STAT 574 HOMEWORK 2 SOLUTION

1 (Identifiability)**.** For each of the following models, determine whether it is identifiable. If the model is not identifiable, provide an example of two parameters yielding the same model.

(a) Random coefficient model with the following parametrization:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{X}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \quad \text{for } i = 1, \ldots, N,$$

where $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{D})$ and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2)$.
Parameters: $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, $\boldsymbol{D}$, σ^2.

(b) Regression with categorical data:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and

$$x_i = \begin{cases} 1 & \text{if person } i \text{ is male} \\ 0 & \text{otherwise} \end{cases}, \quad z_i = \begin{cases} 1 & \text{if person } i \text{ is female} \\ 0 & \text{otherwise} \end{cases}$$

Parameters: β_0, β_1, β_2, σ^2.

(c) Gaussian mixture model:

$$\begin{aligned} z_1, \ldots, z_n &\sim \text{Categorical}(\theta_1, \theta_2, 1 - \theta_1 - \theta_2) \quad \textit{i.i.d.} \\ x_i \mid z_i &\sim \mathcal{N}(\mu_{z_i}, \sigma^2_{z_i}) \quad \text{for } i = 1, \ldots, n, \end{aligned}$$

where $z_1, \ldots, z_n$ take values in $\{1, 2, 3\}$ are latent memberships, and $x_1, \ldots, x_n$ are the observed.
Parameters: θ_1, θ_2, μ_1, μ_2, μ_3, σ_1^2, σ_2^2, σ_3^2.

Solution.

(a) The model is not identifiable because it is invariant under the following transformation:

$$\boldsymbol{\beta} \to \boldsymbol{\beta} + \boldsymbol{\delta}, \quad \boldsymbol{b}_i \to \boldsymbol{b}_i - \boldsymbol{\delta},$$

for any $\boldsymbol{\delta}$ of the same size of $\boldsymbol{b}_i$.

(b) The model is not identifiable because it is invariant under the following transformation:

$$\beta_0 \to \beta_0 + \delta, \quad \beta_1 \to \beta_1 - \delta, \quad \beta_2 \to \beta_2 + \delta,$$

for any $\delta \in \mathbb{R}$.

(c) The model is not identifiable because it is invariant under any permutation of the three labels.

2 (Information matrix)**.** Consider the linear mixed effect model $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$ in the absence of random effects (i.e. $\boldsymbol{D} = \boldsymbol{0}$). Determine the Fisher's information matrix and the variances for the MLEs $\hat{\sigma}^2$ and $\hat{\boldsymbol{\beta}}$.

Solution.
When $\boldsymbol{D} = \boldsymbol{0}$, the model reduces to a linear regression model. The log-likelihood function is

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2}\left\{N_T \log \sigma^2 + \sigma^{-2}\sum_{i=1}^{N} \|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\|^2\right\}.$$

The expected values for second-order partial derivatives are

$$\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^2}\right] = -\sigma^{-2}\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{X}_i$$

$$\mathbb{E}\left[\frac{\partial^2 \ell}{\partial (\sigma^2)^2}\right] = \frac{N_T}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{N} \mathbb{E}\left[\|\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}\|^2\right] = -\frac{N_T}{2\sigma^4}$$

$$\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \sigma^2 \partial \boldsymbol{\beta}}\right] = \mathbf{0}.$$

Therefore, we have the Fisher's information matrix as

$$\mathcal{I}(\boldsymbol{\beta}, \sigma^2) = \begin{bmatrix} \sigma^{-2}\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{X}_i & \mathbf{0} \\ \mathbf{0}^T & \frac{N_T}{2\sigma^4} \end{bmatrix}.$$

Since the variance of MLE is inversely proportional to the information matrix, we have

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{X}_i\right)^T, \quad \mathrm{Var}(\hat{\sigma}^2) = \frac{2\sigma^4}{N_T}.$$

3 (Programming)**.** Use the Weight v.s. Height data. Consider intercept, Height, and Sex in fixed effects, and consider intercept only for random effect. That is:

$$W_{ij} = \beta_0 + \beta_1 H_{ij} + \beta_2 S_{ij} + b_i + \epsilon_{ij}.$$

(a) Fit the model.

(b) Get the estimated random effects $\hat{b}_1, \ldots, \hat{b}_N$. Plot the histogram for the random effects.
Hint: the `coef()` function gives you intercept values for $\beta_0 + b_i$

(c) Test whether the random effect is significant.
Hint: F-test. check code in the lecture note

(d) Obtain the Wald CI for β_1.
Hint: p-values can be read from `summary()` function.

(e) Obtain the PL CI for β_1.
Hint 1: profiled LLH for $\beta_1 = x$ can be calculated as the LLH for the following LME model:

$$W_{ij} - xH_{ij} = y_{ij} = \beta_0 + \beta_2 S_{ij} + b_i + \epsilon_{ij},$$

where y_{ij} is the new observation.
Hint 2: use `uniroot()` function to find zeros of a function.

Solution.

(a) We fit the model with (lmr) function.

```
data = read.table("./Data/MixedModels/Chapter02/Family.txt", header=T,
    stringsAsFactors=F)
head(data)
library(nlme)

fit.lme = lme(fixed=Weight~Height+Sex, random=~1|FamilyID, data=data)
print(fit.lme)
```

The fitted parameters are

$$\hat{\beta}_0 = -54.84, \quad \hat{\beta}_1 = 2.93, \quad \hat{\beta}_2 = 24.17, \quad b_i \sim \mathcal{N}(0, 12.86^2), \quad \epsilon_i \sim \mathcal{N}(0, 24.05^2)$$

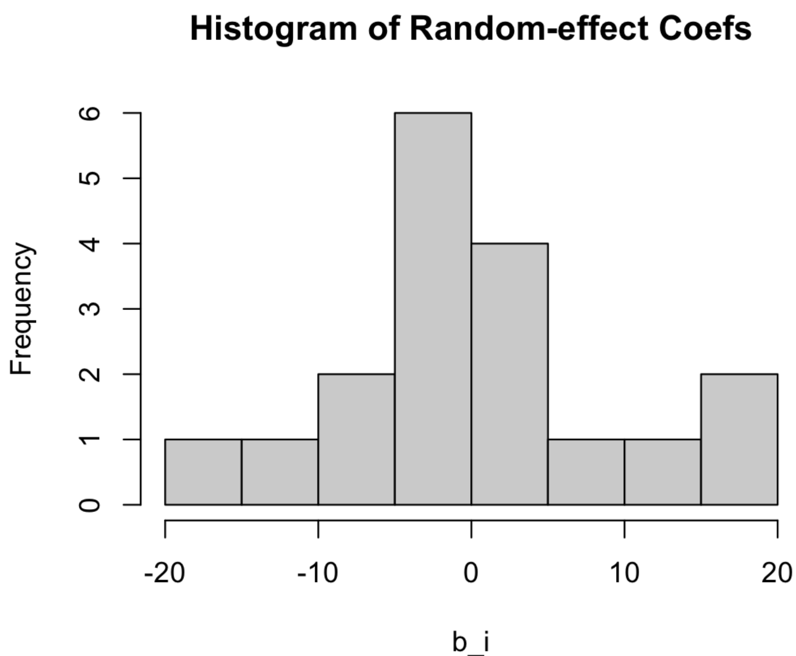
(b) We first get the estimated random effects:

```
print(fit.lme$coefficients$random$FamilyID)
```

The output is

```
    (Intercept)
1    15.7598410
2    -0.9131247
3    12.7375849
4   -18.3523846
5   -11.0695495
6     4.5099808
7    -4.8281441
8    -3.1497331
9     4.2073173
10   -4.0666925
11   -3.1371132
12    0.1708203
13   -5.7063262
14   16.6605668
15    7.9734679
16   -3.9186066
17   -8.0368012
18    1.1588968
```

We can plot the histogram as follows.



(c) We use F-test in the following code.

```
library(Matrix)
Z = as.matrix(bdiag(split(rep(1, dim(data)[1]), data$FamilyID)))
fit0 = lm(data$Weight ~ data$Height + data$Sex)
fit1 = lm(data$Weight ~ 0 + data$Height + data$Sex + Z)
anova(fit0, fit1)
```

The output is

```
Analysis of Variance Table

Model 1: data$Weight ~ data$Height + data$Sex
Model 2: data$Weight ~ 0 + data$Height + data$Sex + Z
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     68 50288
2     51 29784 17     20504 2.0652 0.02384 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Therefore, at 5% significance level, the random effect is present.

(d) The estimate and variance of β_1 can be read from the summary table.

```
Linear mixed-effects model fit by REML
  Data: data
       AIC      BIC    logLik
  661.0278 672.1254 -325.5139

Random effects:
 Formula: ~1 | FamilyID
        (Intercept) Residual
StdDev:    12.86171 24.05039

Fixed effects:  Weight ~ Height + Sex
                Value Std.Error DF    t-value p-value
(Intercept) -54.83871  80.15014 51 -0.6841998  0.4969
Height        2.93276   1.22667 51  2.3908287  0.0205
Sex          24.16578   9.82833 51  2.4587871  0.0174
 Correlation:
       (Intr) Height
Height -0.998
Sex     0.758 -0.788

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-1.9338849 -0.5686201 -0.1252396  0.2917410  3.7230392

Number of Observations: 71
Number of Groups: 18
```

The Wald CI for β_1 is

$$(2.93 - 1.23t_{51,0.975}, 2.93 + 1.23t_{51,0.975}) = (0.47, 5.40)$$

Or it can be directed read from the following code.

```
intervals(fit.lme)
```

```
Approximate 95% confidence intervals

 Fixed effects:
                   lower       est.      upper
(Intercept) -215.7468327 -54.838712 106.069410
Height         0.4701158   2.932763   5.395409
Sex            4.4345769  24.165782  43.896986
```

```
 Random Effects:
  Level: FamilyID
                     lower     est.    upper
sd((Intercept)) 6.651787 12.86171 24.86903

 Within-group standard error:
   lower     est.    upper
19.85338 24.05039 29.13467
```

(e) In order to get the PL CI, we first define the function for the profiled log-likelihood function.

```r
pllh = function(x){
    data$newy = data$Weight - x * data$Height
    return(lme(fixed=newy~Sex, random=~1|FamilyID, method="ML", data)$logLik)
}
```

Then we use `uniroot` function to find the two boundary point of the PL CI.

```r
fit.ml = lme(fixed=Weight~Height+Sex, random=~1|FamilyID, method='ML', data)

obj = function(x){
    return(fit.ml$logLik - 0.5 * qnorm(0.975) ** 2 - pllh(x))
}

beta_hat = fit.ml$coefficients$fixed['Height']
lb = uniroot(obj, c(-10, beta_hat))$root
rb = uniroot(obj, c(beta_hat, 100))$root
print(c(lb, rb))
```

Therefore, the PL CI for β_1 is $(0.46, 5.34)$.