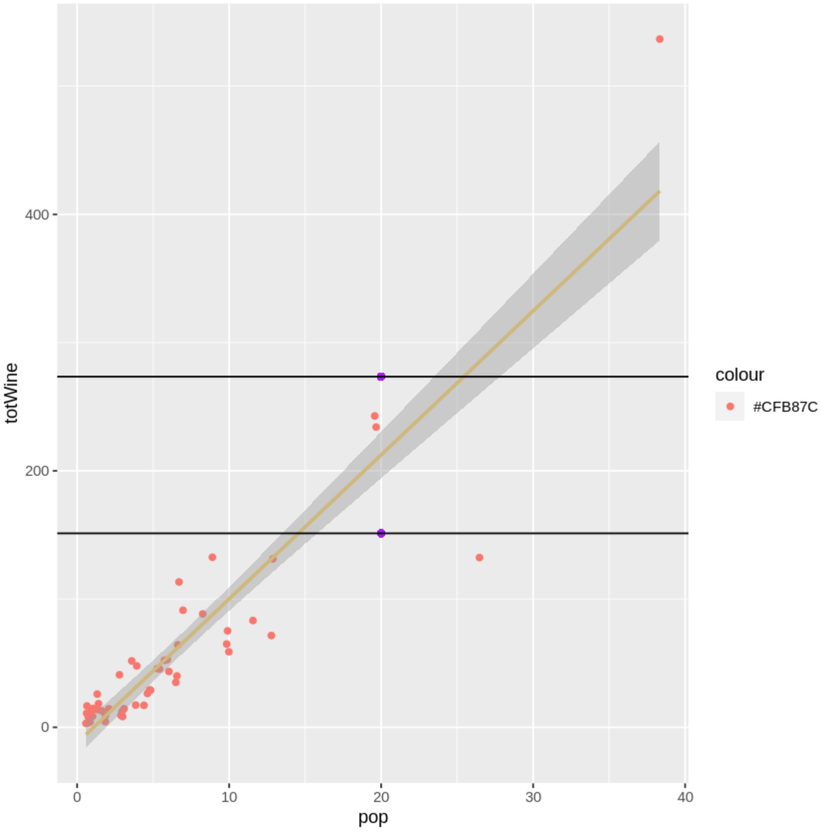
In [53]:

```
# Your Code Here
plot1 = ggplot(wine.data, aes(x = pop, y = totWine, color = '#CFB87C', ymin = 151.3078, ymax = 273.4799)) + geom_point() + geom_smooth(method='lm', col='#CFB87C', se = TRUE, level=0.90)
predict(lmod1, newdata = data.frame(pop = 20), interval = "predict", level = 0.90)
plot1 + geom_point(aes(x=20, y=273.4799), color = "purple") + geom_point(aes(x=20, y=151.3078), color="purple") + geom_hline(yintercept = 151.3078) + geom_hline(yintercept = 273.4799)
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	212.3938	151.3078	273.4799

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



The lower and upper values for the prediction interval at pop = 20 are 151.3078 and 273.4799, respectively. This interval expresses the uncertainty related to the predicted y value (totWine) of the sampled point at pop=20. It also represents the predicted range of values of a future observation based on the initial model - There is a 90% chance that, at a population of 20, there will be approximately between 151 and 273 bottles of wine.

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 $\hat{\varepsilon}_i$ be the residuals of the regression model):

1. $\sum \hat{\varepsilon}_i = 0$: The sum of residuals is 0.
2. $\sum \hat{\varepsilon}_{2i}$ is as small as it can be.
3. $\sum x_i \hat{\varepsilon}_i = 0$
4. $\sum \hat{y}_i \hat{\varepsilon}_i = 0$: 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

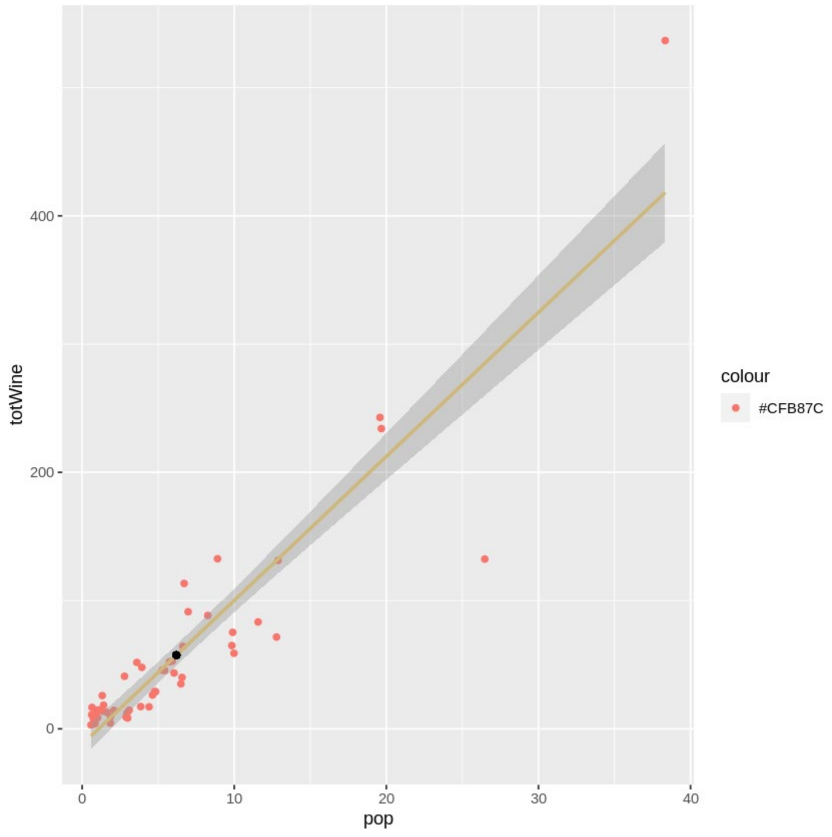
practice, try to prove these facts yourself!

In [62]:

```
# Your Code Here
# 1. The sum of residuals is 0.
sum(resid(lmod1))

# 2. The RSS is as small as it can be.
# This formulation makes sense because, if we fit the correct model, the d
iscrepancy between the data and the model should be as small as possible,
# since the model is supposedly meant to be a 'line-of-best-fit'.
# Since the RSS measures this discrepancy, our model should produce an RSS
 as small as it can be.

# 3. sum(xi*e_i_hat) = 0
sum(resid(lmod1) * wine.data$pop)

# 4. The Residuals are orthogonal to the fitted values.
sum(predict(lmod1) * resid(lmod1))

# 5. The Regression Line always goes through (xbar, ybar).
xbar = mean(wine.data$pop)
ybar = mean(wine.data$totWine)
plot1 + geom_point(aes(x=xbar, y=ybar), color = "black")
```

```
-2.00672811700997e-14

-1.11632925126059e-12

-1.30073729565083e-11
```

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



As shown above, the model verifies consequences 1, 3, 4, and 5. As for consequence 2, the formulation makes sense because the model is meant to be a 'line-of-best-fit'. The residual sum of squares measures the overall discrepancy between each data point and the regression model, which should be as small as it can be for a correct model.

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 from an observational study. We cannot infer causation between population and the amount of wine drank, because that would mean that the ONLY reason why there were x bottles of wines consumed at a certain city is because of that specific population size, without taking into account religion, age, gender, alcohol preference, and many other factors that may influence the number of wines consumed.

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?

In [99]:

```
gen_data <- function(mu1, mu2, var1, var2){
    # Function to generate 20 data points from 2 different normal distribu
tions.
    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

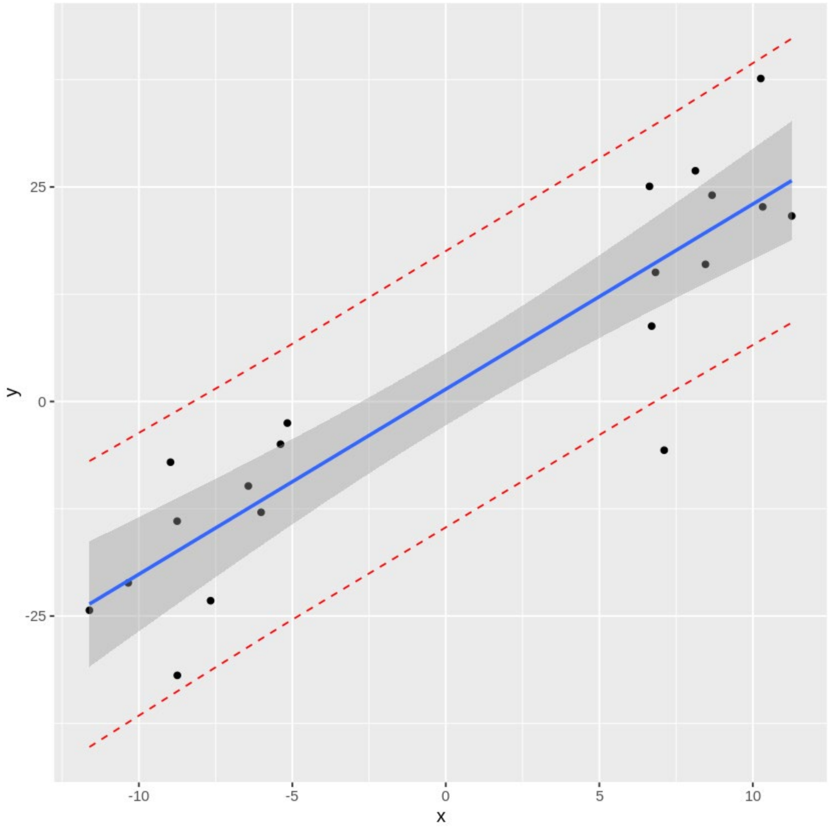
	x	y
	<dbl>	<dbl>
1	-5.474091	-11.1908617
2	-8.652467	-11.5309770
3	-5.340401	-7.3474393
4	-5.455141	-0.8683876
5	-7.170717	-12.9125020
6	-11.079900	-15.1237204

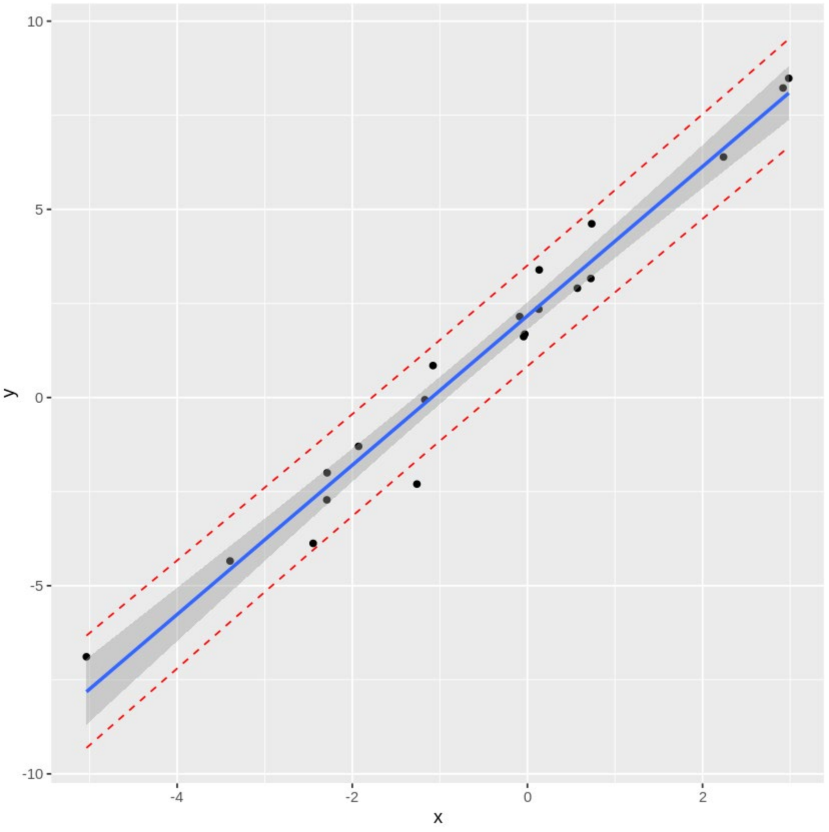
In [107]:

```
# Your Code Here
# Fit model and plot with conf and pred intervals.

df1 = gen_data(-8, 8, 10, 10)
```

```
mod11 = lm(y ~ x, data = df1)
temp1 = predict(mod11, interval = "prediction", level = 0.90)
new_df1 = cbind(df1, temp1)
ggplot(new_df1, 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)


df3 = gen_data(-1, 1, 1, 1)
mod33 = lm(y ~ x, data = df3)
temp3 = predict(mod33, interval = "prediction", level = 0.90)
new_df3 = cbind(df3, temp3)
ggplot(new_df3, 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)
```

```
Warning message in predict.lm(mod11, interval = "prediction", level = 0.9):
“predictions on current data refer to _future_ responses
”
`geom_smooth()` using formula 'y ~ x'

Warning message in predict.lm(mod33, interval = "prediction", level = 0.9):
“predictions on current data refer to _future_ responses
”
`geom_smooth()` using formula 'y ~ x'
```





Setting the means and variances to smaller values creates much narrower confidence and prediction intervals.

3. (b) The Smallest Interval

Recall that the Confidence (Mean) Interval, when the predictor value is x_k, is defined as:

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

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 $MSE \times \left(1n + (x_h - \bar{x})^2 \Sigma (x_i - \bar{x})\right)$ 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?

```
In [130]: # Your Code Here
          # Calculate t-values at 95% confidence level
          t.value = qt(0.975, df = 18)

          # CI with the shortest width is obtained when x_k is equal to x_bar.
          # This is also saying that the predictor value x_k is equal to th sample's
           mean.
          # Then, with a large enough sample size n, the standard error will get clo
          se to 0, regardless of the t-values.
          # This will produce the shortest CI width for any given sample size.
          new_xk = mean(df1$x)
```

```r
# CI width for x_k = x_bar
CI1 = predict(mod11, newdata = data.frame(x = new_xk), interval = "confide
nce", level = 0.95)
CI1[3] - CI1[2]

# CI width for x_k =/= x_bar
CI2 = predict(mod11, newdata = data.frame(x = new_xk-0.1), interval = "con
fidence", level = 0.95)
CI2[3] - CI2[2]

CI3 = predict(mod11, newdata = data.frame(x = new_xk+0.1), interval = "con
fidence", level = 0.95)
CI3[3] - CI3[2]

summary(mod11)
```

8.41494362821628

8.41554303053035

8.41554303053035

```
Call:
lm(formula = y ~ x, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.4574  -3.8028  -0.4931   5.6933  14.0721 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.439      2.004   0.718    0.482    
x              2.157      0.239   9.022 4.25e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.956 on 18 degrees of freedom
Multiple R-squared:  0.8189,	Adjusted R-squared:  0.8089 
F-statistic:  81.4 on 1 and 18 DF,  p-value: 4.245e-08
```

Confidence interval with the shortest width is obtained when $x_k = \bar{x}$, namely, the predictor value is equal to the sample's mean. This is also shown above, where the confidence interval width is smallest at $x_k = \bar{x}$ and gets slightly larger as we move away from the sample's mean at an increment of 0.1. This makes sense because the confidence interval includes the upper and lower values obtained from the variance in confidence. Having the predictor value equal to the sample's mean minimizes this variance. Intuitively, if n is large enough, it is possible for the confidence interval to get close to 0. When we have $x_k = \bar{x}$, the confidence interval width will be the shortest for any given sample size, n.

