Replication of "Regression by Eye: Estimating Trends in Bivariate Visualizations" by Michael Correll and Jeffrey Heer

Tim Brlan, Alix d'Agostino, Mathias Lüthi, Natalia Obukhova,
Alexander Pfyffer, Bianca Stancu
University of Zurich, Zurich, Switzerland
{timothy.brlan, alix.dagostino, mathias.luethi, natalia.obukhova,
alexander.pfyffer, bianca.stancu}@uzh.ch

INTRODUCTION

Finding trends in bivariate data visualizations such as scatter plots is an important analytical task. However, since many charts do not explicitly include trend lines or related statistical summaries, viewers often have to estimate trends of a graph by eye. The question arises, whether the inferences drawn by viewers are reliable. In their original study, Michael Correll and Jeffrey Heer used crowdsourced experiments to assess the accuracy of trends estimated using so called *regression by eye* across a variety of bivariate data visualizations (Correll and Heer 2017). In crowdsourced experiments, quality control is a major challenge because of varying motivation and skill of participants. Furthermore, the experimenter has little or no information regarding the background and profile of the crowdworkers (Gadiraju et al. 2017). For instance, even though participants were selected based on their education level, crowdsourcing workers may have provided wrong information if they have a monetary incentive to do so. Therefore, we want to find out whether the results also hold in a laboratory setting, where we can control for both the correct execution of the experimental tasks and the demographic backgrounds of the participants.

We replicated the first out of the three experiments from the original study, which specifically looks at estimation of the slope in plots. We found it to be paramount to the findings of the entire paper, since the second and third experiments simply added to the first by including bar charts and outliers respectively.

In summary, we contribute:

- Better generalization by using a controlled laboratory setting instead of crowdsourcing participants;
- Gathering more complete demographic background information and including them in our analysis;
- Equal gender ratio compared to the 70% male ratio in the original study;
- Including only participants who are at least undergraduates.

RELATED WORK

The original study from Correll (2017) concluded, that viewers without statistical training accurately estimated trends in many standard visualizations of bivariate data. Specifically, "larger residuals result in less accuracy at regression by eye". Moreover, there was no statistically significant difference in estimation accuracy among linear, quadratic, or trigonometric trends. Finally, there was no statistically significant bias in estimations.

METHOD

The goal of this experiment was to examine how accurate participants are at estimating trends in bivariate visualizations. To measure this we conducted a study, where participants estimated trend lines for generated graphs based on predefined conditions. Participants had to perform these estimation tasks using an interactive software in a lab setting. In line with the original study, the primary dependent variable is accuracy, defined as the error between the user-defined trend line and the correct trend line. The code of the software and the collected data can be found on our Github: https://github.com/mathiasluethi/Replication Experiment-RegressionByEye.

In the original study the effect size was reported using the interquartile mean of the absolute error, because crowdsourced experiments often introduce long-tailed distributions of error (Heer and Bostock 2010). Since this is not an issue in laboratory experiments, we decided to use all collected data in the analysis.

EXPERIMENTAL INTERFACE

We developed a custom interactive software running on a website, that presented the participant with three types of

common bivariate visualizations: scatter plots, line graphs, and area charts. For each plot, they had to estimate the slope of the trend line with a slider to best fit the data, as shown in Figure 1. By adjusting the slider, the graph line could be rotated such that the middle of the graph was always at the center of the plot. The participants confirmed their choice by clicking the "next" button, which led them to the next plot.

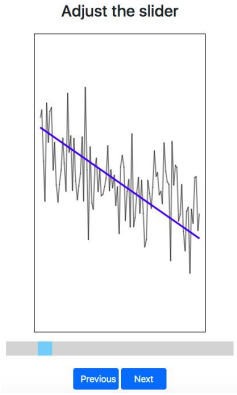

Figure 1: An example estimation task from our experiment. In this case, the participant had to fit the slope of the purple trend to the line graph by adjusting the slider.

DATA GENERATION

We used the same script as the original study to generate the bivariate visualizations (available at https://github.com/uwdata/trend-bias). The standard ordinary least squares (OLS) model in combination with Gaussian sampling was used to generate 100 points for each plot. The script generated all the combinations of three trend types, three chart types, six bandwidths of residuals and eight slopes. The set of generated data includes images with and without trend lines on them. Graphs with the trend lines on them were used for validation tasks. Each participant observed a slightly different visualizations, but the parameters remained the same. All the pregenerated data is also available on our GitHub.

PARTICIPANTS

We randomly selected 19 university students at the University of Zürich. We excluded and replaced one participant due to a technical error during the experiment and excluded the last two participants because we needed a multiple of 8 participants to conduct this study. As a compensation, the participants were offered a chocolate bar after finishing the experiment.

