Origins

* Dr. Pardo
  + new client
  + submission for conference
  + day of, notice a critical error
  + if I had visualized the data…
* Benefits of visualization
  + easily encode large amounts of information
    - Correll Table
  + highlights uncertainty
  + highlights modeling problems
  + easy to understand!
* First teaching graduate students
  + replication crisis was new
  + suggested there were problems with how we did analysis
  + students were not getting it
    - as researchers, they perpetuated misunderstandings
  + Developed “8 steps of data analysis”
    - show eight steps
    - over 1/3 utilize visualizations
    - Two “take-home messages”
      * Raw data should always be plotted
      * every analysis should have a graphic
    - Knew I’d have to tackle SEMs
      * but that’s difficult
  + To understand how to approach, let’s visit regular visualizations

Visualizing Linear Models

* most visualizations are essentially scatterplots
  + simple correlation/regression
    - predictor on x, outcome on y
    - overlay a line that represents the fit of the model
  + No different with categorical variables
    - except we have overlap (before and after jittering, then jitter based on density)
  + With paneling, we can visualize multivariate relationships
    - ancova
    - factorial anova
    - multiple regression
    - even mixed models
  + No matter the situation, we’re always doing the same thing
    - visualizing observed variables
    - overlaying the fit (as a line or symbol)

Visualizing SEMs?

* SEMs are problematic for two reasons
  + SEMs are generally not performed on raw data
    - rather, covariances or correlations
    - how do you plot a scatterplot of this?
  + no clear way to visualize these
    - they are *latent* variables!
    - they cannot be observed
  + these are LINEAR models
    - it’s awkward to even fit nonlinear models, let alone visualize them
* Three approaches mentioned in the literature
  + residual analysis
    - these assume the factor scores are actually meaningful
    - if the model fits poorly, who cares about the residuals?
  + visualizing the factors themselves
    - again, this assumes factor scores are meaningful
    - Essentially, it’s conflating the fit with the observed data
  + Bayesian models
    - so much more flexible
    - can have any form of link function and/or distribution
    - bonus: factor scores are estimated naturally

A Different Approach

* New approach:
  + same as before: scatterplot of raw data, overlay the fit of the model
  + treat observed data as if it’s observed data (these are the dots)
  + treat factor scores as if they’re the fit of the model
    - these are the lines
    - cuz that’s what they are!
    - details:
      * sequence of x scores, generate factor scores over entire range of x
  + latent variables are then like a regression line (show plot)
* Basic idea
  + scatterplot matrix of *items*
  + the “fit” is the predicted factor scores for each person
    - if the fit passes through the data, the model fits
* Other visual aids
  + “residual dependence plots”
    - show what they are in regression
    - you *want* nothing to be left after you fit the model
    - these plots flatten the line and make patterns more apparent
  + We can do the same with SEMs
    - we subtract the fit from the observed
    - there *should* be nothing left if we have a good model
* All this is easier said than done!

Complications

* When we estimate factor scores, they are on a different scale than the raw data
  + might have the same variance, but different means
    - show parallel lines
  + or they might have different variances, but the same means
    - show lines that intersect at the center
  + or they generally have different variances/means
    - show lines that are very different
  + simple solution
    - transform factor scores to have the same mean/variance as y axis variable
* Latent variables correct for measurement error
  + predicted lines are steeper than the actual lines
  + this really annoyed me
  + solution?
    - “uncorrect” for measurement error (\*sqrt(rxx, ryy))
    - mathy part
* New problem: what if the relationship is nonlinear?
  + nonlinear equations may have very complicated slopes
    - e.g., MME equation
    - handle on a case-by-case basis? No!
  + Use an approximation
    - math stuff

In summary…

* favor Bayesian models
  + nonlinear and naturally estimate factor scores
* visualize scatterplots between observed variables
* overlay the fitted factor scores
* weaken the relationship proportional to reliability

So….does it work?

* Three scenarios I will address
  + ideal situation
    - linear
    - model correctly specified
    - what SHOULD it look like?
  + Misspecification
    - linear
    - important crossloading missed
    - what SHOULD it look like?
  + Nonlinear and misspecified

Other visualization tools

* comparing two models
  + overlay two lines AND show difference for each individual in fits
* visualizing latent relationships