Comparing Models for Single-Cell Data

A Study in Flow Cytometry and RNA Sequencing

Lee Panter

University of Colorado Denver

Please don’t “sugarcoat” it, I’m a diabetic.

Do I Have Diabetes?

- As of 2015, 30.3 million Americans (approximately 9.4%) have diabetes
- 84.1 million are prediabetic (approximately 26.1%)
- CDC, National Diabetes Statistics Report, 2017[3]

So HOW Diabetic am I?

- Levels of HbA1c greater than or equal to 6.5% indicate diabetes
- The HbA1c test measures the average glycalated hemoglobin in the blood over the past two to three months.

So what's up doc?

- Questions:
  - How are measurements normalized?
  - How are measurements characterized as abnormal?
- Problems:
  - Each measurement is calculated as a comparison to another individual with different biological characteristics
  - Specifically Red Blood Cell (RBC) Lifespan
    - It is assumed that average RBC lifespan is an accurate estimate of cellular lifespan
    - But older RBCs have increased exposure to blood sugar
- Solutions?
  - Measure cell by cell!
  - Account for cellular age, and re-weight counts according to risk of glycolization
  - Use data sets that are more biologically similar to normalize measurements

What's your problem, buddy?

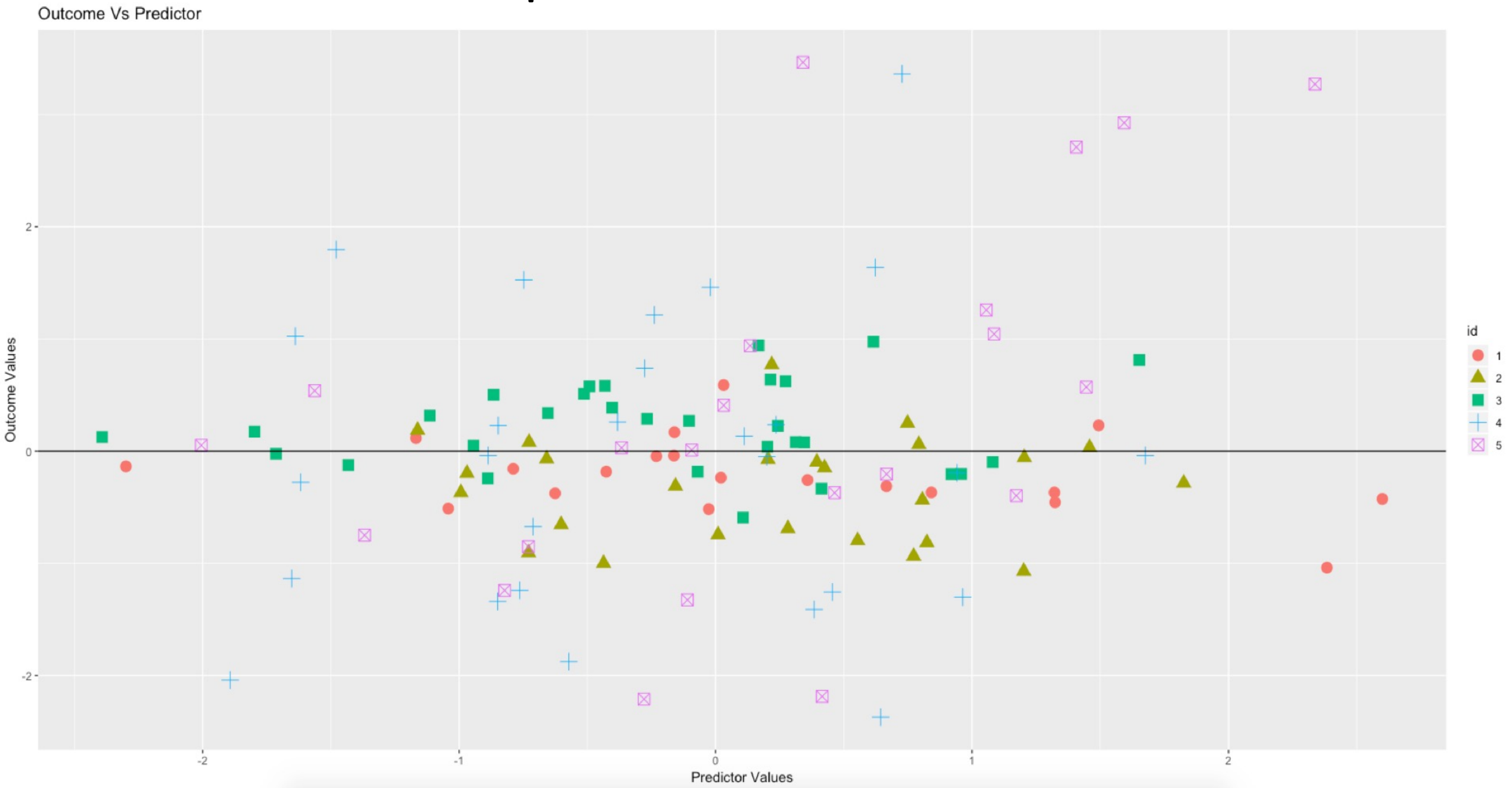
[1]

- **The single-cell data:**
  - 27 Subjects, 384 cells (max) each (observations)
  - Flow Cytometry measurements
    - Cluster of Differentiation (CD) markers for immunophenotyping
    - Forward and side scatter used for shape, size, and calibration (among others)
  - RNA Sequencing (scRNA-seq)
    - ~1 million paired-end reads per cell[2]
    - Reads are mapped to genes to form count data that quantify expression[4]
- **Problems under consideration:**
  - Can CD markers measured using Flow Cytometry be used to model scRNA-seq expression?
  - If so, what modeling methodologies work best, and how well do they work?
  - If not, what other information might be needed?

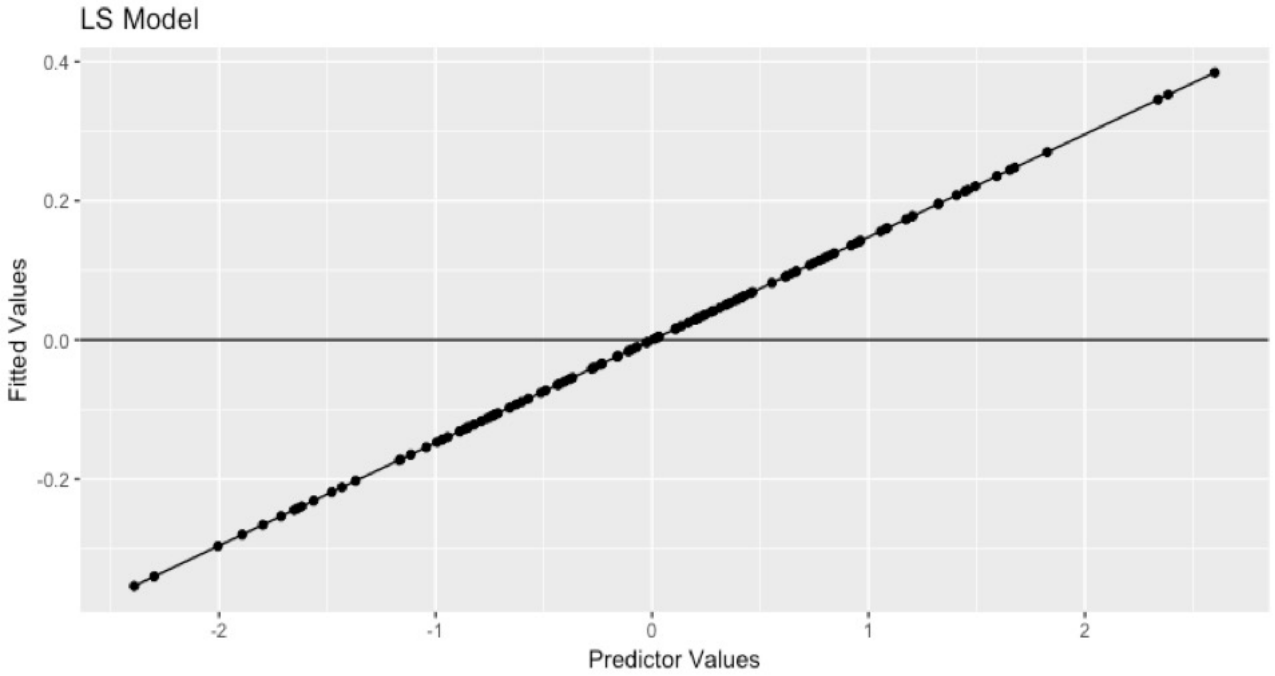
What are you waiting for?

