PLS-SEM Model of Self-Pay Influencing Diabetic Treatment and Outcomes

Team 11

April 2022

1 Introduction

"Decades of research have demonstrated that diabetes affects racial and ethnic minority and low-income adult populations in the U.S. disproportionately," writes Hill-Briggs et al. in their 2020 review, Social Determinants of Health and Diabetes. While systematic discrimination on observable demographic traits by large institutions is well-documented in the US, healthcare provides a particularly challenging setting in which to study bias. Practitioners have the most proximate impact on patient care, as they are responsible for determining prescriptions, behavioral interventions, and recommending operations. In making such decisions, practitioners optimize over a weighted basket of interests that prioritizes client outcomes and employer interests, a reasonable simplification given that the ways in which practitioners directly benefit from client interactions dominantly act through client or employer outcomes. The client optimization problem combines a desire to minimize cost payed out-of-pocket by clients and improvement in client outcomes, which we observe through readmittance rates (within 30 days and after 30 days) and dismissal type which accounts for survival outcomes. The employer optimization exclusively prioritizes profit-maximization, which includes the sum of billable profits and

referral fees, the cost of net unprofitable operations, and the cost of government penalties. Government penalties enable us to connect the story between patient and employer interests. Just as clients do not want to have to return to the hospital, hospitals pay hefty penalties when a client is readmitted, as it reveals unsatisfactory quality of care. A 2019 study of readmissions rates and hospital financial performance in Washington state found that, for patients with diagnoses associated with expensive treatments, a reduction in readmission rates corresponded to a statistically significant increase in operating revenue (Upadhyay et al.). Diabetes is one of the canonical expensive diagnoses in the US. As of 2017, the American Diabetes Association estimates that Americans diagnosed with diabetes incur an average of $9,600 per year in diabetes-associated expenses. At a median household income of $61,423, having one diabetic family member reduces available household income by 16% (Guzman 2018). Similarly, the social cost of paying for diabetes treatment accounts for 10% of all US healthcare spending (ADA 2007). Patients, practitioners, administrators, insurers, and policymakers share a rare, unified goal of jointly minimizing the probability of readmission and minimizing expenditure. The aim of this paper is to exploit the cost- and readmission-minimizing nature of each actor in the treatment pathway in order to explore how the relationship between demographics and costs (explicit and excise) interacts with doctors' recommendations and patient outcomes.

2 Data

The primary dataset is an extract from the Health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States. The The Health Facts data we used was an extract representing 10 years (1999–2008) of care at 130 hospitals throughout the United States: Midwest (18 hospitals), Northeast (58), South (28), and West (16). Most of the hospitals (78) have bed size between 100 and 499, 38 hospitals have bed size less than 100,

and bed size of 14 hospitals is greater than 500. The dataset was extracted by Strack et. al. [1]

The database consists of a total of 117 features. The database includes 74,036,643 unique encounters (visits) that correspond to 17,880,231 unique patients and 2,889,571 providers. Because this data represents integrated delivery network health systems in addition to stand-alone hospitals, the data contains both inpatient and outpatient data, including emergency department, for the same group of patients. However, data from out-of-network providers is not captured.

Encounters were extracted from the datatbase that met the following five requirements:

1. It is an inpatient encounter (a hospital admission).
2. It is a "diabetic" encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
3. The length of stay was at least 1 day and at most 14 days.
4. Laboratory tests were performed during the encounter.
5. Medications were administered during the encounter.

To summarize, our dataset consists of hospital admissions of length between one and 14 days that did not result in a patient death or discharge to a hospice. Each encounter corresponds to a unique patient diagnosed with diabetes, although the primary diagnosis may be different. During each of the analyzed encounters, lab tests were ordered and medication was administered.

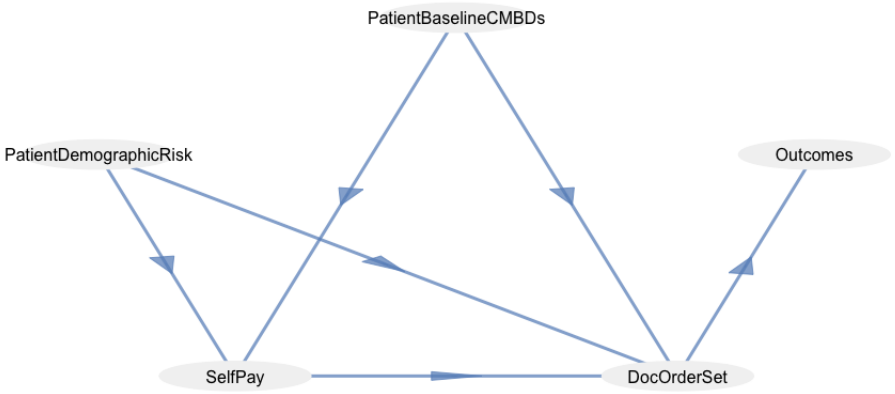

Figure 1: Inner Model

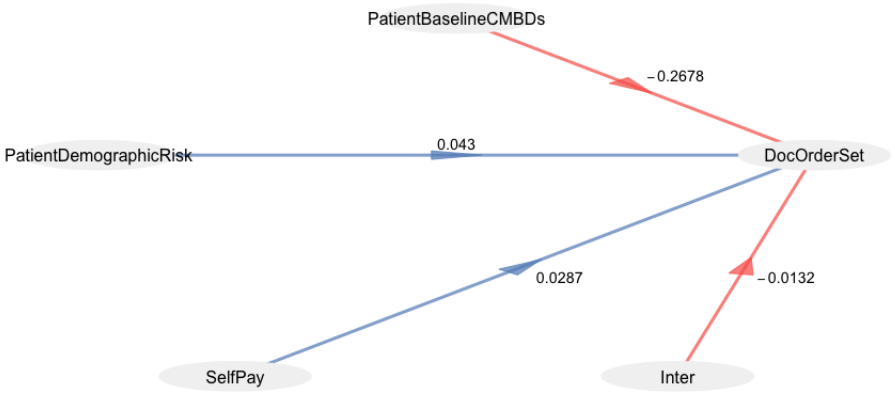

Figure 2: Outer Model

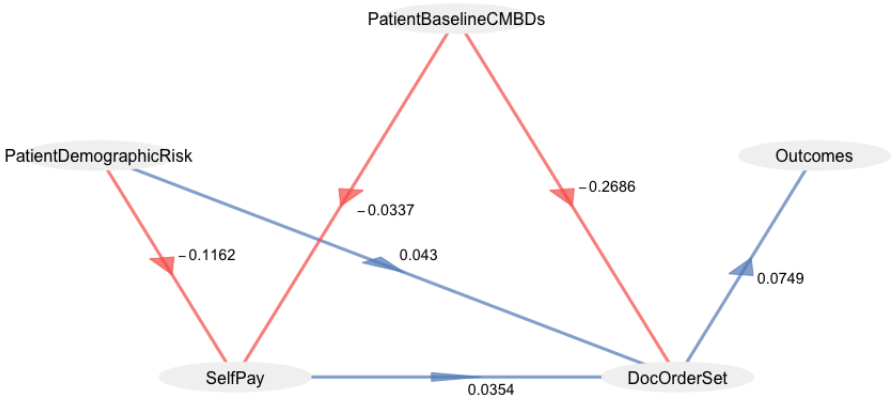

Figure 3: Base model coefficients

3 Model Specification

3.1 Base Path Model

Assumes that self-pay is a causal pathway, like any other latent variable driving the causal chains in the model.

The model that we use to examine the moderating effect of health insurance status on health outcomes is a structural model known as PLS-SEM. A PLS-SEM model is composed of two elements: (1) a *structural* or *inner* model and (2) a *measurement* or *outer* model. The structural model includes latent (unobserved) variables of interest. In our context, these include **Patient Conditions**, **Patient Demographics**, *Self-Pay*, **Interaction**, and **Doctor Order Set**. The **Patient Conditions** and **Patient Deomographics** are deliberately vague latent variables; they are the overall health and demographic characteristics of a particular patient. The **Self-Pay** variable is simply an indicator variable for whether a patient paid out-of-pocket at the healthcare encounter; this is our primary variable of