We analyzed data from 16 participants (8 female, 8 male, M_{age} = 24.2, SD_{age} = 4.6). 9 were bachelor students, 5 were master students and one was pursuing a Ph.D. 11 were studying in the psychological faculty and 5 in the business faculty. Everyone used a computer at least once a week and 11 work with charts and graphs at least once a month. 13 out of the 16 participants were Swiss.

EXPERIMENTAL DESIGN

The goal of this experiment is to examine how accurate participants are at estimating the magnitude (slope, amplitude or curvature) of trends in bivariate visualizations. To measure this, we conducted an experiment, where participants estimated trend lines for generated data visualizations based on predefined conditions.

Each participant viewed a combination of three chart types, eight possible slopes and four bandwidths of Gaussian residuals, for a total of 100 (96 + 4 validations) stimuli. Each participant performed three additional practice trials at the beginning of the session that were excluded from the analysis. The data was pre-generated for each participant. These gave us the following set of independent variables:

- Chart type with three levels: scatter plot, line graph, and area chart;
- Slope with eight levels: $\beta = \{-0.8, -0.4, -0.2, -0.1, 0.1, 0.2, 0.4, 0.8\}$
- Bandwidth of Gaussian residuals with four levels: $\sigma = \{0.05, 0.1, 0.15, 0.2\}$.

These independent variables were tested by a within-subject design, meaning that each participant saw a visualization of every combination of these three independent variables. The chart type variable was randomized, but remains in a block so the participant did not have to switch context between every stimulus. The remaining two variables followed the latin square, each in their own block.

There were three types of trends (linear, quadratic, trigonometric) used as a random factor and each type is assigned 32 times to a stimulus. The slope of each trend type could be adjusted differently with the slider. The trend types were evenly distributed between the 96 stimuli.

Before the start of the experiment, they completed three practice tasks without recording the data in order to exclude any potential misunderstandings. For the practice tasks, one of each trend type and chart type was performed, therefore, variables have been chosen to be the same for all participants and are as follows:

- Scatter plot, slope $\beta = -0.4$, Gaussian residuals $\sigma = 0.15$, trigonometric trend;
- Line graph, slope $\beta = 0.8$, Gaussian residuals $\sigma = 0.1$, quadratic trend;
- Area chart, slope $\beta = 0.2$, Gaussian residuals $\sigma = 0.2$, linear trend.

While the trend line was linear the slider controlled the slope of the rendered trend line. In the case of the quadratic trends, the slider changed the curvature and for trigonometric fits, it changed the positive/negative amplitude of the trend line. We included four validation stimuli for each participant; tasks where the correct trend line was already shown in the chart. These data points were excluded from the analysis.

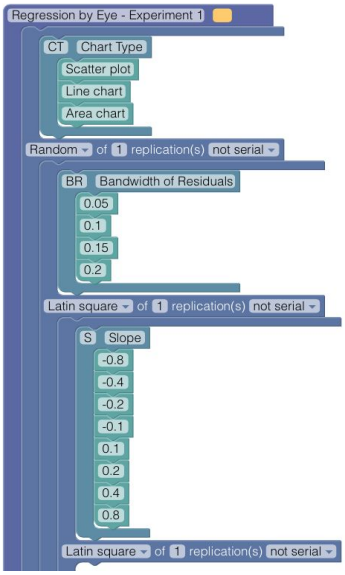

Figure 2: Experimental design made with Touchstone2

To keep interruptions and switching tasks at a minimum, all participants got a general introduction and then worked through all tasks without any significant breaks. We did not impose a time limit. We decided to make three blocks based on the chart type so the participants were not interrupted by the change of context for each task (*Figure 2*). Since we needed a multiple of 8 participants, we decided to recruit 16 to have a full effect coverage. We gave each participant a set of instructions about their tasks and how to use the experiment software.

We chose a random counterbalancing strategy for the chart type and the latin square for bandwidth and the slope. Then, our last factor, the three trend types, were randomly assigned 32 times each to the 96 tasks.

Hypotheses

Our main research question is "How do the type of chart (scatter plot, line graph, area chart), the bandwidth of the residuals and trend type (linear, quadratic, trigonometric) affect the estimated trend of the bivariate visualization by an individual?" In order to answer this question, we tested the following three hypotheses:

1. Estimates are unbiased. This means that the participants will see a balanced set of positive and negative trends with no bias in estimations from under- or overshooting.
2. As the residuals increase, the accuracy decreases, due to the fact that the graphs will be more scattered and less precise, making it harder to estimate accurately,
3. The type of trend has no statistically different effect on participants' estimates. This is a negation of a hypothesis tested in the original study, due to the fact that it was originally rejected.

