Regression by Eye: Estimating Trends in Bivariate Visualizations

Michael Correll
University of Washington
mcorrell@cs.washington.edu

Jeffrey Heer
University of Washington
jheer@cs.washington.edu

ABSTRACT

Observing trends and predicting future values are common tasks for viewers of bivariate data visualizations. As many charts do not explicitly include trend lines or related statistical summaries, viewers often visually estimate trends directly from a plot. How reliable are the inferences viewers draw when performing such *regression by eye*? Do particular visualization designs or data features bias trend perception? We present a series of crowdsourced experiments that assess the accuracy of trends estimated using regression by eye across a variety of bivariate visualizations, and examine potential sources of bias in these estimations. We find that viewers accurately estimate trends in many standard visualizations of bivariate data, but that both visual features (e.g., "within-the-bar" bias) and data features (e.g., the presence of outliers) can result in visual estimates that systematically diverge from standard least-squares regression models.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation (e.g., HCI): Evaluation/Methodology

Author Keywords

Information Visualization; Graphical Perception; Regression

INTRODUCTION

An oft-cited example of the power of data visualization is Anscombe's quartet [2]: a set of four bivariate datasets with identical summary statistics, but with qualitatively different patterns when drawn as four scatter plots (Fig. 1). This example relies on the fact that people have the ability to *perceptually* estimate *statistical* quantities of interest. Visualization users regularly perform statistical tasks — including model selection, identification of outliers, and estimation of summary statistics — entirely through visual inspection. Recent work examines the accuracy of visual estimation of means in scatter plots [14] and time series data [10], and speculates on the affordances of visualizations in general for supporting visual estimation of summary statistics [29].

The estimation of trends in bivariate data is an important analytical task, as it is the basis for many factors relevant to decision-making, such as prediction, imputation, and comparison. However, model information is not always explicitly included by visualization designers. When designers do include trend information (for instance, by annotating a scatter plot with a line of best fit), other statistics relevant to the model (such as *r* values or confidence bands) may be absent. Viewers must therefore perform visual estimation to gain a sense of any relevant statistics not provided.

Even if designers include modeling information, the audience may lack the statistical expertise to interpret these values, or may be misled if the data violate modeling assumptions. As a further complication, the form of visual encoding may influence viewers' inferences. For instance, viewers may be more likely to consider trends with line charts, and to compare individual values with bar charts [33]. Visual design choices can also introduce bias, such as the visual asymmetry of bars causing "within-the-bar" bias [27]. Designers would benefit from guidance regarding how accurately viewers make trend estimations by eye, and to what degree different visualization types might bias these estimations.

Knowing the strengths and limitations of such "estimation by eye" is therefore important for designers of data visualizations seeking to communicate statistical quantities, especially to a general audience. On the one hand, visual estimations may be too *inaccurate* for the use cases intended by a designer, or they may be *biased*, leading to systematic over- or under-estimations. On the other hand, while visual estimation lacks the precision of formal statistics, it may be relatively unencumbered by modeling assumptions.

In this work, we describe a series of crowdsourced experiments on the visual estimation of trends in common bivariate visualizations such as scatter plots, area charts, and line graphs. We present the results of three studies investigating estimation of trend slope, trend intercept, and the effect of outliers. We find that, while in most cases viewers accurately estimate trends, area charts introduce systematic under-estimation of trend intercept, and that viewers give low weights to extreme values when estimating trends. These results suggest that there are several areas where human judgments diverge from the fitted models generated by techniques such as Ordinary Least Squares (OLS), and that the design of bivariate visualizations can introduce additional biases in these judgments.

Designers should therefore make an informed decision between designing for regression by eye and the explicit annota-

CHI 2017, May 6-11, 2017, Denver, CO, USA.
Copyright © 2017 ACM ISBN 978-1-4503-4655-9/17/05 ...$15.00.
http://dx.doi.org/10.1145/3025453.3025922

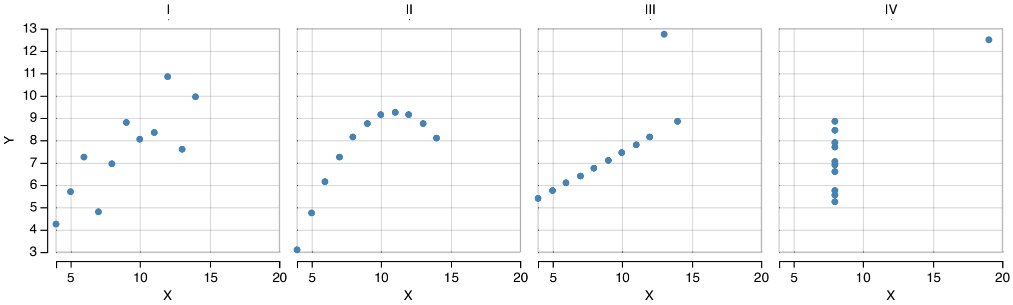


Figure 1. Anscombe's quartet. Each series has nearly identical summary statistics include mean, standard deviation, and linear fit. Yet, through visual inspection, viewers can disambiguate differing patterns and trends. We refer to this visual estimation of trends as "regression by eye."

tion of statistical regression information. Regression by eye affords quick estimations and flexibility in model selections, but is subject to perceptual biases and inaccuracies in estimation. Explicit annotation affords accuracy, but constrains the model space, and adds visual complexity to plots.

RELATED WORK

While there is a great deal of foundational work in visualization and graphical perception dealing with the estimation of individual values in visualizations (such as the height of the bar in a bar chart, or the angle of a line in a line graph), there is comparatively little work on how viewers of visualizations perceive aggregate statistical quantities.

Ariely [3] suggests that, in concert with the perception of individual objects, we also collect information about the *ensemble* properties of visual displays. Szafir et al. [29] note that this *ensemble coding* might afford relatively accurate estimation of *summary statistics* in visualizations. However, visualizations with good performance for summary tasks may not result in good performance at point tasks, and vice versa [1, 13]. A further difficulty is that tasks requiring estimation of values in visualizations (where there is a single correct answer) are qualitatively different from tasks requiring predictions (where differing mental models and priors can result in a multitude of potentially valid responses).

