

## Contents

<b>Introduction</b>	<b>1</b>
<b>Why Visualizations are Critical for LVMs</b>	<b>2</b>
<b>Previous Approaches to Visualizing LVMs</b>	<b>2</b>
<b>Our Approach (Linear LVMs)</b>	<b>2</b>
Diagnostic Plots: Trail Plots . . . . .	2
<b>Disturbance-Dependence Plots</b>	<b>6</b>
<b>Measurement Plots</b>	<b>6</b>
<b>Structural (Beech Ball) Plots</b>	<b>12</b>
<b>References</b>	<b>12</b>

## Introduction

It is currently an unprecedented time in the social sciences; multiple scientific disciplines are reeling from a “replication crisis” (Camerer et al. 2018; Ioannidis 2005; Pashler and Wagenmakers 2012), new norms for credibility are becoming more prevalent (Nelson, Simmons, & Simonsohn, 2018; Nosek, Ebersole, DeHaven, & Mellor, 2018), and the push for open science is accelerating at a rapid pace (Nosek et al. 2018). Amidst this push for open science practices, some have called for greater use of visualization techniques (Fife and Rodgers 2019; Fife 2020b; Tay et al. 2016). As noted by Tay, et al. (2016), “[visualizations]... can strengthen the quality of research by further increasing the transparency of data...” (p. 694). In other words, one of the best, and most efficient ways of making data analysis open and transparent is to display each and every datapoint through visualization techniques.

Not only do visualizations adhere to the principles of openness and transparency, but they offer several additional advantages; they vastly improve encoding of information (Correll 2015), they highlight model misfit (Healy and Moody 2014), and they are an essential component in evaluating model assumptions (Levine 2018; Tay et al. 2016). As such, we (as well as others, e.g., Fife 2019, 2020b; Wilkinson and Task Force on Statistical Inference 1999) insist that every statistical model ought to be accompanied by a graphic.

Unfortunately, this visualization requirement is easier said than done. While visualizing some statistical models is trivial (e.g., regressions, t-tests, ANOVAs, multiple regression), visualizing others is not. One particularly troublesome class of models to visualize is latent variable models (LVMs). While researcher routinely visualize conceptual models (e.g., via path diagrams), visualizing the statistical models is not so easy. The former visualizations are common, while the latter are not (Hallgren et al. 2019). The reason statistical visualizations of LVM are not intuitive is because they rely on unobserved variables (Bollen 1989). If the variables of interest are unobserved, how can we possibly visualize them?

Though it is not, at first glance, easy to visualize unobserved variables, that does not mean visualizing them is any less important. On the contrary, visualizing latent variables is, perhaps more important because

their presence is unobserved. In the following section, we elaborate on why visualizations are particularly crucial for LVMs. We then review previous approaches others have used for visualizing LVMs, and note their strengths and weaknesses. We then introduce our approach and the corresponding R package `flexplavaan`, which allows users to visualize both `lavaan` and `blavaan` objects in R. We then conclude with several examples that highlight how visualizations assisted in identifying appropriate statistical models.

## Why Visualizations are Critical for LVMs

### Previous Approaches to Visualizing LVMs

### Our Approach (Linear LVMs)

#### Diagnostic Plots: Trail Plots

To begin how to conceptualize LVMs, let us first consider how typical linear models are visualized. In a standard regression, each dot in a scatterplot represents scores on the observed variables. Often, analysts overlay additional symbols to represent the fit of the model (e.g., a line to represent the fitted regression model, or large dots to represent the mean). Sometimes additional symbols are overlaid to represent uncertainty (e.g., confidence bands for a regression line or standard error bars). See Figure 1 as an example. In either case, the dots represent observed information, while the fitted information is conveyed using other symbols.

Likewise, visualizing LVMs ought to follow similar conventions; the dots should represent the observed information, as in Bauer (2005). In his visuals, pairwise relationships between observed variables are represented in a scatterplot. However, Bauer’s approach did not overlay a model-implied fit, as we seek to do. When the line represents the model-implied fit, it denotes the trail left behind by the unobserved latent variable. As such, we call these plots “trail plots.” How then does one identify the slope/intercept of the LVM’s model-implied fit? It is quite easy to do so when standard linear LVMs are used. Suppose we have a factor ( $T$ ) with three indicators (e.g.,  $X_1, X_2$  and  $X_3$ ), and we wish to visualize the pairwise trace plot between  $X_1$  and  $X_2$ . To do so, we can simply utilize the model-implied correlation matrix:

$$\beta_{x_1|x_2} = \hat{r}_{x_1,x_2} \frac{s_{x_1}}{s_{x_2}}$$

where  $\hat{r}_{(x_1,x_2)}$  is the model-implied correlation between  $X_1$  and  $X_2$ ,  $s_{x_1}$  and  $s_{x_2}$  are the standard deviations of the two variables. One can then estimate the intercept using basic algebra:

$$b_0 = \bar{X}_1 - \beta_{(x_1|x_2)}\bar{X}_2$$

Figure 2 shows the LVM model-implied fit in red with a regression line in blue for simulated data. Because the regression line minimizes the sum of squared errors, we would hope that the LVM fitted line (red) closely approximate the regression line. In this case, the two overlap quite extensively. On the other hand, if the two lines differ, we can be certain the LVM fails to capture the entire relationship between the two observed variables.

Of course, Figure 2 only shows one pairwise relationship between variables. If we wished to visualize all the variables in our model, we would have to utilize a scatterplot matrix, as in Figure 3. Naturally, this becomes quite cumbersome when users have more than seven or eight variables. In this case, it is best to visualize only a subset of variables. We will later discuss strategies for how best to select appropriate subsets.

The primary advantage of trail plots is that they easily show misfit in LVMs. They do so by showing differences in fit between indicators of different variables. For example, 4 shows a model where  $x_1 - x_2$  load

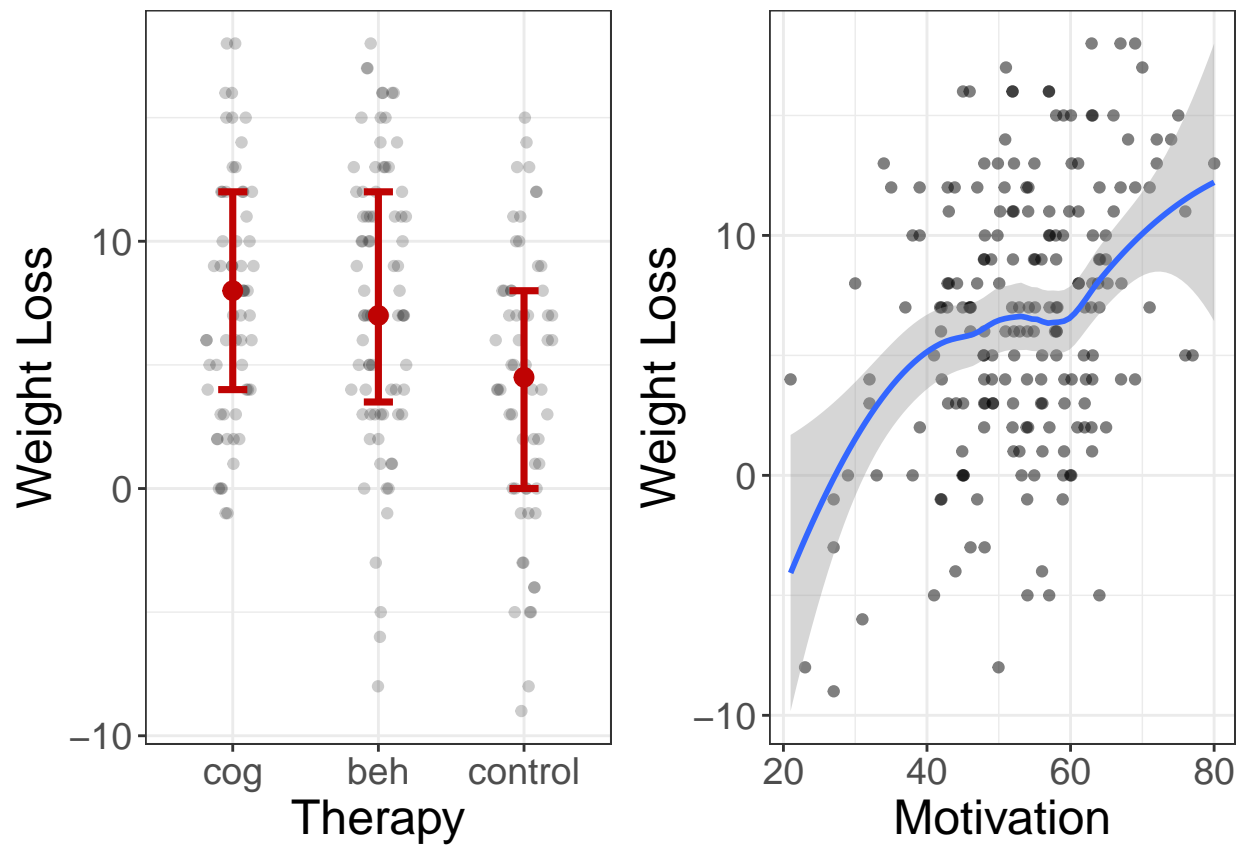


Figure 1: Example figure that shows how standard statistical models are visualized. Dots represent scores on observed variables, while other symbols (e.g., regression line, large dots) represent the fit of the model.

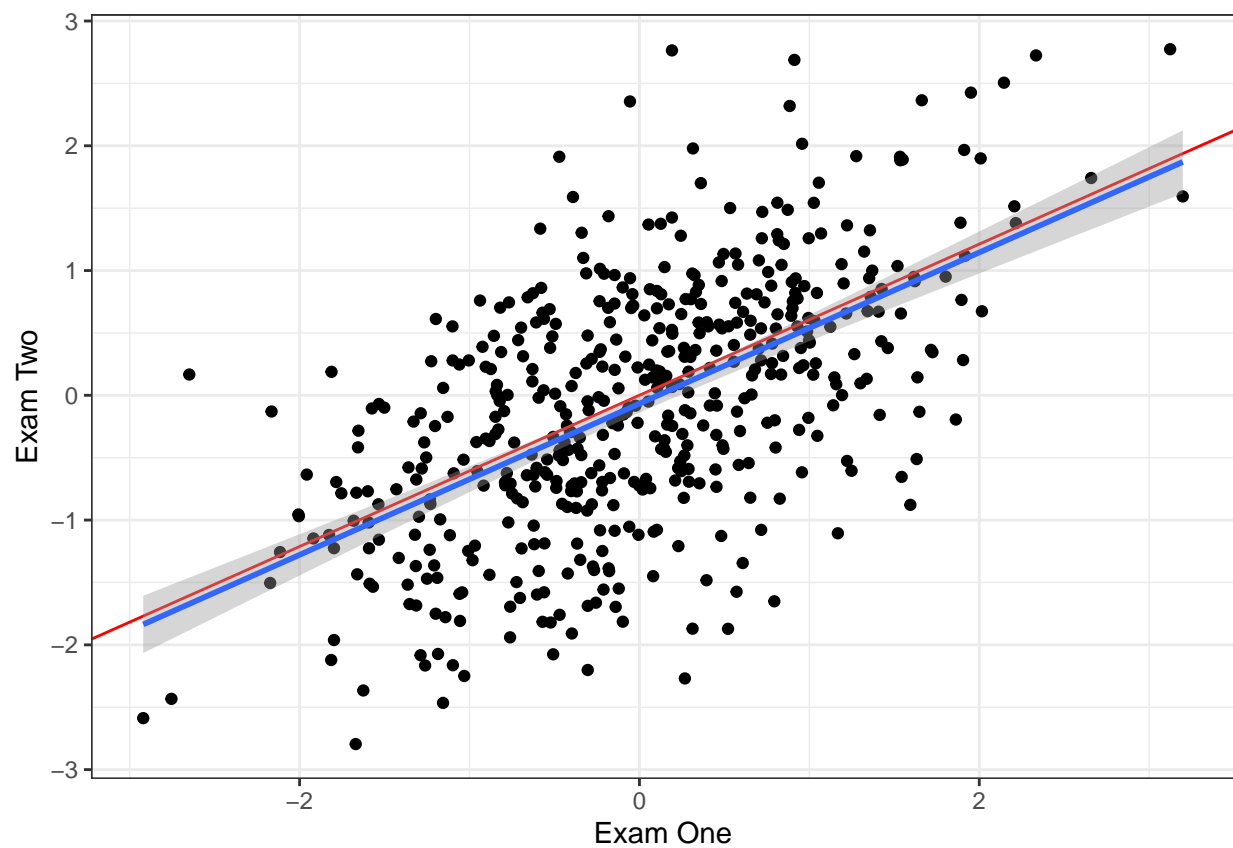


Figure 2: The LVM-implied fit between  $X_1$  and  $X_2$ , shown in red. The blue line represents the regression line between the two variables. The more closely the model-implied fit line resembles the regression line, the better the fit of the LVM.

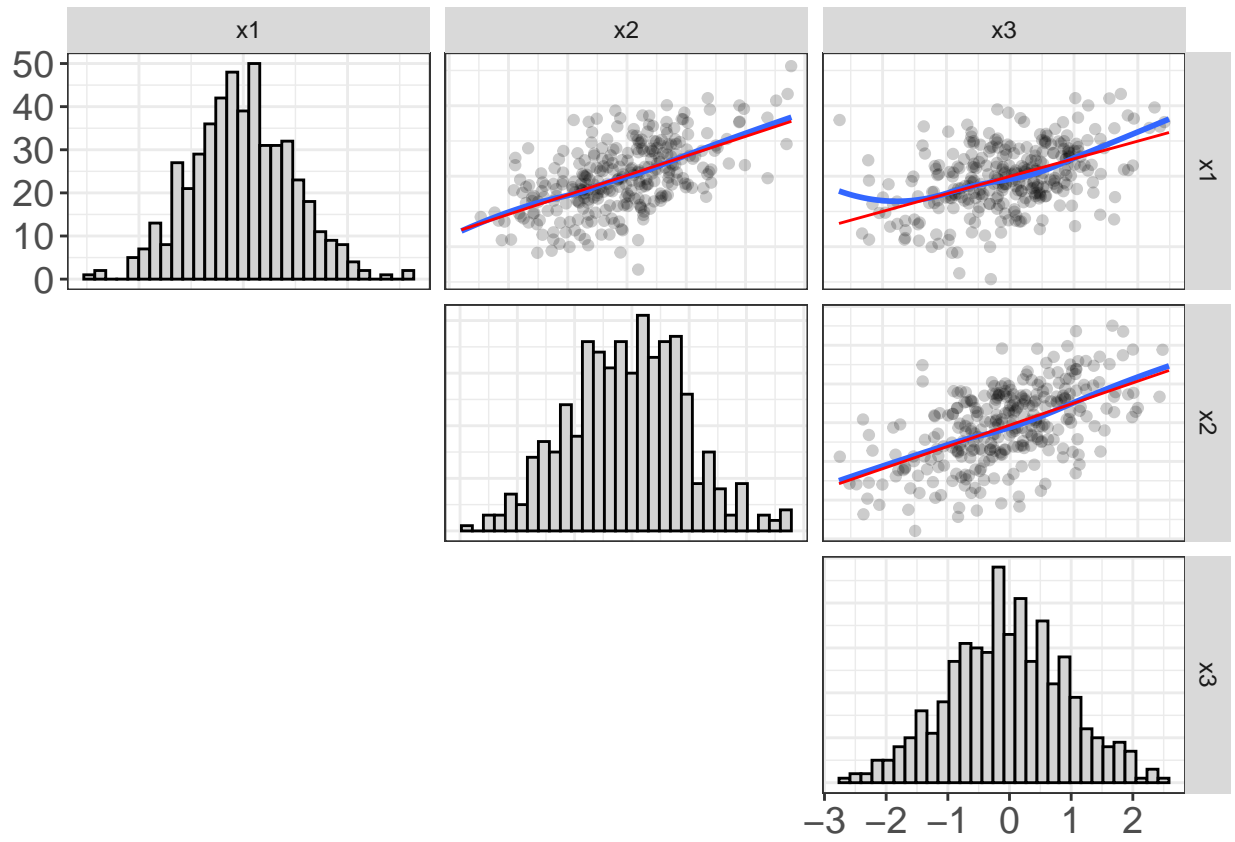


Figure 3: Scatterplot matrix showing the model-implied fit (red) and regression-implied fit (blue) between three simulated indicator variables. The diagonals show the histograms.

onto one latent variable,  $x_4$ - $x_5$  load onto another, and  $x_3$  loads onto both; however, the specified model assumes  $x_3$  only loads onto the first model.

Another advantage of trail plots is they visually (and often times strikingly) show how little information a model might capture. For example, Figure 5 shows a model that, by the fit indices, has remarkably good fit: RMSEA is  $<0.001$ , TLI/CFI/NNFI are all approximately 1, and RMR is approximately zero. However, the visual shows the model actually captures very little information; while the two lines are quite similar, the majority of slopes are very near flat.

## Disturbance-Dependence Plots

One common technique for visualizing the adequacy of statistical models in classic regression is residual-dependence plots. With these graphics, one simply plots the residuals of the model ( $Y$  axis) against the predicted values ( $X$  axis). The rationale behind this is simple: the model should have extracted any association between the prediction and the outcome. The residuals represent the remaining information after extracting the signal from the model. If there is a clear trend remaining in the data (e.g., a nonlinear pattern or a “megaphone” shape in the residuals), this indicates the model failed to capture important information.

Likewise, in LVMs, we can apply this same idea to determine whether the fit implied by the LVM has successfully extracted any association between any pair of predictors. However, in LVMs, residuals refer to the discrepancy between the model-implied and the actual variance/covariance matrix (or correlation matrix). As such, naming these plots “residual-dependence plots” would be a misnomer. Rather, misfit at the raw data level is typically called a disturbance. As such, we call these plots disturbance-dependence plots.

Like trace plots, we visualize disturbance-dependence plots for each pair of observed variables. To do so, `flexplavaan` subtracts the fit implied by the trace plots from the observed scores. For example, a disturbance dependence plot for an  $X_1/X_2$  relationship would subtract the “fit” of  $X_2$  implied by the trace plot from the actual  $X_2$  scores (and vice versa for the  $X_2/X_1$  relationship). If the trace-plot fit actually extracts all association between the pair of observed variables, we would expect to see a scatterplot that shows no remaining association between the two. If there is a pattern in the scatterplot remaining, we know the fit of the model misses important information about that specific relationship. To aid in interpreting these plots, we can overlay the plot with a flat line (with a slope of zero), as well as a regression (or loess) line. The first line indicates what signal should remain after fitting the model, while the second line shows what actually remains. Figure 6 shows an example of trace plots in the upper triangle and disturbance-dependence plots in the lower triangle of a scatterplot matrix. These plots are for the same data shown in the right image of Figure 4. Notice how the plots associated with  $X_3$  all have positive slopes, indicating the model failed to capture important signal remaining in the model.

Together, both of these plots (trace plots and diagnostic-dependence plots) serve as a critical diagnostic check. Both these plots will signal misfit both in the measurement and structural components of the model. Conversely, these models will also help users determine whether the model is to be believed. If they show the model adequately fits the data, the user can then proceed to plot two different types of plots: measurement plots and structural plots.

## Measurement Plots

One of the primary purposes of the diagnostics is to determine whether one’s conceptualization of the latent variables is to be believed. If the trace plots and disturbance dependence plots indicate the LVM is a good representation of the data, one can be more confident the latent variables are properly estimated. If this is the case, we can now make a step toward visualizing the latent variables themselves.

The approach we suggest is to plot the factor scores on the  $Y$  axis and the observed scores on the  $X$  axis. Naturally, this means one could really only visualize one indicator at a time, which is a serious limitation for

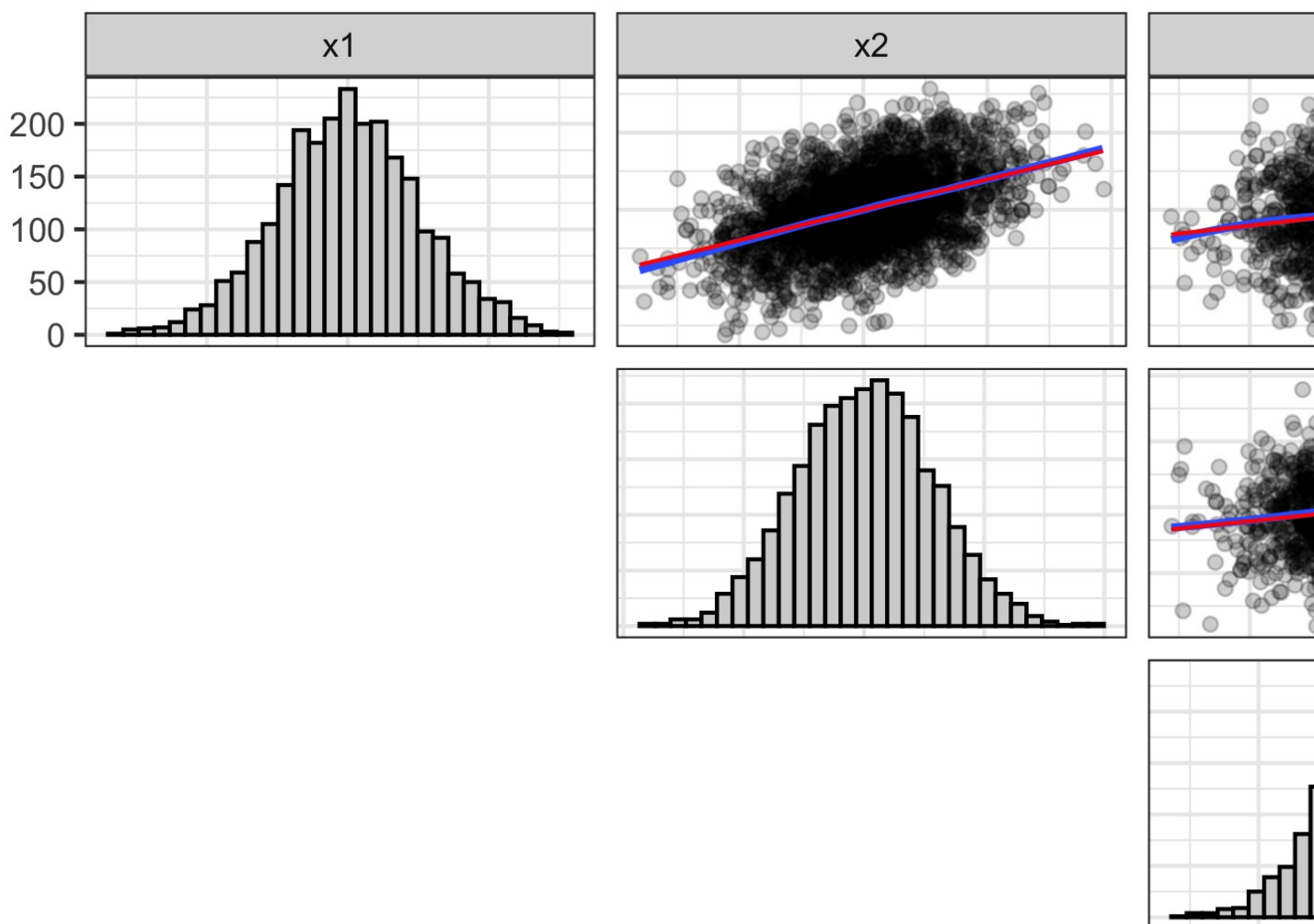


Figure 4: This plot shows structural misfit;  $x_3$  loads onto two factors, but only one is modeled.

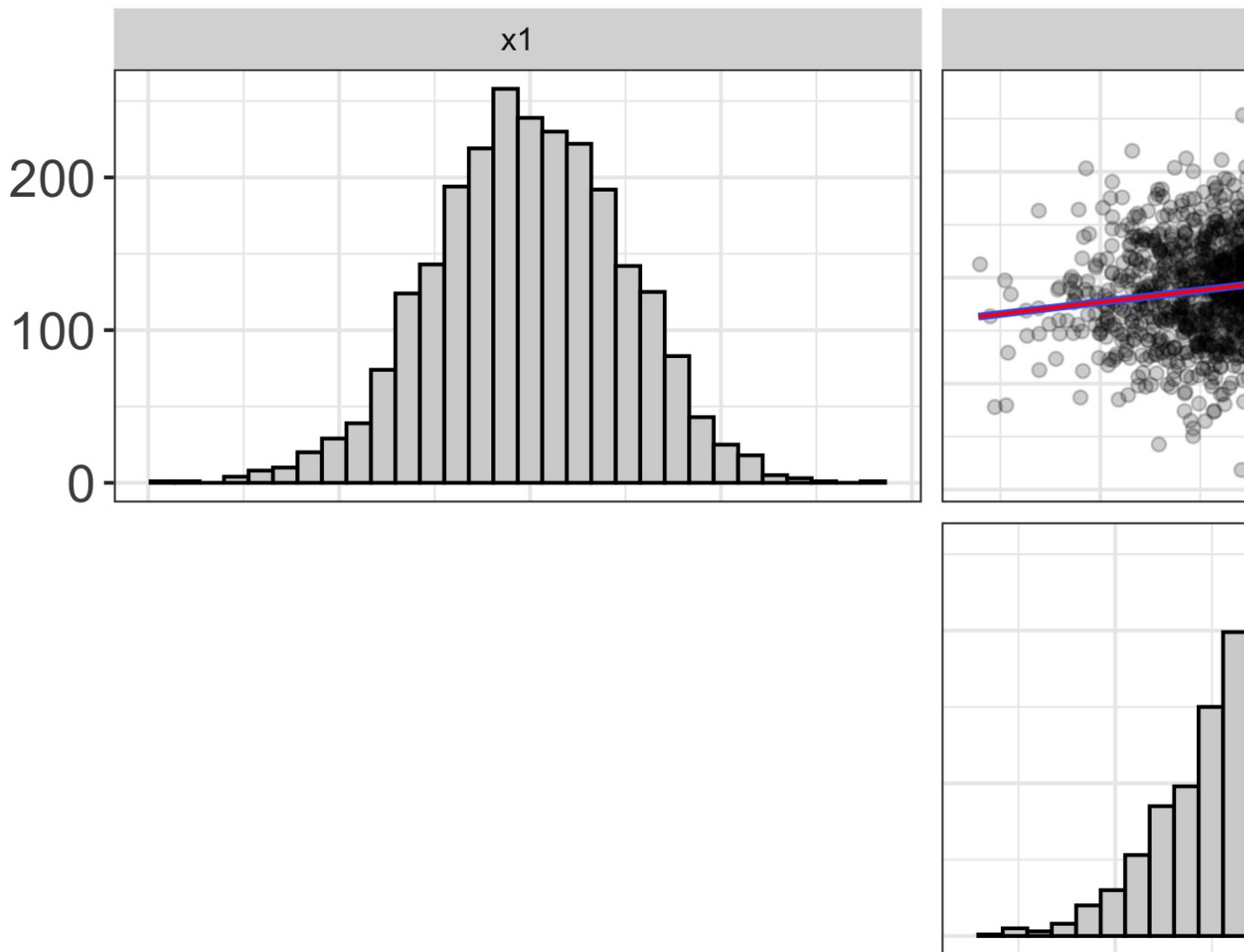


Figure 5: This plot shows a model that, by standard fit indices, fits quite well. The visuals, on the other hand, illustrate that the model isn't capturing



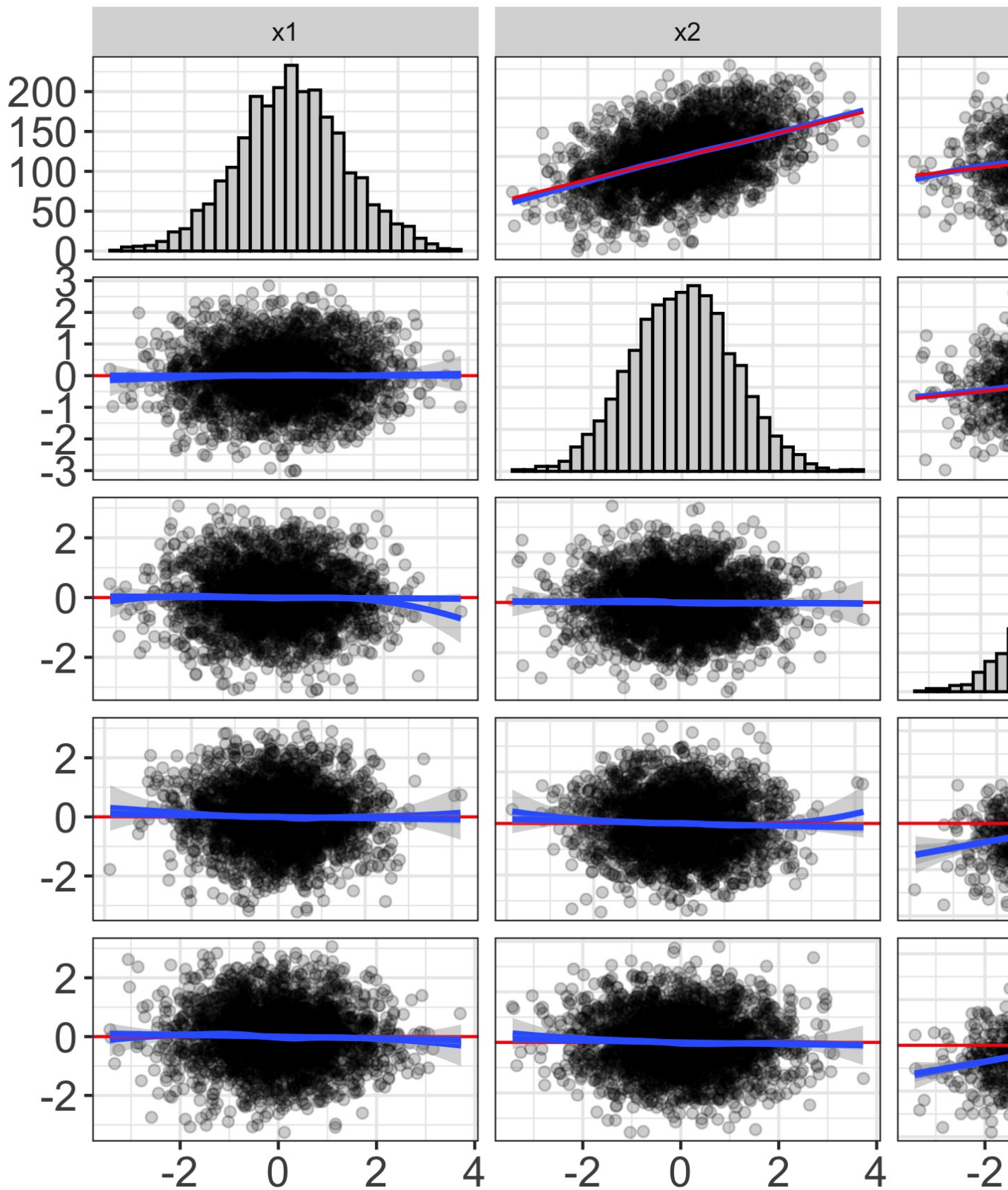


Figure 6: This plot shows a disturbance-dependence plot for the same data visualized in Figure 4 in the lower triangle.

Table 1: A Table of Simulated Data Showing the Data in Wide Format

x1	x2	x3
11	8	10
12	9	8
12	7	10

Table 2: This Table Contains the Same Data in Table 1, but standardized and in “Long” Rather Than “Wide” Format.

Measure	Observed
x1	-1.1547005
x1	0.5773503
x1	0.5773503
x2	0.0000000
x2	1.0000000
x2	-1.0000000
x3	0.5773503
x3	-1.1547005
x3	0.5773503

most (if not all) latent variable models. To overcome this problem, we recommend paneling each indicator, as in Figure 7. To do so, this requires converting our data from “wide” to “long” format (see Table 2), meaning that variables that once occupied separate columns (e.g.,  $x_1$ ,  $x_2$ , and  $x_3$ ; see Table 1) are now collapsed into two columns, one of which contains the values of  $x_1 - x_3$ , the other of which indicates which variable the measurement belongs to (see Table 2). We should also convert each variable to  $z$ -scores to make it easier to compare the observed/latent relationship across indicators, as in Table 2.

Another alteration from a standard scatterplot is the use of vertical bars to indicate uncertainty in estimating the latent variables. In Figure ??, the dots represent the estimated factor scores and the lines indicate  $\pm 1SD$ . The red line is a “ghost line” (Fife 2020a), which simply repeats the line from the second panel ( $x_2$ ) to the other panels. This makes it easier to identify which indicators best load onto the latent variable (i.e., which indicator is most reliable). In Figure 7,  $x_2$  has the largest factor score.

Naturally, there is a great deal of overlapping points/lines in Figure 7. To minimize overlap, we can instead sample datapoints, as in Figure ??.

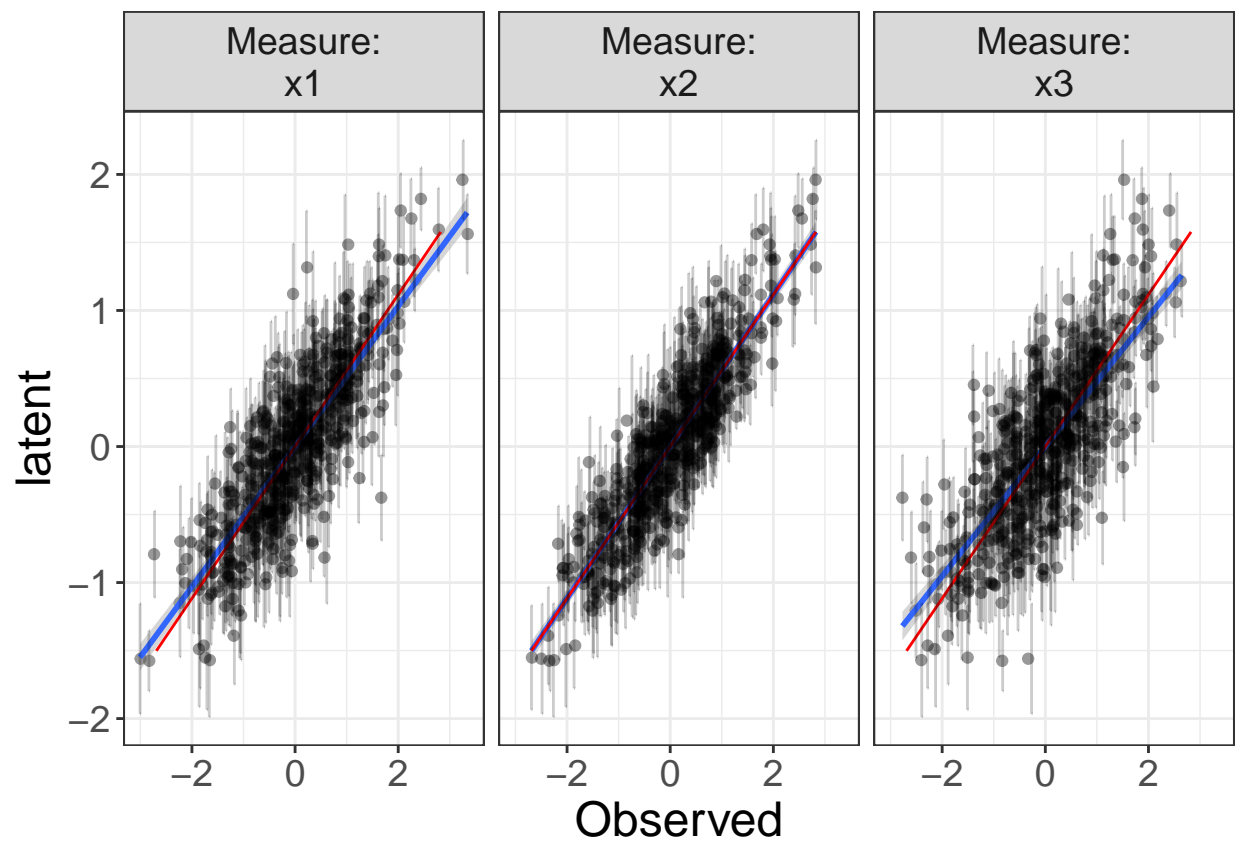
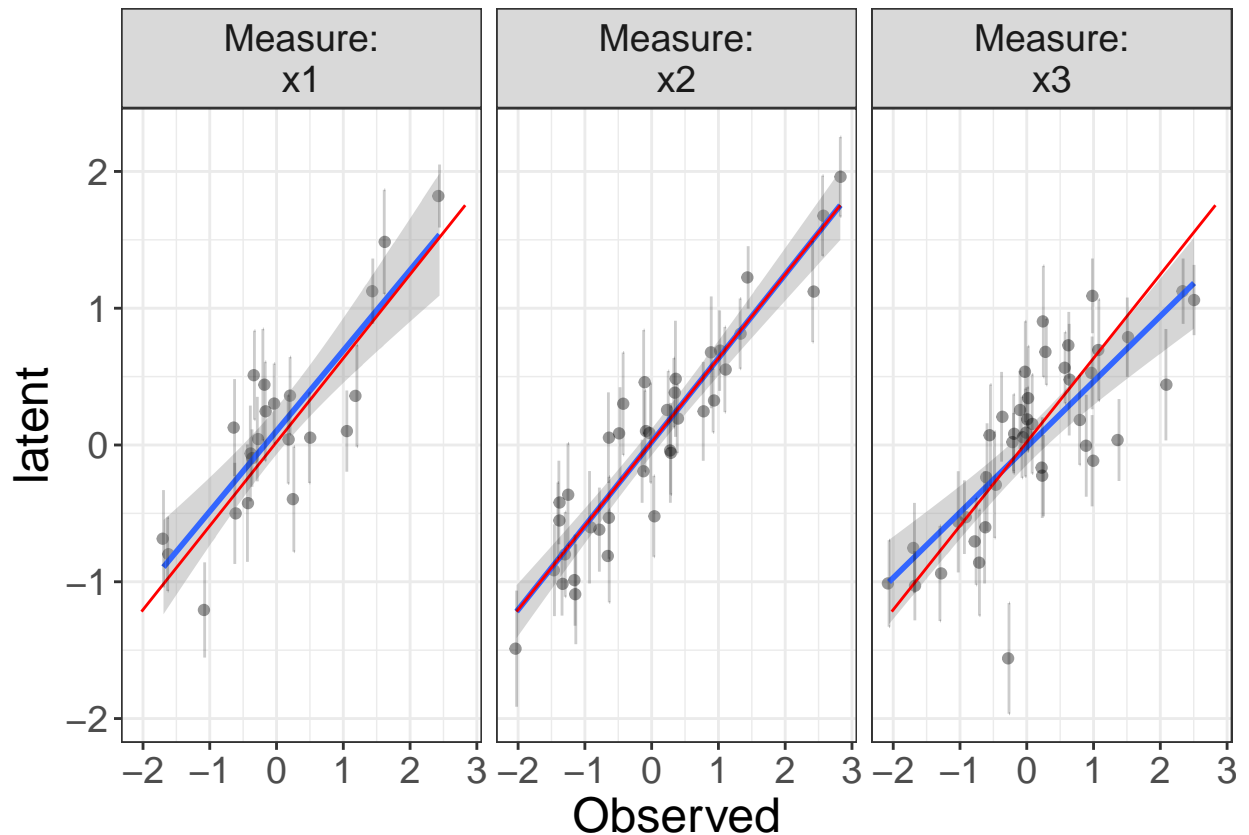


Figure 7: This image, called a measurement plot, shows the relationship between the latent variable (Y axis) and each of the standardized indicators ( $x_1$ ,  $x_2$ , and  $x_3$ ).



## Structural (Beech Ball) Plots

When modeling latent variables, often the visuals of interest are not the observed variables, but the latent variables. In other words, the measurement model is ancillary to the substantive model. Naturally, we might wish to visualize the relationship between the latent variables.

However, as before, we need to have visuals that reflect uncertainty in our estimates of the latent scores. With measurement plots, we needed only to reflect that uncertainty on the Y axis (because the Y axis displayed the latent variable while the X axes displayed the observed variables). When visualizing relationships between latent variables, on the other hand, both axes should reflect uncertainty. In **flexplavaan**, this uncertainty is represented as ellipses. The diameter of the ellipses (for both the X axis and the Y axis) are obtained from prediction intervals for **lavaan** objects or from posterior distributions for **blavaan** objects. Figure 8 shows these plots, which we call “Structural Plots,” or “Beech Ball Plots” (because the ellipses look like beech balls in various stages of compression).

Aside from the beech balls, there is a great deal of flexibility in how one visualizes the structural model. In our example, we only had two variables to visualize, so a simple bivariate plot was most natural. When more variables are included, we might utilize paneling, added variable plots, beeswarm plots, etc. For a review of the types of plots possible, see Fife (2020a).

## References

Bauer, Daniel J. 2005. “The role of nonlinear factor-to-indicator relationships in tests of measurement equivalence.” *Psychological Methods* 10 (3): 305–16. <https://doi.org/10.1037/1082-989X.10.3.305>.

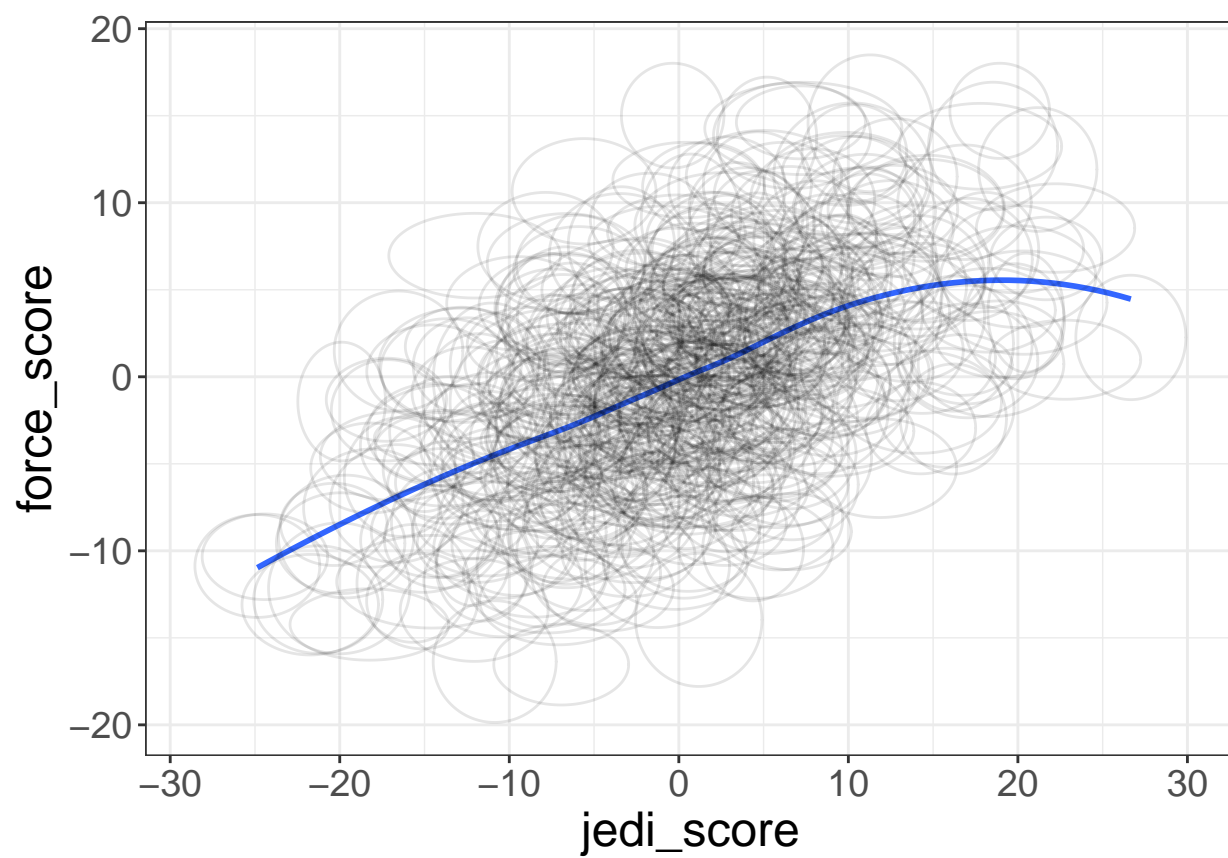


Figure 8: Structural or “Beech Ball” plot of the relationship between the latent variables Force and Jedi. The ellipses represent the prediction intervals for the factor scores of the latent variables.

- Bollen, Kenneth A. 1989. *Structural equations with latent variables*. Wiley Series in Probability and Mathematical Statistics. {Applied} Probability and Statistics Section; 0271-6356. John Wiley & Sons.
- Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2018. “Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015.” <https://doi.org/10.1038/s41562-018-0399-z>.
- Correll, Michael A. 2015. “Visual Statistics.” Doctoral Dissertation, University of Wisconsin-Madison.
- Fife, Dustin A. 2019. “A Graphic is Worth a Thousand Test Statistics: Mapping Visuals onto Common Analyses.” <http://rpubs.com/dustinfife/528244>.
- . 2020a. “Flexplot: Graphical-Based Data Analysis.” *PsyArxiv*. <https://doi.org/10.31234/osf.io/kh9c3>.
- . 2020b. “The Eight Steps of Data Analysis: A Graphical Framework to Promote Sound Statistical Analysis.” *Perspectives on Psychological Science* 15 (4): 1054–75. <https://doi.org/10.1177/1745691620917333>.
- Fife, Dustin A., and Joseph Lee Rodgers. 2019. “Exonerating EDA: Addressing the Replication Crisis By Expanding the EDA/CDA Continuum.” *Unpublished Manuscript*. <http://quantpsych.net/fife-exonerating-eda-draft-oct2019-df-edits/>.
- Hallgren, Kevin A., Connor J. McCabe, Kevin M. King, and David C. Atkins. 2019. “Beyond path diagrams: Enhancing applied structural equation modeling research through data visualization.” *Addictive Behaviors* 94 (March 2018): 74–82. <https://doi.org/10.1016/j.addbeh.2018.08.030>.
- Healy, Kieran, and James Moody. 2014. “Data Visualization in Sociology.” *Annual Review of Sociology* 40 (1): 105–28. <https://doi.org/10.1146/ANNUREV-SOC-071312-145551>.
- Ioannidis, John P. A. 2005. “Why Most Published Research Findings Are False.” *PLoS Medicine* 2 (8): e124. <https://doi.org/10.1371/journal.pmed.0020124>.
- Levine, Sheen S. 2018. “Show us your data: Connect the dots, improve science.” *Management and Organization Review* 14 (2): 433–37. <https://doi.org/10.1017/mor.2018.19>.
- Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. “The preregistration revolution.” *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.1708274114>.
- Pashler, Harold, and Eric-Jan Wagenmakers. 2012. “Editors’ Introduction to the Special Section on Replicability in Psychological Science.” *Perspectives on Psychological Science* 7 (6): 528–30. <https://doi.org/10.1177/1745691612465253>.
- Tay, Louis, Scott Parrigon, Qiming Huang, and James M LeBreton. 2016. “Graphical Descriptives: A Way to Improve Data Transparency and Methodological Rigor in Psychology.” *Perspectives on Psychological Science* 11 (5): 692–701. <https://doi.org/10.1177/1745691616663875>.
- Wilkinson, Leland, and Task Force on Statistical Inference. 1999. “Statistical Methods in Psychology Journals: Guidelines and Explanations.” *American Psychologist* 54 (8): 594–601.