Question 1

(a) The article I find is called "Does Diversity Matter for Health? Experimental Evidence from Oakland". This paper mainly finds and explains the effect of racial diversity of physicians on the demand of preventive health care among the African American men.

(b) Alsan, M., Garrick, O., Graziani, G. (2019). Does Diversity Matter for Health? Experimental Evidence from Oakland. American Economic Review, 109(12), 4071–4111.

https://doi.org/10.1257/aer.20181446.

(c) The paper (Alsan, 2019) has mainly four equations in its model.

$$Y_i = \alpha + \beta_1 \cdot \mathbf{1}_i^{BlackMD} + \beta_2 \cdot \mathbf{1}_i^{\$5} + \beta_3 \cdot \mathbf{1}_i^{\$10} + \Gamma' X_i + \epsilon_i \tag{1}$$

$$\mathbf{1}_i^{RaceMD=k} = \alpha + \beta_1 \cdot \mathbf{1}_i^{RaceResp=k} + \Gamma' X_i + \epsilon_i \tag{2}$$

$$\mathbf{1}_i^{RaceMD=RaceResp} = \alpha + \beta_1 \cdot \mathbf{1}_i^{BlackResp} + \Gamma' X_i + \epsilon_i \tag{3}$$

$$\mathbf{1}_{il}^{RaceMD=RaceResp} = \alpha + \beta_1 \cdot \mathbf{1}_i^{BlackResp} + \lambda_l \cdot \mathbf{1}_l^{Domain} + \Gamma' X_i + \epsilon_{il} \tag{4}$$

For the meaning of variables in the equations, in the equation (1), Y_i refers to the demand of preventive health care for participants; $\mathbf{1}_i^{BlackMD}$, refers to the indicator on whether the participants are assigned with black doctors; $\mathbf{1}_i^{\$5}$ and $\mathbf{1}_i^{\$10}$ are dummy variables that indicate whether participants receive money incentives for the preventives; X_i is a combination of control variables refering to some characteristics of participants including the self-reported health, any health problem, ER visits, nights hospital, medical mistrust, whether has primary care physician, whether uninsured, age, whether married, whether unemployed, education, income and attrition.

The equations from (2) to (4) are a series that explores whether the preference of black men for a black physician is unique for their ethnic group and whether such preference varies across the health care domains. Here the $RaceMD$ and $RaceResp$ refers to the race of doctors and respondents, $BlackResp$ refers to the black respondents, and l refers to indicator of domain categories in health care system. The X_i in these three equations includes the age, education and income of respondents.

(d) The endogenous and exogenous variables in each equation in the model are as followed.

Equations	Endogenous	Exogenous
(1)	$\alpha, \beta_1, \beta_2, \beta_3, \Gamma', \epsilon_i$	$\mathbf{1}_i^{BlackMD}, \mathbf{1}_i^{\$5}, \mathbf{1}_i^{\$10}, X_i, Y_i$
(2)	$\alpha, \beta_1, \Gamma', \epsilon_i$	$\mathbf{1}_i^{RaceResp=k}, X_i, \mathbf{1}_i^{RaceMD=k}$
(3)	$\alpha, \beta_1, \Gamma', \epsilon_i$	$\mathbf{1}_i^{BlackResp}, X_i, \mathbf{1}_i^{RaceMD=RaceResp}$
(4)	$\alpha, \beta_1, \lambda_l, \Gamma', \epsilon_{il}$	$\mathbf{1}_i^{BlackResp}, \mathbf{1}_l^{Domain}, X_i, \mathbf{1}_{il}^{RaceMD=RaceResp}$

(e) Considering all the four equations in the model, the model should be classified as a static, linear and deterministic one.

(f) One variable I think the model is missing is inside the X_i, which refers to the characteristics of participants involved. According to the paper, characteristics are introduced as control variables as mentioned before. However, as a paper relevant to ethnic diversity, the variable on the attitudes of ethnic diversity should be included. For example, the question could be whether they agree on racial diversity in general. Therefore, such variable could have been included in the model.

Another variable I think may have been included is the indicators with incentives in the baseline model. There is no clear proof in the paper on why the incentives are \$5 and \$10. Therefore, it is likely that the threshold of participants to change their behavior is above 10 dollars, and the indicators of \$15 or more should be included.

Question 2

(a) I would like to form a Probit model. The model is described as followed.

$$y_j^* = \sum_{i=0}^{n} \beta_i x_{ij} + \varepsilon_j = \beta X_j + \varepsilon_j \tag{5}$$

$$y_j = \begin{cases} 1, y_j^* > y^* \\ 0, y_j^* <= y^* \end{cases} \tag{6}$$

$$G(z) = \int_{\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}} dv \tag{7}$$

$$Pr(y_j = 1|X_j) = Pr(y_j^* > 0|X_j) = Pr(\beta X_j + \varepsilon_j > 0|X_j) = 1 - G(-\beta X_j) = G(\beta X_j) \tag{8}$$

For the model, y_j denotes whether participants decide to get married (y_j=1) or not (y_j=0). The decision is made by a latent variable y_j^*, which cannot be observed but has a linear relation with factors. There is a bar called y^*, above which people choose to get married. In equation (5), x_{0j} equals to 1 to get the constant value, and X_j includes the key factors I think will influence the decision of marriage. I will discuss it later. Equation (7) and (8) indicate that the probability to get married is following the normal distribution.

In order to better simulate the data, I would like to divide the data into training data and test data. It is like a machine learning method. I will set a threshold on probability of getting married, and in the prediction, those above the threshold will be considered to have the decision to get married.

(b) Since I will make predictions in the test data based on the model, the predicted value will be the output of the model, and y_j in test data is endogenous.

(c) The model is a complete data generating process since if I could get all the data needed, the model could simulate the parameters and make predictions.

(d) I could like to consider three categories of factors to be included. The first one is the demographic features of individuals, including the age, gender, race, income, education level and profession type. The second category is the marriage status of the participants' major family members. The third category considers the sexual orientation of participants.

(e) The reasons why I choose the factors above are as followed. The demographic features should be considered as key factors. For example, if the income level is not high, participants may not afford the daily lives after marriage and thus are not willing to engage into marriage.

The marriage status of family members also has a influence. According to the paper of Cunningham and Thornton (2006), the marriage quality of parents will greatly affect the children's attitudes towards marriage, and thus influence their decisions to get married.

The third category mainly considers the homosexual people. Many countries still forbid homosexual marriage, so the sexual orientation will affect the marriage decisions, especially for the homosexual or LGBTQ people.

(f) One preliminary test on whether the factors are significant is to see the accuracy of the prediction on test data. If the most of the predictions match the real decision, I can say the factors I choose are significant in the real life and have a good prediction power.

References

Alsan, M., Garrick, O., Graziani, G. (2019). Does Diversity Matter for Health? Experimental Evidence from Oakland. American Economic Review, 109(12),4071–4111.https://doi.org/10.1257/aer.20181446.

Cunningham, M., Thornton, A. (2006). The influence of parents' marital quality on adult children's attitudes toward marriage and its alternatives: Main and moderating effects. Demography, 43(4), 659–672. https://doi.org/https://doi.org/10.1353/dem.2006.00.