SM 4337 - Applied Statistics

Tutorial 4

1. 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 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.

   (b) Can you find any obvious irregularities in the data? If you do, take appropriate steps to correct the problems.

   There are missing observations (0) for many variables; change missing values to NA instead of removing errors.

   (c) Redo (a) after correcting the problems.

   (d) Fit a model with the result of the diabetes *test* as the response and all the other variables as predictors. mod ← glm (cbind (test, 1−test) ~ ., family = binomial, data = pima2

   (e) What are the variables that are statistically significant from the model in (d)? glucose, bmi, diabetes

   (f) Find out whether this model fits the data.

   (g) Write down the fitted model.

   (h) Predict the outcome for a woman with predictor values: *pregnant=1, glucose=99, diastolic=64, triceps=22, insulin=76, bmi=27, diabetes=0.25, age=25.* Give a confidence interval for your prediction.

2. The dataset *wbca* comes from a study of breast cancer in Wisconsin. It consists of medical data from 681 women who has potentially cancerous tumors, of which 238 are actually malignant while the remaining 443 are benign. Determining whether a tumor is really malignant is traditionally done 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.

$$\text{logit}(p) = -10.04 + 0.082X_1 + 0.038X_2 - 0.0014X_3 + 0.011X_4 - 0.00083X_5 + 0.071X_6 + 1.141X_7 + 0.034X_8$$

where X_1 : pregnanant

X_2 : glucose

X_3 : diastolic

X_4 : triceps

X_5 : insulin

X_6 : bmi

X_7 : diabetes

X_8 : age

The residual deviance is 89.464 with d.f. 671. The resid. dev. & d.f. has a very big difference when they have to be balanced.

(a) Fit a binomial regression with *Class* as the response and the other nine variables as predictors. Report the residual deviance and associated degrees of freedom. Can this information be used to determine if this model fits the data? Explain.

(b) Use AIC as the criterion to determine the best subset of variables. (Use the *step* function.) The AIC is now minimized from 109.456 to 105.66

(c) Write down the reduced model in (b).

(d) Use the reduced model to predict the outcome for a new patient with predictor variables *Adhes=1, BNucl=1, Chrom=3, Epith=2, Mitos=1, NNucl=1, Thick=4, Ushap=1 and USize=1.* Give a confidence interval for your prediction. $p = 0.9921$; $(0.9757, 0.9975)$

$$\text{logit}(p) = 11.0333 - 0.3984X_1 - 0.4192X_2 - 0.5679X_3 - 0.6456X_4 - 0.2915X_5 - 0.6216X_6 - 0.2541X_7$$

where X_1 : Adhes

X_2 : BNucl

X_3 : Chrom

X_4 : Mitos

X_5 : NNucl

X_6 : Thicc

X_7 : UShap