MAST90104: A First Course in Statistical Learning

Week 9 Practical and Workshop

1 Practical questions

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. The dataset `discoveries` lists the number of great scientific discoveries for the years 1860 to 1959, as chosen by "The World Almanac and Book of Facts", 1975 Edition. Has the discovery rate remained constant over time?

   To answer this question, fit a Poisson regression model with a log link, and use the deviance to compare a null model with models including the year and year squared as predictors.

2 Workshop questions

Note: There will be no workshop class this week due to the AFL Grand Final Eve holiday, the following questions are for students to practice in their own time (optional).

1. The `infert` dataset from the `survival` package presents data from a study of infertility after spontaneous and induced abortion. Using a logistic regression model, analyse and report on the factors related to infertility based on this data. (Don't use the factor stratum, as it is confounded with the other predictors.)

2. The `dvisits` data in the `faraway` package comes from the Australian Health Survey of 1977–78 and consist of 5190 observations on single adults, where young and old have been oversampled.

   (a) Build a Poisson regression model with `doctorco` as the response and `sex, age, agesq, income, levyplus, freepoor, freerepa, illness, actdays, hscore, chcond1` and `chcond2` as possible predictor variables. Considering the deviance of this model, does this model fit the data?

   (b) Plot the response residuals against the fitted values. Why are there lines of observations on the plot?

   (c) Use backward elimination with a critical p-value of 5% to reduce the model as much as possible.

   (d) What sort of person would be predicted to visit the doctor the most under your selected model?

   (e) For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0,1,2, etc. times.

   (f) Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how the Gaussian and Poisson models differ.

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 $\theta = \frac{1}{\mu^2}, \phi = \frac{1}{\lambda}$. Identify the functions a, b, c and find the mean and variance functions as functions of μ, λ.