Stat 624: Project 2

The purpose of this project is to explore some of the wide array of computational topics used for the type of problem solving you may do the rest of the graduate program and in real applications. You can pick one of the following topics or choose one of your own with your own data set. There are two deliverables from this assignment, an in-class presentation and a write-up. The write-up should be composed in Latex and should include the 4 following sections:

1. **Introduction**: The research question should be introduced with some motivations as to why that particular question may be important. The methodology and data used is also introduced along with some other potential areas where the methodology is useful.

2. **Methodology**: The methodology is described in detail. How do you plan to answer the research questions? What other statistical techniques are involved (i.e. optimization, regression, hypothesis testing)?

3. **Simulation Study**: Prove the methodology you proposed in the previous section works the way you say it will work to answer the question.

4. **Results**:Introduce and explore the data. Apply the methodology to the data and report/display your results.

5. **Conclusion** A short rehash of the main ideas from the previous four sections.

The in-class presentation includes a set of slides composed in Beamer and should include elements of each of those 5 sections. The presentations themselves should last between 7 and 10 minutes.

If you choose one of the following projects, the data will be available on the git repositories. If you choose your own data set and problem, you will need to clear it with me. I may add more elements to your idea (or take away some) to match the difficulty level of these other projects.

1 Design

Optical media, such as DVDs and CDs, wear out over time depending on use, temperature and humidity. The rate at which they where out is known by technology specialists to follow what is called the reduced Eyring equation.

$$E(Y_i) = A \exp\left(C/T_i + BH_i\right), \tag{1}$$

where T_i is temperature and H_i is humidity.

Because it would take a long time to determine how long discs are when they fail at normal conditions, they are subjected to extreme conditions and then extrapolated using Equation (1) to normal conditions. Specifically, they are typically tested at temperature/humidity combinations

of $(65, 85), (70, 75), (85, 70), (85, 85)$ and then the model is extrapolated to normal conditions of 25° C and 50% humidity.

Each temperature/humidity combination is tested 5 times and then the fitted model is applied to the standard conditions to predict how long optical media will last.

Uses any

Data 1: Optical media data.

Data 2: Different optical media data.

Research Questions: While not extending past the ranges of temperature and humidity in the test group, is there a better set of 4 conditions to test at that will reduce the bias and MSE of the prediction? Given the 20 data points initially collected, if 5 more tests were to be run at one more condition, what are the optimal temperature and humidity for the additional observations.

2 Markov Chain: Population

Matrix population models are used to to analyze and predict population growth dynamically over time. The general formula is

$$E\begin{pmatrix} S_t \\ M_t \\ L_t \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & 0 \\ 0 & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} S_{t-1} \\ M_{t-1} \\ L_{t-1} \end{pmatrix} \tag{2}$$

which can also be written as $E(\boldsymbol{N}_t) = \boldsymbol{A}\boldsymbol{N}_{t-1}$. S_t, M_t, and L_t represent the counts from a population that are small, medium, and large respectively. The values inside A will control how a population grows. For example, a_{11}, a_{12} and a_{13} represent new individuals in a population, typically through birth. a_{21} and a_{32} represent growth.

Data 1: Black bears in Utah. Estimated sizes of Black Bears in Utah county. We can use the data to determine the probability of an oversized or undersized population of large black bears in the region to help wildlife managers determine control populations.

Data 2: Pink Salmon in Alaskan River. This data can help determine probabilities of different amounts of large Salmon which are fished commercially to help maintain the population.

Research Questions: For a given matrix $\boldsymbol{A}$ and a threshold k, what is the probability that the number of large members of the population at time t exceeds k, i.e. $(Pr(L_t > k)$. What about medium or small? What values inside of the matrix A, when changed, affects this result the most? *Note*: the data is used to find A and then the methods can be used to determine the probabilities.

3 Markov Chain: SIR

An SIR (Susceptible - Infected - Recovered) model is a way that epidemiologists model diseases and their progression over time. The idea is that the proportion in each group evolves over time according to

$$\begin{pmatrix} p_{S,t} \\ p_{I,t} \\ p_{R,t} \end{pmatrix} = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ 0 & a_{32} & 1 \end{pmatrix} \begin{pmatrix} p_{S,t-1} \\ p_{I,t-1} \\ p_{R,t-1} \end{pmatrix} \tag{3}$$

which can also be written as $\boldsymbol{p}_t = \boldsymbol{A}\boldsymbol{p}_{t-1}$. The rows in $\boldsymbol{A}$ must sum to 1 to ensure that $\boldsymbol{p}_t$ sums to 1 as probabilities should. The elements in $\boldsymbol{A}$ are transition probabilities of staying or moving

to different groups. The initial condition for these are typically a $p_{S,0}$ being close to 1, $p_{I,0}$ being close to 0 an $p_{R,0}$ being equal to 0.

Data: Influenza data at an English boarding school. The data reports the number total of 743 individuals at the school were infected with the disease. This data can tell us about how this disease spreads so we can know what to expect if introduced in a different population.

Research Questions: From being given only the total number and the number affected at each time point, how would you construct the SIR model and learn the parameters? For a given matrix $\boldsymbol{A}$ and total size of a population, how could you determine the expected number affected at any given time. What is a reasonable range for the maximum infected. How long does it take until the maximum number of infected is observed? How are these values affected by the parameters in $\boldsymbol{A}$.

4 Kalman Filter

Often times it is important to separate the signal in a time series from random noise. One way to do this is a state space model. An AR(1) model with noise is written as

$$\begin{aligned} y_t &\sim N(x_t, \sigma^2) && (4) \\ x_t &\sim N(\phi x_{t-1}, \tau^2) && (5) \end{aligned}$$

where $-1 < \phi < 1$. The data is $y_1, ..., y_T$. The values of $x_1, ..., x_T$ are parameters that can be estimated using what is called Kalman filter which is a procedure that can be found in a number of sources. The use of this is that x_t is basically a signal without as much additional random noise.

Data: El Nino SOI - An El Niño is associated with when the SOI index is below -1. We can use this method to filter out the noise and better detemine the existence and intesnity of an El Niño event.

Research Question: For a particular threshold k, determine the probability of a false positive, i.e. of the times where y_t exceeds the threshold, how often does x_t still remain below the threshold. Evaluate the effect of ϕ, σ^2, and τ^2 in that result.

5 Poisson Regression

When the response data in a regression model lies in the domain of positive integers, it can fit in the framework of poisson regression. In a poisson regression model

$$\begin{aligned} f(y_i|\lambda_i) &= \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} && (6) \\ \log(\lambda_i) &= \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_p x_{p,i} && (7) \end{aligned}$$

Data: Number of mussels in rivers as a function of a number of variables. Using these methods we can predict how many mussels we'd expect to see under different conditions.

Research Questions: In a linear model, the parameters are fairly easily interpreted, but this is not the case here. Determine some sort of way to display the effects of a one unit change in a covariate. For a certain integer k, determine the probability that a new observation will be greater than or equal to k. Determine a way to do hypothesis testing on the coefficients to determine which are insignificant. Do you determine the same variables are significant or not as if you ignore the count nature of the nature and fit a linear regression model instead?

6 Poisson Process

A poisson process is a process that defines how often an event occurs in a specified time. It is controlled by an intensity function $\lambda(t)$. The number of events that occur between time a and time b is distributed as a Poisson random variable with intensity function $\int_a^b \lambda(t)$, i.e.

$$Pr(N(a,b) = \text{Pois}\left(\int_a^b \lambda(t)\right).$$

This is often simplified by assuming the intensity function is constant, $\lambda(t) = \lambda$. Then the number of events between a and b is

$$Pr(N(a,b) = Pois\left(\lambda(b-a)\right).$$

Suppose that the intensity function is instead piecewise, where there are certain change points. For example,

$$\lambda(t) = \begin{cases} \lambda_1 & t < k_1 \\ \lambda_2 & k_1 < t < k_2 \\ \lambda_3 & t > k_2 \end{cases}$$

Data: 75 years of British accidents. Determine when the change points are to determine when the intensity of accidents changed. Perhaps this can gain some insight into what policies or practices of the different periods caused more frequent accidents.

Research Questions: Determine a way to estimate the intensity function in both the case when it is constant and when it is piecewise with 3 pieces, as above. Determine a way to estimate where the change points are.

7 Pairs Trading

The key idea underlying pairs trading is that the movement of the ratio away from its historical average represents an opportunity to make money. For example, if stock 1 is doing better than it typically does, relative to stock 2, then we should sell stock 1 and buy stock 2. This is called "opening a position." Then, when the ratio returns to its historical average, we should buy stock 1 and sell stock 2. This is called "closing the position." The reasoning is quite simple: when stock 1 is priced sufficiently higher than usual, it is likely to go down in value and the price of stock 2 is likely to go up, at least relative to the price of stock 1, since they are positively correlated. Of course, both could increase, but we are interested in relative change as we are looking at the ratio.

Let m be the historical ratio between two stocks. If the ratio moves above or below the the long term average by k standard deviations, then the strategy would be to buy one and sell the other to make a profit.

Data: Dow Jones (DJIA) and S&P 500 (SP500) stock indices.

Research Question: Determine a way to find the optimal value for k to maximize profits. Is that value for k at all related to the correlation between the two stocks? How does the correlation between the stocks affect the final profit?

8 Baum-Welch Algorithm

A hidden Markov model is a way to analyze hidden states that evolve over time. These unobserved states $x_1, ..., x_T$, drive some observable process, $y_1, ..., y_T$. Imagine if there are K different states, $X_1, ..., X_K$, where there is some probability of transitioning from one state to another, $Pr(x_t = X_i | x_{t-1} = X_j)$. Then there are N different possible observations, $Y_1, ..., Y_N$ that occur with certain probabilities given the hidden state, $Pr(y_t = Y_i | x_t = X_j)$.

The Baum-Welch algorithm is a method of determining the probabilities that define the hidden Markov model.

Data: A luxury car salesman can sell as little as 5 cars a month or as much as 30 cars a month depending on a number of factors that are not always measurable. These factors can deal with the status of the salesman, his/her approach changing over time, or social and economic conditions. The data set is 258 days of data for the number of cars a luxury car salesman sells in 1 day.

Research Question: Reducing the data set to binary outcome (did a car get sold or not in a given day), apply the Baum-Welch algorithm to determine the hidden Markov model for this data. Use 2, 3, and 4 different hidden states. Is there is a way to determine the correct number of hidden states?

9 Expectation Maximization - Missing Data

The multivariate normal distribution with dimension d has a mean function, μ, and a covariance matrix Σ. Estimating these parameters are not terribly difficult. It gets more complicated when some data is missing, though. Suppose some of the d entries in the observations are missing. Three possible methods are

- Throw out all the records with any missing observations in it
- Use the Expectation-Maximization algorithm and iteratively determine the estimate by calculating the conditional expectations and variances of the missing data
- Instead of using the expected values as above, draw samples for the missing values based on conditional normal theory and then estimate the mean. This method will not converge like the E-M algorithm will, but you can collect samples and evaluate the set of samples afterwards. This is a somewhat Bayesian approach (if you included priors it would be fully Bayesian).

Uses multivariate normal data with some missing value.

Data 1: Characteristics of individuals with Hepatitis. Many of the variables are factor variables. Use the continuous or integer variables, which include age, bibirubin, alkphosphate, sgot, albumin, and protime.

Data 2: Automobile sales. Many of the variables are factor variables. Leave out the factor variables such as make, fuel type, aspiration, doors, style, wheels, engine location and type, cylinders and fuel system.

Research Questions: The above algorithms may work differently when there is a reason why data is missing. For example, perhaps the ones that are missing are missing particularly because they were difficult to measure, too high, or too extreme. Determine a method of guessing if the data is missing at random or if there is a pattern to the missingness. Find estimates of the means and covariances between the variables. Determine which variables are most and least correlated.