Data 100, Spring 2023

Discussion #5 Solutions

Data Collection through Sampling

It's time for the Data 100 midterm, and the professors want to estimate the difficulty of the exam. They decide to survey students on the exam's difficulty with a 10-point scale and then use the mean of the students' responses as the estimate.

(a) What is the population the professors are interested in trying to understand?

- ◯ A. Students in Data 100
- ◯ B. Students enrolled in the Data 100 Ed
- ◯ C. Students who attend the Data 100 lectures
- ◯ D. Students who took the Data 100 midterm

Solution: Some students enrolled in Data 100, on Ed, or who attend lectures may not have taken the midterm exam. We only want to survey students who *actually* took the exam, so (D) is the correct choice.

(b) The professors consider a few different methods for collecting the survey data. Which of the following methods is best? (think through which considerations go into "best")

- ◯ A. The professors send a Zoom poll to all students in the first weekly live session following the exam.
- ◯ B. The professors add a question to the homework assignments of a simple random sample of anonymous students within every discussion section.
- ◯ C. The professors make a post on Ed asking students to submit a Google Form containing the survey question.
- ◯ D. The professors choose a simple random sample of discussion sections, go to each selected section, and ask each student in the group as part of the final discussion question.

Solution:

1. It would be a fairly easy method. However, the quality of the data would be suspect. By sampling only from students who attend the Zoom lecture, the professors are restricting students who prefer not to attend live sessions and students who have time conflicts. These can introduce selection bias in the sample.

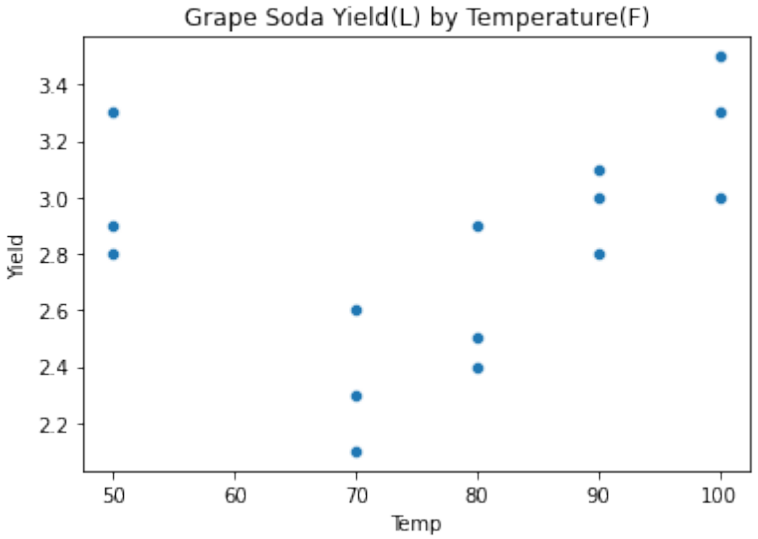
2. This method of sampling works well for this scenario because the professors are sampling randomly from students across discussion times, which takes into account many of the axes along which students can differ. They are also adding the question in such a way that students are required to answer it, reducing non-response bias, and can answer it privately, reducing social pressure.

3. As with (A), there are some students who don't look at Ed or ignore certain posts, which introduces selection bias. Responding to the survey is also optional and easy to forget, introducing non-response bias.

4. The primary issue here is one of social pressure because students are being asked in groups, and this can bias the results of the survey.

Simple Linear Regression

A UC Berkeley College of Chemistry student was watching his favorite chemistry Youtuber NileRed experimenting with turning gloves into grape soda and wanted to try it themselves. The experiment was done at various different temperatures and yielded various different amounts of grape soda. Since this reaction is very costly, they were only able to do it 15 times. This data set of size $n = 15$ (Yield data) contains measurements of yield from an experiment done at five different temperature levels. The variables are y = **yield in liters** and x = **temperature in degrees Fahrenheit**. Below is a scatter plot of our data.



