Problem Set #[1]

MACS 30100, Dr. Evans
Bethany Bailey

Problem 1 Classify a model from a journal.

Part (a). - (b). O'Brien, Rourke L. "Redistribution and the New Fiscal Sociology: Race and the Progressivity of State and Local Taxes." *American Journal of Sociology* 122, no 4 (2017). 1015-1049. https://doi.org/10.1086/690118.

Part (c). This article builds a model to predict the progressivity of U.S. state and local taxes. The model uses data from the Institute for Taxation and Economic Policy (ITEP) to calculate a Suits Index, which is a measure of tax progressivity.

$$\Sigma_s = \beta_0 + \beta_1 \text{Latino}_s + \beta_2 \text{Black}_s + \beta_3 \text{Asian}_s + \beta_4 \text{Foreign}_s + \beta_5 \text{Gini}_s + \beta_6 \text{income}_s + \beta_7 \text{unemp}_s + \beta_8 \text{poverty}_s + \beta_9 \text{lfpr}_s + \beta_{10} \text{party}_s + \epsilon_s \quad (1)$$

Part (d).

Exogenous:

- $Latino_s$: perc Latino
- $Black_s$: perc Black (non-Hispanic)
- $Asian_s$: perc Asian (non-Hispanic)
- $Foreign_s$: perc Foreign Born
- $Gini_s$: State-specific Gini coefficients
- $Income_s$: inflation-adjusted log total income per capital
- $unemp_s$: unemployment rate
- $poverty_s$: poverty rate
- $lfpr_s$: labor force participation rate
- $party_s$: Political party control (Republican control of the state house)

Endogenous:

- Σ_s: Progressivity of state and local taxes, operationalized by Suits index

Part (e).

Static vs. dynamic: This model is static because it is time-invariant; the Suits index is calculated at a static time.

Linear vs. nonlinear: This model is linear.

Deterministic vs. stochastic: This model is stochastic because it includes the ϵ_s error term.

Part (f). As we went over in class, when making models, in order to avoid multicollinearity, overfitting, etc., it is important to "Make things as simple as possible, but no simpler." Thus, I think this model may not need more variables, because it already contains 10 different variables to predict state and local tax code progressivity. However, one variable which could greatly influence the model would be size of tax base; that is, the number of tax-paying individuals in each region. If the tax base was extremely small, the sample used to calculate the variables would be too small to be significant; thus, the results could be greatly skewed. This should be taken into account when analyzing the model.

Problem 2 Make your own model.

Part (a).

The below is a model of whether a U.S. individual decides to get married (models may differ by country because the factors that influence marriage differ by culture).

$$P(married_i = 1|\text{status}_i, ..., \text{legality}_i) = logit^{-1}(\alpha + \beta_1 \text{status}_i + \beta_2 \text{age}_i + \beta_3 \text{religious}_i + \beta_4 \text{SES}_i + \beta_5 \text{location}_i + \beta_6 \text{gen}_i + \beta_7 \text{legality}_i + \epsilon_i) \quad (2)$$

Part (b). In this model, the dependent endogenous variable, $married_i$, is an indicator variable denoting whether or not the person will get married (1=get married, 0=not get married). It is determined using an inverse logit function.

Part (c). This model is a complete data generating process. As long as we have enough well-collected data on all of the exogenous variables, we could create a model to simulate the endogenous variable.

Part (d). The key factors that influence this outcome are the exogenous variables above: $status_i$ is a categorical variable that denotes relationship status (if someone is in a stable relationship, they might be more likely to get married); age_i is a continuous variable that represents the log age of the individual (as a person gets older, they are increasingly likely to get married, but this effect diminishes over time, e.g. older people may be less likely to get married if they are not already); $religious_i$ is an indicator variable that codes whether a person is religious or not (religious people are probably more likely to get married); SES_i is a variable that codes a person's socioeconomic status (people with more stable lives may be slightly less likely to get married); $location_i$ is a categorical variable that denotes whether a person is from an urban, suburban, or rural location (urban populations may be less likely to marry than suburban/rural); gen_i is an indicator variable that denotes male or female; and $legality_i$ is an indicator variable that denotes whether or not marriage is legal for the individual.

Part (e). Though other factors, such as race, educational attainment, whether or not one wants children, etc. also likely effect an individual's likelihood to marry, I did not include these factors because I believe they would influence the model in a way similar to other factors already included in the model; in other words, I believe these factors are closely correlated to other variables, such as SES, location, and religion. Thus, I would like to avoid multicollinearity and do not want to overcomplicate the model. The model above is my attempt to create the simplest possible model with the most predictive power.

Part (f). In order to test whether my chosen factors are significant in real life, I could use data collected from the General Social Survey (GSS) and/or Census data and/or a small-scale survey. I would split the dataset into training and testing data, and run a regression on the training data to determine the model coefficients (hopefully, they would be significantly different than 0, which would indicate that they were significant). Then, I could see if I could predict the outcomes of the testing data using my trained model.