Leveraging Methods in Partial Least Squares Structural Equation Modeling to Analyze Seattle Housing Data

Theory of Statistics I

EN.625.725.FA22

Victoria Rose

1 Executive Summary

Partial Least Squares Structural Equation Modeling (PLS SEM) methods allow for many variables and their complex interactions to be described as a series of simple regression relationships. For this reason, this methodology is very popular in many different domains for its ability to give insight into multiple interactions between many variables.

However, sometimes the relationships between variables is not already known and different models must be studied for their ability to fit and describe available data. In this case, metrics must be used to evaluate the ability of each of the models' performance. Which evaluation statistic is best, though? Metrics such as R^2 and Goodness of Fit are popular amongst other forms of analysis and are commonly used in PLS SEM model evaluation. The authors of *PLS-Based Model Selection: The Role of Alternative Explanations in Information Systems Research* proposed an alternative suite of evaluation statistics to be used to compare the relative performance of these models [9].

In this paper, much of their methodology was reproduced and applied to housing data from the Seattle-area to examine pricing models. To do this, a subset of real housing data was selected and used to generate synthetic data that all models (shown in Figure 2) then were fitted to. Models were then assessed by a variety of statistics–AIC, AICc, BCI, goodness of fit, R^2, and adjusted R^2– which represented a variety of parameter types, including asymptotically efficient and consistent methods. For each iteration of simulated data and parameter adjustment, the model that performed best by each statistical method was noted and recorded in a dataframe. After 100 iterations of each of the 120 parameter conditions, the best performing models by each criteria were noted by two cases: one where the data generating model was included and one where it was not. The results of this simulation are recorded in 3 and 4, respectively.

From these findings, the proposed statistics of AIC and BIC, which were proposed by [9] as improved methods of PLS SEM performance evaluation, were identified as the most successful methods of conservative model evaluation.

2 Project Description

2.1 Overview

Modeling phenomena is an integral part to many different domains of research. Many different models can seek to describe the same phenomena at different resolutions and in different contexts. Sometimes, it is important to consider these alternative models at the same time and to objectively compare them. This can be a part of a scientific investigation to develop new mathematical descriptions entirely or to improve on established models. For the purpose of scientific rigor, it is important to have methodological ways to evaluate the performance of these models. For some subject domains and model types there are very well established and accepted methods of evaluating and comparing models; however, this paper will focus on the evaluation and comparison of Partial Least Squares Structural Equation Modeling (PLS SEM) methods.

PLS models are often used in fields such as social sciences, chemistry, economics, and neuroscience where samples might be limited in quantity but have a very large number of features [4]. PLS SEM allows for the structural modeling of latent variables. This is of particular use in fields such as psychology where variables like intelligence, personality type, or diagnoses are used to clinically describe individuals inner mental states [1]. Additionally, PLS SEM models can be effective in instances where convergence using covariance methods is of concern [3]. One limitation of PLS SEM, however, is that sample variance can introduce notable bias to the model as these models are composites of OLS regressions [8].

Current methodology for comparing PLS SEM models can include metrics such as R^2. However, this introduces a bias towards increasing model complexity as often R^2 can be increased through additional terms [9]. This is undesirable because often times very complex models can overfit to training data and have poor generalizability [7]. For this reason, [9] proposes using alternative statistics to compare the performance of these models

The authors of [9] focused on the domain of information systems; however, the importance

of evaluating models strategically is not limited to this domain. Establishing a rigorous method for selecting a model allows for systematic review of different hypotheses in the wide range of domains where PLS models are used. This project examines the evaluation methods proposed in [9] to evaluate several structural models created from this housing data in largely waterfront King County, Washington (locations of the 21,613 listings are plotted on the provided map).

As noted earlier, R^2, while a frequently used performance metric, often increases with model complexity, something that is not always desirable when developing a generalizable model. Alternatively, [9] states that "[t]he main advantage of the AIC-type criteria is that they can be used to measure the *relative* distances of competing models from the unknown true model, even when the absolute distance to the true model is unknown." For this reason, this paper will explore how these two variables and ones like them assess different exemplar models of varying complexity and accuracy to the sample problem space.

Figure 1: Map of raw data housing locations

[9] examines a series of UTAUT (Unified Theory Of Acceptance And Use Of Technology) models [1] of varying complexity. These structural models included both exogenic and endogenic latent variables. The research team, [9], examined the ability of the selection statistics to assess the performance of known correct and incorrect models of varying complexity with a wide range of simulated samples.

In order to apply this theory beyond the domain of information systems, I examined how effective this methodology is at identifying the "true" house list pricing model based off of variables known about the house. Using the real King County, Washington, housing data,

[1] UTAUT models seek to describe the relationship between user technology usage and user intention [6] within the field of information systems

key population parameters were selected using least squares regression. This was used to design a model that generated several thousand synthetic sample points through the iteration and adjustment of several parameters. This follows directly from the methods employed by [9].

The performance of the seven models, as outlined in Figure 1 of [9], was evaluated by a suite of statistics. Performance statistics fell into three categories: asymptotically efficient, asymptotically consistent, and field-standard PLS model evaluation. From these results, the different categories of statistics will be examined for how effective they were at selecting the most correct available models within both of the cases (Case I: the data-generating model structure (model 5) is included, Case II: the data-generating model structure is not included). These results were then compared against those from the source paper and divergences were examined.

2.2 Implementation

Details on code and data availability can be found in the appendix

2.2.1 Environment

This project was designed in python 3.9.6 and leveraged pandas, numpy, sklearn, and the `semopy` library for model definition and visualization.

2.2.2 Data Generation

Raw housing data contained seventeen variables in addition to price for 21,613 properties in the Seattle area. Least squares analysis was used to identify three input variables – square feet of living area, housing condition, and if it was waterfront property–and one latent variable– housing "grade". Additional regression analysis and correlation yielded the mapping of x1, x2, and x3 to housing condition, living area, and waterfront location, respectively. Samplings of these variables as well as values from the regression models were

then used to generate synthetic output points according to the model below. This model is derived from Model 5 in Figure 2 of this paper and Figure 1 of [9]. Variations on sample size, the assigned weight of living area on house grade, and the fraction of waterfront properties included (this was a very skewed, minority category that had a significant impact on selling price), were used to generate a wide range of synthetic data sets:

1. Sample sizes considered: 50, 100, 150, 200, 250, 500 [2]
2. Modulation in the assigned weight of x2 over a range of +/- 0.1 from its fited OLS regression value at increments of 0.05.
3. Number of waterfront properties (a heavily skewed category) included in the training data: 1, 5, 10, 25

These collectively created 120 conditions for generating synthetic data.

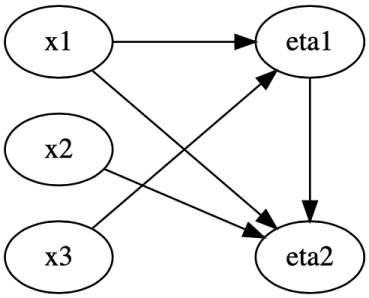
2.2.3 Model Design

As shown in Figure 2, seven models were created for analysis using `semopy` [5]. The structures of these models were taken directly from Figure 1 of [9]. While the models proposed in this paper were selected from stereotypical IS models, the focus of this project is on relative model selection not model design. Additionally, as in [9], these models are not be exhaustive of the space of all possible models, but rather a selection of them that allows for the topics of [9] to be explored. Information on translating them into `semopy` can be found in the appendix.

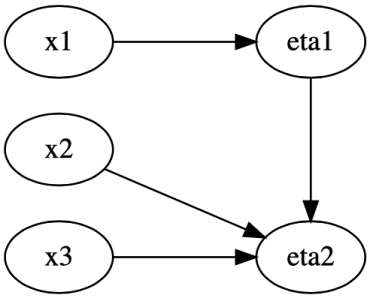
2.2.4 Simulation and Evaluation

Each of the 120 testing conditions were tested one hundred times. After each iteration of data generation, all the models were evaluated by the standards of each performance metric

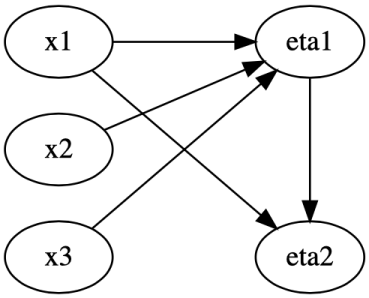
[2] Derived directly from methods of [9]



