**Assignment 1 is due no later than 5pm Friday, 10th of April, 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 `Assign1_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 `Assign1_StNo.tex` and is located in the Assignment 1 folder in the Topic 3 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 1 folder. The second document that you will need to submit is an R document named `Assign1_R_StNo.R` which is located in the Assignment 1 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 `HourlyWage.csv` located in the Assignment 1 folder in the Topic 3 section of LMS. The hourly wage (measured in US dollars) was recorded for 545 full-time working males each year between 1980 and 1987. These males were also classified into one of three racial groups (*Other*, *Black*, *Hispanic*). The variables of interest for Assignment 1 are:

- $Subject$: This is a factor variable that identifies the subject (working male).
- $Year$: This is a continuous variable that identifies the year the observations were recorded.
- *Race*: This is a factor (categorical) variable that identifies the racial group of the subject. It has 3 levels (*Other*, *Black*, *Hispanic*).
- $B$: This is a factor (categorical) variable that identifies whether the subject is Black or not. It has 2 levels (0 = *Not Black*, 1 = *Black*).
- $H$: This is a factor (categorical) variable that identifies whether the subject is Hispanic or not. It has 2 levels (0 = *Not Hispanic*, 1 = *Hispanic*).
- $HW$: The hourly wage which is treated as a continuous variable.

**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.**

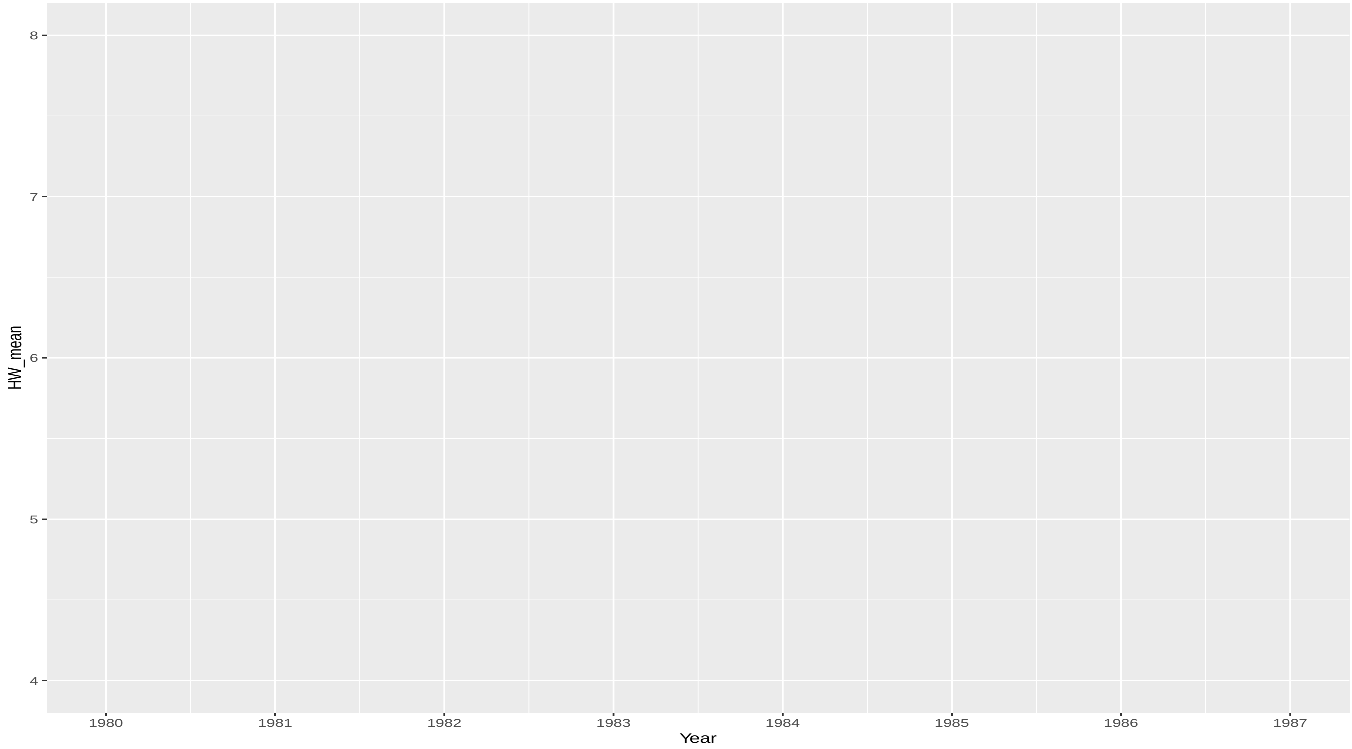
Answer the following questions.

