COLUMBIA UNIVERSITY — Prof. Marco Avella
Statistical Inference and Modeling — TA: Casey Bradshaw, Leon Fernandes
Graded Homework 2 — Spring 2020

You are free to present your solutions to the exercises below in *groups of at most three*. This homework is due on March 9. Buena suerte!

Exercise 1

The data set **transplant.txt** consists of data regarding allotransplant and autotransplant. The first column is the time to death or relapse (in months), the second one shows the types of transplant (1=allogeneic, 2=autologous), and the last column is the survival indicator (0=alive without relapse, 1=dead or relapse). Provide a minimum amount of `R` output to justify your answers.

1. Do you think that it is reasonable to assume that the right censoring in this data set is random? why?
2. Plot the Kaplan-Meier estimators of the survival curves of the two transplant groups. Does the plot seem to indicate a difference between these groups? which type of transplant seems to be more efficient?
3. Fit an exponential model to using the function `survreg`. Do the signs of the fitted parameters agree with your intuition from last point?
4. Using point 3, do we observe a significant difference between the two groups based on the likelihood ratio statistic? Does this conclusion depend on parametric model assumptions?
5. Do a visual model check using the fitted exponential models and the Kaplan-Meier fits. Comment briefly what you observe.
6. Try fitting a Weibull model now. Does this model fit the data better? Does this new model provide evidence against the relevance on an exponential model?

Exercise 2

The purpose of this exercise is to explore the numerical performance of the following approaches to missing data:

(a) Complete case analysis.

(b) Available case analysis.

(c) Mean imputation

(d) Mean imputation with the bootstrap.

(e) The EM-algorithm

We will start by looking at the Student Score Data[1] which has both a small sample size and missing observations. Let λ_1 denote the largest eigenvalue of the population covariance matrix Σ and $\hat{\lambda}_1$ its estimated value using the sample covariance[2]. In the absence of missing data, it can be shown that

$$\sqrt{n}(\log \hat{\lambda}_1 - \log \lambda_1) \xrightarrow[n\to\infty]{\mathcal{D}} N(0, 2).$$

Solutions to 1–3 below should be presented in *no more than two pages*. Provide a minimum amount of `R` output to justify your answers.

1. Estimate the covariance matrix of the 5 variables in the Student Score Data using methods (a)–(e) and comment your results.

2. How would you construct an asymptotic confidence interval leveraging the asymptotic normality of $\hat{\lambda}_1$? Use this result to construct confidence intervals for λ_1 using the estimated covariances from point 1. Comment your findings.

3. Note that the student score data is just a subset of the mathmarks data[3] with artificially generated missing entries. Use the full data to compute the sample covariance and give a confidence interval for λ_1. Compare this with your findings in the previous two questions.

4. Can you derive the form of the EM-algorithm for an i.i.d. normal sample with missing data?

   Remember that with partially observed vectors $X_i = (X_{io}^T, X_{im}^T)^T$ the EM-algorithm simplifies to the iterations:

$$\mu^{(k+1)} : \sum_{i=1}^{n} (\hat{X}_i - \mu) = 0$$
$$\Sigma^{(k+1)} : \sum_{i=1}^{n} \left(\Sigma - (\hat{X}_i - \mu)(\hat{X}_i - \mu)^T - \mathbf{C}_i^{(k)} \right) = 0$$

   with $\hat{X}_{io} = X_{io}$ and

$$\hat{X}_{im} = \mu_{im}^{(k)} + \Sigma_{imo}^{(k)} (\Sigma_{ioo}^{(k)})^{-1} (X_{io} - \mu_{io}^{(k)})$$

[1] Table 1 in Efron (1994), *Journal of the American Statistical Association.*

[2] These quantities are interesting for multivariate analysis and unsupervised learning. For example, in a principal component analysis, the largest empirical eigenvalue is the variance explained by the first principal component

[3] Available in the `R` package "SMPracticals".

and $C_{ijk}^{(k)} = 0$ if X_{ij} or X_{ik} are observed and

$$\mathbf{C}_{imm}^{(k)} = \Sigma_{imm}^{(k)} - \Sigma_{imo}^{(k)}(\Sigma_{ioo}^{(k)})^{-1}\Sigma_{iom}^{(k)}$$

Exercise 3

The data set **CentralPark.csv** consists of precipitation data from weather station at Central Park, New York. The data was collected from the National Oceanic and Atmospheric Administration (NOAA). The variable PRCP shows the observed amount of rain at time t in mm. Consider a first order Markov Chain model with a two dimensional state space corresponding to the states $\{0, 1\} = \{$"rainy day","no rain"$\}$, where we define a rain day as one with a PRCP of at least 1.5 mm. Suppose the estimated transition probability matrices

$$\mathbf{T} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$$

obtained using data collected above.

1. Interpret the meaning of a_i
2. What's the long-term probability of observing a rainy day in Central Park. (Use a_i to express the result)
3. Can you estimate a_i for the month of July using the historical Central Park data?
4. Are the probability laws of "$X_{t+1}|X_t = 1$" and "$1-X_{t+1}|X_t = 0$" significantly different in Central Park?
5. Does a higher order chain improve the fit of the data?

Exercise 4

Consider the following model for a DNA sequencing experiment. Assume that the probability of obtaining one of the four bases A,C,G,T are $p_A = 1-\theta$, $p_C = \theta-\theta^2$, $p_G = \theta^2 - \theta^3$ and $p_T = \theta^3$, where $0 \leq \theta \leq 1$. Further assume that we have n independent realizations where n_A, n_C, n_G, n_T are the observed occurrences for each base and $n_A + n_C + n_G + n_T = n$.

1. Give the joint distribution of (N_A, N_C, N_G, N_T)
2. Show that the MLE of θ is

$$\hat{\theta} = \frac{N_C + 2N_G + 3N_T}{N_A + 2N_C + 3N_G + 3N_T}$$

3. Find the asymptotic distribution of $\hat{\theta}$

4. Find constants a_A, a_C, a_G, a_T such that $T = a_A N_A + a_C N_C + a_G N_G + a_T N_T$ is unbiased for θ.

5. Find the variance of T and five the asymptotic relative efficiency between T and $\hat{\theta}$

6. Compare the two estimators discussed above with the MLE that does not assume that p_A, p_C p_G and p_T depend on a common unknown parameter θ.

7. Using the last point, propose a test statistics for the null hypothesis H_0 : $\mathbf{p} = \mathbf{p}(\theta)$, where $\mathbf{p} = (p_A, p_C, p_G, p_T)^T$ and $\mathbf{p}(\theta) = (1-\theta, \theta-\theta^2, \theta^2-\theta^3, \theta^3)^T$ for some fixed value $\theta \in (0,1)$.

Exercise 5 (Optional bonus question)

Consider once again the data set **transplant.txt** from exercise 1. In this exercise you will use the **log rank test** to check the null hypothesis that the survival function of both groups is the same. Treatment A denotes the allogenic transplant and Treatment B denote the autologous transplant.
Denote by $\tau^{(1)} < \tau^{(2)} < \cdots < \tau^{(K)}$ the K distinct times of observed deaths/relapses. For $1 \le k \le K$, let $n_A^{(k)}$ and $n_B^{(k)}$ be the number of patients at risk undergoing treatment A and treatment B, respectively, at time $\tau^{(k)}$. Also, $n_d^{(k)}$ be the total number of deaths/relapses observed at time $\tau^{(k)}$ across both treatment groups A and B.
Thus, at time $\tau^{(k)}$ there were $n_A^{(k)} + n_B^{(k)}$ patients, and if $y^{(k)}$ is the number of deaths/relapses under treatment A, then $n_d^{(k)} - y^{(k)}$ is the number of deaths/relapses under treatment B.

1. Show that the conditional distribution of $y^{(k)}$ given $(n_A^{(k)}, n_B^{(k)}, n_d^{(k)})$ is a $HyperGeometric(n_A^{(k)} + n_B^{(k)}, n_A^{(k)}, n_d^{(k)})$ random variable under H_0 via the following steps:

   Hint: To calculate $P(y^{(k)} = m, n_A^{(k)} = m_A, n_B^{(k)} = m_B, n_d^{(k)} = m_d)$ *note that at time* $\tau^{(k)}$*, out of the* m_A *at risk from group A,* m *will die/relapse and out of the* m_B *at risk from group B,* $m_d - m$ *will die/relapse. Suppose* $i_1, ..., i_{m_A}$ *are the individuals at risk from group A and* $j_1, ..., j_{m_B}$ *are the individuals at risk from group B. Show that the probability of any* m *out of* $\{i_1, ..., i_{m_A}\}$ *and any* $m_d - m$ *out of* $\{j_1, ..., j_{m_B}\}$ *dying/relapsing does not depend on the value of* m*.*

2. Verify that the conditional mean and variance of $y^{(k)}$ given $(n_A^{(k)}, n_B^{(k)}, n_d^{(k)})$ under H_0 are given by

$$E^{(k)} = \frac{n_A^{(k)} n_d^{(k)}}{n^{(k)}}$$

$$V^{(k)} = \frac{n_A^{(k)} n_B^{(k)} n_d^{(k)} n_s^{(k)}}{(n^{(k)})^2 (n^{(k)} - 1)}$$

where $n^{(k)} = n_A^{(k)} + n_A^{(k)}$ and $n_s^{(k)} = n^{(k)} - n_d^{(k)}$.

3. For each $1 \leq k \leq K$ prove that under H_0, $\text{var}[y^{(k)} - E^{(k)}] = \mathbb{E}[V^{(k)}]$.
Hint: Recall the conditional variance formula for random variables X, Y

$$\text{Var}[X] = \text{Var}[\mathbb{E}[X|Y]] + \mathbb{E}[\text{Var}[X|Y]]$$

4. Show under H_0 that

$$\text{Var}\left[\sum_{k=1}^{K} (y^{(k)} - E^{(k)})\right] = \sum_{k=1}^{K} \text{Var}\left[y^{(k)} - E^{(k)}\right] = \sum_{k=1}^{K} \mathbb{E}[V^{(k)}]$$

5. To test H_0 we use the log rank test statistic:

$$Z = \frac{\sum_{k=1}^{K} \left(y^{(k)} - E^{(k)}\right)}{\left(\sum_{k=1}^{K} V^{(k)}\right)^{1/2}}$$

By using central limit theorem arguments it can be shown that under H_0, Z is asymptotically a standard normal distribution.

Using this fact test the null hypothesis that the two survival functions obtained in Exercise 1.2 above are identical.