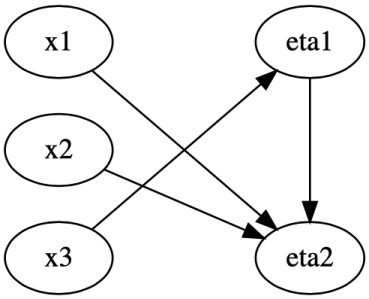
(a) Model 1: Incorrect Model



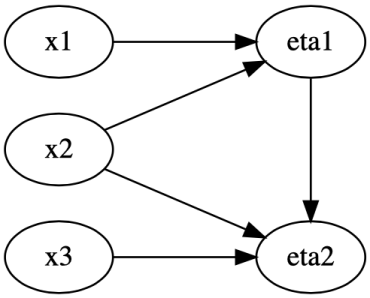
(b) Model 2: Parsimonious Model



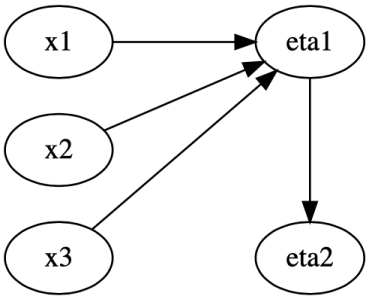
(c) Model 3: Incorrect Model



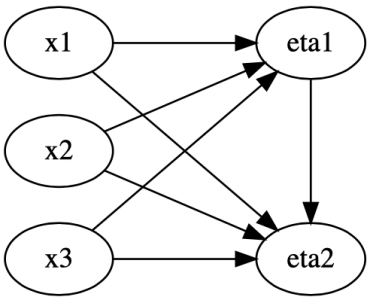
(d) Model 4: Incorrect Model



(e) Model 5: Data Generation Model



(f) Model 6: Incorrect Model



(g) Model 7: Saturated Model

Figure 2: Structure of models from [9]

and the best performing model was selected for each metric. This process was repeated until all iterations were complete.

3 Methods

The performance of the series of models on the synthetic data was evaluated by three primary categories of statistical performance metrics: standard PLS analysis methods, asymptotically efficient, and asymptotically consistent statistics.

[9] notes that there are a set of performance statistics standard to the analysis of PLS models, such as R^2 and adjusted goodness of fit. These statistics are often used because they are frequently used with other modeling approaches and appear intuitive; however, they can often times fail to identify the most correct model. R^2 will almost always increase as the number of parameters in the model increases; similarly, while adjusted R^2 considers the number of parameters included, it is not the most efficient or consistent measure of model performance [9]. Goodness of fit for these models can be considered "as the geometric mean of two types of R^2": "the average communality" and "the average R^2 of the endogenous latent variables" [3]. When calculated for multiple measured variables, this index can be inflated and does not capture models' relative capacity for describing relationships accurately [3].

Alternatively, to standard statistics, there are other asymptotically efficient and consistent methods that are increasingly popular to evaluate and compare model performances. As described in Wasserman, asymptotically efficient statistics minimize variance while asymptotically consistent statistics approach the true population parameter as sample size increases [11]. [10] describes and contrasts two popular statistics in both of these categories: The Akaike Information Criterion (AIC) and AIC corrected (AICc) were two asymptotically efficient methods analyzed. The AIC adjusts the maximum likelihood estimator for the number of parameters included in the model. AIC is effective at identifying correct parsimonious models, though is not as successful when the models being considered have a large number of estimated parameters, and so the AICc adjusts the AIC for cases where the number of parameters estimated are large relative to the amount of sample data. Likewise, the Bayesian information criterion (BIC) is an asymptotically consistent estimator that, with increasingly large sample size, can always statistically always identify the correct model. It does this through bayesian methods in computing the probability distributions for each of the estimated parameters.

4 Results

After simulating 100 iterations of each parameter configuration, the distribution of model selection for each parameter is recorded in the figures below. These values diverge in some ways from the results of [9]. This is not unreasonable as, while constructed similarly, the housing data did likely diverge significantly in distribution of the information system data used in [9]. However, the results did still support many key points made in [9] as well as what was stated about each statistic in the methods section above.

