Problem Set #1

MACS 30000, Dr. Evans
Xingyun Wu

Problem 1 Classify a model from a journal.

Part (a & b). Citation: David Jacobs and Jonathan C. Dirlam, "Politics and Economic Stratification: Power Resources and Income Inequality in the United States," American Journal of Sociology 122, no. 2 (September 2016): 469-500. Link

Part (c). This journal article tests the power resource hypothesis with a pooled time-series analysis of income inequality. The explanatory variables are lagged by at least a year, especially most political variables are lagged by two years, to capture possible lagged effect of changes in public policy or other aspects. Most continuous variables are logged unless the transformation ruins a variable's explanatory power. This is to correct modest nonlinearities. The equation for the statistical model is:

$$
\begin{aligned}
&ln(Theil \quad Index) = b_0 + b_1 ln(\%REPUB.GOV_{t-2}) + b_2 REPUB.LEG_{t-2} \\
&+ b_3 REPUB.PRES_{t-2} + b_4 PRES.ELECT.YR_{t-1} + b_5 ln(\%UNION_{t-1}) \\
&+ b_6 ln(FINANCE.EMP_{t-1}) + b_l n(7\%BLACK_{t-1}) + b_8 ln(\%BLACK \times YR)_{t-1} \\
&+ b_9(\%HISPANIC_{t-1}) + b_{10} INC.PER.CAP_{t-1} + b_{11} INC.PER.CAP^2 \\
&+ b12 ln(\%COLLEGE_{t-1}) + b_{13} ln(\%COLLEGE \times YR)_{t-1} + b_{14} DENSITY_{t-1} \\
&+ b_{15} DENSITY^2_{t-1} + b_{16} \%POVERTY_{t-1} + b_{17} ln(\%RURAL.EMP_{t-1}) \\
&+ b_{18} \%WOMEN.LBR.FRC_{t-1} + b_{19} ln(\%MANU.EMP_{t-1}) + b_{20} STOCK.VAL_{t-1} \\
&+ b_{21} ln(\%UNEMPLOYED_{t-1}) + b_{22} ln(\%UNEMPLOYED \times YR)_{t-1} \\
&+ b_{23} INEQ.MARGNL.TX + b_{24-73}(49 \quad STATE \quad TRNDS) \\
&+ b_{74-123}(49 \quad STATE \quad TRNDS^2) + e
\end{aligned}
$$

Part (d). (1) Exogenous variables are: marginal tax rate, variables indicating government conditions, union strength, higher education rate, financialization, state economic development, the degree to which a state's economy remains rural, population density, employment, women's labor force participation, minority presence.(2) Endogenous variables are: income inequality in the U.S. States from 1978 to 2011. It is measured with the natural log of the Theil inequality measure computed on Internal Revenue Service income data.

Part (e). The model is a dynamic, linear and stochastic model. It is dynamic because it captures changes, interactions and possibilities that might arise in time. It is linear because it represents a linear relationship among variables. And it is stochastic model because the endogenous varible is not fully determined by the exogenous variables and the initial conditions, which is why the equation includes an error term.

Part (f). What I think the model is missing that might be valuable: time invariant explanatory variables, such as State culture differences.

Problem 2 Make my own model.

Part (a - c) My model for how long popular musicians live is written as follows. The dependent endogenous variable is predicted lifespan (in years). The exogenous variables are: birth year of the musicians, gender, dummy variable of whether the musician regularly do physical exercise, dummy variable of existence of strong family tie, dummy variable of whether the musician has drug issue, dummy variable of whether the musician smoke, dummy variable of whether the musician has alcohol abuse, dummy variable of whether the working hours are long, and yearly income. The equation is:

$$PREDICTED \quad LIFESPAN = b_0 + b_1 BIRTH.YEAR + b_2 GENDER + b_3 EXERCISE + b_4 FAMILY + b_5 DRUG + b_6 SMOKE + b_7 DRINK + b_8 LONG.WORKING.HOURS + b_9 ln(YEARLY.INCOME) + e$$

Part (d). Although I believe most of the variables I put in my model are important, some may have more impact on how long popular musicians live. The assumed key factors are yearly income, physical exercise, family, and drug/smoking/drinking issue.

Part(e). (1) Yearly income. It is among key factors because it influences the musicians' relative socioeconomic status, which would set a foundation for the prediction of lifespan. (2) Physical exercise. People regularly doing physical exercise tend to have better health condition and longer life. (3) The variable for family. It is a key factor for people's social support and mental health. And they could significantly influence the musician's lifespan. (4) The variables for drug, smoking and drinking issue. the other key variables are so important because they have critical impact on their physical health.

Part(f). I would gather testing data from the Internet and run my model with the data for the preliminary test. Here are the steps I would take to do the preliminary test:

1. Gather data from the Internet. Since rock & roll is one type of popular music, its relevant data could be used to do the preliminary test. There is a wikipedia page named List of Deaths in Rock and Roll, including data of musician names, age when died, date, location and cause of death. I would randomly generate 50 numbers indicating rows, which would be based on the total amount of records that web page records. I need 50 observations to get some extent of statistical significance, while gathering information of 50 people would not cause too much manual work. Then I would scrape the picked musicians' wikipedia page for the information needed for my model.
2. Data recording and data cleaning in statistical software. Information gathered from wikipedia page would be text of descriptions. Since the observation is only 50, I would manually extract the information from the texts. Then record the data and so data cleaning in R or STATA for the next step.
3. The relationships should be linear. Run the linear regression model in R or STATA and get the results of whether my factors are significant in real life.