MAST90104: Introduction to Statistical Learning

Week 8 Lab and Workshop

The lab problems this week are on binomial regression. The workshop problems are on reparameterisation of less than full rank models - filling in some more challenging aspects than last week's workshop problems and problem 7 is on binomial regression.

1 Lab

1. The dataset `wbca` comes from a study of breast cancer in Wisconsin. There are 681 cases of potentially cancerous tumors of which 238 are actually malignant. Determining whether a tumor is really malignant is traditionally determined by an invasive surgical procedure. The purpose of this study was to determine whether a new procedure called fine needle aspiration, which draws only a small sample of tissue, could be effective in determining tumor status.

   (a) Load the data and read descriptions of the variables using

   ```
   library(faraway)
   data(wbca)
   ?wbca
   ```

   (b) Fit a binary regression model (logistic regression in this case) using `glm`. Include all the variables in your model (shorthand for this in an R model is $\sim$ `.`).

   (c) Use the `step` function to search for a model with minimal AIC. Include all variables in the `scope` (type `?step` to see how to use `step`).

   You should end up with the model `cbind(Class, 1 - Class)` $\sim$ `Adhes + BNucl + Chrom + Mitos + NNucl + Thick + UShap`.

   (d) Using the reduced model, use `predict` to estimate the outcome for a new patient with predictors 1, 1, 3, 1, 1, 4, 1. You will need to put `newdata = list(Adhes=1, BNucl=1, Chrom=3, Mitos=1, NNucl=1, Thick=4, UShap=1)` and `type="response"`.

   To get a 95% CI for your estimate, use `predict` with `type="link"` and `se.fit=TRUE`, to obtain the estimate and its standard error *on the linear scale*. Use these to get a symmetric CI on the linear scale, which you can then transform back to the response scale.

   (e) Suppose that a cancer is classified as benign if $p > 0.5$ and malignant if $p < 0.5$. Compute the number of errors of both types that will be made if this method is applied to the current data with the reduced model.

   (f) Suppose we change the cutoff to 0.9 so that $p < 0.9$ is classified as malignant and $p > 0.9$ as benign. Compute the number of errors in this case.

   Consider how you might determine the cutoff in practice.

2. The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the the dataset `pima`.

   (a) 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.

   There are some obvious irregularities in the data. Take appropriate steps to correct the problems.

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

   Can you tell whether this model fits the data?

   Odds are sometimes a better scale than probability to represent chance. The odds o and probability p are related by

$$o = \frac{p}{1-p} \quad p = \frac{o}{1+o}$$

In a binomial regression model with a logit link we have

$$\text{logit}(p_j) = \log\left(\frac{p_j}{1-p_j}\right) = \eta_j = \beta_0 + \beta_1 x_{1,j} + \cdots + \beta_q x_{q,j}.$$

That is $\log o_j = \eta_j$, where o_j are the odds for the j-th observation.

(c) By what proportion do the odds of testing positive for diabetes change for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Give a confidence interval for this difference.

(d) Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.

(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 confidence interval for your prediction.

3. Consider the binomial regression model with logit link fitted to the Challenger data in class. Using the log likelihood ratio, plot a 95% confidence region for (α, β).

   One way of doing this is to use the function `contour`:

   (a) Let $(\hat{\alpha}^*, \hat{\beta}^*)$ be the MLE, then for a grid of α and β values calculate $2l(\hat{\alpha}^*, \hat{\beta}^*) - 2l(\alpha, \beta)$.

   (b) The contour line with value $\chi^2_2(0.95)$ will delineate the confidence region.

2 Workshop

4. Suppose the less than full rank matrix X is $n \times p$ of rank r and that C is $p \times r$. Suppose further that X has r linearly independent columns and that the corresponding rows of C are also linearly independent. The following parts combine to show that XC is full rank if, and only if, $I_r + DE$ is rank r where, if necessary by reordering the rows and columns of X and the rows of C, X & C have been partitioned as

$$X = \left[\begin{array}{c|c} X_r & X_r D \\ \hline FX_r & FX_r D \end{array}\right] \quad C = \left[\begin{array}{c} C_r \\ \hline EC_r \end{array}\right],$$

   X_r, F, D, C_r, E are respectively $r \times r, n - r \times r, r \times p - r, r \times r \,\&\, p - r \times r$ and X_r, C_r are both rank r.

   (a) Show that the rows and columns of X can be rearranged to achieve the partitions given.

   (b) Show that $r(XC) = r(I_r + DE)$.

   (c) Show that XC is full rank if, and only if, $I_r + DE$ is rank r.

5. For $r = 3$ the matrices D and E and verify that $I_r + DE$ is rank r for C_r from the `contr.treatment` and `contr.sum` matrices in R. Use these to veryify the reparameterisation equations given in notes.

6. Prove Theorem 6.2 using the following steps.

   (a) Show that under the conditions of Theorem 6.1 (question 4 above), the column space of XC is the same as the column space of X.

   (b) Show that if two full-rank linear models have the same column space, the eigenvectors of their hat matrices are the same.

   (c) Hence show that if the column space for two linear models is the same, the fitted values are the same.

   (d) Complete the proof of Theorem 6.2.

7. Verify that for the binomial regression model with logistic link

$$\begin{aligned} \mathbb{E}\frac{\partial l(\boldsymbol{\theta}; \mathbf{Y})}{\partial \theta_i} &= 0 \\ -\mathbb{E}\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{Y})}{\partial \theta_i \partial \theta_j} &= \mathbb{E}\left(\frac{\partial l(\boldsymbol{\theta}; \mathbf{Y})}{\partial \theta_i}\frac{\partial l(\boldsymbol{\theta}; \mathbf{Y})}{\partial \theta_j}\right) \end{aligned}$$