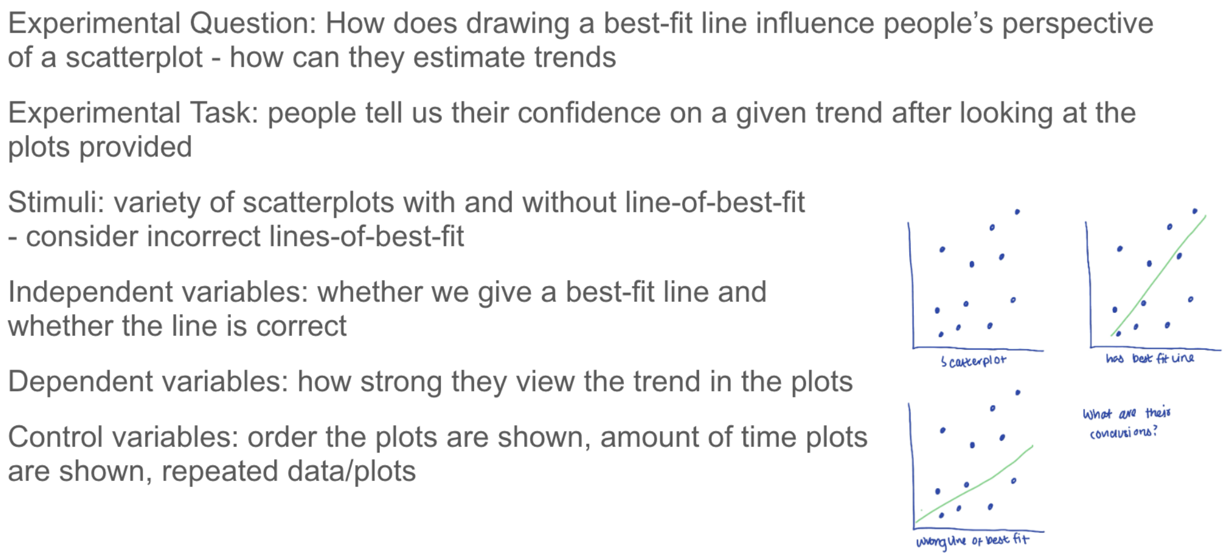
Target Question and Motivation

How do different best fit lines constructed using different regression models on a scatter plot, or the absence of a best fit line entirely, affect the viewer's ability to understand the plot?

We chose this question because lines of best fit are often used in scatterplots, and is presented as a common option in data visualization creation tools (such as Excel). We wanted to see if they had an noticeable impact upon people's understanding, and also see how it differed between graphs with stronger or weaker correlations. Additionally, a line of best fit can be controversial because people may believe it does not accurately model the data or ignore outliers.

Preliminary Design

We created our preliminary design originally in an in-class activity (Week 13 Experimental Design). Below was our slide. We created quick sketch of an example of 3 visualizations (2 with best fit lines and one without) on the same dataset that could be used to probe our target question,



Experimental Question: How does drawing a best-fit line influence people's perspective of a scatterplot - how can they estimate trends

Experimental Task: people tell us their confidence on a given trend after looking at the plots provided

Stimuli: variety of scatterplots with and without line-of-best-fit
- consider incorrect lines-of-best-fit

Independent variables: whether we give a best-fit line and whether the line is correct

Dependent variables: how strong they view the trend in the plots

Control variables: order the plots are shown, amount of time plots are shown, repeated data/plots

