7.4 Mixed-effects Regression Model for Normally Distributed Response

Assume that observations are collected on the same subjects repeatedly over time (**longitudinal data**) or under different conditions (**repeated measures**). In this case, observations for different individuals are modeled as independent but within each individual, observations are modelled as correlated. First we present the model for a normally distributed response variable. Suppose y is the response variable and $x_1, \dots, x_k$ are predictors. Assume also that data are collected at m occasions, at times $t_1, \dots, t_m$. The predictors may be observed once or multiple times. The model that accounts for within-subject correlation is termed **mixed-effects model**. It includes the **fixed-effects predictors** $x_1, \dots, x_k$ as well as additive random terms. The model is also termed **random slope and intercept model**. It is defined as

$$y_{ij} = \beta_0 + \beta_1 \, x_{1ij} + \cdots + \beta_k \, x_{kij} + \beta_{k+1} \, t_j + u_{1i} + u_{2i} \, t_j + \varepsilon_{ij}, \;\; i = 1, \dots, n, \; j = 1, \dots, m,$$

where u_{1i}'s are independent $\mathcal{N}(0, \sigma^2_{u_1})$ **random intercepts**, u_{2i}'s are independent $\mathcal{N}(0, \sigma^2_{u_2})$ **random slopes**, and ε_{ij}'s are independent $\mathcal{N}(0, \sigma^2)$ **errors** that are also independent of u_{1i}'s and u_{2i}'s. It is assumed that $\mathbb{C}ov(u_{1i}, u_{2i}) = \sigma_{u_1 u_2}$, and $\mathbb{C}ov(u_{1i}, u_{2i'}) = 0$ for $i \neq i'$.

It can be shown that $\mathbb{C}ov(y_{ij}, y_{i'j'}) = 0$, $i \neq i'$, meaning that the responses for different individuals are uncorrelated for any time points. It can also be shown that responses between different time points for the same individual are correlated, since $\mathbb{C}ov(y_{ij}, y_{ij'}) = \sigma^2_{u_1} + \sigma_{u_1 u_2} \, (t_j + t_{j'}) + \sigma^2_{u_2} \, t_j t_{j'}$, for $j \neq j'$. In addition, it can be verified that the response variable y_{ij} is normally distributed with the mean

$$\mathbb{E}(y) = \beta_0 + \beta_1 \, x_1 + \cdots + \beta_k \, x_k + \beta_{k+1} \, t$$

and variance $\mathbb{V}ar(y_{ij}) = \sigma^2_{u_1} + 2\sigma_{u_1 u_2} \, t_j + \sigma^2_{u_2} \, t_j^2 + \sigma^2$.

Definition. In the random slope and intercept model, the fitted mean response can be expressed as

$$\widehat{\mathbb{E}}(y) = \widehat{\beta}_0 + \widehat{\beta}_1 \, x_1 + \cdots + \widehat{\beta}_k \, x_k + \widehat{\beta}_{k+1} \, t.$$

All beta parameters along with $\sigma^2_{u_1}, \sigma_{u_1 u_2}, \sigma^2_{u_2}$, and σ^2 are estimated from the data.

Definition. The fitted beta coefficients are interpreted as in the general linear regression model, in terms of an average change in the mean response for a unit-increase in a continuous predictor, or as the difference between mean responses for level 1 and level 0 for a 0-1 predictor, provided all the other predictors stay unchanged.

Definition. For a specified set of values $x_1^0, \dots, x_k^0, t^0$, the predicted response is

$$y^0 = \widehat{\beta}_0 + \widehat{\beta}_1\, x_1^0 + \cdots + \widehat{\beta}_k\, x_k^0 + \widehat{\beta}_{k+1}\, t^0.$$

Missingness in Longitudinal Studies

In longitudinal studies, participants might miss one data collection point due to some unforeseen circumstances (for instance, are hospitalized for some study unrelated reason, or caught common cold and are too sick to venture out of the house). Some participants, however, can drop out of the study altogether (they are termed **lost to follow-up**). Some of the reasons for dropping out might be geographical relocation, incarceration, death or injury in a car accident. All these mentioned missing observations are said to be **missing completely at random** (**MCAR**), when missingness doesn't depend on any characteristic of a person and can essentially happen to anyone. This type of missingness is also termed **ignorable missingness**.

In some cases, however, the cause from dropping out of the study is simply because the treatment doesn't work. This heavily skews the results of the study. Indeed, if everyone for whom the treatment didn't work dropped out, then the results would be skewed towards the efficacy of the study. This type of missingness is called **missing not at random** (**MNAR**) or **non-ignorable missingness**. In this situation, missingness depends on unobservable characteristic of a person (we don't know the true reason for dropping out).

Typically, in longitudinal research, an intensive effort is made to retain participants by offering incentives for every visit and by collecting contact information not just for the participants themselves but for their relatives as well.

Example. Clinical epidemiologists are studying the association between **low-density lipoprotein** (**LDL**), also known as **bad cholesterol**, and stress level. Data are collected on 41 subjects. Their gender and age at the baseline are recorded, as well as stress level LDL level at the baseline, 6, 9, and 24 months. The stress level is measured by a questionnaire. The maximum possible score is 36, indicating the highest stress level. Below we present the analysis done in SAS and R, where we regress LDL on the other variables. Note that some participants missed a visit and some were lost to follow-up.

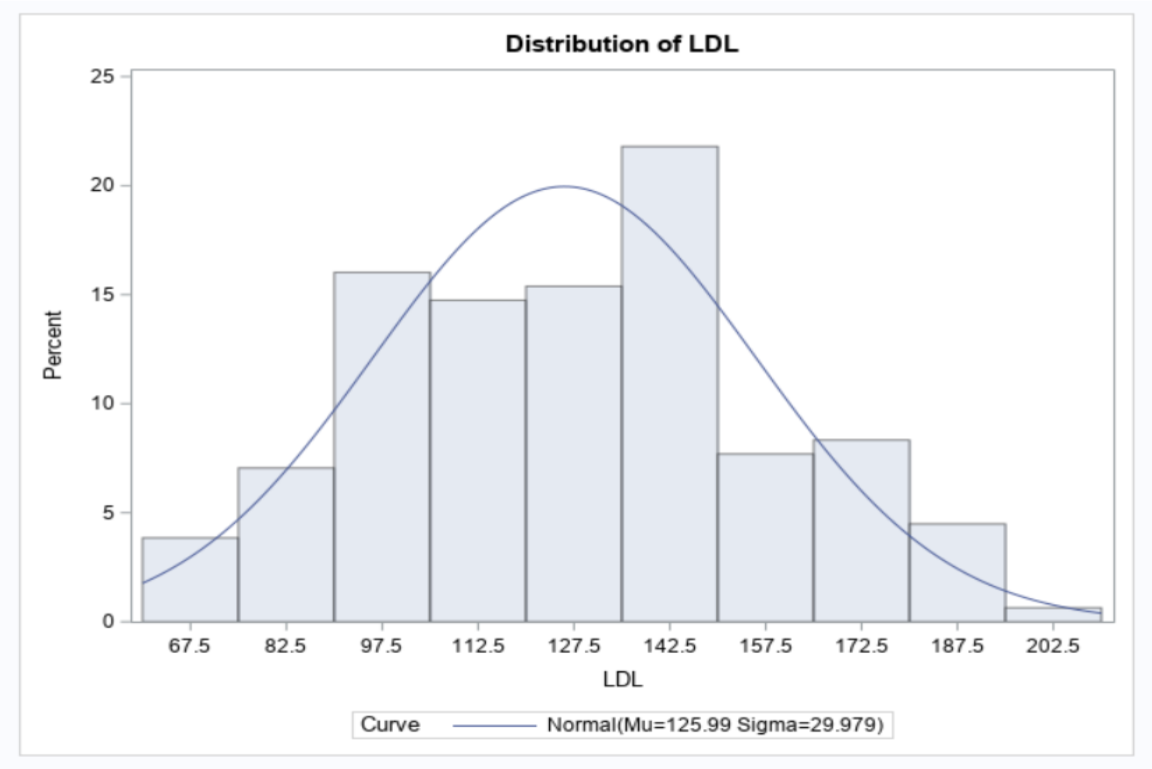
In SAS:

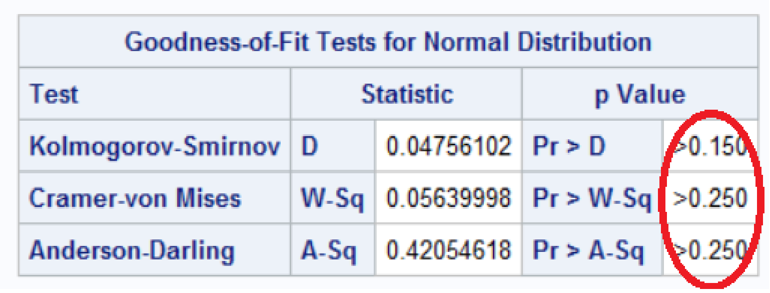
First, we create a long-form data.

```
proc import out=ldldata datafile="./LDLdata.csv" dbms=csv replace;

data longform;
 set ldldata;
   array m[4] (0 6 9 24);
    array s[4] stress0 stress6 stress9 stress24;
  array c[4] LDL0 LDL6 LDL9 LDL24;
       do i=1 to 4;
  month=m[i];
   stress=s[i];
        LDL=c[i];
     output;
    end;
keep id gender age month stress LDL;
run;
```

Next, we show that the variable LDL, which contains all the cholesterol measurements, has a normal distribution. To this end, we construct a histogram with a fitted bell-shaped curve and conduct normality testing.

```
proc univariate;
 var LDL;
  histogram LDL/normal;
run;
```





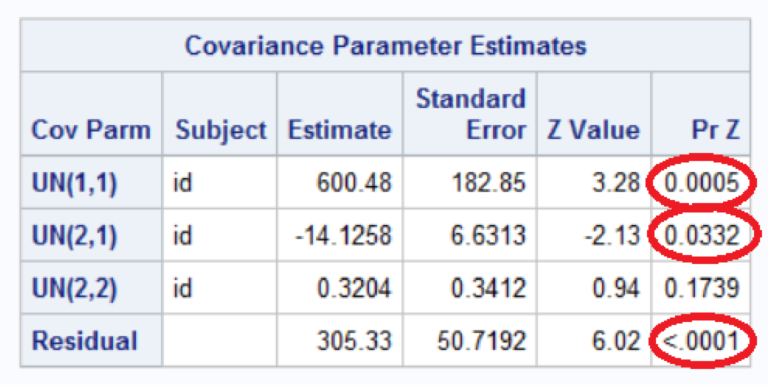
Goodness-of-Fit Tests for Normal Distribution

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.04756102	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.05639998	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.42054618	Pr > A-Sq	>0.250

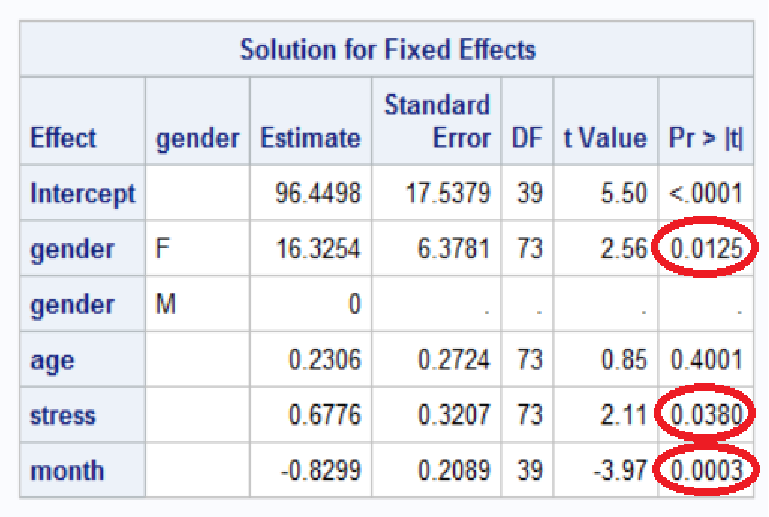
The histogram resembles a bell-shaped curve, and in the normality tests, the p-values exceed 0.05, indicating a good fit. The null hypothesis for these tests is H_0 : $measurements\ follow\ normal\ distribution$. We fail to reject the null and conclude that the data are normally distributed. Our next step is to fit a random slope and intercept model, regressing LDL on the other variables.

```
proc mixed data=longform covtest;
class gender(ref="M");
```

```
  model LDL=gender age stress month/solution;
   random intercept month/subject=id type=un;
run;
```



Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	id	600.48	182.85	3.28	0.0005
UN(2,1)	id	-14.1258	6.6313	-2.13	0.0332
UN(2,2)	id	0.3204	0.3412	0.94	0.1739
Residual		305.33	50.7192	6.02	<.0001

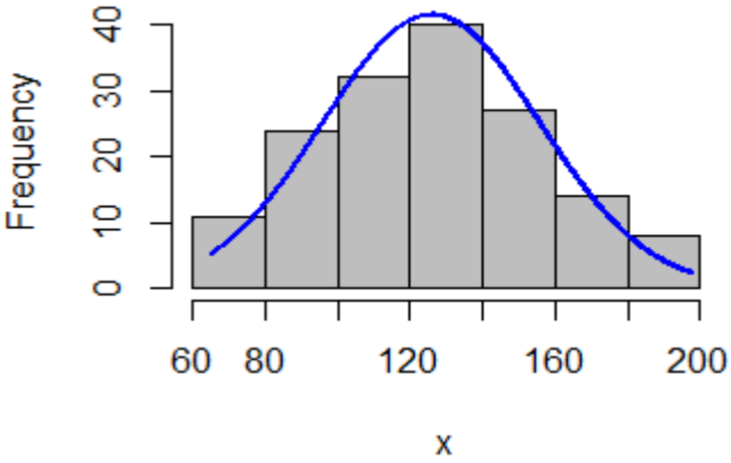


Solution for Fixed Effects						
Effect	gender	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		96.4498	17.5379	39	5.50	<.0001
gender	F	16.3254	6.3781	73	2.56	0.0125
gender	M	0	.	.	.	.
age		0.2306	0.2724	73	0.85	0.4001
stress		0.6776	0.3207	73	2.11	0.0380
month		-0.8299	0.2089	39	-3.97	0.0003

From the output, the variances of the random intercept and error, as well as the covariance term, are all statistically significant, which validates the use of all the random terms in the model. The fitted model can be written as $\widehat{E}(LDL) = 96.4498 + 16.3254 \cdot female + 0.2306 \cdot age + 0.6776 \cdot stress - 0.8299 \cdot month$, with the estimated parameters $\hat{\sigma}^2_{u_1} = 600.48$, $\hat{\sigma}_{u_1 u_2} = -14.1258$, $\hat{\sigma}^2_{u_2} = 0.3204$, and $\hat{\sigma}^2 = 305.33$. At the 5% level, gender, stress, and month are significant. For female patients, the estimated average cholesterol is 16.3254 points higher than that for males. As stress level increases by one point, the estimated average LDL increases by 0.6776 points. As time increases by one month, the estimated average LDL decreases by 0.8299 points.

In R:

```
LDL.data<- read.csv(file="./LDLdata.csv", header=TRUE, sep=",")

#creating long-form data set
library(reshape2)
data1<- melt(LDL.data[,c("id","gender","age","Stress0","Stress6",
"Stress9","Stress24")], id.vars=c("id","gender","age"),
variable.name = "stressmonth", value.name="stress")

data2<- melt(LDL.data[,c("id","LDL0","LDL6","LDL9","LDL24")],id.vars="id",
variable.name="LDLmonth",value.name="LDL")
data2<- data2[c("LDL")]

longform.data<- cbind(data1,data2)

#creating numeric variable for time
month<- ifelse(longform.data$stressmonth=="Stress0", 0,
longform.data$stressmonth=="Stress6", 6,
longform.data$stressmonth=="Stress9", 9, 24)))

longform.data$stress<- as.numeric(longform.data$stress)
longform.data$LDL<- as.numeric(longform.data$LDL)

longform.data<- na.omit(longform.data)

#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

data:  longform.data$LDL
W = 0.98626, p-value = 0.1267
```

```
#specifying reference category
longform.data$gender.rel<- relevel(as.factor(longform.data$gender), ref="M")

#cleaning long-form data
longform.data<- longform.data[!names(longform.data) %in% c("gender", "stress-
month")]

#fitting random slope and intercept model
library(nlme)

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

```
Random effects:
             StdDev    Corr
(Intercept) 24.652449 (Intr)
month        0.638271 -0.937
```

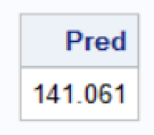
```
Residual    17.197550

Fixed effects:
               Value Std.Error  DF   t-value p-value
(Intercept) 96.11442 17.715999 113  5.425290  0.0000
gender.relF 16.23314  6.456892  38  2.514079  0.0163
age          0.24784  0.275812  38  0.898582  0.3745
stress       0.65098  0.320313 113  2.032316  0.0445
month       -0.83817  0.212233 113 -3.949310  0.0001
```

Note that R produces the same estimates for the beta coefficients as in SAS. As for the rest of the parameters, R estimates $\hat{\sigma}_{u_1} = 24.652449$, $\hat{\sigma}_{u_2} = 0.638271$, $\hat{\rho} = \dfrac{\hat{\sigma}_{u_1 u_2}}{\hat{\sigma}_{u_1}\hat{\sigma}_{u_2}} = -0.937$, and $\hat{\sigma} = 17.197550$.

To predict the LDL level at 3 months for a 60-year-old female patient with the stress score of 25, we compute $LDL^0 = 96.4498 + 16.3254 + 0.2306 \cdot 60 + 0.6776 \cdot 25 - 0.8299 \cdot 3 = 141.0615$.

In SAS:

```
data prediction;
input id gender$ age stress month;
cards;
42 F 60 25 3
;

data longform;
 set longform prediction;
run;

proc mixed;
class gender;
model LDL=gender age stress month/outpm=outdata;
random intercept month/subject=id type=un;
run;

proc print data=outdata (firstobs=165) noobs;
var Pred;
run;
```



Pred
141.061

In R:

```
#using fitted model for prediction
new.data<- rbind(longform.data,data.frame(id="NA", age=60,
stress=25, LDL="NA", month=3, gender.rel="F"))
pred<- predict(fitted.model, fitted.model, new.data, level=0)
pred[length(pred)]
```

```
140.9778
```

□