[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

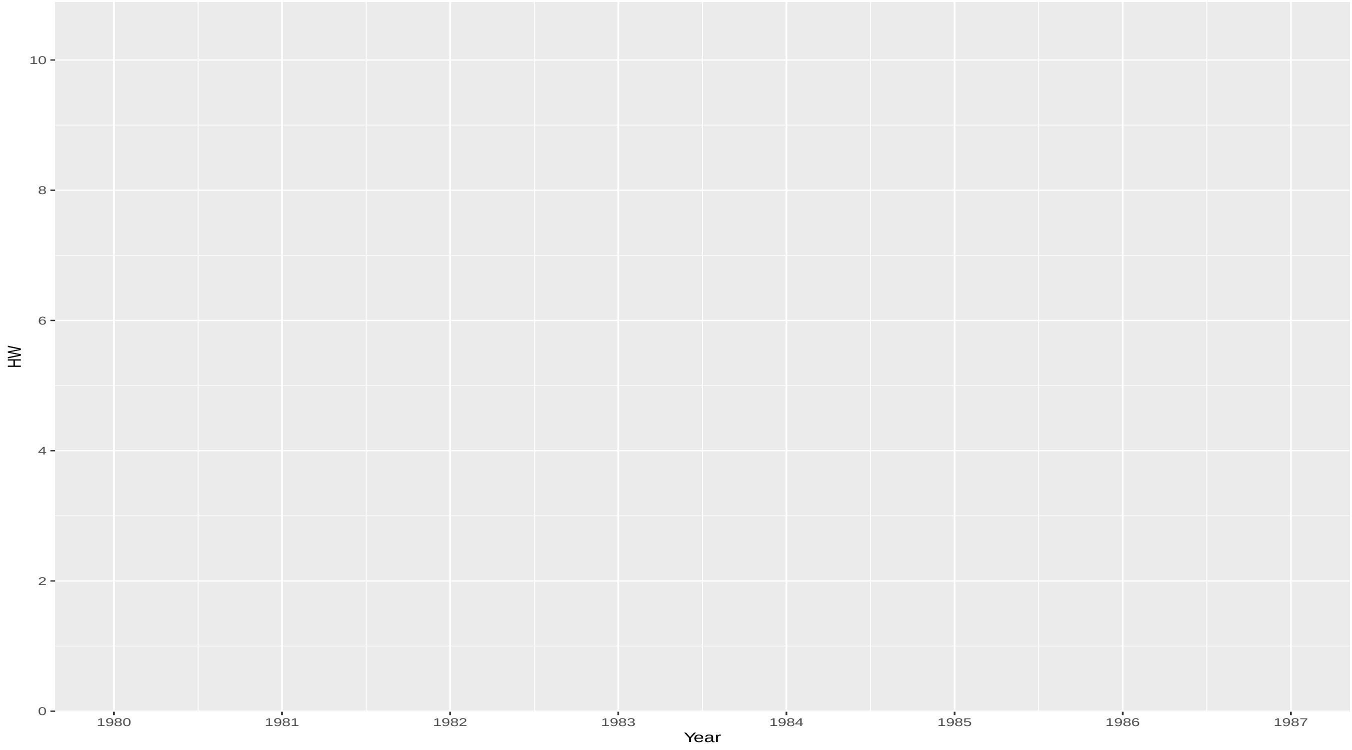
1. Use the R computer package to produce a plot of the mean $HW$ vs $Year$ grouped by $Race$. The scale of the vertical and horizontal axes of your figure should be identical to Figure 1 below (**2 marks**). Do you think that the interaction effect between $Year$ and $Race$ should be included in the linear mixed model? Explain. (**2 marks**)

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



2. Use the R computer package to produce a plot of the values of $HW$ vs $Year$ for the first six subjects (workers). The scale of the vertical and horizontal axes of your figure should be identical to Figure 2 below (**2 marks**). Do you think that the random effect of $Year$ on $HW$ should be included in the linear mixed model? Explain (**2 marks**). Do you think that the random intercept should be included in the linear mixed model? Explain. (**2 marks**)

Figure 2: Plot of $HW$ vs $Year$ for first 6 subjects



3. Use the R computer package to compute an additional variable to your data frame in R, named $Y$, which is equal to the variable $Year$ less 1980. That is $Y = Year - 1980$.

## 2 Describing the model

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

$$HW_{ti} = \beta_0 + \beta_1 Y_{ti} + \beta_2 B_i + \beta_3 H_i + \beta_4 Y_{ti} \times B_i + \beta_5 Y_{ti} \times H_i + \mu_{0i} + \mu_{1i} Y_{ti} + \varepsilon_{ti}, \tag{1}$$

- where $HW_{ti}$ is the hourly wage for subject $i$ ($i = 1, \ldots, 545$) at occasion $t$ ($t = 1, 2, \ldots, 8$),
- $Y_{ti}$ is the year at occasion $t$ for subject $i$, less 1980,
- $B_i = 1$ if the racial group of subject $i$ is *Black*, and 0 otherwise,
- $H_i = 1$ if the racial group of subject $i$ is *Hispanic*, and 0 otherwise,
- $\beta_0$ is the fixed intercept,
- $\beta_1$, $\beta_2$ and $\beta_3$ are the the fixed simple effects of $Y$, $B$ and $H$, respectively,
- $\beta_4$ and $\beta_5$ are the fixed interaction effects of $Y \times B$ and $Y \times H$ respectively,
- $\mu_{0i}$ is the random intercept specific to subject $i$,
- $\mu_{1i}$ is the random effect of $Y$ on $HW$ specific to subject $i$,
- $\varepsilon_{ti}$ is the random error associated with measuring $HW$ at occasion $t$, for subject $i$.

4. The researchers would like to express model (1) in matrix form, $\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{Y}_i$ 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, $\boldsymbol{Z}_i$ is a matrix, for subject $i$, that contains the values of the predictors associated with the random effects of model (1), $\boldsymbol{\mu}_i$ is the random effect vector for subject $i$ and $\boldsymbol{\varepsilon}_i$ is the random error vector for subject $i$. Answer the following questions.

   (a) Write down the response vector, $\boldsymbol{Y}_i$, of model (1), for subject $i$. **(2 marks)**

   (b) Write down the matrix, $\boldsymbol{X}_i$, of model (1), for subject $i$. **(4 marks)**

   (c) Write down the fixed effect vector, $\boldsymbol{\beta}$, of model (1). **(2 marks)**

   (d) Write down the matrix, $\boldsymbol{Z}_i$, of model (1), for subject $i$. **(2 marks)**

   (e) Write down the random effect vector, $\boldsymbol{\mu}_i$, of model (1), for subject $i$. **(2 marks)**

   (f) Write down the random error vector, $\boldsymbol{\varepsilon}_i$, of model (1), for subject $i$. **(2 marks)**

## 3 Testing for random effects

For model (1), the researchers choose an *unstructured* structure for the variance-covariance matrix of the random effect vector, $\boldsymbol{\mu}_i$. That is, the variance-covariance matrix of the random effect vector, $\boldsymbol{\mu}_i$, is

$$\boldsymbol{D} = \begin{bmatrix} \psi_0 & \psi_{01} \\ \psi_{01} & \psi_1 \end{bmatrix},$$

- where $\psi_0$ and $\psi_1$ denotes the variance of the random effects $\mu_{0i}$ and $\mu_{1i}$, respectively,
- $\psi_{01}$ denotes the covariance between the random effects $\mu_{0i}$ and $\mu_{1i}$.

Also for model (1), the researchers choose an *AR(1)* structure for the variance-covariance matrix of the random error vector, $\boldsymbol{\varepsilon}_i$. That is, the variance-covariance matrix of the random error vector, $\boldsymbol{\varepsilon}_i$, is

$$\boldsymbol{R} = \begin{bmatrix} \theta & \theta\rho & \theta\rho^2 & \theta\rho^3 & \theta\rho^4 & \theta\rho^5 & \theta\rho^6 & \theta\rho^7 \\ \theta\rho & \theta & \theta\rho & \theta\rho^2 & \theta\rho^3 & \theta\rho^4 & \theta\rho^5 & \theta\rho^6 \\ \theta\rho^2 & \theta\rho & \theta & \theta\rho & \theta\rho^2 & \theta\rho^3 & \theta\rho^4 & \theta\rho^5 \\ \theta\rho^3 & \theta\rho^2 & \theta\rho & \theta & \theta\rho & \theta\rho^2 & \theta\rho^3 & \theta\rho^4 \\ \theta\rho^4 & \theta\rho^3 & \theta\rho^2 & \theta\rho & \theta & \theta\rho & \theta\rho^2 & \theta\rho^3 \\ \theta\rho^5 & \theta\rho^4 & \theta\rho^3 & \theta\rho^2 & \theta\rho & \theta & \theta\rho & \theta\rho^2 \\ \theta\rho^6 & \theta\rho^5 & \theta\rho^4 & \theta\rho^3 & \theta\rho^2 & \theta\rho & \theta & \theta\rho \\ \theta\rho^7 & \theta\rho^6 & \theta\rho^5 & \theta\rho^4 & \theta\rho^3 & \theta\rho^2 & \theta\rho & \theta \end{bmatrix}.$$

- where $\theta$ denotes the constant variance of the random errors associated with subject $i$,
- $\rho$ denotes the correlation between any two random errors associated with subject $i$.

5. The researchers would like to test whether the random effect of $Y$ should be included in model (1). They decide to test, at the 5% significance level, the null hypothesis $H_0 : \psi_1 = 0$ vs the alternative hypothesis $H_1 : \psi_1 > 0$ using the REML-based likelihood ratio test $p$-value.

   (a) Write down the reference model for this test. **(1 mark)**

   (b) Write down the nested model for this test. **(1 mark)**

   (c) Use the R computer package to perform this test. What is the $p$-value for this test? **(1 mark)**

   (d) Which model would you choose (reference or nested) to continue your analysis? Explain. **(2 marks)**

6. The researchers would like to test whether the random intercept should be included in the model you chose in question 5 part (d) of section 3. They decide to test, at the 5% significance level, the null hypothesis $H_0 : \psi_0 = 0$ vs the alternative hypothesis $H_1 : \psi_0 > 0$ using the REML-based likelihood ratio test $p$-value.

   (a) Write down the reference model for this test. The reference model must be the model you chose in question 1 part (d) of section 3. **(1 mark)**

   (b) Write down the nested model for this test. **(1 mark)**

   (c) Use the R computer package to perform this test. What is the $p$-value for this test? **(1 mark)**

   (d) Which model would you choose (reference or nested) to continue your analysis? Explain. **(2 marks)**

# 4 Testing for fixed effects

7. The researchers would like to test whether the fixed interaction effects of $Y \times B$ and $Y \times H$ should be included in the model you chose for question 6 part (d) of section 3. They decide to test, at the 10% significance level, the null hypothesis $H_0 : \beta_4 = \beta_5 = 0$ vs the alternative hypothesis $H_1 : \beta_4 \neq 0$ or $\beta_5 \neq 0$ using the ML-based likelihood ratio test $p$-value.

   (a) Write down the reference model for this test. The reference model must be the model you chose in question 6 part (d) of section 3. **(1 mark)**

   (b) Write down the nested model for this test. **(1 mark)**

   (c) Use the R computer package to perform this test. What is the $p$-value for this test? **(1 mark)**

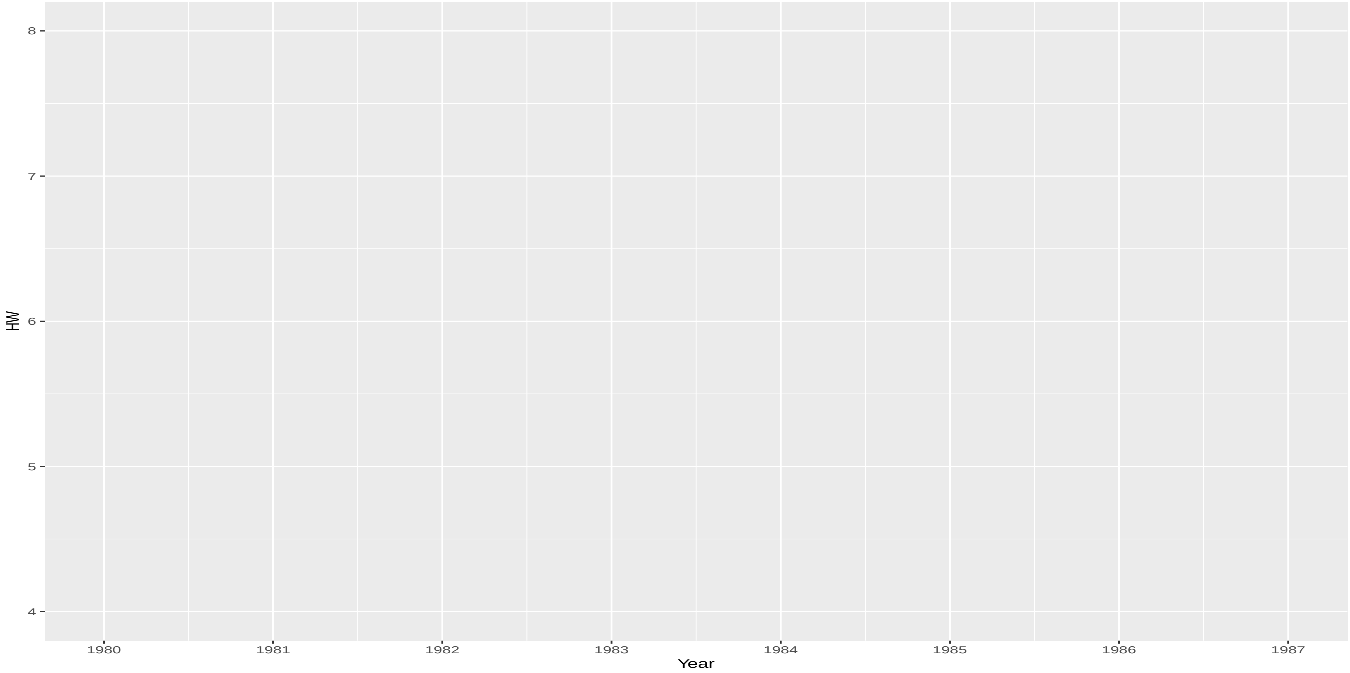
   (d) Which model would you choose (reference or nested) to continue your analysis? Explain. **(2 marks)**

# 5 Diagnostics of your final linear mixed model

Use model (1) as the final linear mixed model to answer the questions in this section and in section 6.

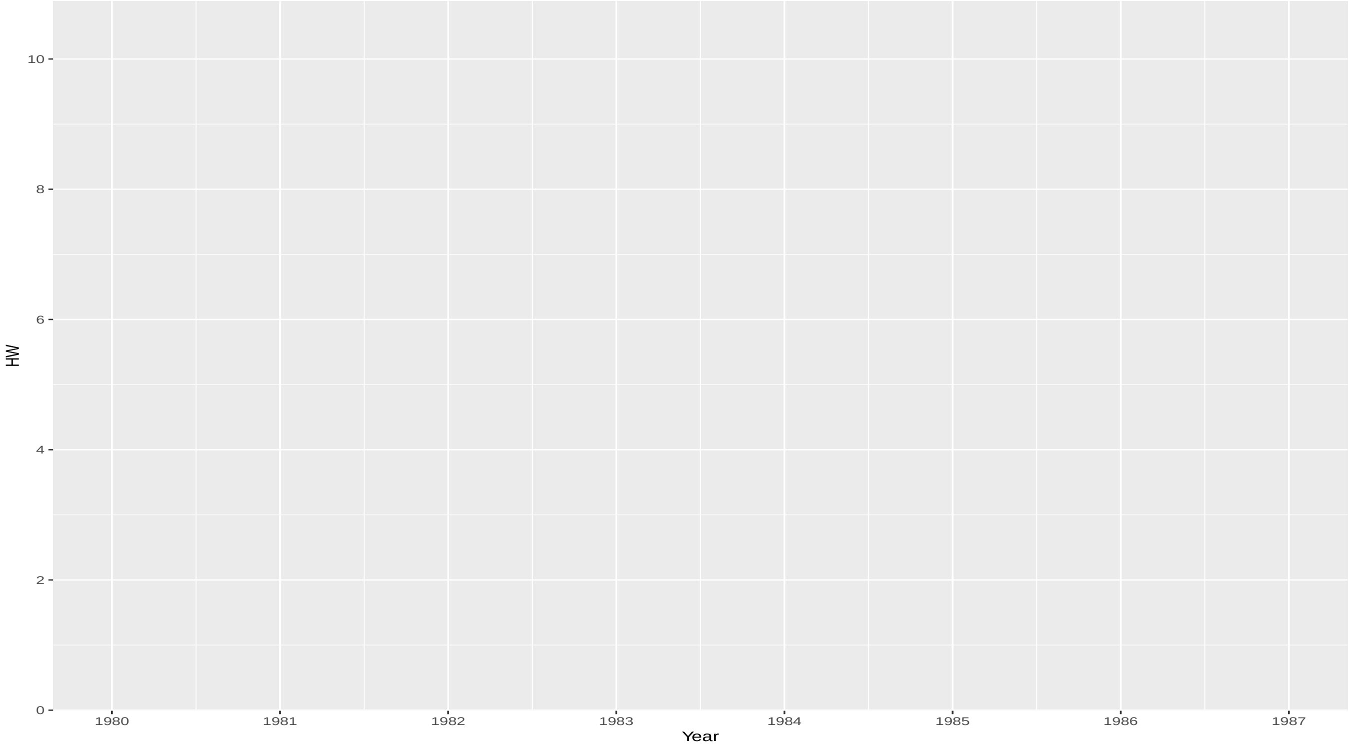
8. Use the R computer package to produce a figure that checks the agreement over time between the predicted marginal values of $HW$ (that come from fitting model (1) to the data) and the observed mean values of $HW$, for each level of *Race*. Present this figure below and make sure the scale of the vertical and horizontal axes of your figure is identical to Figure 3 below. (**2 marks**)

Figure 3: Predicted marginal and observed mean values of $HW$ as a function of $Year$



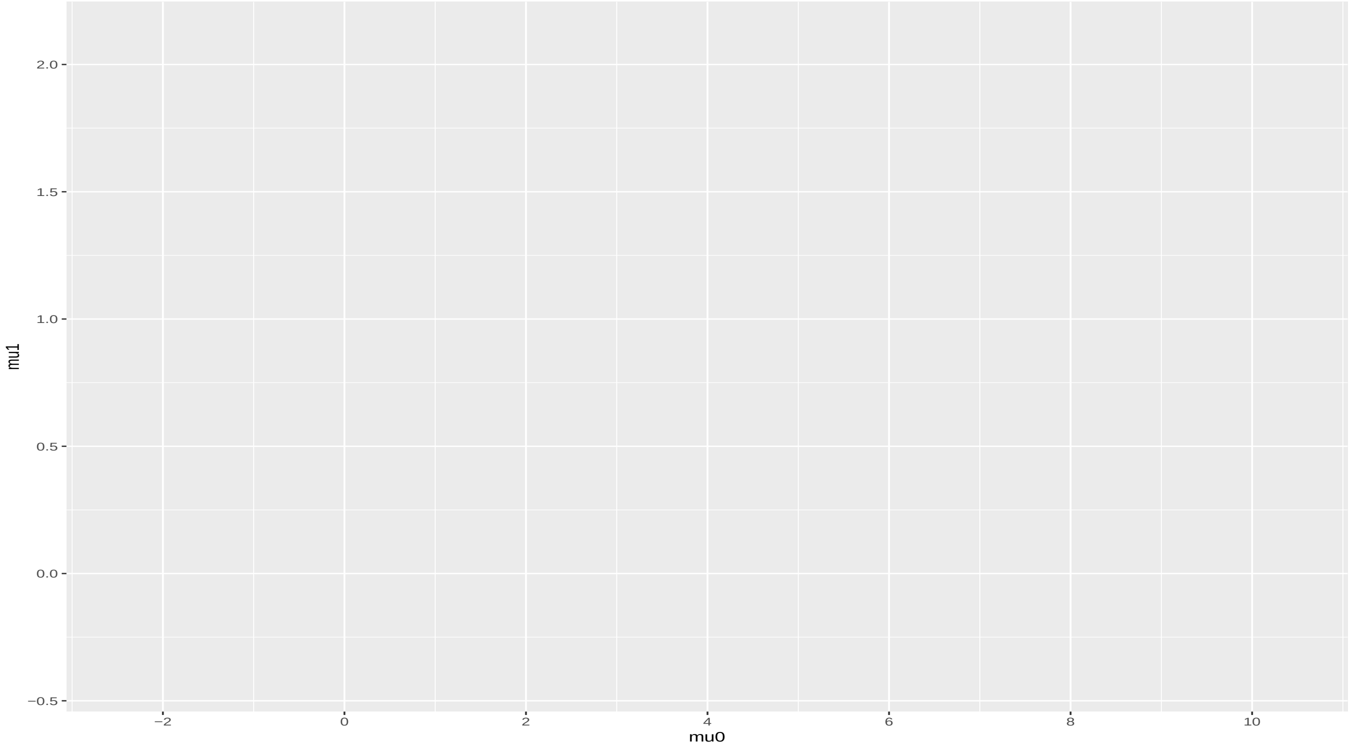
9. Use the R computer package to produce a figure that checks the agreement over time between the predicted conditional values of $HW$ (that come from fitting model (1) to the data) and the observed values of $HW$, for the first six subjects in your data set. Present this figure below and make sure the scale of the vertical and horizontal axes of your figure is identical to Figure 4 below. (**2 marks**)

Figure 4: Predicted conditional and observed values of $HW$ as a function of $Year$ for first 6 subjects