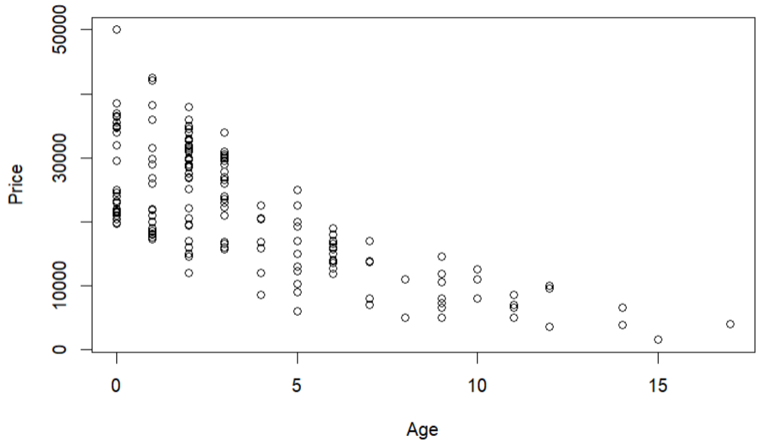
After creating this preliminary design, we sought to adapt it for our final project. We liked the idea of comparing visualizations with and without lines of best fit, as well as our experimental task. However, we decided to not use the wrong line of best fit in our final project, because this is unlikely to occur in the real world. Many tools automatically create a line of best fit upon statistical and mathematical analysis. Instead, we decided to use both linear and nonlinear lines of best fit along with the visualization with no line of best fit. This is the main design choice we made from this preliminary design – to remove the wrong line of best fit, and consider different methods of the line of best fit instead.

Methods

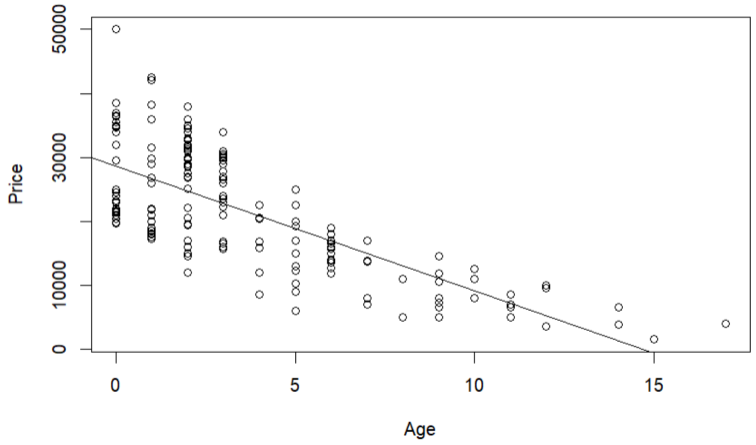
In order to investigate this question, we decided to use an experiment as it would allow us to collect data with a structured environment. We had several independent variables we wanted to consider, such as the correlation of the plot and whether there was a line of best fit or not. We also had several hypotheses on what we would see based upon our own experience encountering lines of best fits. We also wanted to collect more quantitative data than qualitative. After choosing the experiment option, we scoped out our experiment. In order to move from our preliminary design, we needed to add a lot more iterations of the same task, as well as utilize datasets with differing correlation. As well to make the best fit lines accurate and professional, we created visualizations using R and D3, which have packages to accurately draw the lines.

We plan to have people split into 3 different groups, and they will complete either of the 3 surveys independently based upon their group number. There should be an even number of people (we are estimating roughly 5-6 responses per survey) so that we can make better comparisons on how people interpret the graphs depending on the status of the line of best fit. The survey will be 8 sections long, with 4 questions per section.

