BIOSTAT/STAT 570: Coursework 7

To be submitted to the course canvas site by 11:59pm Friday 2nd December, 2022.

1. (a) In this question a simulation study to investigate the impact on inference of omitting covariates in logistic regression will be performed, in the situation in which the covariates are independent of the exposure of interest. Let x be the covariate of interest and z another covariate. Suppose the true (adjusted) model is $Y_i \mid x_i, z_i \sim_{iid}$ Bernoulli(p_i), with

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i + \beta_2 z_i. \tag{1}$$

A comparison with the unadjusted model $Y_i \mid x_i \sim_{iid}$ Bernoulli$(p_i^\star)$, where

$$\log\left(\frac{p_i^\star}{1-p_i^\star}\right) = \beta_0^\star + \beta_1^\star x_i, \tag{2}$$

for $i = 1, \dots, n = 1000$ will be made. Suppose x is binary with $\Pr(X = 1) = 0.5$ and $Z \sim_{iid}$ N$(0, 1)$ with x and z independent. Combinations of the parameters $\beta_1 = 0.4, 1.1$ and $\beta_2 = 0.6, 1.2, 2.5$, with $\beta_0 = -2.5$ in all cases, will be considered.

For each combination of parameters compare the results from the two models, (1) and (2), with respect to:

i. E$[\widehat{\beta}_1]$ and E$[\widehat{\beta}_1^\star]$, as compared to β_1.
ii. The standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_1^\star$.
iii. The coverage of 95% confidence intervals for β_1 and $\beta_1^\star$.
iv. The probability of rejecting $H_0 : \beta_1 = 0$ (the power) under both models using a Wald test.

Based on the results, summarize the effect of omitting a covariate that is independent of the exposure of interest, in particular in comparison with the linear model case.

2. The Pima, or Akimel O'odham, are an indigenous Native American tribe that originates from southern Arizona. In this question you will analyze data on Pima native American women who are at least 21 years of age; these data were originally from the National Institute of Diabetes and Digestive and Kidney Diseases. We will take the aim of the analysis to obtain a model to understand the association between diabetes status and the covariates in the sampled population, using binomial GLMs. The data may be found in the `mlbench` library and are called `PimaIndiansDiabetes2`.

   The variables we will examine as predictors are:

`pregnant`	Number of times pregnant
`glucose`	Plasma glucose concentration (glucose tolerance test)
`mass`	Body mass index (weight in kg/(height in m^2)
`pedigree`	Diabetes pedigree function
`age`	Age (years)

   (a) We will examine models of the form

$$\begin{aligned} Y_i \mid p_i &\sim \text{Binomial}(1, p_i) \\ g(p_i) &= \beta_0 + \beta_1 \times \texttt{pregnant} + \beta_2 \times \texttt{glucose} + \beta_3 \times \texttt{mass} + \beta_4 \times \texttt{pedigree} + \beta_5 \times \texttt{age} \end{aligned}$$

   for $i = 1, \ldots, n$ women, and where the link function $g(\cdot)$ is one of `logit`, `probit`, `cloglog`.

   Form a new dataset containing y and the required x variables, removing the records that contain missing values.

   (b) Fit the three binomial models that correspond to the different link functions, and give a table containing the parameter estimates along with standard errors.

   (c) Which link function provides the most interpretable coefficients? For this link function, provide a brief summary of your fitted model.

   (d) Provide a plot showing the estimated association between diabetes prevalence and pedigree, under the three models. Provide pointwise confidence intervals to your plot, carefully explaining your method. How should one interpret this plot?

   (e) Suppose the aim of the analysis was prediction, rather than understanding associations. How would this affect the way you carry out your analysis, and how would you assess the success of a model? There is no need to do any analyses here, I just want to see an outline of what you would do.