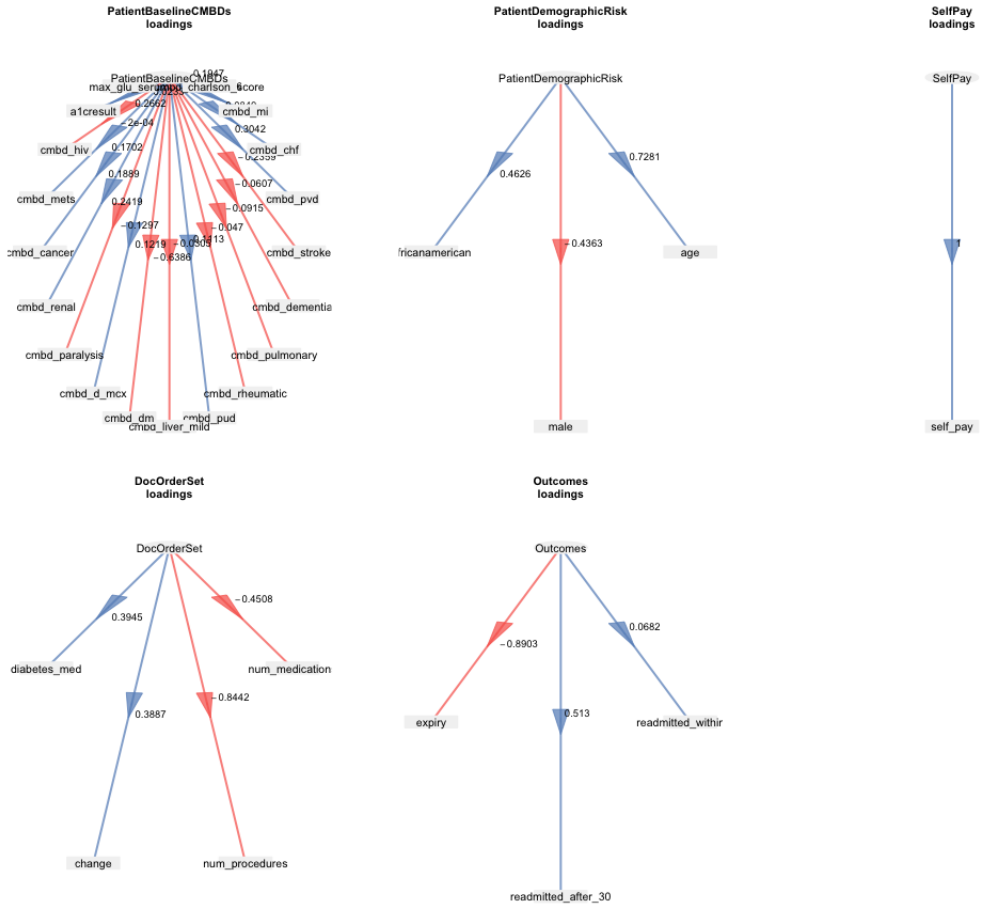

Figure 4: Factor loads

interest. The **Interaction** variable is the interaction of **Self-Pay** with the other latent variables and captures the moderating effect of insurance status on health outcomes. Finally, the **Doctor Order Set** is a proxy for the level of care that a given doctor provides – this is affected by all of the above latent variables and is an input to the final health outcome of a patient.

We include the schematic diagrams describing the causal directions of the latent and observed variables in Figures 1, 2, and 3.

3.2 Moderation Effect Model: Two-Stage Regression Analysis

In PLS-SEM, a moderator variable (M) causes a moderating effect whose variation influences the strength or the direction of a relationship between an independent variable (X) and a dependent variable (Y). Here, we conceive of self-pay as a noisy latent variable that drives a component of variation in the doctor's order sets latent variable. When using Two-Stage Regression Analysis to evaluate the moderation, the first-stage is identical to the base model. The second stage extracts the scores yielded in the first stage, and uses linear regression to analyse them.

Here, once we have obtained the scores for the patient health baseline conditions, patient demographics, self-pay, and doctor order sets latent variables, we interact the scores for the independent variable (patient health and demographics) with the moderator, self-pay. Then we estimate a linear model:

$$DocOrderSets = \beta_1 Demographics + \beta_2 HealthBaseline + \beta_3 SelfPay + \beta_4 (Demographics + HealthBaseline) * SelfPay + \epsilon$$

4 Results

Having validated our measurement model, we move on to assessing the results of the structural model to answer our three questions. First, paying out-of-pocket leads to doc-

tors making decisions which are more detrimental to outcomes than for clients who are paying with insurance. See the positive path coefficient from SelfPay, an indicator for paying out-of-pocket, to DocOrderSet, a latent variable which denotes the decisions doctors make about care in a given interaction in addition to the positive coefficient from DocOrderSet to Outcomes, a representation of the effectiveness of treatment measured through discharge and readmission timeline.

4.1 Basic Model

Table 1: Basic Model

	CMBDs	DemographicRisk	SelfPay	DocOrderSet	Outcomes
1 PatientBaselineCMBDs	0.00	0.00	0.00	0.00	0.00
2 PatientDemographicRisk	0.00	0.00	0.00	0.00	0.00
3 SelfPay	-0.03	-0.12	0.00	0.00	0.00
4 DocOrderSet	-0.27	0.04	0.04	0.00	0.00
5 Outcomes	0.00	0.00	0.00	0.07	0.00

4.2 Specification 2: Second Stage Regression

5 Validation and Goodness of Fit

Examining the validity of a PLS-SEM model is a two-stage process due to the latent nature of the model. First, the validity of the outer measurement model must be confirmed so that the observed variables are correctly allocated to latent variables. With this established, the validity of the inner model can be assessed and conclusions regarding the relationships between latent variables can be drawn.

Assessing the validity of the measurement model amounts to checking that, broadly speaking, observed variables have been allocated to the correct latent variable and that the observed variables assigned to each latent variable have strong mutual association.

Table 2: Second Stage Regression

	Dependent variable:
	DocOrderSet
PatientBaselineCMBDs	-0.268^{***}
	(0.004)
PatientDemographicRisk	0.043^{***}
	(0.004)
SelfPay	0.029^{***}
	(0.004)
Inter	-0.013^{***}
	(0.003)
Observations	59,758
R^2	0.074
Adjusted R^2	0.074
Residual Std. Error	0.962 (df = 59754)
F Statistic	$1{,}190.312^{***}$ (df = 4; 59754)
Note:	*p<0.1; **p<0.05; ***p<0.01

	var	Mode	MVs	C.alpha	DG.rho	eig.1st	eig.2nd
1	PatientBaselineCMBDs	A	19	0.12	0.43	1.85	1.26
2	PatientDemographicRisk	A	3	0.00	0.00	1.16	1.04
3	SelfPay	A	1	1.00	1.00	1.00	0.00
4	DocOrderSet	A	4	0.52	0.73	1.70	1.23
5	Outcomes	A	3	0.00	0.01	1.27	1.03

5.1 Unidimensionalty

To test for mutual association between the observed variables, we test for *unidimensionality* within each block of observed variables associated to a particular latent variable. Roughly speaking, unidimensionality means that the observed variables lie in the same direction as each other with respect to the latent variable. For instance, if the general latent variable **Health Outcome** improves, all of the measurements associated with the latent variable (**Expiry**, **Not Readmitted**) should improve as well.

We focus on two commonly used tests for unidimensionality: Dillon-Goldstein's Rho and the spectral gap of the correlation matrix of observed variables for each block. In Table 5.1, we report these statistics for each block of observed variables. As a rule of thumb, values of Dillon-Goldstein's Rho larger than 0.7 indicate sufficient levels of unidimensionality. For the spectral analysis of the correlation matrix, one expects the first eigenvalue of the correlation matrix of unidimensional variables to be much larger than 1 and the second eigenvalue to be much smaller than 1.

5.2 Communality and Cross-Loadings

Now that we have validated the mutual association within each block of observed variables associated to a given latent variable, we look to check that the observed variables are allocated correctly to latent variables. This amounts to verifying that a given latent variable explains a large portion of each associated observed variable and that each observed variable is not highly associated with other latent variables.

Recall that we model an observed variable X_{ij} associated to latent variable Z_i as

$$X_{ij} = \beta_{ij} Z_i + \varepsilon_{ij}$$

where ε_{ij} is a normally distributed error term representing the part of the observed variable that is not explained by the latent variable. Our model implicitly assumes that the magnitude of ε_{ij} is small relative to $\beta_{ij} Z_i$. We justify this assumption by computing the *communality* of each observed variable. The communality of X_{ij} is defined as

$$\text{Com}(Z_i, X_{ij}) = \rho(Z_i, X_{ij})^2 = \beta_{ij}^2,$$

where ρ is the standard Pearson correlation coefficient. This is simply the square of the loading and measures the proportion of the variance of X_{ij} that is reproducible from Z_i. We include the communalities of every observed variable in Table 1. As a rule of thumb, we want each communality to be larger then 0.5 so that at least half the variance of each observed variable is explained by the associated latent variable.

Now we look to establish that observed variables are not closely related to latent variable aside from the one we associate to them. To test this attribute of the model, we compute *cross-loadings* between different blocks of observed variables. We include the numerical values of the cross-loadings in Table 1. We are looking to verify that the cross-loading of each observed variable with its associated latent variable is larger than the cross-loading with other latent variables.

5.3 Inner Model Goodness of Fit

With the measurement model correctly specified, we can turn to assessing the validity and goodness of fit of the inner model, which contains our variables of interest. This step is very similar to validating a standard linear regression model. Each regression in

the structural model has its own R^2 value, which we assess as we would in any linear regression analysis. The breadth of the PLS-SEM framework means that there is no single global statistic to judge the goodness of fit of the entire model. An often-used metric is the aptly named GoF index, which is the geometric mean of the average communality (across all blocks of observed variables) and the average R^2 value. This is a rough metric for judging the explanatory power of a model because it is somewhat comparing apples to oranges. However, it does provide a single number to judge in the context of all PLS-SEM models – a naive rule of thumb is that a value larger than 0.7 is considered good. The GoF index for our model is approximately 0.81.

References

[1] Beata Strack et al. "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records". In: *BioMed research international* 2014 (Apr. 2014), p. 781670. DOI: 10.1155/2014/781670.