DATA ANALYSIS

Before testing the hypotheses, a series of preprocessing steps were undertaken. These include eliminating participants whose average of absolute errors on validation tasks was greater than 0.2 (none in our case), eliminating the validation rounds, proper encoding of the variables, as well as calculating the error as: *error = correct_slope - participant_slope*. An overview of the results per participant can be seen in *Figure 3*.

Our first regression model was created with the following code:

```
m_regression1 <- aov(unsignedError ~ sigma *
graphtype * type + id + m, data)
```

We added the sign to the slope (m) and factored type, graphtype, slope, and id. Then, we ran a Three-Way Anova. As independent variables we used the residual bandwidth (sigma), graph type and trend type. As for the dependent variable we used the unsigned error and participant ids and as covariates we used the slope.

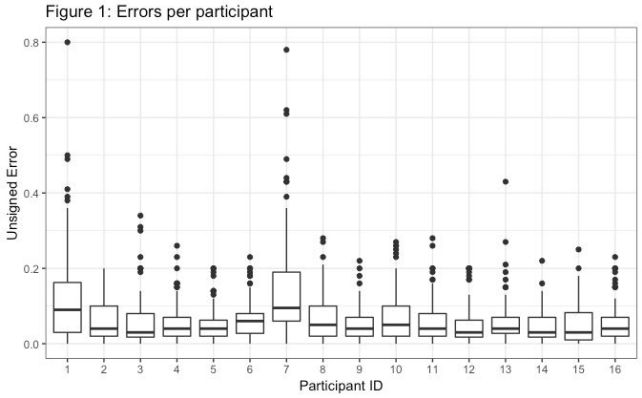

Figure 3. Error box plots for unsigned errors per participant

This model provided us with similar results to our source study, when utilising the identical dataset from the source study. We did, however, encounter substantial differences.

For hypothesis 1, our degrees of freedom are 4519 compared to 4554 and the main variable is sigma which has an F-value of 1157 compared to 950 from the original study; although, the hypothesis is supported either way.

For hypothesis 2, this is where the deviation is an issue. The degrees of freedom have the identical difference, while the main variable is "type" which has an F-value of 6.3 compared to 2.6 from the study. The issue is that our variable is significant and the one in the study is not at a 5 percent level. We therefore knew there must be a difference between the original model and ours. We contacted our professor and he provided us with the following model:

```
m_regression2 <- lmer(unsignedError ~ sigma * graphtype * type + (1|id) + m, data)
```

This second regression model contains two major differences. The choice of model is the first difference, as the first model used an Analysis of Variance Model based on a linear model with fixed effects, while model two used a linear mixed-effect model categorised by 'lmer'. The reason behind this change is that in our first model we assume all observations to be independent from each other. However, this might not be the case since each participant has multiple observations, which can create a dependency (Winter 2013). Hence, a mixed-effect model was utilised with both fixed and random effects, where the random effect creates interdependence of specific observations (Winter 2013), as seen in *Figure 3*.

The second difference is the replacement of 'id' with '(1|id)', where random effect terms are notated by vertical bars (Bolker 2013). What this term signifies is that the intercept is different for each subject and '1' represents the intercept (Winter 2013). This directly alleviates the issue of non-interdependence mentioned above, since now we factor in the effect of multiple responses from each participant (Winter 2013).

Hypothesis 1

In order to test the first hypothesis, we performed a Student T test on the signed mean $mu_{error} = 0$. In order to confirm our hypothesis, we would expect to see the mean of the signed error to be close to 0.

Hypothesis 2

For this hypothesis, a three-way analysis of covariances (ANCOVA) was performed. The goal was to identify whether there is a significant statistical effect of the bandwidth on the unsigned error. The equation of the model used is the following: *unsignedError ~ sigma * graphtype * type + (1|id) + m*. The model is also assessing the effect of the graph type and trend type.

Hypothesis 3

For this hypothesis, the same model was used to interpret whether there is a significant effect of the trend type. Post-hoc tests using Tukey's Honest Significant Difference were also undertaken in order to check for the existence of pairwise interactions, after eliminating the interaction factors from the model.

RESULTS

Hypothesis 1

Unlike the original study, we do not confirm the first hypothesis. The Student's T-test rejected the null hypothesis that $mu_{error} = 0$ ($t(1535) = -3.3823$, $p = 0.0007$).

The mean error was –0.0087, whose absolute value is almost 10 times larger than the absolute value of the mean error in the original study. However, the value is still more than 10 times lower than the slider fidelity. A graphical representation of the errors can be seen in *Figure 4*.