σ_x	σ_y	r	$\bar{x}$	$\bar{y}$
17.20	0.38	0.30	78.00	2.83

(a) Given the above statistics, **calculate** the slope ($\hat{\theta}_1$) and y-intercept ($\hat{\theta}_0$) of the line of best fit using Mean Squared Error (MSE) as our loss function and **plot the line on the graph above**:

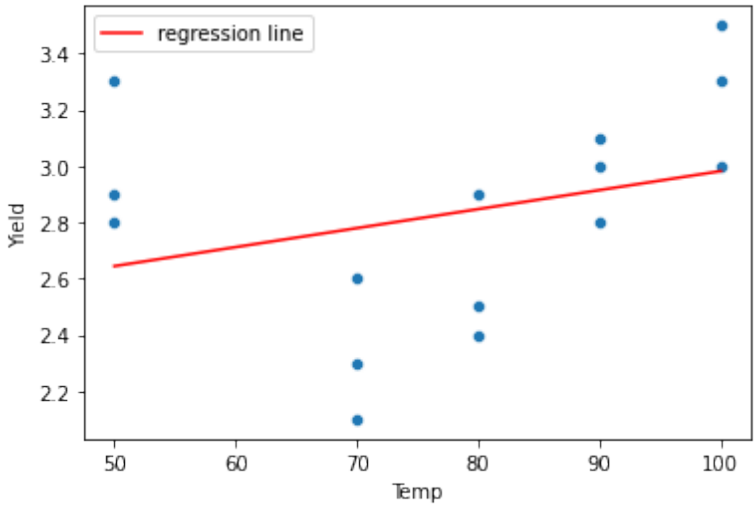
$$y = \hat{\theta}_0 + \hat{\theta}_1 x$$

Solution: $\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}$

$\hat{\theta}_1 = 0.3\frac{0.38}{17.2} = 0.0068$

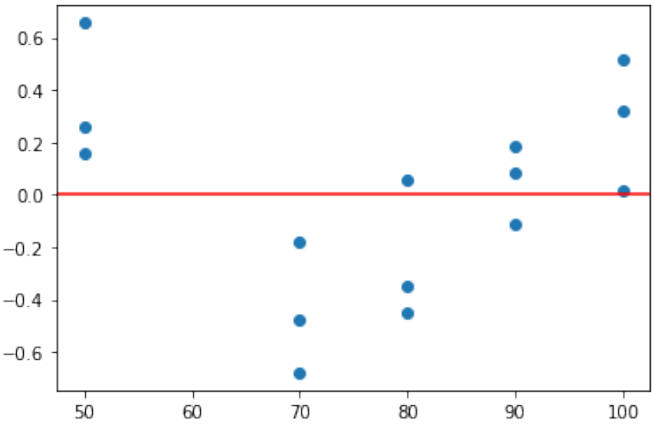
$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1\bar{x}$

$\hat{\theta}_0 = 2.83 - 78.00 * 0.0068 = 2.31$



(b) Plot the residuals from the line of best fit you calculated in part (a). What does the residual plot tell us about the relationship between x and y?

Solution:



The plot of the residuals is no equally variable across all values of x. This means that there is heteroscedasticity in our residuals. Thus, the relationship between x and y is likely not linear. y is likely not linear in terms of x.

(c) Which of the following relations most closely represent the relationship we see between Temperature (x) and Yield (y)?

◯ A. $y = \theta_2 x^2$

◯ B. $y = \theta_2 x^2 + \theta_1 x + \theta_0$

◯ C. $y = \theta_1 \log x + \theta_0$

◯ D. $y = \theta_1 x + \theta_0$

MAE Minimization

In the lecture, we derived the estimating equations for SLR, which we obtained by differentiating the MSE with respect to θ_0 and θ_1. Now, suppose that we choose the same model, $\hat{y} = \theta_0 + \theta_1 x$, but we minimize the L1 loss instead. That is, we minimize the *Mean Absolute Error (MAE)*, defined as

$$\hat{R}(\theta_0, \theta_1) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \theta_0 - \theta_1 x_i|.$$

