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.

Solution

It is clear that the average price oscillates (non-stationarity), so it makes no sense to model the average price in each region. In contrast, the weekly average margin of profit appears steady, even if there are disparities between regions.

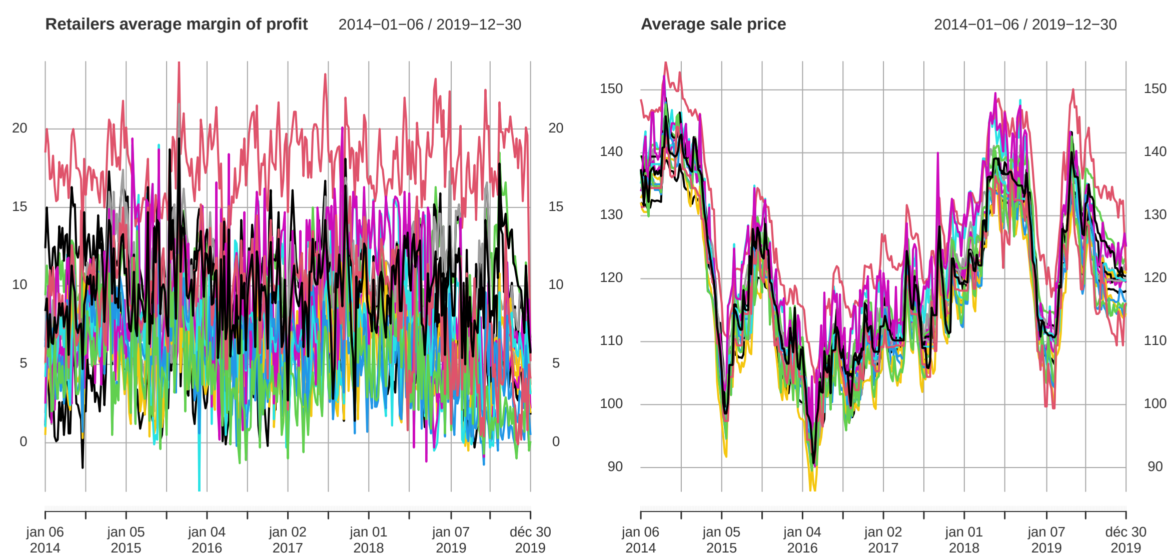

Figure 1: Weekly average retailers margin of profit (in CAD cents) and weekly average per region of gasoline sale price (in CAD cents) according to Régie de l'énergie data.

(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.

Solution

There is clear evidence that records are not independent from one week to the next (likelihood ratio tests shown in the default SAS output attest this). The unstructured model cannot be fitted because there are 313 time observations for each of the 17 series, meaning more parameters than the total number of observations recorded. The first-order autoregressive model is the only model that makes sense in the present context.

(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.

Solution

It is easy to obtain the standard errors by putting the Gaspé as reference category. We get the ratio $(0.1672/0.2625)^2 = 0.4057$. Correlated observations carry less information than independent measurements.

(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%?

Solution

There is evidence of heteroscedasticity between regions, so one should fit a different AR(1) model to each region (akin to a Welch test since each variance is computed separately, but we need to account for the time dependence).

- If we assume the parameters are all the same, then the only pairwise difference which is not statistically difference relative to the Gaspésie-îles-de-la-Madeleine is Côte-Nord (p value of 0.2008).
- If we fit a different AR(1) model, all but Côte-Nord (p value of 0.3566)have lower margins; for these two regions, we cannot conclude that the retailer margins differ.

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.

Solution

A one-way ANOVA tests whether the mean in each group is the same to make sure that no group is stronger/weaker before the experiment starts.

(b) Fit a one-way ANOVA for `dpp` = `mpost` − `mpre` with `group` as factor (Model 5.2.1). Is there a difference between groups?

