Ideas:

- A quick description of causality and what needs to be true for us to infer causality.
- Make then describe the difference between a point estimate and a prediction interval.
- We could suggest that students practice interpreting those CIs and PIs because they will be asked about the interpretations on the peer review assignment.

In [ ]:

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 [1]:
```
# This cell loads the necesary libraries for this assignment
library(tidyverse)
```

```
Registered S3 methods overwritten by 'ggplot2':
  method         from
  [.quosures     rlang
  c.quosures     rlang
  print.quosures rlang
Registered S3 method overwritten by 'rvest':
  method            from
  read_xml.response xml2
── Attaching packages ─────────────────────────────────────── tidyve
rse 1.2.1 ──
✔ ggplot2 3.1.1       ✔ purrr   0.3.2
✔ tibble  2.1.1       ✔ dplyr   0.8.0.1
✔ tidyr   0.8.3       ✔ stringr 1.4.0
✔ readr   1.3.1       ✔ forcats 0.4.0
── Conflicts ────────────────────────────────────────── tidyverse_co
nflicts() ──
✖ 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).

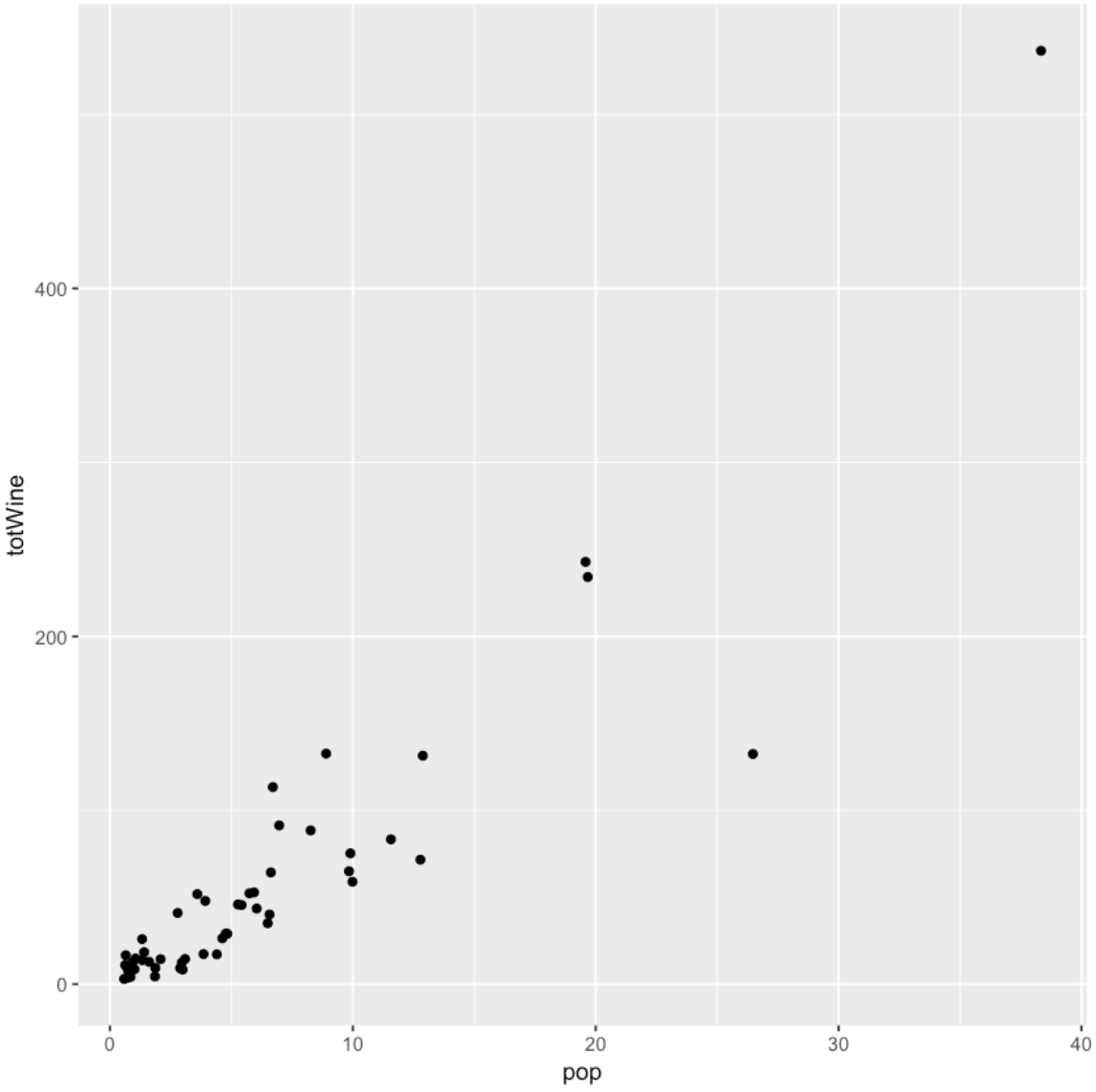
1. (a) Initial Inspections

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 [18]:
```