(a) Show that the partial derivatives of $|y_i - \theta_0 - \theta_1 x_i|$ with respect to θ_0 and θ_1 are:

$$\frac{\partial}{\partial \theta_0} |y_i - \theta_0 - \theta_1 x_i| = -\text{sign}(y_i - \theta_0 - \theta_1 x_i)$$

$$\frac{\partial}{\partial \theta_1} |y_i - \theta_0 - \theta_1 x_i| = -x_i \cdot \text{sign}(y_i - \theta_0 - \theta_1 x_i),$$

and undefined if $y_i = \theta_0 + \theta_1 x_i$, where sign is the sign function defined by

$$sign(z) = \begin{cases} +1 & \text{if } z > 0 \\ -1 & \text{if } z < 0 \\ 0 & \text{if } z = 0 \end{cases}$$

Note: Although the derivative of the MAE function at 0 is undefined, you can use the default value of 0 for the partial derivative of MAE(0) with respect to both θ_0 and θ_1.

Solution: We can split the absolute value into three cases and analyze them separately:

If $y_i > \theta_0 + \theta_1 x_i$, then $|y_i - \theta_0 - \theta_1 x_i| = y_i - \theta_0 - \theta_1 x_i$

so the derivative with respect to θ_0 and θ_1 are -1 and $-x_i$.

If $y_i < \theta_0 + \theta_1 x_i$, then $|y_i - \theta_0 - \theta_1 x_i| = -(y_i - \theta_0 - \theta_1 x_i)$

so the derivative with respect to θ_0 and θ_1 are $+1$ and $+x_i$.

The graph of the absolute value function has a "fold" at $y_i = \theta_0 + \theta_1 x_i$. While the derivative at the "fold" is usually considered undefined, it is often convenient and reasonable to give it a default value of 0.

Finally, we can summarize the above case with the $sign$ function.

(b) (Bonus) Show that the number of points with negative residuals is the same as the number of points with positive residuals.
Hint: Find the partial derivative of the loss function with respect to θ_0 and set it to 0. What do you observe?

Solution: We can simply sum up the previous part across i:

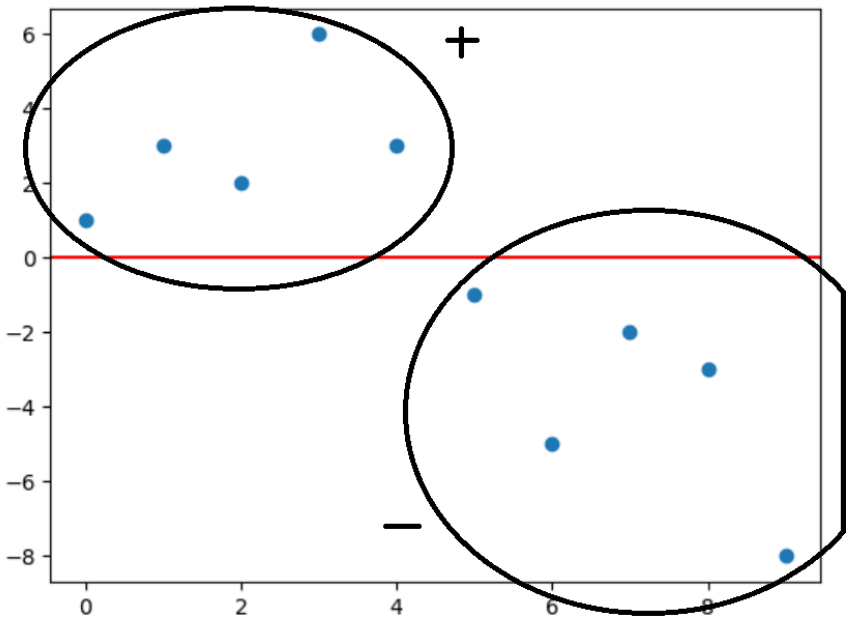
$$\frac{\partial}{\partial \theta_0} \hat{R}(\theta_0, \theta_1) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_0} |y_i - \theta_0 - \theta_1 x_i|$$

$$= \frac{1}{n} \sum_{i=1}^{n} -sign(y_i - \theta_0 - \theta_1 x_i)$$

$$= (1 + 1 + ... + 1) + (-1 - 1... - 1) + (0 + 0 + ... + 0) = 0$$

Where the number of $+1$ is the number of points that $y_i - \theta_0 - \theta_1 x_i < 0$ or negative residuals. Similarly, number of -1 is the number of points that $y_i - \theta_0 - \theta_1 x_i > 0$ or positive residuals. Since the derivative must be 0 at the optimum, we can conclude that the number of points with negative residuals is the same as the number of points with positive residuals.

Note: what this is saying is that there are an equal number of data points above and below the $y = 0$ line in the residual plot.



Similarly (not needed for this question),

$$\frac{\partial}{\partial \theta_1}\hat{R}(\theta_0, \theta_1) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \theta_1}|y_i - \theta_0 - \theta_1 x_i|$$

$$= \frac{1}{n}\sum_{i=1}^{n} -x_i \times sign(y_i - \theta_0 - \theta_1 x_i)$$

The interpretation of this is beyond the scope of this class.

(c) (Bonus) We can often check how sensitive a model is based on estimating equations. That is, how will the estimators θ_0 and θ_1 change when an outlier is introduced. Based on the estimating equations of the linear regression with L2 loss from class:

$$\sum_{i=1}^{n} y_i - \hat{y_i} = 0 \quad \sum_{i=1}^{n}(y_i - \hat{y_i})x_i = 0 \tag{1}$$

What do you think will happen to θ_0 and θ_1 if we take a single y_i value and make it larger and larger while holding all the other x and y values fixed? Consider the three options for what might happen:

1. $\hat{\theta}_0$ and $\hat{\theta}_1$ would not change at all.
2. $\hat{\theta}_0$ and $\hat{\theta}_1$ would change somewhat, but there is a limit to how much they change.
3. $\hat{\theta}_0$ and $\hat{\theta}_1$ can change more and more as we change y_i more and more, without any limit.

You may assume that $x_i \neq \bar{x}$. Explain your answer. If which one of (i)-(iii) happens depends on the data y and x, explain what it depends on (but you do not have to give necessary and sufficient conditions).

Solution:

The answer is (iii). We can tell this from the explicit formulae for $\hat{\theta}_0$ and $\hat{\theta}_1$ from class, or from the estimating equations. In particular, all the residuals have to average out to zero. So if y_i gets larger and larger, then its prediction will have to get larger and larger too, or else its residual will eventually be larger than all the other residuals combined. Making the i^{th} prediction bigger and bigger without any limit will mean changing either $\hat{\theta}_0$ or $\hat{\theta}_1$ (or more likely both) without any limit.

(d) (Bonus) On the other hand, what do you think would happen to $\hat{\theta}_0$ and $\hat{\theta}_1$ for L1-loss linear regression if we take a single value of y_i and make it larger and larger while holding all the other x and y values fixed? Among the three options for what might happen listed above, what do you think would happen? Explain your answer. If which one of (i)-(iii) happens depends on the data y and x, discuss what it depends on (but you do not have to give necessary and sufficient conditions). The estimating equations are written below again for reference:

$$\frac{1}{n}\sum_{i=1}^{n} -sign(y_i - \theta_0 - \theta_1 x_i) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} -x_i \times sign(y_i - \theta_0 - \theta_1 x_i) = 0$$

Solution: The estimating equations, and the derivatives of the empirical risk, do not involve y_i at all, as long as its residual is positive or negative. If the residual for y_i is already positive, then increasing y_i will not change anything about the fit. If the residual is negative, then increasing y_i will potentially change the fit at the point where y_i hits the regression line. Then, as y_i continues to increase, the prediction for y_i might increase along with it. But there will be a limit to how long the line can track y_i. If the line stays touching y_i as y_i goes off to infinity, it would need to either leave all rest of the points behind (in which case the first condition would eventually be violated) or become more and more vertical (in which case the second would eventually be violated). Thus, eventually,

the line has to stop following y_i, and as soon as y_i has a positive residual, the line will stop changing.