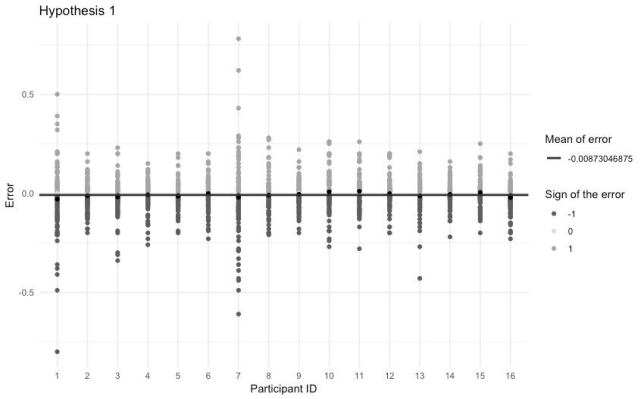

Figure 4. No bias in estimations from under- or overshooting

Hypothesis 2

The results support the second hypothesis. We observed a significant effect of the bandwidth ($F(3, 1478) = 96.20$, $p < 0.001$). *Figure 5* showcases how the mean of the absolute error increases as the bandwidth of the residuals increases. It starts from 0.02 when sigma is 0.05 and increases up to 0.1. The results support the findings of the replicated study.

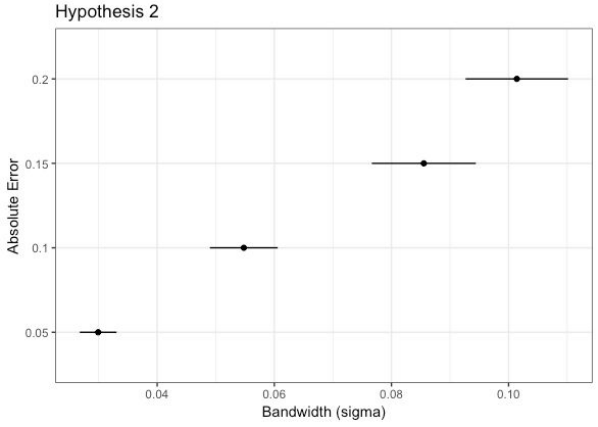

Figure 5. The effect of increased residual bandwidth on error

Hypothesis 3

The results support the third hypothesis. The effect of the trend type was not significant ($F(2, 1478) = 1.86$, $p > 0.1$). Furthermore, the post-hoc tests did not identify any significant pairwise interactions. Hence, as seen in *Figure 6*

the more complex trend types do not result in a significantly higher error.

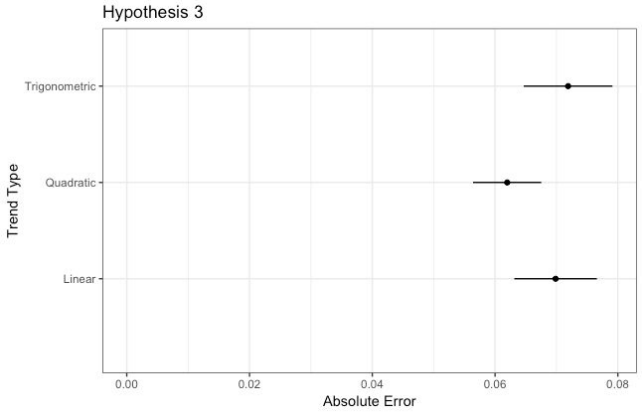

Figure 6. The effect of trend type on signed error

DISCUSSION

Although we tried to replicate the original study as close as possible there were some key differences in our experimental setup. The software used during the experiment was as similar as possible and differences should be negligible. The experiments were done in a more controlled environment where all participants did their tasks in the same setting, using the same hardware.

While the original study was based on crowdsourced experiments and used Amazon's Mechanical Turk to find participants for their study, we recruited participants by approaching people at a university building. There were only 16 participants tested who all had similar demographic backgrounds. This had effects on both our data processing and analysis.

While the original study used the interquartile mean for their data analysis to remove outliers, there was no need to do this for the replication. The full dataset could be used, because the gathered data in this controlled experiment was of higher quality. The data quality was tested with the four control tasks each participant had to complete, this showed that all participants were in the allowed tolerance level and no data had to be filtered out.

Hypothesis 1

We were not able to replicate the same results for hypothesis 1 as in the original study. Our mean error for under or overestimating trend lines was almost 10 times larger than in the original study. Participants in our study slightly underestimated the trend lines regardless of the graph type.

