Generalized Linear Models

Lecture 7. Models with binary response II

Existence of estimates

Lemma (Claudia Czado, München, 2004)

The log-likelihood $\ln L(\boldsymbol{\beta})$ in logistic regression is strict concave in $\boldsymbol{\beta}$ if $rank(\mathbf{X}) = p$.

This implies that the score equations can have at most one solution

$\Rightarrow$ if a ML estimate of $\boldsymbol{\beta}$ exists, it is unique and it is a solution to score equations

Existence of estimates, R

```
> sep1=glm(y~x1+x2,data=separ,family="binomial")
Warning message:
glm.fit:  fitted probabilities numerically 0 or 1 occurred
```

What's going on?

What is the cause?

How to proceed?

Infinite parameter estimates (1)

Parameter estimates $\hat{\beta}$ are found using ML method

Estimates for parameters exist $\Leftrightarrow$ iteration converges

The existence of a MLE depends on points in the observation space, i.e. data (Albert & Anderson, 1984)

ML estimates exist if there exists no hyperplane separating the values of the response

Three possible scenarios

- complete separation
- quasi-complete separation
- overlap

Separation – a covariate or a set of covariates determine the response ($y_i = 0$ or $y_i = 1$)

Large standard errors of parameters are an indication of possible separation issues

Infinite parameter estimates (2)

Complete separation

Arguments can divide the response values to exact groups (prediction in each group exactly 1 or 0)

ML estimates do not exist, log-likelihood tends to zero when the number of iterations increases

Quasi-complete separation

For at least one subject, it is not exactly fixed to which response group it belongs

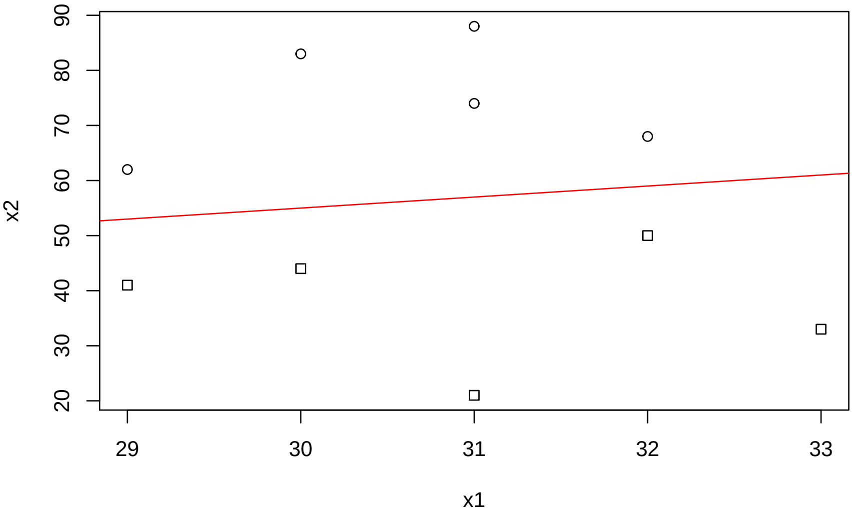
ML estimates do not exist, log-likelihood does not tend to zero, but the information matrix is unbounded and the inverse does not exist

Overlap

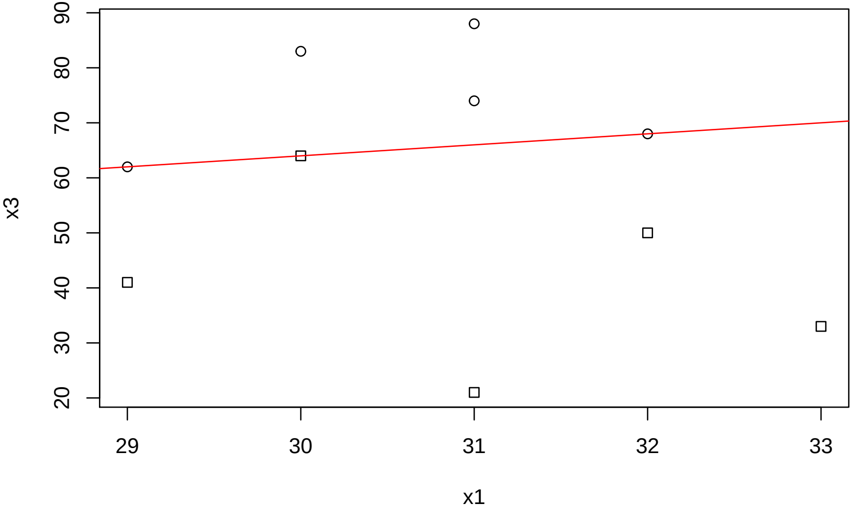
If there is no separation, then there is overlap

