Making Predictions with SEMs

Timothy Jordan Regis

12/19/2022

Intro & Motivation

The purpose of this paper is to investigate the capabilities of using structural equation models for predictions of observed variables. For many years the general consensus was that SEMs were unsuitable for prediction estimation, and that the more common regression tools were far better when it comes to this task. However, recent literature has begun to challenge this notion as researchers have explored numerous methods to extract the necessary information to generate accurate predictions.

Within this literature is the recent article that this paper pulls most of its inspiration from. (Rooij et. al. 2022) developed a mathematical method for transforming the structure of an SEM to fit the familiar regression framework for out-of-sample predictions, and tested these results up against standard multiple linear regression across multiple datasets and simulations. This paper aims to take this work further by reproducing similar results across simulated data, and examining additional model frameworks not seen in the paper to test more of the method's limits.

Prediction methods are often at the forefront of any data analysis as they determine how the created models generalize and perform on unseen data. This is undoubtedly helpful in any field as it can allow researchers to put their findings into action, whether it be through a new drug or medical procedure, thus having stronger and more accurate predictions is a primary goal in these studies. This is where SEM comes in as we hypothesize that it in fact can produce predictions and even score higher in comparison to multiple linear regression. The ability to use structural equation models for prediction can provide a significant benefit as it allows us to, in addition to the main tools of prediction through linear regression, incorporate any assumptions we make about the covariances between the observed variables in the data, in turn allowing for stronger and more consistent predictions in theory. For example, if we have some prior knowledge or a clear understanding of what variables are likely to share a correlation, and which variables are likely to be independent, using an SEM lets us set these exactly, rather than in simple linear regression where we must rely on the model to realise this on its own. Ultimately, this difference can allow SEMs to develop a better estimate of the causal relationships through these direct and indirect effect incorporations. Furthermore, within structural equation models, we allow for the incorporation of the effects of measurement error on all of our variables, whereas in regular linear models, this measurement error can cause serious problems. By ignoring this error in linear regression, we can end up with misleading results in our regression parameter estimates and oversized type 1 error probabilities, which ultimately hurt its predictive power. In turn, we believe that the use of structural equation models will benefit our results again here by taking these factors into account first.

Methods and Model

For the purposes of this paper, we will be using the function created in (Rooij et al 2022), predicty.lavaan(), to test our SEM's predictive performance. To begin, we will briefly explain the method that was developed.

Firstly, given the nature of a structural equation model, the variables are assumed to follow a joint multivariate normal distribution. This method makes use of the mean vector and covariance matrix of the observed variables from this distribution as computed by the SEM, partitioning them between endogenous and exogenous variables for the regression model as follows:

$$\mu = \begin{bmatrix} \mu_x & \mu_y \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_{yy} \end{bmatrix}$$

The goal of the method is to obtain an estimate for θ the parameter vector of means, factor loadings, and regression weights, and this is done through maximum likelihood estimation through the typical minimization of a loss function: $\hat{\theta} = argmin_\theta \mathcal{L}(\mathbf{X}, \mathbf{Y}, \theta)$. The authors note that as SEMs estimate the joint distribution of the endogenous and exogenous variables, they mimic the familiar methods of generative models, in which predictions can be generated through the predictive distribution: $p(y|x_0)$. To generate their predictions from the structural model, the authors estimate the predictive distribution of $\hat{\theta}$ with mean $\mu_{y|x_0}(\hat{\theta})$ and $\Sigma_{y|x_0}(\hat{\theta})$ through the following functions:

$$\begin{aligned} \hat{\mu}_{y|x_0} &= \hat{\mu}_y + \hat{\Sigma}_{xy}^T \hat{\Sigma}_{xx}^{-1}(x_0 - \hat{\mu}_x) \\ \Sigma_{y|x_0} &= \hat{\Sigma}_{yy} - \hat{\Sigma}_{xy}^T \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xy} \end{aligned}$$

