

Problem 2.**Part (a)-(c).**

Note: Since the problem requires an output of either 0 or 1, a binary logistic regression with a floor function is applied here. My model includes the following variables: **Y** (output, either 0 or 1), **age** (by years), **gender** (male/female), **race** (Nonhispanic African Americans vs. Other), **income** (logarithm of annual individual income), **BMI** (body mass index), **location** (state where an individual currently lives), **Recession** (the occurrence of an economic recession), **LGBT** (gay, lesbian, bisexual, or transgender). The model can be expressed as follows:

$$Y_i = \lfloor \pi_i + 0.5 \rfloor = \frac{1}{1 + e^{-\lambda}}$$

where

$$\lambda = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{gender}_i + \beta_3 \text{race}_i + \beta_4 \text{income}_i + \beta_5 \text{BMI}_i + \beta_6 \text{location}_i + \beta_7 \text{Recession}_i + \beta_8 \text{LGBT}_i + \epsilon_i$$

Part (d)-(e).

The key factors in this model are:

Demographics: Goodwin, McGill, and Chandra (2009)¹ discovers that age, gender, and race groups are three important factors affecting the odds of getting married. The effect has been repeatedly tested and significant across all subgroups.

Income: Building a family requires a lot of resources. Goodwin, et al. (2009) discovers that a white woman who is not poor is 550% more likely to get married than a poor African American woman.

Location: According to the US Census (2003)², D.C. has the lowest marriage rate (24.9%) in the US, which is almost half of the other 42 states. Weather, local cultures, and costs of living all play a role when it comes to the decision of getting married.

The following factors are also important but not as important as the above variables:

BMI: BMI is not an accurate representation of physical attractiveness; the effect of BMI may be partly explained by **age** and **location** (the tastes are different among different age groups and locations).

LGBT: The effect of LGBT is largely moderated by local culture (whether other people care) and state laws (whether they can lawfully get married). Therefore, the effect of **LGBT** can be partly explained by the effect of **Location**.

Recession: Since recession can impact people's annual income, the effect of **Recession** may be partly explained by **income**.

Part (f).

To test my model, I decide to get the data from NHANES. The NHANES includes almost all of the variables in my model. Moreover, NHANES is conducted every two years, with over 8,000 effective observations per cycle. Once I get the data, I am going to conduct a stepwise regression to check if my variables are significant in the model. I will also add the variables that are not in my model to test whether the effects of my variables are mediated or moderated by other variables. Finally, I would like to use a hierarchical design: Adding variables to a blank model. I would like to see how the explanatory power of my model changes after I add the variables one-by-one. I will try to strike a balance between effectiveness and parsimony.

¹Paula Goodwin, et al. (2009): Who Marries and When? Age at First Marriage in the United States. *Centers for Disease Control and Prevention/National Center for Health Statistics*, Data Brief 19

²Rose Kreider and Simmons, Tavia (2003): *Marital Status: 2000: Census Brief*