BIOSTAT/STAT 570: Coursework 2

To be submitted to the course canvas site by 11:59pm Monday 18th October, 2021.

1. In this question we will begin to investigate the robustness of the OLS estimator to non-normality of the errors.

   Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, ..., n, \tag{1}$$

   where the error terms ϵ_i are such that $\mathrm{E}[\epsilon_i] = 0$, $\mathrm{var}(\epsilon_i) = \sigma^2$ and $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0, \; i \neq j$. In the following you will consider x_i with: $x_i \sim_{iid} N(20, 3^2)$, with $\beta_0 = 2$ and $\beta_1 = -2.5$ and $n = 15, 30$.

   Simulate from model (1) with the error terms ϵ_i i.i.d. from the distributions:

   - The normal distribution with mean 0 and variance 2^2.
   - The uniform distribution on the range $(-r, r)$ for $r = 2$.
   - A skew normal distribution[1] with $\alpha = 5$, $\omega = 1$ and ξ chosen to give a distribution with mean zero.

   (a) Confirm numerically that the bias is zero.

   (b) Compare the variance of the estimator as reported by least squares, with that which follows from the sampling distribution of the estimator (which you can estimate from the simulations).

   (c) Examine the distribution of the resultant estimators (across simulations) of β_0 and β_1, in particular with respect to normality. For each parameter find the coverage probability of a 95% confidence interval, that is the proportion of times that the confidence intervals contain the true value.

   (d) **Bonus:** Can you "break" least squares? i.e., find a distribution of the errors (with mean zero) that produces poor confidence interval coverage?

[1] If $\phi(x)$ and $\Phi(x)$ are the density function and distribution function of a standard normal then the skew normal distribution with location ξ and scale ω is

$$f(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{x-\xi}{\omega}\right)\right).$$

The mean of the distribution is $\mathrm{E}[X] = \xi + \omega\delta\sqrt{\frac{2}{\pi}}$, where $\delta = \frac{\alpha}{\sqrt{1+\alpha^2}}$.

[Note: You should simulate the set of x values once, and then use in all subsequent simulations.]

[Hint: to simulate from a skew normal distribution you may use the function `rsn` in the `sn` package.]

2. Consider the exponential regression problem with independent responses

$$p(y \mid \lambda_i) = \lambda_i e^{-\lambda_i y}, \quad y > 0$$

and $\log \lambda_i = \beta_0 + \beta_1 x_i$ for given covariates x_i, $i = 1, \ldots, n$. We wish to estimate the 2×1 regression parameter $\boldsymbol{\beta} = [\beta_0, \beta_1]^\intercal$ using MLE.

(a) Find expressions for the likelihood function $L(\boldsymbol{\beta})$, log likelihood function $l(\boldsymbol{\beta})$, score function $\boldsymbol{S}(\boldsymbol{\beta})$ and Fisher's information matrix $\boldsymbol{I}(\boldsymbol{\beta})$.

(b) Find expressions for the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$. If no closed form solution exists, then instead provide a functional form that could be simply implemented for solution.

(c) For the data in Table 1, numerically maximize the likelihood function to obtain estimates of $\boldsymbol{\beta}$. These data consist of the survival times (y) of rats as a function of concentrations of a contaminant (x). Find the asymptotic covariance matrix for your estimate using the information $\boldsymbol{I}(\boldsymbol{\beta})$. Provide a 95% confidence interval for each element of β_0 and β_1.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	6.1	4.2	0.5	8.8	1.5	9.2	8.5	8.7	6.7	6.5	6.3	6.7	0.2	8.7	7.5
y_i	0.8	3.5	12.4	1.1	8.9	2.4	0.1	0.4	3.5	8.3	2.6	1.5	16.6	0.1	1.3

Table 1: Survival times y_i and concentrations of a contaminant x_i for $i = 1, ..., 15$.

(d) Plot the log-likelihood function $l(\beta_0, \beta_1)$ and compare with the log of the asymptotic normal approximation to the sampling distribution of the MLE.

(e) Find the maximum likelihood estimate $\hat{\beta}_0$ under the null hypothesis $H_0 : \beta_1 = 0$.

(f) Perform score, likelihood ratio, and Wald tests of the null hypothesis $H_0 : \beta_1 = 0$ with $\alpha = 0.05$. In all cases explicitly indicate the formula you use to compute the test statistic.

(g) Summarize the results of the estimation and hypothesis testing presented above in a manner that would address the question of whether increasing concentrations of the contaminant had an effect on a rat's life expectancy.