They then propose the use of the mean of the predictive distribution, $\mu_{y|x_0}(\hat{\theta})$ as a way of obtaining prediction estimates of $\hat{y}$, and transform this function to match the familiar regression format, $y = \alpha + \beta x$ as follows:

$$\begin{aligned} \hat{y} &= \mu_{y|x_0}(\hat{\theta}) \\ \hat{y} &= \hat{\mu}_y + \hat{\Sigma}_{xy}^T \hat{\Sigma}_{xx}^{-1}(x_0 - \hat{\mu}_x) \\ \hat{y} &= \hat{\mu}_y + \hat{\Gamma}(x_0 - \hat{\mu}_x) \\ \hat{y} &= \hat{\mu}_y - \hat{\Gamma}\hat{\mu}_x + \hat{\Gamma}x_0 \\ \hat{y} &= \hat{\alpha} + \hat{\Gamma}x_0 \\ Where: \hat{\alpha} &= \hat{\mu}_y - \hat{\Gamma}\hat{\mu}_x \text{ and } \hat{\Gamma} = \hat{\Sigma}_{xy}^T \hat{\Sigma}_{xx}^{-1} \end{aligned}$$

Where $\hat{\alpha}$ is the vector of intercepts of each response estimate, and $\hat{\Gamma}$ is the matrix of estimated regression weight vectors.

As we can see, this very closely mirrors the familiar simple linear regression model, when comparing the frameworks, we can see that in the MLE of the estimates we normally expect from linear regression there is an almost exact match to $\hat{\alpha}$ and $\hat{\Gamma}$ seen here in terms of contents, but the methods of estimation behind these values are where the difference lies. Importantly, in structural equation modelling, we can apply assumptions

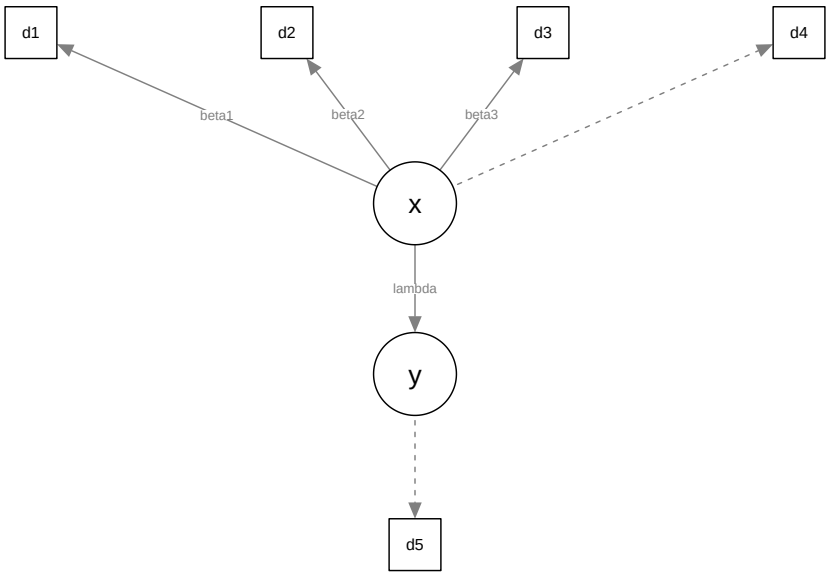
to the covariance matrix of the observed variables and make restrictions on it, in turn altering the makeup of the parameters of the regression equation, diverging it from the ordinary linear regression framework, which the authors believe can potentially improve prediction estimates.

First Simulations

To begin with testing the method, we start with a very simple model with few latent variables and a lot of structure and symmetry in its parameters.

Model 1:

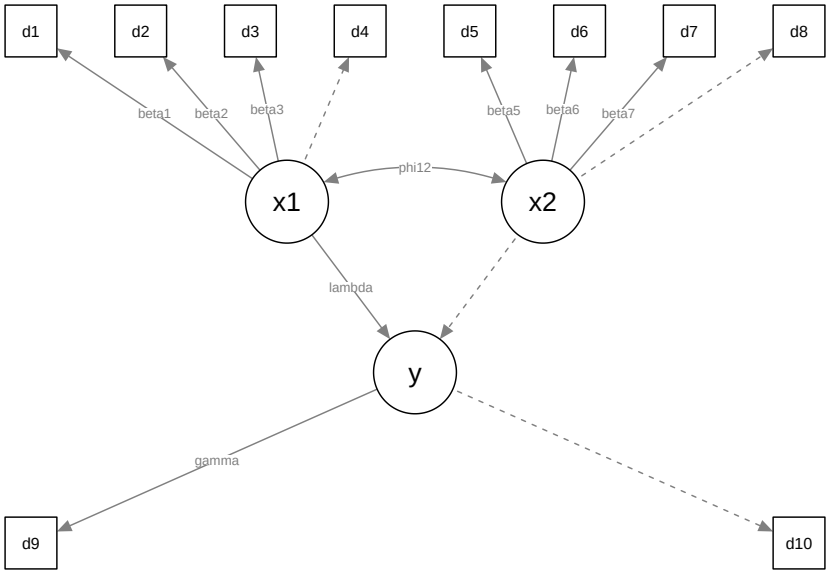
The first model is defined by the following path diagram:



