STA 711 Homework 6

Due: Friday, March 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.

Central limit theorem with estimated variance

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 $\sigma^2 = Var(Y_i)$. This limiting distribution is useful when we want to construct confidence intervals and tests for μ, but it requires us to know σ^2. When σ^2 is unknown, we replace it with an estimate. Two possible estimators of σ^2 are:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$$

1. Our goal is to show that using $\widehat{\sigma}^2$ or s^2 in place of σ^2 does not change our limiting normal distribution. For the purposes of this problem, suppose that $Y_1, Y_2, ...$ is a sequence of iid random variables, and that the moment generating function of Y_i exists in a neighborhood of 0.

   (a) Show that $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$ and $s^2 \xrightarrow{p} \sigma^2$.

   (b) Show that

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

   and

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

   (c) If both $\widehat{\sigma}^2$ and s^2 can be used as estimates of the population variance σ^2, why do we have two estimates? The reason is that $\widehat{\sigma}^2$ is a *biased* estimator of σ^2 (that is, $\mathbb{E}[\widehat{\sigma}^2] \neq \sigma^2$), whereas s^2 is *unbiased* (that is, $\mathbb{E}[s^2] = \sigma^2$). Later in the course we will discuss the bias of estimators in more detail.

   Calculate $\mathbb{E}[\widehat{\sigma}^2]$ and $\mathbb{E}[s^2]$.

Wald testing in practice

In the second part of this assignment, you will work with a dataset from DrivenData, an online data competition site that hosts competitions aimed at improving education, health, safety, and general well being for individuals around the world.

Our data come from the 2015 Gorkha earthquake in Nepal. 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.

2. (a) Fit the logistic regression model in R, and interpret the estimated slope $\widehat{\beta}_1$ in terms of the *odds* of damage.
   (b) Calculate the estimated probability of damage for a 50 year old building with surface condition = t.
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 hypothesis 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 test whether there is a difference in the probability of damage between surface `o` and surface `t`, for a 25 year old building. Use a hypothesis 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.