5.1 **Gas fixing cartel in the Gaspe Peninsula**: Many mayors and prefects of the Gaspe Peninsula asked the *Régie de l'énergie*, the Quebec government agency in charge of overseeing energy prices, to investigate potential overpricing of gasoline in their region. To replicate their analysis, the following data were scraped from the organism's website for the period 2014–2019. The data include

- `region`: Quebec administrative region, one among Bas-Saint-Laurent (1), Saguenay-Lac-Saint-Jean (2), Capitale-Nationale (3), Mauricie (4), Estrie (5), Montréal (6), Outaouais (7), Abitibi-Témiscamingue (8), Côte-Nord (9), Nord-du-Québec excluding Nunavik (10), Gaspésie-Îles-de-la-Madeleine (11), Chaudière-Appalaches (12), Laval (13), Lanaudière (14), Laurentides (15), Montérégie (16) et Centre-du-Québec (17).
- `date`: day of measurements of minimum weekly retail price and average price for fuel, formatted `yyyy-mm-dd`.
- `pmin`: minimum sale price calculated by Régie de l'énergie, including taxes and transportation costs.
- `ave`: average retail price of retailers, survey-based.

Perform a longitudinal data analysis to assess whether the retailers margin of profit for Gaspésie-Îles-de-la-Madeleine is significantly higher than elsewhere through the use of a one-way ANOVA model that accounts for within-region correlation. *Indication: in SAS , use the option `ddfm=satterth` for the model degrees of freedom with the `mixed` procedure.*

(a) Plot the time series of (a) the average price and (b) the difference between average price and minimum retail price for each region and comment on the observed differences.

(b) Select an appropriate covariance model to account for the within-region dependence. Potential choices are (a) independence (diagonal), (b) AR(1), (c) compound symmetry and (d) unstructured. Justify your choice.

(c) Report the standard errors for the estimed mean retail margin of the Gaspésie-Îles-de-la-Madeleine region for the ordinary linear regression assuming independence and the first-order autoregressive model. State which is highest and the underlying reason for this.

(d) Compute pairwise differences of retailers margin of profit between Gaspésie-Îles-de-la-Madeleine and each other region accounting for within-region correlation; which are statistically significant at level 5%?

5.2 **Teaching to read**: the data used in this study is from

> J. Baumann, N. Seifert-Kessell, L. Jones (1992), *Effect of Think-Aloud Instruction on Elementary Students' Comprehension Monitoring Abilities*, Journal of Reading Behavior, **24** (2), pp. 143–172.

Researchers conducted a study to determine the efficiency of three learning methods for reading. The sample consists of 66 fourth-grade students from an elementary school. The students, 32 girls and 34 boys, were randomly split between three groups. Interest lies in improvement over the default method, directed reading (DR). Two tests were administered before and after the experiment to monitor the effectiveness of the methods; to make these comparable, they were rescaled so that a total of 1 means perfect score.

The data contains information about

- `group`: experimental group, one of directed reading-thinking activity (`DRTA`), think-aloud (`TA`) and directed reading group (`DR`).
- `mpre`: average pre-test prediction score (standardized) for average of standardized error detection test and comprehension monitoring questionnaire.
- `mpost`: same as `mpre`, but post-test score.
- `dpp`: difference between post-intervention results and pre-intervention results, `mpost-mpre`.

In this first part, we are interested in the improvement in scores and two models are fitted to assess this.

(a) In their paper, Baumann *et al.* run a one-way analysis of variance (ANOVA) for "pre-tests" `mpre` with the `group` factor. Explain what is the purpose of doing such a test in the context of the study.

(b) Fit a one-way ANOVA for `dpp` = `mpost` − `mpre` with `group` as factor (Model 5.2.1). Write down the model equation in terms of `mpost` and show that it is a special case of a linear regression with an offset.

(c) Compare the one-way ANOVA model for `dpp` with `group` to a linear regression model with `mpost` as response and `mpre` and `group` as covariates (Model 5.2.2). Given the output of the latter, is the ANOVA model adequate?

Compiled 15/10/2020 at 10:05

Justify your answer.

(d) Transform the dataset from wide to long-format; the latter is more suitable for longitudinal studies. In addition to `group`, your data should contain the following columns
- `id`: unique student identification number.
- `score`: average result for evaluation.
- `test`: categorical variable, one of `mpost` or `mpre` indicating whether the score reported is pre-test or post-test.

Table 1 contains the first 10 lines of the expected dataset.

group	test	score	id
DR	mpre	0.23	1
DR	mpost	0.27	1
DR	mpre	0.35	2
DR	mpost	0.42	2
DR	mpre	0.41	3
DR	mpost	0.24	3
DR	mpre	0.57	4
DR	mpost	0.39	4
DR	mpre	0.67	5
DR	mpost	0.56	5

Table 1: First ten lines of the Baumann data in long format

Compare two models for `score` as function of `group` and `test` and an interaction term between the two, but with different covariances for the two scores of students:
- Model 5.2.3 assumes a compound symmetry model;
- Model 5.2.4 assumes an unstructured covariance model.

Explain what is the fundamental difference between Model 5.2.2 and Model 5.2.3-5.2.4.

(e) Write down the covariance matrix implied by Model 5.2.3 and report the estimated correlation between the pre-test and the post-test scores for any student.

(f) Using the output of Models 5.2.3 and 5.2.4, test whether the variability of the mean pre-test and post-tests is the same. Specifically, write down the name of the test, the numerical value of the statistic and the p-value before concluding.

(g) Since the data are longitudinal, one could consider fitting, in addition to Model 5.2.3–5.2.4, a first-order autoregressive covariance model, AR(1). Would it be useful in this case? Justify your answer.

(h) Up till now, we assumed that the covariance matrix of the pre- and post-intervention scores is the same for all students in Models 5.2.3 and Models 5.2.4. One could however postulate that the parameters of the covariance matrix in Model 5.2.3 differs from one reading teaching method to the next. Is this hypothesis supported by the data?

(i) Use Model 5.2.4 to determine if the teaching methods `DRTA` and `TA` significantly improve over the standard teaching method of directed reading `DR`.

5.3 **Tolerance of teenagers towards delinquency**: The data come from the American National Longitudinal Survey of Youth, which started in 1997. This longitudinal study follows a sample of young Americans born between 1980 and 1984. A total of 8984 participants aged 12 to 17 were interviewed for the first time in 1997 and the cohort has been followed up 15 times till now.

We consider 16 individuals who responded to the first five interview waves between age 11 to 15 years old, with annual follow-up. Of particular interest are questions related to attitude towards delinquency. Teens were asked to indicate their attitude towards (a) cheating on an exam (b) purposely destroying someone's goods (c) smoking

marijuana (d) stealing an object worth less than five dollars (e) hitting or threatening someone without reason (f) drug consumption (g) break in a building or a vehicle to steal (h) sell hard drugs and (i) steal items worth more than \$50. Each score was recorded on a Likert scale of four ranging from very bad (1) to completely acceptable (4). The provided data, `tolerance`, includes the following variables,

- `id`: integer for identification of the participant.
- `age`: age of the participant at follow-up.
- `tolerance`: average score for the nine questions on tolerance towards delinquency.
- `sex`: binary indicator, unity for men and zero for women.
- `exposure`: average score of participant at age 11 to delinquent behaviour among acquaintances, an estimate of the participation of friend(s) in each activity (a) to (i) described above.

(a) Report and interpret descriptive statistics for the variables `tolerance`, `sex` and `exposure`.

(b) Evaluate graphically the relationship between `tolerance` and each of `sex`, `exposure` and `age`. Briefly describe your findings.

(c) Produce a spaghetti plot of tolerance to delinquency trajectory as a function of the age of participants and comment the output.

(d) Using only the data for age 11, fit a linear regression model explaining tolerance to delinquency behaviour as a function of sex and exposure, i.e,

$$\texttt{tolerance}_i = \beta_0 + \beta_1 \texttt{sex}_i + \beta_2 \texttt{exposure}_i + \varepsilon_i, \qquad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), i = 1, \ldots, 16. \tag{M$_1$}$$

Comment on the effect of each of these variables.

(e) Using the full data set and assuming that the observations are independent, fit the model

$$\texttt{tolerance}_{ij} = \beta_0 + \beta_1 \texttt{sex}_{ij} + \beta_2 \texttt{exposure}_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), i = 1, \ldots, 16; j = 1, \ldots, 5. \tag{M$_2$}$$

(f) Taking into account the within-subject correlation assuming a constant correlation between every year, fit the model

$$\begin{aligned} \texttt{tolerance}_{ij} &= \beta_0 + \beta_1 \texttt{sex}_{ij} + \beta_2 \texttt{exposure}_{ij} + \beta_3 \texttt{age}_{ij} + \varepsilon_{ij}, \\ \boldsymbol{\varepsilon}_i &\overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_5, \Sigma_i), \boldsymbol{\Sigma}_i \sim \mathsf{CS}, i = 1, \ldots, 16; j = 1, \ldots, 5. \end{aligned} \tag{6.3.1}$$

i. Interpret the effect of the variables in the model and comment on your results.

ii. Compute the estimated correlation between individual `tolerance` scores between measurements at age 11/12 and 11/15.

(g) Taking into account the within-subject correlation between the five measurements by assuming that the errors follow a first order autoregressive process, fit the model

$$\begin{aligned} \texttt{tolerance}_{ij} &= \beta_0 + \beta_1 \texttt{sex}_{ij} + \beta_2 \texttt{exposure}_{ij} + \beta_3 \texttt{age}_{ij} + \varepsilon_{ij}, \\ \boldsymbol{\varepsilon}_i &\overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_5, \Sigma_i), \boldsymbol{\Sigma}_i \sim \mathsf{AR}_1, i = 1, \ldots, 16; j = 1, \ldots, 5. \end{aligned} \tag{6.3.2}$$

i. Interpret the parameters and comment on your results.

ii. Write down the equation of the fitted model 6.3.2.

iii. Identify all of the parameters of the covariance matrix matrix

iv. Compute the estimated correlation between individual `tolerance` scores between measurements at age 11/12 and 11/15. Compare your results with the estimated correlation of the compound symmetry model.

(h) Assuming measurements at every time point are independent, fit the model

$$\texttt{tolerance}_{ij} = \beta_0 + \beta_1 \texttt{sex}_{ij} + \beta_2 \texttt{exposure}_{ij} + \beta_3 \texttt{age}_{ij} + \varepsilon_{ij}, \qquad \boldsymbol{\varepsilon}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_5, \sigma^2 \mathbf{I}_5), i = 1, \ldots, 16; j = 1, \ldots, 5, \tag{M_5}$$

and interpret the effect of the variables.

(i) Assuming measurements at every time point are independent, fit the model

$$\begin{aligned} \texttt{tolerance}_{ij} &= \beta_0 + \beta_1 \texttt{sex}_{ij} + \beta_2 \texttt{exposure}_{ij} + \beta_3 \texttt{age}_{ij} + \varepsilon_{ij}, \\ \boldsymbol{\varepsilon}_i &\overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_5, \sigma^2 \mathbf{I}_5), i = 1, \ldots, 16; j = 1, \ldots, 5, \end{aligned} \tag{6.3.3}$$

Which of Models 6.3.1,6.3.2 and 6.3.3 would you choose? Justify your answer.

(j) List all of the hypothesis of Models 6.3.1,6.3.2 and 6.3.3.

Compiled 15/10/2020 at 10:05