HW Question Week 4

STA2101F 2021

Due October 14 2021 11.59 pm

Homework to be submitted through Quercus

Part 1: Data set for project

Please submit details about the data you will use for your project. Ideally the data will have a single response or outcome variable of interest, and several potential explanatory variables. You should submit with this homework:

(1) the data source: both bibliographic and a web link
(2) the number of observations and the number of potential explanatory variables
(3) a description of the response variable
(4) a description of the potential explanatory variables
(5) the scientific question(s) of interest

When you submit the final project, it will consist of the parts listed in Slide 3 on October 6.

Part 2: Question(s) for marking

There has been a lot of talk this week about rapid testing in the schools. On one hand there seems no harm in using rapid antigen tests on a regular basis, but on the other hand if a lot of the tests give incorrect results, especially flagging as covid-related too often, then children will unnecessarily miss school. This seems to be the main concern from the public health officials who are cautioning a slower approach.

Tests for Covid19 (or any screening for that matter), are assessed by their false positive and false negative rates, or equivalently by their sensitivity and specificity. Sensitivity of the test is the *true positive rate*, i.e. $Pr(T+ \mid C+)$, and 1 minus sensitivity is the *false negative rate* $Pr(T- \mid C+)$. Specificity of the test is the *true negative rate*, i.e. $pr(T- \mid C-)$, and 1 minus specificity is the *false positive rate*. (My source is Wikipedia.)

(a) If a given student tests positive, compute the probability that s/he has Covid19 using Bayes theorem.

(b) Find some information on the true positive and true negative rates for the rapid antigen tests used in Ontario. Give the estimated rates and your source for this information.

(c) Compute the so-called "positive predictive value", i.e. $Pr(C+ \mid T+)$ using the rates from (b) and a range of plausible values for the background rate of Covid. Present the

results in a table.

(d) Your neighbour's child has just tested positive with a rapid antigen test. Explain in non-technical language how likely or not it is that the child does indeed have Covid19. (Assume the test has the same parameters as you found for (b).)

(e) Still with Covid-19, but a different question: A friend of mine noted in the summer that according to some data in the news, about 25% of the current cases of Covid19 in the hospital were in fully vaccinated people. He said "I thought the vaccine was 95% effective. What gives?" What would you tell him?

(f) *Bonus/PhD: Measurement error in regression*

Suppose y depends on x in a simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_j, \quad i = 1, \ldots, n;$$

where we assume now that $x_i \sim (\mu_x, \sigma_x^2)$, and as usual that $\epsilon_i \sim (0, \sigma_e^2)$, independently across i, with also x independent of ϵ.

Instead of observing x_i, we are only able to observe a corrupted value $w_i = x_i + u_i$, where u_i is independent of x_i, and $u_i \sim (0, \sigma_u^2)$, independently across i. The least squares estimator of β_1 is then

$$\hat{\beta}_1 = \Sigma(y_i - \bar{y})(w_i - \bar{w})/\Sigma(w_i - \bar{w})^2.$$

(i) Give an expression for the variance of $\hat{\beta}_1$.

(ii) Find an expression for the limit in probability of $\hat{\beta}_1$ and thus deduce that $\hat{\beta}_1$ will tend to be an underestimate of the true regression coefficient β_1. In what special circumstance will it be consistent for β_1?

Usually this result is relied on to argue that if there is uncertainty in the x's used in a given regression, the association with the response will be attenuated, i.e. less likely to be significantly different from zero. A closely-related discussion, somewhat more elaborate, is discussed in LM-2 7.1 and LM-1 5.1.