library(RCurl) #a package that includes the function getURL(), whic
h allows for reading data from github.
library(ggplot2)
url = getURL(paste0("https://raw.githubusercontent.com/bzaharatos
/",
                    "-Statistical-Modeling-for-Data-Science-Applica
tions/",
                    "master/Modern%20Regression%20Analysis%20/Datas
ets/wine_state_2013.csv"))
wine.data = read.csv(text = url, sep = ",")
head(wine.data)
dim(wine.data)
library(ggplot2)
ggplot(wine.data, aes(x = pop, y = totWine)) +
    geom_point()
```

State	pcWine	pop	totWine
Alabama	6.0	4.829479	28.976874
Alaska	10.9	0.736879	8.031981
Arizona	9.7	6.624617	64.258785
Arkansas	4.2	2.958663	12.426385
California	14.0	38.335203	536.692842
Colorado	8.7	5.267603	45.828146

51 4



1. (b) Confidence Intervals

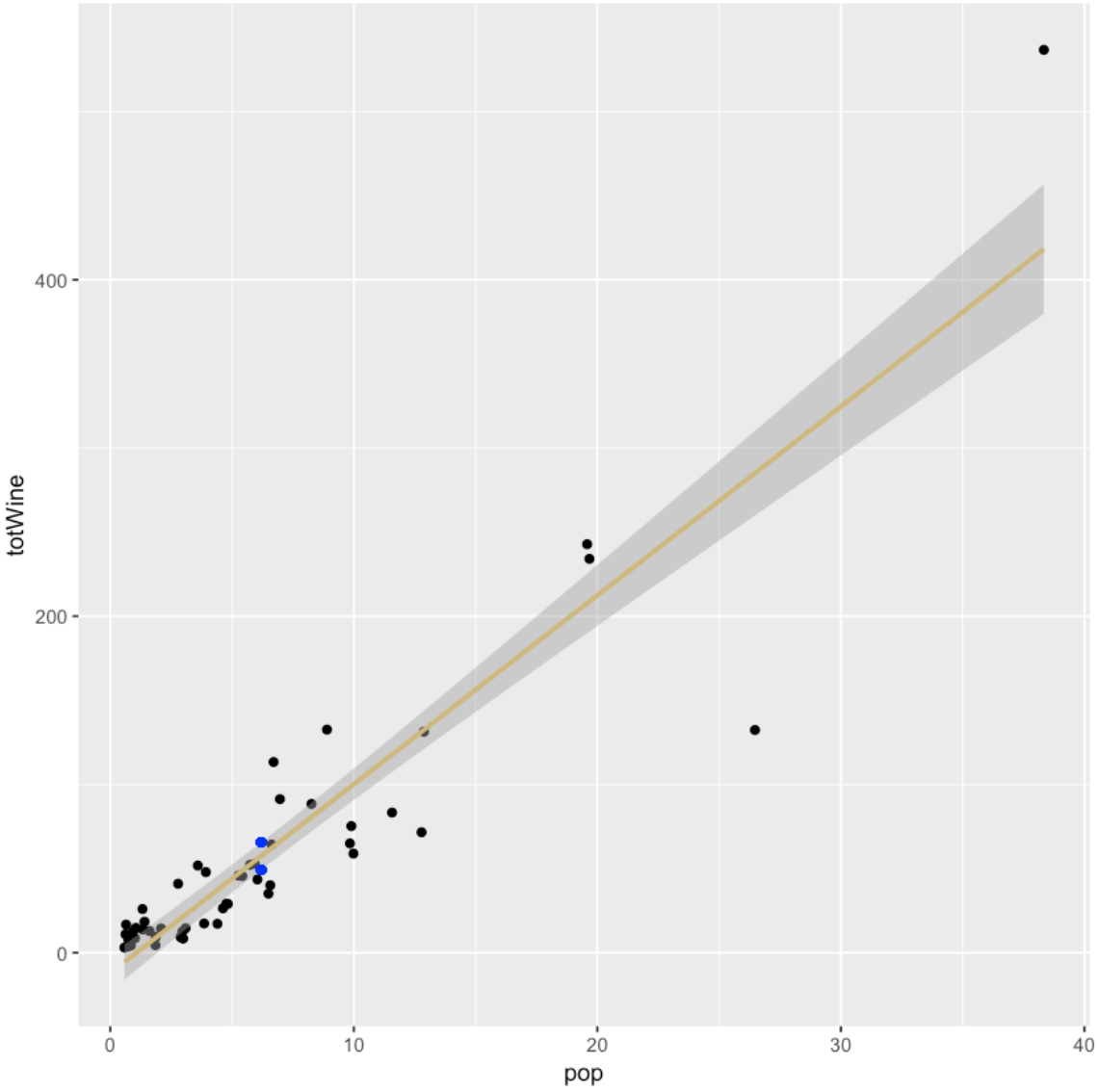
Fit a linear regression with `totWine` as the response and `pop` as the predictor. Add the regression line to your scatterplot. For fun, set its color to gold with `col=#CFB87C` . Add the 90% Confidence Interval for the regression line to the plot.

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.

In [21]:
```
lm_wine = lm(totWine ~ pop, data = wine.data)
x_new = data.frame(pop = mean(wine.data$pop)); x_new
predict(lm_wine, newdata = x_new, interval = "confidence", level =
0.9)

