Statistical Modeling 2
Weekly assignments 4 2021

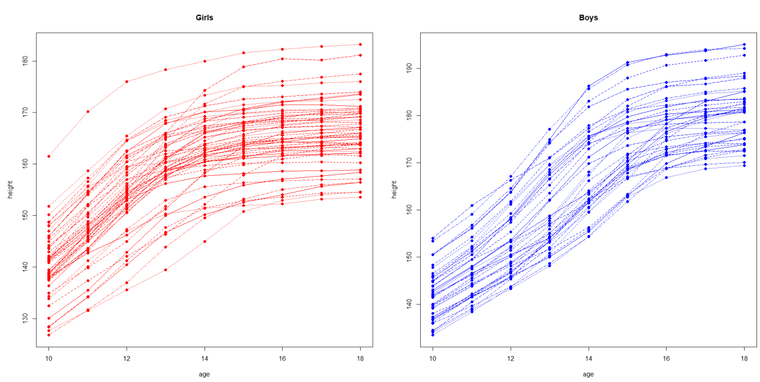
1. Consider the data set growthheight.txt related to growth profiles of sample of boys and girls:

```
> head(data)
       gender   Y10   Y11   Y12   Y13   Y14   Y15   Y16   Y17   Y18
girl01   girl 138.6 146.8 153.1 156.2 157.7 158.2 158.6 158.7 158.9
girl02   girl 140.9 146.1 152.9 159.5 162.6 165.0 165.6 166.1 166.0
girl03   girl 148.8 155.6 159.6 160.3 161.6 161.7 161.9 161.7 162.2
girl04   girl 143.0 148.5 154.8 161.2 165.2 166.5 167.2 167.4 167.8
girl05   girl 141.0 147.0 153.0 161.0 166.0 168.0 169.0 170.0 170.0
> tail(data)
      gender   Y10   Y11   Y12   Y13   Y14   Y15   Y16   Y17   Y18
boy35    boy 154.0 160.9 166.3 171.0 177.3 183.4 186.2 186.7 188.0
boy36    boy 143.0 148.0 153.6 162.1 171.2 176.8 179.1 180.1 180.8
boy37    boy 148.3 154.2 161.8 171.1 177.9 181.3 182.2 183.1 183.7
boy38    boy 147.8 153.5 161.5 169.7 175.6 178.3 179.2 179.8 180.7
boy39    boy 139.8 145.1 150.2 156.7 164.1 171.0 174.7 176.1 176.4

A list containing the heights of 39 boys and 54 girls from
age 10 to 18 and the ages at which they were collected.

gender - values of boy and girl
Y10, Y11, ..., Y18 - the variables giving the heights in centimeters of
                     children at ages 10,11,12,...18.

Tuddenham, R. D., and Snyder, M. M. (1954)
"Physical growth of California boys and girls from birth to age 18",
University of California Publications in Child Development, 1, 183-364.
```



Denote variables as following $Y_1 = \text{Y10}, Y_2 = \text{Y11}, \ldots, Y_9 = \text{Y18}$ and $X = \text{gender}$ with index variable j associated to its values. Let us assume that the random vector (i.e, random vector for each row i) $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{i9})$ follows the normal distribution $\mathbf{y}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where the expected value vector $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{i9})'$ is modeled by the growth curve model

$$\mathcal{M}: \quad \boldsymbol{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \mu_{i3} \\ \vdots \\ \mu_{i8} \\ \mu_{i9} \end{pmatrix} = \begin{pmatrix} 1 & t_1 & t_1^2 & t_1^3 \\ 1 & t_2 & t_2^2 & t_2^3 \\ 1 & t_3 & t_3^2 & t_3^3 \\ & \vdots & & \\ 1 & t_8 & t_8^2 & t_8^3 \\ 1 & t_9 & t_9^2 & t_9^3 \end{pmatrix} \begin{pmatrix} \theta_{0_j} \\ \theta_{1_j} \\ \theta_{2_j} \\ \theta_{3_j} \end{pmatrix} = (\mathbf{1} : \mathbf{t} : \mathbf{t}^2 : \mathbf{t}^3) \begin{pmatrix} \theta_{0_j} \\ \theta_{1_j} \\ \theta_{2_j} \\ \theta_{3_j} \end{pmatrix} = \mathbf{T}\boldsymbol{\theta}_j,$$

where the time values are $\mathbf{t} = (t_1, t_2, t_3 \ldots, t_8, t_9)' = (10, 11, 12, \ldots, 17, 18)'$.

(a) The model $\mathcal{M}$ can be written for the whole data as $\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta}\mathbf{T}' + \mathbf{E}$, where, for each row of $\mathbf{E}$, it is assumed $\varepsilon_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\theta}_1' \\ \boldsymbol{\theta}_2' \end{pmatrix}$. Find an estimate $\widehat{\boldsymbol{\Theta}}$ for the parameters $\boldsymbol{\Theta}$.

(1 point)

(b) For the girls, find the estimate to the expected value μ_{i10} at the age $t_{10} = 19$.

(1 point)

(c) Predict the value of the new observation $\mathbf{y}_f$ in both cases of a child being girl (i.e., $x_f = 1$) and a boy (i.e., $x_f = 0$). Create also a graph that displays the predicted profiles for a girl and a boy in the same graph.

(1 point)

(d) Test at 5% significance level, is the explanatory variable $X = \texttt{gender}$ statistically significant variable in the model. Calculate the value of the test statistic.

(1 point)

(e) Test the hypotheses

$$H_0 : \boldsymbol{\mu}_i = (\mathbf{1} : \mathbf{t}) \begin{pmatrix} \theta_{0_j} \\ \theta_{1_j} \end{pmatrix}, \qquad H_1 : \ \boldsymbol{\mu}_i = (\mathbf{1} : \mathbf{t} : \mathbf{t}^2 : \mathbf{t}^3) \begin{pmatrix} \theta_{0_j} \\ \theta_{1_j} \\ \theta_{2_j} \\ \theta_{3_j} \end{pmatrix}.$$

(1 point)

(f) Consider the partitioned random vector

$$\mathbf{y}_f = \begin{pmatrix} \mathbf{y}_{f1} \\ \mathbf{y}_{f2} \end{pmatrix} = \begin{pmatrix} \mathbf{T}_1 \\ \mathbf{T}_2 \end{pmatrix} \boldsymbol{\Theta}'\mathbf{x}_f + \begin{pmatrix} \varepsilon_{f1} \\ \varepsilon_{f2} \end{pmatrix},$$

where $\mathbf{y}_{f1}$ contains the random variables $\mathbf{y}_{f1} = (y_{f1}, y_{f2}, y_{f3})'$. Predict the value of the random vector $\mathbf{y}_{f2} = (y_{f4}, y_{f5}, y_{f6}, y_{f7}, y_{f8}, y_{f9})'$ when

```
gender   Y10   Y11   Y12
  girl 148.7 156.6 164.7
```

Create also 80 % simultaneous (asymptotic) prediction intervals for elements of the random vector $\mathbf{y}_{f2}$.

(1 point)

2. Consider the dataset stageforest.txt:

```
> data<-read.table("stageforest.txt", sep="\t", dec=".", header=TRUE)
> head(data)
  Tree.ID Age  Forest.ID dbhib.cm height.m
1       1  55 Clearwater   37.084 21.76272
2       1  45 Clearwater   31.496 18.71472
3       1  35 Clearwater   22.352 12.22248
4       1  25 Clearwater   17.780  8.71728
5       1  15 Clearwater   10.160  5.97408
6       2 107 Clearwater   50.800 31.51632
> tail(data)
    Tree.ID Age Forest.ID dbhib.cm height.m
537      85  68   Wallowa   36.322 29.77896
538      85  58   Wallowa   31.750 28.98648
539      85  48   Wallowa   24.892 25.72512
540      85  38   Wallowa   17.526 18.31848
541      85  28   Wallowa   11.684 11.24712
542      85  18   Wallowa    6.604  6.18744

Description

The data are stem measures from 66 trees.
The trees were selected as having been dominant throughout their
lives with no visible evidence of damage or forks.
The trees came from stands throughout the inland range of the species.

A data frame with 542 observations on the following variables.
Tree.ID - A factor uniquely identifying the tree.
Age - Age of tree
Forest.ID - The national forest in which the tree was.
dbhib.cm - Diameter (cm.) at 1.37 m (4'6?).
height.m - Height of tree (m)

The national forests are: Kaniksu, Coeur d'Alene, St. Joe, Clearwater, Nez Perce,
Clark Fork, Umatilla, Wallowa, and Payette.

Note that values are strangely in wrong order respect the age variable.
```

Denote the variables as following

$$Y_1 = \text{dbhib.cm}, \qquad Y_2 = \text{height.m}, \quad T = \text{Age}.$$

For the tree i, consider the multivariate linear mixed effects model

$$\begin{aligned} \mathcal{M}: \quad y_{i1t} &= \beta_{0_1} + \beta_{1_1} t_i + \beta_{2_1} t_i^2 + b_{i0_1} + b_{i1_1} t + \varepsilon_{i1t}, \\ y_{i2t} &= \beta_{0_2} + \beta_{1_2} t_i + \beta_{2_2} t_i^2 + b_{i0_2} + b_{i1_2} t + \varepsilon_{i2t}, \end{aligned}$$

which can be written also as

$$\mathcal{M}: \quad \mathbf{Y}_i = \mathbf{X}_i \mathbf{B} + \mathbf{Z}_i \mathbf{R}_i + \mathbf{E}_i, \qquad \mathrm{vec}(\mathbf{E}_i) \sim N(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I}),$$

where

$$\mathbf{Y}_i = (\mathbf{y}_{i1} : \mathbf{y}_{i2}), \quad \mathbf{X}_i = (\mathbf{1} : \mathbf{t}_i : \mathbf{t}_i^2), \quad \mathbf{Z}_i = (\mathbf{1} : \mathbf{t}_i), \quad \mathrm{vec}(\mathbf{R}_i) \sim N(\mathbf{0}, \mathbf{F}).$$

(a) Calculate the estimate for the fixed parameters $\mathbf{B}$.

(2 points)

(b) Calculate the estimate for the covariance matrix $\boldsymbol{\Sigma}$.

(1 point)

(c) Predict the value of the new observation $\mathbf{y}_f = \begin{pmatrix} y_{f1t} \\ y_{f2t} \end{pmatrix}$, when

`Tree.ID=67, Age=97.`

Create also 80 % simultaneous (asymptotic) prediction intervals for elements of the random vector $\mathbf{y}_f$.

(2 points)

(d) Predict the value of the new observation y_{f2t}, i.e., predict the value of the variable $Y_2 = \textsf{height.m}$ when we know that

`Tree.ID=67, Age=97, dbhib.cm=35.5,`

where the value dbhib.cm=35.5 means that at the age $t = 97$ observed value $y_{f1t} = 35.5$ is known.

(1 point)

3. (a) Consider the following longitudinal dataset:

```
> library(geepack)
> data(ohio)
> head(ohio)
  resp id age smoke
1    0  0  -2     0
2    0  0  -1     0
3    0  0   0     0
4    0  0   1     0
5    0  1  -2     0
6    0  1  -1     0
> tail(ohio)
     resp  id age smoke
2143    1 535   0     1
2144    1 535   1     1
2145    1 536  -2     1
2146    1 536  -1     1
2147    1 536   0     1
2148    1 536   1     1

The ohio data frame has 2148 rows and 4 columns.
The dataset is a subset of the six-city study,
a longitudinal study of the health effects of air pollution.

resp - an indicator of wheeze status (1=yes, 0=no)
id - a numeric vector for subject id
age - a numeric vector of age, 0 is 9 years old
smoke - an indicator of maternal smoking at the first year of the study
```

Denote the variables as following

$$Y = \textsf{resp}, \qquad X = \textsf{smoke}, \quad T = \textsf{age},$$

with index i related to the values of id. Model the expected value of the random variable y_{it} by the interaction effect model

$$g(\mu_{it}) = \beta_0 + \beta_1 x_{it} + \beta_2 t + \beta_3 x_{it} t,$$

with taking into account that most likely, for each i, the values $\mathbf{y}_i$ are correlated. Calculate the estimate for the expected value μ_{it} when

```
 id    age smoke
536   1.25     1
```

(3 points)

(b) In a growth curve model

$$\mathcal{M}: \ \mathbf{Y} = \mathbf{X}\boldsymbol{\Theta}\mathbf{T}' + \mathbf{E},$$

parameter matrix $\boldsymbol{\Theta}$ can be estimated by the formula

$$\widehat{\boldsymbol{\Theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}.$$

The growth curve model $\mathcal{M}$ can also be written in a form

$$\mathcal{V}: \ \mathrm{vec}(\mathbf{Y}) = (\mathbf{T} \otimes \mathbf{X})\,\mathrm{vec}(\boldsymbol{\Theta}) + \mathrm{vec}(\mathbf{E}).$$

Show that the ordinary least squares estimate for $\mathrm{vec}(\boldsymbol{\Theta})$

$$\min_{\mathrm{vec}(\boldsymbol{\Theta})} \left[\mathrm{vec}(\mathbf{Y}) - (\mathbf{T} \otimes \mathbf{X})\,\mathrm{vec}(\boldsymbol{\Theta})\right]' \left[\mathrm{vec}(\mathbf{Y}) - (\mathbf{T} \otimes \mathbf{X})\,\mathrm{vec}(\boldsymbol{\Theta})\right]$$

is containing the same estimates as $\widehat{\boldsymbol{\Theta}}$. Note that in a classical linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the ordinary least squares estimate is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

(3 points)