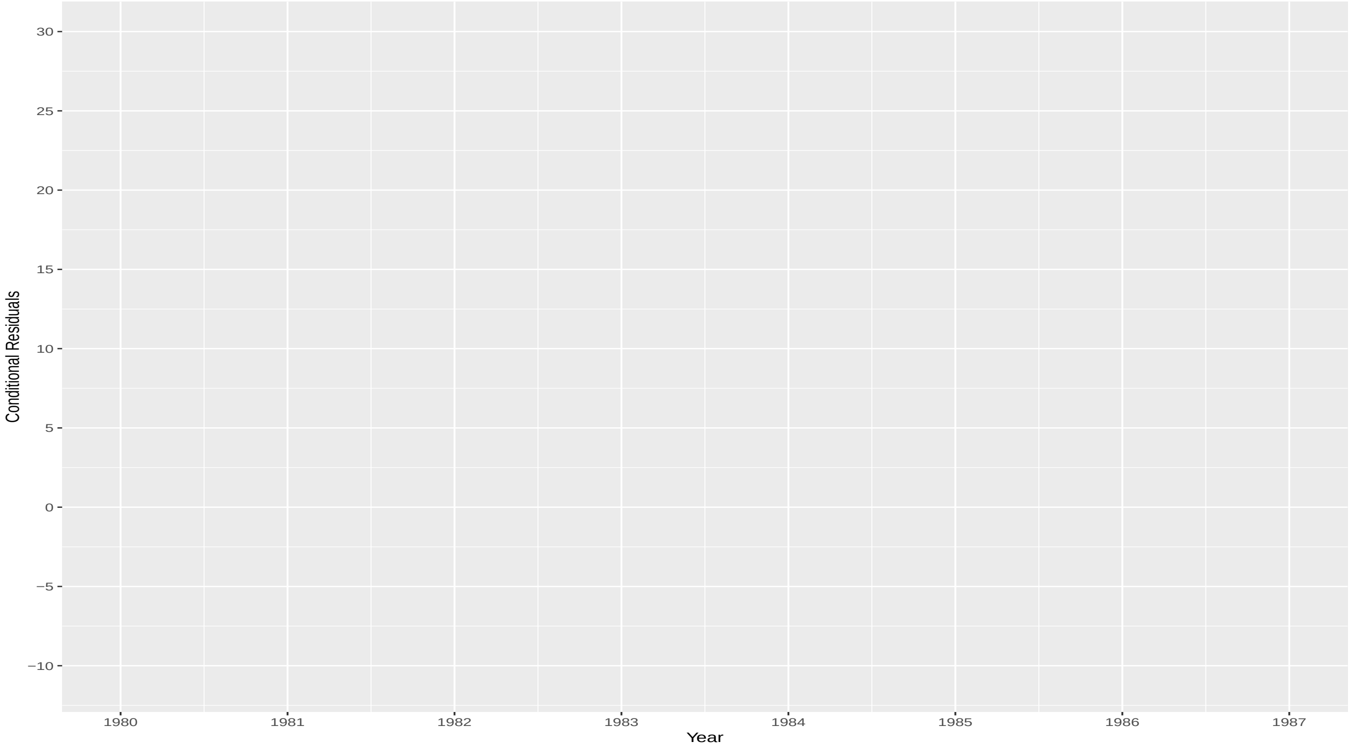
10. Use the R computer package to produce a figure that checks the distribution of the random effect vector, $\boldsymbol{\mu}_i$. Present this figure below and make sure the scale of the vertical and horizontal axes of your figure is identical to Figure 5 below. (**2 marks**)

Figure 5: Hexagonal 2-D plot of the predicted random effect vector $\hat{\boldsymbol{\mu}}_i$



11. Use the R computer package to produce a figure that checks the constant variance assumption of the random errors. Present this figure below and make sure the scale of the vertical and horizontal axes of your figure is identical to Figure 6 below. (**2 marks**)

Figure 6: Scatter plot of conditional residuals vs $Year$



# 6 Fixed effect estimates for your final linear mixed model

12. Use the R computer package to produce a table that lists the estimates of the fixed effects in model (1), together with their corresponding standard errors, degrees of freedom, observed test statistics and $p$-values. Present this table below. (**1 mark**)

13. Use the R computer package to produce a table that provides the approximate 95% confidence intervals for the fixed effects in model (1). Present this table below. (**1 mark**)

14. Interpret the estimates of $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$ and $\beta_5$. (**12 marks**)

15. Do you think that there is sufficient statistical evidence to suggest that the linear effect of $Year$ on mean hourly wage is lower for *Black* male workers than *Other* male workers? Explain by referring to the appropriate $p$-value and confidence interval you computed in questions 12 and 13, respectively. (**3 marks**)

16. Use the R computer package to calculate the estimate of the difference in mean hourly wage, in 1987, between *Other* and *Black* male workers. 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 17. (**1 mark**)

17. Do you think that there is sufficient statistical evidence that the mean hourly wage for *Other* male workers in 1987 is higher than the mean hourly wage for *Black* male workers in 1987? Explain by referring to the estimate and the $p$-value you computed in question 16. (**3 marks**)

18. Use the R computer package to calculate the estimate of the difference in mean hourly wage, in 1987, between *Hispanic* and *Black* male workers. 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 19. (**1 mark**)

19. Do you think that the mean hourly wage for *Hispanic* male workers in 1987 is higher than the mean hourly wage for *Black* male workers in 1987? Explain by referring to the estimate and the $p$-value you computed in question 18. (**3 marks**)