ggplot(wine.data, aes(x = pop, y = totWine)) +
    geom_point() +
    geom_smooth(method = "lm", col = "#CFB87C", level = 0.9) +
    geom_point(aes(x=6.2, y= 49.31087), colour="blue") +
    geom_point(aes(x=6.2, y= 65.64838), colour="blue")
```

pop
6.200096

fit	lwr	upr
57.47962	49.31087	65.64838



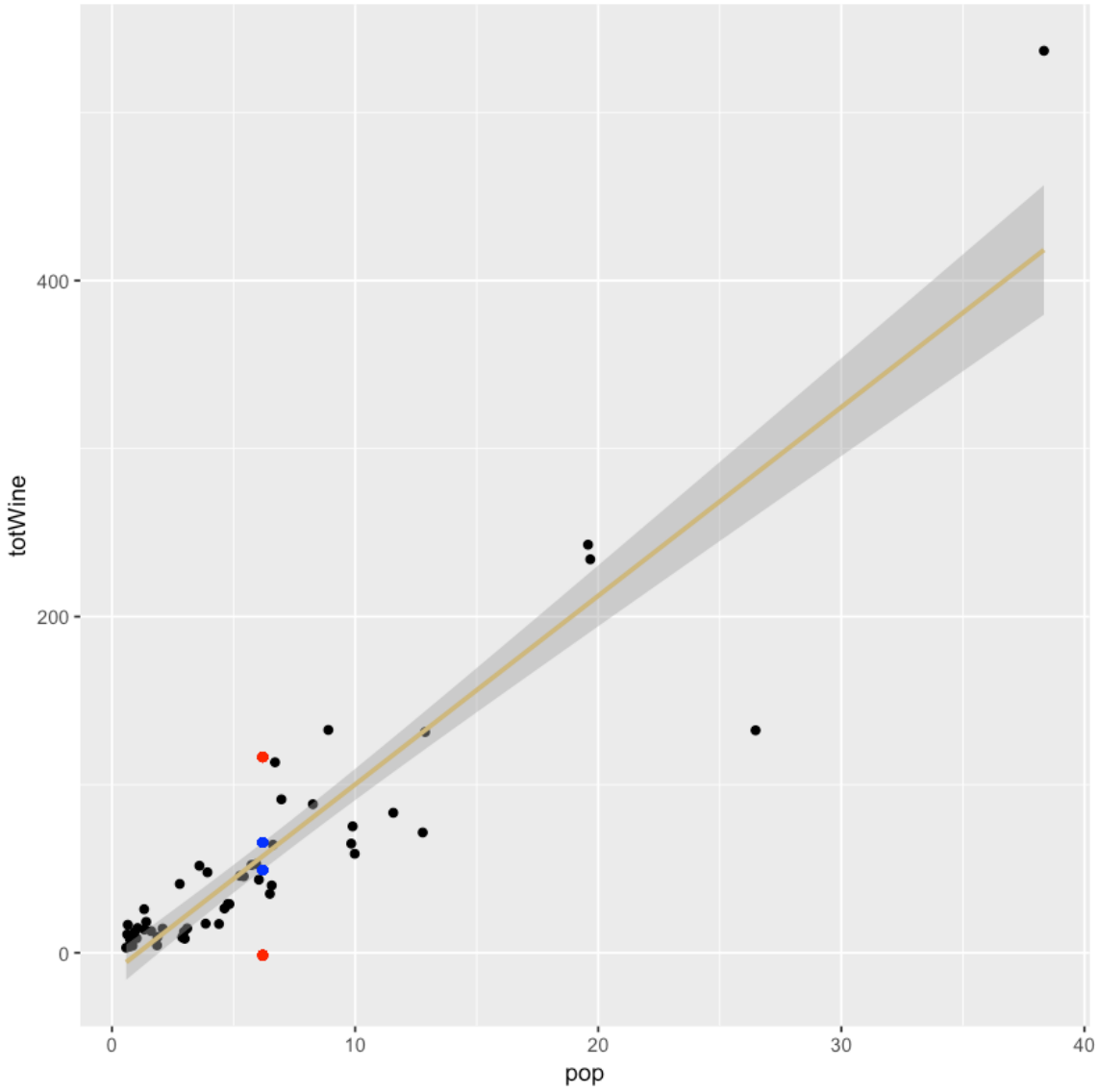
If we resampled the response many times at the same values of `pop`, and refit the line each time, 90% of those lines would fall between our two blue points (the upper and lower bound of the CI).

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 [27]:
```
predict(lm_wine, newdata = x_new, interval = "prediction", level =
0.9)

