ACTIVITY 10

PICKING THE “BEST” LINE

# DIRECTIONS

Have one group member upload this docx file to your Google Drive and share it with your groupmates and your instructor. Name your document “**GroupX\_Activity10”** (where X is your group number). Work together to type up your responses to each question. Download your document as a PDF and submit this to Canvas individually.

# PART 1: FINDING THE BEST LINE

In your pre-work assignment, you examined linear relationships by drawing a line on a scatterplot to summarize the relationship. However, there is quite a bit of subjectivity when it comes to drawing a line in this manner. This begs the question: how do we know **which straight line to draw?**

We will re-examine the data set on age and sight distance from that assignment in the **driversight.tp** data set.

# BRAINSTORMING

1. Re-create your scatterplot from the pre-work with age on the x-axis and sight distance on the y-axis. Place your line of best fit (reference line diagonal) again on the scatterplot similar to how you did it in the pre-work and paste that plot below. Explain what led you to picking this line.
2. Below is a scatterplot with two different straight lines placed on it. Which line do you think better represents the relationship between age and sight distance? That is, which line better “fits” the data? Explain your reasoning.



1. Now consider the plot and the two lines shown below. Can you determine which line is the better fit to the data shown? What methods can we use to help us evaluate how well a line fits? Brainstorm some ways you could measure or quantify how well a line fits the data.



**STOP HERE!**

Wait for your instructor to begin a class discussion on your ideas on how to evaluate how well a line fits. Until then, keep brainstorming other ideas for determining how well a line fits.

# fINDING THE BEST LINE



1. Use the ruler tool to measure the distance between a dot of your choice and the line. How far away is this dot from the line? Is this gap between the dot and the line measured in terms of sight distance or age? Explain what this difference is measuring based on the line.



1. What is the absolute sum of differences for our starting line? Copy the plot with all the differences measured below.
2. Drag around your line now and observe changes in the total absolute difference. **Be sure to drag your line in all possible ways: up and down by dragging from the middle, and tilting it by dragging it from the sides!** What do you notice about the sum of the differences? Describe what happens to the sum as you move the line around.
3. If our goal is to find the line that best fits the data, what should our goal be with the value of the sum of absolute differences? Do we want this value to be small or large? Explain your reasoning.
4. Using the sum of absolute differences, find the best fitting line that you can. Paste an image of the plot with your line below. What is your sum? Check with two other groups around you, what sum of differences did they get?
5. Compare the line you have found based on the sum of absolute differences to the one you initially created in question 1. How are they similar/different? How does this new “best” line surprise you?

# PART 2: USING OUR LINE OF BEST FIT

Now that we have decided on a single line that we believe best fits the data, we can use it to help make predictions for our y-variable based on a fixed value of the x-variable. In this way, we consider the y-variable the **response variable**, and it typically reflects the primary focus in our research question. The x-variable we use is the **explanatory variable**, as we see this variable as one that affects (or “explains”) the response variable.

1. In this study, what is the response variable and the explanatory variable?
2. Give an interpretation of what the slope value means in the context of the problem. How does it help describe the relationship between age and sight distance?
3. Using the equation for the line of best fit, calculate the predicted sight distance for a driver that is 34 years old?

As you might notice, our line doesn’t necessarily go through any of the points. A straight line cannot possibly go through all of the points, which means that if we tried to predict the sight distance for any of our original 60 participants, it would be wrong. The distance between a data point’s predicted and actual value is called a **residual**. We calculate a residual by taking the actual observation of a response variable and subtract the predicted value based on the line of best fit. We can visualize residuals as the lengths of the **vertical black lines between the dots and the line** in the plot that you created in question 8. Observations above the line will have a positive residual and below the line will have a negative residual.

1. In the data set, there is one 34-year-old driver that was studied. What was their actual sight distance? Calculate the residual for this driver using the prediction you calculated in the previous question.
2. Interpret what this residual means – would this person be happy or upset to see their residual be positive/negative?



1. What is the value that you get for your line when using the mean of differences option? Interpret what this value means in terms of the accuracy of making a prediction with your line of best fit.

# PART 3: LEAST SQUARES LINE

As noted in your class discussion, minimizing the absolute deviations can be a bit problematic in finding a single “best” fitting line. In practice, statisticians use the least squares method for finding the line of best fit, which minimizes the sum of the squared residuals for the line. Unfortunately, we cannot replicate this process visually like we did for minimizing absolute deviations, but TinkerPlots can calculate the slope and intercept of the line directly. To do this, create two new columns in your data table and apply these formulas to each of those columns:

linRegrSlope(Age, Distance)
linRegrIntercept(Age, Distance)

These formulas can be found in the formula editor under **Functions -> Statistical -> Two Attributes**. Be careful to choose the correct slope function, there is another function **linRegrSESlope** which is not the correct calculation!

1. What is the equation of the least squares line for age and sight distance?
2. Using your scatterplot with the line we found by minimizing the absolute deviations, add a new diagonal reference line to the plot, and place it so it is as close to the equation you found in the previous question as possible. Paste that plot with both lines on it below.
3. Compare and contrast the two lines in the plot from the previous question. Focus on the slope and relative position of the lines in your comparison.

# PART 4: EVALUATING THE STRENGTH OF THE ASSOCIATION

Now that we have analyzed the line of best fit itself, we may be curious as to how useful the line itself is for making predictions as we did earlier. We calculated the average residual size before to give some idea of this, but this measure is in the units of our response variable, making it difficult to compare with other potential variables. To measure the **strength** of the association, we can use the **correlation**, often denoted by ***R***. To calculate the correlation of two variables in TinkerPlots, we can use the following function found under **Functions -> Statistical -> Two Attributes**:

correlation(Age, Distance)

The value you get from this function will always be between -1 and 1, and the value indicates two ideas: **strength** and **direction**. Positive correlation values indicate an upward sloping relationship, where negative correlation values indicate a downward sloping relationship. Values near 0 indicate that the relationship is not strong, with points scattered far from the line. The closer the correlation value is to -1 or 1, the closer the points are to the line, or the **stronger the relationship is.**

1. Find the correlation value of age and sight distance. What does this value indicate about the direction of the relationship between age and sight distance? Does this agree with the slope you found for the line of best fit?

The ***R*** or **correlation** value gives a relative idea of how strong a relationship is, but we can’t easily interpret exactly what this means in terms of the data itself. However, the square of this value (***R2***), known as the **coefficient of determination**, is more easily interpretable. To find this value, you can simply square the value from the correlation function in TinkerPlots, or you can find it more directly with the following function found under **Functions -> Statistical -> Two Attributes**:

rSquared(Age, Distance)

This value will always be between 0 and 1, and is interpreted as a percentage. Specifically, it is the *percentage of variance in our response variable that can be explained by its linear relationship with the explanatory variable.*

1. Find the correlation of determination (*R2­­*) and interpret what this means in the context of the problem. (Hint: the italicized text above gives a general interpretation, just add this data’s contextual details.)