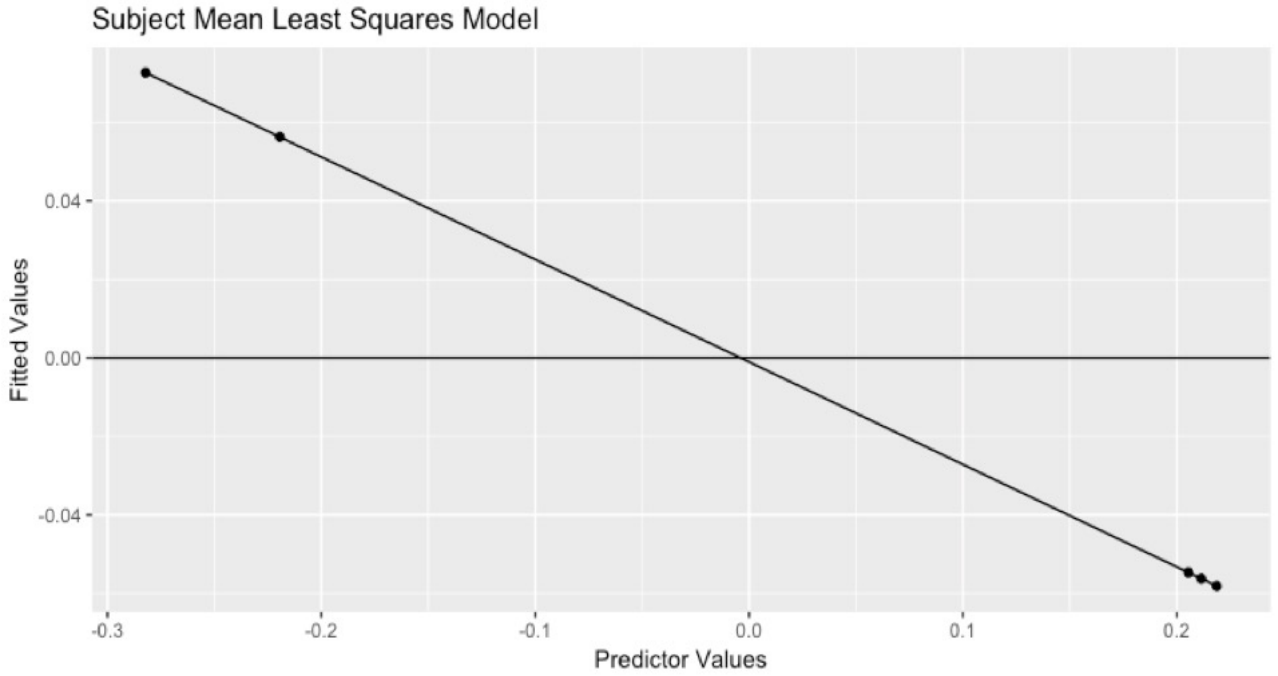
- Model scRNA-seq outcomes using *appropriate* cytometric CD marker measurements
- Modeling approaches:
  - Ordinary Least Squares
  - Regression on mean effects of each individual
  - Linear regression with fixed effects (main effects, interaction)
  - Linear mixed effects models (random individual and interactive effects)
  - Generalized Estimating Equations (main effects, interaction)
- Modeling Analysis
  - Nested Model Comparisons
  - Model Selection (if applicable)
  - Quantifying explanation
  - Diagnostics

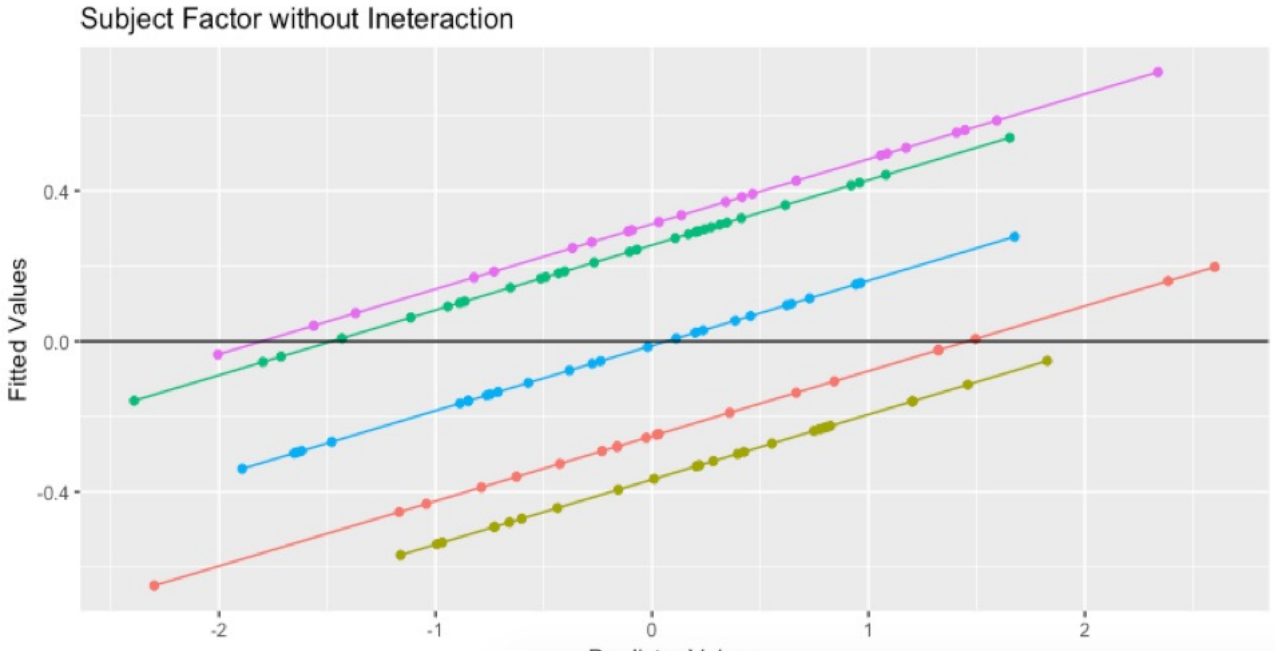
Example Data (that's not so exemplary)

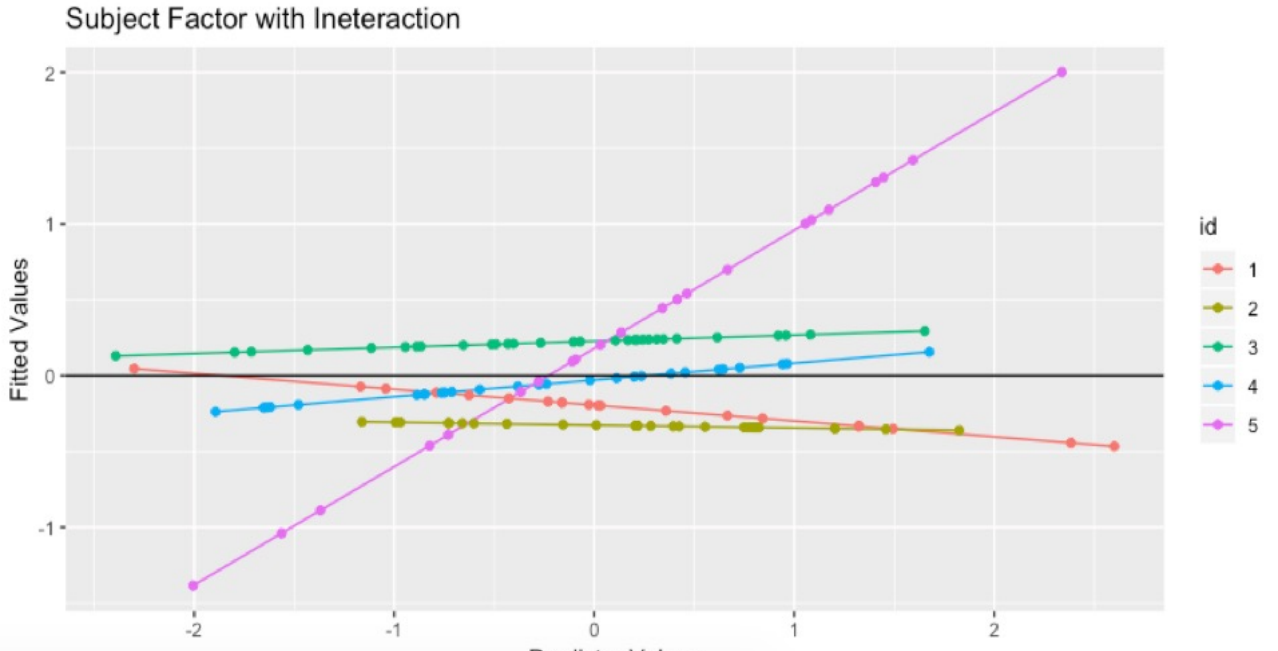


Least Squares Models

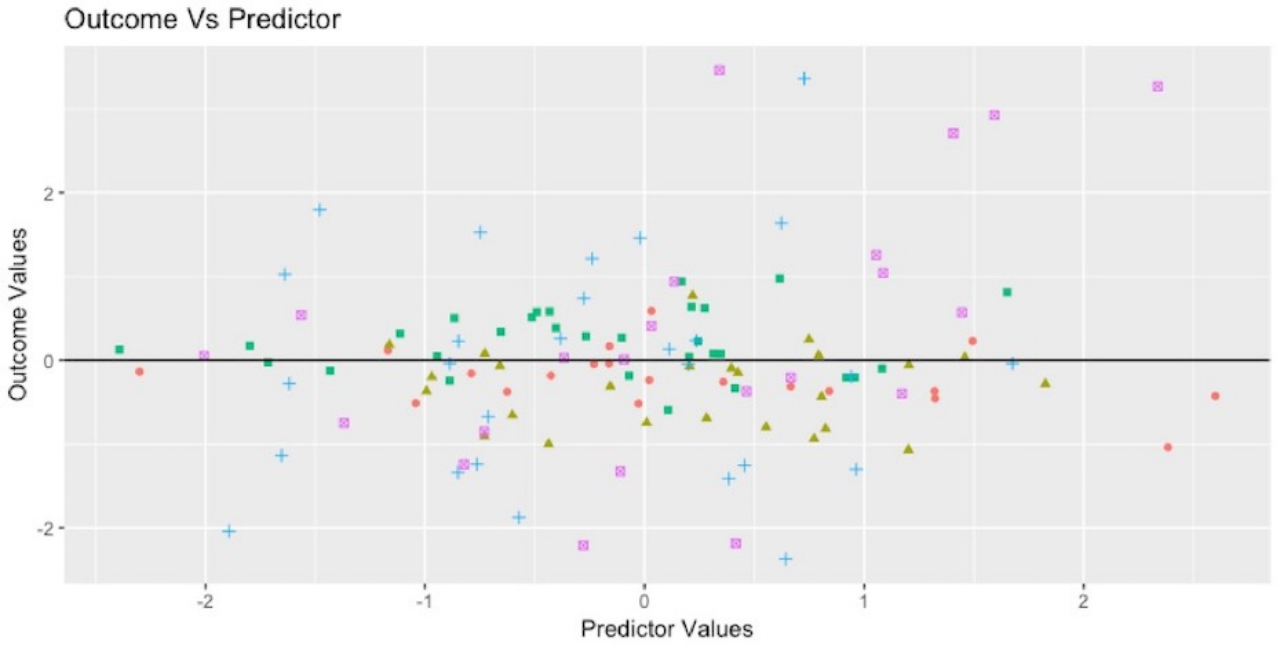


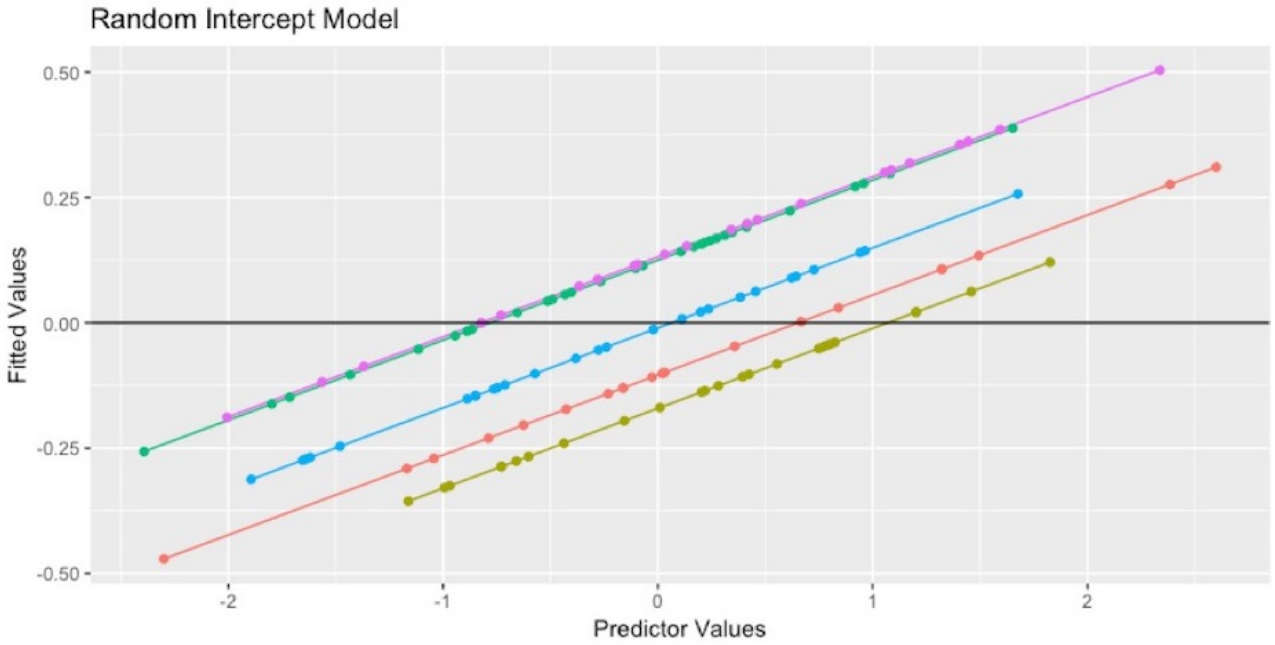


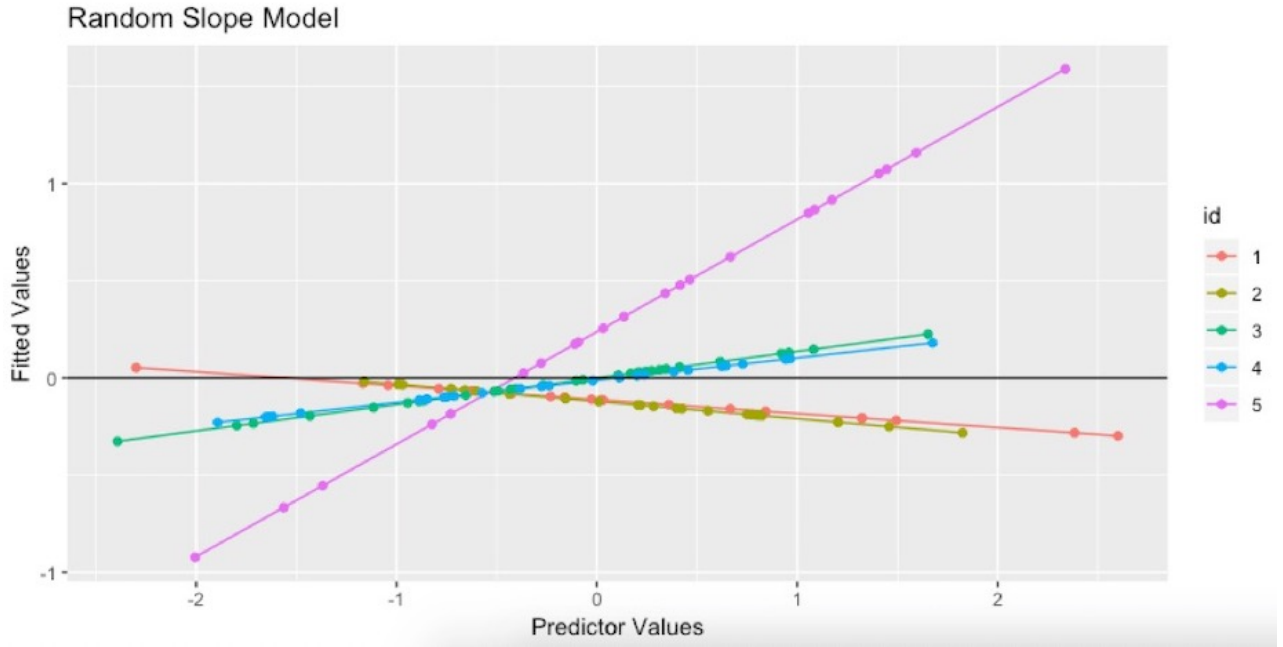




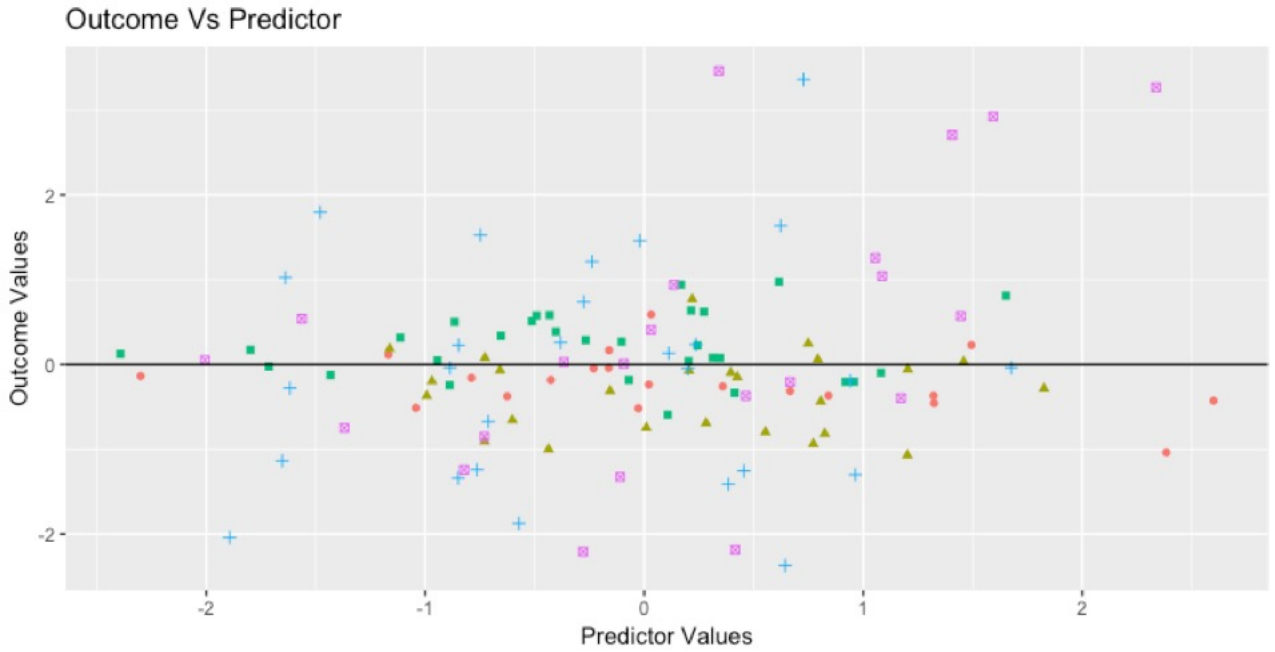
Linear Mixed Effects Models

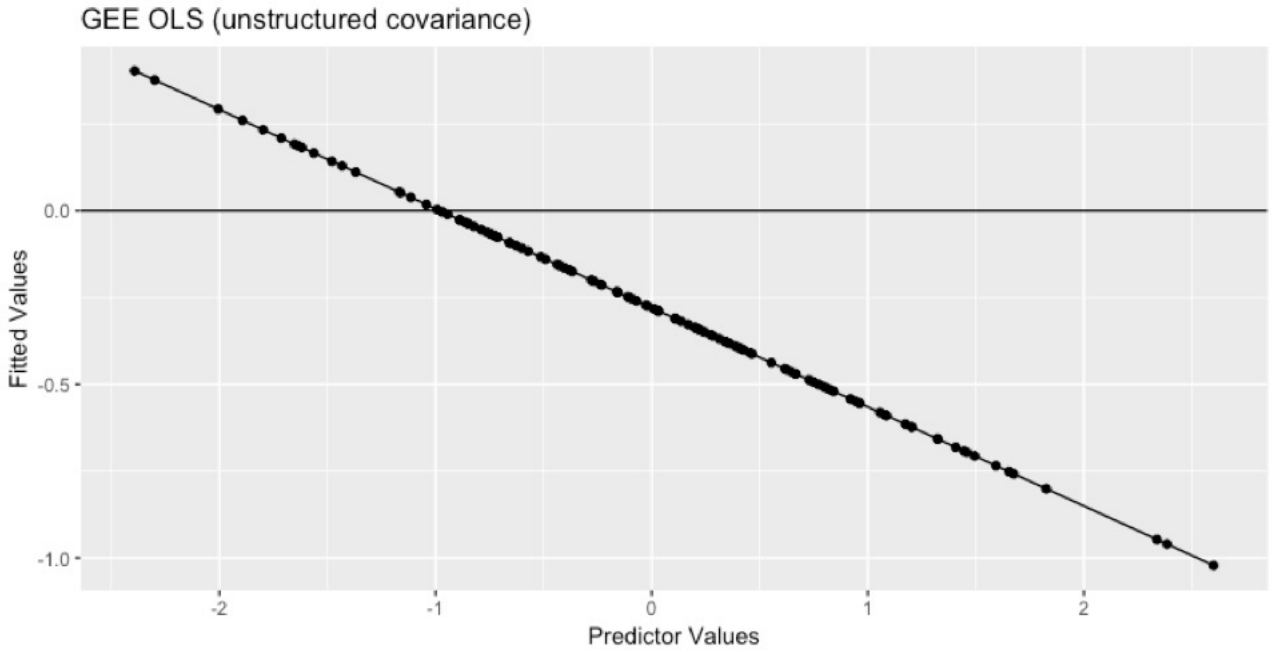


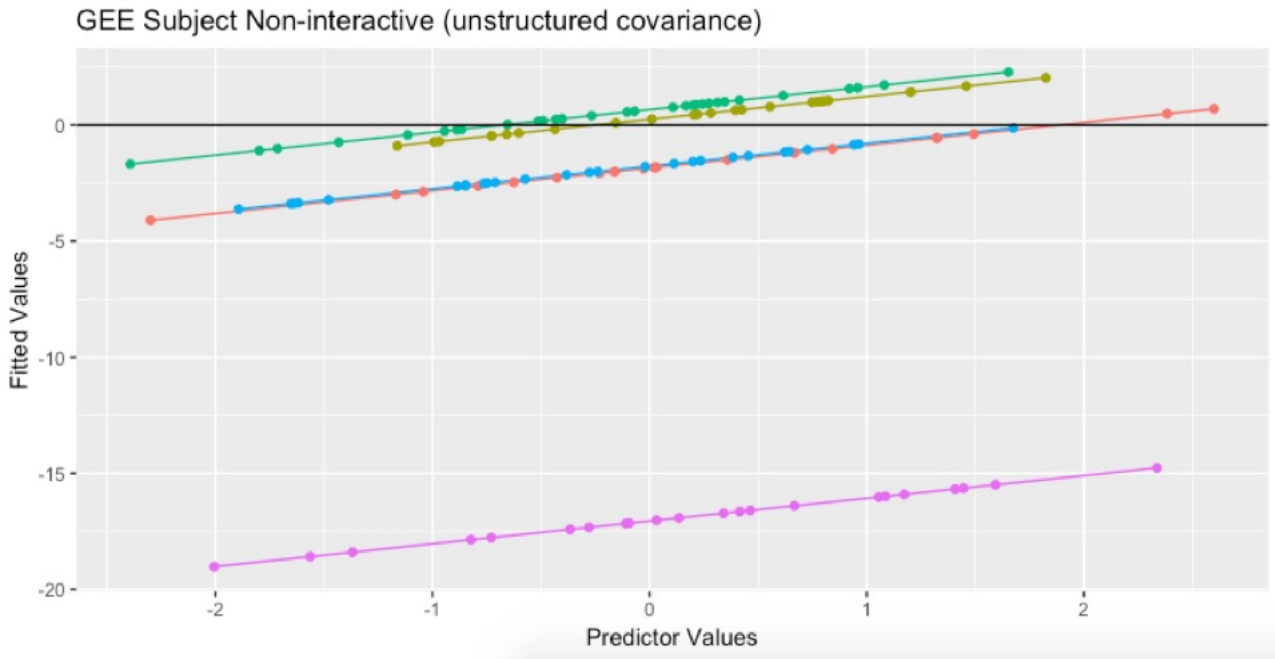


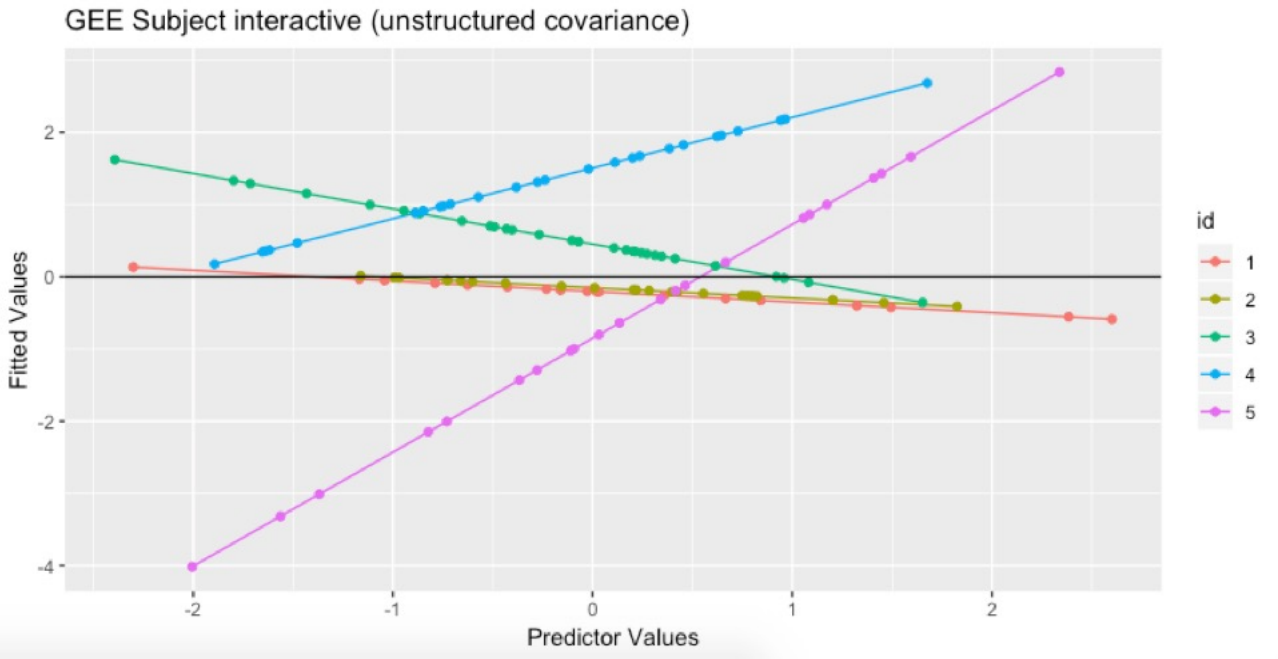


GEE Models


Outcome Vs Predictor
Outcome Values
Predictor Values
2
0
-2
-2
-1
0
1
2


GEE OLS (unstructured covariance)
Fitted Values
Predictor Values
0.0
-0.5
-1.0
-2
-1
0
1
2


GEE Subject Non-interactive (unstructured covariance)
Fitted Values
Predictor Values
0
-5
-10
-15
-20
-2
-1
0
1
2


GEE Subject interactive (unstructured covariance)
Fitted Values
Predictor Values
2
0
-2
-4
-2
-1
0
1
2
id
1
2
3
4
5

So, did you just *"forget"* the problem?

- What I've been doing:
  - Data quality control including:
    - Normalization of cell-specific biases
    - Corrections for batch-effects
    - Protocol & measurement biases
    - "Anchoring" for integration and comparison with other data[5]
  - Developing error and correlation structure
    - Error structure for GLM link functions
    - Correlation specified for GEE
- What's next
  - Initial models using updated information

OK, wrap it up!

- Questions, comments, concerns, and feedback
  - lee.panter@ucdenver.edu

- Special Thanks to:

Audrey E Hendricks, PhD
University of Colorado Denver
Department of Mathematical and Statistical Sciences

I don’t know, I just work here

References

[1] *Cellometer k2 fluorescent viability cell counter.*

[2] A. Arazi, D. A. Rao, C. C. Berthier, A. Davidson, Y. Liu, P. J. Hoover, A. Chicoine, T. M. Eisenhaure, A. H. Jonsson, S. Li, et al., *The immune cell landscape in kidneys of lupus nephritis patients*, bioRxiv, (2018), p. 363051.

[3] C. for Disease Control, Prevention, et al., *New cdc report: More than 100 million americans have diabetes or prediabetes. july 18, 2017*, 2017.

[4] A. T. Lun, D. J. McCarthy, and J. C. Marioni, *A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor*, F1000Research, 5 (2016).

[5] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, *Comprehensive integration of single-cell data*, Cell, (2019).