ggplot(wine.data, aes(x = pop, y = totWine)) +
    geom_point() +
    geom_smooth(method = "lm", col = "#CFB87C", level = 0.9) +
    geom_point(aes(x=6.2, y= 49.31087), colour="blue") +
    geom_point(aes(x=6.2, y= 65.64838), colour="blue") +
    geom_point(aes(x=6.2, y= -1.426109), colour="red") +
    geom_point(aes(x=6.2, y= 116.3854), colour="red")
```

fit	lwr	upr
57.47962	-1.426109	116.3854



The prediction interval gives a range of plausible values for a prediction of `totWine` at `pop = 6.2`.

Or a more detailed interpretation:

1. fix the predictors in the training data, and resample the response many times;
1. fit the model to each resample of the training data;
1. compute the prediction interval at the same values of the predictor, namely, `pop = 6.2`.

Among these prediction intervals, 90% would cover the true value of the response.

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 because that is how we "chose" $\hat{\beta}_0$ and $\hat{\beta}_1$.
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 each of the above "consequences", excluding the second one.

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 [33]:
```
r = resid(lm_wine); sum(r)
yhat = fitted(lm_wine); #1
sum(wine.data$pop*r) #3
sum(yhat*r) #4

predict(lm_wine, newdata = x_new) - mean(wine.data$totWine) #5