Solution

Yes, the change in mean pre- versus post-experiment is non-zero; the type 3 F-test for checking whether the mean increases are the same for all three groups is 15.986, to be compared to a $\mathsf{F}(2,63)$ distribution; this yields a negligible p-value.

(c) Write down the model equation in terms of `mpost` and show that it is a special case of a linear regression with an offset. Hence 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). Is the one-way ANOVA an adequate simplification of Model 5.2.2? Justify your answer.

Solution

The model equation is

$$\texttt{mpost} = \beta_0 + \beta_1\texttt{DRTA} + \beta_2\texttt{TA} + \texttt{mpre} + \varepsilon \tag{5.2.1}$$

$$\texttt{mpost} = \beta_0 + \beta_1\texttt{DRTA} + \beta_2\texttt{TA} + \beta_3\texttt{mpre} + \varepsilon \tag{5.2.2}$$

If we set $\beta_3 = 1$ in Model 5.2.1, we recover Model 5.2.2: we can test whether $\beta_3 = 1$ by looking at the confidence interval. Since the 95% Wald confidence interval, $0.6017 \pm 1.96 \times 0.08327 = [0.438, 0.765]$., does not include 1, we reject the null that the one-way ANOVA (Model 5.2.1) is an adequate simplification of the linear regression model (Model 5.2.2).

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 dataset in long format.

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

We will consider additionally 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.

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

Solution

The difference between Model 5.2.3–5.2.4 and Model 5.2.2 is that now we consider `mpre` as random. These scores are used to estimate the (common) variance of the observations. We also explicitly model the correlation within-student.

(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.

Solution

Specifically, the covariance is between `mpre` and `mpost` is

$$\mathsf{Cov}\,(\boldsymbol{\varepsilon}_i) = \begin{pmatrix} \sigma^2+\tau & \tau \\ \tau & \sigma^2+\tau \end{pmatrix}, \qquad \widehat{\mathsf{Cor}}(\boldsymbol{\varepsilon}_i) = \begin{pmatrix} 1 & \dfrac{\widehat{\tau}}{\widehat{\sigma}^2+\widehat{\tau}} \\ \dfrac{\widehat{\tau}}{\widehat{\sigma}^2+\widehat{\tau}} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.6715 \\ 0.6715 & 1 \end{pmatrix}.$$

(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.

Solution

- We use a likelihood ratio test for the comparison, since the model with a compound symmetry and unstructured covariance within individual are nested.
- Twice the negative log-likelihood -2ℓ for Model 5.2.3 is -174.1, that of Model 5.2.4 is -175.7, corresponding to a difference of 1.6.
- The null distribution is χ^2_1; we reject the null of homoscedasticity if $D > 3.84$, i.e. if the likelihood ratio test exceeds the 95% percentile of the null distribution (the corresponding p-value is 0.206).

(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, $\mathsf{AR}(1)$. Would it be useful in this case? Justify your answer.

Solution

No, because it would yield the same covariance as a compound symmetry Model 2.3 (the within-group covariance matrix is 2×2).

(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?

Solution

No. We can fit a different compound symmetry model in each group. The compound symmetry model 5.2.3 is nested in this model, since we can obtain it by setting $\mathscr{H}_0 : \tau_{\texttt{DR}} = \tau_{\texttt{TA}} = \tau_{\texttt{DRTA}}$ and $\sigma^2_{\texttt{DR}} = \sigma^2_{\texttt{TA}} = \sigma^2_{\texttt{DRTA}}$. The log-likelihood of the complex model is -175.7; the likelihood ratio statistic is 1.6, to be compared with a χ^2_4 distribution. We fail to reject the null hypothesis that the joint compound symmetry model is an adequate simplification of the model in which the parameters of the compound symmetry model differ between groups, meaning that the added complexity does not lead to a significant improvement in fit.

(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`.

Solution

This amounts to testing if the change pre- versus post-intervention is the same for all group. According to the model, the interaction term `test*group` is statistically significant. The improvement for `DRTA` versus `DR` method for post-intervention score is 0.1457 (0.032) and that of the `TA` method (relative to the `DR` method) is 0.1656 (0.032). The predicted score for `DR`, `DRTA` and `TA` for post-intervention scores are respectively 0.3629, 0.4786 and 0.475. The `DR` method score drops between pre- and post-intervention score by 0.141, whereas the other methods yield equivalent scores.

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`.

Solution

Both `exposure` and `sex` are fixed for all time points, so descriptive statistics must be calculated for a single age.

- `sex`: 43.75% of the candidates are male
- `exposure`: average (std. error) of 1.19 (0.08), values ranging between 0.81 and 1.99.

Tolerance changes over time, so we use all of the 80 measurements to compute descriptive statistics. The average (std. error) of tolerance is 1.62 (0.055) and the values range from 1 to 3.46.

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

Solution

The plots are presented in Figure 2. There ain't much difference between men and women, but the former are on average more tolerant. Two of the reported tolerance are abnormally high. Tolerance seems to increase linearly with age; it also increases with exposure, but the effect is more pronounced between the small and higher exposure levels. There is more heterogeneity at age 15 than at other ages.

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

Solution

For most individuals, the tolerance paths in Figure 3 increases over time. The spaghetti plot shows that the two largest tolerance scores, which lie outside of the whiskers in the left panel of Figure 2, belong to the same man.

(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,

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

Comment on the effect of each of these variables.

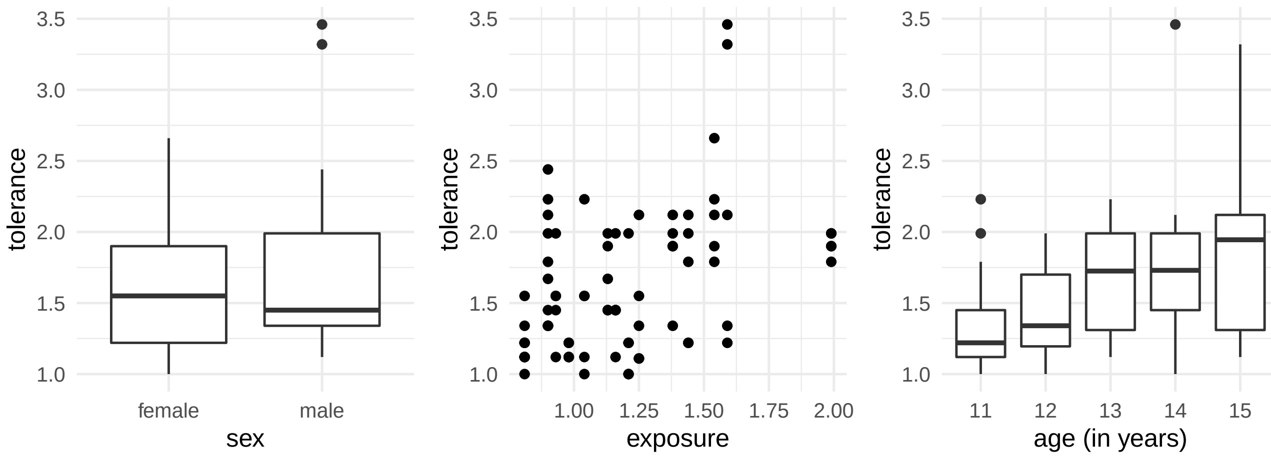

Figure 2: Box-and-whiskers plots and scatterplot of response `tolerance` as a function of explanatories.

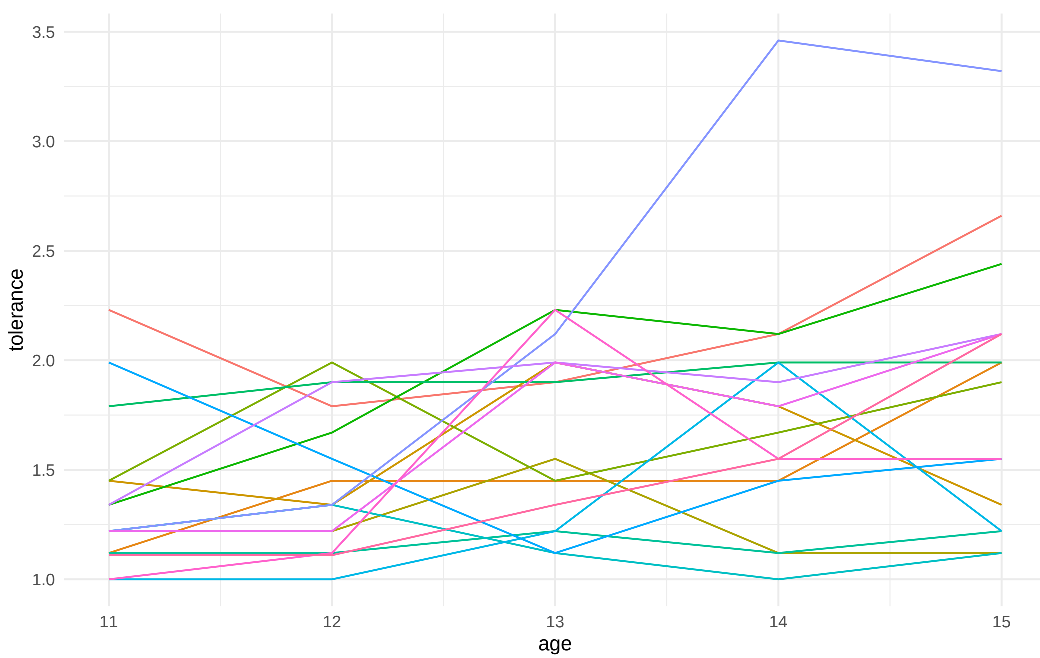

Figure 3: Spaghetti plot of trajectories for tolerance score of 16 individuals in study.

Solution

- $\widehat{\beta}_0 = 0.91(0.38)$ is the average at age 11 of women who are not exposed to delinquent behaviour.
- The tolerance score is $\widehat{\beta}_1 = 0.05(0.18)$ units lower for men than women, but this different is not significant.
- When the average exposure score increases by one, the estimated increase in average tolerance is $\widehat{\beta}_2 = 0.36(0.29)$.

The differences between sex and due to increased exposures are not significant. This is due to the small sample size, which results in low power.

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

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

Solution
The parameter estimates are $\widehat{\beta}_0 = 0.63(0.21)$, $\widehat{\beta}_1 = 0.23(0.1)$ and $\widehat{\beta}_2 = 0.74(0.16)$. The individual t-tests for $\beta_i = 0$ ($i=0,1,2$) are invalid, even if the associated p-values are less than 0.05. This is due to false assumption of independence, whereas the correlation between observations is strong.

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

$$\begin{aligned}\texttt{tolerance}_{ij} &= \beta_0 + \beta_1 \texttt{sex}_{ij} + \beta_2 \texttt{exposure}_{ij} + \beta_3 \texttt{age}_{ij} + \varepsilon_{ij},\\ \boldsymbol{\varepsilon}_i &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_5, \Sigma_i), \boldsymbol{\Sigma}_i \sim \mathsf{CS}, i=1,\ldots,16; j=1,\ldots,5.\end{aligned} \tag{5.3.1}$$

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.

Solution
The coefficient estimates are $\widehat{\beta}_0 = -1.07(0.45)$, $\widehat{\beta}_1 = 0.23(0.13)$, $\widehat{\beta}_2 = 0.74(0.20)$, $\widehat{\beta}_3 = 0.13(0.03)$, $\widehat{\sigma}^2 = 0.127$ and $\widehat{\tau} = 0.035$.

- The interpret is meaningless in the context because it would correspond to age at birth.
- The estimated effect of sex is a 0.229 increase for men relative to women *ceteris paribus*, but it is not significant.
- The effect of exposure is statistically significant (with a p-value of 0.0024). People more exposed to deliquency behaviour are more tolerant, with an estimate of 0.745 increase when the average exposure score increases by one, age and sex being constant.
- The average tolerance score, everything else held constant, increases by about 0.13 each year. This effect is strongly significant.
- The correlation between individual (assumed constant for each year) is $\widehat{\rho} = 0.22$ and is statistically significant (the likelihood ratio test comparing Models 5.3.1 and 5.3.3 amounts to testing $\tau = 0$); the p-value is 0.0257, indicating that the compound symmetry model leads to significantly better fit than the vanilla multiple linear regression model.

(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

$$\begin{aligned}\texttt{tolerance}_{ij} &= \beta_0 + \beta_1 \texttt{sex}_{ij} + \beta_2 \texttt{exposure}_{ij} + \beta_3 \texttt{age}_{ij} + \varepsilon_{ij},\\ \boldsymbol{\varepsilon}_i &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_5, \Sigma_i), \boldsymbol{\Sigma}_i \sim \mathsf{AR}_1, i=1,\ldots,16; j=1,\ldots,5.\end{aligned} \tag{5.3.2}$$

i. Interpret the parameters and comment on your results.
ii. Write down the equation of the fitted model 5.3.2.
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.

Solution
The interpretation of the β's is the same as for the compound symmetry model.

The fitted model is

$$\widehat{\texttt{tolerance}} = -1.05 + 0.23\texttt{sex} + 0.77\texttt{exposure} + 0.13\texttt{age}.$$

The estimated lag-one correlation is $\widehat{\rho} = 0.54$ and the estimated variance $\widehat{\sigma}^2 = 0.17$. The correlation decays geometrically in the AR(1) model, contrary to the compound symmetry model. The correlation between measurements at age 11 and 12 is $\widehat{\rho} = 0.54$ and $\widehat{\rho}^4 = 0.086$ between measurements at age 11 and 15.

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

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

and interpret the effect of the variables.

Solution
The interpretation is again the same as for the compound symmetry model.

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

$$\begin{aligned}\texttt{tolerance}_{ij} &= \beta_0 + \beta_1\texttt{sex}_{ij} + \beta_2\texttt{exposure}_{ij} + \beta_3\texttt{age}_{ij} + \varepsilon_{ij},\\ \boldsymbol{\varepsilon}_i &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_5, \sigma^2\mathbf{I}_5), i = 1,\ldots,16; j = 1,\ldots,5,\end{aligned} \tag{5.3.3}$$

Which of Models 5.3.1,5.3.2 and 5.3.3 would you choose? Justify your answer.

Solution
Both 5.3.1 and 5.3.2 are superior to the multiple linear regression model 5.3.3 based on likelihood ratio tests. To compare 5.3.1 and 5.3.2, which are not nested, we can look at information criteria. The value of AIC (BIC) for 5.3.1 is 88.9 (90.4) and that of 5.3.2 is 74.5 (76.1), indicating that the AR(1) model is preferable to the compound symmetry structure.
The only differences between these models are in terms of within-group correlation:

- 5.3.1 assumes that observations from the same individuals are correlated in time and this correlation, $\tau/(\sigma^2+\tau)$, is constant.
- 5.3.2 assumes that observations from the same individuals are correlated in time and this correlation follows a first-order autoregressive process, meaning that ρ^h for two observations separated by h years, corresponding to a geometric decay.
- 5.3.3 assumes observations at each age for individuals are independent from one another.

(j) List all of the hypothesis of Models 5.3.1,5.3.2 and 5.3.3.

Solution
The assumptions are given mathematically on each equation. Common assumptions include:

- The equation for the mean model includes `sex`, `exposure` and `age` and their effect is assumed linear.
- All of the observations are assumed independent between groups.
- All of the observations have the same variance, σ^2.