Question 1. You are regressing y on 4 mutually exclusive and exhaustive dummy variables: a dummy for high school dropout, a dummy for high school graduate, a dummy for high school graduate plus some university, a dummy for university graduate or more. How is each coefficient interpreted? What is the formula for each estimated OLS coefficient in this regression?

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

$$x_1 = \begin{cases} 1 & \text{High school dropout} \\ 0 & \text{otherwise} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{High school graduate} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{High school graduate + University} \\ 0 & \text{ otherwise} \end{cases} \qquad x_4 = \begin{cases} 1 & \text{University graduatue or more} \\ 0 & \text{ otherwise} \end{cases}$$

Each coefficients are interpreted as the effect of x_i on y.

Question 2. Consider the classic test of parameter stability of a linear regression model across two subsamples I and II, of size N_I and N_{II} respectively, where:

$$y_I = x_I \beta_I + \epsilon_I \quad \text{and} \quad y_{II} = x_{II}\beta_{II} + \epsilon_{II}$$

and the hypothesis of stability corresponds to $H_0 : \beta_I = \beta_{II}$. Define two dummy variables of length $N_I + N_{II}$ as follows:

$$d_{Ii} = \begin{cases} 1 & \text{if observation belongs to subsample I} \\ 0 & \text{otherwise} \end{cases}$$

and

$$d_{IIi} = \begin{cases} 1 & \text{if observation belongs to subsample II} \\ 0 & \text{otherwise} \end{cases}$$

Use these dummy variables to derive a test of H_0 using an F-statistic.

[**Answer**] Let's use the Chow test.

$$F = \frac{RSS_C - (RSS_I + RSS_{II})/(k)}{(RSS_I + RSS_{II})/(N_I + N_{II} - 2k)} \sim F(k, N_I + N_{II} - 2k)$$

If identical, both betas will be the same ($H_0 : \beta_I = \beta_{II}$)

$$\text{Unrestricted RSS} : y = d_{Ii} x_I \beta_I + d_{IIi} x_{II} \beta_{II} + \epsilon$$
$$\text{Restricted RSS} : y = \beta(d_{Ii} x_I + d_{IIi} x_{II}) + \epsilon$$

If F-statistic > ciritical value, then reject H_0

Question 3. An investigator considers the linear model:

$$\log Earnings_i = \gamma_1 + \beta_1 Age_i + \beta_2 FullTimeEmployed_i + \beta_3 Tenure_i + \gamma_2 Education_i + \gamma_3 White_i + \gamma_4 Female_i + \epsilon_i$$

where $FullTimeEmployed_i$ is a dummy variable indicating whether individual i was employed full time in period t, $White_i$ is a dummy indicating whether individual i is of white race, and $Female_i$ is a dummy taking the value 1 if individual i is female. The available sample of a cross-section of individuals is indexed by $i = 1, \cdots, N$.

1. Explain what would happen if instead of the variables $White_i$ and $Female_i$ were to use the complementary variables $Non - White_i$ and $Male_i$ defined in the obvious way.

   If instead of the variables $White_i$ and $Female_i$ were to use the complementary variables $Non - White_i$ and $Male_i$, the values and meanings of the coefficients change.
   The meanings of γ_3 and γ_4 change as follows:

   - γ_3 : effect of $Non - White_i$ on $Earnings_i$
   - γ_4 : effect of $Male_i$ on $Earnings_i$

   However, RSS of the model remains the same.

2. Suppose we define the interaction variables $Z_i \equiv Age_i \times White_i$ and $W_i \equiv Age_i \times Female_i$. What would we achieve by introducing these two variables as additional regressors?

   We would achieve the effect of white's age on $Earnings_i$ by interaction variable Z_i and the effect of women's age on $Earnings_i$ by interaction variable W_i.

3. Discuss possible reasons why the regressors may violate exogeneity assumptions with respect to the error term, i.e., regressors and disturbances may be statistically related. Which regressors do you consider the most suspect in this regard?

   $Education_i$ seems to be statistically related to ϵ_i because omitted variables such as parent's Earning, parent's Education are related to $Education_i$. If parents are rich, they can get a better education, and if parents are smart, their children are also likely to be genetically smart. Also, since $Education_i$, $White_i$, and $Female_i$ affect $Tenure_i$, there seems to be a multicollinearity problem.