```

5.28466159721575e-14

-5.96522831131097e-13

-3.33955085807247e-12

1: 7.105427357601e-15

Note that these values are computationally zero. We note that, for part 2, least squares minimizes the sum of the squares of the residuals, and $\sum \widehat{\varepsilon}_i^2$ is exactly that with the minimizers plugged in.

Problem 2: Explanation

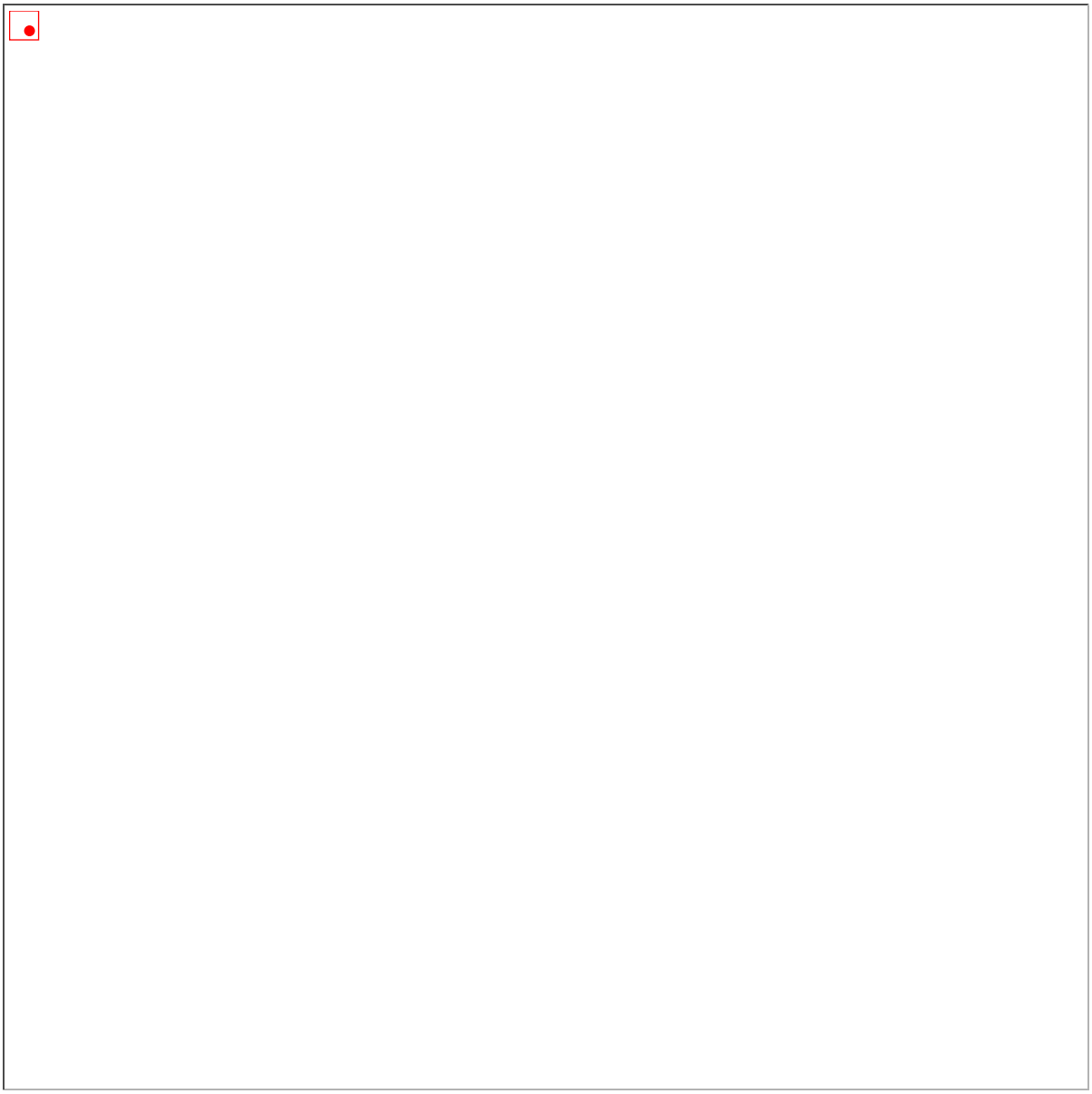


Image Source: https://xkcd.com/552/ (https://xkcd.com/552/)

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?

These data come from an observational study. No experimental "treatment" was manipulated; rather, states were observed for their wine consumption. For that reason, strictly speaking, we cannot infer a causal relationship between population and the amount of wine drank (although it does seem quite reasonable that a larger population would be one causal factor in more wine consumption).

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?

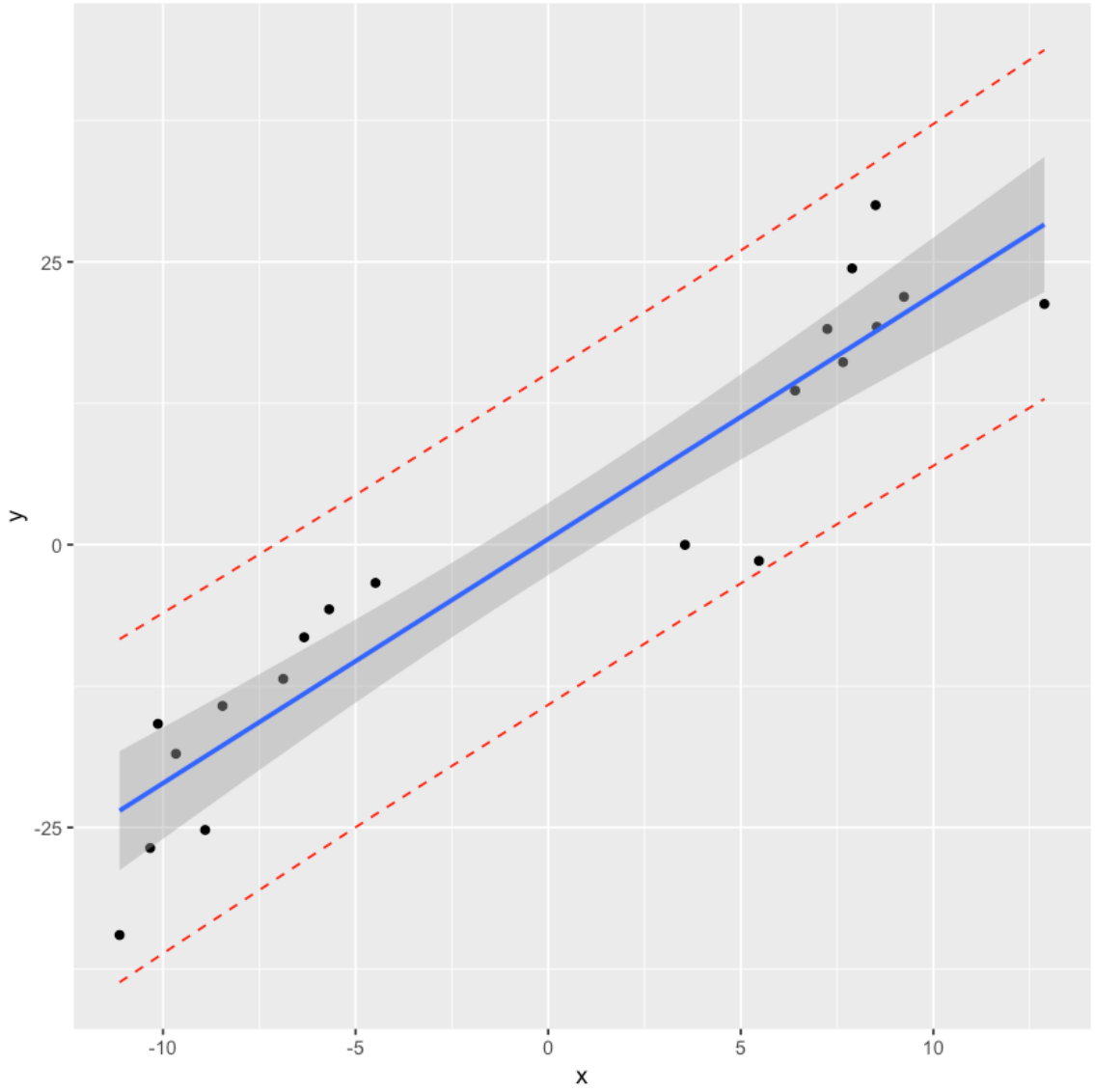
In [56]:
```
gen_data <- function(mu1, mu2, var1, var2){
    # Function to generate 20 data points from 2 different normal d
istributions.
    x.1 = rnorm(10, mu1, 2)
    x.2 = rnorm(10, mu2, 2)
    y.1 = 2 + 2*x.1 + rnorm(10, 0, var1)
    y.2 = 2 + 2*x.2 + rnorm(10, 0, var2)