Where we have 4 observed variables and 1 latent, with the intention of predicting 'd5' from the values of d1 to d4 with our model.

Model 2:

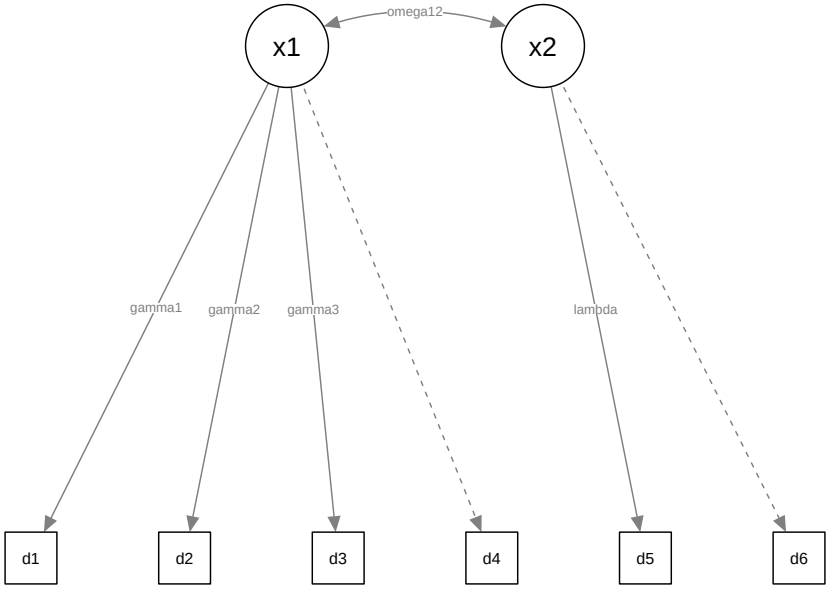
Then, adding another layer of complexity through an additional exogenous latent variable and covariance structure, we estimate the following model:



Where we now attempt to predict d10 and d9 from d1 through d8.

Model 3:

And finally, we will test a model where there is no formal regression between the latent variables, and only a covariance to rely on, as follows:



Where we aim to predict d5 and d6 from d1 to d4. This is an especially interesting case as I believe the lack of a formal regression-like link between x1 and x2 will be quite problematic for linear regression.

Firstly, before testing any models, we must note that if we choose to specify and estimate a full covariance matrix between all observable predictor variables into our SEM and data, without any restrictions, we find that the proposed SEM method produces the exact same results to linear regression. This makes sense given the styles of the two prediction methods, with a full covariance matrix, the SEM prediction matches the typical linear regression framework exactly. The authors also note this as a result, explaining that this can in turn lead to predictions with a higher bias, but lower variance than simple linear regression modelling, due to the calculation of MSEs. However, as mentioned, the benefit to SEMs is that we can restrict this covariance matrix, to be pretty much anything we'd like, thus turning it away from the same framework as simple linear regression.

To make things easy to start, we begin with the simplest parameters, with multiple factor loadings set equal to each other, and assuming a zero covariance between the predictor variables. We will then assess how the model can perform in comparison to linear regression on the same data including all parameters. To compare between the linear model and the SEM, we will be measuring each model's mean squared deviation of its fitted response values and the true values, and we will simulate this over 100 trials, displaying the results in the next section.

