BIOSTAT/STAT 570: Coursework 4

To be submitted to the course canvas site by 11:59pm Friday 28th October, 2022.

1. This question is intended to illustrate how OLS estimation is affected by misspecification of the variance-covariance of the error terms.

   Consider the simple linear regression model

$$\begin{aligned} Y_i &= \mu_i + \epsilon_i \\ &= \beta_0 + \beta_1(t_i - \bar{t}) + \epsilon_i, \end{aligned} \tag{1}$$

   where t_i represents time and the error terms ϵ_i are normal and are such that $\mathrm{E}[\epsilon_i] = 0$, $i = 1, \ldots, n$. In the following assume that $n = 5, 10, 20, 30, 40, 50$, t_i equally spaced in $(-2, 2)$, $\beta_0 = 4$, $\beta_1 = 1.75$ and $\sigma^2 = 1$.

   Consider the following three forms for the variance-covariance:

   I. var$(\epsilon_i) = \mu_i \sigma^2$, and cov$(\epsilon_j, \epsilon_k) = 0$ for $j \neq k$.

   II. var$(\epsilon_i) = \mu_i^2 \sigma^2$, and cov$(\epsilon_j, \epsilon_k) = 0$ for $j \neq k$.

   III. var$(\epsilon_i) = \sigma^2$, and cov$(\epsilon_j, \epsilon_k) = \sigma^2 \rho^{|t_j - t_k|}$ with $-1 < \rho < 1$.

   (a) Simulate data from the above models and estimate β_0 and β_1 using OLS (use the values $\rho = 0.1, 0.5, 0.9$). Examine the 95% confidence interval coverage for β_0 and β_1.

   (b) Summarize your conclusions.

   [Hint: For model III, note that the marginal distribution of $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\mathsf{T}$ is an n-dimensional zero mean normal with the covariance matrix taking the form

$$\text{var}(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & \delta\sigma^2 & \delta^2\sigma^2 & \cdots & \cdots & \delta^{n-1}\sigma^2 \\ \delta\sigma^2 & \sigma^2 & \delta\sigma^2 & \cdots & \cdots & \delta^{n-2}\sigma^2 \\ \vdots & \vdots & \vdots & \cdots & \cdots & \vdots \\ \delta^{n-1}\sigma^2 & \cdots & \cdots & \cdots & \cdots & \sigma^2 \end{bmatrix},$$

   the parameter δ takes some suitable value corresponding to the values of ρ and n, which will help with data simulation.]

2. In this question we will consider inference when the sampling model is multivariate hypergeometric. Suppose a population contains objects of K different types, with $X_1, \ldots, X_K$ being the number of each type, $\sum_{k=1}^{K} X_k = N$. A simple random sample of size n is taken and the number of each type, $Y_1, \ldots, Y_K$, is recorded (so that $\sum_{k=1}^{K} y_k = n$).

   An obvious model for $Y_1, \ldots, Y_k$, is the multivariate hypergeometric distribution:

$$\Pr(Y_1 = y_1, \ldots, Y_K = y_K | x_1, \ldots, x_K) = \frac{\prod_{k=1}^{K} \binom{x_k}{y_k}}{\binom{N}{n}},$$

   with means and variances:

$$\begin{aligned} \mathsf{E}[Y_k | x_k] &= n \frac{x_k}{N} && (2) \\ \mathsf{var}(Y_k | x_k) &= n \frac{x_k}{N} \left(1 - \frac{x_k}{N}\right) \frac{N-n}{N-1} && (3) \end{aligned}$$

   Suppose we take a sample from a population of K distinct objects, and record $y_1, \ldots, y_K$, but the numbers $X_1, \ldots, X_K$ are unknown (but N is known).

   (a) Using (2), write down an estimator for X_k, $k = 1 \ldots, K$. We will refer to this as a method of moments estimator. Using (3) give a from for the variance of this estimator, along with an estimator of this variance.

   (b) We now consider a Bayesian approach to inference. Consider a multinomial distribution for counts $X_1, \ldots, X_K$,

$$\Pr(X_1 = x_1, \ldots, X_K = x_K | p_1, \ldots, p_K) = \frac{N!}{\prod_{k=1}^{K} x_k!} \prod_{k=1}^{K} p_k^{x_k}, \quad (4)$$

   with $p_k > 0$ and $\sum_{k=1}^{K} p_k = 1$. Show that the Dirichlet:

$$\pi(p_1, \ldots, p_K) = \frac{\Gamma(\alpha_+)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \quad (5)$$

   where $\alpha_k > 0$, $k = 1, \ldots, K$, and $\alpha_+ = \sum_{k=1}^{K} \alpha_k$, is the conjugate distribution to the multinomial sampling model.

   [One interpretation of this set up is that $p_1, \ldots, p_K$ are the proportions in each category in a hypothetical infinite population of objects.]

   (c) The compound multinomial distribution, CMult$(N, \boldsymbol{\alpha})$, is defined as

$$\Pr(X_1 = x_1, \ldots, X_K = x_K) = \frac{N! \Gamma(\alpha_+)}{\Gamma(N + \alpha_+)} \prod_{k=1}^{K} \frac{\Gamma(x_k + \alpha_k)}{x_k! \Gamma(\alpha_k)},$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$. Show that the prior predictive distribution, obtained as the marginal distribution when the likelihood is (4) and the prior is (5), is of compound multinomial form with parameters that you should identify.

(d) Find the mean E$[X_k]$ and variance var(X_k), $k = 1, \dots, K$, of a compound multinomial distribution.

[Hint: you may quote without proof the means and variances of the multinomial and Dirichlet distributions.]

(e) Let $W_k = X_k - y_k$ represent the unobserved counts, $k = 1, \dots, K$. Show that the posterior distribution $\Pr(W_1, \dots, W_K | y_1, \dots, y_K)$ is compound multinomial CMult$(N - n, \boldsymbol{\alpha} + \boldsymbol{y})$, where $\boldsymbol{y} = (y_1, \dots, y_K)$.

(f) Write down the posterior mean and posterior variance of X_k, $k = 1, \dots, K$. Comment on the case when $\alpha_k = 0$, $k = 1, \dots, K$.

A certain infectious disease can be caused by one of three different pathogens, A, B, or C. Over a 1 year period population surveillance is carried out, and 750 individuals are observed to be infected. A random sample of 65 cases is selected for lab testing, i.e., to determine the pathogen responsible. Of these 65 selected cases, the numbers who were infected by pathogens A, B, C, were 44, 21, 0, respectively.

We wish to estimate the numbers of the total population of cases that were infected by each of the pathogens.

(g) Calculate the method of moments estimators of X_k, $k = 1, \dots, K$, and the associated standard errors.

(h) Calculate the Bayesian posterior mean and posterior standard deviation of X_k, with prior specification, $\alpha_k = 1$, $k = 1, \dots, K$. Which estimates are the most reasonable?

(i) Devise a sampling-based method for sampling from the posterior, and present histogram representations of the posterior distributions of $X_k | \boldsymbol{y}$, $k = 1, \dots, K$.