STA 711 Homework 5

Due: Friday, March 8, 11:00am on Canvas.

Instructions: Submit your work as a single PDF. For this assignment, you may include written work by scanning it and incorporating it into the PDF. Include all R code needed to reproduce your results in your submission.

Convergence of random variables

In this section, you will practice proving limits for sequences of random variables. As a reminder, here are some of the common techniques for proving convergence:

- For convergence in probability: if you have a sequence of means, try to apply the WLLN
- For convergence in probability: if you can easily calculate a mean or variance, try bounding probabilities with Markov's or Chebyshev's inequality
- For convergence in probability: if calculating means or variances is hard, try calculating the probabilities directly for the convergence
- For convergence in distribution to a normal or χ^2: check if the central limit theorem applies
- For convergence in distribution: if CLT is not the right strategy, try calculating the cdfs directly

1. For each of the following sequences $\{Y_n\}$, show that $Y_n \xrightarrow{p} 1$. Then write a simulation in R demonstrating the convergence.

   (a) $Y_n = 1 + nX_n$, where $X_n \sim Bernoulli(1/n)$

   (b) $Y_n = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^2$, where $X_i \overset{iid}{\sim} N(0,1)$

2. Suppose that $Y_1, Y_2, ... \overset{iid}{\sim} Beta(1, \beta)$. Find a value of ν such that $n^{\nu}(1 - Y_{(n)})$ converges in distribution. Then write a simulation in R demonstrating the convergence. (*Hint:* if you are struggling to find ν, starting with simulations may be helpful)

3. In this problem, we will prove part of the continuous mapping theorem. Let $\{Y_n\}$ be a sequence of real-valued random variables such that $Y_n \xrightarrow{p} Y$ for some random variable Y. Let g be a continuous function; recall that g is continuous if for all $\varepsilon > 0$, there exists some $\delta > 0$ such that $|g(x) - g(y)| < \varepsilon$ whenever $|x - y| < \delta$. Prove that $g(Y_n) \xrightarrow{p} g(Y)$.

4. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

   where the X_i are known constants, and the ε_i are iid with $\mathbb{E}[\varepsilon_i] = 0$ and $Var(\varepsilon_i) = \sigma^2$. It can be shown that the least squares estimate of β_1 is

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum\limits_{i=1}^{n}(X_i - \overline{X}_n)\varepsilon_i}{\sum\limits_{i=1}^{n}(X_i - \overline{X}_n)^2}.$$

Show that if $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \to \infty$ as $n \to \infty$, then $\widehat{\beta}_1 \xrightarrow{p} \beta$. (Note: no distribution for ε_i or Y_i has been assumed, so $\widehat{\beta}_1$ cannot be treated as a maximum likelihood estimator).

Multivariate normal distributions

The multivariate normal distribution will appear frequently in 711, for example as the asymptotic distribution of our coefficient estimates $\widehat{\beta}$. The purpose of this section is to derive a basic property of the multivariate normal distribution that we use regularly, for example in constructing our Wald test statistic.

One way to define a multivariate normal distribution is with its *moment generating function* (MGF). Let $X \in \mathbb{R}^k$ be a random vector. The (multivariate) moment generating function $M_X(t)$ of X is defined by

$$M_X(t) = \mathbb{E}[e^{t^T X}],$$

where $t \in \mathbb{R}^k$. As with univariate MGFs, if $M_X(t) = M_Y(t)$ for all t, then the two random variables X and Y have the same distribution.

We say that the random vector $X \in \mathbb{R}^k$ follows a multivariate normal distribution with mean $\mu \in \mathbb{R}^k$ and variance matrix $\Sigma \in \mathbb{R}^{k \times k}$, and write $X \sim N(\mu, \Sigma)$, if

$$M_X(t) = e^{t^T \mu} e^{\frac{1}{2} t^T \Sigma t}.$$

We will also make use of the following properties:

- For any random vector X, the covariance matrix $\Sigma = Var(X)$ is positive semi-definite (you may use this without proof)
- If Σ is a positive semi-definite matrix, then there exists a unique positive semi-definite matrix $\Sigma^{\frac{1}{2}}$ such that $\Sigma = \Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}$ (you may use this without proof)
- $Z \sim N(0, \mathbf{I})$ if and only if $Z = (Z_1, ..., Z_q)^T \overset{iid}{\sim} N(0,1)$ (you may use this without proof).

6. An important property of multivariate normal random variables is that if $X \sim N(\mu, \Sigma)$, then

$$\boldsymbol{a} + \boldsymbol{B}X \sim N(\boldsymbol{a} + \boldsymbol{B}\mu, \boldsymbol{B}\Sigma\boldsymbol{B}^T),$$

where $\boldsymbol{a} \in \mathbb{R}^m$ and $\boldsymbol{B} \in \mathbb{R}^{m \times k}$. Our goal is to use MGFs to prove this property.

   (a) Show that for any random vector X in $\mathbb{R}^k$, the MGF of $Y = \boldsymbol{a} + \boldsymbol{B}X$ is given by

   $$M_Y(t) = e^{t^T \boldsymbol{a}} M_X(\boldsymbol{B}^T t).$$

   (b) Using (a), show that if $X \sim N(\mu, \Sigma)$, then $\boldsymbol{a} + \boldsymbol{B}X \sim N(\boldsymbol{a} + \boldsymbol{B}\mu, \boldsymbol{B}\Sigma\boldsymbol{B}^T)$.

7. Now let us prove some further properties of multivariate normal distributions:

   (a) Show that if $X \sim N(\mu, \Sigma)$, then $\Sigma^{-\frac{1}{2}}(X - \mu) \sim N(0, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix.

   (b) Show that $X \sim N(\mu, \Sigma)$ if and only if $X = \mu + \Sigma^{\frac{1}{2}} Z$ where $Z \sim N(0, \mathbf{I})$.

(c) Let $X \sim N(\mu, \Sigma)$, where $X \in \mathbb{R}^q$. Suppose that for some $1 \leq p < q$, Σ can be partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0_{p\times(q-p)} \\ 0_{(q-p)\times p} & \Sigma_{22} \end{pmatrix},$$

where Σ_{11} is $p \times p$, Σ_{22} is $(q-p)\times(q-p)$, and $0_{m\times n}$ denotes the matrix of zeros of the specified dimensions. Similarly partition

$$X = \begin{pmatrix} X_{(1)} \\ X_{(2)} \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} \end{pmatrix},$$

into vectors of length p and $q-p$. Prove that

$$X_{(1)} \sim N(\mu_{(1)}, \Sigma_{11}), \qquad X_{(2)} \sim N(\mu_{(2)}, \Sigma_{22}),$$

and $X_{(1)}$ and $X_{(2)}$ are independent.

(d) Using (c), conclude that if $X = (X_1, ..., X_q)^T \sim N(\mu, \Sigma)$, then the entries X_i and X_j are independent *if and only* if $\Sigma_{ij} = Cov(X_i, X_j) = 0$.

Normal distributions and the Wald test

Suppose that $\widehat{\theta}$ is some estimator of a parameter of interest $\theta \in \mathbb{R}^d$. We want to test the hypotheses

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0.$$

If $\widehat{\theta}$ is approximately normal, then we can use the Wald test (often $\widehat{\theta}$ will be the MLE, but the Wald test can be applied to any asymptotically normal estimator, not just to the MLE). Formally, suppose that

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, V),$$

and let $\widehat{V}$ be some estimator of the covariance matrix V, such that $\widehat{V} \xrightarrow{p} V$. Then the Wald test statistic is

$$W = n(\widehat{\theta} - \theta_0)^T \widehat{V}^{-1} (\widehat{\theta} - \theta_0).$$

The goal of this section is to verify that $W \xrightarrow{d} \chi^2_d$ if H_0 is true.

8. Now let's derive the relationship between the normal distribution and the χ^2 distribution.

(a) Let $Z \sim N(0,1)$ be a standard normal variable. Show that $Z^2 \sim \chi^2_1$ (a χ^2 distribution with 1 degree of freedom), by proving that the pdf of $Y = Z^2$ is

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2}.$$

(b) Suppose that $Z_1, Z_2, ..., Z_q \overset{iid}{\sim} N(0,1)$. Show that $\sum_{i=1}^{q} Z_i^2 \sim \chi^2_q$ (a χ^2 distribution with q degrees of freedom). You may use the mgf of a χ^2 distribution without proof.

(c) Let $\theta \in \mathbb{R}$ be a parameter of interest, and $\widehat{\theta}_n$ the maximum likelihood from a sample of size n. Let

$$Z_n = \sqrt{n\mathcal{I}_1(\theta)}(\widehat{\theta}_n - \theta).$$

Asymptotic normality of the MLE tells us that $Z_n \xrightarrow{d} N(0,1)$. Show that $Z_n^2 \xrightarrow{d} \chi^2_1$.

9. Finally, let's connect the multivariate normal with the χ^2.

   (a) Show that if $X \sim N(\mu, \Sigma)$, then $(X-\mu)^T \Sigma^{-1} (X - \mu) \sim \chi^2_q$.

   (b) Suppose that $\widehat{\theta}$ is some estimator of $\theta \in \mathbb{R}^d$, and $\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, V)$. Let $\widehat{V}$ be an estimator of V such that $\widehat{V} \xrightarrow{p} V$, and let $W = n(\widehat{\theta} - \theta_0)^T \widehat{V}^{-1} (\widehat{\theta} - \theta_0)$. Prove that $W \xrightarrow{d} \chi^2_d$ if the null hypothesis $H_0 : \theta = \theta_0$ is true.

Logistic regression with earthquake data

In the second part of this exam, you will work with a dataset from DrivenData, an online data competition site that hosts competitions aimed at improving education, health, safety, and general well being for individuals around the world.

Our data come from the 2015 Gorkha earthquake in Nepal. After the earthquake, a large scale survey was conducted to determine the amount of damage the earthquake caused for homes, businesses and other structures. This is one of the largest post-disaster surveys in the world, and researchers are interested in which building characteristics are associated with earthquake damage.

You will work with a subset of the earthquake data, consisting of 211774 buildings, containing the following variables:

- `Damage`: whether the building sustained any damage (1) or not (0)
- `Age`: the age of the building (in years)
- `Surface`: a categorical variable recording the surface condition of the land around the building. There are three different levels: `n`, `o`, and `t`. (The researchers who collected the data anonymized the level names to protect inhabitants' privacy).

You can load the data into R by

```
earthquake <- read.csv("https://sta711-s24.github.io/homework/earthquake_small.csv")
```

You will work with the following logistic regression model (you may assume all assumptions are met; no transformations or diagnostics are needed):

$$Damage_i \sim Bernoulli(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Age_i + \beta_2 SurfaceO_i + \beta_3 SurfaceT_i + \beta_4 Age_i \cdot SurfaceO_i + \beta_5 Age_i \cdot SurfaceT_i$$

where $SurfaceO$ and $SurfaceT$ are indicator variables for whether surface is o or t, respectively.

10. (a) Fit the logistic regression model in R, and interpret the estimated slope $\widehat{\beta}_1$ in terms of the *odds* of damage.

    (b) Calculate the estimated probability of damage for a 50 year old building with surface condition = t.

11. The researchers want to know whether the relationship between Age and the probability of damage is the same for buildings in all three surface conditions. Use a hypothesis test to address the researchers' question; you should state the hypotheses in terms of one or more model parameters, calculate a test statistic and p-value, and make a conclusion.