ML estimate $\hat{\beta}$ exists and is unique

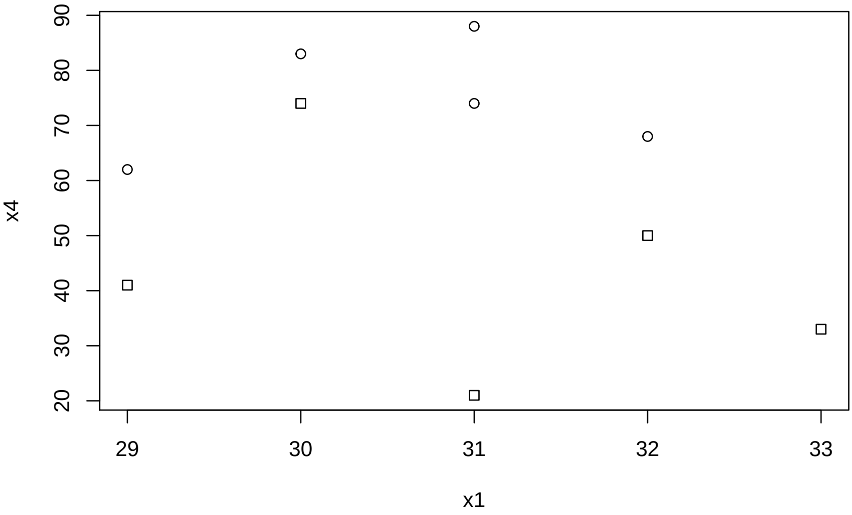
One can describe overlap as a number observations that can be removed to reach partial separation, i.e. situation when the parameters can no longer be estimated. In Vaso example, overlap is only 3 observations



Complete separation



Quasi-complete separation



Overlap

How to solve the separation issue?

How to proceed?

- Find out which arguments cause the problem, leave some arguments (or observations) out, or re-code
- *Penalized likelihood* (Firth, 1993) – add an adjustment term (which result skewed estimates), used for continuous arguments
- *Exact logistic regression* – used when number of parameters is small, samples are small and arguments are discrete

Firth's method (Firth, 1993)

Idea: add an adjustment (penalty) term to the log-likelihood and maximize the penalized log-likelihood. Information matrix remains unchanged

Method is asymptotically consistent, i.e. the estimate converges to ML estimate

Idea is similar to *ridge regression*, which is used in case of multicollinearity

In R: function `logistf` (package `logistf`)

```
> logistf(y~x1+x2,data=separ)
logistf(formula = y ~ x1 + x2, data = separ)
Model fitted by Penalized ML
Confidence intervals and p-values by Profile Likelihood

                  coef    se(coef)   lower 0.95 upper 0.95       Chisq
(Intercept) -1.7984331 22.19477577 -52.60524094 65.3741588 0.006847239
x1          -0.1642127  0.72706874  -2.78341758  1.2547411 0.053029128
x2           0.1216285  0.06997688   0.02404056  0.3756169 7.380218208
                      p
(Intercept) 0.934051880
x1          0.817873779
x2          0.006594517

Likelihood ratio test=7.570012 on 2 df, p=0.02270872, n=10
```

Classification problem

A GLM (eventually) predicts the probabilities of an event (or nonevent)

If we need to classify the results, how to do that?

Simplest way is to say that $pi_i \leq 0.5$ means 0 and $\pi_i > 0.5$ means 1, but is it actually the best option? Maybe another cut-off point is better? Which one? How to compare the classification ability of models using different cut-offs?

Confusion matrix:

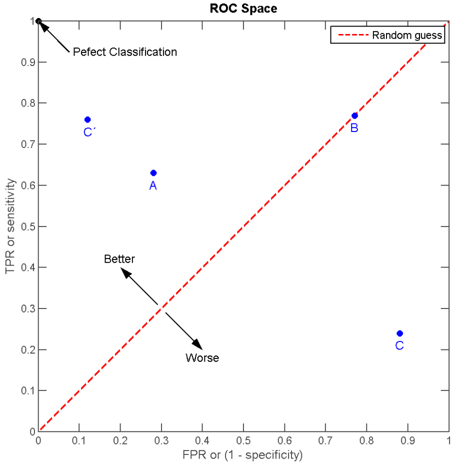
		Actual (true) value	
		1	0
Predicted value	1	True positives (TP)	False positives (FP)
	0	False negatives (FN)	True negatives (TN)

Now

- $TP + FN = P$ – total actual positives
- $TN + FP = N$ – total actual negatives
- $\frac{TP}{P}$ – true positive rate (also sensitivity)
- $\frac{TN}{N}$ – true negative rate (also specificity)
- $\frac{TP+TN}{P+N}$ – accuracy of the model

Receiver operating characteristic

ROC curve allows us to compare models based on their classification ability
The bigger area under ROC curve (AUC), the better



ROC Space. Source: Wikipedia

Decisions

Which cut-off point (probability) to choose?

Possible options:

- the point closest to perfect classification
- the point farthest from the random guess line
- the point that maximizes accuracy
- the point that minimizes the cost (individual costs are specified by a cost matrix)
- the point where sensitivity = specificity

In R: library `ROCR` or `pROC`

A nice example by Arthur Charpentier:
`https://freakonometrics.hypotheses.org/48285`