Scatter plots are a standard means of visualizing bivariate data, with a multitude of design parameters that affect their suitability for aggregate tasks [9]. Prior work has confirmed that viewers can make use of scatter plots to perform prediction tasks (which tacitly rely on trend estimation) in ways that are robust to both noise [16] and problem frame [22]. However, the heuristics used to perform these prediction tasks (such as the anchor-and-adjust method [5]) can introduce biases.

Similarly, visual design choices (for instance, the decision to encode class membership with color or with shape) can also impact performance at aggregate tasks [14, 23]. The aspect ratio of graphs can also bias judgments about trends [4]: narrow aspect ratios can result in overestimation in severity of trends, and wider aspect ratios in underestimation of severity. Another bias in prediction tasks is the "within-the-bar" bias [27]: for visually asymmetric visualizations such as bar charts, points contained within the visual area of the bar glyph are perceived as likelier than those outside the glyph.

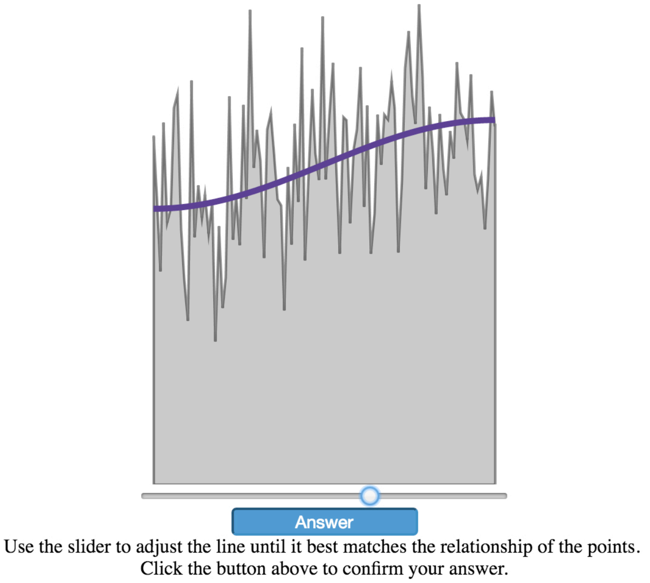
Recent work in the visualization community has focused on the perception of correlation in scatter plots. Rensink et al. [28] show that viewers can estimate correlation with some accuracy in scatter plots. Harrison et al. [15] extend this finding to other visualization types, and a re-analysis by Kay & Heer [20] indicates that, for many of the more esoteric bivariate visualizations, performance at this task is poor. Estimation of correlation can also be biased by the choice of axis scales [7]: the whitespace changes introduced by expanding the scale of the axes can cause over-estimation of correlation.

Our task of trend estimation combines elements of both prediction and correlation estimation tasks. As with prediction, there is not necessarily an unambiguous correct estimation of trend (different modeling and regression methods can produce different trend lines). As with correlation, viewers must make holistic judgments about the dataset in a way that (as per Harrison et al. [15]) likely relies on a set of visual proxies. That is, while viewers are unlikely to directly estimate correlation *per se*, they are estimating correlation through the perception of visual features such as the envelope of the points, point dispersion, or some other visual feature(s) highly correlated with the statistic of interest.

Prior work in visualization has not directly addressed the capabilities of regression by eye, instead assuming as a given that visual estimates of trend are sufficiently accurate. Work in the perception of summary statistics lends credences to this assumption, but we believe that trend estimation can be biased through conscious or unconscious design choices. For instance, within-the-bar bias may result in under-estimation of trends in bar and area charts, and factors that affect perception of correlation (such as aspect ratio and whitespace) may result in up- or down-weighting of outliers.

GENERAL EXPERIMENTAL METHODS

In order to assess the ability of visualization viewers to estimate trends, we conducted a series of three crowdsourced experiments on Amazon's Mechanical Turk. We designed these experiments to establish a performance baseline for regression by eye, and to examine potential sources of bias. Crowdsourced graphical perception experiments have been found to produce results that are largely in keeping with prior, lab-based work [17, 30]. In this section, we describe experimental design aspects shared across all experiments. Data



Use the slider to adjust the line until it best matches the relationship of the points.
Click the button above to confirm your answer.

Figure 2. An example estimation task from our experiments. Here, the participant must adjust the amplitude of the purple trigonometric function until it best matches a particular set of bivariate data.

tables, stimuli, and experimental instruments are available online at https://github.com/uwdata/trend-bias.

In these experiments, we report effect size using the interquartile mean (or midmean) of absolute error. The interquartile mean discards points in the first and fourth quartiles before averaging, trimming the tails of distributions. Cleveland & McGill [8] used the interquartile mean to provide a more robust measure of central tendency for responses from graphical perception studies, where participant error can create long tailed distributions of error. This is especially the case for crowdsourced studies, where correcting for data quality issues is perceived as more difficult than in laboratory settings [17]. However, the use of interquartile means can violate the assumptions involved in performing null hypothesis significance tests (as sample means can have non-normal distributions). Our statistical analyses were therefore performed using standard group means.

Experimental Interface

We investigated three types of common bivariate visualizations: scatter plots, line graphs, and area charts (see Fig. 3). In each experiment, we presented participants with a visualization and asked them to interactively adjust trend lines to best fit the data. Participants responded using a slider without tick marks, in order to limit anchoring effects [24]. Moving the slider adjusted a purple trend line by modifying its slope (in experiments 1 and 3) or its y-intercept (in experiment 2). Fig. 2 shows an example experimental task. After moving the slider, participants had to confirm their choice of trend line. The primary dependent measure is accuracy (e.g., the difference between the subject-specified slope and the slope of the series if the residuals were removed). We chose this design over a standard binary forced choice design for its greater expressiveness, as well as for the interactive feedback of adjusting the fit by hand. While other studies investigate the use of more expressive designs (such as free-form drawing) [21], we sought to disambiguate the model selection task (e.g., the selection of a linear or non-linear model) from the model fitting task (adjusting the parameters of the selected model). Our design also reduces the impact of factors such as noise from motor movements or heterogeneous input devices.

