STA 711 Homework 8

Due: Monday, April 3, 12:00pm (noon) 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.

Simulation study with the central limit theorem

The central limit theorem tells us that if $Y_1, Y_2, ...$ is a sequence of iid random variables, then

$$\frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1),$$

where $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $\mu = \mathbb{E}[Y_i]$, and $\sigma^2 = Var(Y_i)$. Using the central limit theorem, the Wald test rejects $H_0 : \mu = \mu_0$ in favor of $H_A : \mu \neq \mu_0$ when

$$|Z_n| = |\sqrt{n}(\overline{Y} - \mu_0)/\sigma| > z_{\alpha/2}$$

The goal of this section is investigate how large n needs to be before the normal approximation from the central limit theorem is reasonable.

1. Choose a non-normal distribution (e.g., a Bernoulli, a Poisson, a Gamma, etc.). Let μ_0 be the mean of your chosen distribution, and σ^2 the variance. Begin with $n = 5$.
   (a) Sample $Y_1, ..., Y_n$ iid from your chosen distribution. Calculate $Z_n = \sqrt{n}(\overline{Y} - \mu_0)/\sigma$.
   (b) Repeat (a) many times, and make a plot comparing the distribution of your simulated Z_n to a $N(0,1)$ distribution (e.g, a quantile-quantile plot).
   (c) If we were testing $H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \mu_0$ at level $\alpha = 0.05$, for what fraction of the simulated tests in (b) do you reject H_0 (i.e., what is the type I error)?
   (d) For the same chosen distribution, repeat (b) and (c) for $n = 10, 15, 20, 30, 50, 75, 100$. Make two plots: one comparing the distribution of your test statistics to a $N(0,1)$ for each n, and one plotting the type I error as a function of n.
   (e) Using the plots in (d), how large does n need to be before the normal approximation seems reasonable?
2. Repeat question 1 for at least three other population distributions. Experiment in particular with population distributions which are very different from the normal distribution (e.g. discrete, or strongly skewed, or multimodal). Use your simulations to provide a rough guide for how large n needs to be for the normal approximation to be reasonable.

Likelihood ratio tests with logistic regression

In this part of the assignment, you will revisit the 2015 Gorkha earthquake data from HW 6.

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-s23.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.

3. 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 likelihood ratio 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.
4. Now the researchers want to know whether there is *any* relationship between Age and damage, after accounting for surface condition. Use a likelihood ratio 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.

Power calculation

Suppose we are working with researchers interested in the relation between caffeine intake and insomnia. The researchers conduct a sleep study with a set of n subjects, all of whom have reported difficulty sleeping.

In the study, subjects are given a warm cup of coffee at 10pm, and then asked to go to bed directly after drinking the coffee. The subjects receive coffees containing different quantities of caffeine, with $n/5$ patients randomly assigned to each of 5 treatment groups: 0mg caffeine, 25mg, 50mg, 75mg, and 100mg caffeine (for reference, a normal cup of coffee contains about 100mg of caffeine).

For each subject, the researchers record whether they fell asleep in the first hour after consuming the coffee (*note: time-to-event analysis is probably better here, but that is outside the scope of this course*). We plan to fit the following logistic regression model:

$$Sleep_i \sim Bernoulli(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Caffeine_i$$

where $Sleep_i = 1$ if subject i fell asleep during the first hour, and $Caffeine_i$ is the quantity of caffeine consumed (in mg) by subject i.

To test for a relationship, the researchers plan to test $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ using a Wald test, rejecting if the p-value is < 0.05. We also know that from prior observation, there is a 40% probability that a subject drinking decaf coffee directly before bed will fall asleep within the first hour.

5. Recall that if $\theta \in \mathbb{R}^q$ is a parameter of interest, and $\widehat{\theta}$ is the maximum likelihood estimator, then $\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, \mathcal{I}_1^{-1}(\theta))$, and the Wald statistic is

$$W = (\widehat{\theta} - \theta_0)^T \mathcal{I}(\widehat{\theta})(\widehat{\theta} - \theta_0),$$

where $\mathcal{I}_1(\theta)$ is the Fisher information for a single observation, and $\mathcal{I}(\theta) = n\mathcal{I}_1(\theta)$. Under H_0, $W \xrightarrow{d} \chi^2_q$.

With some rearrangement, it can be shown (see the class notes from March 22) that if the true value of θ is $\theta_1 \neq \theta_0$, then $W \approx \chi^2_q(\lambda)$ (the non-central χ^2_q distribution with non-centrality parameter λ), where $\lambda = (\theta_1 - \theta_0)^T \mathcal{I}(\theta_1)(\theta_1 - \theta_0)$. Given n and θ, the approximate power of the Wald test is then

$$P(\chi^2_q(\lambda) > \chi^2_{q,\alpha}),$$

where $\chi^2_{q,\alpha}$ is the upper α quantile of a χ^2_q distribution. In R, the `pchisq(...)` function allows you to specify the noncentrality parameter (`ncp`).

(a) Suppose we observe 50 subjects, and we believe that a one-mg increase in caffeine is associated with a decrease of 0.02 in the log-odds of sleep within the first hour. What is the approximate power of the Wald test for $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$?

(b) Suppose we believe that a one-mg increase in caffeine is associated with a decrease of 0.02 in the log-odds of sleep within the first hour. How many subjects n do we need to observe for the approximate power of the Wald test to be at least 0.8?

(c) Suppose we have 50 subjects. How small can the true effect of caffeine be (i.e., how small can $|\beta_1|$ be) if we want our Wald test to have an approximate power of at least 0.8?