C1M4_peer_reviewed

March 4, 2025

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.

[4]:

```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`.

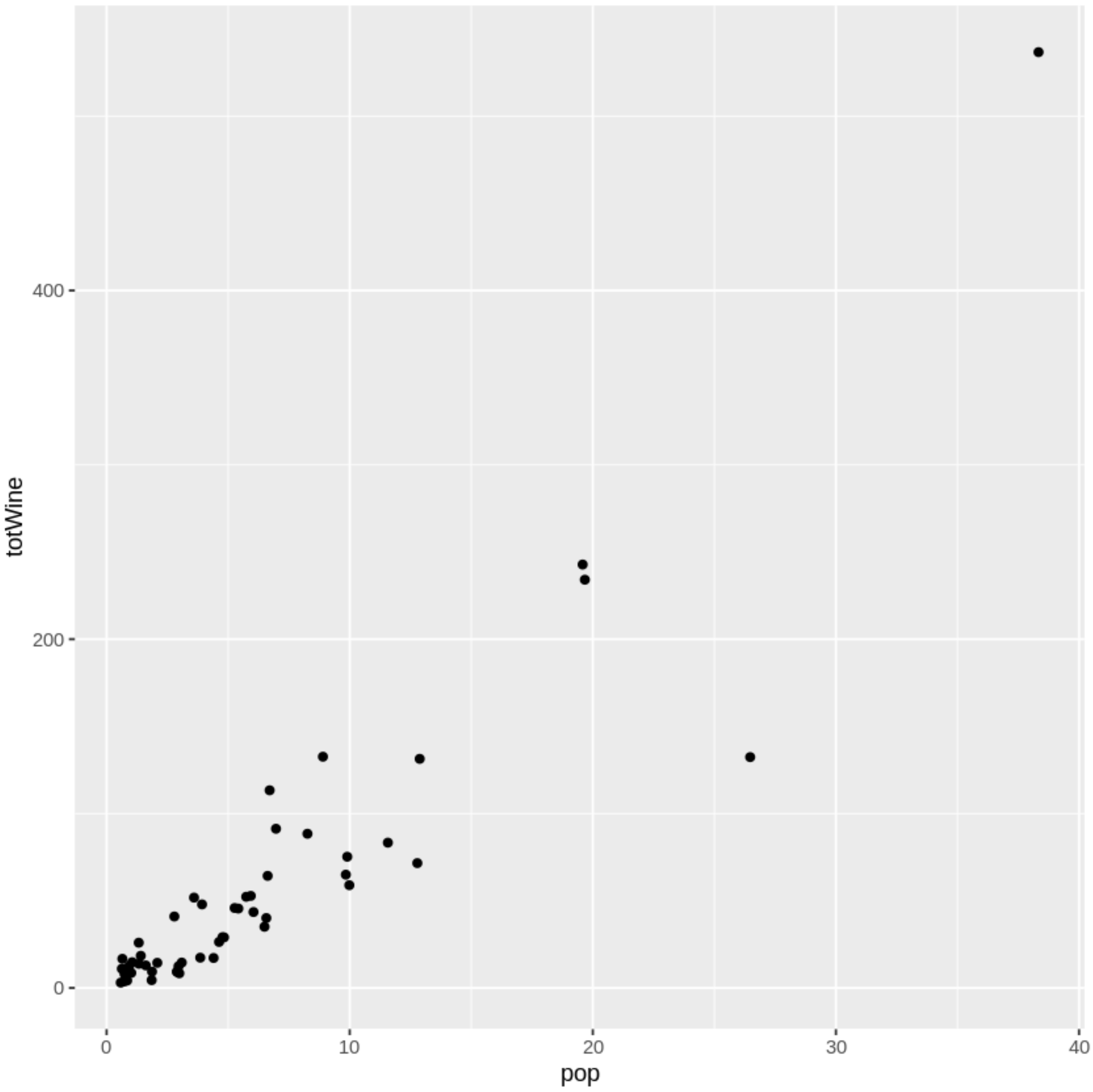
[5]:

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

set.seed(10000)
colnames(wine.data)

ggplot(wine.data, aes(x = pop, y = totWine), color = "#CFB87C") + geom_point()
```

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. 'State' 2. 'pcWine' 3. 'pop' 4. 'totWine'



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.

[6]:
```
lm_wine <- lm(totWine ~ pop, data = wine.data)
lm_wine

df.new <- data.frame(pop = mean(wine.data$pop))
df.new
```

```
pop.new <- as.data.frame(df.new)$pop
pop.new
```

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

Coefficients:
(Intercept)          pop
     -12.12        11.23
```

A data.frame: 1 × 1

pop <dbl>
6.200096

6.20009623529412

[7]:
```
pred <- predict(lm_wine, newdata = df.new, interval = "confidence", level = 0.9)
pred

upper <- pred[3]
lower <- pred[2]

upper
lower

ggplot(wine.data, aes(x = pop, y = totWine), color = "#CFB87C") +
geom_point() +
geom_smooth(method = "lm", col = "#CFB87C", level = 0.9) +
geom_point(aes(x = pop.new, y = upper), color = "orange") +
geom_point(aes(x = pop.new), y = lower, color = "red")
```

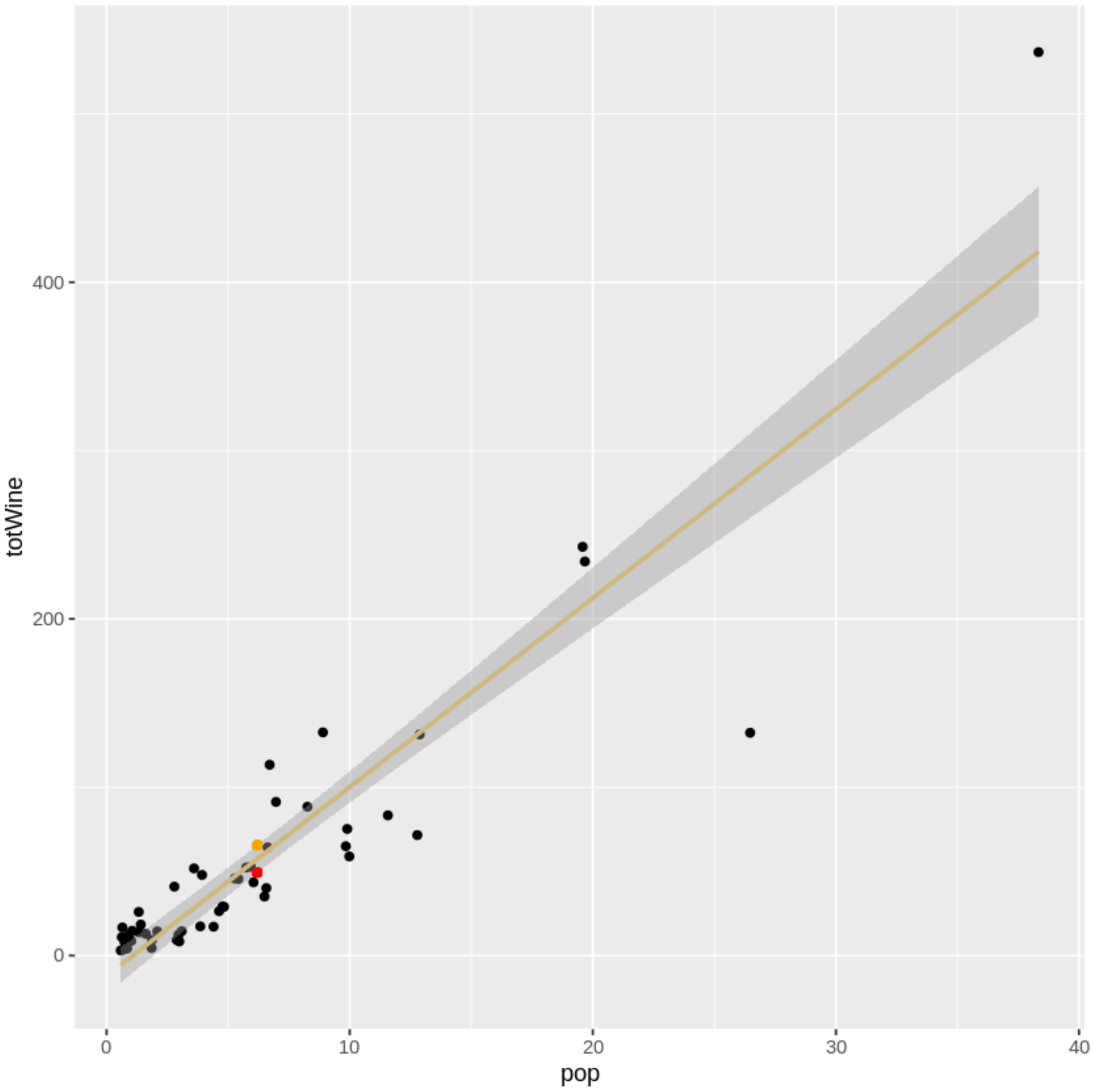
A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	57.47962	49.31087	65.64838

65.6483803100355

49.3108692625135

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