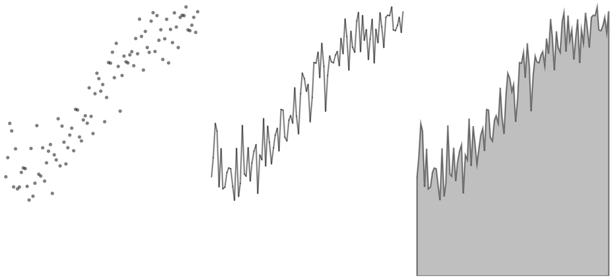


Figure 3. The three types of bivariate visualizations explored in this work: scatter plots, line graphs, and area charts. The density of points made bar charts and area charts visually similar, and so we excluded bar charts from our experiments. Likewise, the error in estimating trends from heatmaps [13] made them unsuitable for the experimental task.

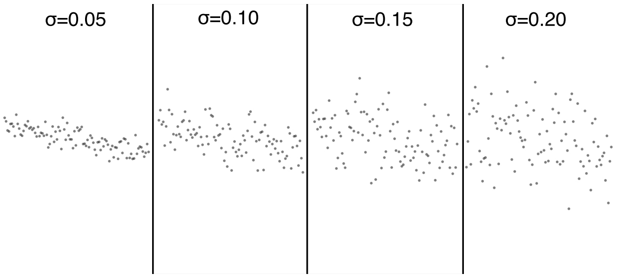


Figure 4. The four different Gaussian bandwidths used to generate residuals in this study. We produced stimuli by creating points on a target trend line (in this case, $f(x) = -0.2x + 0.6$), and then adding residuals drawn evenly from a Gaussian with a given bandwidth σ. Larger bandwidths mean more dispersion and so weaker fits.

Data Generation

For each stimulus we wished to have precise, independent control over relevant statistics such as noise, slope, and sample size, while maintaining the appearance of a "natural" distribution of points. Existing methods for generating points for related experimental tasks such as estimation of correlations (e.g., Harrison et al. [15]) do not afford independence: e.g., Pearson's r is correlated with the slope, as points with identical residuals but different slopes of their linear fits will have differing r values. We also wished to have a fair comparison between visual estimation and the results of ordinary least squares (OLS) regression. We therefore used the standard model of OLS to generate points, namely that $y = \beta x + \varepsilon$, where ε is a normally-distributed error term.

For each point set, we placed 100 points along a particular trend, regularly spaced in x. This both affords the use of line graphs and area charts for displaying this data (as they require that points be one to one), and aligns well with time series

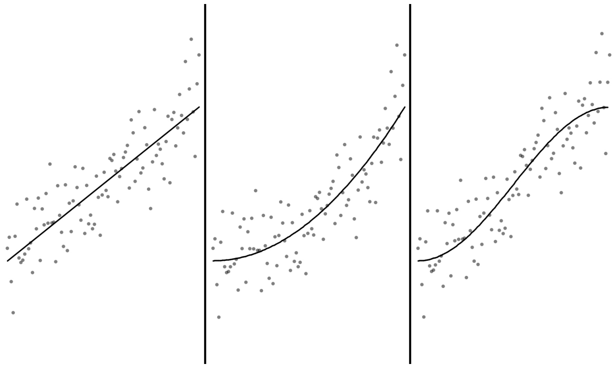


Figure 5. The three different trend types in this study: linear, quadratic, and trigonometric trends. In Experiment 1, participants estimated the slope of the linear fits, curvature of the quadratic fits, and amplitude of the trigonometric fits. In Experiment 2, they estimated the y-intercept of these fits. In Experiment 3, they estimated only the slope of linear fits. Trend lines are displayed here for reference.

data, a common domain for estimating (temporal) trends and relationships. For each point set, we created a set of residual values, sampled evenly from a Gaussian. The bandwidth of this Gaussian controls both goodness of fit and dispersion of points (see Fig. 4). We permuted this set of residuals, and then applied them to the original points. As heteroskedasticity introduced through permutation could alter the trend away from the target trend, we performed rejection sampling to ensure that the slope of the trend of the resulting points was within 10^{-7} of the target. We reused these residuals across all different trend types (linear, quadratic, or trigonometric). Fig. 5 shows these different trend types.

Except where noted, we selected trend lines that were centered in the image: that is, for a horizontal data extent of $[0,1]$, $f(0.5) = 0.5$. In all experiments, we desired control over the direction of the trend. For linear fits, this is the slope of the trend line. For quadratic fits, this was the curvature, as controlled by the coefficient of the second degree term. For trigonometric fits, this was the amplitude of a cosine function (with negative amplitude corresponding to negative slopes). These terms create similar relationships, such that a value of 0 is the line $f(x) = 0.5$, a value of 1 goes from $f(0) = 0$ to $f(1) = 1$, and -1 has the inverse relationship.

Participants

We limited the Mechanical Turk participant pool to subjects from within the United States, with a prior task approval rating of at least 90%. For validation, additional trials, without ambiguities such as outliers or design differences, contained the actual (OLS) trend line, which subjects then needed to simply match. We excluded responses from participants with high average error (greater than 0.2) on these validation stimuli. Across the three reported experiments and pilots, we performed 7 such exclusions. We recruited additional participants to replace these excluded subjects. Based on timings from internal piloting, we paid each participant \$2 for their participation, for a target rate of \$8/hour.

We analyzed data from 48 participants for each experiment (excluding rejections), for a total of 144 participants (98 male, 43 female, 3 who declined to state; $M_{age} = 33.2$, $SD_{age} = 8.8$). Across the subject pool, 10 reported having graduate or professional degrees, 69 college degrees, 40 at least some college, and 25 high school diplomas. After completing the main experimental task, participants were asked to self-assess their familiarity with charts and graphs on a 5 point Likert scale. The plurality (64) rated themselves as “3. Some familiarity,” and none rated themselves with the maximum rating of “5. A great deal of experience.”

