Eye Fitting Straight Lines in the Modern Era

Emily A. Robinson 1
Department of Statistics, University of Nebraska - Lincoln
and
Susan VanderPlas 2
Department of Statistics, University of Nebraska - Lincoln
and
Reka Howard 3
Department of Statistics, University of Nebraska - Lincoln

September 6, 2021

Abstract

Fitting lines by eye through a set of points has been explored since the 20th century. Common methods of fitting trends by eye involve maneuvering a string, black thread, or ruler until the fit is suitable, then drawing the line through the set of points. In 2015, the New York Times introduced an interactive feature, called 'You Draw It'. Readers are asked to input their own assumptions about various metrics and compare how these assumptions relate to reality. The New York Times team utilizes Data Driven Documents (D3) that allows readers to predict these metrics by drawing a line on their computer screen with their computer mouse. In my research, I established 'You Draw It' as a method for graphical testing by adapting the New York Times feature. I recruited participants via crowdsourcing websites and replicated the study found in Eye Fitting Straight Lines (Mosteller et al., 1981). Participants were directed to an RShiny application link and shown points following a linear trend and asked to draw a line through the data points using their computer mouse; task plots were generated using the r2d3 package in R statistical software. Results from my study were consistent with those found in the previous study; when shown points following a linear trend, participants tended to fit the slope of the first principal component over the slope of the least-squares regression line. This trend was most prominent when shown data simulated with larger variances. The reproducibility of these results serves as evidence of the reliability of the you draw it method. Future work is necessary to implement the 'You Draw It' tool as a method of testing graphics. [200 word limit]

Keywords: Graphics, Regression, Graph Perception, Scatterplot, Cognitive Bias

1 Introduction

Advanced technology and computing power have promoted data visualization as a central tool in modern data science. Unwin (2020) defines data visualization as the art of drawing graphical charts in order to display data. Graphics are useful for data cleaning, exploring data structure, and have been an essential component in communicating information for the last 200 years (Lewandowsky & Spence 1989). Although statistical graphics have become widely used and valued in science, business, and in many other aspects of life, as creators of graphics, we are too accepting of them as default without asking critical questions about the graphics we create or view (Unwin, 2020).

1.1 Graph Perception

1.2 Testing Statistical Graphics

One way in which we evaluate the effectiveness of charts is through the use of graphical tests. We could ask participants to identify differences in graphs, read information off of a chart accurately, use data to make correct real-world decisions, or predict the next few observations. All of these types of tests require different levels of use and manipulation of the information being presented in the chart.

- Psychophysics (speed and accuracy)
- Lineups

1.3 Trend Judgement

- Finney (1951)
- Mosteller et al. (1981)
- Ciccione & Dehaene (2021)

Initial studies in the 20th century explored the use of fitting lines by eye through a set of points (Finney 1951, Mosteller et al. 1981). Common methods of fitting trends by eye involve maneuvering a string, black thread, or ruler until the fit is suitable, then drawing the line through the set of points. In Finney (1951), it was of interest to determine the effect

of stopping iterative maximum likelihood calculations after one iteration. Many techniques in statistical analysis are performed with the aid of iterative calculations such as Newton's method or Fisher's scoring. Guesses are made at the best estimates of certain parameters and these guesses are then used as the basis of a computation which yields a new set of approximation to the parameter estimates; this same procedure is the performed on the new parameter estimates and the computing cycle is repeated until convergence, as determined by the statistician, is reached. The author was interested in whether one iteration of calculations was sufficient in the estimation of parameters connected with dose-response relationships. One measure of interest is the relative potency between a test preparation of doses and standard preparation of does; relative potency is calculated as the ratio of two equally effective doses between the two preparation methods. **??** shows a pair of parallel probit responses in a biological assay. The x-axis is the $\log_{1.5}$ dose level for four dose levels (for example, doses 4, 6, 9, and 13 correspond correspond to equally spaced values on a logarithmic scale, labeled 0, 1, 2, and 3) and the y-axis is the corresponding probit response as calculated in Finney & Stevens (1948); circles correspond to the test preparation method while the crosses correspond to the standard preparation method. For these sort of assays, the does-response relationship follows a linear regression of the probit response on the logarithm of the dose levels; the two preparation methods can be constrained to be parallel (**?**), limiting the relative potency to one consistent value. In this study, twenty-one scientists were recruited via postal mail and asked to "rule two lines" in order to judge by eye the positions for a pair of parallel probit regression lines in a biological assay. The author then computed one iterative calculation of the relative potency based on starting values as indicated by the pair of lines provided by each participant and compared these relative potency estimates to that which was estimated by the full probit technique (reaching convergence through multiple iterations). Results indicated that one cycle of iterations for calculating the relative potency was sufficient based on the starting values provided by eye from the participants.

Thirty years later, Mosteller et al. (1981), sought to understand the properties of least squares and other computed lines by establishing one systematic method of fitting lines by eye. The authors recruited 153 graduate students and post doctoral researchers in

Introductory Biostatistics. Participants were asked to fit lines by eye to four scatterplots using an 8.5 x 11 inch transparency with a straight line etched completely across the middle. A latin square design (Giesbrecht & Gumpertz 2004) with packets of the set of points stapled together in four different sequences was used to determine if there is an effect of order of presentation. It was found that order of presentation had no effect and that participants tended to fit the slope of the first principal component (error minimized orthogonally, both horizontal and vertical, to the regression line) over the slope of the least squares regression line (error minimized vertically to the regression line).

In 2015, the New York Times introduced an interactive feature, called You Draw It (Aisch et al. 2015, Buchanan et al. 2017, Katz 2017). Readers are asked to input their own assumptions about various metrics and compare how these assumptions relate to reality. The New York Times team utilizes Data Driven Documents (D3) that allows readers to predict these metrics through the use of drawing a line on their computer screen with their computer mouse. (Katz 2017) is one such example in which readers are asked to draw the line for the missing years providing what they estimate to be the number of Americans who have died every year from car accidents, since 1990. After the reader has completed drawing the line, the actual observed values are revealed and the reader may check their estimated knowledge against the actual reported data.

Major news and research organizations such as the New York Times, FiveThirtyEight, Washington Post, and the Pew Research Center create and customize graphics with Data Driven Documents (D3). In June 2020, the New York Times released a front page displaying figures that represent each of the 100,000 lives lost from the COVID-19 pandemic until this point in time (Barry et al. 2020); this visualization was meant to bring about a visceral reaction and resonate with readers. During 2021 March Madness, FiveThirtyEight created a roster-shuffling machine which allowed readers to build their own NBA contender through interactivity (Ryanabest 2021). Data Driven Documents (D3) is an open-source JavaScript based graphing framework created by Mike Bostock during his time working on graphics at the New York Times. For readers familiar with R, it is notable to consider D3 in JavaScript equivalent to the ggplot2 package in R (Wickham 2016).

1.4 Research objective

The goal of this paper is to establish you draw it as a tool for measuring predictions of trends fitted by eye and a method for testing graphics. In order to validate you draw it as a method for testing graphics, the first sub-study, referred to as Eye Fitting Straight Lines in the Modern Era, replicated the experiment and results found in Mosteller et al. (1981).

2 Study Development

2.1 Interactive Plot Development

2.2 Data Generation

All data processing was conducted in R before being passed to the D3.js source code. A total of $N = 30$ points $(x_i, y_i), i = 1, ...N$ were generated for $x_i \in [x_{min}, x_{max}]$ where x and y have a linear relationship. Data were simulated based on linear model with additive errors:

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{1}$$
$$\text{with } e_i \sim N(0, \sigma^2).$$

The parameters β_0 and β_1 are selected to replicate Mosteller et al. (1981) with e_i generated by rejection sampling in order to guarantee the points shown align with that of the fitted line. An ordinary least squares regression is then fit to the simulated points in order to obtain the best fit line and fitted values in 0.25 increments across the domain, $(x_k, \hat{y}_{k,OLS}), k = 1, ..., 4x_{max} + 1$. The data simulation function then outputs a list of point data and line data both indicating the parameter identification, x-value, and corresponding simulated or fitted y value. The data simulation procedure is described in Algorithm 1.

Simulated model equation parameters were selected to reflect the four data sets (F, N, S, and V) used in Mosteller et al. (1981) (Table 1). Parameter choices F, N, and S simulated data across a domain of 0 to 20. Parameter choice F produces a trend with a positive slope and a large variance while N has a negative slope and a large variance. In comparison, S shows a trend with a positive slope with a small variance and V yields a

Algorithm 1 Eye Fitting Straight Lines in the Modern Era Data Simulation

• **Input Parameters:** $y_{\bar{x}}$ for calculating the y-intercept, β_0; slope β_1; standard deviation from line σ; sample size of points $N = 30$; domain x_{min} and x_{max}; fitted value increment $x_{by} = 0.25$.

• **Output Parameters:** List of point data and line data each indicating the parameter identification, x value, and corresponding simulated or fitted y value.

1: Randomly select and jitter $N = 30$ x-values along the domain, $x_{i=1:N} \in [x_{min}, x_{max}]$.

2: Determine the y-intercept, β_0, at x = 0 from the provided slope (β_1) and y-value at the mean of x ($y_{\bar{x}}$) using point-slope equation of a line.

3: Generate "good" errors, $e_{i=1:N}$ based on $N(0, \sigma)$ by setting a constraint requiring the mean of the first $\frac{1}{3}$N errors $< |2\sigma|$.

4: Simulate point data based on $y_i = \beta_0 + \beta_1 x_i + e_i$

5: Obtain ordinary least squares regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$, for the simulated point data using the lm function in the stats package in base R.

6: Obtain fitted values every 0.25 increment across the domain from the ordinary least squares regression $\hat{y}_{k,OLS} = \hat{\beta}_{0,OLS} + \hat{\beta}_{1,OLS} x_k$.

7: Output data list of point data and line data each indicating the parameter identification, x value, and corresponding simulated or fitted y value.

Table 1: Eye Fitting Straight Lines in the Modern Era simulation model parameters

Parameter Choice	$y_{\bar{x}}$	β_1	σ
S	3.88	0.66	1.30
F	3.90	0.66	1.98
V	3.89	1.98	1.50
N	4.11	-0.70	2.50

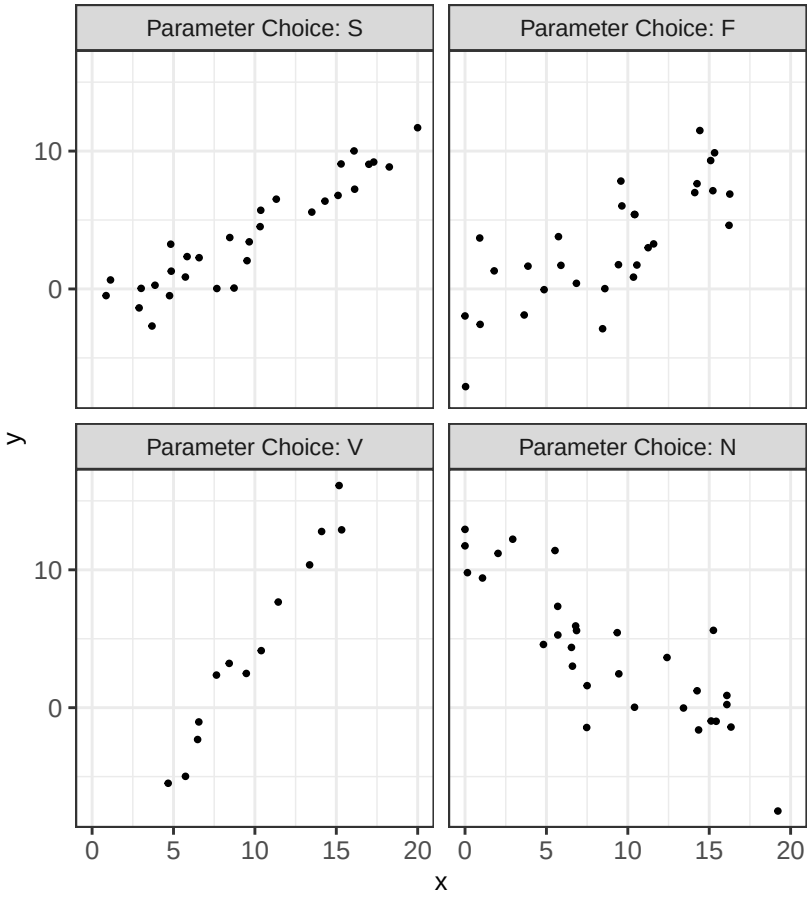

Figure 1: Eye Fitting Straight Lines in the Modern Era Simulated Data Example

steep positive slope with a small variance over the domain of 4 to 16. Fig. 1 illustrates an example of simulated data for all four parameter choices intended to reflect the trends seen in **??**. Aesthetic design choices were made consistent across each of the interactive you draw it plots; the aspect ratio, defining the x to y axis ratio was set to one and the y-axis range extended 10% beyond (above and below) the range of the simulated data points to allow for users to draw outside the simulated data set range.

2.3 Study Design

Participants completed the experiment using a RShiny application found here. During May 2021, participants were recruited through Twitter, Reddit, and direct email. A total of 39 individuals completed 256 unique you draw it task plots; all completed you draw it task

plots were included in the analysis.

3 Results

In addition to the participant drawn points, $(x_k, y_{k,drawn})$, and the ordinary least squares (OLS) regression fitted values, $(x_k, \hat{y}_{k,OLS})$, a regression equation with a slope based on the first principal component (PCA) was used to calculate fitted values, $(x_k, \hat{y}_{k,PCA})$. For each set of simulated data and parameter choice, the PCA regression equation was determined by using the princomp function in the stats package in base R to obtain the rotation of the coordinate axes from the first principal component (direction which captures the most variance). The estimated slope, $\hat{\beta}_{1,PCA}$, is determined by the ratio of the axis rotation in y and axis rotation in x of the first principal component with the y-intercept, $\hat{\beta}_{0,PCA}$ calculated by the point-slope equation of a line using the mean of of the simulated points, $(\bar{x}_i, \bar{y}_i)$. Fitted values, $\hat{y}_{k,PCA}$ are then obtained every 0.25 increment across the domain from the PCA regression equation, $\hat{y}_{k,PCA} = \hat{\beta}_{0,PCA} + \hat{\beta}_{1,PCA} x_k$. Fig. 2 illustrates the difference between an OLS regression equation which minimizes the vertical distance of points from the line and a regression equation with a slope calculated by the first principal component which minimizes the smallest distance of points from the line.

For each participant, the final data set used for analysis contains $x_{ijk}, y_{ijk,drawn}, \hat{y}_{ijk,OLS}$, and $\hat{y}_{ijk,PCA}$ for parameter choice $i = 1, 2, 3, 4$, j $= 1, ...N_{participant}$, and x_{ijk} value $k = 1, ..., 4x_{max} + 1$. Using both a linear mixed model and a generalized additive mixed model, comparisons of vertical residuals in relation to the OLS fitted values ($e_{ijk,OLS} = y_{ijk,drawn} - \hat{y}_{ijk,OLS}$) and PCA fitted values ($e_{ijk,PCA} = y_{ijk,drawn} - \hat{y}_{ijk,PCA}$) were made across the domain. Fig. 3 displays an example of all three fitted trend lines for parameter choice F.

Using the lmer function in the lme4 package (Bates et al. 2015), a linear mixed model (LMM) is fit separately to the OLS and PCA residuals, constraining the fit to a linear trend. Parameter choice, x, and the interaction between x and parameter choice were treated as fixed effects with a random participant effect accounting for variation due to participant. The LMM equation for each fit (OLS and PCA) residuals is given by:

$$y_{ijk,drawn} - \hat{y}_{ijk,fit} = e_{ijk,fit} = [\gamma_0 + \alpha_i] + [\gamma_1 x_{ijk} + \gamma_{2i} x_{ijk}] + p_j + \epsilon_{ijk} \tag{2}$$

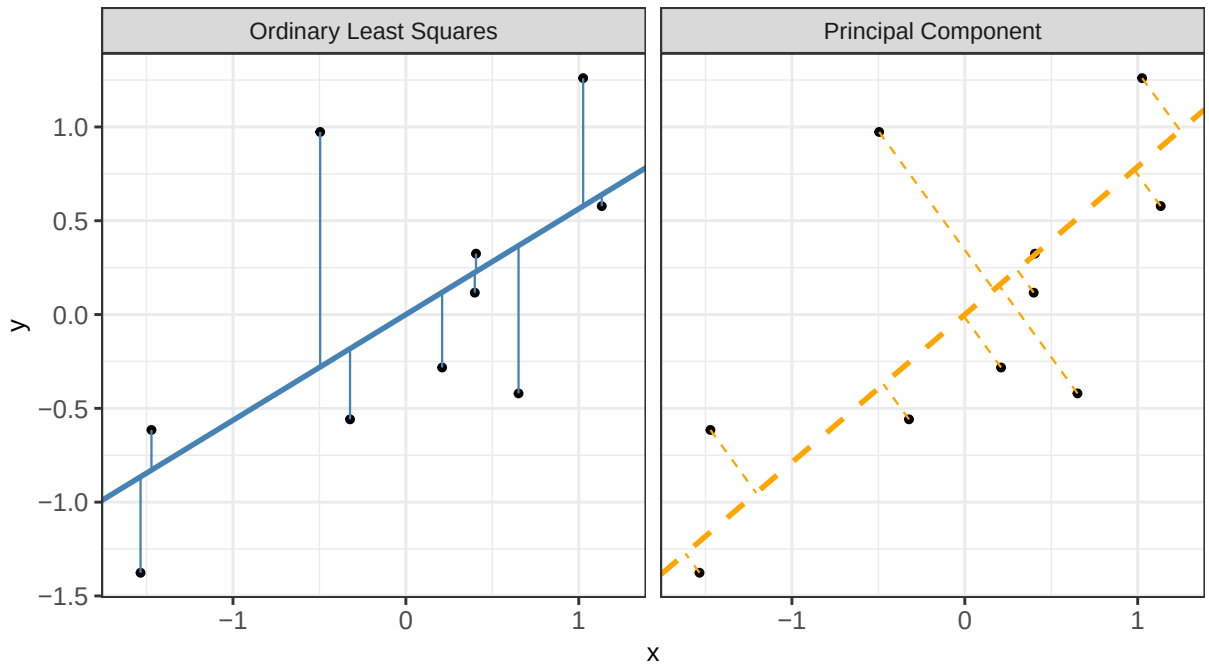

Figure 2: OLS vs PCA Regression Lines

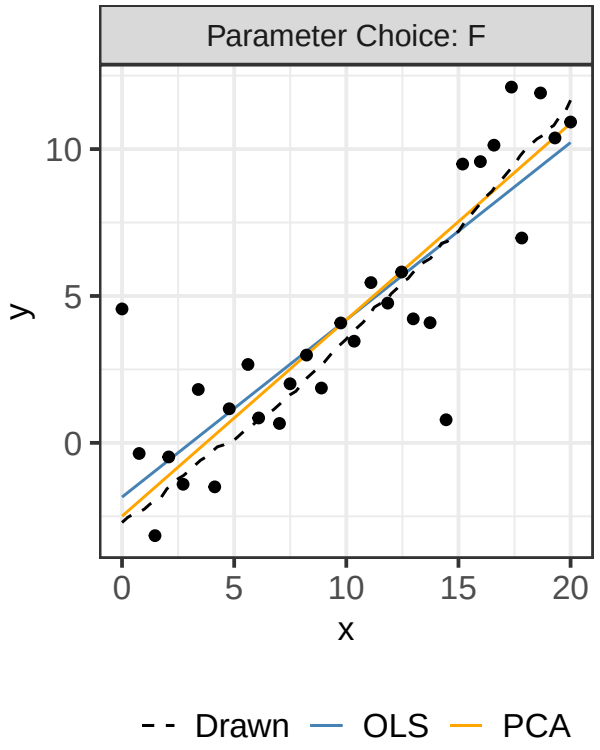

Figure 3: Eye Fitting Straight Lines in the Modern Era Example

where

- $y_{ijk,drawn}$ is the drawn y-value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x-value
- $\hat{y}_{ijk,fit}$ is the fitted y-value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x-value corresponding to either the OLS or PCA fit
- $e_{ijk,fit}$ is the residual between the drawn and fitted y-values for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x-value corresponding to either the OLS or PCA fit
- γ_0 is the overall intercept
- α_i is the effect of the i^{th} parameter choice (F, S, V, N) on the intercept
- γ_1 is the overall slope for x
- γ_{2i} is the effect of the parameter choice on the slope
- x_{ijk} is the x-value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment
- $p_j \sim N(0, \sigma^2_{participant})$ is the random error due to the j^{th} participant's characteristics
- $\epsilon_{ijk} \sim N(0, \sigma^2)$ is the residual error.

Eliminating the linear trend constraint, the bam function in the mgcv package (Wood 2011, Wood et al. 2016, Wood 2004, 2017, 2003) is used to fit a generalized additive mixed model (GAMM) separately to the OLS and PCA residuals to allow for estimation of smoothing splines. Parameter choice was treated as a fixed effect with no estimated intercept and a separate smoothing spline for x was estimated for each parameter choice. A random participant effect accounting for variation due to participant and a random spline for each participant accounted for variation in spline for each participant. The GAMM equation for each fit (OLS and PCA) residuals is given by:

$$y_{ijk,drawn} - \hat{y}_{ijk,fit} = e_{ijk,fit} = \alpha_i + s_i(x_{ijk}) + p_j + s_j(x_{ijk}) \tag{3}$$

where

- $y_{ijk,drawn}$ is the drawn y-value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x-value
- $\hat{y}_{ijk,fit}$ is the fitted y-value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x-value corresponding to either the OLS or PCA fit

- $e_{ijk,fit}$ is the residual between the drawn and fitted y-values for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x-value corresponding to either the OLS or PCA fit
- α_i is the intercept for the parameter choice i
- s_i is the smoothing spline for the i^{th} parameter choice
- x_{ijk} is the x-value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment
- $p_j \sim N(0, \sigma^2_{participant})$ is the error due to participant variation
- s_j is the random smoothing spline for each participant.

Fig. 4 and Fig. 5 show the estimated trends of residuals (vertical deviation of participant drawn points from both the OLS and PCA fitted points) as modeled by a LMM and GAMM respectively. Examining the plots, the estimated trends of PCA residuals (orange) appear to align closer to the $y = 0$ horizontal (dashed) line than the OLS residuals (blue). In particular, this trend is more prominent in parameter choices with large variances (F and N). These results are consistent to those found in Mosteller et al. (1981) indicating participants fit a trend line closer to the estimated regression line with the slope of the first principal component than the estimated OLS regression line.

4 Discussion and Conclusion

The intent of this paper was to establish you draw it as a tool. Eye Fitting Straight Lines in the Modern Era replicated the results found in Mosteller et al. (1981). When shown points following a linear trend, participants tended to fit the slope of the first principal component over the slope of the least squares regression line. This trend was most prominent when shown data simulated with larger variances. The reproducibility of these results serve as evidence of the reliability of the you draw it method.

5 Future Work

Use the method. *Should probably elaborate here..*

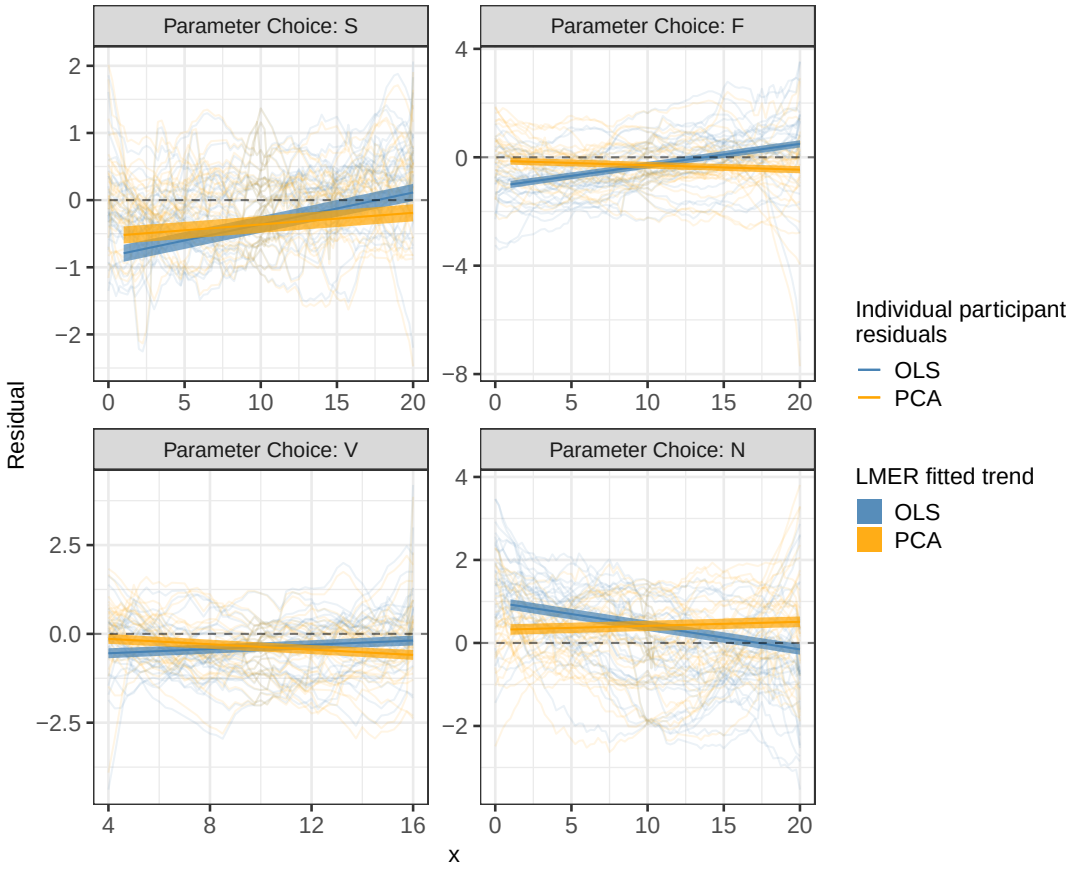

Figure 4: Eye Fitting Straight Lines in the Modern Era LMM results

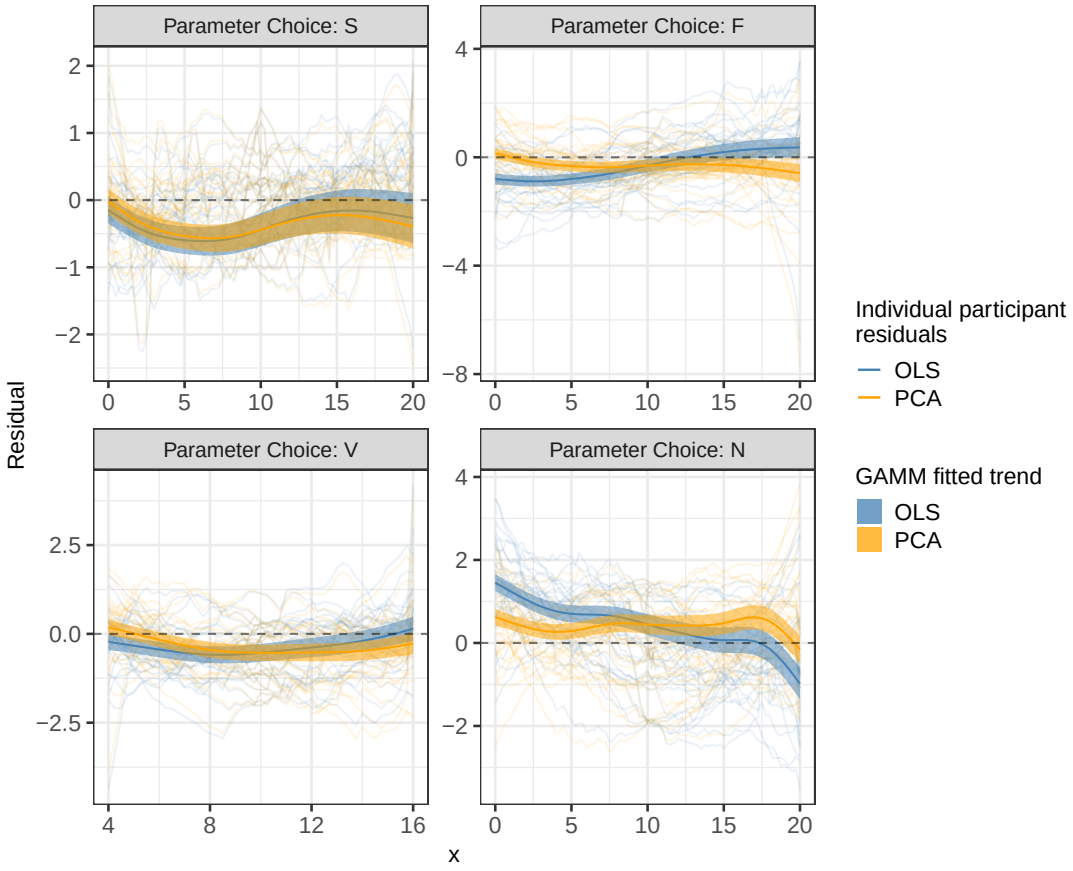

Figure 5: Eye Fitting Straight Lines in the Modern Era GAMM results

References

Aisch, G., Cox, A. & Quealy, K. (2015), 'You draw it: How family income predicts children's college chances'.
URL: *https://www.nytimes.com/interactive/2015/05/28/upshot/you-draw-it-how-family-income-affects-childrens-college-chances.html*

Barry, D., Buchanan, L., Cargill, C., Daniel, A., Delaquérière, A., Gamio, L., Gianordoli, G., Harris, R., Harvey, B., Haskins, J. & et al. (2020), 'Remembering the 100,000 lives lost to coronavirus in america'.
URL: *https://www.nytimes.com/interactive/2020/05/24/us/us-coronavirus-deaths-100000.html*

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015), 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software* **67**(1), 1–48.

Buchanan, L., Park, H. & Pearce, A. (2017), 'You draw it: What got better or worse during obama's presidency'.
URL: *https://www.nytimes.com/interactive/2017/01/15/us/politics/you-draw-obama-legacy.html*

Ciccione, L. & Dehaene, S. (2021), 'Can humans perform mental regression on a graph? accuracy and bias in the perception of scatterplots', *Cognitive Psychology* **128**, 101406.

Finney, D. (1951), 'Subjective judgment in statistical analysis: An experimental study', *Journal of the Royal Statistical Society: Series B (Methodological)* **13**(2), 284–297.

Finney, D. J. & Stevens, W. (1948), 'A table for the calculation of working probits and weights in probit analysis', *Biometrika* **35**(1/2), 191–201.

Giesbrecht, F. G. & Gumpertz, M. L. (2004), *Planning, construction, and statistical analysis of comparative experiments*, Vol. 405, John Wiley & Sons.

Katz, J. (2017), 'You draw it: Just how bad is the drug overdose epidemic?'.
URL: *https://www.nytimes.com/interactive/2017/04/14/upshot/drug-overdose-epidemic-you-draw-it.html*

Lewandowsky, S. & Spence, I. (1989), 'The perception of statistical graphs', *Sociological Methods & Research* **18**(2-3), 200–242.

Mosteller, F., Siegel, A. F., Trapido, E. & Youtz, C. (1981), 'Eye fitting straight lines', *The American Statistician* **35**(3), 150–152.

Ryanabest (2021), 'Build an nba contender with our roster-shuffling machine'.
URL: *https://projects.fivethirtyeight.com/nba-trades-2021/*

Unwin, A. (2020), 'Why is data visualization important? what is important in data visualization?', *Harvard Data Science Review* **2**(1).

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
URL: *https://ggplot2.tidyverse.org*

Wood, S. (2017), *Generalized Additive Models: An Introduction with R*, 2 edn, Chapman and Hall/CRC.

Wood, S. N. (2003), 'Thin-plate regression splines', *Journal of the Royal Statistical Society (B)* **65**(1), 95–114.

Wood, S. N. (2004), 'Stable and efficient multiple smoothing parameter estimation for generalized additive models', *Journal of the American Statistical Association* **99**(467), 673–686.

Wood, S. N. (2011), 'Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models', *Journal of the Royal Statistical Society (B)* **73**(1), 3–36.

Wood, S., N., Pya & Saefken, B. (2016), 'Smoothing parameter and model selection for general smooth models (with discussion)', *Journal of the American Statistical Association* **111**, 1548–1575.