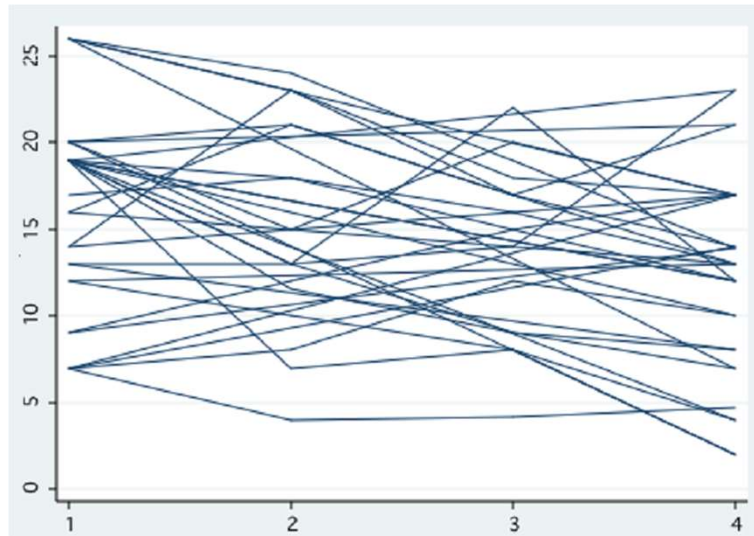
Modeling Normally Distributed Data with Repeated Measures

by

Olga Korosteleva, Ph.D.
CSULB



February 9, 2021, OCRUG

ABOUT ME

- *BS in Mathematics, Wayne State University, Detroit, MI, 1996*
- *MS in Statistics, Purdue University, West Lafayette, IN, 1998*
- *Ph.D. in Statistics, Purdue University, West Lafayette, IN, 2002*
- *Professor of Statistics, CSU, Long Beach, 2002-present*

SCHEDULE

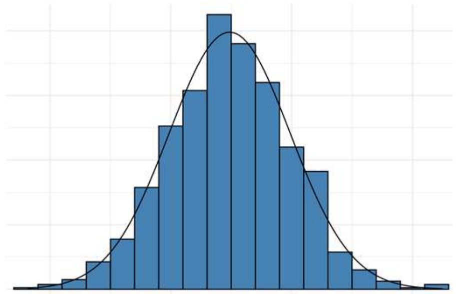
- *6:40PM-7:30PM Mixed-effects Model for Normal Response, Example*
- *7:30PM-7:50PM Mixed-effects Model Exercise*
- *7:50PM-8:00PM Mixed-effects Model Exercise Solution*
- *8PM-8:10PM Break*
- *8:10PM-8:30PM Generalized Estimating Equations (GEE) Model for Normal Response, Example*
- *8:30PM-8:50PM GEE Exercise*
- *8:50PM-9:00PM GEE Exercise Solution*
- *9:00PM-9:30PM Additional Exercise + Solution*
- *9:30PM-9:45PM Wrap-up*

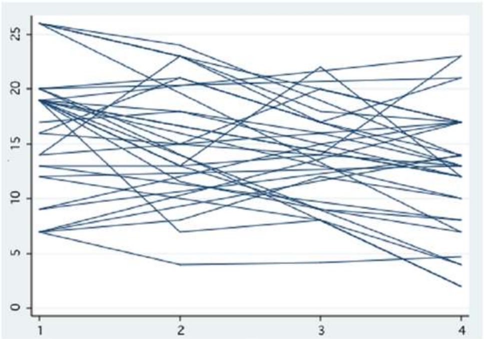
Greek Letters

- ❑ Alpha α
- ❑ Beta β
- ❑ Epsilon ε
- ❑ Rho ρ
- ❑ Sigma σ

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE:
Setting Explained

- Measurements are collected on individuals at different time points (*longitudinal data*), or under several conditions (*repeated measures*).
- The response variable y is normally distributed.
- The predictor variables $x_1, x_2, \dots, x_k$ may or may not depend on time (condition).
- Observations for different individuals are independent for any time point (or condition).
- Observations within each individual are modeled as correlated.





MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: Mathematics Explained

- Measurements are collected on n individuals at times $t_1, t_2, \ldots, t_p$ (or under conditions $1, 2, \ldots, p$). $Times\ (conditions)\ are\ used\ as\ continuous\ variables.$
- For the ith individual at time t_j, the response is y_{ij} and predictors are $x_{1ij}, x_{2ij}, \ldots, x_{kij}$. The model is

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \cdots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j + \varepsilon_{ij}$$

where $u_{1i} \sim N\left(0, \sigma_{u_1}^2\right)$ is $random\ intercept$, $u_{2i} \sim N\left(0, \sigma_{u_2}^2\right)$ is *random slope*, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ is $random\ error$. Random intercepts are independent, random slopes are independent, and random errors are independent. Covariance between u_{1i} and u_{2i} is $\sigma_{u_1 u_2}$.

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: Mathematics Explained (Continued)

- In the model $y_{ij} = \beta_0 + \beta_1 x_{1ij} + \cdots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j + \varepsilon_{ij}$, the terms $\beta_1 x_{1ij}, \ldots, \beta_k x_{kij}$, and $\beta_{k+1} t_j$ are called *fixed-effect terms*, u_{1i}, and $u_{2i} t_j$ are called *random-effect terms*, so overall, the model is called a *mixed-effects model*.

- It can be shown that for two different individuals, the responses are independent: $Cov(y_{ij}, y_{i'j'}) = 0$ for any $i \neq i'$.

- It can be shown that observations within the same individual are correlated: for any given i and $j \neq j'$, $Cov(y_{ij}, y_{ij'}) = \sigma_{u_1}^2 + \sigma_{u_1 u_2}(t_j + t_{j'}) + \sigma_{u_2}^2 t_j t_{j'}$.

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: Mathematics Explained (Continued)

- In this model, y is a normally distributed random variable with mean $Ey = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} t$ and variance $Var(y) = \sigma^2_{u_1} + 2\sigma_{u_1 u_2} t + \sigma^2_{u_2} t^2 + \sigma^2$.

- Parameters are $\beta_0, \beta_1, \ldots \beta_{k+1}$, $\sigma^2_{u_1}, \sigma^2_{u_2}$, $\sigma_{u_1 u_2}$, and σ^2.

- Fitted model has $\hat{E}y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t$, and the estimated parameters $\hat{\sigma}^2_{u_1}, \hat{\sigma}^2_{u_2}$, $\hat{\sigma}_{u_1 u_2}$, and $\hat{\sigma}^2$. R outputs $\hat{\sigma}_{u_1}, \hat{\sigma}_{u_2}$, $\hat{\rho} = \frac{\hat{\sigma}_{u_1 u_2}}{\hat{\sigma}_{u_1} \hat{\sigma}_{u_2}}$, and $\hat{\sigma}$.

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: Mathematics Explained (Continued)

- **Interpretation of fitted coefficients**:

  - If x_1 is continuous, $\hat{\beta}_1$ represents the change in the estimated mean of y for a one-unit increase in x_1, provided all the other variables are unchanged. Indeed,
$$\hat{E}y|_{x_1+1} - \hat{E}y|_{x_1} = \hat{\beta}_0 + \hat{\beta}_1(x_1 + 1) + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t\right) = \hat{\beta}_1.$$

  - If x_1 is a 0 - 1 variable, $\hat{\beta}_1$is interpreted as the difference of the estimated means of y for $x_1 = 1$ and $x_1 = 0$, controlling for the other predictors. Indeed,
$$\hat{E}y|_{x_1=1} - \hat{E}y|_{x_1=0} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t) = \hat{\beta}_1.$$

- **Prediction**: For a given set of predictors $x_1^0, x_2^0, \ldots, x_k^0, t^0$, the predicted response y^0 is computed as:
$$y^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots \hat{\beta}_k x_k^0 + \hat{\beta}_{k+1} t^0.$$

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXAMPLE

- In a clinic, doctors are testing a certain cholesterol lowering medication. Patients' gender and age at the beginning of the study are recorded for 27 patients. The low-density lipoprotein (LDL) cholesterol levels are measured in all the patients at the baseline, and then at 6-, 9-, and 24-month visits. We use these data to develop a regression model that relates LDL level to the gender, age, and months into the study.

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXAMPLE

- We create a long-form data.

```
cholesterol.data<- read.csv(file="C:/./LDLData.csv", header=TRUE, sep=",")

#creating long-form data set
library(reshape2)
longform.data<- melt(cholesterol.data, id.vars=c("id", "gender", "age"),
variable.name = "LDLmonth", value.name="LDL")

#creating numeric variable for time
month<- ifelse(longform.data$LDLmonth=="LDL0", 0, ifelse(longform.data$LDLmonth
=="LDL6", 6, ifelse(longform.data$LDLmonth=="LDL9",9,24)))
```

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXAMPLE

```
> longform.data
    id gender age LDLmonth LDL
1    1      M  50     LDL0  73
2    2      F  72     LDL0 174
3    3      M  46     LDL0  85
4    4      F  71     LDL0 172
5    5      F  75     LDL0 186
       < rows omitted >
104 23      M  62    LDL24  94
105 24      F  77    LDL24 155
106 25      M  55    LDL24  78
107 26      F  74    LDL24 111
108 27      F  79    LDL24 145
```

```
#creating numeric variable for time
month<- ifelse(longform.data$LDLmonth=="LDL0",
0, ifelse(longform.data$LDLmonth
=="LDL6", 6,
ifelse(longform.data$LDLmonth=="LDL9",9,24)))

> month
[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[24] 0 0 0 0 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
[47] 6 6 6 6 6 6 6 6 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
[70] 9 9 9 9 9 9 9 9 9 9 9 9 24 24 24 24 24 24 24 24
[93] 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24
```

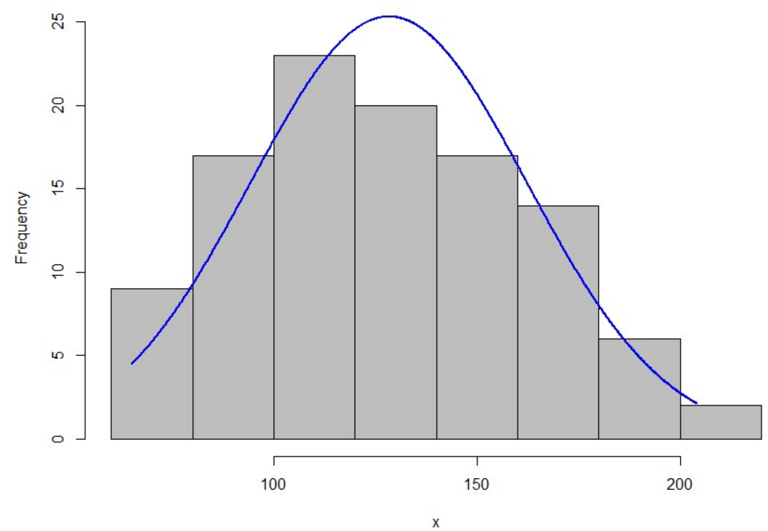
MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXAMPLE

❑ We plot a histogram and conduct the normality test.

```
#plotting histogram with fitted normal density
library(rcompanion)
plotNormalHistogram(longform.data$LDL)

#testing for normality of distribution
shapiro.test(longform.data$LDL)
```

```
Shapiro-Wilk normality test
W = 0.97668, p-value = 0.05449
```



Testing H_0:normal vs. H_1: non-normal.
Since p-value > 0.05, fail to reject H_0 and conclude normality.

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXAMPLE

- We fit the model.

```
#fitting random slope and intercept model
library(nlme)

summary(fitted.model<- lme(LDL ~ gender+age+month,
random =~ 1+month|id, control=lmeControl(opt="optim"), data=longform.data))
```

```
Random effects:
             StdDev Corr
(Intercept) 22.807 (Intr)
month         0.886 -0.812
Residual      8.358
```

```
Fixed effects:
              Value Std.Error DF t-value p-value
(Intercept)  94.827 23.379 80  4.056  0.0001
genderM     -29.811  6.972 24 -4.276  0.0003
age           0.920  0.337 24  2.732  0.0116
month        -1.096  0.193 80 -5.671  0.0000
```

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXAMPLE

- We write the fitted model.

$$\hat{E}(LDL) = 94.827 - 29.811 \cdot male + 0.920 \cdot age - 1.096 \cdot month,$$

and $\hat{\sigma}_{u_1} = 22.807$, $\hat{\sigma}_{u_2} = 0.886$, $\hat{\rho} = \frac{\hat{\sigma}_{u_1 u_2}}{\hat{\sigma}_{u_1}\hat{\sigma}_{u_2}} = -0.812$, and $\hat{\sigma} = 8.358$.

Since all the p-values are less than 0.05, all predictors are statistically significant.

WHAT DOES THIS ALL MEAN?

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXAMPLE

$\hat{\sigma}_{u_1} = 22.807$, $\hat{\sigma}_{u_2} = 0.886$, $\hat{\rho} = \frac{\hat{\sigma}_{u_1 u_2}}{\hat{\sigma}_{u_1}\hat{\sigma}_{u_2}} = -0.812$, and $\hat{\sigma} = 8.358$

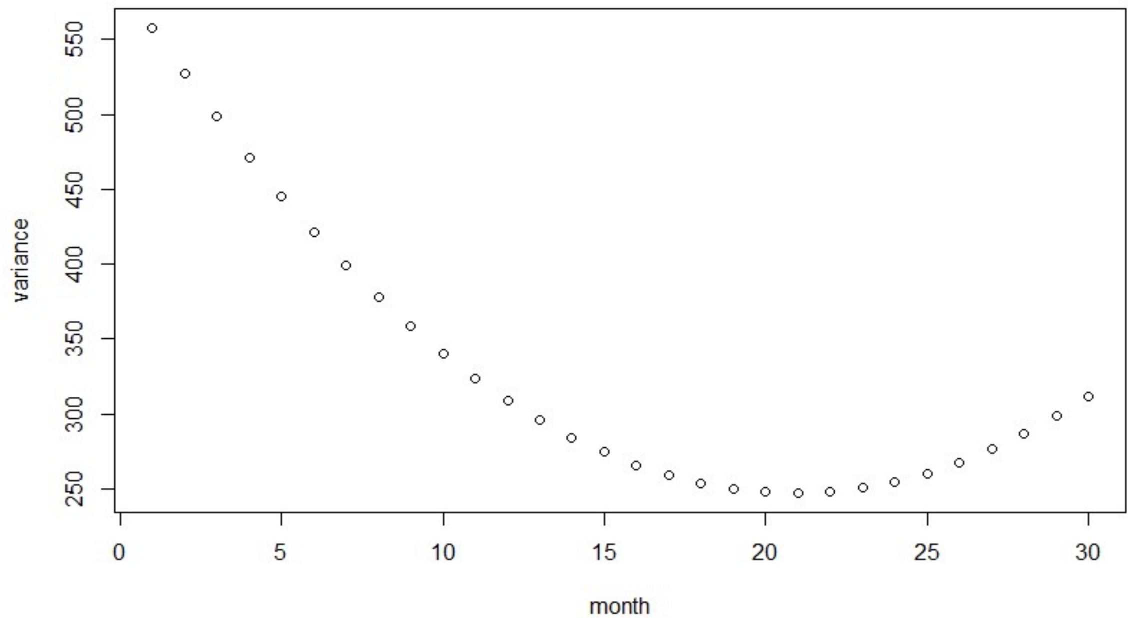
- IT MEANS THAT: The LDL measurement has a normal distribution with the estimated mean $\hat{E}(LDL) = 94.827 - 29.811 \cdot male + 0.920 \cdot age - 1.096 \cdot month$, and variance $\widehat{Var}(LDL) = \hat{\sigma}_{u_1}^2 + 2\hat{\sigma}_{u_1 u_2} month + \hat{\sigma}_{u_2}^2 month^2 + \hat{\sigma}^2$

$= (22.807)^2 + (2)(-0.812)(22.807)(0.886) month + (0.886)^2 month^2 + (8.358)^2 = 590.015 - 32.816\, month + 0.785 month^2.$

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: : EXAMPLE

- We plot the variance against month.

```
variance<- function(t) {
      590.015-32.816*t+0.785*t^2
      }
t<- 1:30
plot(t,variance(t), xlab="month",
ylab="variance")
```



- We see that variance decreases between 0 and 24 months.

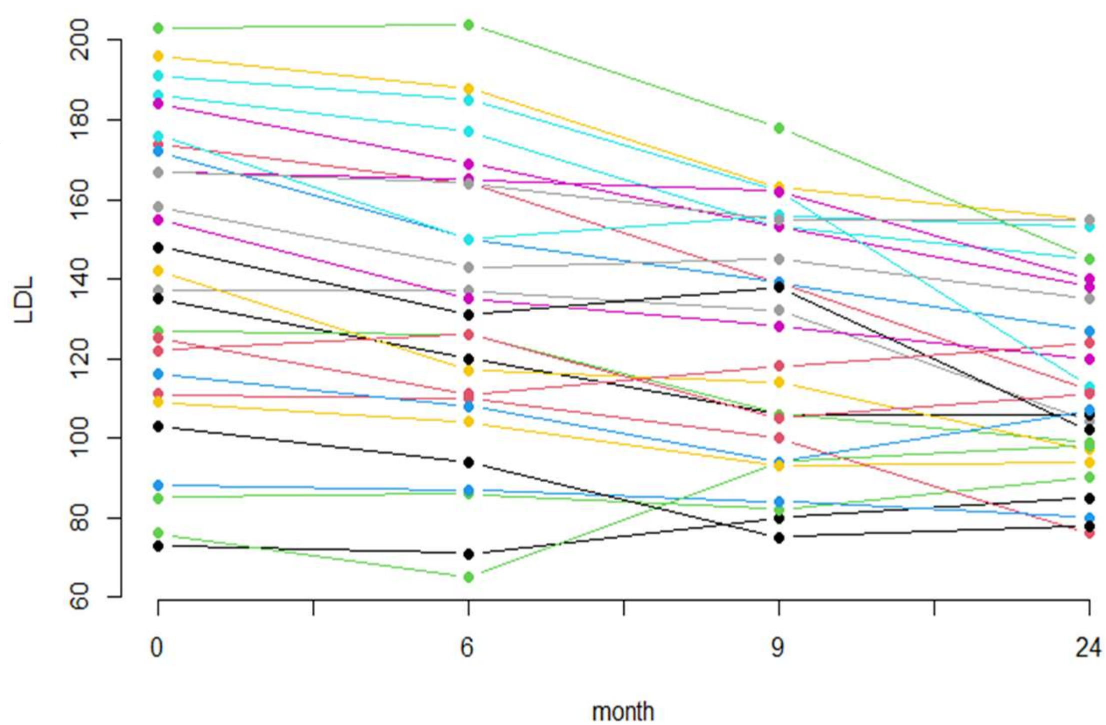
MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXAMPLE

- We plot individual profiles (LDL against months for each of 27 patients).

```
tr.data<- t(cholesterol.data)[-(1:3),]

matplot(tr.data, type="b", pch=16, lty=1,
col=1:27, axes=FALSE, ylab="LDL",
xlab="month")

xticks=c("0", "", "6", "", "9", "", "24")
axis(1,at=seq(1,4,0.5),labels=xticks)
axis(2)
```



- We see that variance decreases between 0 and 24 months.

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXAMPLE

$$\hat{E}(LDL) = 94.827 - 29.811 \cdot male + 0.920 \cdot age - 1.096 \cdot month$$

- We interpret the estimated regression coefficients.

- Gender: The estimated mean LDL for men is 29.811 points smaller than that for women.
- Age: With a one-year increase in age, the estimated mean LDL increases by 0.92 points.
- Month: For every additional month in the study, the estimated mean LDL is reduced by 1.096 points.

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXAMPLE

$$\hat{E}(LDL) = 94.827 - 29.811 \cdot male + 0.920 \cdot age - 1.096 \cdot month$$

- We use the fitted model for prediction of the LDL level for a 48-year old female patient 3 months into the study.

- By hand: $LDL^0 = 94.827 - 29.811 \cdot 0 + 0.920 \cdot 48 - 1.096 \cdot 3 = 135.699.$

- In R:

```
> predict(fitted.model, data.frame(gender=0, age=48, month=3), level=0)
  135.7156
```

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXERCISE

- Measurements were taken on 20 people involved in a physical fitness course. The data contain participants' gender, age, oxygen intake (in ml per kg body weight per minute), run time (time to run 1 mile, in minutes), and pulse (average heart rate while running). The running was done under three different conditions: the first one on a treadmill, the second one on an indoor running track, and the third one on an outdoor running track. Use the longform data to answer the following questions:
- (a) Check that pulse has a normal distribution. Construct a histogram and conduct normality tests.
- (b) Run a random slope and intercept regression model for pulse. Write down the fitted model.
- (c) Discuss significance of predictors at the 5% level. Interpret estimated significant regression coefficients.
- (d) Predict an average heart rate for a 36-year-old woman who is running on a treadmill, if her oxygen intake is 40.2 units, and her run time is 10.3 minutes per mile.

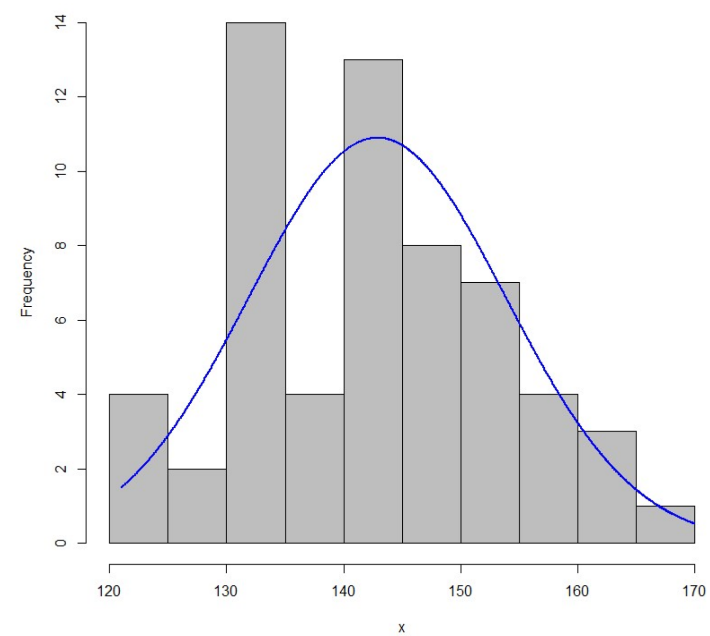
MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXERCISE SOLUTION

- (a) Check that pulse has a normal distribution. Construct a histogram and conduct normality tests.

```
library(rcompanion)
plotNormalHistogram(longform.data$pulse)


shapiro.test(longform.data$pulse)
```

```
Shapiro-Wilk normality test
W = 0.98398, p-value = 0.6173
```



- (b) Run a random slope and intercept regression model for pulse.

```
library(nlme)
summary(fitted.model<- lme(pulse ~ gender + age + oxygen
+ runtime + condition, random = ~ 1 + condition | id,
control=lmeControl(opt="optim"), data=longform.data))
```

```
Random effects:
            StdDev  Corr
(Intercept) 8.008   (Intr)
condition   6.091   -0.999
Residual    3.939
```

Fixed effects:

	Value	Std.Error	DF	t-value	p-value
(Intercept)	174.492	10.075446	37	17.318583	0.0000
genderM	-4.782	1.856487	17	-2.575887	0.0196
age	-0.198	0.124495	17	-1.589534	0.1304
oxygen	-0.909	0.167780	37	-5.419243	0.0000
runtime	0.614	0.591748	37	1.037967	0.3060
condition	6.194	1.531663	37	4.043907	0.0003

Significant are: gender, oxygen, and condition.

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXERCISE SOLUTION

- Write down the fitted model.

$\hat{E}(pulse) = 174.492 - 4.782 \cdot male - 0.198 \cdot age - 0.909 \cdot oxygen + 0.614 \cdot runtime + 6.194 \cdot$

$condition$ and $\hat{\sigma}_{u_1} = 8.008$, $\hat{\sigma}_{u_2} = 6.091$, $\hat{\rho} = \frac{\hat{\sigma}_{u_1 u_2}}{\hat{\sigma}_{u_1}\hat{\sigma}_{u_2}} = -0.999$, and $\hat{\sigma} = 3.939$.

- (c) Discuss significance of predictors at the 5% level. Interpret estimated significant regression coefficients.

  - Gender: For male runners, the estimated average pulse is 4.782 units lower than that for female runners.
  - Oxygen: As oxygen intake increases by one unit, the estimated mean pulse decreases by 0.909 units.
  - Condition: As the condition number increases by one, the estimated mean pulse increases by 6.194 units.

MIXED-EFFECTS MODEL FOR NORMAL RESPONSE: EXERCISE SOLUTION

$$\hat{E}(pulse) = 174.492 - 4.782 \cdot male - 0.198 \cdot age - 0.909 \cdot oxygen + 0.614 \cdot runtime + 6.194 \cdot condition$$

- (d) Predict an average heart rate for a 36-year-old woman who is running on a treadmill (condition=1), if her oxygen intake is 40.2 units, and her run time is 10.3 minutes per mile.

- By hand: $pulse^0 = 174.492 - 4.782 \cdot 0 - 0.198 \cdot 36 - 0.909 \cdot 40.2 + 0.614 \cdot 10.3 + 6.194 \cdot 1 = 143.3404.$

- In R:

```
print(predict(fitted.model, data.frame(gender=0, age=36, oxygen=40.2,
runtime=10.3, condition=1),level=0))
```

```
143.3374
```

GENERALIZED ESTIMATING EQUATIONS MODEL FOR NORMAL RESPONSE: Mathematics Explained

- In Generalized Estimating Equations (GEE) model, for each individual, y is a normally distributed random variable with mean $Ey = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} t$ and correlation matrix (called *working correlation matrix*) $\boldsymbol{R}$ of the form:

$$\begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \dots & \alpha_{1p} \\ \alpha_{12} & 1 & \alpha_{23} & \dots & \alpha_{2p} \\ \alpha_{13} & \alpha_{23} & 1 & \dots & \alpha_{3p} \\ \dots & \dots & \dots & \dots & \dots \\ \alpha_{1p} & \alpha_{2p} & \alpha_{3p} & \dots & 1 \end{bmatrix}$$

- Unstructured ($\frac{p(p-1)}{2}$ parameters)

<u>Meaning</u>: Correlations at different time points are all different.

$$\begin{bmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{p-1} \\ \alpha & 1 & \alpha & \dots & \alpha^{p-2} \\ \alpha^2 & \alpha & 1 & \dots & \alpha^{p-3} \\ \dots & \dots & \dots & \dots & \dots \\ \alpha^{p-1} & \alpha^{p-2} & \alpha^{p-3} & \dots & 1 \end{bmatrix}$$

- Autoregressive (1 parameter)

<u>Meaning</u>: Measurements are less correlated for time points further apart.

GENERALIZED ESTIMATING EQUATIONS MODEL FOR NORMAL RESPONSE: Mathematics Explained

$$\begin{bmatrix} 1 & \alpha & \alpha & \dots & \alpha \\ \alpha & 1 & \alpha & \dots & \alpha \\ \alpha & \alpha & 1 & \dots & \alpha \\ \dots & \dots & \dots & \dots & \dots \\ \alpha & \alpha & \alpha & \alpha & 1 \end{bmatrix}$$

- Compound symmetric or exchangeable (1 parameter)

Meaning: Better works for conditions rather than time points.

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

- Independent (0 parameters)

Meaning: not correlated.

- The model that fits the data the best is chosen according to the *quasi-likelihood under the independence* (*QIC*) criterion. The model with the smallest QIC value is the winner. If there is a tie, pick either model.
- Once the best-fitted model is chosen, we work with the estimated mean response for interpretation and prediction.

GEE MODEL FOR NORMAL RESPONSE: EXAMPLE

- In our example, we use the GEE model to regress LDL on gender, age, and months into the study.

```
library(geepack)

#fitting GEE model with unstructured working correlation matrix

summary(un.fitted.model<- geeglm(LDL ~ gender + age + month, data=longform.data, id=id,
family=gaussian(link="identity"), corstr="unstructured"))
```

```
 Coefficients:
            Estimate  Std.err   Wald Pr(>|W|)
(Intercept)  83.8023  22.0269 14.475 0.000142 ***
genderM     -34.3149   6.5082 27.800 1.35e-07 ***
age           1.0077   0.2935 11.786 0.000597 ***
month        -0.4788   0.4071  1.383 0.239578
```

```
Estimated Correlation Parameters:
          Estimate Std.err
alpha.1:2   1.1383 0.08577
alpha.1:3   0.6437 0.14531
alpha.1:4   0.1415 0.32543
alpha.2:3   0.6076 0.12888
alpha.2:4   0.1527 0.19510
alpha.3:4   0.4274 0.12394
```

$$\begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \dots & \alpha_{1p} \\ \alpha_{12} & 1 & \alpha_{23} & \dots & \alpha_{2p} \\ \alpha_{13} & \alpha_{23} & 1 & \dots & \alpha_{3p} \\ \dots & \dots & \dots & \dots & \dots \\ \alpha_{1p} & \alpha_{2p} & \alpha_{3p} & \dots & 1 \end{bmatrix}$$

Model is not reliable because one estimated correlation is above 1.

GEE MODEL FOR NORMAL RESPONSE: EXAMPLE

```
#fitting GEE model with autoregressive working correlation matrix

summary(ar.fitted.model<- geeglm(LDL ~ gender + age + month, data=longform.data, id=id,
family=gaussian(link="identity"), corstr="ar1"))
```

```
Coefficients:
            Estimate Std.err Wald Pr(>|W|)
(Intercept)   90.171  22.964 15.4  8.6e-05 ***
genderM      -36.463   6.942 27.6  1.5e-07 ***
age            1.032   0.319 10.4   0.0012 **
month         -0.926   0.173 28.5  9.3e-08 ***
```

$$
\boldsymbol{R} = \begin{bmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{p-1} \\ \alpha & 1 & \alpha & \dots & \alpha^{p-2} \\ \alpha^2 & \alpha & 1 & \dots & \alpha^{p-3} \\ \dots & \dots & \dots & \dots & \dots \\ \alpha^{p-1} & \alpha^{p-2} & \alpha^{p-3} & \dots & 1 \end{bmatrix}
$$

```
Estimated Correlation Parameters:

      Estimate Std.err
alpha    0.701  0.0827
```

```
library(MuMIn)

QIC(ar.fitted.model)

123
```

GEE MODEL FOR NORMAL RESPONSE: EXAMPLE

```
#fitting GEE model with compound symmetric (exchangeable) working correlation matrix

summary(cs.fitted.model<- geeglm(LDL ~ gender + age + month, data=longform.data, id=id,
family=gaussian(link="identity"), corstr="exchangeable"))
```

```
Coefficients:
            Estimate Std.err  Wald Pr(>|W|)
(Intercept)   88.917  25.341 12.31  0.00045 ***
genderM      -37.405   7.297 26.28  3.0e-07 ***
age            1.069   0.352  9.22  0.00239 **
month         -1.096   0.190 33.39  7.5e-09 ***
```

$$
\boldsymbol{R} = \begin{bmatrix} 1 & \alpha & \alpha & \dots & \alpha \\ \alpha & 1 & \alpha & \dots & \alpha \\ \alpha & \alpha & 1 & \dots & \alpha \\ \dots & \dots & \dots & \dots & \dots \\ \alpha & \alpha & \alpha & \alpha & 1 \end{bmatrix}
$$

```
Estimated Correlation Parameters:

      Estimate Std.err
alpha    0.582  0.0939
```

```
QIC(cs.fitted.model)
126
```

GEE MODEL FOR NORMAL RESPONSE: EXAMPLE

```
#fitting GEE model with independent working correlation matrix

summary(ind.fitted.model<- geeglm(LDL ~ gender + age + month, data=longform.data, id=id,
family=gaussian(link="identity"), corstr="independence"))
```

```
Coefficients:
            Estimate Std.err  Wald Pr(>|W|)
(Intercept)   88.917  25.341 12.31  0.00045 ***
genderM      -37.405   7.297 26.28  3.0e-07 ***
age            1.069   0.352  9.22  0.00239 **
month         -1.096   0.190 33.39  7.5e-09 ***
```

$$\boldsymbol{R}=\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

```
QIC(ind.fitted.model)
126
```

GEE MODEL FOR NORMAL RESPONSE: EXAMPLE

- We fit the autoregressive model (it is the best-fitted model).

```
summary(geeglm(LDL ~ gender + age + month, data=longform.data, id=id,
family=gaussian(link="identity"), corstr="ar1"))
```

```
 Coefficients:
             Estimate Std.err Wald Pr(>|W|)
(Intercept)    90.171  22.964 15.4  8.6e-05 ***
genderM       -36.463   6.942 27.6  1.5e-07 ***
age             1.032   0.319 10.4   0.0012 **
month          -0.926   0.173 28.5  9.3e-08 ***

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.701  0.0827
```

The fitted model has the estimated mean $\hat{E}(LDL) = 90.171 - 36.463 \cdot male + 1.032 \cdot age - 0.926 \cdot month$ and the estimated working correlation matrix

$$\hat{\boldsymbol{R}} = \begin{bmatrix} 1 & 0.701 & 0.491 & 0.344 \\ 0.701 & 1 & 0.701 & 0.491 \\ 0.491 & 0.701 & 1 & 0.701 \\ 0.344 & 0.491 & 0.701 & 1 \end{bmatrix}.$$

$(0.701)^2 = 0.491, (0.701)^3 = 0.344.$

All predictors (gender, age, and month) are statistically significant at the 5% level.

GEE MODEL FOR NORMAL RESPONSE: EXAMPLE

$$\hat{E}(LDL) = 90.171 - 36.463 \cdot male + 1.032 \cdot age - 0.926 \cdot month$$

- We interpret the estimated regression coefficients.

- Gender: The estimated mean LDL for men is 36.463 points lower than that for women.
- Age: With a one-year increase in age, the estimated mean LDL increases by 1.032 points.
- Month: For every additional month in the study, the estimated mean LDL is reduced by 0.926 points.

GEE MODEL FOR NORMAL RESPONSE: EXAMPLE

$$\hat{E}(LDL) = 90.171 - 36.463 \cdot male + 1.032 \cdot age - 0.926 \cdot month$$

- We use the fitted model for prediction of the LDL level for a 48-year old female patient 3 months into the study.

- By hand: $LDL^0 = 90.171 - 36.463 \cdot 0 + 1.032 \cdot 48 - 0.926 \cdot 3 = 136.929.$

- In R:

```
print(predict(ar.fitted.model, data.frame(gender="F", age=48, month=3)))
```

137

GEE MODEL FOR NORMAL RESPONSE: EXERCISE

- Use the longform data in the fitness exercise to answer the following questions:
- (a) Run GEE models with unstructured, autoregressive, compound symmetric, and independent working correlation matrices. Output QICs.
- (b) Find the optimal model according to the QIC criterion.
- (c) For the optimal model, write down the fitted model, estimating <u>all</u> parameters.
- (d) Discuss significance of predictors and interpret significant estimated regression coefficients.
- (e) Predict an average heart rate for a 36-year-old woman who is running on a treadmill, if her oxygen intake is 40.2 units, and her run time is 10.3 minutes per mile.

GEE MODEL FOR NORMAL RESPONSE: EXERCISE SOLUTION

- Run GEE models.

```
#fitting GEE model with unstructured working correlation matrix
summary(un.fitted.model<- geeglm(pulse ~ gender + age + oxygen + runtime +condition,
data=longform.data,  id=id, family = gaussian(link="identity"), corstr = "unstructured"))
```

```
Coefficients:
            Estimate Std.err  Wald Pr(>|W|)
(Intercept)  156.204  15.968 95.69  < 2e-16 ***
genderM       -6.991   2.703  6.69   0.0097 **
age           -0.232   0.132  3.08   0.0795 .
oxygen        -0.373   0.217  2.95   0.0861 .
runtime        0.134   0.628  0.05   0.8315
condition      7.431   1.404 28.03  1.2e-07 ***
```

```
Estimated Correlation Parameters:
          Estimate Std.err
alpha.1:2    0.237  0.0946
alpha.1:3   -0.270  0.1600
alpha.2:3   -0.138  0.1881
```

```
QIC(un.fitted.model)
72.91
```

GEE MODEL FOR NORMAL RESPONSE: EXERCISE SOLUTION

```
#fitting GEE model with autoregressive working correlation matrix
summary(ar.fitted.model<- geeglm(pulse ~ gender + age + oxygen + runtime + condition,
data=longform.data, id=id, family = gaussian(link="identity"), corstr = "ar1"))
```

```
Coefficients:
            Estimate Std.err   Wald Pr(>|W|)
(Intercept)  159.809  15.190 110.68  < 2e-16 ***
genderM       -7.041   2.446   8.28    0.004 **
age           -0.213   0.123   3.02    0.082 .
oxygen        -0.438   0.218   4.04    0.044 *
runtime        0.104   0.566   0.03    0.855
condition      6.951   1.571  19.57  9.7e-06 ***
```

```
Estimated Correlation Parameters:
      Estimate Std.err
alpha   0.0272  0.0688
```

```
QIC(ar.fitted.model)
72.2
```

GEE MODEL FOR NORMAL RESPONSE: EXERCISE SOLUTION

```
#fitting GEE model with compound symmetric (exchangeable) working correlation matrix
summary(cs.fitted.model<- geeglm(pulse ~ gender + age + oxygen + runtime + condition,
data=longform.data,id=id, family = gaussian(link="identity"), corstr = "exchangeable"))
```

```
Coefficients:
            Estimate Std.err   Wald Pr(>|W|)
(Intercept)  159.541  15.374 107.68  < 2e-16 ***
genderM       -6.974   2.398   8.45   0.0036 **
age           -0.216   0.119   3.29   0.0696 .
oxygen        -0.441   0.206   4.59   0.0322 *
runtime        0.149   0.617   0.06   0.8085
condition      6.951   1.551  20.10  7.4e-06 ***
```

```
Estimated Correlation Parameters:
      Estimate Std.err
alpha  -0.0638  0.0947
```

```
QIC(cs.fitted.model)
72.2
```

GEE MODEL FOR NORMAL RESPONSE: EXERCISE SOLUTION

```
#fitting GEE model with independent working correlation matrix
summary(ind.fitted.model<- geeglm(pulse ~ gender + age + oxygen + runtime + condition,
data=longform.data,id=id, family = gaussian(link="identity"), corstr = "independence"))
```

```
Coefficients:
            Estimate Std.err   Wald Pr(>|W|)
(Intercept)  159.859  15.209 110.47  < 2e-16 ***
genderM       -7.042   2.429   8.41   0.0037 **
age           -0.213   0.122   3.06   0.0803 .
oxygen        -0.440   0.215   4.19   0.0407 *
runtime        0.108   0.575   0.04   0.8510
condition      6.948   1.567  19.67  9.2e-06 ***
```

```
QIC(ind.fitted.model)
72.2
```

Either of the three models (autoregressive, compound symmetric , or independent) are the best-fitted models. Independent is the simplest.

GEE MODEL FOR NORMAL RESPONSE: EXERCISE SOLUTION

- We fit the independent model.

```
summary(geeglm(pulse ~ gender + age + oxygen + runtime + condition, data=longform.data,
id=id, family = gaussian(link="identity"), corstr = "independence"))
```

```
Coefficients:
            Estimate Std.err   Wald Pr(>|W|)
(Intercept)  159.859  15.209 110.47  < 2e-16 ***
genderM       -7.042   2.429   8.41   0.0037 **
age           -0.213   0.122   3.06   0.0803 .
oxygen        -0.440   0.215   4.19   0.0407 *
runtime        0.108   0.575   0.04   0.8510
condition      6.948   1.567  19.67  9.2e-06 ***
```

The fitted model has $\hat{E}(pulse) = 159.859 - 7.042 \cdot male - 0.213 \cdot age - 0.440 \cdot oxygen + 0.108 \cdot runtime + 6.948 \cdot condition$ and the working correlation matrix

$$\widehat{\boldsymbol{R}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Predictors that are statistically significant at the 5% level are gender, oxygen intake, and condition.

GEE MODEL FOR NORMAL RESPONSE: EXERCISE SOLUTION

$$\hat{E}(pulse) = 159.859 - 7.042 \cdot male - 0.213 \cdot age - 0.440 \cdot oxygen + 0.108 \cdot runtime + 6.948 \cdot condition$$

- Give interpretation of estimated significant regression coefficients.

- Gender: For male runners, the estimated average pulse is 7.042 units lower than that for female runners.
- Oxygen: As oxygen intake increases by one unit, the estimated mean pulse decreases by 0.440 units.
- Condition: As the condition number increases by one, the estimated mean pulse increases by 6.948 units.

GEE MODEL FOR NORMAL RESPONSE: EXERCISE SOLUTION

$$\hat{E}(pulse) = 159.859 - 7.042 \cdot male - 0.213 \cdot age - 0.440 \cdot oxygen + 0.108 \cdot runtime + 6.948 \cdot condition$$

- Predict an average heart rate for a 36-year-old woman who is running on a treadmill (condition=1), if her oxygen intake is 40.2 units, and her run time is 10.3 minutes per mile.

- By hand: $pulse^0 = 159.859 - 7.042 \cdot 0 - 0.213 \cdot 36 - 0.440 \cdot 40.2 + 0.108 \cdot 10.3 + 6.948 \cdot 1 = 142.5634.$

- In R:

```
print(predict(ind.fitted.model, data.frame(gender="F", age=36,
oxygen=40.2, runtime=10.3, condition=1)))
```

143

ADDITIONAL EXERCISE

- A health center conducted a study on efficacy of an intervention on weight loss. The intervention consisted of a lecture on proper nutrition and importance of exercising, followed by a cooking class. The study had a wait list control group. For each of the 34 study participants, the investigators recorded the group (intervention or control), gender (F/M), the typical length of daily exercise in the past week (in minutes), and BMI (in kg/m^2) at the beginning of the study, and at 1 and 3 months afterwards. Use the data set "WeightLossData.csv" to analyze the data.

- (a) Verify normality of the response variable BMI by plotting the histogram and carrying out the normality test.
- (b) Fit the random slope and intercept model. Present the fitted model and specify all estimated parameters. Discuss significance of the parameters at the 5% significance level.
- (c) Give interpretation of the estimated significant beta coefficients. Is the intervention efficient?
- (d) Compute the predicted BMI at 3 months for an intervention group male participant, if he exercises for 1 hour every day.

ADDITIONAL EXERCISE

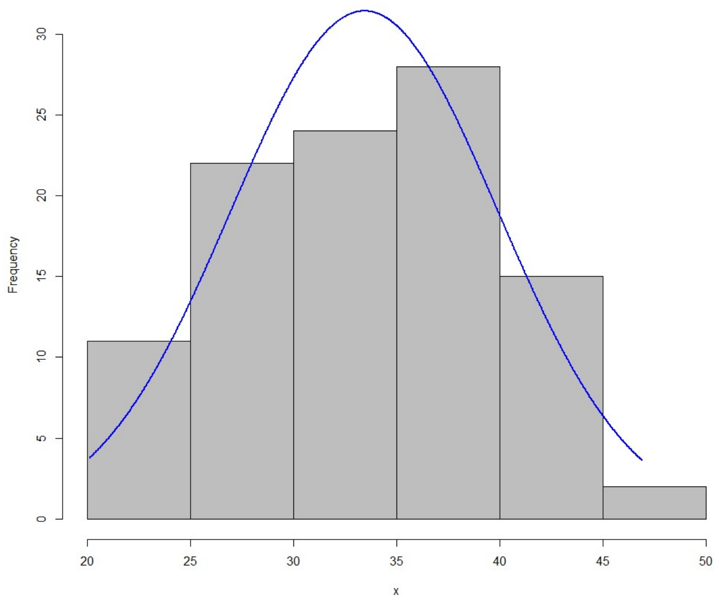
- (e) Fit the GEE models with unstructured, autoregressive, compound symmetric, and independent working correlation matrices of the response variable BMI.
- (f) Choose the best-fitted model with respect to the QIC criterion.
- (g) For the best-fitted model, write down the fitted model. Estimate all parameters. Discuss what predictors are significant at the 5% level.
- (h) Interpret the estimated significant regression coefficients. Is the intervention efficient?
- (i) Compute the predicted BMI at 3 months for an intervention group male participant, if he exercises for 1 hour every day.

ADDITIONAL EXERCISE SOLUTION

- (a) Verify normality of the response variable BMI by plotting the histogram and carrying out the normality test.

```
library(rcompanion)
plotNormalHistogram(longform.data$BMI)

shapiro.test(longform.data$BMI)
Shapiro-Wilk normality test
W = 0.98317, p-value = 0.2216
```



ADDITIONAL EXERCISE SOLUTION

- (b) Fit the random slope and intercept model. Present the fitted model and specify all estimated parameters. Discuss significance of the parameters at the 5% significance level.

```
library(nlme)
summary(fitted.model<- lme(BMI ~ group + gender + exercise + month,
random = ~ 1 + month | id, data=longform.data))
```

```
            StdDev    Corr
(Intercept) 5.4112519 (Intr)
month       0.5749658 0.821
Residual    1.8535182
```

```
Fixed effects:
               Value Std.Error DF   t-value p-value
(Intercept) 35.78162 1.5680426 66 22.819288  0.0000
groupInt    -1.19608 1.8719456 31 -0.638949  0.5275
genderM      1.23698 1.8969761 31  0.652082  0.5192
exercise    -0.03974 0.0112112 66 -3.544454  0.0007
month       -0.84454 0.2028822 66 -4.162726  0.0001
```

The fitted model is of the form $\hat{E}(BMI) = 35.78162 - 1.19608 \cdot intervention + 1.23698 \cdot male - 0.03974 \cdot$ *exercise* $- 0.84454 \cdot month$. The estimates of the other model parameters are $\hat{\sigma}_{u_1} = 5.411, \hat{\sigma}_{u_1} = 0.575$, $\hat{\rho}_{u_1 u_2} = 0.821$, and $\hat{\sigma} = 1.854$. Typical length of daily exercise and month into the study are significant predictors.

ADDITIONAL EXERCISE SOLUTION

- (c) Give interpretation of the estimated significant beta coefficients. Is the intervention efficient?

- Exercise: As the length of daily exercise increases by one minute, the estimated average BMI decreases by 0.03974 units.
- Month: It is estimated that the average BMI decreases by 0.84454 units for every additional month in the study.
- Group is not a significant predictor, thus from the statistical point of view, the intervention is not efficient.

ADDITIONAL EXERCISE SOLUTION

- (d) Compute the predicted BMI at 3 months for an intervention group male participant, if he exercises for 1 hour every day.

- Predicted BMI is
- By hand: $BMI^0 = 35.78162 - 1.19608 \cdot 1 + 1.23698 \cdot 1 - 0.03974 \cdot 60 - 0.84454 \cdot 3 = 30.9045.$
- In R:

```
print(predict(fitted.model, data.frame(gender=1, group=1, exercise=60,
month=3),level=0))
```

30.90464

ADDITIONAL EXERCISE SOLUTION

- (e) Fit the GEE models with unstructured, autoregressive, compound symmetric, and independent working correlation matrices of the response variable BMI.

```
#fitting GEE model with unstructured working correlation matrix

summary(un.fitted.model<- geeglm(BMI ~ group + gender + exercise + month, data=longform.data, id=patid,
family=gaussian(link="identity"), corstr = "unstructured"))
```

```
Coefficients:
             Estimate   Std.err    Wald Pr(>|W|)
(Intercept) 37.406687  1.796953 433.336  < 2e-16 ***
groupInt    -3.898733  2.105776   3.428   0.0641 .
genderM      0.748395  2.029100   0.136   0.7123
exercise    -0.048367  0.009774  24.486 7.48e-07 ***
month       -0.766805  0.153825  24.850 6.20e-07 ***
```

```
Estimated Correlation Parameters:

          Estimate Std.err
alpha.1:2   0.7600 0.12095
alpha.1:3   0.8422 0.12092
alpha.2:3   1.0617 0.05563
```

Unreliable model.

ADDITIONAL EXERCISE SOLUTION

```
#fitting GEE model with autoregressive working correlation matrix

summary(ar.fitted.model<- geeglm(BMI ~ group + gender + exercise + month, data=longform.data, id=id,
family=gaussian(link="identity"), corstr = "ar1"))
```

```
Coefficients:
            Estimate  Std.err   Wald Pr(>|W|)
(Intercept) 37.01575  1.73804 453.58  < 2e-16 ***
groupInt    -4.09328  2.04978   3.99    0.046 *
genderM      0.90402  1.97902   0.21    0.648
exercise    -0.02297  0.00489  22.09  2.6e-06 ***
month       -0.96053  0.16435  34.16  5.1e-09 ***
```

```
Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.912  0.0556
```

```
QIC(ar.fitted.model)
115
```

ADDITIONAL EXERCISE SOLUTION

```
#fitting GEE model with compound symmetric working correlation matrix

summary(cs.fitted.model<- geeglm(BMI ~ group + gender + exercise + month, data=longform.data, id=id,
family=gaussian(link="identity"), corstr = "exchangeable"))
```

```
Coefficients:
            Estimate  Std.err   Wald Pr(>|W|)
(Intercept) 36.78987  1.77243 430.84  < 2e-16 ***
groupInt    -3.57677  2.06691   2.99    0.084 .
genderM      0.99212  1.98382   0.25    0.617
exercise    -0.02777  0.00574  23.37  1.3e-06 ***
month       -0.95011  0.20191  22.14  2.5e-06 ***
```

```
Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.887  0.0651
```

```
QIC(cs.fitted.model)
115
```

ADDITIONAL EXERCISE SOLUTION

```
#fitting GEE model with independent working correlation matrix

summary(ind.fitted.model<- geeglm(BMI ~ group + gender + exercise + month, data=longform.data, id=id,
family=gaussian(link="identity"), corstr = "independence"))
```

```
 Coefficients:
            Estimate Std.err   Wald Pr(>|W|)
(Intercept)  36.3214  1.8095 402.92  < 2e-16 ***
groupInt     -4.6539  1.9882   5.48    0.019 *
genderM       1.0787  1.9477   0.31    0.580
exercise      0.0250  0.0186   1.82    0.177
month        -1.4161  0.2930  23.36  1.3e-06 ***
```

```
QIC(ind.fitted.model)
120
```

ADDITIONAL EXERCISE SOLUTION

- (f) Choose the best-fitted model with respect to the QIC criterion.

The models with the autoregressive and compound symmetric working correlation matrices have the smallest QIC value and thus has the best fit.

- (g) For the best-fitted model, write down the fitted model. Estimate all parameters. Discuss what predictors are significant at the 5% level.

We pick the GEE model with the compound symmetric working correlation matrix (it is simpler). The fitted model is $\hat{E}(BMI) = 36.78987 - 3.57677 \cdot intervention + 0.99212 \cdot male - 0.02777 \cdot exercise - 0.95011 \cdot month$, with the estimated working correlation matrix $\widehat{\boldsymbol{R}} = \begin{bmatrix} 1 & 0.887 & 0.887 \\ 0.887 & 1 & 0.887 \\ 0.887 & 0.887 & 1 \end{bmatrix}$. Typical length of daily exercise and month into the study are significant predictors.

ADDITIONAL EXERCISE SOLUTION

$$\hat{E}(BMI) = 36.78987 - 3.57677 \cdot intervention + 0.99212 \cdot male - 0.02777 \cdot exercise - 0.95011 \cdot month$$

- (h) Interpret the estimated significant regression coefficients. Is the intervention efficient?

- Exercise: As the length of daily exercise increases by one minute, the estimated average BMI decreases by 0.02777 units.
- Month: It is estimated that the average BMI decreases by 0.95011 units for every additional month in the study.
- Group is not a significant predictor at the 5% level, thus from the statistical point of view, the intervention is not efficient. It is significant at the 10% level.

ADDITIONAL EXERCISE SOLUTION

- (i) Compute the predicted BMI at 3 months for an intervention group male participant, if he exercises for 1 hour every day.

- Predicted BMI is
- By hand: $BMI^0 = 36.78987 - 3.57677 \cdot 1 + 0.99212 \cdot 1 - 0.02777 \cdot 60 - 0.95011 \cdot 3 = 29.68869.$
- In R:

```
print(predict(cs.fitted.model, data.frame(group="Int", gender="M",
exercise=60, month=3)))
29.7
```


Thank you !