Homework 4: Miscellaneous and Project Pre-Study

Reno Malanga, Christian Arndt
Behavioral Data Science

April 19, 2025

Problem 1: True-False Questions

Mark all statements which are FALSE.

- Path coefficients in structural equation models (SEMs) are analogous to regression coefficients from ordinary least squares (OLS) regression.
- Latent variables in SEM can serve as both predictors and outcomes.
- Sample size does not affect the reliability of SEM estimates if the model is properly specified.
- An underidentified SEM model cannot produce meaningful results.
- Goodness-of-fit indices like RMSEA, CFI, and TLI can quantify how well the model reproduces the latent variables.
- Correlations among observed variables undermine the estimation of SEM parameters.
- The variance of a Wiener process with scale coefficient $\sigma = 1$ at time t is t^2.
- The standard Drift-Diffusion Model (DDM) assumes that evidence about a dominant alternative accumulates in discrete chunks over time.
- The Euler-Maruyama method can only be used to simulate linear stochastic differential equations (SDEs).
- The drift diffusion model has three core parameters: (relative) starting point, boundary separation, and drift rate.

Problem 2: Reflection

Describe and elaborate on three personal takeaways related to the topics of the course.

Reno:

1. One key takeaway from the course is the nature of behavioral experiments and how researchers have previously interpreted these results, which have caused problems in the social sciences and potentially led to the replication crisis. For example, looking back at the Big 5 dataset from earlier this semester, we have found that there do not seem to be five personality features, but rather two. Seeing this in person has been very impactful because my research will include behavioral data. This course has emphasized the need to be diligent in analyzing data so that I can effectively explain the results of my experiments.

2. Along with the replication crisis and how to properly analyze data, the other main takeaway was seeing the math behind all of the machine learning algorithms we looked at. I am not only referring to the forward and backpropogation, but also to the matrix operations and how data is oriented for it to be processed. Although it is more difficult to understand than the concept (at least for me), it was very interesting learning how everything comes together to create the models that we now use.

3. The third takeaway from this class that I got was learning how to implement git and the command terminal (more specifically, Anaconda Prompt) into my workflow. Before taking this class, I wanted nothing to do with git and the terminal as they were rather intimidating and looked like a nightmare to work with. I heard from many peers that git is a powerful tool and a desirable skill for those in the workforce, so I begrudgingly looked into them and was met with confusion and an intense lack in graphics (still not sure if the aesthetics were a problem for me). After the growing pains of seeing what git is and how it works, I started to better understand *what* it is, and it became more bearable. The Anaconda prompt is clearer and more welcoming than I initially thought, becoming a tool I feel that can be worked with in the future. Despite these two components not being something learned explicitly in class, I have worked with them on the side and in the background of my work. I found them to be essential in my workflow, and I am more open to working with them in the future. Without this class, I would not have had the hands-on experience that I needed to establish this connection and I am grateful.

Chris:

1. The first thing that comes to mind when I think about my takeaways from this class is a general dislike for traditional frequentist statistical methods.

   This feeling has been growing for a long time. It probably began when I took a classical statistics course and had trouble remembering the names and applications of all the various tests and metrics; I tend to have an easier time remembering something if it comes packaged with a principles-based motivation or a useful application, but many things from that class felt somewhat arbitrary, so they just didn't stick. I assumed for a while that my brain just

doesn't work well with those techniques, and figured I could just google anything I don't understand as I read academic literature.

These feelings grew during the Bayesian Data Analytics course I took last fall. That course showed me what a different statistics could look like. It was a little bit tricky to wrap my brain around initially, but after the first couple of a-ha moments, things quickly became intuitive to the point where memorization wasn't really required. I remember you saying early in the semester, when we were continually referring to p-values as probabilities, that people intuitively to think about statistics in terms of distributions. I don't know if that's true for everyone else, but I've found it particularly true for myself.

The final evolution of these feelings of disdain for traditional frequentist statistical methods came this semester, when I saw first-hand just how insufficient those methods are at extracting useful information from behavioral data. A large part of the problem comes from the relatively low signal-to-noise ratio that behavioral data seems plagued by, but I was also continually struck by how easy it was to get seemingly meaningful frequentist metrics out of what my models seemed to think was a solid nothingburger.

To synthesize these three things together: I find a large subset of classical frequentist statistics to be abstruse, which is a very undesirable trait for a branch of mathematics primarily concerned with presenting summary information. There are alternative methods for collecting summary information that are not nearly so abstruse. And finally, at least in fields with a low signal-to-noise ratio, these frequentist methods are not dependable, and can be gamed by a bad actor without much difficulty. Basically, they seem kind of bunk.

2. I mentioned this in the first point, but another takeaway I have from this class is an increased skepticism of basically any claim made by statistical distillations from behavioral data. Obviously we've all heard about the replicability crisis, but it hit different when I was the one trying (and mostly failing) to extract useful information out of a dataset. This is not to say that I think behavioral data science is futile; on the contrary, I think having well-qualified, responsible data scientists is all the more important to figure out exactly how much of what we're measuring is noise. But it seems like there are a lot of papers that make it to publication by standing on some seriously iffy claims, backed up by opaque methods, that very few people can really intuit about.

3. A final takeaway I had from this class was a realization of the importance of using generative performance as a metric for model quality. Again, this is something that everyone sort of knows, as it's discussed a lot. Still, something really clicked when I was the one creating the model. The moment I remember the most was comparing the error our linear regression autism predictor produced between the training data and the test data; even with a simple binary prediction, there was a pretty significant loss increase.

Problem 6 (3): Project Pre-Study

What problem are you considering? What is the type of data? What is the modeling task (e.g., regression, classification, latent variable modeling, parameter estimation...)? What existing models

have been applied to tackle the problem? What would be an adequate model in your case? What would be its strengths and limitations? How would you criticize the model(s) in your intended project? What is the metric of success?

We began by brainstorming what kinds of research topics are both:

- in the realm of behavioral science, and
- of some personal interest to us.

Chris has no experience in behavioral science, but Reno's main research interest is on stress. This gave us a pretty well-defined direction to explore, and we went on Kaggle to find datasets.

We narrowed our search down to a few final candidate datasets and had a meeting to discuss what kinds of questions they might allow us to explore. We ended up settling on a dataset containing information about the lifestyle and stress levels of 5000 teenagers. The lifestyle metrics contain self-reported data on things like sleep schedule, physical activity, and screen time, while the stress metrics contain both a self-reported field and also some data provided by a 'wearable device'. That dataset can be found here.

Our goal with this dataset is to explore the relationship between lifestyle and stress. This goal has two implicit subgoals:

1. We need to understand what kinds of lifestyles these teenagers are living.
2. Once we have an established understanding of the teenagers' lifestyles, we need to figure out how lifestyle and stress are related.

To develop our understanding of student lifestyle, our plan is to perform an exploratory factor analysis on the lifestyle features. The idea here is that lifestyle features which are highly correlated would group into latent variables representing different classes of teenager lifestyle.

The next step, exploring the relationship between lifestyle and stress, involves building some regressor models. The specifics here would likely depend on the results of our factor analysis. If there are no obvious lifestyle groupings, our best option would be to regress from the features in the original dataset. If, on the other hand, the lifestyle features are capable of explaining the majority of the information present in the dataset, we might ignore the original features and regress from the latent variables onto stress. The most likely scenario, though, is that we build regressor models from both sets of features onto stress to compare. We'll also try a number of different regression techniques, from linear regression through to deep neural networks, to see how much complexity our model needs to capture the relationship of interest.

Both factor analysis and regression models come with built-in metrics for model quality. For factor analysis, we can examine the loadings, the RMSEA, and the CFI and TLI. For the regression models, calculating the RMSE for a generative test set is the most obvious metric.

It seems likely—almost obvious, in fact—that there is a relationship between lifestyle and stress. Our main concern is about the quality of the data. The dataset provides a pretty paltry amount of information about the data sourcing. Much of the data is self-reported, and the documentation doesn't even describe what 'wearable device' was used to measure stress. Also, a quick visual inspection of the data shows that most of the fields are roughly uniform in distribution. This isn't an explicit problem, but we both expected the data to fall into a fairly obvious normal distribution. Still, we're hopeful; our chosen dataset has a high usability score. Also, in the discussion section, the author claims to have used AI to create it, which can only be good.