These results might be caused by the lower number and the different demographic background of our participants. We performed a linear regression to test the effects of the demographic variables on the error and found the graph use variable to have a small coefficient of -0.011, significant at 10%. We then split the data into two subsets based on the binary "graph use" variable. New Student's T-test were executed, again with a confidence level of 95% and the null hypothesis $mu_{error} = 0$. For the participants who are analyzing graphs more than once a month the null hypothesis was rejected ($t(1055) = -3.5564$, $p = 0.0003782$), however, it was not for those who analyze graphs less than once a month ($t(479) = -0.41803$, $p = 0.6761$). This shows that people who analyze graphs more often slightly underestimate trends compared to the ones who do not. The former's results are more precise, as the p-value for the graph users is inferior to 0.001 and as we can see in *Figure 7* with the confidence intervals (7.6% smaller confidence intervals). This graph analysis variable could partially explain the differences in our results, yet to form better conclusions this effect would have to be studied by future research.

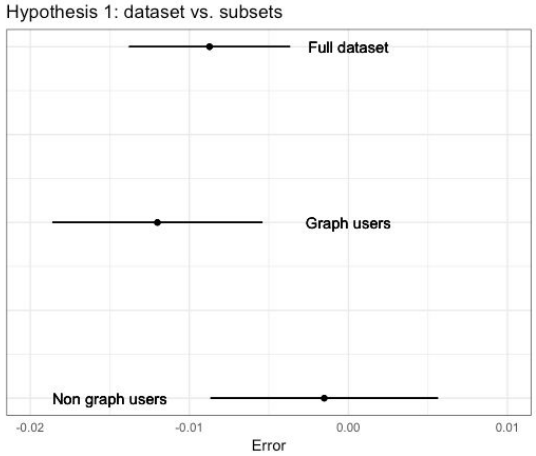

Figure 7. The comparison of the Student's T-tests between the full dataset, the graph users subset with the non-graph users subset

This finding that experienced people were underestimating trends more frequently also coincides with the fact that the participants of this replication were generally higher educated than the participants of the original study. 100% of our participants have a degree or study in tertiary education, while only 83% of the participants of the original study studied at an American college level or higher.

Hypothesis 2 and 3

Hypothesis 2 and 3 were confirmed by our experiment. This shows that the findings in the original crowdsourced study can be replicated in a controlled laboratory experiment.

LIMITATIONS

Much of the experiment setup was identical to the original study. Therefore, we share many of the limitations from the original setup. Most notably it concerns the simple experimental design, with only one parameter to change. In most actual regressions by eye, the viewer makes different

estimations simultaneously, such as the type of fit and the estimation of the model's parameters while ignoring outliers. Errors in any one of these steps could add up, resulting in performance worse than our measures.

Limitations of our specific setup are all related to the lower number of participants. The experiment was done with only 16 participants most of which have a similar educational background. In order to address these similarities, we collected additional demographic information about each participant to check for biases in our sample.

CONCLUSION

In this replication study we examined the ability of viewers to estimate regressions of bivariate visualizations in a lab setting. Firstly, we were able to replicate and show that as the bandwidth of residuals in a visualization increases, regression lines are harder to estimate by eye. Secondly, we showed that the type of trend has no significant effect on the people's ability to estimate trend lines.

However, we were not able to replicate the results that showed that trend estimates are unbiased. Additionally, our data showed that participants who have experience with working with graphs were underestimating the slope of trend lines.

REFERENCES

1. Bolker, Ben. 2013. "Fit Linear Mixed-Effects Models." R Documentation. July 10, 2013. https://www.rdocumentation.org/packages/lme4/versions/1.1-21/topics/lmer.

2. Correll, Michael, and Jeffrey Heer. 2017. "Regression by Eye: Estimating Trends in Bivariate Visualizations." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1387–96. CHI '17. New York, NY, USA: ACM.

3. Gadiraju, Ujwal, Sebastian Möller, Martin Nöllenburg, Dietmar Saupe, Sebastian Egger-Lampl, Daniel Archambault, and Brian Fisher. 2017. "Crowdsourcing Versus the Laboratory: Towards Human-Centered Experiments Using the Crowd." In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*, edited by Daniel Archambault, Helen Purchase, and Tobias Hoßfeld, 10264:6–26. Lecture Notes in Computer Science. Cham: Springer International Publishing.

4.. Heer, Jeffrey, and Michael Bostock. 2010. "Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 203–12. CHI '10. New York, NY, USA: ACM.

5. Winter, Bodo. 2013. "A Very Basic Tutorial for Performing Linear Mixed Effects Analyses (Tutorial 2)." Bodowinter.com. January 19, 2013. http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf.