

MAST30027: Modern Applied Statistics

Week 2 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) ~ 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 dataset `pima`. Read the help file (`?pima`) to get a description of the predictor and response variables. There are missing observations for many variables, which have been recorded as zeros. The easiest (not necessarily the best) way to deal with these is to remove the relevant observations from the data set.

```
> missing <- with(pima, missing <- glucose==0 | diastolic==0 | triceps==0 | bmi == 0)
> pima <- pima[!missing,]
```

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