# 7 Symbols, equations, matrices, tables and figures in LaTeX

Table 1: Data

| $P$ | $T$ | $M$ | $E$ |
|---|---|---|---|
| 1 | 0 | 0 | 18.00 |
| 1 | 0 | 1 | 17.00 |
| 1 | 0 | 2 | 18.00 |

$$\begin{aligned} Vsae_{ti} &= \beta_0 + \beta_1\, Age_{ti} + \beta_2\, Sicdegp_i \\ &+ \beta_3\, Age_{ti} \times Sicdegp_i \\ &+ \mu_{0i} + \varepsilon_{ti}. \end{aligned} \tag{2}$$

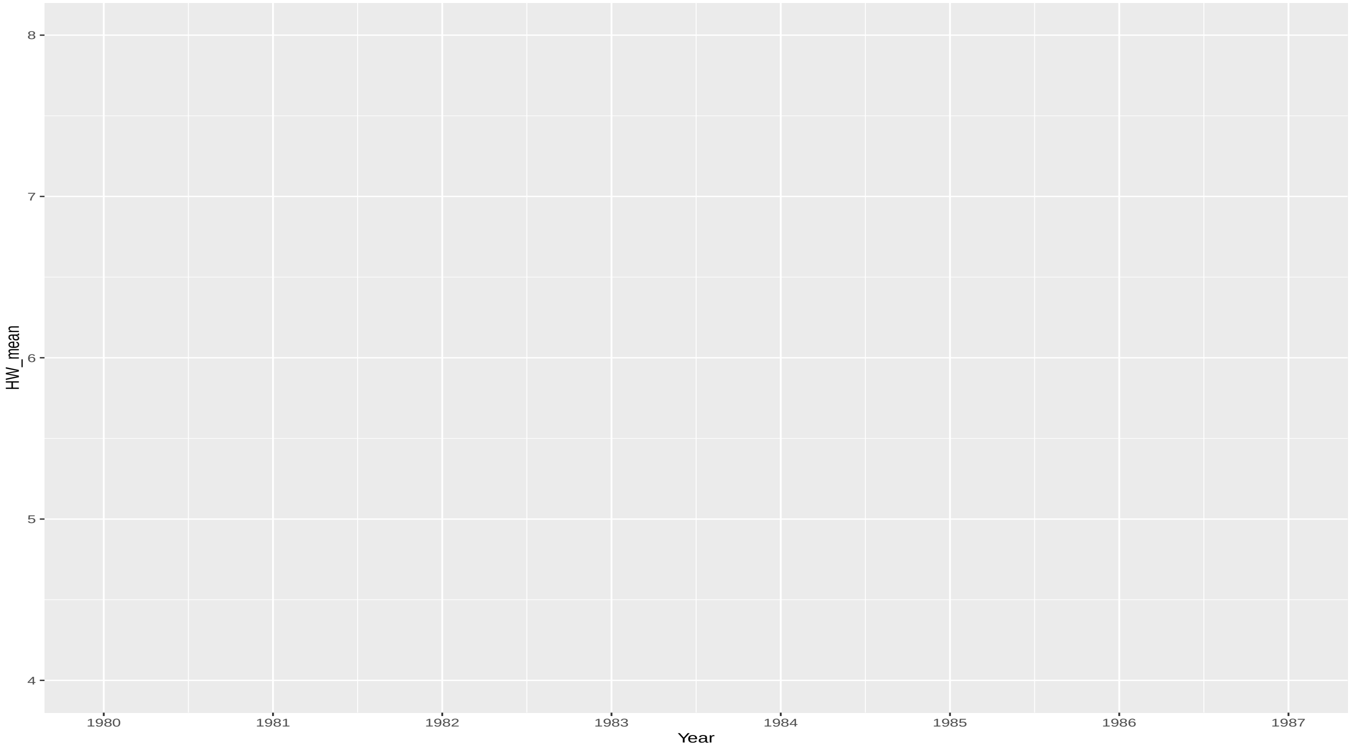
$$\boldsymbol{Y}_i = \boldsymbol{X}_i\, \boldsymbol{\beta} + \boldsymbol{Z}_i\, \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i,$$

$$\boldsymbol{Y}_i = \begin{bmatrix} Vsae_{1i} \\ Vsae_{2i} \\ Vsae_{3i} \\ Vsae_{4i} \\ Vsae_{5i} \end{bmatrix}.$$

$$\boldsymbol{Z}_i = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}.$$

$\boldsymbol{\mu}_i \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{D})$

Figure 7: Plot of mean $HW$ vs $Year$ grouped by $Race$



**Do not include this section (Section 7) in your solutions.**