    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))
```

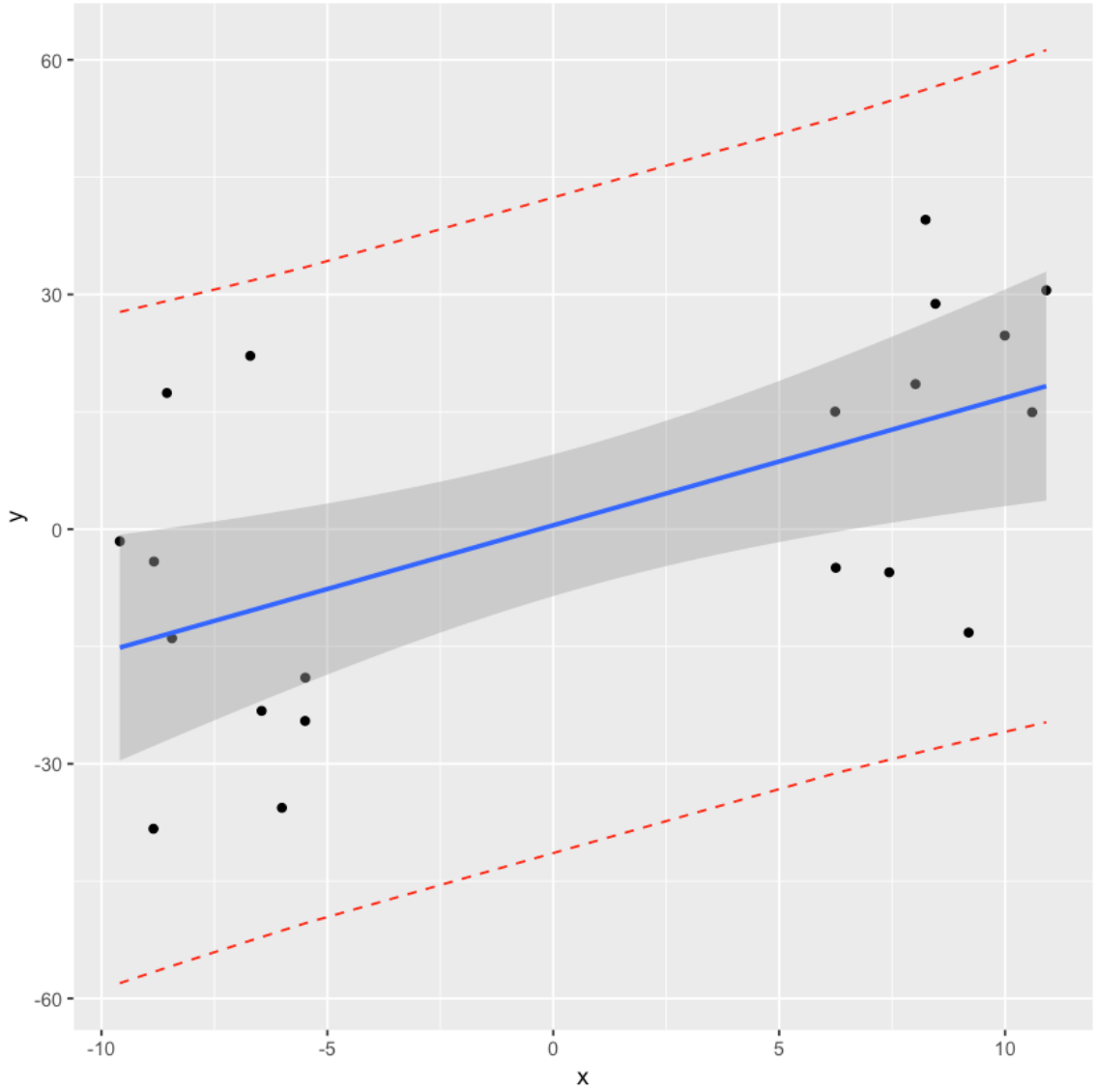
x	y
-5.474091	-11.1908617
-8.652467	-11.5309770
-5.340401	-7.3474393
-5.455141	-0.8683876
-7.170717	-12.9125020
-11.079900	-15.1237204

In [57]:

```
df = gen_data(-8, 8, 10, 10)
lm_gen = lm(y ~ x, data = df)

# Predicton on the training set
df.pred = predict(lm_gen, interval="prediction", level=0.95)
df.fit = df.pred[,1]
df.upper = df.pred[,3]
df.lower = df.pred[,2]

ggplot(df, aes(x, y))+
    geom_point() +
    geom_line(aes(y=df.lower), color = "red", linetype = "dashed")+
    geom_line(aes(y=df.upper), color = "red", linetype = "dashed")+
    geom_smooth(method=lm, se=TRUE)

df2 = gen_data(-8, 8, 20, 20)

lm_gen2 = lm(y ~ x, data = df2)


df2.pred = predict(lm_gen2, interval="prediction", level=0.95)
df2.fit = df2.pred[,1]
df2.upper = df2.pred[,3]
df2.lower = df2.pred[,2]

ggplot(df2, aes(x, y))+
    geom_point() +
    geom_line(aes(y=df2.lower), color = "red", linetype = "dashe
d")+
    geom_line(aes(y=df2.upper), color = "red", linetype = "dashe
d")+
    geom_smooth(method=lm, se=TRUE)
```

```
Warning message in predict.lm(lm_gen, interval = "prediction", level
= 0.95):
“predictions on current data refer to _future_ responses
”Warning message in predict.lm(lm_gen2, interval = "prediction", lev
el = 0.95):
“predictions on current data refer to _future_ responses
”
```





The prediction and confidence intervals become wider with larger variances and smaller with smaller variances (all else equal).

3. (b) The Smallest Interval

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

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

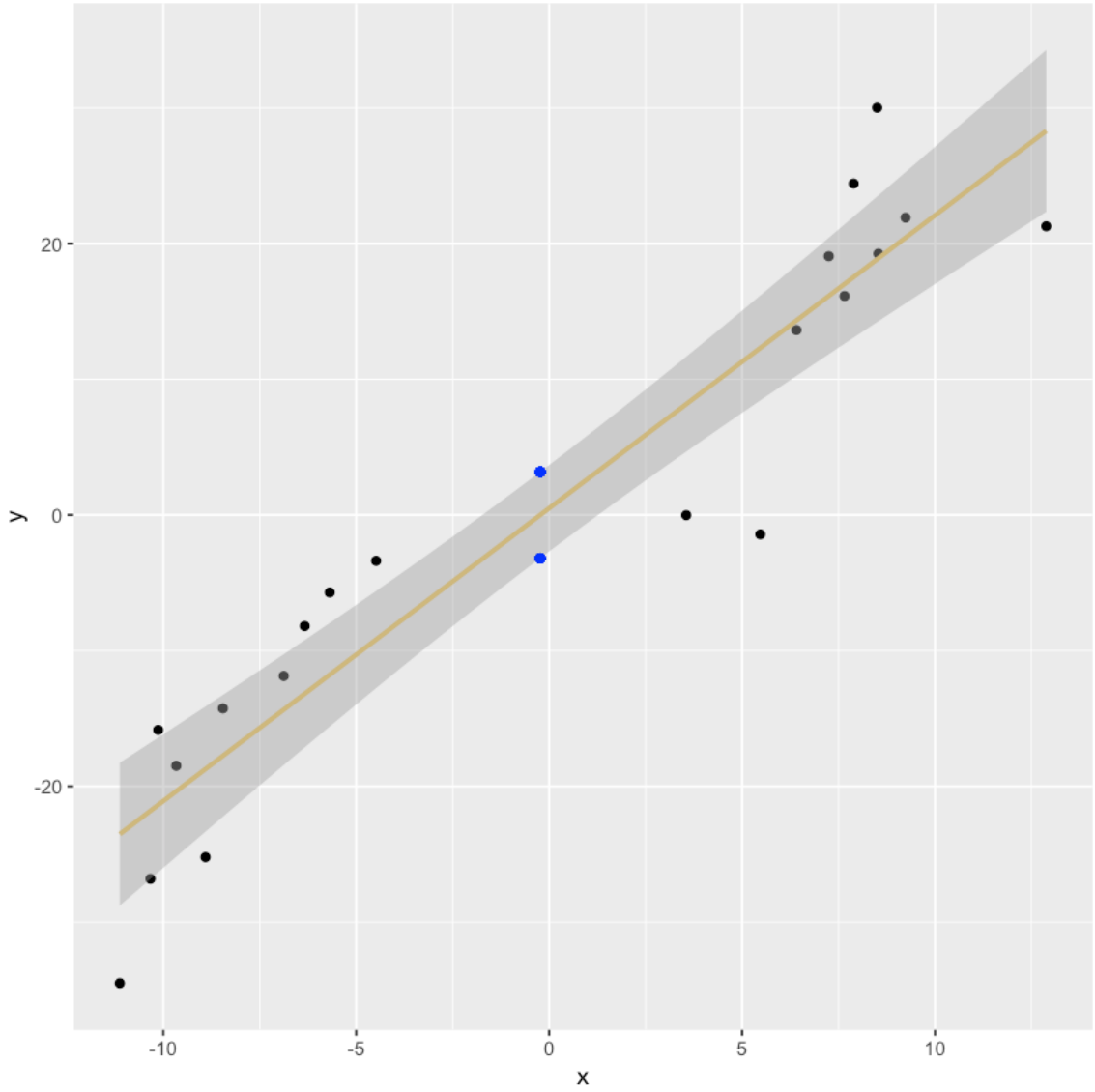
where $\hat{y}_h$ is the fitted response for predictor value x_h, $t_{\alpha/2,n-2}$ is the t-value with $n-2$ degrees of freedom and $MSE \times \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})}\right)$ is the standard error of the fit.

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?

In [73]:
```
x_new = data.frame(x = mean(df$x)); x_new
predict(lm_gen, newdata = x_new, interval = "confidence", level =
0.95)

ggplot(df, aes(x = x, y = y)) +
    geom_point() +
    geom_smooth(method = "lm", col = "#CFB87C", level = 0.95) +
    geom_point(aes(x=-0.2312037, y= -3.181803), colour="blue") +
    geom_point(aes(x=-0.2312037, y= 3.17982), colour="blue")
```

x
-0.2312037

fit	lwr	upr
-0.0009917047	-3.181803	3.17982



$x_k = \bar{x}$ would result in the CI with the shortest width. We can see this verified for the generated data on the plot above.