```
[8]: print("The 90% confidence interval is represented between the orange and red␣
      ↪lines.")
```

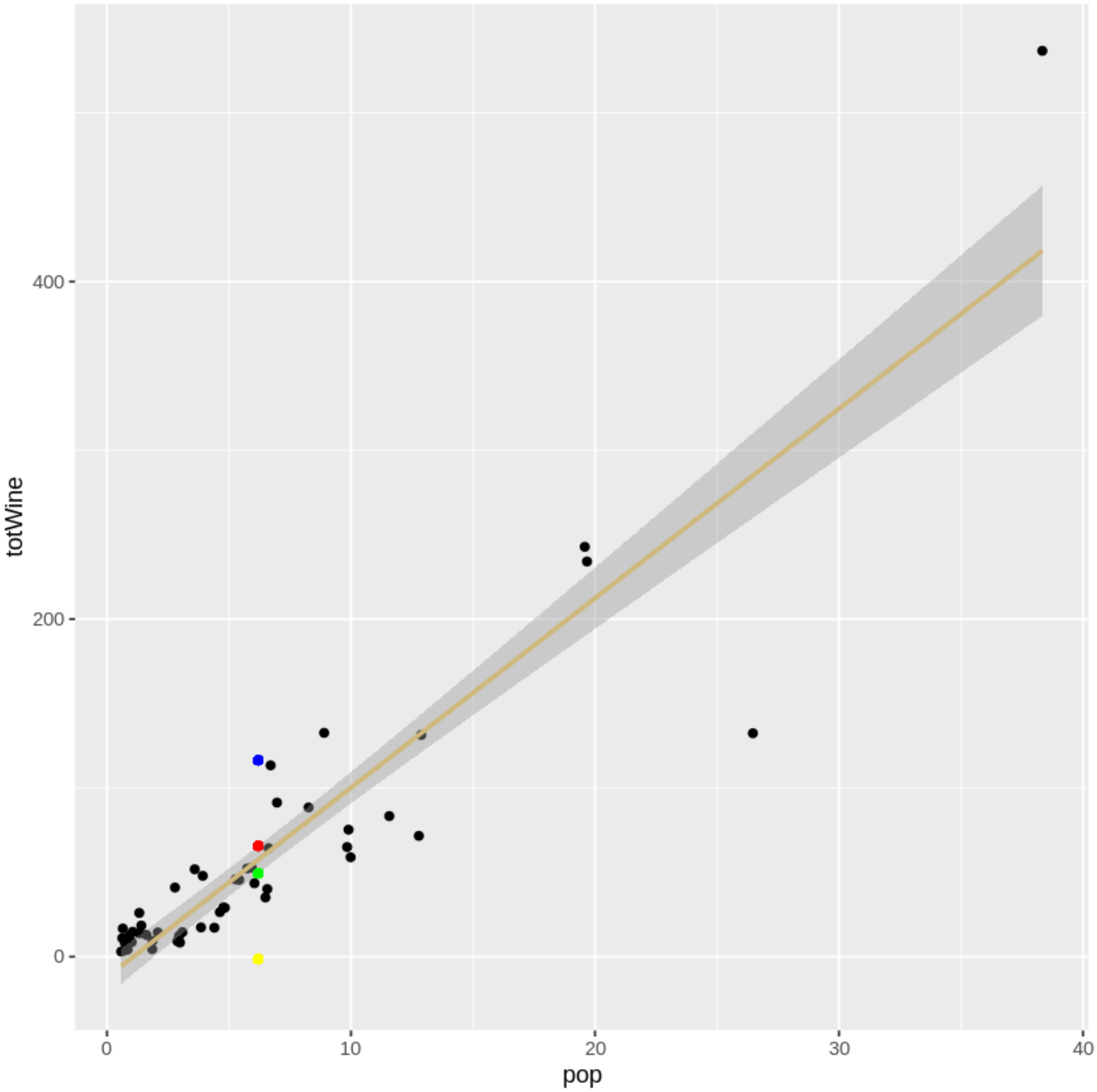
```
[1] "The 90% confidence interval is represented between the orange and red
lines."
```

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.

