Instructions:

- Assignment 2 must be submitted no later than **April 11 at 11:55 PM (Montreal time)**.
- Please submit your assignment as a SINGLE PDF file via Zone Cours. The file should include a coverpage with the full names of all team members (as it appears on HEC en ligne) and ID numbers. In addition, provide an evaluation of the breakdown of every team members' contribution to the work (in terms of a %).
- Your solution must include an appendix containing your SAS/R code (only show the code here, not the output).
- For each question, be sure to show all relevant work, clearly define all variables, describe the method/approach used, and include any **relevant** SAS/R output.
- Note that plagiarism in ANY form will NOT be tolerated (will result in a grade of 0)!
- This assignment is graded on a total of **145 points**. The points for each question are detailed below. **Failure to follow the above instructions will result in points being deducted!**

The `data_wages` dataset consists of a longitudinal study on the hourly wages of individuals working in the USA between the years 1980 and 1987. In particular, the data includes the following variables

id	individual identification number
wage	hourly wage on the log scale (i.e., log of hourly wage)
yr	year (ranging from 0 to 7, where 0 represents 1980 and 7 represents 1987)
exper	years of experience
school	years of schooling
union	binary variable indicating whether the individual is part of a union (i.e., union=1 if the individual is part of a union, and =0 otherwise)
married	binary variable indicating whether the individual is married (i.e., married=1 if the individual is married, and =0 otherwise)
manufacturing	binary variable indicating whether the individual works in manufacturing (i.e., manufacturing=1 if the individual works in manufacturing, and =0 otherwise)

The dataset consists of a modified version of data from the National Longitudinal Survey (NLS Youth Sample) and is available in the `Ecdat` *library in R.*

Throughout, let Y_{ij} represent the log hourly wage for individual i in year j, and let $\mathbf{X}_i$ denote the corresponding design matrix for individual i. For the categorical variables, you may directly use the indicator variables described above. (That is, level 0 should be the reference levels for the variables union, married and manufacturing, respectively). Treat all other variables (i.e., yr, exper and school) as continuous covariates.

The goal of the analyses considered here is to understand the factors which affect an individual's hourly wage. In particular, the main goals are to:

(i) assess whether an individual's years of experience has an effect on their hourly wage, and whether this effects depends on

  (a) whether or not they are part of a union;

  (b) whether or not they are married;

(ii) assess whether hourly wages vary with time

(iii) assess whether hourly wages are impacted by the individual's years of schooling

(iv) assess whether hourly wages differ for those in the manufacturing industry

We will consider various **linear** regression models for explaining hourly wages (on the log scale) in terms of the covariates provided. Throughout, consider the same mean specification (in terms of the fixed effects), allowing to address the research goals (i) through (iv). For all questions, **be sure to show all your work / justify your answer.**

Question 1

[15 points] Begin by carrying out an exploratory data analysis. In particular, answer the following questions, providing graphical or tabular summaries as necessary.

(a) **[1 pts]** Verify whether or not each individual responded to the survey at each of the 8 time measurements (i.e., 1980 through 1987).

(b) **[3 pts]** Summarize each covariate at each distinct point in time (i.e., in each of the years 1980 through 1987).

(c) **[3 pts]** Provide a plot depicting the evolution of individuals' wages over time by plotting the trajectory of the variable `wage` as a function of `yr` for the first 10 individuals. Comment on the results.

(d) **[8 pts]** Create a **single** graph allowing to visualize the evolution of individuals' wages (on the log scale) across time, distinguishing between those who are unionized and those who are not, and according to whether or not they work in manufacturing. That is, the graph should allow to visualize individuals' hourly wages as a function of time, while simultaneously depicting the differences between those who are unionized vs. not, and between those who work in manufacturing vs. not.

Question 2

[60 points] Begin by considering linear models with fixed effects only (i.e., no random effects), allowing to address the research goals previously listed. For all questions, **be sure to show all your work / justify your answer.**

(a) **[6 pts]** Fit a linear regression model with an exchangeable correlation structure for the random errors. Provide the estimated regression coefficients and the estimated covariance parameters. Based on the fitted model, provide an estimate for $Corr(Y_{i1}, Y_{i3}|\mathbf{X}_i)$, and an estimate for $Corr(Y_{i1}, Y_{r2}|\mathbf{X}_i, \mathbf{X}_r)$ for $i \neq r$ (be sure to show your work).

(b) **[6 pts]** Fit a linear regression model (including the same covariates), this time with an AR(1) correlation structure for the random errors. Provide the estimated regression coefficients and the estimated covariance parameters. Based on the fitted model, provide an estimate for $Cov(Y_{i1}, Y_{i3}|\mathbf{X}_i)$, and an estimate for $Cov(Y_{i1}, Y_{r2}|\mathbf{X}_i, \mathbf{X}_r)$ for $i \neq r$ (be sure to show your work).

(c) **[6 pts]** Fit a linear regression model (including the same covariates), this time with an ARH(1) correlation structure for the random errors. Provide the estimated regression coefficients and the estimated covariance parameters. Based on the fitted model, provide an estimate for $Cov(Y_{i1}, Y_{i2}|\mathbf{X}_i)$, and an estimate for $Cov(Y_{i1}, Y_{i4}|\mathbf{X}_i)$ (be sure to show your work).

(d) **[6 pts]** Carry out a likelihood ratio test to compare the model with an ARH(1) correlation structure to an ordinary linear regression model (i.e., assuming an independent correlation structure). What can you conclude?

(e) **[3 pts]** Based on AIC, which correlation structure, between independence, exchangeable, AR(1) or ARH(1), is most appropriate for the data? Justify your answer.

(f) **[3 pts]** Based on the model chosen in (e), interpret the intercept term.

(g) **[4 pts]** Based on the model chosen in (e), interpret the effect of being part of a union on an individual's hourly wage.

(h) **[9 pts]** According to the model chosen in (e), carry out an appropriate statistical test (using $\alpha = 1\%$) to address each of the research goals (i) (a) and (b).

(i) **[5 pts]** According to the model chosen in (e), carry out an appropriate statistical test (using $\alpha = 1\%$) to address research goal (iii).

(j) **[6 pts]** According to the model chosen in (e), does being married have a significant impact on an individual's hourly wage (recall, `married` is involved in an interaction)? Carry out a likelihood ratio test using $\alpha = 1\%$.

(k) **[3 pts]** How do the standard errors of the regression coefficients in the independent model compare with those obtained using the sandwich estimation approach? Comment.

(l) **[3 pts]** How do the standard errors of the regression coefficients in the model with an ARH(1) correlation structure for the random errors compare with those obtained using the sandwich estimation approach? Comment.

Question 3

[60 points] Continuing with the same fixed effects as previously explored, now consider linear mixed models (i.e., linear models with random effects) to explore the evolution of an individual's hourly wage over time. For all questions, **be sure to show all your work / justify your answer.**

(a) **[6 pts]** Begin by fitting a model with a random intercept at the individual level (assuming independent random errors). Provide the estimated fixed effects, the estimated conditional covariance matrix $\widehat{Cov}(\mathbf{Y}_i|\mathbf{X}_i, b_i)$, and the estimated marginal covariance matrix $\widehat{Cov}(\mathbf{Y}_i|\mathbf{X}_i)$ for individual $i = 1$ (here, `id`=13). How do the results here compare with those obtained from the linear regression model with an exchangeable correlation structure for the random errors explored in question 2? Explain / justify any similarities, or differences.

(b) **[8 pts]** Now consider a model which includes both a random intercept as well as a random effect for years of schooling, again at the individual level (assuming independent random errors and assuming independent random effects). Provide the estimated fixed effects and the estimated covariance parameters. Will the estimated conditional covariance, $\widehat{Cov}(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{b}_i)$, and the estimated marginal covariance, $\widehat{Cov}(\mathbf{Y}_i|\mathbf{X}_i)$, be the same for all individuals, i.e., for all $i = 1, \ldots, m$, here? Explain/justify your answer in a few sentences by deriving an expression for $Cov(Y_{ij}, Y_{ik}|\mathbf{X}_i, \mathbf{b}_i)$ and $Cov(Y_{ij}, Y_{ik}|\mathbf{X}_i)$ for arbitrary $i \in \{1, \ldots, m\}$ and $j, k \in \{1, \ldots, n_i\}$.

(c) **[5 pts]** Based on the model in part (b), is there significant variation in the individual-level schooling effect on a person's hourly wage? Carry out an appropriate statistical test, using $\alpha = 1\%$. What can you conclude, in the context of the problem, with regards to the effect of schooling on an individual's hourly wage?

(d) **[8 pts]** Now consider a model which includes both a random intercept as well as a random effect for the year (`yr`), again at the individual level (assuming independent random errors and assuming independent random effects). Provide the estimated fixed effects and the estimated covariance parameters for the random effects. Will the estimated conditional covariance, $\widehat{Cov}(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{b}_i)$, and the estimated marginal covariance, $\widehat{Cov}(\mathbf{Y}_i|\mathbf{X}_i)$, be the same for all individuals, i.e. for $i = 1, \ldots, m$, here? Explain/justify your answer by deriving an expression for $Cov(Y_{ij}, Y_{ik}|\mathbf{X}_i, \mathbf{b}_i)$ and $Cov(Y_{ij}, Y_{ik}|\mathbf{X}_i)$ for arbitrary $i \in \{1, \ldots, m\}$ and $j, k \in \{1, \ldots, n_i\}$.

(e) **[4 pts]** Provide the estimated marginal covariance matrix $\widehat{Cov}(\mathbf{Y}_i|\mathbf{X}_i)$ for $i = 1$ (here, `id`=13) for the model in (d). In a few sentences, describe and explain the pattern seen along the first row and first coloumn of the matrix.

(f) **[8 pts]** Provide the estimated marginal correlation matrix $\widehat{Corr}(\mathbf{Y}_i|\mathbf{X}_i)$ for $i = 1$ (here, `id`=13) for the model in (d). In a few sentences, describe and explain the pattern seen. How does this compare to the estimated marginal correlation obtained in the model with fixed-effects only and an AR(1) correlation structure considered in question 2 (b)? Explain any similarities or differences by deriving a general expression for $Corr(Y_{ij}, Y_{ik}|\mathbf{X}_i)$ according to model 2(b) and according to model 3 (d), for arbitrary $i \in \{1, \ldots, m\}$ and $j, k \in \{1, \ldots, n_i\}$.

(g) **[3 pts]** Accorrding to the predicted random effects, which individual has an hourly wage which increases most markedly with time? What about the least? Justify your answer in a few sentences.

(h) **[5 pts]** Now consider a model which includes both a random intercept as well as a random effect for the year (`yr`), again at the individual level (assuming independent random errors), but this time allowing for correlated random effects. Use a likelihood ratio test to compare the model in part (h) to that considered in part (d). What can you conclude?

(i) **[3 pts]** Based on model (h), what is next year's (i.e., 1988) predicted salary for individual `id`=13, supposing that they will then have 8 years of experience, but all other covariates remain the same (i.e., 14 years of schooling, still unmarried, and still working for an non-unionized company, not in manufacturing)?

(j) **[3 pts]** Based on model (h), what would the predicted salary be for an individual entering the job market in the next year (i.e., 1988) after having completed 5 years of schooling, given that this person has no work experience, is not married, and will be working for a manufacturing company where the employees are part of a union?

(k) **[7 pts]** Based on model (h), interpret the main effect of the variable `union` and the effect of the variable `yr`.

Question 4

[10 points] Find a research article from a peer-reviewed journal that involves the analysis of correlated data in a specific context. Answer the questions below. A maximum of 2 pages is allotted for this question.

(a) Provide a high level summary of the paper.

(b) Discuss the main advantages or strong points of the methodology considered in the paper, given the specific context.

(c) Discuss any disadvantages or possible limitations of the work.