EXPERIMENT 1: SLOPE ESTIMATION

We designed our first experiment to examine how accurate participants were at estimating the magnitude (slope, amplitude, or curvature) of trends in bivariate visualizations. We examined three types of bivariate visualizations: scatter plots, line graphs, and area charts (with the filled area below the line). In addition to linear trends, we examined more complex relationships such as quadratic and trigonometric functions.

We presented participants with a series of bivariate visualizations, who adjusted a slider to fit the perceived trend. The slider parameterized one of three types of trend: linear, quadratic, or trigonometric. For each stimulus, participants adjusted a slider that controlled the slope of a rendered trend line. In the case of quadratic trends, this slope was the curvature; for trigonometric fits, the positive/negative amplitude.

Participants saw one of each combination of 3 chart types (scatter plot, line graph, or area chart), 8 possible slopes $\beta = \pm\{0.1, 0.2, 0.4, 0.8\}$, and 4 bandwidths of Gaussian residuals $\sigma = \{0.05, 0.1, 0.15, 0.2\}$, for a total of 96 stimuli. We also included an additional 4 validation stimuli otherwise excluded from analysis. The type of trend (linear, quadratic, or trigonometric) was a random factor, with 32 stimuli of each factor level randomly assigned.

Hypotheses

We had three hypotheses for the first experiment:

1. **As the bandwidth of the residuals increased, accuracy would decrease**. Increasing the bandwidth of the residuals results in a lower correlation coefficient and higher perceived noise in the bivariate data. Prior work indicates that these related measures correspond to decreased accuracy for aggregate tasks in bivariate visualizations [1, 15].

2. **More complex relationships would result in lower accuracy**. Quadratic and trigonometric relationships are visually more complex than linear relationships, and often require more complex statistical methods to analyze. We anticipated that estimation of these less familiar relationships would therefore be more difficult than the linear case.

3. **Estimations would be unbiased**. That is, there would be no systematic over- or under-estimation of trends.

Results

We performed a three-way analysis of covariance (ANCOVA) of the effect of residual bandwidth, graph type and trend type on error in estimation of trend lines. We included participant ids and the actual slope of the trend line as covariates. We

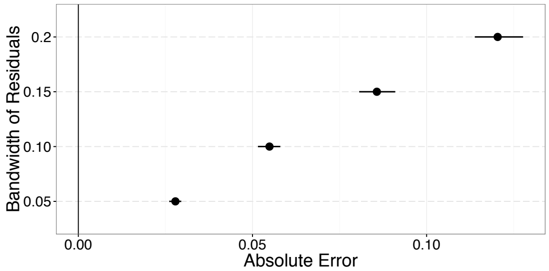


Figure 6. The effect of increased residual bandwidth on error in Experiment 1. See Fig. 4 for example stimuli at each factor level. Absolute error at estimating trend monotonically increases as the goodness of fit decreases. Confidence intervals represent bootstrapped 95% CIs of interquartile means.

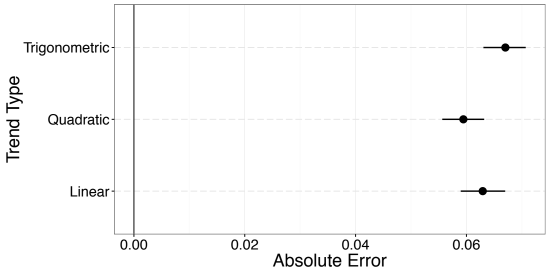


Figure 7. The effect of different trend types on error in Experiment 1. See Fig. 5 for example stimuli at each factor level. Despite the differing complexity of these types of fit, participant estimates were similar, indicating that regression by eye is capable of non-linear estimates. Confidence intervals represent bootstrapped 95% CIs of interquartile means.

defined error as the absolute difference between the slope of the OLS trend line, and the slope of the participant estimate.

Our results support our first hypothesis: **larger residuals result in less accuracy at regression by eye**. We observed a significant main effect for the bandwidth of the Gaussian used to generate residuals ($F(1,4554) = 950$, $p < 0.001$). The interquartile mean of absolute error increased monotonically with this bandwidth, from 0.02 of the actual slope of the trend line when the bandwidth was 0.05, to 0.12 when the bandwidth was 0.20. Figure 6 illustrates this result.

Our results fail to support our second hypothesis: **there was no statistically significant difference in estimation accuracy among linear, quadratic, or trigonometric trends**. Fit type was not a significant main effect ($F(2,4554) = 2.60$, $p = 0.074$), and post-hoc tests (using Tukey's Honest Significant Difference) did not identify any significant pairwise interactions. The interquartile mean of absolute error of was 0.06 for linear and quadratic fits, and 0.07 for trigonometric fits (see Fig. 7). This suggests that the relative unfamiliarity of non-linear trends does not result in poorer performance at regression by eye.

Our results support our third hypothesis: **there was no statistically significant bias in estimations**. Participants saw a balanced set of positive and negative trends. If estimates of these trends were unbiased, we would expect the average signed error to be close to zero. The average signed error was 0.0008, far less than the fidelity of the slider used to input guesses ($\Delta = 0.01$). A Student's T-test failed to support the hypothesis that $mu_{error} \neq 0$ ($t(4590) = 0.39$, $p = 0.70$).

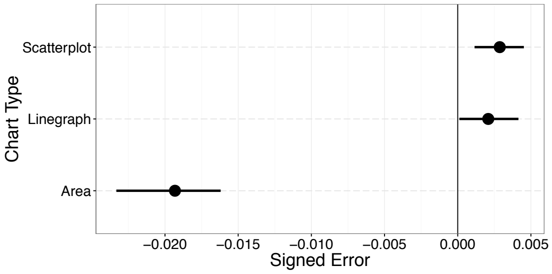


Figure 8. The effect of chart type on signed error in Experiment 2. See Fig. 4 for example stimuli at each factor level. Area charts are visually asymmetric, with the area below the line filled in with a color. This visual asymmetry results in a form of within-the-bar bias [27], where values in the filled region are perceived as likelier than values outside of it. This bias manifests as a consistent under-estimation in the intercept of trend lines. Other chart types we examined do not have this bias. Confidence intervals represent bootstrapped 95% CIs of interquartile means.

