Homework 4 - BIOS 6643 Analysis of Longitudinal Data

Dominic Adducci

2023-09-20

Question 1

Complex MLE estimation in LMM requires computational optimization approaches, the goal here is to implement a basic Newton-Raphson algorithm in R. The following data are an i.i.d sample from a Cauchy(θ,1) distribution: 1.77, -0.23, 2.76, 3.80, 3.47, 56.75, -1.34, 4.24, -2.44, 3.29, 3.71, -2.40, 4.53, -0.07, -1.05, -13.87, -2.53, -1.75, 0.27, 43.21.

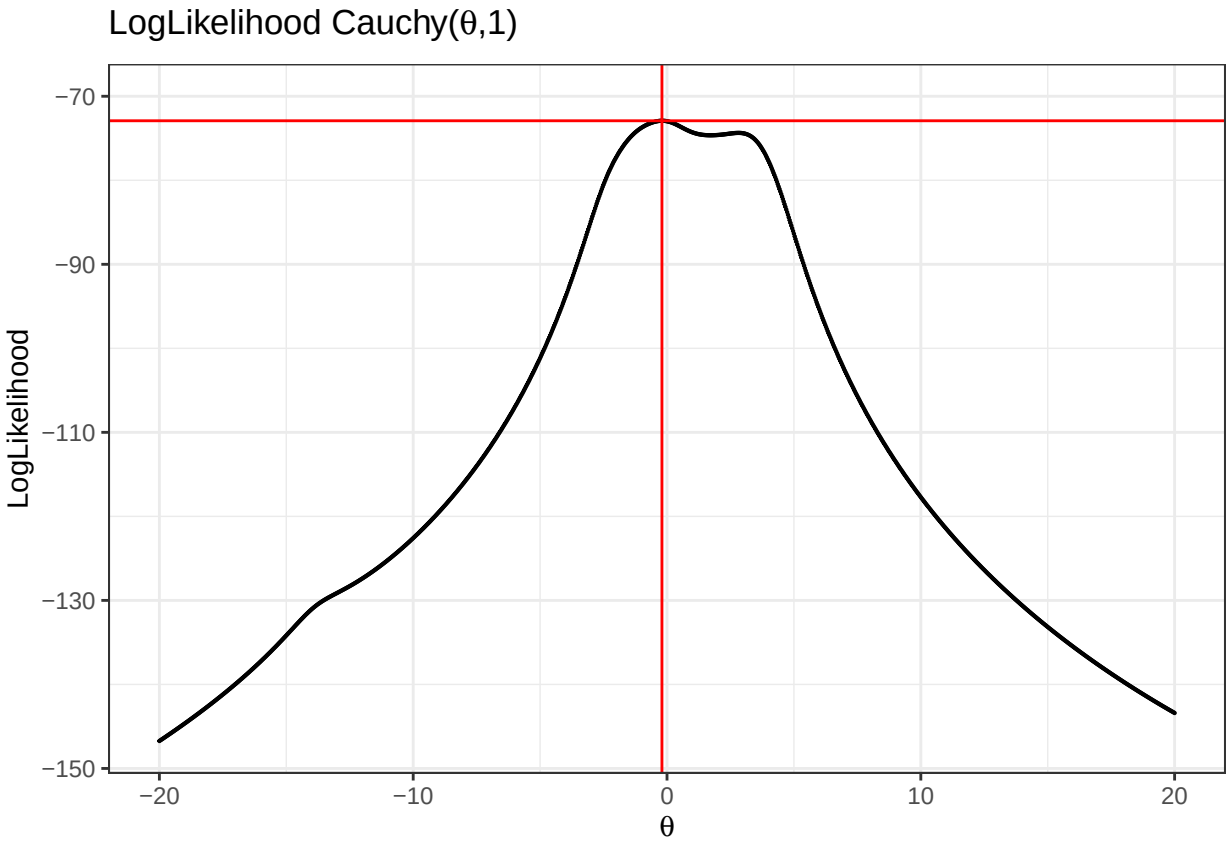
Part A: Graph the log likelihood function.

The likelihood function of the Cauchy(θ,1) distribution is:

$$\frac{1}{\pi^n}\frac{1}{\prod[1+(x_i-\theta)^2]}$$

and the loglikelihood is:

$$-nlog(\pi)-\sum_i^n log(1+(x_i-\theta)^2)$$



From the plot of log-likelihood the maximum log-likelihood values is -72.92, and the optimal theta is -0.2.

Part B

Find (and write an R program) to find the MLE for θ using the Newton-Raphson method.

The equation for the Newton-Raphson method in general form is as follows:

$$x_i = x_{i-1} - \frac{f(x_{i-1})}{f'(x_{i-1})}$$

Translating this to finding the MLE of θ:

$$\hat{\theta}_i = \hat{\theta}_{i-1} - \frac{logL'(\hat{\theta}_{i-1})}{logL''(\hat{\theta}_{i-1})} = \hat{\theta}_{i-1} - \frac{S(\hat{\theta}_{i-1})}{I(\hat{\theta}_{i-1})}$$

What finding an MLE the score $S(\theta)$, the derivative of the log-likelihood, should equal 0, $S(\theta) = 0$. $I(\theta)$ is the Fisher observed information from the data.

Check the code appendix for the function.

```
##    Estimate Difference
## 1  2.938901  -1.061099
## 2  2.834520  -1.165480
## 3  2.817895  -1.182105
## 4  2.817472  -1.182528
## 5  2.817472  -1.182528
```

```
## 6  2.817472  -1.182528
## 7  2.817472  -1.182528
## 8  2.817472  -1.182528
## 9  2.817472  -1.182528
## 10 2.817472  -1.182528
## 11 2.817472  -1.182528
## 12 2.817472  -1.182528
## 13 2.817472  -1.182528
## 14 2.817472  -1.182528
## 15 2.817472  -1.182528
## 16 2.817472  -1.182528
## 17 2.817472  -1.182528
## 18 2.817472  -1.182528
## 19 2.817472  -1.182528
## 20 2.817472  -1.182528
```

Part C Try all of the following starting points: -11,-1,0,1.5,4,4.7,7,8,and 38.

Question 2

In a paragraph explain the difference between a general linear model or multiple regression (GLM; not a generalized linear model like a logistic regression or (GLM; not a generalized linear model like a logistic regression or Poisson, which will be discussed later) and a linear mixed model (LMM).

A general linear model only has fixed effects, while a linear mixed model includes both fixed and random effects. The random effects of the linear mixed model allows the model to account for differences between subjects. A simple example would be measuring something like cholesterol levels through time. A general linear model (fixed effects only) may include covariates such as time, BMI, smoking status, sex, race, etc. Every individual will follow the same regression line based on those covariates. In a linear mixed model random effect which account for subject differences, such as someone tending to have higher or lower cholesterol at start (random intercept) can be included to better model change over time. A better fit may be found by including random slopes as well for each subject if cholesterol trajectory is found to be different between subjects.

Question 3

In a short paragraph, explain the difference between a profiled likelihood and a restricted likelihood for a linear mixed model, and how and why they are used.

In a profile likelihood you maximize the likelihood by fixing every other parameter and only allowing one to vary. Doing this maximizes the single parameter you allowed to vary. You can then repeat this process for every other parameter incrementally, plugging in the estimates for parameters which have already been maximized. The downside to this method is that variance estimates are biased downward. The restricted likelihood (REML) allows for estimating parameters which are not biased regardless of sample size. The downside for the REML method compared to profile likelihood is that you can only compare REML model using a likelihood ratio of both models have the same set of fixed effects.

Profile likelihood - maximize likelihood by fixing every other parameter and only allow one to vary and maximize that using a grid search. After getting that maximum repeat each process for each parameter in the set of likelihood parameter.

Restricted maximum likelihood - REML has property that your standard error are unbiased regardless of sample size, so generally is in situations where were trying to get unbiased estimates of errors when sample

sizes are small. Loglikelihood are not valid for reduced in full REML models. Can only compare REML models that have the same set of fixed effects.

Part 2ish

Investigator wants to understand whether Cortisol (a stress hormone) secretion differs in women suffering from depression. Cortisol was measured every 10 minutes for a period of 24 hours starting at 9 am. 26 patients and 26 controls were collected in the study. Although the data were collected every 10 minutes for a period of 24 hours on each subject (144 observations), the investigators were interested in differences in the circadian pattern between the groups. Data was divided into 6 blocks of 4 hours and averaged to obtain a set of "block means".

Question 4

Part A Fir a multiple linear regression to investigate how mean cortisol values change over the day (categorical time) and how the average cortisol levels differ by group (no interaction for this model). This will be used to anchor the comparisons later in the assignment.

Table 1: MLR Cortisol

Term	Estimate	Std.Error	95% Conf.Low	95% Conf.High	P-Value
(Intercept)	3.0501	0.5088	2.0489	4.0513	5.7451e-09
timeTime2	3.9697	0.6662	2.6589	5.2806	6.9911e-09
timeTime3	10.0791	0.6662	8.7682	11.3899	$<$ 2.22e-16
timeTime4	6.1979	0.6662	4.8870	7.5087	$<$ 2.22e-16
timeTime5	4.4042	0.6662	3.0934	5.7151	1.7017e-10
timeTime6	1.7518	0.6662	0.4410	3.0627	0.0089779
casecontrolp	0.9211	0.3846	0.1643	1.6779	0.0172285

Table 1 shows the output of the multiple linear regression where mean cortisol is the outcome and time and casecontrol are covariates. Time is factored into 6 different levels, where time 1 is the reference.

Part B Provide a table of mean differences from the 6th time period along with SE's of the differences. Interpret two of the coefficients. You do not need to conduct inference.

Table 2: MLR Cortisol - Time 6 Reference

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	4.8020	0.5088	3.8008	5.8031	$<$ 2.22e-16
timeTime1	-1.7518	0.6662	-3.0627	-0.4410	0.00897790
timeTime2	2.2179	0.6662	0.9071	3.5287	0.00097714
timeTime3	8.3273	0.6662	7.0164	9.6381	$<$ 2.22e-16
timeTime4	4.4461	0.6662	3.1352	5.7569	1.1726e-10
timeTime5	2.6524	0.6662	1.3415	3.9632	8.5630e-05
casecontrolp	0.9211	0.3846	0.1643	1.6779	0.01722849

Table 2 shows the output of the multiple linear regression where time 6 is the reference group. For term timeTime1 this is the mean difference between cortisol at time 1 and cortisol at time 6. Interpreting this it means cortisol is 1.7518 units higher at time 6 compared to time 1. For the term casecontrolp this is the mean difference between the control (reference) group and women with depression. Interpreting this it means cortisol is 0.9211 units higher for women with depression compared to women without depression (control).

Part C Will these standard error be too big or small and why?

Standard error is found with the equation:

$$Std.Error = \frac{\sigma}{\sqrt{n}}$$

For a study at a single instance n is generally the number of subjects. For this study because subjects each have 6 time points (assuming balanced data) n refers to all instances where cortisol was measured for each subject, 52×6, which mean $n = 312$.

Table 3: MLR Cortisol - Time 6 Ref. Variances

Intercept	Time 1	Time 2	Time 3	Time 4	Time 5	Group P
1.294317	2.218829	2.218829	2.218829	2.218829	2.218829	0.7396097

Table 3 shows the variances for each estimate. Given that each estimate is relatively low compared to the variances the standard errors (which were used to calculate variance) may be too large to make a meaningful interpretation, as subjects may have significantly different mean cortisol measurements between each other.

```
# Extracting covariance matrix from gls
#lmm <- gls()
#lmm_summ <- summary(lmm)
#varcov <- lmm_summ$modelStruct$corStruct
```