1. **Best linear unbiased prediction and empirical Bayes.** In this problem, you will demonstrate several assertions in the notes having to do with the **empirical Bayes estimator/best linear unbiased predictor** for $b_i$,

\[
\hat{b}_i = \hat{D}Z_i^T V_i^{-1}(\xi, x_i)\{Y_i - X_i\hat{\beta}\}
\]

in (6.55) in the usual form of the linear mixed effects model

\[
Y_i = X_i\beta + Z_ib_i + e_i,
\]

where $b_i$ and $e_i$ are taken to be mutually independent and independent of $x_i$ for each $i$, and iid across $i = 1, ..., m$ with

\[
b_i \sim \mathcal{N}(0, D), \quad e_i \sim \mathcal{N}(0, R_i(\gamma)), \quad Y_i|x_i, b_i \sim \mathcal{N}(X_i\beta + Z_ib_i, R_i(\gamma)).
\]

Write $R_i = R_i(\gamma)$ and $V_i = V_i(\xi, x_i) = Z_iDZ_i^T + R_i$ for brevity.

(a) Treating $\beta$, $D$, and $\gamma$ as fixed and known, show that the posterior density

\[
p(b_i|y_i, x_i; \beta, \gamma, D) = \frac{p(y_i|x_i, b_i; \beta, \gamma) p(b_i; D)}{\int p(y_i|x_i, b_i; \beta, \gamma) p(b_i; D) db_i}
\]

in (6.53) is normal with mean

\[
DZ_i^T V_i^{-1}\{y_i - X_i\hat{\beta}\}
\]

in (6.54) and find its covariance matrix by evaluating (1) (including the integral in the denominator).

(b) Treating $D$ and $\gamma$ as known, show that, conditional on $\tilde{x}$,

\[
\hat{b}_i = DZ_i^T V_i^{-1}(\xi, x_i)\{Y_i - X_i\hat{\beta}\}
\]

in (6.56) has mean zero and covariance matrix

\[
\text{var}(\hat{b}_i|\tilde{x}) = DZ_i^T \left\{ V_i^{-1} - V_i^{-1}X_i \left( \sum_{i=1}^{m} X_i^TV_i^{-1}X_i \right)^{-1} X_i^TV_i^{-1} \right\} Z_i D
\]

as in (6.57).

(c) Show that for $\hat{b}_i$ in (2),

\[
\text{var}(\hat{b}_i - b_i|\tilde{x}) = D - DZ_i^T \left\{ V_i^{-1} - V_i^{-1}X_i \left( \sum_{i=1}^{m} X_i^TV_i^{-1}X_i \right)^{-1} X_i^TV_i^{-1} \right\} Z_i D
\]

as in (6.58).
2. Theophylline pharmacokinetics for a single subject. In the file theo5.dat on the course webpage are concentration-time data for one of the participants in the pharmacokinetic study of theophylline in Example 4 of Chapter 1. This subject was given a dose of \( D = 5.86 \) mg/kg of theophylline orally at time 0, and blood samples were taken at the \( n = 10 \) times \( t_j, j = 1, \ldots, n \) (in hours) in the first column of the data set; the corresponding theophylline concentrations (mg/L) are in the second column.

As discussed in the Chapter 1, a standard model for individual pharmacokinetics following oral administration of theophylline is the one compartment open model with first order absorption and elimination given in (1.3). In this problem, as conventional, we will take the bioavailability \( F = 1 \), so that the model involves the three parameters \( k_a \), the fractional rate of absorption; \( Cl \), clearance rate; and \( V \), the volume of distribution. Letting \( x_j = (D, t_j), j = 1, \ldots, n \), assume that

\[
E(Y_j | x_j) = f(x_j, \beta) = \frac{De^{\beta_1}}{e^{\beta_3} - e^{\beta_2}/e^{\beta_3}} \{ \exp(-e^{\beta_2}t_j/e^{\beta_3}) - \exp(-e^{\beta_1}t_j) \}, \quad \beta = (\beta_1, \beta_2, \beta_3)^T,
\]

so that \( k_a = \exp(\beta_1) \), \( Cl = \exp(\beta_2) \), and \( V = \exp(\beta_3) \). The parameterization of the model in (3) is different from that in (7.4) and is often adopted to enforce positivity of the pharmacokinetic parameters and for reasons we will discuss in Chapter 9. Assume further that we adopt the power of the mean variance model

\[
\text{var}(Y_j | x_j) = \sigma^2f^{2\delta}(x_j, \beta)
\]

for \( \delta > 0 \) and that the sampling times \( t_j \) are sufficiently intermittent that serial correlation among the \( Y_j \) is negligible.

(a) Under this model, use the generalized least squares algorithm (as implemented in the R function \texttt{gls()} in the \texttt{nlme} package or using SAS (e.g., the macro using \texttt{proc nlin} on the course website) to fit this model to these data and obtain estimates of the parameters \( k_a \), \( Cl \), and \( V \).

\textit{Hint: Starting values} are required for \( \beta \) and \( \delta \). Finding starting values for nonlinear models is model-dependent and a bit of an art form. Here, you can derive a starting value for \( k_a \), and thus \( \beta_1 \), by considering the fact that the early concentrations (up to the peak concentration) are dominated by the absorption process. Similarly, a starting value for \( k_e \) can be derived by noting that the later concentrations (after the peak) are dominated by the elimination process. A starting value for \( V \), and thus \( \beta_3 \), can be derived by noting that, given \( k_a \) and \( k_e \) and a concentration value at any time, (3) can be “solved” for \( V \). Starting values for \( Cl \) and thus \( \beta_2 \) follow.

(b) Obtain approximate standard errors to accompany your estimates of \( k_a \), \( Cl \), and \( V \).

\textit{Hint:} The model (3) is parameterized in terms of \( \beta \) such that \( k_a = \exp(\beta_1) \), \( Cl = \exp(\beta_2) \), and \( V = \exp(\beta_3) \), so the standard errors output by the software are standard errors for \( \beta \), not for \( k_a \), \( Cl \), and \( V \). Thus, you need approximate standard errors for a \textit{transformation} of the original parameters.

3. Iterative algorithm for solution of general estimating equations. Suppose we have a general set of estimating equations of the form in (7.38) of the notes, i.e.,

\[
\sum_{j=1}^{n} D_j^T(\eta)V_j^{-1}(\eta)(s_j(\eta) - m_j(\eta)) = 0,
\]

(4)
Suppose that $\eta$ can be partitioned as $\eta = (\beta^T, \theta^T)^T$ and that $s_j(\eta)$ depends on $\eta$ only through the expression $\{Y_i - f(x_i, \beta)\}^2$. By taking a linear Taylor series of the left hand side of (4) about $\eta = \eta^*$, where $\eta^*$ is “close to” $\eta$, derive an iterative algorithm for solving (4) similar to that in (7.24). That is, defining

$$ s(\eta) = \{s_1^T(\eta), ..., s_m^T(\eta)\}^T, \quad m(\eta) = \{m_1^T(\eta), ..., m_m^T(\eta)\}^T, $$

$$ \mathcal{V}(\eta) = \text{block diag}\{\mathcal{V}_1(\eta), ..., \mathcal{V}_m(\eta)\}, \quad \mathcal{D}(\eta) = \{\mathcal{D}_1(\eta), ..., \mathcal{D}_m(\eta)\}, $$

show that the $(a + 1)$th iterate can be obtained as

$$ \eta_{a+1} = \eta_a + \{\mathcal{D}_a^{-1} \mathcal{V}_a^{-1} \mathcal{D}_a\}^{-1} \mathcal{D}_a^{-1} \mathcal{V}_a^{-1} \{s_a - m(a)\}, $$

where $\mathcal{D}(a) = \mathcal{D}(\eta(a))$ and similarly for the other components. Be sure to justify any terms you regard as negligible in your derivation.

4. **Alternative form of the quadratic estimating equation for $\xi$.** Suppose $Y_i = (Y_{i1}, Y_{i2}, Y_{i3})^T$, where each $Y_{ij}$ is a count and the observations are taken at equally-spaced times $(t_1, t_2, t_3)$ for all $i = 1, ..., m$. Suppose further that there are no among- or within-individual covariates, and that we assume the population mean model

$$ E(Y_{ij}) = \exp(\beta_1 + \beta_2 t_j) = f(t_j, \beta), \quad \beta = (\beta_1, \beta_2)^T, $$

population variance model

$$ \text{var}(Y_{ij}) = \sigma^2 f(t_j, \beta), $$

and that the overall correlation structure of $Y_i$ is AR(1) with lag-1 correlation parameter $\alpha$.

Consider solving the quadratic estimating equation (8.17) to estimate the covariance parameter $\xi = (\sigma^2, \alpha)$. Give the vectors $u_i$ and $v_i(\beta, \xi)$, the gradient matrix $\mathbf{E}_i(\beta, \xi)$, and the “covariance matrix” $\mathbf{Z}_i(\beta, \xi)$ under the “Gaussian working assumption” such that the quadratic estimating equation (8.17) can be written in the alternative form (8.22).

5. **Skin cancer prevention trial.** The data in the file `skincancer.dat` on the course webpage are from a clinical trial to study the effectiveness of beta-carotene to prevent non-melanoma skin cancer in high risk subjects with previous skin cancer. The data are from $m = 811$ subjects recruited from 4 medical centers who were randomized to 50 mg beta-carotene per day (coded as 1) or placebo (coded as 0), each of whom miraculously returned to the clinic at years 1, 2, 3, 4, and 5 after starting on his/her assigned treatment. Several baseline variables were collected, as shown below; the variable exposure is a count of the number of previous skin cancers experienced by each subject at baseline. The outcome is the binary indicator of presence of skin cancer. The data set has the following columns:

1. subject ID
2. center (1, 2, 3, 4)
3. age at baseline (years)
4. skin type (1 = burns, 0 = no burns)
5. gender (0 = female, 1 = male)
6. exposure
7. presence of skin cancer (outcome, 0 = no, 1 = yes)
8. treatment (0 = placebo, 1 = beta-carotene)
9. year (1, 2, 3, 4, 5)
The investigators are interested in the following questions: (i) Is there evidence that the probability of new skin cancers changes over the study period for either treatment? (ii) Does beta-carotene lead to a lower probability of new skin cancers than placebo after five years in this population? What is the probability of experiencing new skin cancers at year 5 for each treatment? (iii) Is there evidence that the probability of experiencing new skin cancers over the study period is associated with age, gender, previous cancer, or center?

Using methods in Chapter 8 of the notes, carry out analyses to address these questions and write a brief report summarizing what you did and the results. Be sure to describe how you interpreted and formalized the questions of interest. Comment on how confident you feel about the reliability of the inferences and conclusions.

Please turn in your code and output along with your report (you can edit the output to include only the portions that pertain directly to your report and embed it in your report if you like).