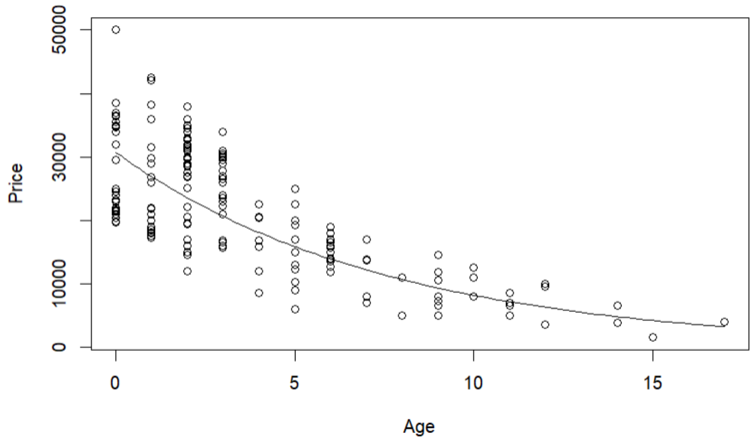
We are splitting people into 3 groups because we are using 8 different datasets. We utilized different parts of the energy dataset from the last module, as well as some others (at the bottom of this section). The specific variables shown in the dataset are not important – rather we want to find variables that correlate to create some sort of trend. For each dataset, we are creating a scatterplot with no line of best fit, a linear line of best fit, and a nonlinear line of best fit. The nonlinear lines of best fit were either logarithmic or quadratic. Within our 8 datasets, 4 of them have strong correlations while 4 have weaker correlations. We were able to split this evenly upon surveys, although for survey 1 and 3, there were 5 scatter plots with lines of best fit, and for survey 2, 6 lines. This was to be expected as we have 24 visualizations in total: 12 with strong correlations, 12 with weak correlations, 8 with no line of best fit, 8 with a linear line of best fit and 8 with a nonlinear line of best fit. Here are the final visualizations for one dataset with a strong correlation:



No line of best fit



Linear line of best fit



Nonlinear line of best fit

Within the 8 section survey, people will see only one of the three graphs of each dataset, to determine that each of the graphs are independently considered. We plan for participants to have 1.5 minutes per section of the survey, which should be ample time for them to consider the questions. The first three questions of the survey will be required (these are the quantitative portions) and the final question will be qualitative and not required unless someone has observations to note.

We used datasets from multiple sources. We utilized the energy dataset from the module 3 assignment, as well as these sources:

- https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/StateSAT.csv
- https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Turtles.csv
- R Stat2Data library
- UsedCar spreadsheet from Jessica's STOR455 class

The independent variables in our study will be:

- The different visualizations the participant will see – our participants will see scatterplots without best fit lines, ones with a linear best fit line and ones with an nonlinear best fit line.
- The ordering of the scatterplots – the participants will be divided into 3 different groups, and they will see different trend lines or no trend lines on each data set. They will not see the same data set with a different scatterplot model.
- The content of the data sets – the data sets that we construct visualizations out of could have a weak, positive, negative, uncertain/neutral type of correlation. This will be reflected in the visualizations we present to the participants.

The dependent variables in our study will be:

- What participants believe the overall trend of the scatterplot is, and how accurately they are able to do so. This will be measured with multiple choice options.
- How accurately participants can predict a data point from a scatterplot that could or could not have a best fit line – they will be given an x, and be asked to estimate a y. This will be measured with a short answer question expected to produce a numerical value.
- How confident participants are in their y value prediction (considering we are manipulating both the correlation of the graph and no best fit line/type of best fit line). This will be on a scale of 1 (not confident) to 10 (very confident).
- What observations participants had about the data shown in the visualization (for instance, did they see any outliers, multiple trends in one graph, or anything unusual?) This will be a long answer question and qualitative .

The control variables in our study will be:

- The number of visualizations each group is given – everyone will have the same number of questions on the Google form.
- The amount of time to look at each visualization – we will have a timer during the experiment and have participants advance to the next section of questions every 1.5 minutes.
- The survey will be the same for each participant group.