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.

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. In short, the fundamental problem of causal inference is: at most one of two (or many, in a multilevel treatment experiment) potential outcomes can be observed for a given experimental unit, and thus, causal effects cannot be measured directly. This is a problem for a certain way of approaching science and statistics - sometimes called empiricism - that posits that observations drive the generation of knowledge!
2. Because we cannot observe counterfactuals, we cannot measure the response in counterfactual situation. Close substitutes means that we try to measure the response for an individual sufficiently similar to the one in question.
3. A theory about the nature of causality is *deterministic* if the effect *necessarily* follows from the case. On deterministic accounts of causality, object b necessarily moves y meters when struck by object a. It would be impossible for a to strike b (at the same speed) and not have b move y meters. A theory about the nature of causality is *probabilistic* if the existence of a cause impacts the probability of an effect. For example, the existence of smoking increases the probability of having cancer.

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 (https://lpeproject.org/blog/law-liberation-and-causal-inference/) 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 (https://statmodeling.stat.columbia.edu/2020/07/06/statistical-controversy-on-racial-bias-in-the-criminal-justice-system/)
2. Can Racial Bias in Policing Be Credibly Estimated Using Data Contaminated by Post-Treatment Selection? (https://dcknox.github.io/files/KnoxLoweMummolo_PostTreatmentSelectionPolicing.pdf)
3. A Causal Framework for Observational Studies of Discrimination (https://5harad.com/papers/post-treatment-bias.pdf)

Please read Lily Hu's blog post (https://lpeproject.org/blog/law-liberation-and-causal-inference/) and Andrew Gelman's blog post "Statistical controversy on estimating racial bias in the criminal justice system" (https://statmodeling.stat.columbia.edu/2020/07/06/statistical-controversy-on-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 importan 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?

In [ ]: