Assignment 3 is due no later than 5pm Friday, 29th of May, 2020. In submitting your work, you are consenting that it may be copied and transmitted by the University for the detection of plagiarism. Submission is your guarantee that the below statement of originality is correct.

"This is my own work. I have not copied any of it from anyone else."

NAME: **Replace this text with your name.**
STUDENT NUMBER: **Replace this text with your student number.**

Instructions for assignment: You will need to submit two documents for this assignment. The first document is a pdf document named `Assign3_StNo.pdf` which will provide all your analysis and solutions for this assignment. To produce this pdf document you will need to use LaTeX. The LaTeX document which was used to produce this assignment is named `Assign3_StNo.tex` and is located in the Assignment 3 folder in the Topic 9 section of LMS. You can use LaTeX online via Overleaf which is a website dedicated to producing documents from LaTeX. To use LaTeX, follow the instructions in the `Overleaf.pdf` document located in the Assignment 3 folder. The second document that you will need to submit is an R document named `Assign3_R_StNo.R` which is located in the Assignment 3 folder. This document should provide the R code you used to perform all your data manipulation and analysis.

Assessment information for assignment: There are a total of 100 marks for this assignment.

Description of assignment: The data and information presented in this assignment is adapted from Vella and Verbeek (1998)[1] and is stored in the file named `WageRace.csv` located in the Assignment 3 folder in the Topic 9 section of LMS. The hourly wage (measured in US dollars) was recorded for 148 full-time working males each year between 1984 and 1987. These males were also classified into one of two racial groups (*Black*, *Hispanic*). The variables of interest for Assignment 3 are:

- $Subject$: This is a factor variable that identifies the subject (working male).
- $Year$: This is a factor (categorical) variable that identifies the year the observations were recorded. It has 4 levels (*1984*, *1985*, *1986*, *1987*).
- *Race*: This is a factor (categorical) variable that identifies the racial group of the subject. It has 2 levels (*Black*, *Hispanic*).
- H: This is a factor (categorical) variable that identifies whether the subject is Hispanic or Black. It has 2 levels (0 = *Black*, 1 = *Hispanic*).
- HW: The hourly wage which is treated as a continuous variable.
- $Y2$: This is a factor (categorical) variable that identifies whether the hourly wage was recorded in 1985 or not. It has 2 levels (0 = *not 1985*, 1 = *1985*).
- $Y3$: This is a factor (categorical) variable that identifies whether the hourly wage was recorded in 1986 or not. It has 2 levels (0 = *not 1986*, 1 = *1986*).
- $Y4$: This is a factor (categorical) variable that identifies whether the hourly wage was recorded in 1987 or not. It has 2 levels (0 = *not 1987*, 1 = *1987*).

2 marks are allocated for each question that requires the use of the R computer package. These marks are awarded using the following criterion:

1. **R code that accurately produces the analysis/output required in the question.**

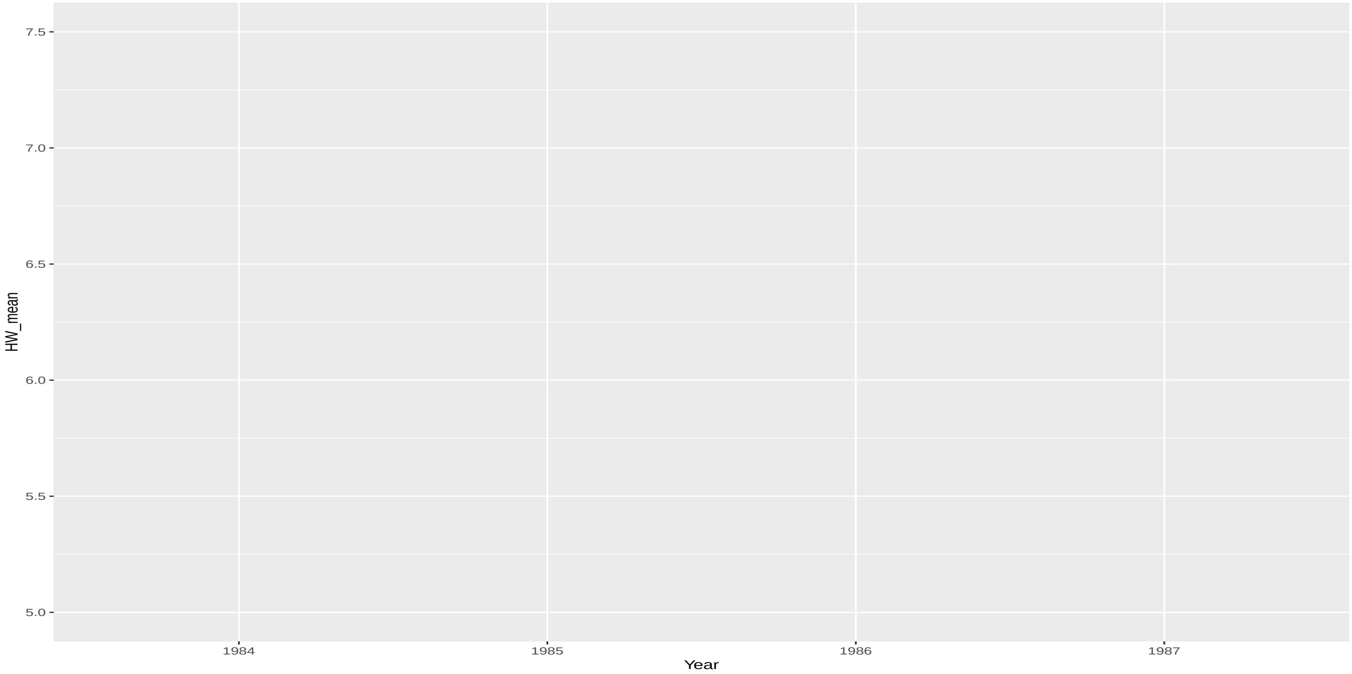
Answer the following questions.

[1] Vella, F., and Verbeek, M. (1998). Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. *Journal of Applied Econometrics*, **13**: 163–183.

1 Graphical analysis

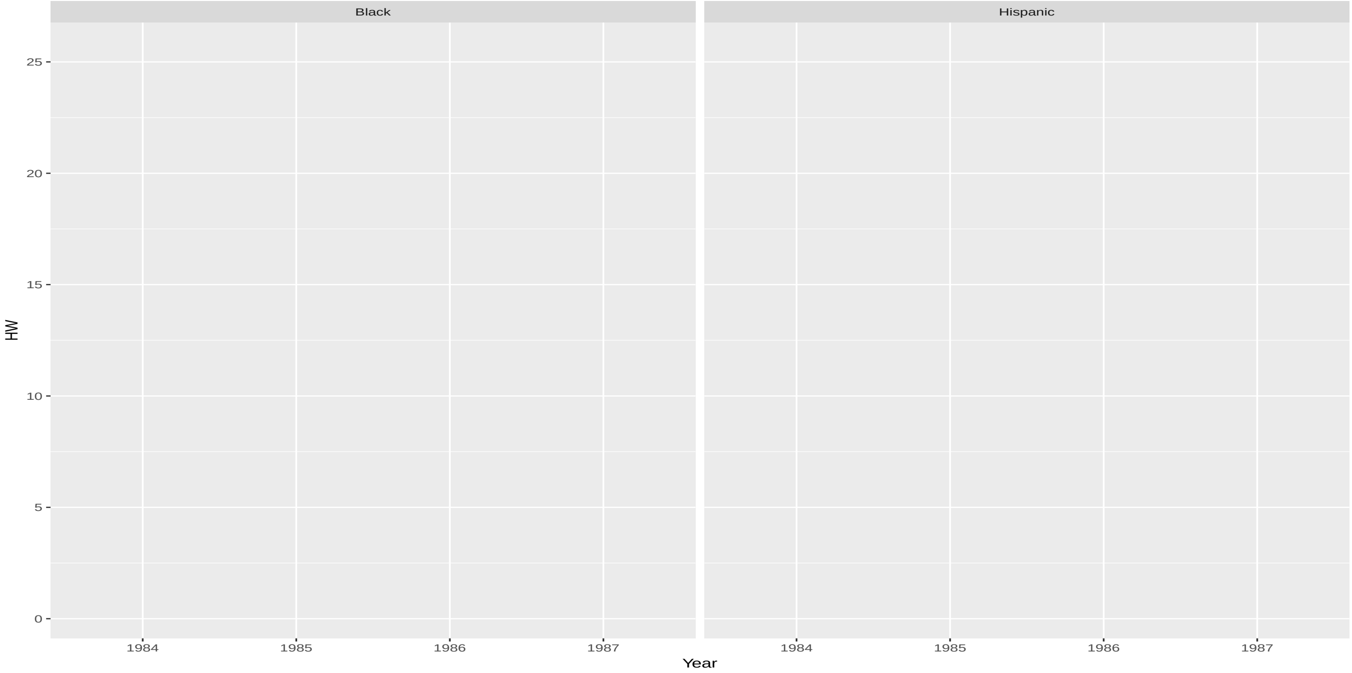
1. Use the R computer package to produce a plot of the mean HW vs $Year$ grouped by *Race*. The scale of the vertical and horizontal axes of your figure should be identical to Figure 1 below. Make sure that the $Year$ variable is treated as discrete in your plot rather than continuous (**3 marks**). Do you think that the interaction effect between $Year$ and *Race* should be included in the marginal model? Explain. (**3 marks**)

Figure 1: Plot of mean HW vs $Year$ grouped by *Race*



2. Use the R computer package to produce a scatter plot of the values of HW vs $Year$ grouped by *Race*. This plot should be like the plot presented on slide 6 of the topic 10 lecture material however the data in your plot should be grouped by *Race*, and your plot should not include the mean values of HW. The scale of the vertical and horizontal axes of your figure should be identical to Figure 2 below. Make sure that the $Year$ variable is treated as discrete in your plot rather than continuous (**3 marks**). Do you think that the variability of HW changes across the years? Explain. (**3 marks**) Do you think that the variability of HW in each year changes between the *Black* and *Hispanic* races? Explain. (**3 marks**)

Figure 2: Scatter plot of HW vs $Year$ grouped by *Race*



2 Describing the model

The researchers in the study set up the following marginal model to analyze their research questions.

$$HW_{ti} = \beta_0 + \beta_1 Y2_{ti} + \beta_2 Y3_{ti} + \beta_3 Y4_{ti} + \beta_4 H_i + \beta_5 Y2_{ti} \times H_i + \beta_6 Y3_{ti} \times H_i + \beta_7 Y4_{ti} \times H_i + \varepsilon_{ti}, \qquad (1)$$

- where HW_{ti} is the hourly wage for subject i ($i = 1, \ldots, 148$) at occasion t ($t = 1, 2, 3, 4$),
- $Y2_{ti} = 1$ if the year at occasion t, for subject i is 1985, otherwise $Y2_{ti} = 0$,
- $Y3_{ti} = 1$ if the year at occasion t, for subject i is 1986, otherwise $Y3_{ti} = 0$,
- $Y4_{ti} = 1$ if the year at occasion t, for subject i is 1987, otherwise $Y4_{ti} = 0$,
- $H_i = 1$ if the racial group of subject i is *Hispanic*, otherwise $H_i = 0$,
- β_0 is the fixed intercept,
- β_1, β_2, β_3 and β_4 are the fixed simple effects of $Y2$, $Y3$, $Y4$ and H, respectively,
- β_5, β_6 and β_7 are the fixed interaction effects of $Y2 \times H$, $Y3 \times H$ and $Y4 \times H$, respectively,
- ε_{ti} is the random error associated with measuring HW at occasion t, for subject i.

3. Interpret the intercept β_0. **(3 marks)**
4. Interpret the simple effects β_1, β_2, β_3 and β_4. **(12 marks)**
5. Interpret the interaction effects β_5, β_6 and β_7. **(9 marks)**
6. The researchers would like to express model (1) in matrix form, $\mathbf{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$, where $\mathbf{Y}_i$ represents the response vector for subject i, $\boldsymbol{X}_i$ represents a matrix, for subject i, that contains the values of the predictors associated with the fixed effects of model (1), $\boldsymbol{\beta}$ is the fixed effect vector and $\boldsymbol{\varepsilon}_i$ is the random error vector for subject i. Answer the following questions.
   (a) Write down the response vector, $\mathbf{Y}_i$, of model (1), for subject $i = 1$ (1st subject in data set). **(2 marks)**
   (b) Write down the matrix, $\boldsymbol{X}_i$, of model (1), for subject $i = 1$ (1st subject in data set). **(4 marks)**
   (c) Write down the fixed effect vector, $\boldsymbol{\beta}$, of model (1). **(2 marks)**
   (d) Write down the random error vector, $\boldsymbol{\varepsilon}_i$, of model (1), for subject i. **(2 marks)**

3 Choosing the appropriate R matrix

Define model (1A) as model (1) with R_A matrix structure,

$$\boldsymbol{R}_A = \begin{bmatrix} \theta & \theta_{12} & \theta_{13} & \theta_{14} \\ \theta_{12} & \theta & \theta_{23} & \theta_{24} \\ \theta_{13} & \theta_{23} & \theta & \theta_{34} \\ \theta_{14} & \theta_{24} & \theta_{34} & \theta \end{bmatrix}$$

- where $\theta = Var(\varepsilon_{ti})$ for all $t = 1, 2, 3, 4$,
- $\theta_{tt'} = Cov(\varepsilon_{ti}, \varepsilon_{t'i})$ for all $t = 1, 2, 3, 4$; $t' = 1, 2, 3, 4$ and $t \neq t'$.

Define model (1B) as model (1) with R_B matrix structure,

$$\boldsymbol{R}_B = \begin{bmatrix} \theta & \theta\rho & \theta\rho^2 & \theta\rho^3 \\ \theta\rho & \theta & \theta\rho & \theta\rho^2 \\ \theta\rho^2 & \theta\rho & \theta & \theta\rho \\ \theta\rho^3 & \theta\rho^2 & \theta\rho & \theta \end{bmatrix}$$

- where $\theta = Var(\varepsilon_{ti})$ for all $t = 1, 2, 3, 4$,
- $\rho = Corr(\varepsilon_{ti}, \varepsilon_{t'i})$ for all $t = 1, 2, 3, 4$; $t' = 1, 2, 3, 4$ and $t \neq t'$.

Define model (1C) as model (1) with R_C matrix structure,

$$\boldsymbol{R}_C = \begin{bmatrix} \theta & \theta\rho & \theta\rho & \theta\rho \\ \theta\rho & \theta & \theta\rho & \theta\rho \\ \theta\rho & \theta\rho & \theta & \theta\rho \\ \theta\rho & \theta\rho & \theta\rho & \theta \end{bmatrix}$$

- where $\theta = Var(\varepsilon_{ti})$ for all $t = 1, 2, 3, 4$,
- $\rho = Corr(\varepsilon_{ti}, \varepsilon_{t'i})$ for all $t = 1, 2, 3, 4$; $t' = 1, 2, 3, 4$ and $t \neq t'$.

7. The researchers decide to use the AIC and BIC criteria to choose the model with the best fit out of the models (1A), (1b) and (1C) defined above.
   (a) Use the R computer package to compute the AIC criterion for each of the models (1A), (1b) and (1C). What is the AIC value for each of the three models? **(3 marks)**
   (b) Use the R computer package to compute the BIC criterion for each of the models (1A), (1b) and (1C). What is the BIC value for each of the three models? **(3 marks)**
   (c) Which model would you choose to continue your analysis? Explain by making reference to parts (a) and (b). **(3 marks)**

4 Fixed effect estimates for your final marginal model

As their final marginal model the researchers choose model (1A). Use this model to answer the questions in this section.

8. Use the R computer package to produce a table that lists the estimates of the fixed effects in model (1A), together with their corresponding standard errors, degrees of freedom, observed test statistics and p-values. Present this table below. **(2 marks)**
9. Do you think that there is sufficient statistical evidence to suggest that the difference in mean hourly wage between *Hispanic* and *Black* male workers significantly increased from 1984 to 1985? Explain by referring to the appropriate fixed effect estimate and p-value you computed in question 8. **(5 marks)**
10. Do you think that there is sufficient statistical evidence to suggest that the difference in mean hourly wage between *Hispanic* and *Black* male workers significantly increased from 1984 to 1986? Explain by referring to the appropriate fixed effect estimate and p-value you computed in question 8. **(5 marks)**
11. Let τ_1 denote the difference in mean hourly wage between *Hispanic* and *Black* male workers in 1986 and let τ_2 denote the difference in mean hourly wage between *Hispanic* and *Black* male workers in 1985. Use the R computer package to calculate the estimate of $\tau_1 - \tau_2$. Write down this estimate. To receive full marks for this question you must use the `glht` command in R. This command will also produce a p-value which you will need to use for your answer in question 12. **(3 marks)**
12. Do you think that there is sufficient statistical evidence to suggest that the difference in mean hourly wage between *Hispanic* and *Black* male workers significantly decreased from 1985 to 1986? Explain by referring to the estimate and the p-value you computed in question 11. **(5 marks)**
13. Let η_1 denote the difference in mean hourly wage between 1987 and 1985 for *Hispanic* male workers and let η_2 denote the difference in mean hourly wage between 1987 and 1985 for *Black* male workers. Use the R computer package to calculate the estimate of $\eta_1 - \eta_2$. Write down this estimate. To receive full marks for this question you must use the `glht` command in R. This command will also produce a p-value which you will need to use for your answer in question 14. **(3 marks)**
14. Do you think that there is sufficient statistical evidence to suggest that the increase in mean hourly wage from 1985 to 1987 was significantly higher for *Hispanic* male workers than those workers who are *Black*? Explain by referring to the estimate and the p-value you computed in question 13. **(5 marks)**