- The survey will be conducted in the same environment (Google forms), and will format the same overall for all groups.
- The same four questions are asked over all surveys for each visualization, three required and one non-required.
- The x value we will ask participants to calculate the y for will be the same for each dataset.
- The same ordering of multiple choice options for all participants

Below are our hypotheses about what we will see:

- We anticipate that people will follow the best fit line when it is present, irrespective of the type of best fit line that is there (linear or nonlinear). This is because the human mind has evolved to find patterns even when none exist. A line of best fit is the most obvious pattern people will see, even if inaccurate. People will feel most confident and be more accurate about predicting data points on graphs with a line of best fit and determining the overall trend.
- Similarly, when there are no lines of best fit, we anticipate people will largely be able to fill in the gaps and identify the trend in the data (i.e. where the line of best fit will be). If there is no correlation, they will still attempt to find a trend for the reasons stated above (humans always seek patterns). However, we believe that people will feel less confident about predicting data points on graphs with no line of best fit and do so less accurately. They may also have a harder time predicting the overall trend.
- People will also have an easier time predicting data points and understanding the overall trend of data on graphs where data is showing a strong correlation, no matter if there is a line of best fit or not. They may struggle with these two tasks on ones that show a weak correlation, because they have to observe the data more carefully and make judgment calls, especially within a time limit.

Our experimental tasks will be that participants will complete a Google form with 8 different graphs corresponding to the survey number they were assigned. Since Google forms does not have an inbuilt timer to advance each section, we will ask each participant to advance to the next section upon our timer (1.5 minutes per question). On the Google form, for each section, there will be a picture of the visualization at the top, then 4 questions, the first three of which are required to advance. The four questions will be:

1. What trend does this graph show?
    a. Multiple choice: positive, negative, neutral, unsure
2. Given x = {a value}, what value do you expect for y?
    a. The x value asked for will be constant among the same data set – all 3 iterations of the visualization for the same data set will ask for the same, but different data sets will give a different x value relevant to the graph

    b. A short answer question, expecting numerical values
3. On a scale of 1-10, how confident are you on this answer?
    a. On a scale of 1 (not confident at all) to 10 (certain)
4. Did you notice anything interesting or strange about the data in this graph?
    a. Long answer and not required, this is particularly to see how people feel about the visualization or data that could have affected their response to the past 3 questions

Here are the three different surveys we created. Advancing through the surveys will showcase all of the different visualizations.

- Survey 1
- Survey 2
- Survey 3

Analysis and Discussion

To analyze our data, we organized all of the 8 datasets separately – compiling and then comparing the answers for the visualization with no best fit line, the visualization with the linear best fit line and the visualization with the nonlinear best fit line. After conducting analysis upon our results, we found our hypotheses to mostly hold. We included both graphs with obvious correlations, and others without. We decided to split up our analysis between these two.

When dealing with graphs with a noticeable correlation (positive or negative), we noticed that people felt about the same confidence in their answers – even if there was a line of best fit or not. The confidence averaged to around 5 to 6 out of 10. We noticed that the y values they predicted were accurate, or close to accurate. When the trendline is linear, people's guesses of the y-value become even more precise. Something interesting we noticed was that on plots with a linear trendline, people's answers contained decimal points and ended in numbers other than 0 and 5. They were more precise. This is in contrast to the scatterplots with no line of best fit – people's answers usually end in 0 or 5. When the trend was noticeable, people also correctly selected the trend, and there were few unsure or neutral answers. This is in line with our hypothesis that people will have an easier time predicting data values and determining the trend with a strong correlation. It was surprising to see that a line of best fit did not affect how confident people felt about graphs with strong correlation, but simply helped them be more accurate.

When dealing with graphs without a noticeable correlation, the answers were more mixed. When there is no line of best fit, people were answering that the trend was neutral or that they were unsure. When there was a line of best fit, people would answer in accordance with the shape of the line. People's confidence levels also varied. We found that people were most confident here predicting data values when there was a linear line of best fit (3x as much as no line of best fit), then when there was a nonlinear line of best fit (2x as much as no line of best fit) and least when there was no line of best fit. This definitely fit in with our hypothesis; although we did not realize that whether the trend line was linear or nonlinear made a difference. In terms of the data values that were predicted, when there was a line, people tended to follow the line, although some people did not (perhaps implying that they believed the line of best fit did not accurately capture the trend. When there was no line, people's answers were much varied with a larger range. In particular, one participant tried to answer this sequence by giving a sequence of different values.

