Problem Set #[1]

MACS 30100, Dr. Evans
Liqiang Yu

Problem 1 Find a theoretical or statistical model from a recently published article (no earlier than 2013) in either the American Economic Review, American Journal of Political Science, or the American Journal of Sociology.

Part (a). I found a statistical model from a paper called "How Do Hours Worked Vary with Income? Cross-Country Evidence and Implications." on American Economic Review 2018, 108(1). I am particularly interested in the relation between working hours and salary across or within countries.

Part (b). Ref: Bick, Alexander, Nicola Fuchs-Schndeln, and David Lagakos. 2018. "How Do Hours Worked Vary with Income? Cross-Country Evidence and Implications." American Economic Review, 108(1): 170-99.

Part (c). There are several linear regression models in this paper, corresponding to different subquestions. The one which I am especially interested in is the measurement of individual hours-wage elasticities (Figure 1) by country. The regression model is

$$\log(h_i) = \alpha + \beta_w \log(w_i) + \delta_1 age_i + \delta_2 age_i^2 + \epsilon_i.$$

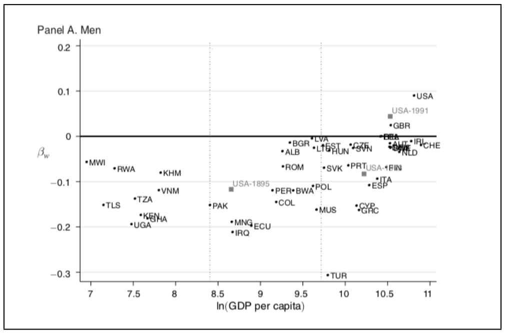
Part (d). Types of variables are categorized below:

- Exogenous variables: $\log(h_i)$, $\log(w_i)$, age_i, age_i^2.
- Endogenous variables: α, β_w, δ_1, δ_2, ϵ_i.

Part (e). This is a linear, static and stochastic model. It is linear since it is a linear approach between a dependent variable and several independent variables. There is no time involving so it is static. Stochasticity comes from the error term ϵ_i, which makes the model stochastic.

Part (f). A categorical variable indicating education level could have been introduced to this model, to evaluate individual hours-wage elasticities within groups with similar education level.

Figure 1: Elasticities for men across countries



Problem 2 Decision on getting married model.

Part (a). This is a prediction problem whose output set is {0,1}, so a logistic regression would be appropriate for this problem. To obtain the data, we can design a survey study which asks if the participant wants to get married or not and other demographic information. To obtain the model, there would be two steps. First, we want to use our training data to learn the parameters θ_i:

$$h_\theta(x) = g(\theta^T x),$$
$$where\ y = 0\ when\ h_\theta(x) < 0.5\ and\ y = 1\ for\ h_\theta(x) > 0.5$$

Here, g is the sigmoid function. After training, we then obtain the parameters θ:

$$\theta^T x = \theta_0 + \theta_1 x_1 + ... + \theta_n x_n.$$

For our interest, we want to focus on several exogenous variables:

$$\theta^T x = \theta_0 + \theta_1 sex + \theta_2 age + \theta_3 education\_level + \theta_4 income$$

Part (b). The output set of the dependent endogenous variable is $y = 1$ for getting married and $y = 0$ for not getting married.

Part (c). This model satisfies a complete data generating process, that is, given all parameter θ_i, we can obtain a binary classification by simulating data points for exogenous independent variables and putting them into our model.

Part (d). The possible key factors that may influence our outcomes are:

- Sex, Age, Education level, Income.

Part (e). I first came up with many factors by brainstorm, such as age, salary, family location, etc. Then I searched online articles or resorted to trustworthy data source to filter out valuable factors and lowly-correlated variables.

- Sex: From the Estimated Median Age of First Marriage by Gender, we can observe a gap in age between men and women when first getting married.
- Age: As above, the mean age of men for first marriage is 29.2 in 2015, so we anticipate men are prone to getting married when they are above a certain age.
- Education level: "College-educated adults are more likely to be married than less-educated adults.", Wendy Wang says.
- Income: "marriage has sharply declined among people without college degrees, while staying steady among college graduates with higher incomes.", said by Claire C. Miller.

Part (f). For practical significance, we can first use common knowledge to eliminate unreasonable factors and "compute predicted/expected values for hypothetical cases" or calculate "marginal effects", as mentioned in How to calculate the practical significance of citation impact differences?, to evaluate practical significance.

For statistical significance, since we are using logistic regression, we can calculate the p-value for each θ as well as looking at the R-square value. Besides, by modifying the number of features, we can compare between models and choose the significant one. Beyond manually select factors, we can apply a Model Selection on our model.