Prior work in graphical perception often measures performance as the absolute log error: e.g., Cleveland & McGill [8] calculate their performance metric as $\log_2(|\text{error}| + \frac{1}{8})$. As a point of comparison, the interquartile mean of the absolute log error across all conditions in this experiment was 2.4. Differences in methodology and measurement discourage statistical inferences comparing this value with those in similar experiments. Nevertheless, error rates for regression by eye are at least in the same magnitude as those observed in other graphical perception tasks in both lab and crowdsourced studies. For instance, Heer & Bostock [17] report an absolute log error of 2.5 for proportional judgments in treemaps, and Cleveland & McGill [8] report an absolute log error of 2.4 for proportional judgments in stacked bars (although compare to a log error of estimation of relative lengths of lines of 1.1). This similarity suggests that, despite requiring estimation of aggregate statistical information, regression by eye results in judgments that are accurate enough for many practical purposes, comparable with accuracy in comparing individual values in visualizations.

EXPERIMENT 2: “WITHIN-THE-BAR” BIAS

“Within-the-bar” bias is a known perceptual bias involving bar charts, in which points contained in the visual area of the glyph of the bar are deemed likelier than points outside of the glyph. Newman & Scholl [27] encountered this bias for a sampling task: “how likely is this point to have been drawn from the distribution represented by this bar?” Correll & Gleicher [11] likewise encountered this bias for inferential tasks: “how likely is the population mean to take a particular value, given the sample represented by this bar?”

We hypothesized that this bias would likewise occur in regression by eye when using visually asymmetric visualizations such as area charts: a “within-the-area” bias. The slope estimation task in the previous experiment would not capture this bias, as there is no method for participants to indicate a *uniform* under-estimation in trends; decreasing the slope would cause under-estimation at the beginning of the plot but not the end, and vice versa. We therefore designed this experiment to elicit estimates of the y-intercepts of trends. A within-the-area bias would then appear as systematic under-estimation of intercept in area charts.

As with the previous experiment, we presented participants with a series of bivariate visualizations. However, instead of estimating the *slope* of the points, participants estimated the *y-intercept* of the trend line. For each trial, we added a uniform offset to the points in the bivariate visualization in the data range $[-0.25, 0.25]$. The rendered trend line was initially placed with the correct slope, and such that $f(0.5) = 0.5$ in data space. Participants adjusted a slider controlling the vertical offset of this trend line.

The plots had the same factor levels as the previous experiment, for a total of 96 stimuli per participant, with an additional 4 validation stimuli otherwise excluded from analysis. The uniform offset was an additional random factor for each stimulus.

Hypotheses

We had one hypothesis for the second experiment:

1. **Area charts would be subject to within-the-area bias.** That is, participants would estimate lower values of y-intercepts of trends in area charts, as opposed to line graphs and scatter plots.

Results

We performed a one-way ANCOVA of the effect of graph type on *signed* error. We included residual bandwidth as a covariate, and participant ids as a random factor.

Our results support our first hypothesis: **participants systematically underestimated the intercept of trends in area charts, but not in scatter plots or line graphs**. We observed a significant main effect of graph type on signed error ($F(2,2431) = 27$, $p < 0.001$). A post-hoc Tukey's HSD confirmed significant differences in error between the area chart and the other two chart types, but not between scatter plots and line graphs. The interquartile mean of the signed error of estimations made with area charts was an under-estimation of -0.02, compared to an interquartile mean of 0.002 for the other two conditions. This under-estimation corresponds to unsigned errors more than twice as large in area charts (interquartile mean of 0.04 for intercepts that ranged from $\{-0.25, 0.25\}$, compared to 0.02 for line graphs and scatter plots). Figure 8 illustrates this result.

EXPERIMENT 3: ESTIMATION INVOLVING OUTLIERS

OLS regression operates under the assumption that there is a unimodal, symmetric distribution of residuals surrounding the line of best fit. Extreme outliers violate this assumption, and can result in trend lines that are substantially different from those produced by more robust methods. Visual inspection can accurately identify certain classes of outliers [1]. Therefore, regression by eye may afford the estimation of both outlier-robust and outlier-sensitive trends. However, cognitive biases [31] such as anchoring (the tendency to overweight the first set of information), availability (the tendency to overweight more recent or extreme information), and the hot-hand fallacy (the tendency to assume that runs of high or low values

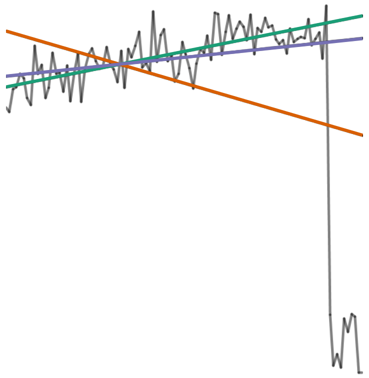


Figure 9. An example stimulus from Experiment 3. We replaced the final 10 points of this dataset with extreme values. The overlaid green trend line represents a robust fit (ignoring the outlier values), while the overlaid orange line represents the standard OLS fit with all points included. The purple line represents the average participant response on this stimulus. In general, participants' estimates of trend lines were closer to the robust than the non-robust trend; regression by eye tends to downweight outliers compared to OLS.

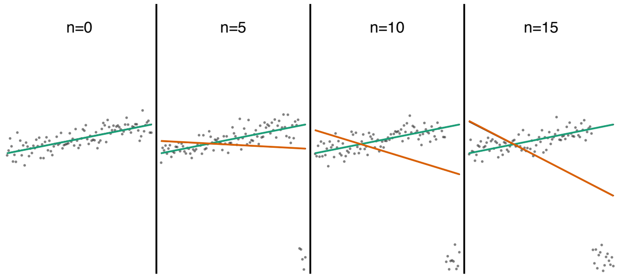


Figure 10. The four different outlier numerosities tested in Experiment 3. The overlaid green trend line represents a robust fit (ignoring the outlier values), while the overlaid orange line represents the standard OLS fit with all points included. More outliers results in larger divergence between these two types of fit.

will continue) can impact how viewers categorize and utilize outliers in prediction tasks [6,19]. In other words, regression by eye may not uniformly weight outliers, depending on their position in the plot.