We also had a qualitative question about whether there was anything interesting or strange about the graph. This was not a required question, and while we did get many answers near the beginning of the survey, by the end, there were fewer. People mentioned if they noticed if the points were clustered or super spread out. Additionally, people disliked the weak correlation plots with no lines. We also asked for participants' observations about the whole activity right after the experiment. People commented that they trusted the trendline, and it made them more confident in their answers. This was reflected in the data we collected. In graphs with no clear correlation, people noted they were less confident about their answers. These initial reactions matched with our more detailed analysis, as well as with our hypotheses overall.

Outside Voices

To get opinions from those without a background in data visualization, we recruited 3 other participants to be a part of our study. They are:

- Jesse Wei, a graduate student in Computer Science and TA for COMP 311 at UNC-CH
- Eric Schneider, a graduate student studying Computer Science at UNC-CH
- Christine Mendoza, an undergraduate Computer Science student at UNC-CH

Their responses are marked on the spreadsheet with their names specified in the "Outside Voice" column. None of these outside respondents produced quantifiably different survey responses. Additionally, all 3 opined that the existence of trendlines in the survey they completed heavily influenced their opinion on the trend that they believed the dataset to possess, consistent with both our hypothesis and the responses of other participants.

Respondent Jesse Wei, deviating from other responders, stated that they felt that the phrase "trend" when asking the participant for their opinion on the dataset was too ambiguous and unclear. Instead, he argued, we should have used the phrase "negative or positive correlation."

Recommendations and Summary

To begin our activity, we assigned participants into either group 1, 2, or 3. There were supposed to be 5 people completing survey 1, 5 people completing survey 2 and 5 people completing survey 3. We told participants to scan the QR code corresponding to their number – the three . After submissions, we noticed that 5 people completed survey 1, 7 completed survey 2 and only 3 people completed survey. To help get some additional data on survey 3, the additional voices we incorporated were sent survey 1 and/or survey 3. Within the experiment time, we should have enforced people completing the correct survey so we would get an even spread of results.

We also attempted to enforce the timer - 1.5 minutes per question. However, it was difficult to mandate this because Google forms does not have a timer, and any extensions that provided a time only did so for the entire form. Instead, we opted to have a timer on screen, set for 1.5 minutes per question. For the first section, we found it difficult to set up the timer, which may have slightly affected the results. However, we did get the timer working for all other sections. We did get several responses earlier than the expected end time. In another iteration of our experiment, we would make sure to mandate the timing, so people cannot skip ahead on sections. The timing was intended to be a variable that we controlled, but we were not able to fully do so during the experiment.

In a future iteration, we would make sure that everyone is completing the survey they are assigned – perhaps participants could be emailed the link. Since we had all of the QR codes on the screen, it's likely people clicked the first one that popped up. Additionally, we would utilize a different tool that advanced to the next section on a timer – this would likely be paid software. In our analysis, we also noted that people wrote fewer answers to the qualitative portion of the survey – this was not required. However, we could have made this question required so that all graphs were given an equal amount of consideration by participants. One of our outside voices, Jesse, also commented that the phrase "trend" was too ambiguous, and we should have reworded the question to ask for "negative or positive correlation". This perhaps would better help participants understand their task.

In summary, we found that people better understood the plot when there was a trend line, particularly one that is linear. They are more accurate and confident in their predictions. People remain confident, but not as accurate with interpreting data from visualizations with a strong correlation, but no best fit line. However, when there is a scatter plot with weaker correlation and no best fit line, people generally struggle to interpret the data, whereas a best line provides guidance in interpretation. We would add the refinements of adding a strict timer, enforcing what survey participants, turning the last question required in order, and rewording the first question to facilitate better data collection in future runs of the experiment.