Applied Machine Learning in Health Sciences 2023
-
Logistic regression

Peter Mondrup Rasmussen

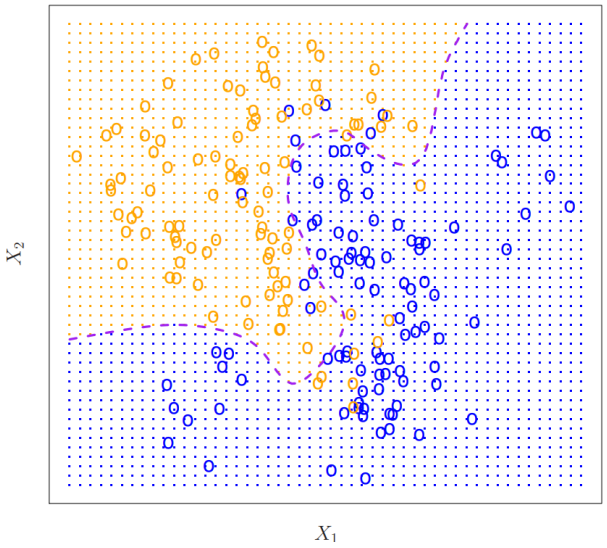
CFIN

pmr@cfin.au.dk

Logistic regression

Coding of response variables in classification

- In classification we have a categorical output/response variable e.g. {control, disease}, {child, teenager, adult, senior}.
- In many software packages incl. Matlab you can directly use such categorical responses when building models.
- Alternatively a generic numerical coding can also be used e.g. {0,1} {-1,1} {1,2,3,4}.



Logistic regression

- As for linear regression we here use the simplifying term *logistic regression* to cover both simple- and multiple logistic regression.
- Logistic regression is used to model the *probability* that an observation belongs to a particular category in a binary classfication task.
- Without loss of generality we can use the generic coding {0,1} of the response Y.
- As a model, logistic regression uses the logistic function

$$p(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

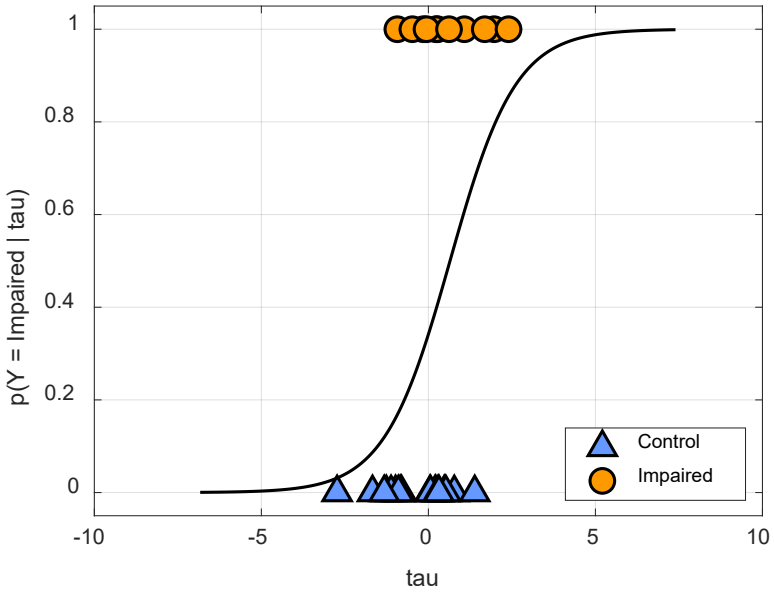
  to model the probability that an observation X belongs to class 1.
- The probability that an observation belongs to class 0 is simply

$$p(Y = 0|X) = 1 - p(Y = 1|X)$$

Logistic regression

$$p(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

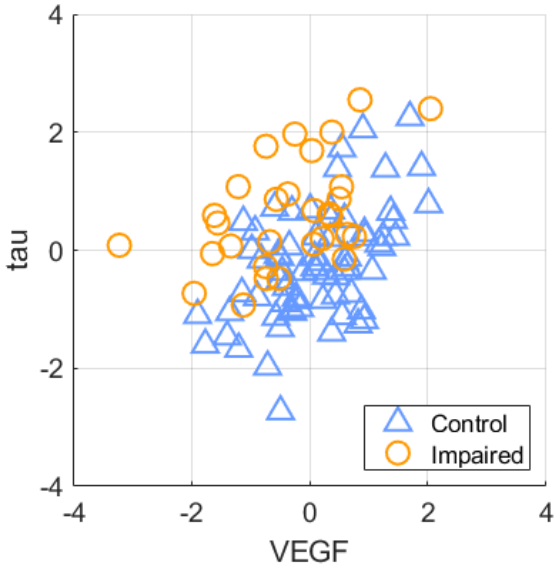
- The logistic regression model produces *sigmoid* outputs

