Applied Statistics - Exercises 10

1. Unbiased estimators (Theory)

Consider a random sample $X_1, \ldots, X_n$ from a uniform distribution in the interval $[-\theta, \theta]$, where θ is an unknwon parameter. You are interested in estimating the values of θ.

a. Show that
$$\hat{\Theta} = \frac{2}{n}(|X_1| + |X_2| + \cdots + |X_n|)$$
is an unbiased estimator for θ. *Hint*: you may need to use the *change of variable* formula (cfr. Chapter 7 of the book).

b. Consider instead the problem of estimating θ^2. Show that
$$T = \frac{3}{n}(X_1^2 + X_2^2 + \cdots + X_n^2)$$
is an unbiased estimator for θ^2

c. Is $\sqrt{T}$ an unbiased estimator for θ? If not, discuss whether it has positive or negative bias.

2. Maximum likelihood estimator for geometric random variables (Theory)

The geometric random variable, as presented in the textbook, has the following probability mass function
$$Pr[X = k] = (1-p)^{k-1} \cdot p$$
which can be described as the probability of requiring k trials to obtain the first success in a sequence of Bernoulli trials. For this random variable, we have seen that the maximum likelihood estimator for the parameter p is
$$\hat{p} = \frac{n}{\sum_{i=0}^{n} x_i} = \frac{1}{\bar{x}}$$

However, in some contexts [1] a slightly different definition of geometric random variable is used:
$$Pr[X = k] = (1-p)^{k} \cdot p$$
This second formulation can be described as the probability of experiencing k consecutive failures before the first success.

We shall see, with this exercise, that this small change leads to a different maximum likelihood estimator for p!

a. Derive the loglikelihood function $\ell(p)$
b. Compute the derivative $\ell'(p)$ of the loglikelihood function

[1] Including the R implementation of the geometric random distribution

c. Show that the maximum likelihood estimator for p is

$$\hat{p} = \frac{n}{n + \sum_{i=0}^{n} x_i}$$

Therefore, *pay attention* to the distribution you are dealing with, always read carefully the definitions and the documentation!

3. Maximum likelihood estimator for geometric random variables (R)

In Problem 2, you showed that the geometric distribution defined as

$$Pr[X = k] = (1 - p)^k \cdot p$$

has the following maximum likelihood estimator for p:

$$\hat{p}^* = \frac{n}{n + \sum_{i=0}^{n} x_i}$$

This definition of geometric random variable is the one use by R, as state at the beginning of the "Details" section of `help(rgeom)`. In this exercise you will verify that, in this case, using the inverse of the sample mean as the estimator for p leads to heavily biased estimations.

Let $n = 200$. First of all, define a function `estimate_p` that, given the realization of a random sample of n elements it returns the estimate of p using the estimator $\hat{p}^* = \frac{n}{n+\sum_{i=0}^{n} x_i}$.

Then, define $p = 0.3$, and take a random sample of n elements using the `rgeom` function. From this random sample, estimate p using first the estimator $\hat{p} = \frac{1}{\bar{x}}$ and then using the estimator $\hat{p}^* = \frac{n}{n+\sum_{i=0}^{n} x_i}$. Compute the two values $\hat{p} - p$ and $\hat{p}^* - p$. What do the resulting numbers suggest?

Repeat the above sampling and estimation procedure 1000 times, accumulating the values $\hat{p} - p$ and $\hat{p}^* - p$ in two separate lists. Plot the two distributions, possibly overlaying them on the same plot. What can you conclude by observing the plot?

4. Linear regression model and residuals (R)

Let us take a look at the `Cars93 (MASS)` data set.

(a) Plot the mileage `MPG.highway` in the function of `Horsepower`. Compute the least-squares estimate for the regression line and add it to the plot.

(b) What the predicted mileage for a car with 225 horsepower?

(c) Compute and plot the residuals in the function of horsepower. On the basis of the residuals, is the linear model assumption reasonable?

5. Linear models (Theory)

In some situations we may know that the linear model should have some peculiarities, like having no slope, or having intercept equals to zero[2]. Answer to the two following separate questions (i.e. the answer to one doesn't depend on the answer to the other). Let U_i be random variables with expectation zero and variance σ^2.

[2]For instance we may know that when one quantity of the bivariate dataset is 0 then the other *must* be zero.

a. Consider the case $\alpha = 0$. The model then becomes $Y_i = \beta x_i + U_i$, for $i = 1, 2, \ldots, n$. Find the least squares estimate $\hat{\beta}$ for β.
b. Consider the case $\beta = 0$. The model is then $Y_i = \alpha + U_i$, for $i = 1, 2, \ldots, n$. Find the least squares estimate $\hat{\alpha}$ for α.