Simple logistic regression

April 19, 2020

Applied STAtistics group at AAU

Department of Mathematical Sciences

Aalborg University

AALBORG UNIVERSITY
DENMARK

Introduction

Outline of session:

- Data
- Model
- Inference

Lecturer for this session is Ege Rubak, Dept. of Math. Sciences, AAU

Data

Wisconsin Breast Cancer Database covers 683 observations of 10 variables in relation to examining tumors in the breast.

- Nine clinical variables with a score between 0 and 10.
- The binary variable `Class` with levels `benign/malignant`.

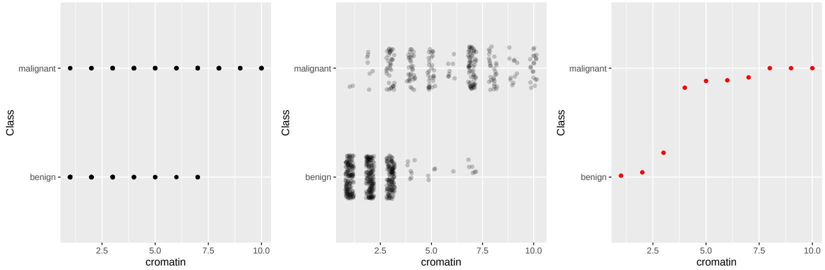
We will use 4 of the predictors, where 2 have been discretized.

id	nuclei	cromatin	Size.low	Size.medium	Shape.low	Class
1	1	3	TRUE	FALSE	TRUE	benign
2	10	3	FALSE	TRUE	FALSE	benign
...	...	...	...	...	...	...
25	7	3	TRUE	FALSE	FALSE	malignant
26	1	2	TRUE	FALSE	TRUE	benign
...	...	...	...	...	...	...
682	4	10	FALSE	FALSE	FALSE	malignant
683	5	10	FALSE	FALSE	FALSE	malignant

Plot of data

Three different plots of the same data, where from left to right:

- many points are plotted on top of each other
- points are plotted as semi-transparent and "jittered"
- fractions are plotted instead of "0"s and "1"s.



Binary response

- We consider a binary response y with outcome 1 or 0, e.g. malignant or beneign.
- Furthermore, we are given an explanatory variable x, which is numeric, e.g. score.
- We shall study models for

  $$P(y = 1 \mid x)$$

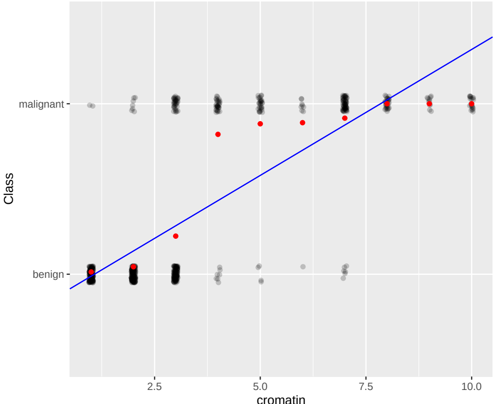
  e.g. the probability that a tumor with score x is malignant.
- We shall see methods for determining whether or not score actually influences the probability, i.e. is y independent of x?

A linear model

- The simple linear model is often inappropriate.

$$P(y = 1 \mid x) = \alpha + \beta x$$

- If β is positive and x sufficiently large, then the probability exceeds 1.



Logistic model

Instead we consider the **odds** that the tumor is malignant

$$\texttt{Odds}(y = 1 \mid x) = \frac{P(y = 1 \mid x)}{P(y = 0 \mid x)} = \frac{P(y = 1 \mid x)}{1 - P(y = 1 \mid x)}$$

which can have any positive value.

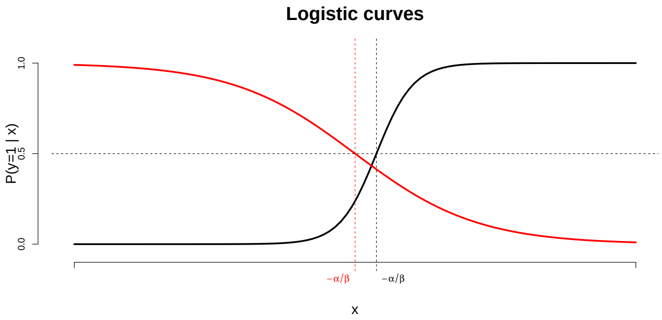
The logistic model is defined as:

$$\texttt{logit}(P(y = 1 \mid x)) = \log(\texttt{Odds}(y = 1 \mid x)) = \alpha + \beta x$$

The function $\texttt{logit}(p) = \log(\frac{p}{1-p})$ - i.e. **log of odds** - is termed **the logistic transformation**.

Remark that log odds can be any number, where zero corresponds to $P(y = 1 \mid x) = 0.5$. Solving $\alpha + \beta x = 0$ shows that for the score $x_0 = -\alpha/\beta$ the tumor has fifty-fifty chance of being malignant.

Simple logistic regression



Examples of logistic curves. The black curve has a positive β-value (=10), whereas the red has a negative β (=-3). Note that:

- Increasing the absolute value of β yields a steeper curve.
- When $P(y = 1 \,|\, x) = \frac{1}{2}$ then logit is zero, i.e. $\alpha + \beta x = 0$.

Odds-ratio

Interpretation of β:

What happens to odds, if we increase x by 1?

Consider the so-called **odds-ratio**:

$$\frac{\mathtt{Odds}(y = 1 \mid x + 1)}{\mathtt{Odds}(y = 1 \mid x)} = \frac{\exp(\alpha + \beta(x + 1))}{\exp(\alpha + \beta x)} = \exp(\beta)$$

where we see, that $\exp(\beta)$ equals the odds for score $x + 1$ relative to odds for score x.

This means that when x increases by 1, then the relative change in odds is given by $100(\exp(\beta) - 1)\%$.

Inference

- For the cancer data the estimates are

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.28	0.3919	-13.47	2.201e-41
cromatin	1.365	0.1173	11.64	2.624e-31

- With $\hat{\alpha} = -5.28$ and $\hat{\beta} = 1.37$ we see that a score of $-\hat{\alpha}/\hat{\beta} = 3.87$ corresponds to fifty/fifty risk of malignant tumor.
- Since $\exp(\hat{\beta}) = 3.92$, increasing the score by 1 increases the risk of malignant tumor by 292%.
- Null hypothesis of no relation between score and class of tumor is

$$H_0 : \quad \beta = 0$$

  with the alternative $\beta \neq 0$.
- $\hat{\beta}$ is 11.6 standard errors away from zero, so H_0 is clearly rejected with a p-value of practically zero.

Confidence interval for odds ratio

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.28	0.3919	-13.47	2.201e-41
cromatin	1.365	0.1173	11.64	2.624e-31

From the summary:

- Standard error on $\hat{\beta}$ is 0.12 and hence a 95% confidence interval for log-odds ratio is $\hat{\beta} \pm 1.96 \times 0.12 = (1.14, 1.6)$.
- Corresponding interval for odds ratio:
  $(\exp(1.14), \exp(1.6)) = (3.11, 4.93)$,

  i.e. the relative increase in odds is - with confidence 95% - between 211% and 393%.

Plot of model predictions against actual data