14.380: Statistics

Problem Set 5

due Thursday, October 17, 2019

Problem 1 is required. All other problems are for your own practice.

1. *(Required problem)* Suppose that the random variables $Y_1, ..., Y_n$ satisfy

$$y_i = \beta x_i + e_i, \qquad i = 1, ..., n,$$

where $x_1, ..., x_n$ are fixed constants and $e_1, ..., e_n$ are i.i.d. normals with mean 0 and unknown variance σ^2. Assume that the hypothesis of interest is $H_0 : \beta = 0$.

(a) Write the likelihood function (treating both β and σ^2 as unknown). Write down score and information matrix.

(b) Find the unrestricted maximum likelihood estimator. Write the Wald test for the null hypothesis.

(c) Solve the restricted maximization problem. Write the Lagrange Multiplier test.

(d) Write down the LR test.

(e) Introduce sample correlation between y_i and x_i:

$$\widehat{r} = \frac{\sum_{i=1}^n y_i x_i}{\sqrt{\sum_{i=1}^n y_i^2}\sqrt{\sum_{i=1}^n x_i^2}}.$$

Re-write the Wald, LM and LR statistics as functions of $\widehat{r}^2$ and n only.

(f) Notice, that for any $x < 1$ the following inequality holds: $x < -ln(1-x) < \frac{x}{1-x}$. This implies some ordering of the statistics discussed above. What is it? Apparently, it holds in general for linear hypothesis in OLS models.

2. Assume that n_1 people are given treatment 1 and n_2 people are given treatment 2. Let X_1 be the number of people on treatment 1 who respond favorably to

the treatment and let X_2 be the number of people on treatment 2 who respond favorably. Assume that $X_1 \sim \text{Binomial}(n_1, p_1)$, $X_2 \sim \text{Binomial}(n_2, p_2)$ and $n_1 = \gamma n$, $n_2 = (1-\gamma)n$. Let $\psi = p_1 - p_2$.

(a) Assume that the unknown parameters are (p_1, p_2), write the likelihood. Find the MLE estimator of ψ. Would your answer change if you write likelihood in terms of parameters (ψ, p_2)?

(b) Find the Fisher information matrix $I(p_1, p_2)$.

(c) Use the multiparameter delta-method to find the asymptotic variance of $\hat{\psi}$ assuming that $n \to \infty$.

(d) Construct a Wald confidence set for ψ. Is it asymptotic? Why or why not?

(e) Test the null hypothesis $\psi = 0$ using the asymptotic LR test. Describe all the details.

3. Suppose that $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$ with known σ^2. Find a minimum value of n to guarantee that a 0.95 confidence interval for μ will have length no more than $\frac{\sigma}{4}$.

4. (Computer exercise) Suppose we observe an i.i.d. sample $(X_1, \ldots, X_n)$ from the $\text{U}[0, \theta]$ (uniform on interval $[0, \theta]$) distribution. Our goal is to produce a 95% CI for θ. We will consider two options:

(a) Finite sample distribution theory. The likelihood is

$$L(\theta) = \frac{1}{\theta^n}.$$

The MLE is: $\hat{\theta}_{\text{MLE}} = \max\{X_1, \ldots, X_n\}$. Consider the likelihood ratio (LR) test statistic

$$LR_n(\theta) = \frac{L(\theta)}{L(\hat{\theta}_{\text{MLE}})}.$$

It can be shown that under the true value θ

$$\mathbb{P}(LR_n(\theta) > \alpha) = 1 - \alpha.$$

Thus the following is a $(1-\alpha)\%$ confidence interval: $[\hat{\theta}_{\text{MLE}}, \hat{\theta}_{\text{MLE}} \cdot \alpha^{-1/n}]$.

(b) The bootstrap. For each bootstrap sample (drawn from observed sample at random with replacement), compute

$$\hat{\theta}^* = \max\{X_1^*, \ldots, X_n^*\}. \tag{1}$$

Use the bootstrap distribution of this statistic to approximate the finite sample distribution of $\hat{\theta}_{\text{MLE}}$ and therefore conduct inference on θ. That is, you want to approximate $P\{\hat{\theta}_{\text{MLE}} < x\}$ with $P^*\{\hat{\theta}_* < x\}$.

In this assignment, we will check the finite sample performance of these two approaches by performing a Monte Carlo simulation.

Suppose the true upper bound is $\theta = 5$.

i. Repeat the following $S = 1000$ times:
   (a) Generate a dataset of size n from the above model.
   (b) Construct a 95 % confidence interval using the finite sample CI described above.
   (c) Construct a 95 % confidence interval using the bootstrap percentile interval, with $B = 200$ bootstrap replications.
   (d) For each CI, record whether that CI contains the true θ.
   (e) For each CI, record the length of that CI.

ii. For each CI, estimate the finite sample coverage probability.

iii. For each CI, estimate the average length of the CI. Repeat the above procedure for $n = 25, 100, 400, 1600$. Report your results. After doing this, set $n = 25$, $S = 1$, and $B = 20,000$ and plot the histogram of your B bootstrap estimates $\hat{\theta}^{*,B}, \ldots, \hat{\theta}^{*,B}$. Use $400 = B/50$ bins. Discuss all of your findings.