STA 711 Homework 6

Due: Friday, March 22, 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.

1. Suppose that $Y_1, ..., Y_n$ are identically distributed with $\mathbb{E}[Y_i] = \mu$, $Var(Y_i) = \sigma^2$, and covariances

$$Cov(Y_i, Y_{i+j}) = \begin{cases} \rho\sigma^2 & |j| \leq 2 \\ 0 & |j| > 2 \end{cases},$$

   where $\rho \in [-1, 1]$ and $\rho \neq 0$. Show that $\overline{Y}_n \xrightarrow{p} \mu$ as $n \to \infty$. (Note: you may not directly use the version of the WLLN stated in class, because it assumes iid data).

2. Let $X_1, ..., X_n$ be an iid sample from a population with mean μ and variance σ^2, and suppose that σ^2 is known. We wish to test the hypotheses $H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \mu_0$. Suppose that $\alpha = 0.05$, $\mu_0 = 0$ and $\sigma^2 = 1$. What is the minimum sample size n needed such that $\beta(0.5) > 0.7$?

3. (Global F-test for linear regression) Suppose that $V_1 \sim \chi^2_{d_1}$ and $V_2 \sim \chi^2_{d_2}$ are independent χ^2 random variables. Then $F = \dfrac{V_1/d_1}{V_2/d_2} \sim F_{d_1,d_2}$, where F_{d_1,d_2} denotes the F-distribution with numerator degrees of freedom d_1 and denominator degrees of freedom d_2.

   The F-distribution is important for hypothesis testing in linear regression models. Suppose we observe independent data $(X_1, Y_1), ..., (X_n, Y_n)$, where $Y_i = \beta^T X_i + \varepsilon_i$, with $\beta = (\beta_0, ..., \beta_k)^T$ and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We wish to test the hypotheses

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \qquad H_A : \text{at least one of } \beta_1, ..., \beta_k \neq 0.$$

   The F-test for these hypotheses is based on the F-statistic

$$F = \frac{(SSTO - SSE)/k}{SSE/(n-k-1)},$$

   where $F \sim F_{k,n-k-1}$ under H_0, and

$$SSTO = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \qquad SSE = \sum_{i=1}^{n} (Y_i - \widehat{\beta}^T X_i)^2$$

   The goal of this problem is to demonstrate that, indeed, $F \sim F_{k,n-k-1}$ under H_0.

   (a) Show that under H_0, $\frac{1}{\sigma^2} \sum\limits_{i=1}^{n} (Y_i - \beta_0)^2 \sim \chi^2_n$.

   (b) Find symmetric matrices A_1, A_2, A_3 such that under H_0,

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0)^2 = Z^T A_1 Z + Z^T A_2 Z + Z^T A_3 Z$$

   where $Z \sim N(0, I)$, $\frac{1}{\sigma^2} SSE = Z^T A_1 Z$, and $\frac{1}{\sigma^2}(SSTO - SSE) = Z^T A_2 Z$.

(c) Using the matrices A_1, A_2, A_3 from part (b), show that $rank(A_1) = n-k-1$, $rank(A_2) = k$, and $rank(A_3) = 1$.

(d) By applying Cochran's theorem, show that $F = \dfrac{(SSTO - SSE)/k}{SSE/(n-k-1)} \sim F_{k,n-k-1}$ under H_0.

Nonparametric estimation

So far, we have focused on estimated parameters in parametric distributions. But what if we want to estimate a distribution without assuming any parametric family? Let $X_1, ..., X_n$ be iid from some distribution with cdf F. The *empirical distribution function* F_n is a *nonparametric* estimate of F defined by

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\}.$$

The goal of this section is to show that F_n is a reasonable estimate of F.

4. Let $\{Y_n^{(1)}\}, \{Y_n^{(2)}\}, ..., \{Y_n^{(k)}\}$ be k sequences of random variables, such that each $Y_n^{(i)} \xrightarrow{p} Y^{(i)}$ (here the superscript (i) is used to distinguish the ith sequence; it does not denote an exponent, derivative, or order statistic). Show that

$$\max_{i=1,...,k} |Y_n^{(i)} - Y^{(i)}| \xrightarrow{p} 0.$$

5. Show that for each t, $F_n(t) \xrightarrow{p} F(t)$. (In other words, the empirical distribution function converges pointwise to the true cdf).

6. Let $t \in \mathbb{R}$ be given. Suppose for this specific t, we want to test the hypotheses

$$H_0 : F(t) = p_0 \qquad H_A : F(t) \neq p_0.$$

Derive a Wald test using the empirical distribution function F_n; you should state the test statistic, demonstrate that it has the desired asymptotic distribution, and specify when the test will reject the null hypothesis.

7. Let $X_1, ..., X_n$ be iid continuous random variables with cdf F. Show that

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{p} 0$$

(in other words, F_n converges *uniformly* to F in probability; a slightly weaker version of the Glivenko-Cantelli theorem). *Hint:* Begin by choosing $a_0, a_1, ..., a_k$ such that

$$-\infty = a_0 < a_1 < \cdots < a_k = \infty$$

and $F(a_i) - F(a_{i-1}) = \frac{1}{k}$ for $i = 1, ..., k$ (your proof should explain why you can do this).