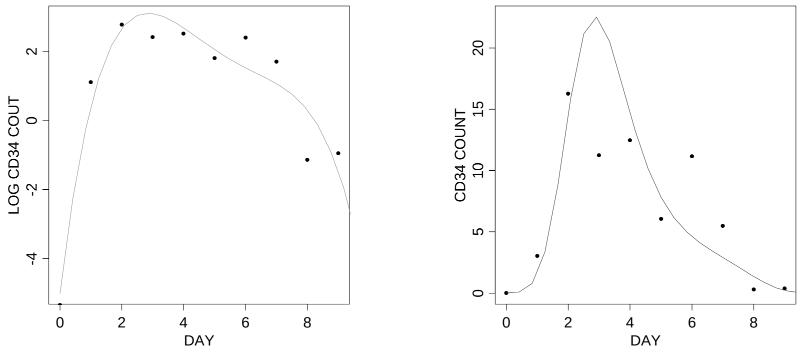
STA 214 – Midterm 2

1. In this problem we consider expected utility optimization in a longitudinal data model with random effects. The application concerns optimal scheduling of stem cell collections for chemotherapy patients.

Background: (from Palmer & Mueller 1994) Reconstitution of white blood cell components after high-dose chemoradiotherapy is an essential need of patients undergoing that type of treatment for cancer. This reconstitution can be achieved either by bone marrow transplant or by transfusion of stem cells (CD34) collected from the circulating blood. Stem cell transfusion has the advantage of having a more rapid reconstitution of white blood cells, thus reducing the length of time the patient is at high risk for potentially lethal infection. It is also an option for patients considered ineligible for bone marrow transplant. In order to maximize the usefulness of blood-derived stem cells, however, two issues must be resolved: the timing and number of collection procedures for these stem cells.
Pre-treatment regimens for patients scheduled for high-dose chemoradiotherapy often result in a transient 10-fold or more increase in the concentration of circulating blood stem cells, which can then be collected from the patient's blood.



Typical profile of CD34 counts for a patient. The left panel shows the log counts as a function of days after pre-treatment. The right panel shows the actual counts.

During the stem cell collection process (apheresis) blood is drawn from the patient then centrifuged and its components separated. The white blood cell components which contain the stem cells are collected and remaining cells are returned to the patient. The collected stem cells are frozen for later use by the patient after high-dose chemoradiotherapy treatment for cancer. Each apheresis procedure requires about three hours, although its duration may vary, based on the fraction of the patient's blood volume which is processed. The initial apheresis is currently scheduled to occur on the third day after an initial pre-treatment regimen and additional aphereses are then scheduled each day for five days unless the target number of cells is collected earlier.
The total blood stem cell yield per apheresis may depend on variables such as: blood volume

processed, separation efficiency of the device used, centrifuge speed, hematocrit and white blood cell counts before apheresis, disease status at time of apheresis and response to pre-treatment. The highest concentration of circulating blood stem cells is most often observed in patients who have received minimal or no prior chemoradiotherapy.

Model: We will consider a stylized version of this problem. Let y_{ij} denote the log CD34 (stem cell) count of patient i on day j, $i = 1, \ldots, I$, $j = 1, \ldots, J$ with $I = 15$, $J = 10$. Let $x_{ij} = j$ denote the day of the j-th observation on the i-th patient (in this example have data for each patient for each day, i.e. $x_{ij} = j$; in general it could be different).

Consider the following normal/linear model for y_{ij}:

$$\begin{aligned} y_{ij} &= \beta_{i0} + \beta_{i1} x_{ij} + \beta_{i2} x_{ij}^2 + \beta_{i3} x_{ij}^3 + \beta_{i4} x_{ij}^4 + \epsilon_{ij}, \\ & \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \\ \beta_i &\sim N(\mu, \Sigma), \\ p(\mu) &\propto 1 \end{aligned}$$

with $\sigma^2 = 1.0$ and $\Sigma = I_5$.

Data: The data y_{ij} is linked to the 214 homepage. The file is a (15×10) matrix with y_{ij} in the i-th row and j-th column.

1a. Find $p(\beta|\mu, y)$ and $p(\mu|y)$. Let $\bar{\beta}_{ij}$ denote the marginal posterior mean of β_{ij}, $i = 1, \ldots, I$, $j = 0, \ldots, 4$. Show some summaries of the $\bar{\beta}_{ij}$'s (like histograms, summary statistics, etc.).

1b. Plot $E(y_{I+1,j}|y)$ as a function of $x_{I+1,j} = t$, $t = 0, \ldots, 10$.
Note: It is probably easiest to do this by simulation, i.e., simulate a large number, say $M = 100$, of samples from $p(y_{I+1,j}|y)$, and then just take sample averages over the simulations to approximate the posterior predictive mean function $E(y_{I+1,j}|y)$.

1c. We want to determine the optimal apheresis schedule for a future patient $i = I + 1$ (before observing any data on the patient). We consider designs d of the type: "schedule stem cell collections starting on day d with additional apheresis on each following day until either the target number z^* is collected, or day 10 (whichever comes first)." Assume $z^* = 40$ is the target number. Possible designs are $d \in \{1, 2, \ldots, 10\}$.
To determine an optimal schedule we define a utility function: Assume we use design d and observe data y_{I+1}. Assume that m is the number of resulting stem cell collections (easy to determine if you have d and $y_{I+1,j}, j = 1, \ldots, 10$). The utility is

$$u(d, y_{I+1}) = -m + C \cdot I\left(\sum_{j=d}^{d+m-1} exp(y_{I+1,j}) > z^*\right).$$

where $I(A)$ is an indicator for event A and $C = 100$ is a reward for collecting the targeted amount z^* (remember that y_{ij} are log counts).
For each $d \in \{1, \ldots, 10\}$ evaluate the expected utility

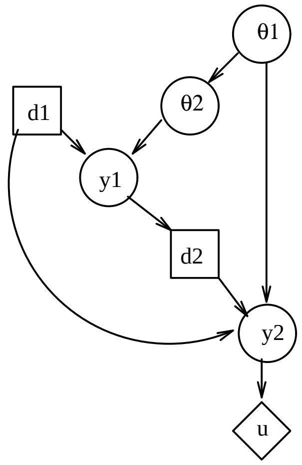
$$U(d) = \int u(d, y_{I+1}) p(y_{I+1}|y) dy_{I+1}.$$

To evaluate $U(d)$ you can use the posterior predictive simulations from **1b.** Denote with y_{I+1}^t, $t = 1, \ldots, M$ the t-th predictive simulation. You can approximate $U(d)$ by

$$\hat{U}(d) = \frac{1}{M} \sum_{t=1}^{M} u(d, y_{I+1}^t.$$

Plot $\hat{U}(d)$ as a function of d to find the optimal design.

2. The oil wildcatter problem (Raiffa, 1968). The value (u) in this problem is a function of the decision maker's profit (y_2), which in turn depends on the decision to drill ($d_2 \in \{0, 1\}$), the presence of oil to be found ($\theta_1 \in \{0, 1\}$), and whether or not a test is performed ($d_1 \in \{0, 1\}$). At the time the wildcatter must choose whether to drill, s/he knows the results of the test s/he ordered ($y_1 \in \{0, 1, 2\}$). Here $y_1 = 0$ is defined as the value of y_1 when no test was ordered. If a test was ordered then $y_1 \in \{1, 2\}$ is a (noisy) determination of the seismic structure ($\theta_2 \in \{1, 2\}$), so the result y_1 depends both on the testing decision and the seismic structure. The seismic structure itself is dependent on the presence of oil (θ_1). Finally, there is no observation available before the testing decision must be made. The following ID represents the wildcatter problem:



2a. In the above formulation the decision d_2 depends on the outcome y_1. Replace the influence diagram (ID) by an equivalent non-sequential problem.

2b. Describe how you would solve the ID from **2a** by a sequence of "averaging out" (= marginalization), "arc reversal" (= Bayes' theorem) and maximization. I.e., discuss how to find the optimal decision for d_1 and d_2. Draw a sequence of ID's (like on page 13 of Shenoy's paper).[1]

[1]Since we don't have the required distributions etc. we cannot actually find the solution.