Problem Set 1

Problem 1

Classify a model from a journal.

a) I chose an article from *American Economic Review*, which is *The Expanding Gender Earning Gap: Evidence from the LEHD-2000 Census*. Full citation of this article is listed in the following question.

b) Detailed citation of this article in Chicago style is listed as above: Goldin, Claudia, Sari Pekkala Kerr, Claudia Olivetti, and Erling Barth. "The expanding gender earnings gap: Evidence from the LEHD-2000 Census." *American Economic Review* 107, no. 5 (2017): 110-14.

c) There are two models in this article:

$$\ln(y_{ijt}) = \alpha_{jt} + \sum_{k} \beta_{kjt} X_{ijkt} + \varphi_{jt} F_{ij} + \varepsilon_{ijt} \tag{1}$$

where in equation (1), y is the mean quarterly earnings for individual i of education level j in year t. X is a vector of k individual demographic characteristics, including race and exact education which are invariant with time, exact age and state which can change through time.

Here is the second equation, which is an extended version of equation (1):

$$\ln(y_{ijt}) = \alpha_{jt} + \sum_{k} \beta_{kjt} X_{ijkt} + \gamma_{jt} \ln(MEE)_{ijt} + \xi_{jt} I_{ijt} + \omega_{jt} O_{ij} + \varphi_{jt} F_{ij} + \varepsilon_{ijt} \tag{2}$$

where MEE denotes mean establishment earnings, I is the three-digit industry code of the establishment, and O is the three-digit occupation dummy.

d) Generalized from the above models, y, X_{ijkt}, F_{ij}, MEE, I_{ijt}, and O_{ij} are exogenous variables. These variables are the input of the model and are obtained from different datasets.

α_{jt} (which is the constant term in this model), β_{kjt}, γ_{jt}, ξ_{jt}, ω_{jt}, φ_{jt}, and the error term ε_{ijt} are endogenous variables. They are the outputs of the model.

e) The two models in this article are both static because they do not include lags of the dependent variable among the regressors. Hence they do not capture dynamic relationships between variables of interest.

The two models are log-linear, which is hard to determine whether they are completely linear or completely non-linear. They can be regarded as non-linear models because the dependent variable y and one of the independent variables MEE are taking the form of

log. However, they can, to some degree, be considered as linear models cause they are linear combinations of parameters of interest. Therefore, these models can be judged as linear models in a more general way.

The models are deterministic not stochastic, since the output of the model are fully determined by the inputs.

f) There is one variable that the model is missing: the individual's marriage situation. To fully understand and interpret gender earning gap, the authors need to include marriage situation, which is whether the person is married or not, into his or her demographic background. Required by social gender norm, married women might devote more time to their family than to their work. Therefore, the gender earning gap might be larger for married people compared with single people.

Problem 2

Make your own model.

a) I use logistic model to predict whether the person decide to get married or not. The dependent endogenous variable *married* is determined by the probability of getting married.

$$married = \begin{cases} 1 & \text{if } Pr(mar) > 0.5; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

where

$$Pr(mar) = \frac{1}{1 + e^{-\mathbf{X}}} \tag{4}$$

where

$$\mathbf{X} = \beta_0 + \beta_1 female + \beta_2 inc + \beta_3 edu + \beta_4 age + \beta_5 child + \beta_6 work + \vartheta \tag{5}$$

d) In my model, variables in equation (5) are the key factors that influence people's marriage decisions. In equation (5), $female$ is a dummy variable that denotes the individual's gender; *inc* is one's income; *edu* denotes one's educational level; *age* denotes one's age; *child* denotes the number of children one has; and *work* denotes whether the person has a job or not.

e) There are other variables that may also influence people's marriage decisions, including their characteristics, their parents' relationship, their religion background, also their sexual identity as well as sexual orientation. Some of these factors cannot be easily and accurately measured, such as one's characteristics and their parents' relationship. Other

variables, such as one's religion background and information regarding to one's sexual orientation, are quite sensitive hence not easy to obtain. Therefore, I only take consideration of a certain number of key factors in my model. Information of these variable are well accessible.

f) I would use CPS (Current Population Survey) data to do a preliminary test on my model. This dataset is conducted by the U.S. National Bureau of Census, thus is reliable in the first place. CPS data also contains detailed information about earnings, labor force, and demographic characteristics. Firstly I will randomly assign a fraction of CPS data to a training set and obtain the estimation parameters in my model. Then I will apply my model to the data remained to see whether the model is valid or not.