Results

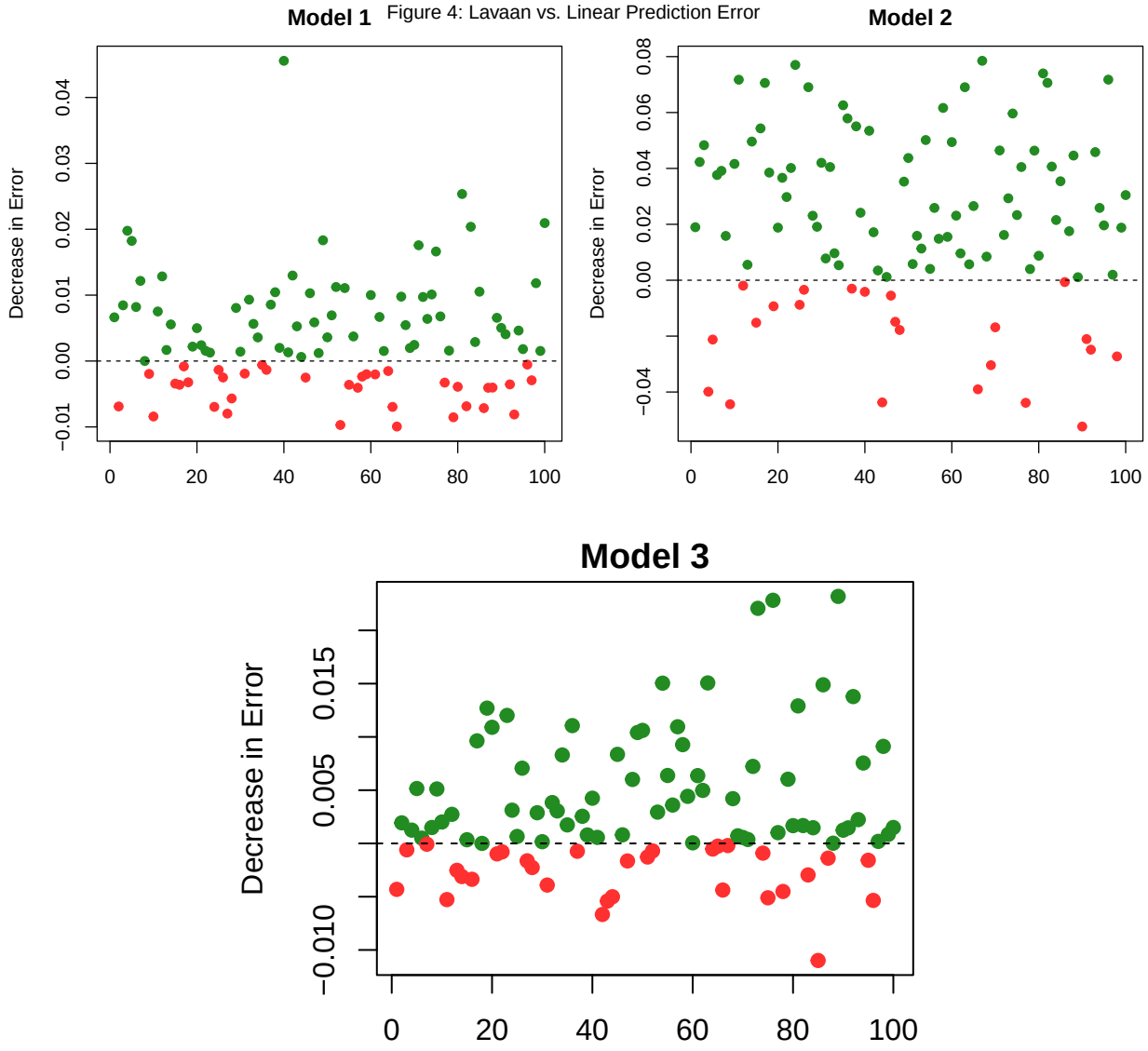


Figure 4: Lavaan vs. Linear Prediction Error

Table 1: Simulation Set 1 Summary

Model	Wins	Avg_Lavaan_Error	Avg_Linear_Error	Avg_Decrease
1	64	1.022	1.026	0.004
2	77	2.626	2.647	0.021
3	69	0.819	0.822	0.003

To compare the results of each prediction method, we compute the number of times the SEM method scored a lower mean squared prediction error than the linear method, as well as the average error in each model and

the change in prediction error seen in **Table 1**. From the preceding plots in **Figure 4**, we can also get a look at the raw differences between the SEM and linear models on the simplest parameters after 100 simulations.

