MAST90104: A First Course in Statistical Learning

Week 10 Lab and Workshop

1. We revisit the `pima` dataset in Week 9. Remember that the data may be found in the package `faraway`.

   (a) This question use a data set in package `faraway`. Load the package and read the help file (`?pima`) to get a description of the predictor and response variables, then use `pairs` and `summary` to perform simple graphical and numerical summaries of the data.

   Use the same set codes in Q2(a) Week 9 to remove observations with missing values.

   (b) Fit a probit regression model with `test` as the response and all the other variables as predictors.

   Answer the following questions using your fitted probit regression model.

   (c) Is the diastolic blood pressure significant in the regression model? Use your R output to evaluate its significance at 10% significance level.

   (d) Write down the formula for the fitted regression equation using your R output.

   (e) Predict the outcome for a woman with predictor values 1, 99, 64, 22, 76, 27, 0.25, 25 (same order as in the dataset). Give a 95% confidence interval for your prediction. Explain why the confidence is not symmetric about the estimated probability.

2. In this question, we will generate simulated data using a probit model.

   (a) Write a function in R with argument n that sets the random seed as `set.seed(n)` and generates independent draws $\{y_i\}_{i=1}^n$, where each y_i is drawn as

   $$y_i \sim \text{Bin}(6, \Phi(-0.5 + 0.1x_{i1} - 0.2x_{i2}))$$

   and each $\mathbf{x}_i = (x_{i1}, x_{i2})$ are drawn from a bivariate normal distribution with mean $\mathbf{0}$ and identity covariance matrix.

   (b) Use the function in part (a) to generate a dataset of size $n = 30$.

   (c) Use the simulated dataset from part (b) to fit the binomial probit model:

   $$y_i \sim \text{Bin}(6, \Phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}))$$

   (d) Using your fitted model in part (c), construct a 90% confidence interval for

   $$\Phi(\beta_0 - 0.5\beta_1 - 0.5\beta_2).$$

1 Workshop questions

1. Suppose $Y_i, i = 1, \cdots, n$ are from a generalised linear model so they are independent from an exponential family:
$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$
with the parameter ϕ constant and supposed known but θ_i varies. Recall that
$$\begin{aligned}\mu &= \mathbb{E}Y = b'(\theta)\\ V(\mu) &= \operatorname{Var} Y = b''(\theta)a(\phi)\\ v &= b'' \circ (b')^{-1}\end{aligned}$$
and that there is a link function, g, so that $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ are the parameters of interest, $\mu_i = \mathbb{E}Y_i$ and $\mathbf{x}_i$ is a vector of explanatory variables (this is the ith row of the predictor matrix X). In answering the questions below, you will establish that the Newton-Raphson method with Fisher scoring is the same as the iteratively weighted least squares algorithm introduced in lectures.

   (a) Write down the log likelihood as a function of $\boldsymbol{\beta}$ and show that its derivative, $U(\beta_j)$, with respect to β_j may be written as:
$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)}.$$

   (b) Hence show that
$$Cov(U(\beta_j)U(\beta_k)) = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{V(\mu_i)(g'(\mu_i))^2}.$$

   (c) Find the Fisher information and show that it is $X^T W(\boldsymbol{\beta})X$ where $W(\boldsymbol{\beta})$ is a diagonal matrix whose ith diagonal entry is
$$\frac{1}{V(\mu_i)(g'(\mu_i))^2}.$$

2. Suppose that students answer questions on a test and that a specific student has an aptitude T. A particular question might have difficulty d_i and the student will get the answer correct only if $T > d_i$. Consider d_i fixed and $T \sim N(\mu, \sigma^2)$, then the probability that a randomly selected student will get the answer wrong is $p_i = \mathbb{P}(T < d_i)$.

   Show how you might model this situation using a probit regression model.

3. Show that the Gamma density, f, in the form
$$f(y;\lambda,\alpha) = \frac{1}{\Gamma(\alpha)}\lambda^\alpha y^{\alpha-1}e^{-\lambda y}$$
is an exponential family with $\theta = -\frac{\lambda}{\alpha}, \phi = \frac{1}{\alpha}$. Identify the functions a, b, c and find the mean and variance functions as functions of θ.

4. Show that the inverse Gaussian density, f, in the form
$$f(y;\mu,\lambda) = \frac{\lambda}{\sqrt{2\pi y^3}}e^{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}}$$
is an exponential family with $\theta = \frac{-1}{2\mu^2}, \phi = \frac{1}{\lambda}$. Identify the functions a, b, c and find the mean and variance functions as functions of μ, λ.