EX4

roi hezkiyahu

27 4 2022

```
# imports
library(tidyverse)
library(glue)
library(nlme)
```

Q1

Question 1.

Consider the first order autoregressive process AR(1): $y_i=\rho y_{i-1}+\varepsilon_i$, i=1,...,n, where $y_0=0$, $\varepsilon_i \sim N(0,\sigma^2)$ and i.i.d.

1. Write down the likelihood function for the data.
2. Find the MLEs of ρ and σ^2.

$$y_i|y_{i-1} \sim N(\rho y_{i-1}, \sigma^2) \text{ as a sum of normal variables}$$

$$f(y_i|y_{i-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\rho y_{i-1})^2}{2\sigma^2}}$$

$$L(\rho,\sigma^2) = f(y_1,\dots,y_n) = f(y_n|y_{n-1},\dots,y_0)f(y_{n-1},\dots,y_0) = \dots = \Pi_{i=1}^n f(y_i|y_{i-1},\dots y_0)f(y_0) = \Pi_{i=1}^n f(y_i|y_{i-1},\dots,y_0)$$

$$\text{notice that: } f(y_i|y_{i-1},\dots,y_0) = f(y_i|y_{i-1}) \text{ thus we get:}$$

$$L(\rho,\sigma^2) = \Pi_{i=1}^n f(y_i|y_{i-1}) = \Pi_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\rho y_{i-1})^2}{2\sigma^2}}$$

$$l(\rho,\sigma^2) = Log(L(\rho,\sigma^2)) = \sum_{i=1}^n ln(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\rho y_{i-1})^2}{2\sigma^2}}) = C + nln(\sigma) - \sum_{i=1}^n (\frac{(y_i-\rho y_{i-1})^2}{2\sigma^2})$$

$$\frac{\partial l}{\partial \rho} = \frac{1}{2\sigma^2}\sum_{i=1}^n 2y_i y_{i-1} - 2\rho y_{i-1}^2 := 0 \Rightarrow \hat{\rho}_{MLE} = \frac{\sum_{i=1}^n y_i y_{i-1}}{\sum_{i=1}^n y_{i-1}^2}$$

$$\frac{\partial l}{\partial \sigma} = \frac{n}{\sigma} - \sum_{i=1}^n \frac{(y_i-\rho y_{i-1})^2}{\sigma^3} = \frac{n\sigma^2 - \sum_{i=1}^n (\hat{\varepsilon}_i^2)}{\sigma^3} := 0 \Rightarrow \hat{\sigma}^2_{MLE} = \frac{\sum_{i=1}^n (\hat{\varepsilon}_i^2)}{n}; \hat{\varepsilon}_i = y_i - \hat{\rho}_{MLE} y_{i-1}$$

Q2

Question 2.

The file Dyestuff.dat (Dyestuff.dat) contains the dyestuff data. The object of experiment was to learn to what extent batch to batch variation in a certain raw material was responsible for variation in the final product yield. Five samples from each of six randomly chosen batches of raw material were taken and two laboratory determinations of product yield were made in two different laboratories for each of the resulting thirty samples.

1. Define a proper model to describe the data.
2. Does batch variation in the raw material strongly affects variation in the product yield?
3. Are there systematic differences between test results performed in different laboratories?

$$\text{lets define a LMM:}$$

$$Y = \text{vector of product yield}$$

$$X_1 = \underline{1}$$

$$X_2 = \text{vector of lab number}$$

$$Z = \text{vector of batch number}$$

$$Y_i|\gamma = X\beta + Z\gamma + \varepsilon \; ; \; \varepsilon \sim N(0,\sigma_\varepsilon^2 I), \; \gamma \sim N(0,\sigma_\gamma^2)$$

$$\text{thus: } Y \sim N(X\beta,\Sigma); \Sigma = diag(V); V_{ij} = \begin{cases} \sigma_\varepsilon^2 + \sigma_\gamma^2\,, & i=j \\ \sigma_\gamma^2\,, & i \neq j \end{cases}$$

```
dyestuff <- as_tibble(read.table("Dyestuff.dat",header = T)) %>%
  mutate(across(c(batch,laboratory),factor))
model_dye <- lme(yield~laboratory,random =~ 1|batch,data = dyestuff, method = "ML")
summary(model_dye)
```

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: dyestuff
##       AIC      BIC    logLik
##   632.243 640.6204 -312.1215
##
## Random effects:
##  Formula: ~1 | batch
##         (Intercept) Residual
## StdDev:    40.14192 38.87142
##
## Fixed effects: yield ~ laboratory
##                 Value Std.Error DF  t-value p-value
## (Intercept) 1527.1667  18.16387 53 84.07718  0.0000
## laboratory2   -8.6667  10.20814 53 -0.84900  0.3997
##  Correlation:
##             (Intr)
## laboratory2 -0.281
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -1.98686140 -0.78162560  0.01580612  0.85124201  2.42323899
##
## Number of Observations: 60
## Number of Groups: 6
```

we can see that batch variation has a strong affect on product yield due to the fact that ~ 1/2 of the variance can be explained by batch variation

we can also see that laboratory 2 has (in average) 9 less product yield but it is not significant so we can't conclude that there is a difference

Q3

Question 3.

The file Urine.dat (Urine.dat) gives ratios u_t of fluid intake to urine output over five consecutive 8-hour periods ($t=1,\ldots,5$) for 19 babies divided into two groups (G). The twelve babies in Group 1 received a surfactant treatment. The seven babies in Group 2 were given a placebo and constitute a control group.

1. Define a proper model expressing u_t as a linear function of t for both groups (for simplicity assume that the covariance matrix is the same for both groups). Fit the model.
2. Test the hypotheses that the linear trend is the same for both groups.
3. Do you think that a straight line is an appropriate model for the trend? If not, suggest way(s) to improve your original model.

$$\text{lets define a LMM:}$$
$$\mu_t(i) = \text{fluid intake to urine output for baby i at time t}$$
$$X_1 = \underline{1}$$
$$X_2 = \text{vector of group number}$$
$$X_3 = \text{vector of times}$$
$$Z_1 = \text{vector of baby ID}$$
$$Z_2 = \text{vector of times}$$
$$\mu_t(i)|\gamma = X\beta + Z\gamma + \varepsilon \; ; \; \varepsilon \sim N(0, \sigma_\varepsilon^2 I), \; \gamma \sim N(0, D)$$
$$\mu_t(i) \sim N(X\beta, ZDZ^t + \sigma_\varepsilon^2 I)$$

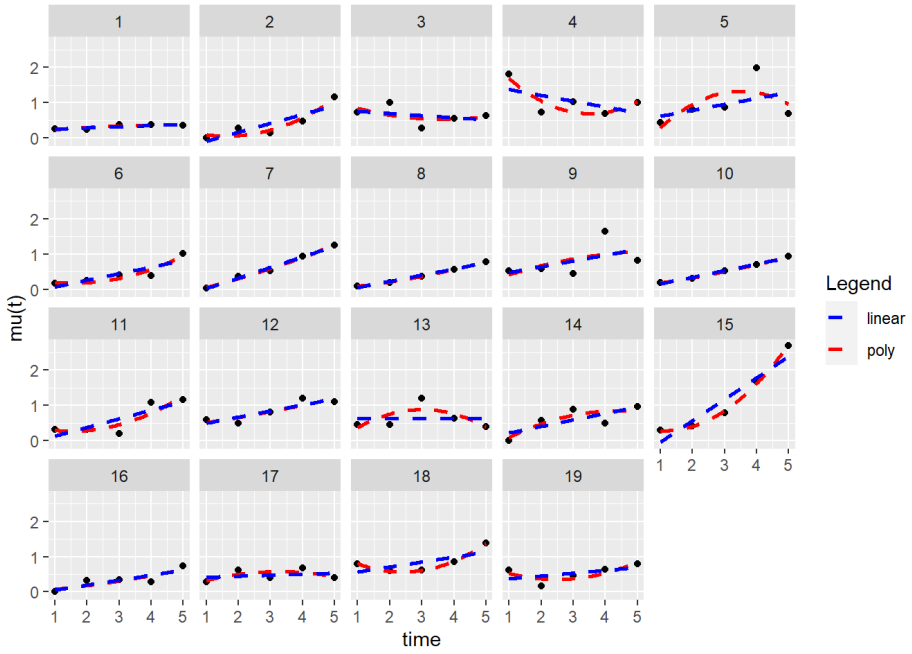
```
urine <- as_tibble(read.table("Urine.dat",header = T)) %>%
  mutate(across(G,factor))
urine_model_mat <- tibble("G"=rep(urine$G,5),
                          "baby" = factor(rep(1:19,5)),
                          "urine_time" = c(rep(1,19),rep(2,19),rep(3,19),rep(4,19),rep(5,19)),
                          "mu_t" = c(urine$u_1,urine$u_2,urine$u_3,urine$u_4,urine$u_5))%>%
  arrange(baby)
urine_model <- lme(mu_t ~ G + urine_time ,random =~ urine_time|baby ,data = urine_model_mat, method = "ML")
summary(urine_model)
```

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: urine_model_mat
##        AIC      BIC    logLik
##   91.81605 109.6932 -38.90802
##
## Random effects:
##  Formula: ~urine_time | baby
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 0.3613975 (Intr)
## urine_time  0.1241735 -0.858
## Residual    0.2941297
##
## Fixed effects: mu_t ~ G + urine_time
##                  Value  Std.Error DF  t-value p-value
## (Intercept) 0.18821835 0.11811787 75 1.593479  0.1153
## G2          0.01055021 0.11130662 17 0.094785  0.9256
## urine_time  0.15010526 0.03616861 75 4.150153  0.0001
##  Correlation:
##            (Intr) G2
## G2         -0.347
## urine_time -0.820  0.000
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -1.8016226 -0.4120285 -0.1289265  0.4092662  3.2587880
##
## Number of Observations: 95
## Number of Groups: 19
```

the linear model is the same for both groups if $\beta_1 = 0$ in the summary we can see that β_1 has a pvalue of 0.9256 thus not significant meaning that the linear trend is the same for both groups

```
urine_model_mat %>%
  ggplot(aes(x = urine_time, y = mu_t))+
  geom_point()+
  facet_wrap(urine_model_mat$baby)+
  geom_smooth(method='lm',formula = y~poly(x,2),se = F,aes(color = "poly"),lty =2)+
  geom_smooth(method='lm',se = F,aes(color = "linear"),lty =2)+
  labs(x = "time",
       y = "mu(t)",
       color = "Legend")+
  scale_color_manual(values = c("poly" = "red","linear" = "blue"))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



we can see that while most babies does have a linear relation, some babies have a more cubic relation so adding $time^2$ may improve our