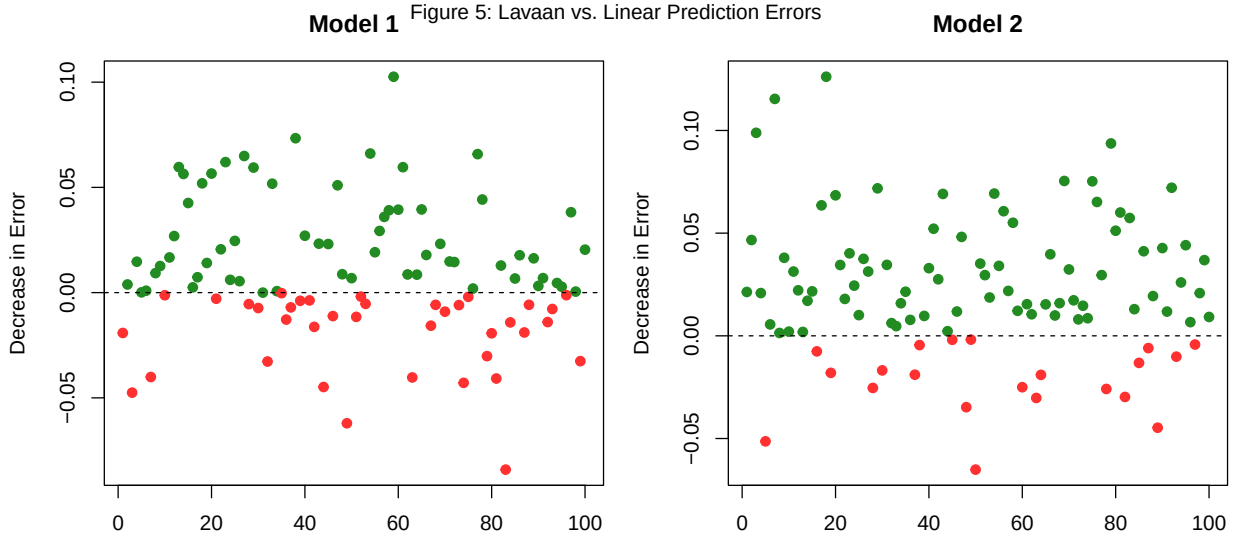
In model 1, we see a total number of wins of *64* out of 100 for the SEM model, with an average decrease in mean squared error of *0.004*. In model 2, as compared to the linear model, the SEM scored a lower prediction error in *77* out of the 100 tests ran, seeing a decrease on average of 0.021. And finally, in model 3, the SEM outperformed the linear model again in *69* out of the 100 trials, with an average decrease of *0.003*. As we can see, with a relatively small magnitude, the SEM method of predictions performed significantly better than the linear model, resulting in lower mean squared prediction errors in a majority of the 100 runs, across all three models. Furthermore, of the wins and losses seen we also tend to see greater magnitudes of decrease on wins than we do on losses. This suggests to us that there is some strength behind using structural equation models for predictions and that they can potentially be even more accurate than linear regression models as opposed to what we have believed in the recent past.

Second Simulations

Now to apply further testing to the method, we will use the same simulation parameters as before, but add a bit more complexity into our data and model through covariances and the restrictions we put on the SEM. This can potentially provide an even larger benefit to the SEM method of prediction as we can attempt to mimic the covariance matrix of the data with the structure within the lavaan framework, restricting certain values and setting the necessary constraints, unlike in our linear regression method where we will be using the exact same model. The path diagrams from each model are included at the end of the report.

Results

Figure 5: Lavaan vs. Linear Prediction Errors



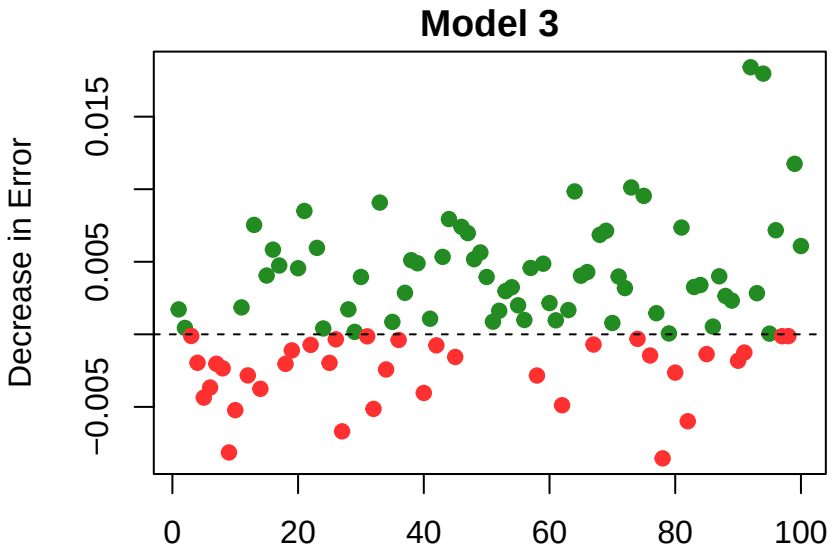


Table 2: Simulation Set 2 Summary

Model	Wins	Avg_Lavaan_Error	Avg_Linear_Error	Avg_Decrease
1	62	4.467	4.476	0.009
2	79	2.685	2.708	0.022
3	64	0.827	0.829	0.002

Now, with the increased complexity and restrictions placed on the covariance matrix and factor loadings of the variables, we can again see that the SEM-based prediction method does a better job of estimating the outcome of new data as compared to rudimentary multiple linear regression across all models in **Table 2**. Here we are using the same criteria of mean squared prediction error to classify model performance. Following a very similar trend, firstly, in model 1, the SEM prediction method beat the linear prediction in *62* cases, with an average decrease in mean squared prediction error of *0.009*. Next, in model 2, the SEM method outperformed the linear model in *79* out of the 100 tests, with an average decrease in prediction error of *0.021*. Lastly, in model 3, we see the SEM scoring higher than the prediction method in *64* trials, with an average decrease in error of *0.002*. This result again provides us with significant evidence that structural equation models can in fact be used for predictions, and more importantly, that they can perform better than simple linear regression.

Discussion

As we saw in the results of these methods, predictions are not only possible when using structural equation models, they even have the capability to significantly outperform simple multiple linear regression models.

Within the results, we saw a fairly similar trend across the datasets without any covariances on the observed variables as well as the ones with multiple covariances specified. We found the strongest results in model 2, where we had two latent exogenous variables impacting y, as we saw the largest average decreases in error with the SEM method, and greatest number of wins. While it is difficult to dissect exactly why this is the case, we believe it is likely due to the structure of the covariances between the observed exogenous variables, while in linear regression it will predict a covariance matrix between all 8 of these, our data was designed so that the variables from x1 and x2 are independent. This in turn adds many constraints onto the data that can only be captured if properly specified in a structural equation model using lavaan. The next most significant results tended to appear from model 1, where we had a very simple latent structure with only one latent exogenous variable impacting y. The SEM method here had a much smaller average decrease in error than in model 2, and taking our earlier reasoning, we believe this occurs because the SEM structure is much more similar to the linear model as there are fewer covariances we can put restrictions on. Finally, and somewhat as expected, we found the least significant results most commonly coming from the third model, where there was no formal regression between the latent variables, and only a covariance for the data to be linked through. Here, even when taking a summary of the linear model we found it struggled to produce results with significant coefficients which was representative of the true lack of a connection between the variables, which we believe to be the primary cause of the lack of a significant difference seen between the models. While the improvements seen were not massive, they do point in a direction which ultimately suggests that SEMs are in fact quite useful for predictions, as was the goal of this report.

Limitations and Future Work

While our results are promising, there are many areas at which we were limited and where future work can take over. Firstly, a key area of prediction methods is the ability to create prediction intervals for your estimates. We believe that this may be possible under an SEM framework, but we have been unable to prove so. Current work has resulted in standard error estimates that are all almost 0, and when using the typical method for calculation $\hat{y} \pm t_{1-\alpha/2} * \sqrt{MSE(1 + x_0(X^T X)^{-1} x_0^T)}$, we find extremely wide intervals. We have included this current work in the appendix, but future work could greatly benefit in solving this issue as it would allow for even better precision for our predictions, and bring it into better competition with the current methods.

Another highly important factor for this method is the use of real data. While it is good to have a model that works on the data designed for it, the true test of this model's ability would come through its application to real data, where parameters are unclear or completely unknown, as they will push the limits of how accurate this method can be. In turn, work in these areas can rapidly expand the possibilities in this relatively new field, and eventually present structural equation models as a suggestible alternative to many other prediction methods.

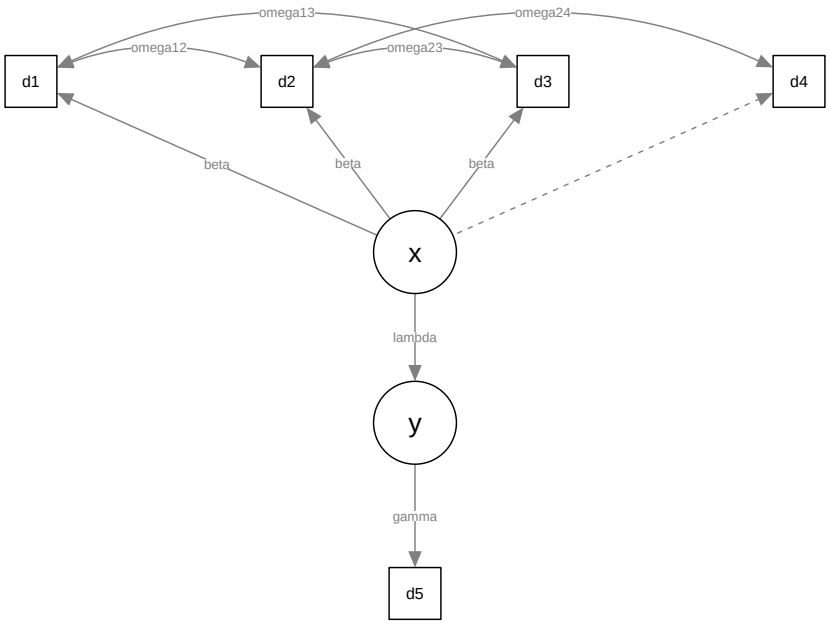
References

Mark de Rooij, Julian D. Karch, Marjolein Fokkema, Zsuzsa Bakk, Bunga Citra Pratiwi & Henk Kelderman (2022): SEM-Based Out-of-Sample Predictions, Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2022.2061494

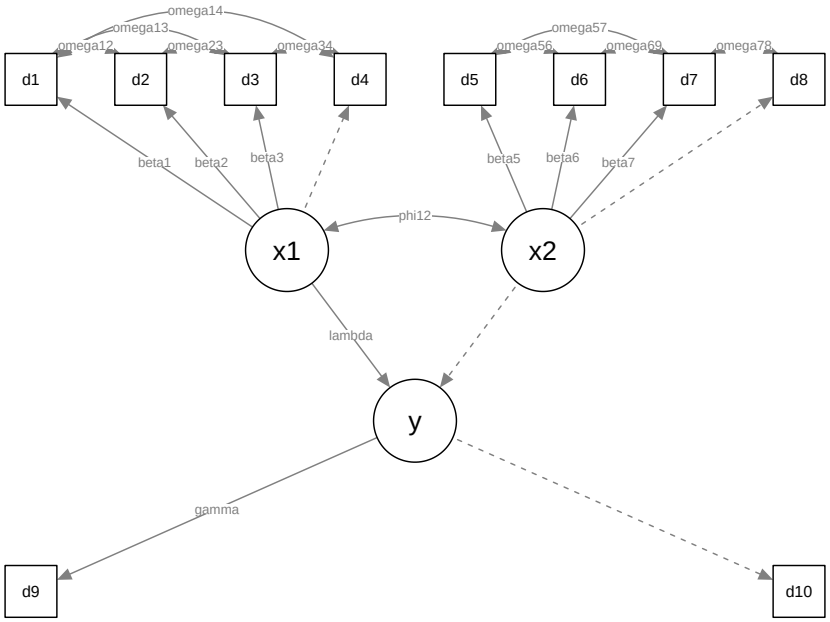
Jerry Brunner (2022): Structural Equation Models: An Open Textbook Edition 0.10: Department of Statistical Sciences, University of Toronto, https://www.utstat.toronto.edu/~brunner/2053f22/textbook/OpenSEM_0.10L-1.pdf

Second Simulation Path Models

Model 1



Model 2



Model 3