This experiment was largely identical to Experiment 1, except we designated 0, 5, 10, or 15 points at the very beginning, first third, or end of the series as outliers. We vertically positioned outliers within the top or bottom 10% of the visual area of the visualization (whichever was farthest from the trend line), and added random jitter. We then calculated the outlier-sensitive line of best fit, as well as the intersection between this new trend line and the original, outlier-less trend line. We refer to the original trend line, estimated without the presence of outliers, as the **robust trend line**. We refer to the re-estimated trend, which takes into account the added outliers, as the **OLS trend line**. Figure 10 shows examples of this process, and the corresponding changes in trend line.

As in Experiment 1, the participants controlled the slope of a rendered trend line with a slider. However, rather than being placed such that $f(0.5) = 0.5$, we offset the trend line to the intersection of the robust (without outliers) and non-robust (including outliers) trend line. This allowed participants to express both types of fit with the same slider interaction, while affording estimations beyond a simple interpolation of both trends. We limited the stimuli to linear fits only, excluding quadratic and trigonometric trends from this experiment.

Participants saw one of each combination of factors: the three graph types, eight slopes, and four Gaussian residual bandwidths. The four outlier quantities $\{0, 5, 10, 15\}$ were an additional factor. To maintain a manageable number of stimuli (piloting showed evidence of fatigue for more than 100 stimuli per trial), both the sign of the trend line (positive or negative), and the location of the outliers (beginning, first third, or end) were random factors, with each level randomly apportioned to half and one third of the stimuli respectively. This resulted in 96 total stimuli, with an additional 4 validation stimuli otherwise excluded from analysis, in line with prior experiments.

Hypotheses

We had three hypotheses for the third experiment:

1. **Participant estimation would hew closer to a robust trend line ignoring outliers, rather than the OLS trend line.** We assumed that in general, participants would ignore or downweight outliers when performing regression by eye.

2. **As the number of outliers increased, estimations would be closer to the non-robust OLS fit.** We speculated on the existence of a "tipping point" of outlier density, beyond which participants would avoid robust fits, and interpolate between the robust and OLS fits.

3. **Outliers at the end of the chart would result in estimations closer to the OLS fit than outliers in other locations on the plot.** Prior work on cognitive biases in predictions suggests that viewers may give more credence to more recent outliers (as indicative as new anchor points, or of an emerging "streak") [12]. We therefore believed that these points would be more heavily weighted, and the resulting estimation closer to the outlier-sensitive OLS line, than outliers "earlier" in the plot.

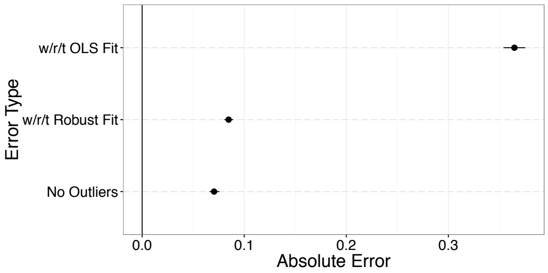


Figure 11. The absolute error of trend estimates in Experiment 3, measured as difference from the OLS slope (which includes outliers, row 1) or from the robust slope (which excludes outliers, row 2). Participants were significantly closer to the robust fit, indicating that they were largely insensitive to outliers. We also include the absolute error for trials with no outliers for reference. Error to the robust line is higher than this standard, indicating that participants were at least performing some interpolation between OLS and robust fits, even if they largely favored the robust fit. Confidence intervals represent bootstrapped 95% CIs of interquartile means.

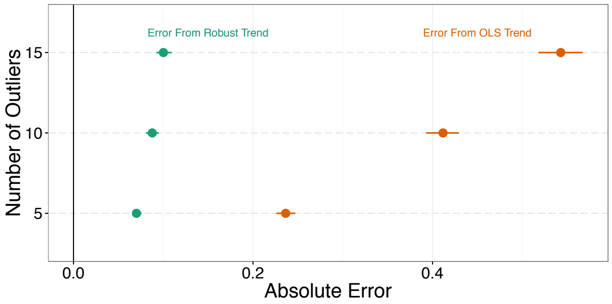


Figure 12. The effect of numerosity of outliers on absolute error in Experiment 3. See Fig. 10 for example stimuli at each factor level. As the number of outliers increases, there is increasing divergence between the robust line of best fit (which ignores outliers), and the standard OLS line of best fit (which is sensitive to outliers). Estimates become increasingly dissimilar to the OLS fit, indicating that participants down-weight outliers when making estimates of trend. However, the increasing error even from the robust trend line indicates that participants still perform some interpolation between the two types of fits. Confidence intervals represent bootstrapped 95% CIs of interquartile means.

Results

Our results support our first hypothesis. **Participant estimates were closer to the robust trend line than the trend line that included outliers**. Excluding conditions with zero outliers (and so the two trend lines would be identical), the interquartile mean of absolute error was over 4 times higher when calculated as a comparison to the OLS trend line (0.36) as opposed to when calculated from the robust trend line (0.08). A Student's T-test confirmed that the error as defined by the OLS line was significantly higher than the robust line ($t(6670) = 43$, $p < 0.001$). However, a Student's T-test found a significant difference in unsigned error when compared to the robust trend line between estimates where no outliers were present, and those with any amount of outliers ($\mu = 0.07$ vs. $\mu = 0.08$, $t(3808) = 9.8$, $p < 0.001$). Figure 11 illustrates this result. This indicates that participants are not entirely ignoring outliers, although the small effect size suggests that participants are giving outliers significantly less weight than OLS regression. Figure 9 shows an comparison of robust, participant, and OLS trend lines.

Our results fail to support our second hypothesis. **While participant estimates diverge from the robust line as the number of outliers increases, the OLS trend line diverges faster than participant estimates**. There was a significant positive interaction between outlier count and absolute error with respect to the robust trend line ($R^2 = 0.02$, $F(1,4591) = 110$, $p < 0.001$). However, there was also a significant positive interaction between outlier count and absolute error with respect to the OLS line ($R^2 = 0.38$, $F(1,4591) = 2800$, $p < 0.001$). Additionally, while the range of errors with respect to the robust line remained small (interquartile means of 0.07 with 0 outliers, monotonically increasing to 0.10 with 15 outliers), absolute error with respect to the OLS line monotonically increased almost 7-fold, from 0.07 with 0 outliers to 0.54 with 15 outliers. These results indicate that, while increased numbers of extreme values may result in participants increasing the weight of these outliers, this increase in weight is slower than the sensitivity of OLS to outliers. Figure 12 illustrates this result.

Our results also fail to support our third hypothesis: **There was no significant impact of outlier location on performance**. There is a confound between outlier location and OLS line: outliers on the ends of the data ranges create larger shifts in slope than outliers in the center, which may only create shifts in intercept. To evaluate our third hypothesis, we therefore defined a parameterized performance function $\frac{\text{Estimated Slope}-\text{Robust Slope}}{\text{OLS Slope}-\text{Robust Slope}}$ that contextualizes a participant estimate as an interpolation between the robust slope (value of 0) and the OLS slope (value of 1). A one way ANCOVA of this interpolation value as an effect of outlier location, with outlier count as a covariate, and participant id as a random factor, found no significant effect of outlier location on interpolative value ($F(2,3441) = 0.63$, $p = 0.53$).

DISCUSSION

In many cases, designers do not explicitly encode regression information. In other cases, viewers may not have the statistical or graphical expertise to interpret such information, even when it is supplied. Yet, our results point to regression by eye as a robust and reasonably accurate method for estimating trends in bivariate data. Participants from a variety of backgrounds and levels of self-reported graphical and statistical expertise were capable of estimating both the slope and intercept of both linear and non-linear trends. That is, viewers of visualization are largely trustworthy when estimating the relationship between two variables in a plot.

However, this general accuracy of trend estimation is not universal. Area charts are visually asymmetrical: the area below the line is filled in, and the area above it is not. We found that this asymmetry creates a “within-the-area” bias: a systematic under-estimation of the intercept of trends, due to the perceived higher likelihood of points in the filled-in area. Designers hoping to rely on regression by eye should avoid such asymmetries in their bivariate visualizations.

Likewise, while viewers do not give the same weight to outliers as OLS regression, they do not ignore them either. For noisy data, this robustness may be beneficial: people can be relied on to perform filtering operations without explicit guidance. In other cases, this insensitivity may be undesirable. For instance, more complicated automatic analytical processes may use OLS regression or other outlier sensitive techniques as part of their calculations, resulting in a disconnect between human and statistical conclusions at later stages of the sensemaking process. Similarly, the human tendency to discount or downweight outliers may result in slow adaptation to new data that does not fit into existing patterns.

Implications for Design

The estimation of trends, both for the prediction of values, and the imputation of missing values, is a common and important task in data analysis. We believe that our results suggest two main recommendations for designers of information visualizations that include time series or bivariate data:

1. *Designers do not need to annotate bivariate visualizations with trend lines in many cases.* Regression by eye, without explicit trend lines, is comparable in accuracy to other types of graphical perception tasks (such as comparison of value), even for a wide class of models. The presence of outliers is an exception to this general advice: human estimation of trends does not give much weight to outliers. For data where outliers are important, and expected to contribute to important trends, designers ought to visualize the outlier sensitive models directly.

2. *Designers should avoid area charts when the perception of trends is important.* The visual asymmetry caused by filling in the area under the line results in a corresponding asymmetry in judgment of trend: an underestimation in perceived trend. This within-the-area bias is undesirable in many cases, but can be countered by visual encodings (such as scatter plots and line charts) that have no such visual asymmetry.

Limitations & Future Work

Our experimental setup was intentionally simple, affording only a single free parameter for each experiment. In actual regression by eye, the viewer may simultaneously engage in multiple types of estimation: choosing a particular type of fit, ignoring outliers, and estimating the parameters of the chosen model. Errors in any one of these steps could compound, resulting in performance worse than our measures, where many of these decisions are fixed *a priori*. It is also likely that there are many visual features, such as the convex hull of points, or the numerosity of point clouds, that are acting as visual proxies for estimation of the trend. Given the unimodal error function and uniform density of our experimental stimuli, these proxies are useful for estimating the trend line. However, in datasets where internal densities of points may be skewed, or the envelope of point clouds non-informative for estimating central tendency, these visual proxies may introduce biases and confounds in visual estimations.

There are also a number of design decisions not considered in this study that could impact regression by eye. For example, Wood et al. [32] have shown that "sketchiness" in visualizations can result in skepticism of the data and design. It is possible that outliers could receive less visual (and so statistical) weight as a response to this sort of skepticism. Our stimuli likewise contained a sufficient number of points that bar charts were not a feasible choice of visualization. Given the propensity of bar charts to encourage comparison of individual, rather than aggregate quantities [33], it is possible that bar charts of smaller scale bivariate data could promote fitting of local rather than global trends.

Finally, we focused on a simple form of regression, ordinary least squares (OLS), as our standard for measuring accuracy. While our data were constructed to satisfy the assumptions of OLS (with the intentional exception of our outlier experiment), in most real world scenarios OLS is just one tool of many, and analysts must exercise judgment when determining how to fit their data. More complex models may not have ready visual analogues, and data concerning regression by eye may not extend to cover these cases.

Our future work is focused in three areas. First, we wish to examine the impact of annotations relevant to regression (such as confidence bands, error bars, and curve boxplots [26]) on regression by eye. Can sufficient information promote caution in judgments of trend? Second, we wish to examine the impact of different rhetorical framings and presentations on regression by eye. Language from semiotics and rhetoric can provide testable structures for how visualizations are consumed [18]: these framings could similarly impact statistical judgments. For instance, outliers from a credible source could be weighted higher than outliers with more suspect provenance. Likewise, preconceptions about the volatility of data from certain domains may color how experts estimate trends. Lastly, we wish to examine techniques for overcoming bias in regression by eye and similar visual estimations of statistical quantities. Cognitive and perceptual biases are difficult to overcome, and may require exploration of new visualization designs [25].

Conclusion

In this paper, we examine the ability of visualization viewers to perform *regression by eye*: the visual estimation of trends in bivariate visualizations. We show that viewers without statistical training can reliably estimate both linear and non-linear trends in charts such as scatter plots and line graphs. However, area graphs are subject to a "within-the-area" bias, leading to estimated trends with lower intercepts than other bivariate visualizations. Regression by eye is also less sensitive to outliers than standard least-squares regression. This decreased sensitivity results in a divergence between trends estimated by viewers, which do not give much weight to extreme outliers, and those trends calculated by regression methods, that can be heavily influenced by a few extreme values.

ACKNOWLEDGMENTS

This work was supported by a Moore Foundation Data-Driven Discovery Investigator Award.

REFERENCES

1. Danielle Albers, Michael Correll, and Michael Gleicher. 2014. Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 551–560.
2. Francis J Anscombe. 1973. Graphs in statistical analysis. *The American Statistician* 27, 1 (1973), 17–21.
3. Dan Ariely. 2001. Seeing sets: Representation by statistical properties. *Psychological science* 12, 2 (2001), 157–162.
4. Vivien Beattie and Michael John Jones. 2002. The impact of graph slope on rate of change judgments in corporate reports. *Abacus* 38, 2 (2002), 177–199.
5. Fergus Bolger and Nigel Harvey. 1993. Context-sensitive heuristics in statistical reasoning. *The Quarterly Journal of Experimental Psychology* 46, 4 (1993), 779–811.
6. Sean D Campbell and Steven A Sharpe. 2009. Anchoring bias in consensus forecasts and its effect on market prices. *Journal of Financial and Quantitative Analysis* 44, 02 (2009), 369–390.
7. William S Cleveland, Persi Diaconis, and Robert McGill. 1982. *Variables on scatterplots look more highly correlated when the scales are increased.* Technical Report. DTIC Document.
8. William S Cleveland and Robert McGill. 1984a. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association* 79, 387 (1984), 531–554.
9. William S Cleveland and Robert McGill. 1984b. The many faces of a scatterplot. *J. Amer. Statist. Assoc.* 79, 388 (1984), 807–822.
10. Michael Correll, Danielle Albers, Steven Franconeri, and Michael Gleicher. 2012. Comparing averages in time series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1095–1104.
11. Michael Correll and Michael Gleicher. 2014. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2142–2151.
12. Rachel Croson and James Sundali. 2005. The gamblerâĂŹs fallacy and the hot hand: Empirical data from casinos. *Journal of risk and uncertainty* 30, 3 (2005), 195–209.
13. Johannes Fuchs, Fabian Fischer, Florian Mansmann, Enrico Bertini, and Petra Isenberg. 2013. Evaluation of alternative glyph designs for time series data in a small multiple setting. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3237–3246.
14. Michael Gleicher, Michael Correll, Christine Nothelfer, and Steven Franconeri. 2013. Perception of average value in multiclass scatterplots. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2316–2325.

15. Lane Harrison, Fumeng Yang, Steven Franconeri, and Remco Chang. 2014. Ranking visualizations of correlation using weber's law. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 1943–1952.

16. Nigel Harvey Teresa Ewart Robert West. 1997. Effects of data noise on statistical judgement. *Thinking & Reasoning* 3, 2 (1997), 111–132.

17. Jeffrey Heer 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*. ACM, 203–212.

18. Jessica Hullman and Nick Diakopoulos. 2011. Visualization rhetoric: Framing effects in narrative visualization. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2231–2240.

19. Li-Jun Ji, Richard E Nisbett, and Yanjie Su. 2001. Culture, change, and prediction. *Psychological Science* 12, 6 (2001), 450–456.

20. Matthew Kay and Jeffrey Heer. 2016. Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation. *IEEE transactions on visualization and computer graphics* 22, 1 (2016), 469–478.

21. Yea-Seul Kim, Jessica Hullman, and Katharina Reinecke. 2017. Explaining the Gap: Visualizing OneâĂŹs Predictions Improves Recall and Comprehension of Data. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM. To appear.

22. Stephan Lewandowsky. 2011. Popular consensus climate change is set to continue. *Psychological Science* (2011).

23. Stephan Lewandowsky and Ian Spence. 1989. Discriminating strata in scatterplots. *J. Amer. Statist. Assoc.* 84, 407 (1989), 682–688.

24. Justin Matejka, Michael Glueck, Tovi Grossman, and George Fitzmaurice. 2016. The Effect of Visual Appearance on the Performance of Continuous Sliders and Visual Analogue Scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5421–5432.

25. Luana Micallef, Pierre Dragicevic, and Jean-Daniel Fekete. 2012. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2536–2545.

26. Mahsa Mirzargar, Ross T Whitaker, and Robert M Kirby. 2014. Curve boxplot: Generalization of boxplot for ensembles of curves. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2654–2663.

27. George E Newman and Brian J Scholl. 2012. Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic bulletin & review* 19, 4 (2012), 601–607.

28. Ronald A Rensink and Gideon Baldridge. 2010. The perception of correlation in scatterplots. In *Computer Graphics Forum*, Vol. 29. Wiley Online Library, 1203–1210.

29. Danielle Albers Szafir, Steve Haroz, Michael Gleicher, and Steven Franconeri. 2016. Four types of ensemble coding in data visualizations. *Journal of vision* 16, 5 (2016), 11–11.

30. Justin Talbot, Vidya Setlur, and Anushka Anand. 2014. Four experiments on the perception of bar charts. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2152–2160.

31. Amos Tversky and Daniel Kahneman. 1975. Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making*. Springer, 141–162.

32. Jo Wood, Petra Isenberg, Tobias Isenberg, Jason Dykes, Nadia Boukhelifa, and Aidan Slingsby. 2012. Sketchy rendering for information visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2749–2758.

33. Jeff Zacks and Barbara Tversky. 1999. Bars and lines: A study of graphic communication. *Memory & Cognition* 27, 6 (1999), 1073–1079.