```
[9]: pred_2 <- predict(lm_wine, newdata = df.new, interval = "prediction", level = 0.
      ↪9)
     lower2 <- pred_2[2]
```

```
upper2 <- pred_2[3]

ggplot(wine.data, aes(x = pop, y = totWine), color = "#CFB87C") +
geom_point() +
geom_smooth(method = "lm", col = "#CFB87C", level = 0.9) +
geom_point(aes(x = pop.new, y = upper), color = "red") +
geom_point(aes(x = pop.new), y = lower, color = "green") +
geom_point(aes(x = pop.new, y = upper2), color = "blue") +
geom_point(aes(x = pop.new), y = lower2, color = "yellow")


# Your Code Here
```

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



[10]:
```
print("The resampled value is fit within the green, blue, red, and yellow␣
↪points.")
```

```
[1] "The resampled value is fit within the green, blue, red, and yellow points."
```

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!

[11]:
```
r <- resid(lm_wine)
sum(r)

print("least squares minimizes the sum of squares for residuals")

sum(wine.data$pop*r)

yhat <- mean(wine.data$totWine)
sum(yhat*r)

predict(lm_wine, newdata = df.new) - mean(wine.data$totWine)
```

-2.00672811700997e-14

```
[1] "least squares minimizes the sum of squares for residuals"
```

-1.11632925126059e-12

-2.32702745961433e-12

1: 7.105427357601e-15

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?

[12]:

```
print("observational study - we cannot infer causation")
```

```
[1] "observational study - we cannot infer causation"
```

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?

[13]:

```
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

[29]:

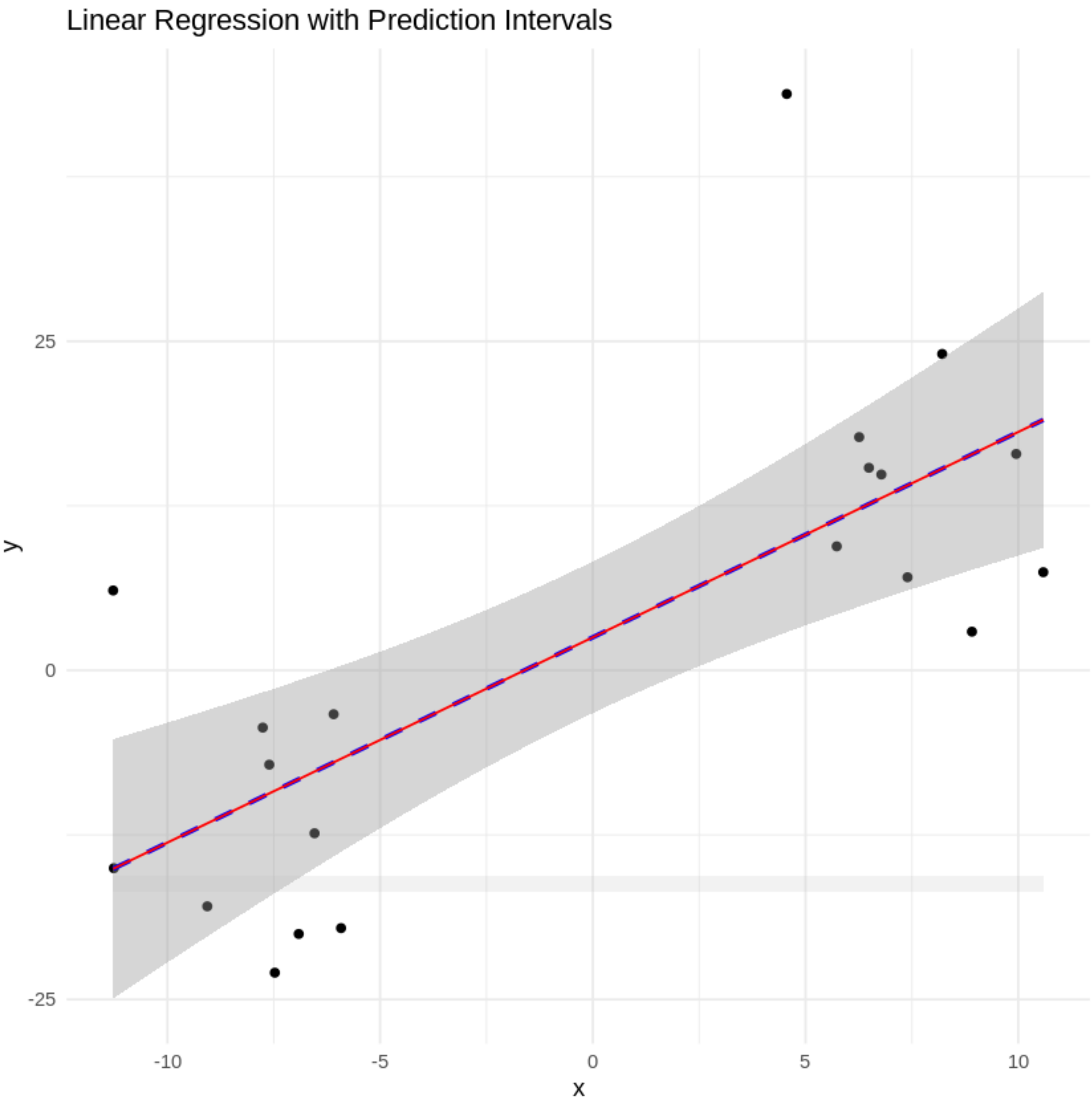
```
gen_data2 <- gen_data(-8, 8, 10, 10)
lm_gen <- lm(y ~ x, data = gen_data2)
pred <- predict(lm_gen, newdata = gen_data2, interval = "prediction", level = 0.95)
df <- data.frame(pred)
df_gen <- gen_data2

gen_data2$fit <- pred[,"fit"]
gen_data2$lower <- pred[,"lwr"]
gen_data2$upper <- pred[, "upr"]


lm_gen
```

```
Call:
lm(formula = y ~ x, data = gen_data2)

Coefficients:
(Intercept)            x
      2.511        1.560
```

[33]:

```
ggplot(gen_data2, aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "blue", linetype = "dashed") +
  geom_line(aes(y = fit), color = "red") +
  geom_ribbon(aes(ymin = pred_lower, ymax = pred_upper), alpha = 0.2, fill = "grey") +
  labs(title = "Linear Regression with Prediction Intervals",
       x = "x", y = "y") +
  theme_minimal()


print("Increasing the variances will increase the width of the intervals - more uncertainty in predictions.")
```

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


[1] "Increasing the variances will increase the width of the intervals - more
uncertainty in predictions."
```



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

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

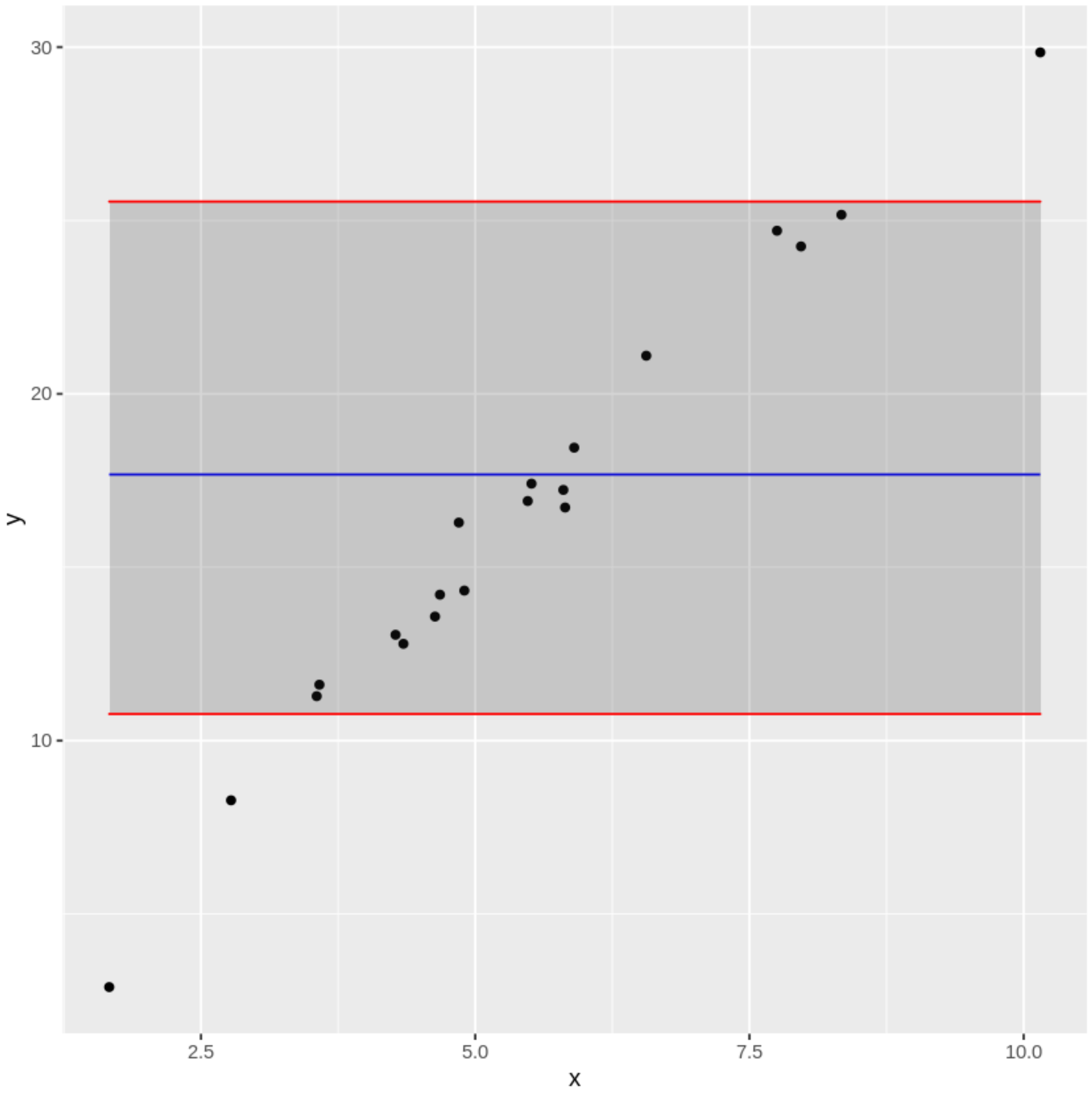
where $\hat{y}_h$ is the fitted response for predictor value x_h, $t_{\alpha/2,n-2}$ is the t-value with $n-2$ degrees of freedom and $MSE \times \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})}\right)$ is the standard error of the fit.

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?

[50]:

```
set.seed(7282)
n <- 20
x <- rnorm(n, mean = 5, sd = 2)
y <- 3 * x + rnorm(n, mean = 0, sd = 1)
data <- data.frame(x = x, y = y)

model <- lm(y ~ x, data = data)
x_bar <- mean(data$x)
x_grid <- seq(min(data$x), max(data$x), length.out = 20)
pred <- predict(model, newdata = data.frame(x_grid), interval = "confidence",␣
 ↪level = 0.95)


ggplot(data, aes(x = x, y = y)) +
  geom_point() +
  geom_line(aes(x = x_grid, y = pred[1]), color = "blue") +
  geom_ribbon(aes(x = x_grid, ymin = pred[2], ymax = pred[3]), alpha = 0.2,␣
 ↪color = "red")



print("The width is narrowest around x = x_bar.")

      # Your Code Here
```

```
[1] "The width is narrowest around x = x_bar."
```



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.

```
[51]: print("The intervals are wider for the mean predictions as x moves away from␣
      ↪the mean.")
```

```
[1] "The intervals are wider for the mean predictions as x moves away from 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 we can only observe one outcome for a given unit in an experiment. It is not possible to observe both the treatment and control in an experiment for the same subject and so causal inferences cannot be measured.
2. The use of close substitutes as a solution to the fundamental problem of causal inference assumed that the groups have similar outcomes without the treatment. This helps estimate a causal effect because it compares the treatment and control groups. It is used as a sort of control for existing variables and allows for comparisons.
3. The deterministic theory of causality states that a specific cause leads to a specific effect. If B is present then A must have occurred. The probabilistic theory of causality states that the cause effects the likelihood of an event occurring but does not necessarily effect it directly. If C occurs, then A most likely occurred prior.

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?

[ ]:

[ ]: