‘You Draw It’: Implementing human interaction in R shiny with r2d3

2022 SDSS

Emily A. Robinson

Department of Statistics, University of Nebraska - Lincoln

emily.robinson@huskers.unl.edu

www.emilyarobinson.com

earobinson95

Slides: https://bit.ly/3xav05Y

University of Nebraska Lincoln

Outline

- Background of graphical testing and measuring patterns
- Integration of D3.js and R Shiny
- Test your drawing skills!
- Validation and Application
- Future Work

Background

Testing statistical graphics

- Evaluate design choices and understand cognitive biases through the use of visual tests.
- Researchers conduct studies in which human subjects are asked to conduct tasks related to the perception of statistical charts such as differentiation, prediction, estimation, and extrapolation.

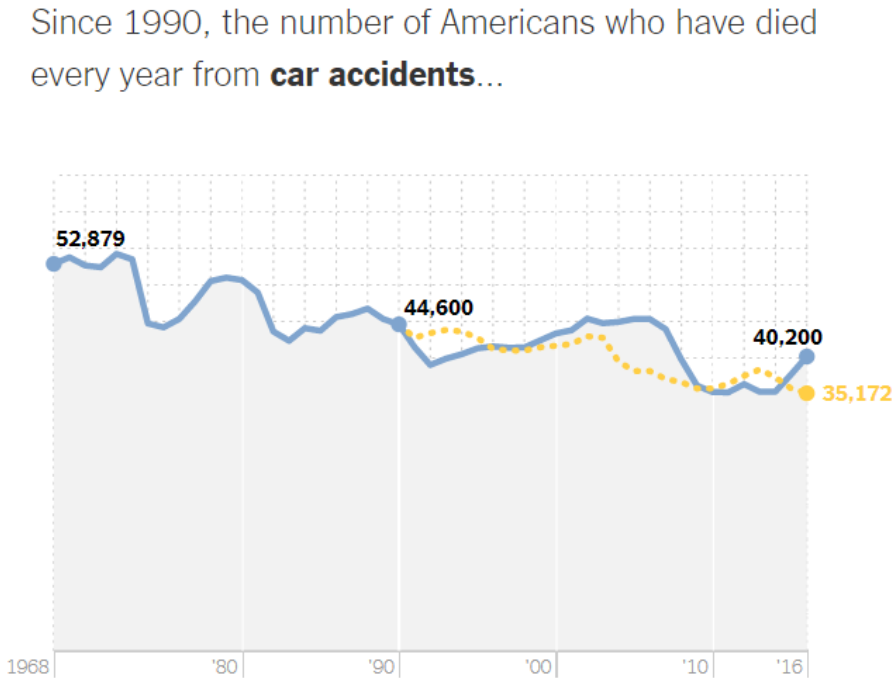
Measuring Patterns & Trends

- Our visual system is naturally built to look for structure and identify patterns.
- Physical manipulation: maneuvering a string, black thread, or ruler until the fit is suitable, then drawing the line through the set of points (Mosteller, Siegel, Trapido, et al., 1981; Finney, 1951)
- Ciccione and Dehaene (2021) conducted a comprehensive set of studies based on psychophysical approaches.

How can we compare our intuitive visual sense of patterns to those determined by statistical methods?

'You Draw It' feature

(New York Times, 2015)



(Katz, 2017)

Readers are asked to input their own assumptions about various metrics and compare how these assumptions relate to reality.

- Family Income affects college chances (Aisch, Cox, and Quealy, 2015)
- Just How Bad Is the Drug Overdose Epidemic? (Katz, 2017)
- What Got Better or Worse During Obama's Presidency (Buchanan, Park, and Pearce, 2017)

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 mouse.

Background of D3

Who? Mike Bostock created D3 during his time working on graphics at the New York Times.

What? Open-source JavaScript based graphing framework

- D3 = "Data Driven Documents"
- `D3` is to JavaScript as `ggplot2` is to R
- Framework for binding objects and layers to plotting area
    - framework for movement and user interaction

When? D3 v1.0 released in 2011.

Where? The internet!

Why? Advantages of using D3 include animation and allowing for movement and user interaction.

How? `r2d3`!

Relationship between D3 and R

The `r2d3` package (Strayer, Luraschi, and Allaire, 2020) in R provides an efficient integration of D3 visuals and R by displaying them in familiar formats:

- RMarkdown with HTML output
- Shiny applications (amazing!)

`r2d3` makes it easy to do your data processing in R, then apply D3.js code to visualize that data!

How?

- Converts data in R to JSON that can be interpreted by JavaScript
- Sources D3 code library
- Creates plot container (svg)
- Renders plot using source code

```
r2d3(data = data,
     script = "d3-source-code.js",
     d3_version= "5")
```

Getting started with D3

`D3.js` is to JavaScript as `ggplot2` is to R

JS Codecademy: Introduction to JavaScript

Understand SVG elements: inspect elements in web browser!

Amelia Wattenberger's Full Stack D3 and Data Visualization Book

Build a basic graphic using r2d3

Modify `D3.js` code until it does what you want!

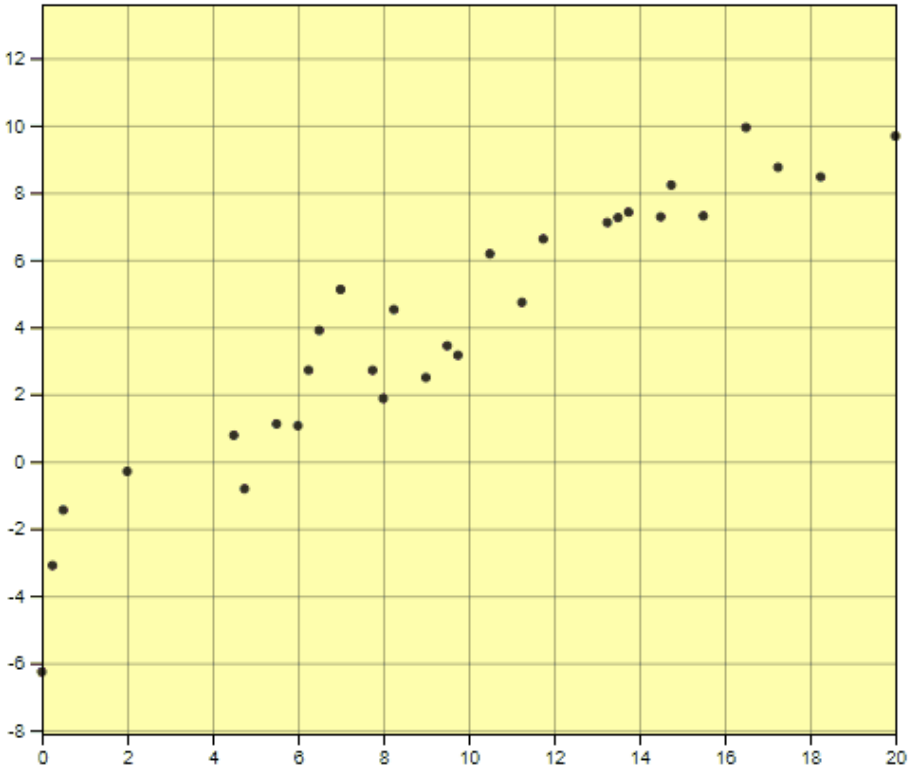
Additional Resources

How to learn D3 with no coding experience

Amelia Wattenberger on Twitter

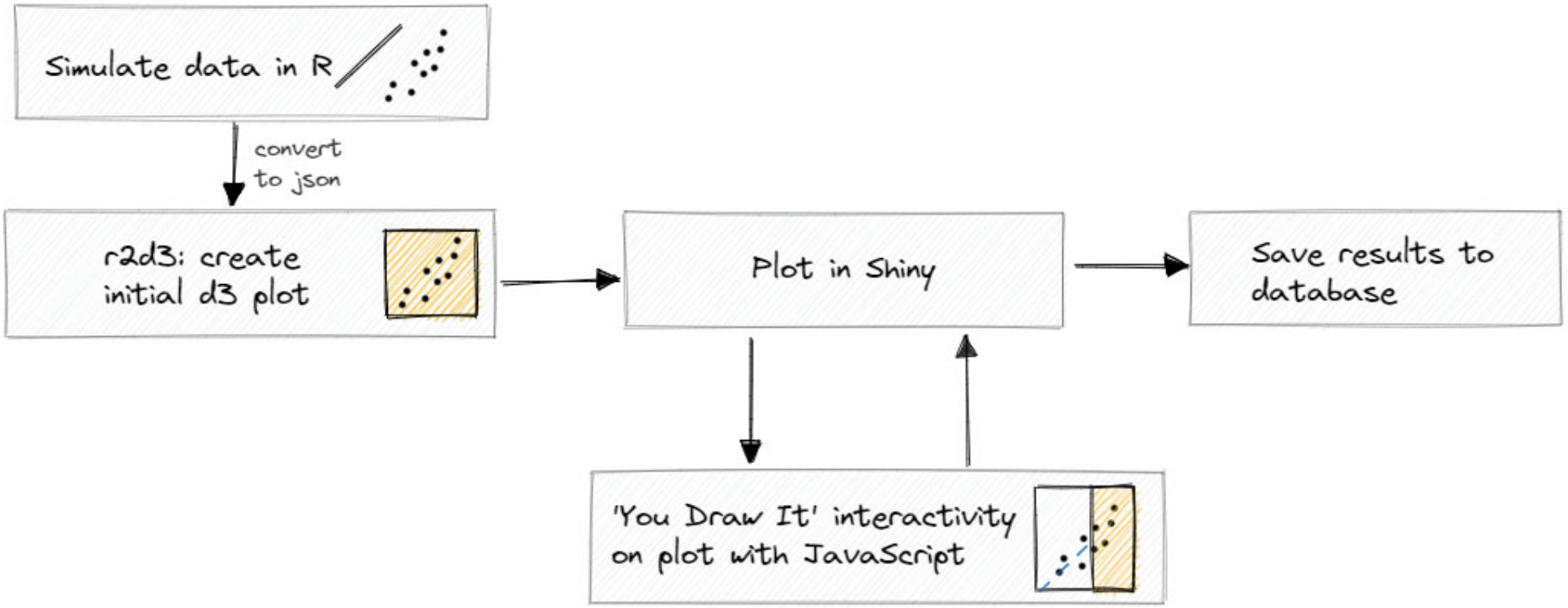
'You Draw It' task plot

Prompt: *Use your mouse to fill in the trend in the yellow box region.*



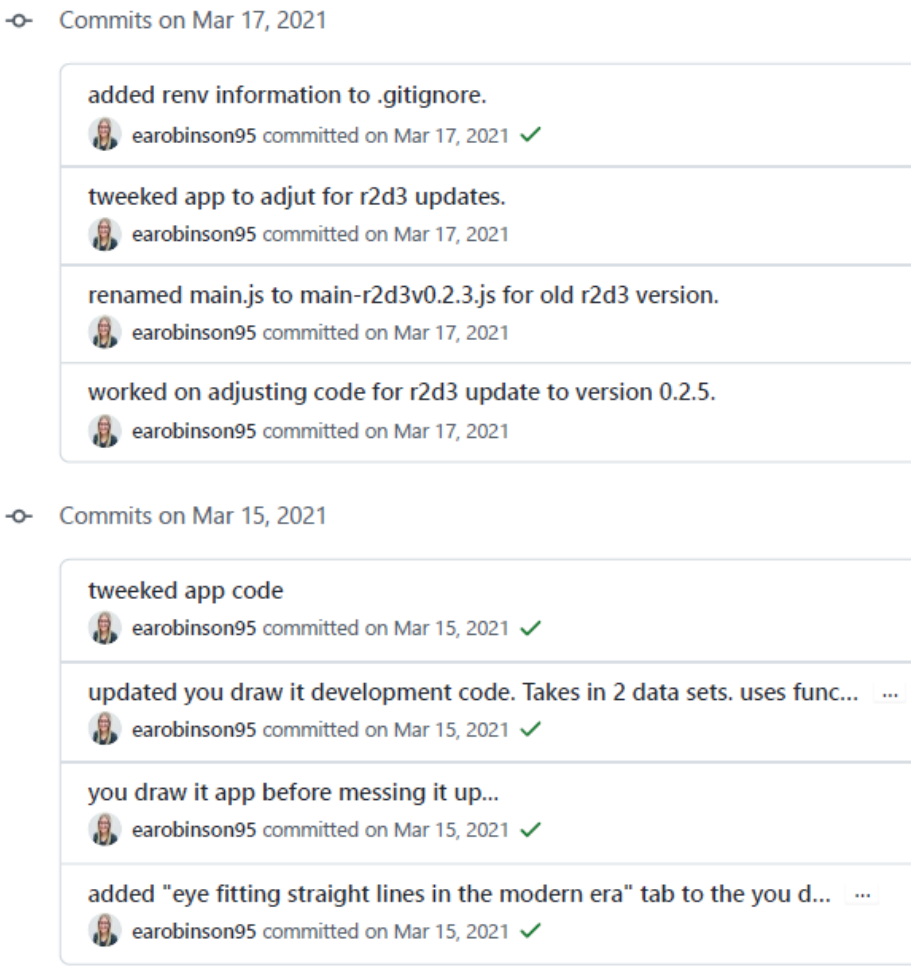
Test out your drawing skills by scanning the QR code on your mobile device or visit emily-robinson.shinyapps.io/can-you-draw-it on a laptop.

Code sketch


Simulate data in R
convert to json
r2d3: create initial d3 plot
Plot in Shiny
Save results to database
'You Draw It' interactivity on plot with JavaScript

Challenges

- Converting between pixels and coordinate points.
- Passing two data sets into `r2d3()` requires converting to json first.
- Layering elements with opacity.
- D3 version matters!
- Only works with a one-to-one function.



Validation study

Replicated *Eye Fitting Straight Lines* by Mosteller, Siegel, Trapido, et al. (1981)

- **Big Idea:** Students fitted lines by eye to four sets of points.
- **Method:** 8.5 x 11 inch transparency with a straight line etched across the middle.
- **Sample:** 153 graduate students and post docs in Introductory Biostatistics.
- **Experimental Design:** Latin square.
- **Findings:** Students tended to fit the slope of the first principal component.

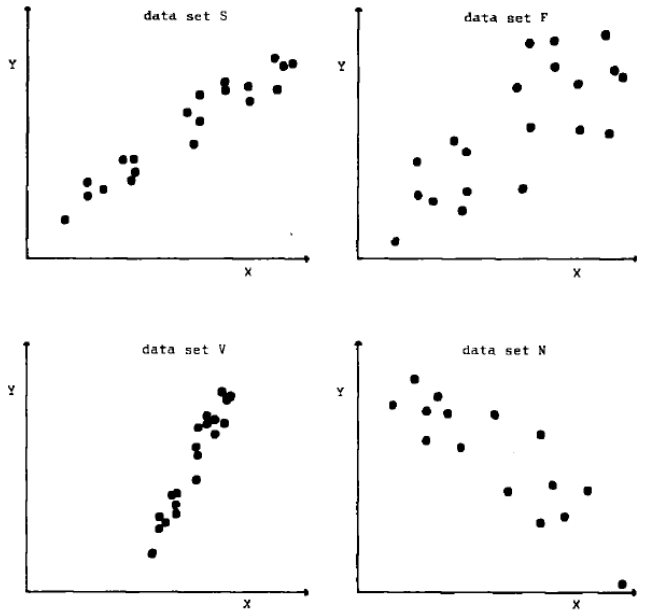
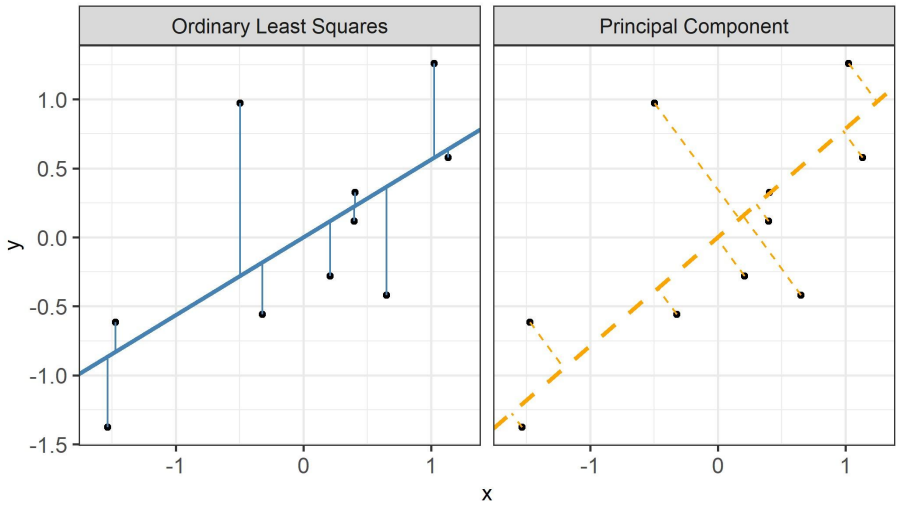


Figure 1. The Data Sets of S, F, V, and N

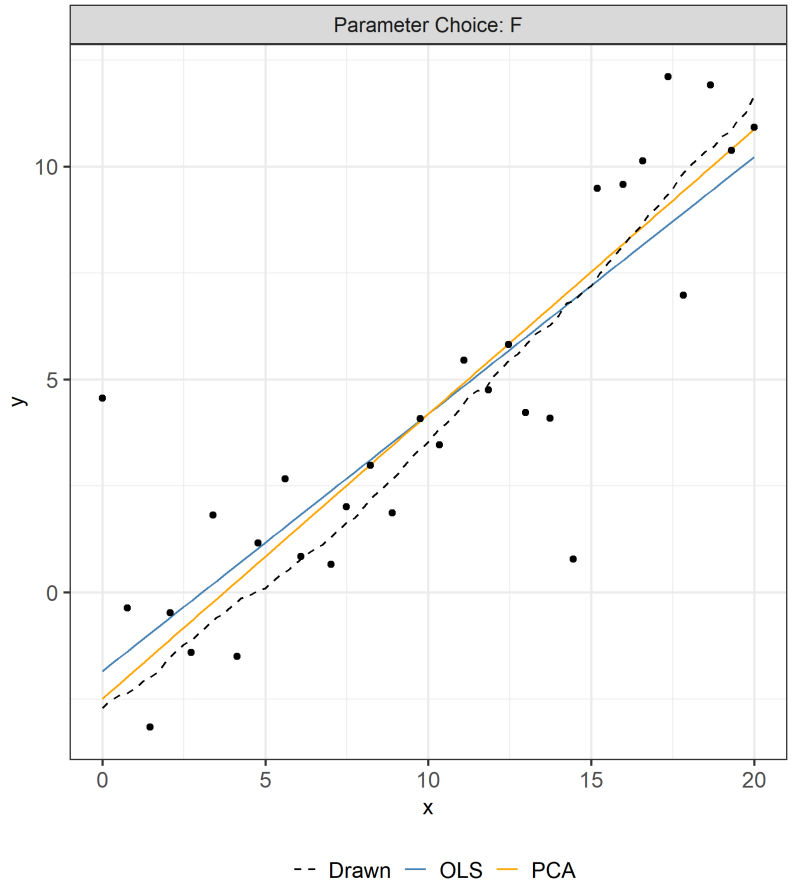
Experiment conducted and distributed through an R Shiny application found at emily-robinson.shinyapps.io/you-draw-it-validation-applet/ in May 2021 with 35 participants completing 119 unique 'You Draw It' task plots.

Validation study

Compare participant drawn line to **statistical regression results**.

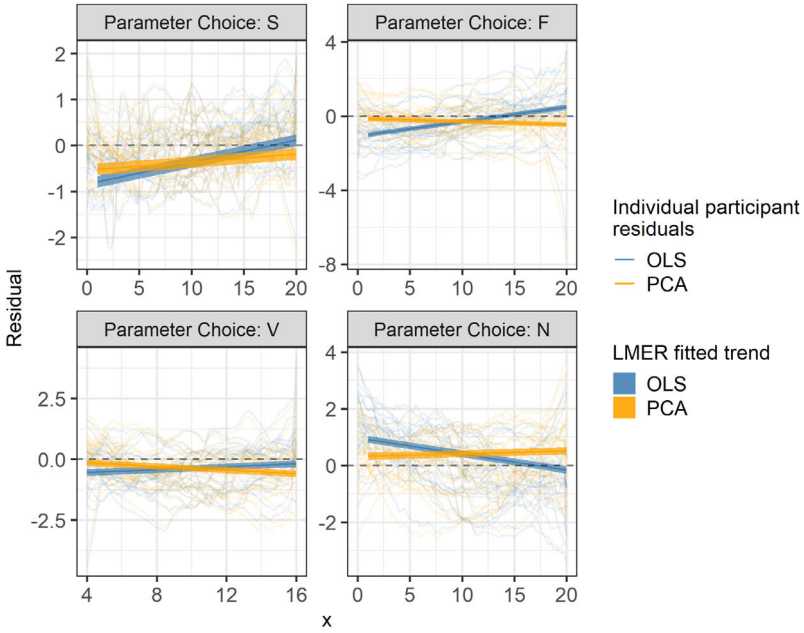


Interested in the **vertical residuals** between the drawn and fitted values.

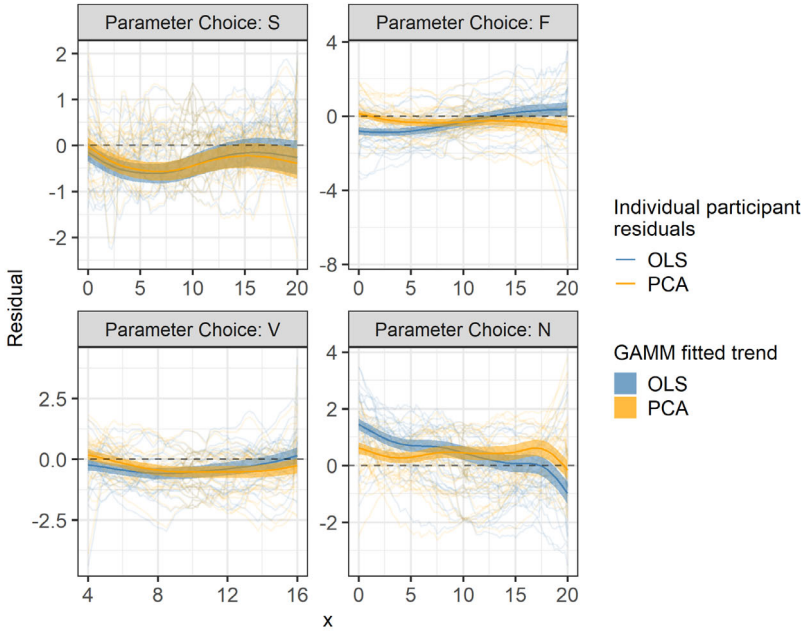


Validation study

Linear Trend Constraint



Generalized Additive Mixed Model (GAMM)



Coming soon for 'You Draw It'!

Software Development

R package designed for easy implementation of ‘You Draw It’ task plots.

Future Applications

Evaluate human's ability to make future predictions from trends and to fit curved trend lines.

Apply 'You Draw It' to other fields in order to evaluate science communication using graphics.

Publications

[Under Review] Robinson, E. A., VanderPlas, S., Howard, R., (2021) Eye Fitting Straight Lines in the Modern Era. Submitted to *Journal of Computational and Graphical Statistics*. Access on GitHub.

References

Aisch, G., A. Cox, and K. Quealy (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.

Buchanan, L., H. Park, and A. Pearce (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. and S. Dehaene (2021). "Can humans perform mental regression on a graph? Accuracy and bias in the perception of scatterplots". In: Cognitive Psychology 128, p. 101406.

Finney, D. (1951). "Subjective judgment in statistical analysis: An experimental study". In: Journal of the Royal Statistical Society: Series B (Methodological) 13.2, pp. 284–297.

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.

Mosteller, F., A. F. Siegel, E. Trapido, et al. (1981). "Eye fitting straight lines". In: The American Statistician 35.3, pp. 150–152.

Strayer, N., J. Luraschi, and J. Allaire (2020). r2d3: Interface to 'D3' Visualizations. R package version 0.2.5. URL: https://CRAN.R-project.org/package=r2d3.

Thank you!

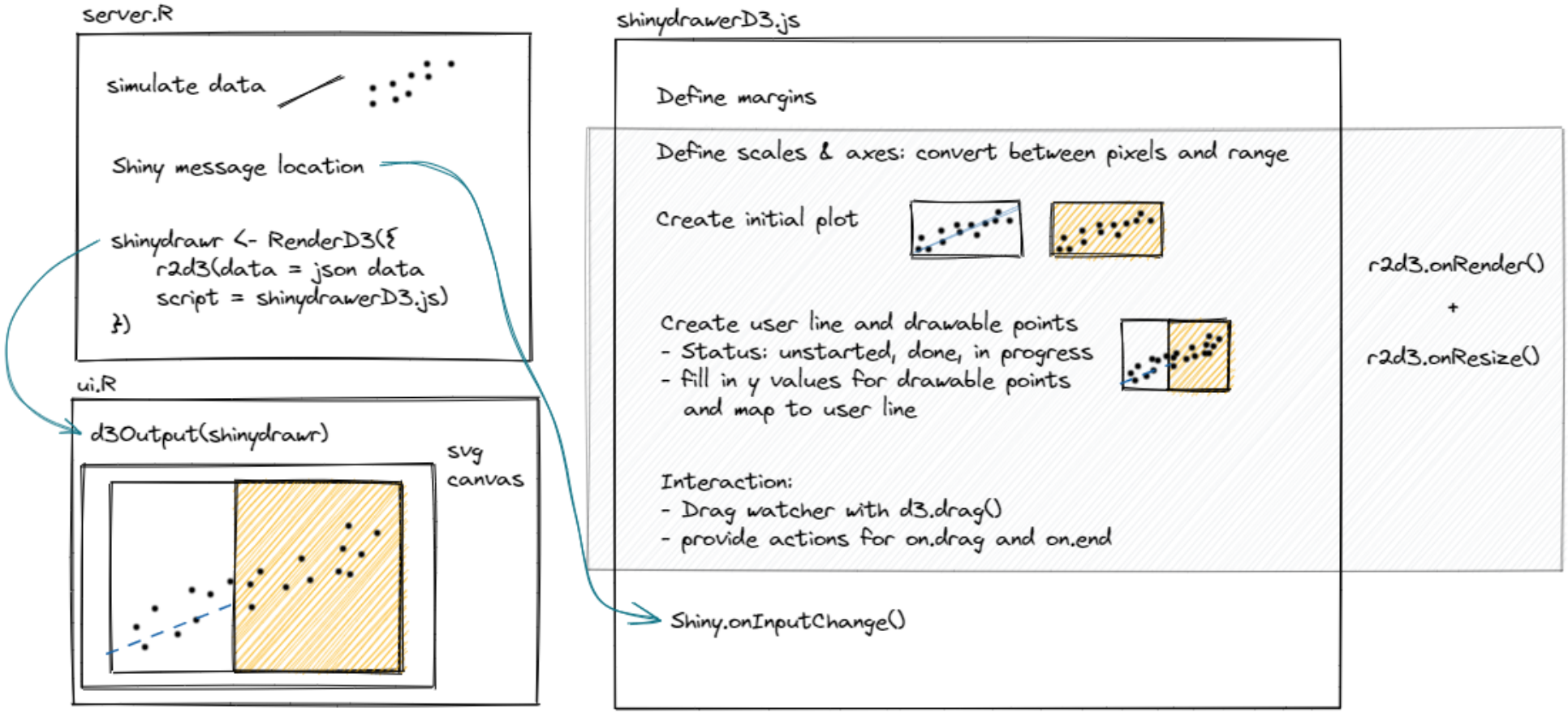
Acknowledgments: Dr. Susan VanderPlas and Dr. Reka Howard

emily.robinson@huskers.unl.edu

www.emilyarobinson.com

earobinson95

Code Sketch (detailed)



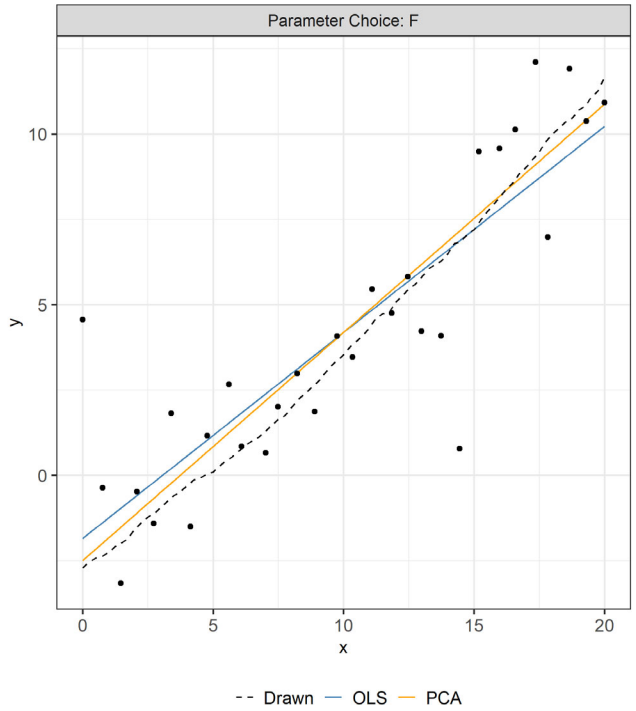
Feedback Data Notation

For each participant, the final data set used for analysis contains:

- $x_{ijk}, y_{ijk,drawn}, \hat{y}_{ijk,OLS}, \hat{y}_{ijk,PCA}$

for

- parameter choice $i = 1, 2, 3, 4,$
- participant j = $1, \ldots N_{participant}$
- x_{ijk} value corresponding to increment $k = 1, \ldots, 4x_{max} + 1.$



Vertical residuals between the drawn and fitted values were calculated as:

- $e_{ijk,OLS} = y_{ijk,drawn} - \hat{y}_{ijk,OLS}$
- $e_{ijk,PCA} = y_{ijk,drawn} - \hat{y}_{ijk,PCA}.$

LMM Equation

The **Linear Mixed Model** equation for each fit (OLS and PCA) residuals is given by:

$$e_{ijk,fit} = [\gamma_0 + \alpha_i] + [\gamma_1 x_{ijk} + \gamma_{2i} x_{ijk}] + p_j + \epsilon_{ijk}$$

where

- $e_{ijk,fit}$ 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.

GAMM Equation

The **Generalized Additive Mixed Model** equation for each fit (OLS and PCA) residuals is given by:

$$e_{ijk,fit} = \alpha_i + s_i(x_{ijk}) + p_j + s_j(x_{ijk})$$

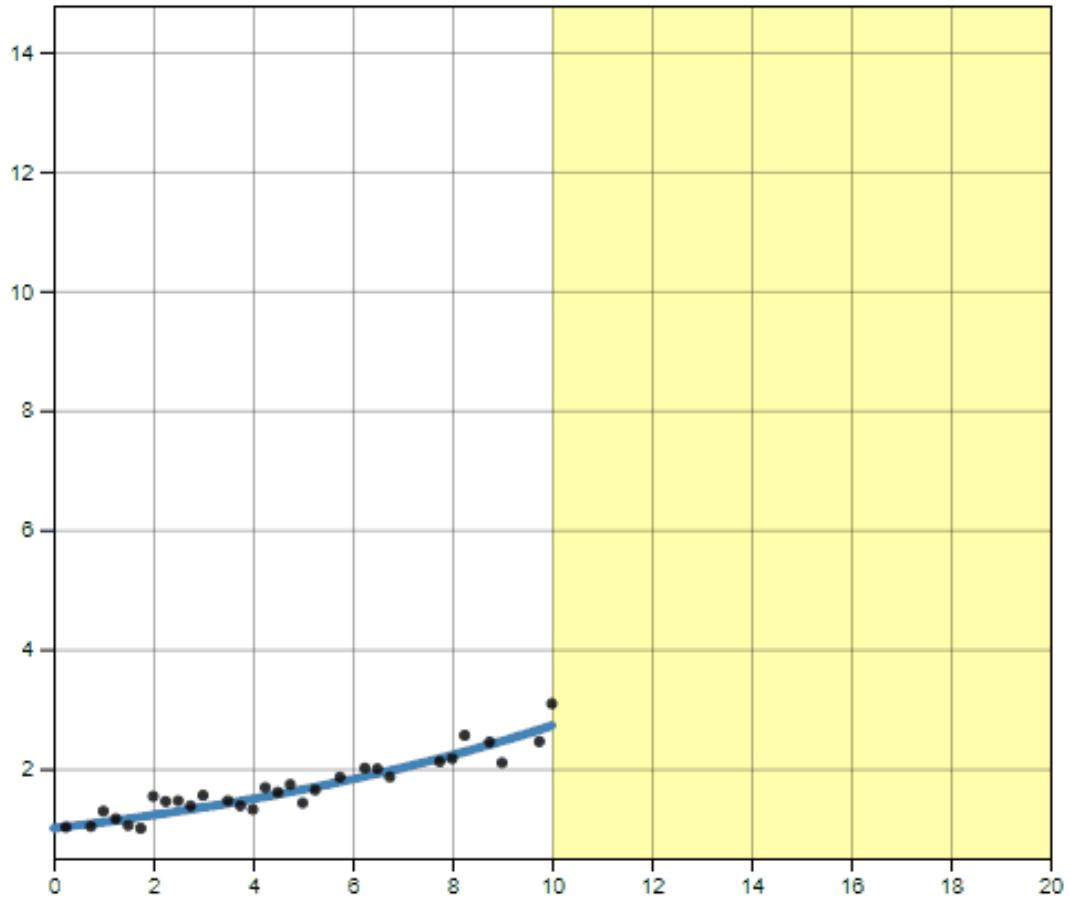
where

- $e_{ijk,fit}$ 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.

Exponential Growth Study

Exponential Growth 'You Draw It'

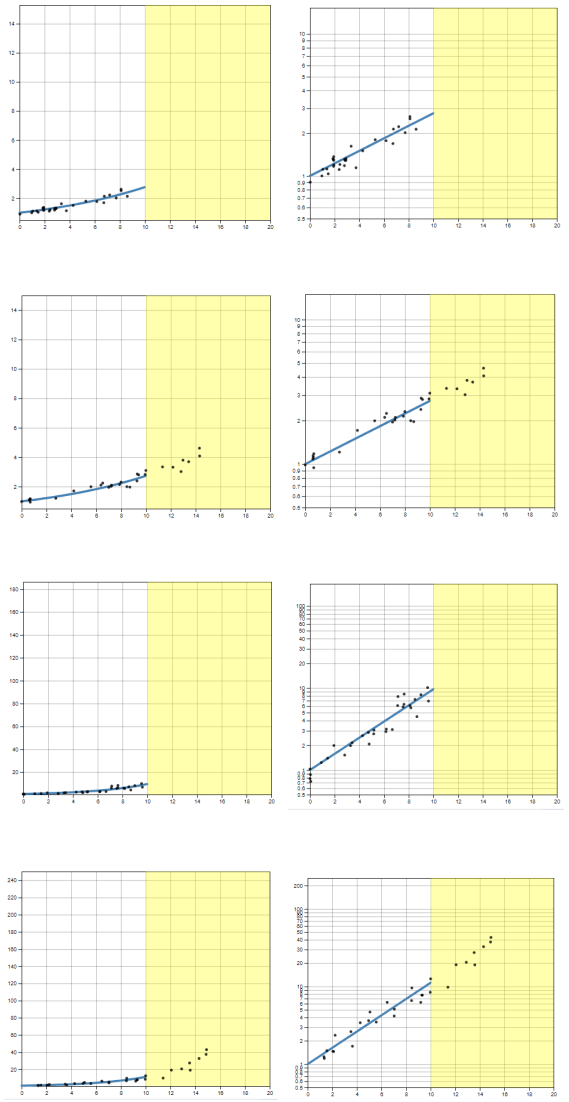
Study Participant Prompt: *Use your mouse to fill in the trend in the yellow box region.*



Treatment design

2 x 2 x 2 factorial:

- **growth rate:** low and high.
- **points truncated:** 50% and 75% of the domain.
- **scale:** log and linear.



Feedback data

For each participant, the final data set used for analysis contains:

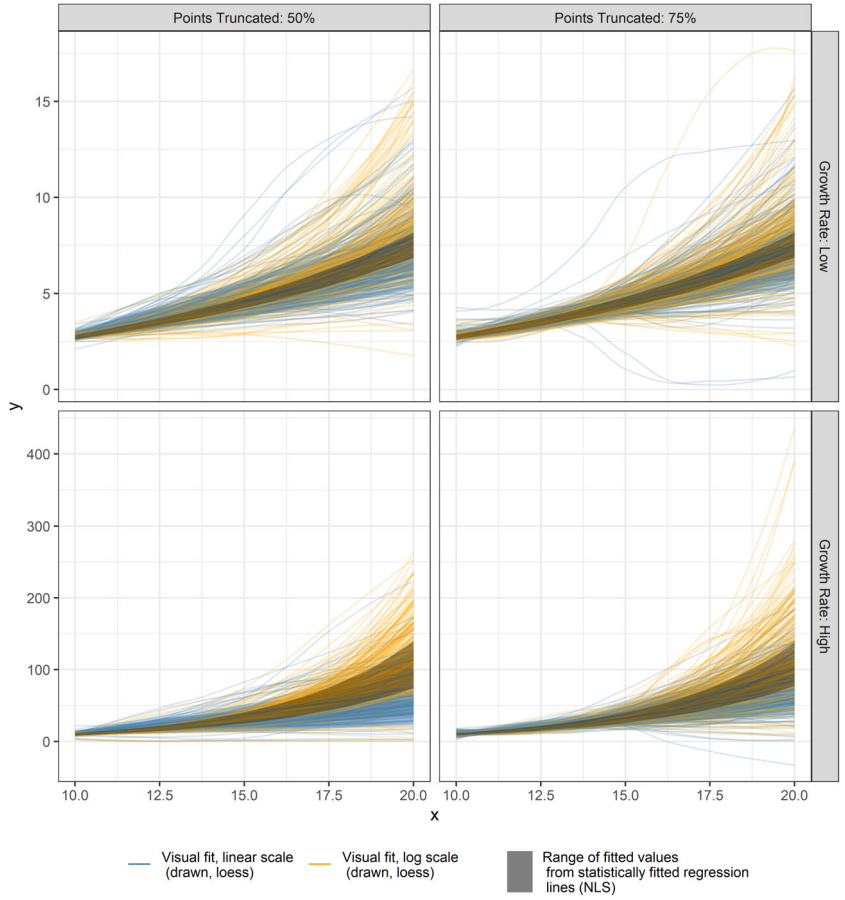
- x_{ijklm}, $y_{ijklm,drawn}$, and $\hat{y}_{ijklm,NLS}$

for:

- growth rate $i = 1, 2,$
- point truncation $j = 1, 2,$
- scale $k = 1, 2,$
- participant $l = 1, \ldots N_{participant}$, and
- x_{ijklm} value $m = 1, \ldots, 4x_{max} + 1$.

Vertical residuals between the drawn and fitted values were calculated as:

- $e_{ijklm,NLS} = y_{ijklm,drawn} - \hat{y}_{ijklm,NLS}$.



Generalized Additive Mixed Model

The GAMM equation for residuals is given by:

$$e_{ijklm,nls} = \tau_{ijk} + s_{ijk}(x_{ijklm}) + p_l + s_l(x_{ijklm})$$

where

- $e_{ijklm,NLS}$ is the residual between the drawn y-value and fitted y-value for the l^{th} participant, m^{th} increment, and ijk^{th} treatment combination
- τ_{ijk} is the intercept for the i^{th} growth rate, j^{th} point truncation, and k^{th} scale treatment combination
- s_{ijk} is the smoothing spline for the ijk^{th} treatment combination
- x_{ijklm} is the x-value for the l^{th} participant, m^{th} increment, and ijk^{th} treatment combination
- $p_l \sim N(0, \sigma^2_{participant})$ is the error due to the l^{th} participant's characteristics
- s_l is the random smoothing spline for the l^{th} participant.

GAMM results