3. (c) Interviewing the Intervals

Recall that the Prediction Interval, when the predictor value is x_k, is defined as:

$$\hat{y} \ \pm t \qquad MSE \ 1 + 1n + (x \ - \bar{x})^2 \sum (x \ - \bar{x})$$

$h \quad \alpha/2,n-2 \sqrt{} \quad (\quad k \quad i \quad)$

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

No, the width does not change at different population values. Expanding out the terms under the square root shows that even if the sample size n approaches infinity, the terms collectively do not converge to a single, smaller value, barely affecting the width of the prediction interval.

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?

1. Describe the use of "close substitutes" as a solution to the fundamental problem of causal inference. How does this solve the problem?

1. 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, when chosen option A out of two possible choices A and B, we will only be able to observe the outcome of A and never know that of B. The counterfactual definition of causality, according to the lecture slide, states that "C causes E if, in the absence of C, E would not have occurred (or would have been less likely to occur)." This is similar to saying that, from the option A-B example, choosing option A caused the respective outcome. However, the issue is that we have no way of knowing if choosing option B may have caused the same outcome, in which case the very cause-effect relationship may potentially have come from somewhere else.
2. The method of 'close substitutes' refers to finding subjects that are similar in relevant characteristics with respect to a certain cause-effect event and attempting to answer causal claims. This way, the cause-effect event can be replicated while keeping all conditions as similar as possible, allowing for the observation of the 'other' outcome, i.e. outcome from choice B.
3. Deterministic: An effect must necessarily follow from a cause. Probabilistic: The probability of an effect changes due to the existence of a cause.

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?

1. What are some "possible distortionary effect[s] of using arrest data from administrative police records to measure causal effects of race"?

1. What role do assumptions (both statistical and otherwise) play in this debate? To what extent are assumptions made by different researchers falsifiable?

The fundamental problem of causal inference is the idea that there is no way of knowing the outcome of an alternative decision once a certain decision is made. Lily Hu addresses this idea and shares her thoughts regarding the issue by bringing in discussions and debates from professionals and peers. In her article, she explains the disparity between statistical/empirical methods and their outcomes and the society's/government's take on racial bias in the criminal justice system.

She acknowledges that deciding on a set "standard" of ethical and just laws is a difficult task by saying "this burden of normativity is so onerous that efforts at alleviating any of it are always being offered up, and taken up, in legal analysis." The article immediately illustrates how statistical methods have transformed this process in areas such as economics, but cautions that, when it comes to the criminal justice system, it takes both scrutiny and actionable guidance from the statistics arising from "complex social systems". She takes the debate regarding "possible distortionary effect[s] of using arrest data from administrative police records to measure causal effects of race" as the pillar of her discussion. According to the article, the central issue of the debate is that the statistical methods used to conduct research on racial discrimination in the criminal justice system may or may not be flawed based on the data.

Upon further examining the debate, Lily concludes that the dispute ultimately is about the assumptions underlying the statistical inference process. This ties into the fundamental problem of causal inference in the sense that, because of the observable outcome from an assumed cause, there is no way to find out the other outcome that would have arose from other sets of variables except from assumptions. The article highlights this fact by saying "assumptions are assumptions are just assumptions!".

Despite the tremendous growth in data size and complexity, statistical inference and conclusions appear quite simple. After all, one of the goals of empirical research is to bring concreteness to otherwise vague and abstract ideas. The article addresses the risk and disputes that may arise when assumptions become less intuitive or perhaps too self-serving, cautioning readers to be wary before selecting/performing statistical outcomes.

```
In [ ]:
```