Strathmore UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES
MASTER OF SCIENCE IN STATISTICAL SCIENCE
END OF SEMESTER EXAMINATION
STA 8404: LONGITUDINAL DATA ANALYSIS

Date: 30th August, 2021 Duration: 3 Hours

Instructions: Attempt Question ONE and any other two questions

Question ONE (30 marks)

a) Briefly describe the following types of longitudinal data.

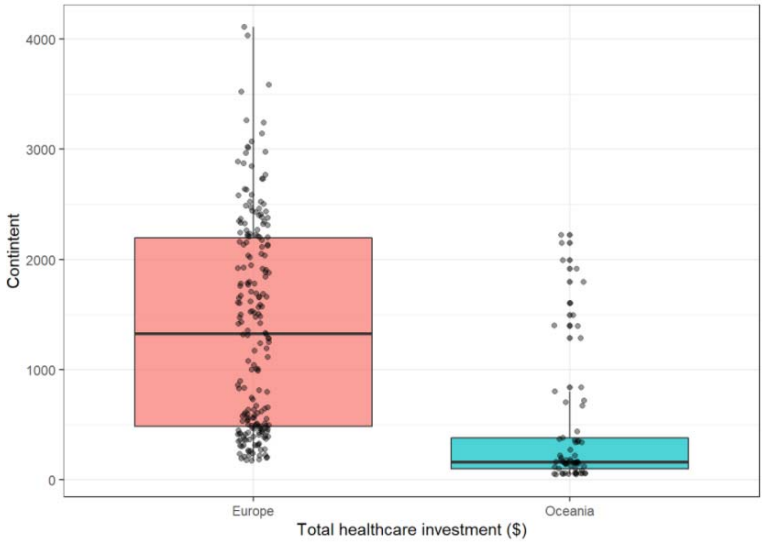
(i) Repeated cross-sections (2 marks)

(ii) Time Series (2 marks)

(iii) Panel data (2 marks)

(iv) Clustered data (2 marks)

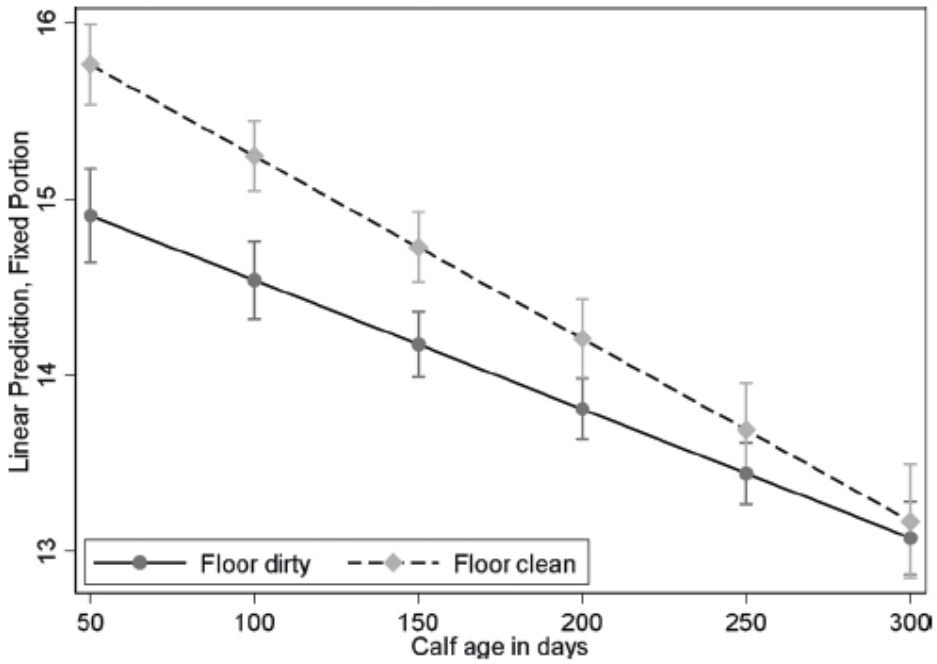
b) Consider boxplot of the Healthcare Expenditure in Europe and Oceania



Briefly describe the best linear mixed effect model for analyzing the data (4 marks)

c) Explain two reasons for doing Exploratory Data Analysis in Longitudinal studies (4 marks)

d) The figure below is an iinteraction plot of predicted daily lying time and 95%CI for calf age and floor cleanliness of the housing based on the final model of 187 calves in 150 Kenyan smallholder farms. This is adopted from Preventive Veterinary Medicine 189 (2021) 105296



i) Is the *daily lying time* significantly different by *floor type*? (2 marks)

ii) What is you comment on the variation between floor types by calf age? (3marks)

iii) Describe the best longitudinal model for analyzing this kind of data. (4 marks)

e) Let Y1 and Y2 be random variables with means μ1 and μ2, variances $\sigma_1^2$and σ_2^2 , and covariance σ12, respectively. Let c1 and c2 be constants.

i) Show that (3 marks)

$$var(c_1Y_1 + c_2Y_2) = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + 2c_1c_2\sigma_{12}.$$

ii) Let $Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$. Write down the covariance matrix Σ of Y. (2 marks)

.

Question TWO (15 marks)

Consider the Reaction times in a sleep deprivation study. A data frame with 180 observations with outcome **average reaction time** per day (in milliseconds); and predictors (1) **Days**, Number of days of sleep deprivation and (2) **Subject,** Subject number on which the observation was made.

Consider stage 1 model

$$Y_{sd} = \beta_{0s} + \beta_{1s} X_{sd} + e_{sd}$$

And stage two model:

$$\beta_{0s} = \gamma_0 + S_{0s}$$

$$\beta_{1s} = \gamma_1 + S_{1s}$$

With variance- covarance compnents

$$\langle S_{0s}, S_{1s} \rangle \sim N\left(\langle 0, 0 \rangle, \boldsymbol{\Sigma}\right)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \tau_{00}{}^2 & \rho\tau_{00}\tau_{11} \\ \rho\tau_{00}\tau_{11} & \tau_{11}{}^2 \end{pmatrix}$$

$$e_{sd} \sim N\left(0, \sigma^2\right)$$

Consider also the following possible model formulas for the sleep data

	Model	R Syntax
1		Reaction ~ days_deprived + (1 | Subject)
2		Reaction ~ days_deprived + (1 + days_deprived | Subject)
3		Reaction ~ days_deprived + (days_deprived | Subject)
4		Reaction ~ days_deprived + (0 + days_deprived | Subject)
5		Reaction ~ days_deprived + (days_deprived || Subject)

iii) write down the resulting the resulting final model after applying stage 2 modelling trick to stage 2. (4 marks)

iv) is there any difference between model 2 and 3? Explain (2 marks)

v) is there any difference between model 3 and 4? explain (2 marks)

vi) construct the variance-covariance matrices they construct for each of the 5 models (7 marks)

Qquestion THREE (15 marks)

a) Explain when and why researchers would prefer longitudinal studies over cross-sectional studies. (4 marks)

b) Consider the linear mixed effects model (in matrix notation), $y = X\beta + Zu + \varepsilon,$ expressed two level hierarchical model:

$$y|u \sim N(X\beta + Zu, R)$$

$$u \sim N(0, R)$$

i) Re-write the marginal model version of this model (4 marks)

ii) Derive an expression for the Maximum Likelihood Estimator or weighted Least Squares Estimator of β. (3 marks)

iii) Derive an expression for the best linear unbiased estimator of u (4 marks)

Question FOUR (15 marks)

a) Describe each of the following covariance pattern models, explaining how they arise and how they can be employed in longitudinal data analysis:

(i) Compound Symmetry Structure (2 marks)

(ii) First-Order Autoregressive Structure (2 marks)

(iii) Toeplitz or Banded Structure (2 marks)

(iv) Unstructured Form (2 marks)

b) Suppose we have the following statistical model

```
library(tidyverse)
tolerance <- read.csv("tolerance1.csv")
head(tolerance, n = 6)
```

```
##      id tol11 tol12 tol13 tol14 tol15 male exposure
## 1     9  2.23  1.79  1.90  2.12  2.66    0     1.54
## 2    45  1.12  1.45  1.45  1.45  1.99    1     1.16
## 3   268  1.45  1.34  1.99  1.79  1.34    1     0.90
## 4   314  1.22  1.22  1.55  1.12  1.12    0     0.81
## 5   442  1.45  1.99  1.45  1.67  1.90    0     1.13
## 6   514  1.34  1.67 2.23  2.12  2.44    1     0.90
```

(i) What does the following block of R code supposed to do? (2 marks)

```
tolerance_pp %>%
  distinct(id) %>%
  count()
```

(ii) What does the following block of R code supposed to do? (2 marks)

```
cor(tolerance[ , 2:6]) %>%
  round(digits = 2)
```

(iii) What kind of graphic(s) do you think the following block of R code will produce? (3 marks)

```
tolerance_pp %>%
  # we'll want to add that `tol` prefix back to the `age` values
  mutate(age = str_c("tol", age)) %>%
  # this variable is just in the way. we'll drop it
  select(-time) %>%
  # here's the main action
  pivot_wider(names_from = age, values_from = tolerance)
```