APTS Statistical Modelling: Assessment

Helen Ogden

April 2023

Students should talk to their supervisors to find out whether or not their department requires this work as part of any formal accreditation process (APTS itself has no resources to assess or certify students). It is anticipated that departments will decide on the appropriate level of assessment locally, and may choose to drop some (or indeed all) of the parts, accordingly.

0. If you have not already done so, complete both APTS week practical sessions.
1. (a) Show that AIC for a normal linear model with n responses, p explanatory variables and unknown σ^2 may be written as

   $$n \log \hat{\sigma}^2 + 2p + c$$

   where $\hat{\sigma}^2 = RSS/n$ is the maximum likelihood estimate of σ^2 and c is a constant which does not depend on the model under consideration, so may be omitted without affecting model selection.

   (b) If $\hat{\sigma}_0^2$ is the unbiased estimate under some fixed 'correct' model with $q > p$ covariates, show that AIC is equivalent to using

   $$n \log \left\{ 1 + (\hat{\sigma}^2 - \hat{\sigma}_0^2)/\hat{\sigma}_0^2 \right\} + 2p$$

   as a model comparison criterion, and that this is approximately equal to

   $$C_p = n \left(\hat{\sigma}^2/\hat{\sigma}_0^2 - 1 \right) + 2p,$$

   a quantity known as Mallows' C_p. Deduce that model selection using Mallows' C_p approximates that using AIC.

   (c) In the same context as (b), show that $C_p = (q-p)(F-1) + p$ where F is the F-statistic for comparison of the models with p and $q > p$ covariates. Deduce that if the model with p covariates is correct then $E(C_p) \doteq p$ but that otherwise $E(C_p) > p$.
2. The data frame `bacteria` are discussed in Chapter 10 of *Modern Applied Statistics with S (4th edition)* by Venables and Ripley (Springer, 2002). They are available in R by loading the library `MASS`. The response `y` indicates presence or absence of a particular bacteria when assessed on 50 individuals (`ID`) at each of up to 6 time points (`week`). Each individual has received one of three treatments (`trt: placebo/drug/drug+`).

Model the dependence of `y` on `trt` and week using binary GLMs and GLMMs (to account for intra-subject dependence in the response), fitted by maximum likelihood and associated approximations. Functions which you might wish to investigate for doing this include `glmmPQL` (from the `MASS` library) and `glmer` (from the `lme4` library). Use the library documentation provided to learn about the required arguments of these functions. Compare the inferences obtained by different fitting methods (quadrature, Laplace, PQL).

3. Suppose that we have binary data $Y_1, \dots, Y_n$, and a single explanatory variable x_i, which we model by using a logistic regression model

$$Y_i \sim \text{Bernoulli}(\mu_i), \quad \text{logit}(\mu_i) = \beta_0 + \beta_1 x_i. \tag{1}$$

In reality, suppose these binary variables have been generated according to whether an unobserved continuous variable Y_i^* exceeds 0, that is

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that

$$Y_i^* = \beta_0^* + \beta_1^* x_i + \epsilon_i,$$

where ϵ_i are independent and identically distributed error terms with $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = 1$. We consider three possibilities for the error distribution:

- $\epsilon_i \sim N(0, 1)$.
- ϵ_i have logistic distribution with mean zero and scale parameter $s = \sqrt{3}/\pi$, with cumulative distribution function $F(x) = \text{logit}^{-1}(x/s)$.
- ϵ_i have uniform distribution between $-\sqrt{3}$ and $\sqrt{3}$.

(a) In each case, find an expression for $\mu(x) = E(Y_i|x_i = x)$ in terms of x and the parameters β_0^* and β_1^*, and state whether or not the model (1) is correctly specified.

(b) Now suppose the data are generated with $\beta_0^* = 0$ and $\beta_1^* = 1$. In each case, make plots of $\mu(x)$ and $\text{logit}(\mu(x))$ for x between -2 and 2. For which true error distribution do you think the model misspecification will be most serious?

(c) Suppose that the observed explanatory variables x_i are uniformly distributed between -1 and 1. Based on your plots of $\text{logit}(\mu(x))$, make a guess about the approximate limiting values of $\hat{\beta}_0$ and $\hat{\beta}_1$ as $n \to \infty$ in each case.

(d) In each case, generate data in `R` according to the true data generating process with a large n (e.g. $n = 10\,000$), fit a logistic regression model to your simulated data, and check whether your estimates are close to the limiting values you guessed in part (c).

(e) We are often interested in the log odds ratio, which we can think of as the derivative of $\text{logit}(\mu(x))$ with respect to x. We estimate the log odds ratio as $\hat{\beta}_1$. In each misspecified case, what is the range of true log odds ratio (for x in the range -1 to 1)? Where is the error in the estimated log odds ratio greatest?