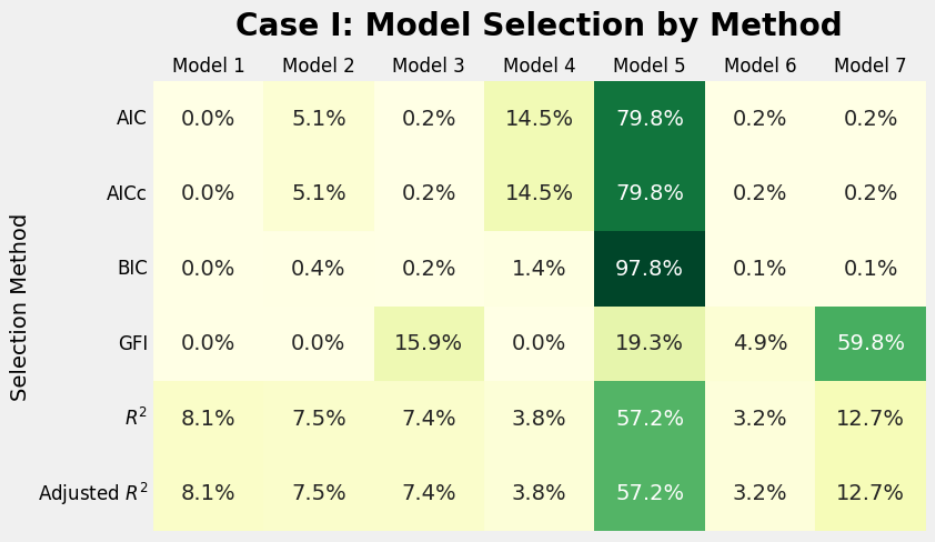


Figure 3: Case I Results

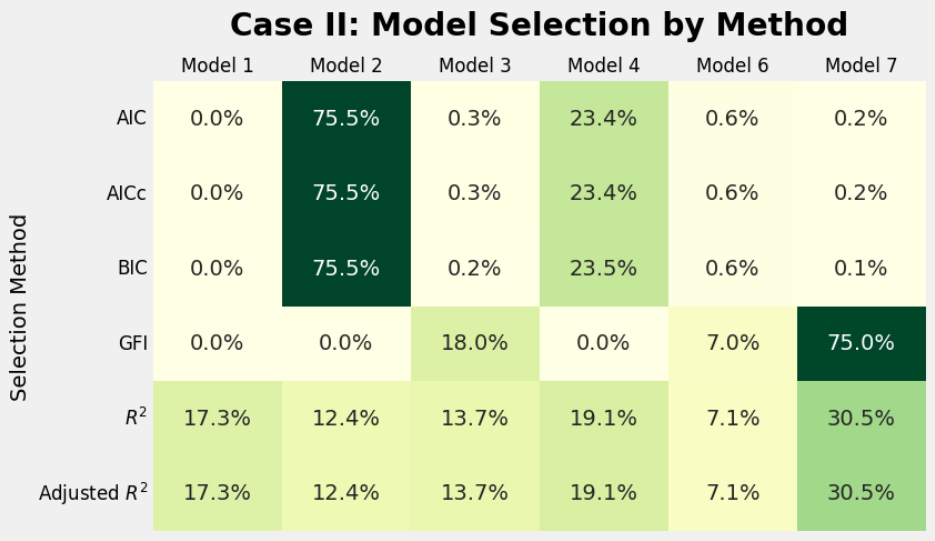


Figure 4: Case II Results

This divergence can serve as the be the guide for analysis. Firstly, the selection distribution of models for each statistic was quite different from tables 1A and 1B in their paper. Notably, in case I, the data generating model (model 5) was preferred by most of the statistics over the parsimonious model (model 2). This, however, did support what was stated by [10] that AIC-based statistics and BIC are both highly successful at consistently identifying the most correct (i.e. data generating) or parsimonious model. Additionally, when the data generating model was included, BIC was more effective than AIC at correctly identifying the data generating model, which was also stated in [10]. This aligns with the asymptotic consistency of BIC discussed in the methods section. Furthermore, the goodness of fit index also consistently favored the saturated model in both cases, as seen in figures 1A and 1B of [9].

The distribution of the R^2 based statistics for cases diverged as well from figures 1A and 1B. In the original study [9], R^2 favored the saturated model more, whereas in both cases of this simulation it was much more inconclusive. Furthermore, both of the adjusted statistics AICc and adjusted R^2 did not improve in performance from the base statistics AIC and R^2, respectively. This is an interesting point of divergence from the original study, but is likely due to the different spread of data in this housing study. Future variations of this experiment could include a scaled selection method where all models are awarded points for their performance by each statistical criteria for each iteration, which could allow better insight into relative performance.

5 Conclusions

Simplified PLS SEM methods are likely not going to replace currently popular housing price predictions used by companies like Zillow that incorporate dozens of variables [2, 12]. One key feature of these methods that this model lacked was more sophisticated handling of location-based data. A path-based model could be implemented in very interesting ways

to study the combined impact of several different factors in housing prices ranging from details about the house itself to location based data on local schools, public resources, and neighboring properties.

While this data diverged from that found in [9], it still provides an additional lens to further appreciate the strengths and shortcomings of each of these model selection methods as well as limits of PLS SEM itself. One limit that was found while designing and running this experiment was the impact of highly skewed data. While papers like [3] noted that this methodology was successful at handling skewed data, a limit to this was found during preliminary model design with the housing data. Initially, the three variables selected by just considering the significance of their regression coefficient through OLS methods were square feet of basement, square feet of living space, and the waterfront Boolean.

One issue that was quickly identified with this selection was that both square feet of basement and the waterfront Boolean were highly skewed variables. This introduced a data resolution issue where all variations of the linear regression-based models tested struggled to fit the data. For this reason, the skewed variable of square feet of basement was replaced by housing condition. In future studies, revisiting the model's fitting and performance issues when given such limited data would be a good point of focus. This could possibly be done by examining models that have many more channels of information or possibly implementing a data transformation such as a log transformation that allows for the model to better interpret this information.

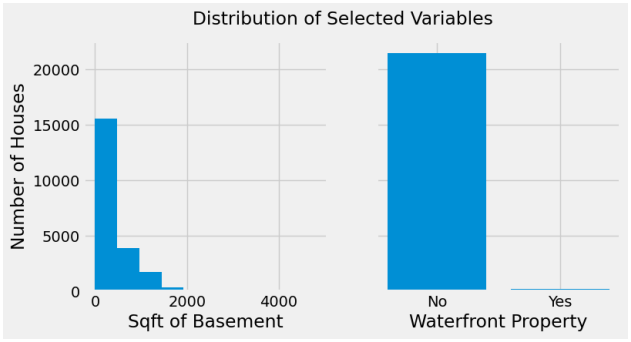


Figure 5: Skewed features

While this project was a simple introduction study of PLS SEM models using just three parameters about the properties; completing this project has been a very educational

experience for me to learn about this methodology and different analysis techniques that accompany it. As stated in the paragraphs above, there is much room to investigate more complex models to further model house pricing, and after this introductory exploration, I understand this method much better and feel excited to explore latent variables, specifically, in more depth through future work. Synthetic data generation is another point of interest for further future research. I think that investigating this as a methodology to help produce synthetic data for instances of limited data or, more interestingly, instances such as health data were raw original data might be of concern on the account of personal privacy concerns with PID.

6 Appendix

6.1 Tables

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
AIC	0.000000	0.050917	0.002417	0.145250	0.797917	0.001750	0.001750
AICc	0.000000	0.050917	0.002417	0.145250	0.797917	0.001750	0.001750
BIC	0.000000	0.004083	0.001750	0.013833	0.978167	0.001250	0.000917
GFI	0.000000	0.000000	0.159167	0.000000	0.193417	0.049000	0.598417
R^2	0.081333	0.074833	0.074333	0.038250	0.572167	0.032333	0.126750
Adjusted R^2	0.081333	0.074833	0.074333	0.038250	0.572167	0.032333	0.126750

Table 1: Case I: The data generating model (model 5) is included.

	Model 1	Model 2	Model 3	Model 4	Model 6	Model 7
AIC	0.00000	0.754583	0.003083	0.234250	0.006333	0.001750
AICc	0.00000	0.754583	0.003083	0.234250	0.006333	0.001750
BIC	0.00000	0.755083	0.002500	0.234917	0.006333	0.001167
GFI	0.00000	0.000000	0.180250	0.000000	0.070167	0.749583
R^2	0.17325	0.123583	0.136833	0.190917	0.070833	0.304583
Adjusted R^2	0.17325	0.123583	0.136833	0.190917	0.070833	0.304583

Table 2: Case II: The data generating model is not included.

6.2 Code Availability

All code used can be found on github at :

`https://github.com/ttorir/theoryOfStatistics1`

6.3 Data Sources

King County, Washington, housing data:

`https://raw.githubusercontent.com/rashida048/Datasets/master/home_data.csv`

6.4 Models Evaluated

6.4.1 Models Evaluated

The models were sourced from [9]. The figures below were generated with the online graphviz tool `https://dreampuf.github.io/GraphvizOnline/`.

6.4.2 Model Representation with semopy

For additional reference of methods used in this paper, additional explanation and reference on the models created using semopy[3] are also shared.

```
"""
Definition of the model equation

in semopy path models must be defined through a series of equations
for sampling, this is how model 1 can be defined and created using the
    library
"""
from semopy import Model, gather_statistics

model_eqn = '''eta1 ~ x1 + x3
          eta2 ~ x1 + x2 + eta1'''

# a semopy model object is then created
model = Model(model_eqn)

# the model can then be fit using data
model.fit(synthetic_data)

# once fit the model can then be used to predict values
model.predict(test_values)

# information on how well the model fit the data can also be gathered
gather_statistics(model)
```

[3]Complete semopy documentation can be found at `https://semopy.com/`

References

[1] Brannick, M. T. (n.d.). Structural Equation Modeling (SEM). Structural equation modeling (SEM). Retrieved November 10, 2022, from http://faculty.cas.usf.edu/mbrannick/regression/SEM.html

[2] Guerrieri, V., Hartley, D., Hurst, E. (2013). Endogenous gentrification and housing price dynamics. In Journal of Public Economics (Vol. 100, pp. 45–60). Elsevier BV. https://doi.org/10.1016/j.jpubeco.2013.02.001

[3] Henseler, J., Sarstedt, M. (2012). Goodness-of-fit indices for partial least squares path modeling. In Computational Statistics (Vol. 28, Issue 2, pp. 565–580). Springer Science and Business Media LLC. https://doi.org/10.1007/s00180-012-0317-1

[4] Höskuldsson, A. (1988), PLS regression methods. J. Chemometrics, 2: 211-228. https://doi.org/10.1002/cem.1180020306

[5] Igolkina, A. A., Meshcheryakov, G. (2020). semopy: A Python Package for Structural Equation Modeling. In Structural Equation Modeling: A Multidisciplinary Journal (Vol. 27, Issue 6, pp. 952–963). Informa UK Limited. https://doi.org/10.1080/10705511.2019.1704289

[6] Marikyan, D. Papagiannidis, S. (2021) Unified Theory of Acceptance and Use of Technology: A review. In S. Papagiannidis (Ed), TheoryHub Book. http://open.ncl.ac.uk

[7] Myung, J. I., Pitt, M. A. (2009). Optimal experimental design for model discrimination. Psychological Review, 116(3), 499–518. https://doi.org/10.1037/a0016104

[8] Rönkkö, M., McIntosh, C. N., Antonakis, J. (2015). On the adoption of partial least squares in psychological research: Caveat emptor. In

Personality and Individual Differences (Vol. 87, pp. 76–84). Elsevier BV. https://doi.org/10.1016/j.paid.2015.07.019

[9] Sharma, Pratyush Sarstedt, Marko Shmueli, Galit Kim, Kevin Thiele, Kai. (2018). PLS-Based Model Selection: The Role of Alternative Explanations in Information Systems Research. Journal of the Association for Information Systems. 10.17705/1jais.00538.

[10] Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychological Methods, 17(2), 228–243. https://doi.org/10.1037/a0027127

[11] Wasserman, L. (2010). All of statistics : a concise course in statistical inference. New York: Springer. ISBN: 9781441923226 1441923225

[12] Zillow home value forecasts – zillow help center. Zillow. (n.d.). Retrieved December 8, 2022, from https://zillow.zendesk.com/hc/en-us/articles/203512180-Zillow-Home-Value-Forecasts