Name: _____

Instructions: This is a take-home exam. No collaboration with other people is allowed. You may consult books or online resources, but you must cite them. You must write all statistical algorithms yourself. You may use the built-in optimization routine `nlm` to check your work, but you must be the author of all routines used for the solutions you turn in. If you have a question about whether a built-in function is acceptable, email me. Turn in a paper copy of your exam, including derivations, code, output, and images.

Academic honesty: Sign on the line below when you are finished with the exam. Turn in this coversheet with your exam.

I have neither given nor received aid in this exam.

_____ _____

Signature Date

Problem 1 Misreporting counts

When survey respondents report counts, they often round their true count up or down to the nearest 5 or 10. Suppose we administer a survey to n people. Let the true count for subject i be

$$X_i \sim \text{Poisson}(\lambda_i)$$

where $\log \lambda_i = \alpha + \beta Z_i$, with α and β scalars, and Z_i is a scalar covariate value. Unfortunately subjects do not report X_i. Instead, they report a rounded count Y_i as follows:

$$Y_i = \begin{cases} X_i - (X_i \mod 10) & \text{with probability } p \\ X_i + (10 - (X_i \mod 10)) & \text{with probability } q \\ X_i & \text{with probability } r \end{cases}$$

where $p+q+r = 1$ and "mod" is the modulus operator. Let $\theta = (\alpha, \beta, p, q, r)$ be the unknown parameter vector. The data, consisting of the pairs (Z_i, Y_i) for each i, are in the file `misreported_counts.csv`.

a. **[5 points]** Derive $\Pr(Y_i = y | X_i = x, Z_i = z)$ and $\Pr(X_i = x | Y_i = y, Z_i = z)$.

b. **[15 points]** Develop an algorithm to estimate θ by maximum likelihood. Show any derivations and describe how your algorithm works. Turn in your code. Report your estimates $\hat{\theta}$ and asymptotic standard errors.

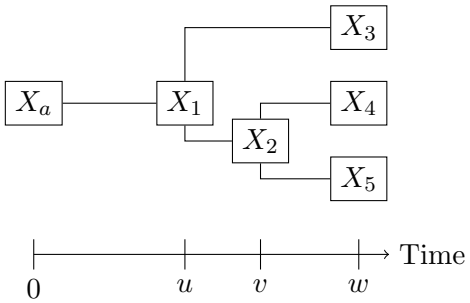
c. **[5 points]** Using $\hat{\theta}$, invent a method to find the most likely value of X_i for each $i = 1, \ldots, n$. Call this estimate of the true count $\widehat{X}_i$. Compare $\widehat{X}_i$ to Y_i in a scatterplot.

Problem 2 Brownian motion on a phylogenetic tree

Let $X(t)$ be the trait value of an organism at time t. It is common in evolutionary biology to model $X(t)$ using a stochastic process known as *Brownian motion*. One useful property of Brownian motion with variance σ^2 is that $X(t)$ has *indenpendent increments*: when $0 \leq s < t$,

$$X(t) - X(s) \sim \text{Normal}\big(0, (t-s)\sigma^2\big).$$

A phylogenetic tree describes the relationship between species over evolutionary timescales. Here is a phylogenetic tree for three species.



Each horizontal line is called a *lineage*. Time goes from left to right, $0 < u < v$ are the times of the splitting events, and w is the time at which we observe the trait values of all extant lineages. You can think of w as the present day. When a lineage splits into two, its *daughter* lineages inherit its trait value and evolve independently thereafter. For example, the ancestral lineage a has trait value X_a at time 0. This trait evolves for a time $u > 0$. At time u its value is X_1, and it splits into two identical lineages, each of which inherit the trait value X_1 and evolve independently thereafter. When you know the value of a trait at the beginning and end of a lineage, the distribution of the change in a trait value over time is easy to calculate: for example, $X_1 - X_a \sim \text{Normal}(0, u\sigma^2)$.

Suppose we observe $X_a = 0$, $X_3 = 3.807184$ $X_4 = 2.688513$ $X_5 = 2.628231$ and splitting times $u = 1.074132$ $v = 1.373923$ $w = 1.501105$.

a. **[10 points]** Write the joint likelihood of the unknowns (X_1, X_2, σ^2).

b. **[15 points]** Let σ^2 have inverse gamma prior distribution. Estimate the joint posterior distribution of (X_1, X_2, σ^2). [Hint: you do not need prior distributions for X_1 and X_2]