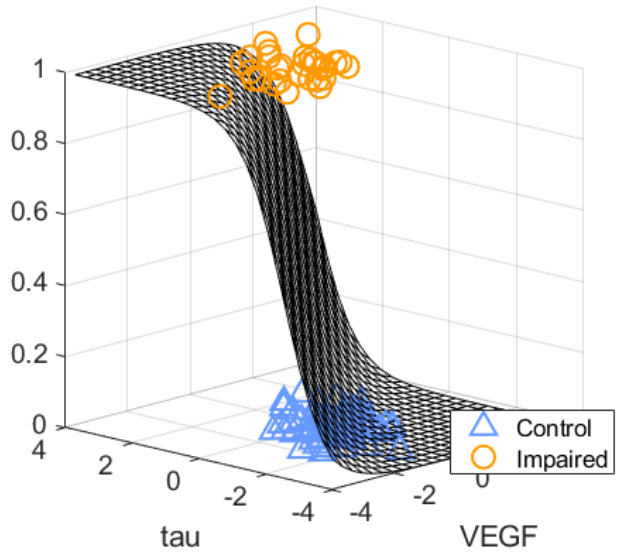


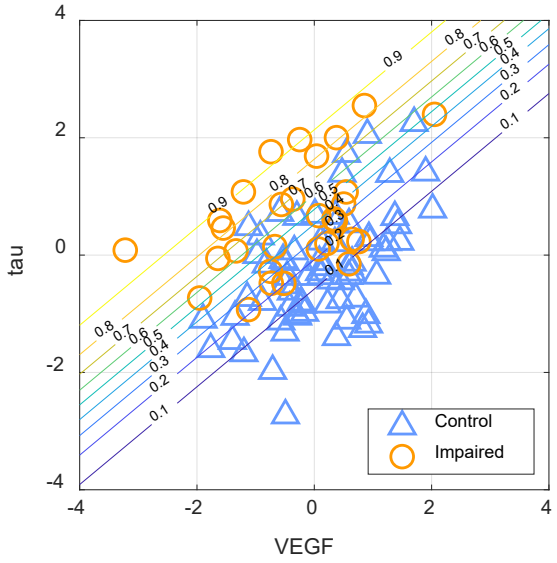
Logistic regression

$$p(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- The logistic regression model produces *sigmoid* outputs





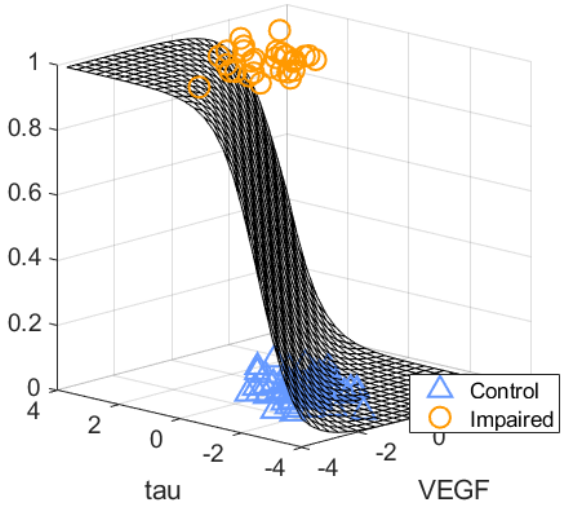


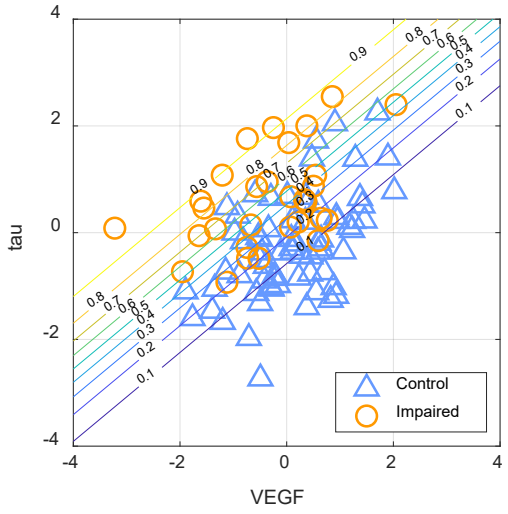
Logistic regression

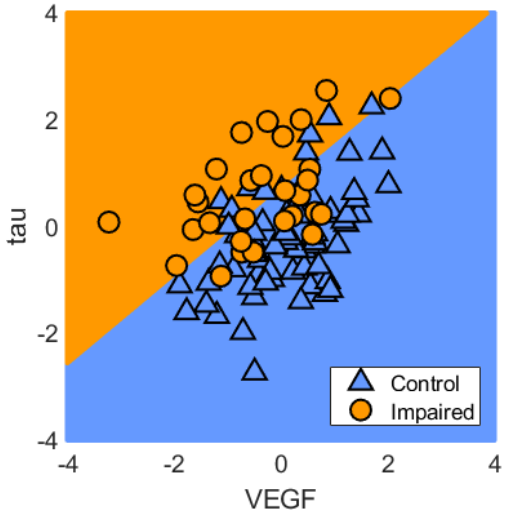
We can rewrite the logistic regression model

$$log\left(\frac{p(Y=1|X)}{1-p(Y=0|X)}\right) = \beta_0 + \beta_1 x_1 + \cdots \beta_P x_P$$

The lhs. is called the *log odds* or *logit*. Logistic regression has an output that is nonlinear in X, but the log odds is linear in X.

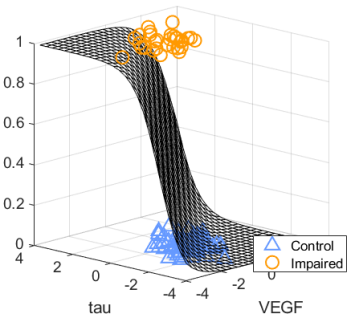


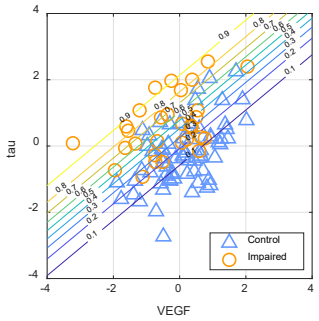


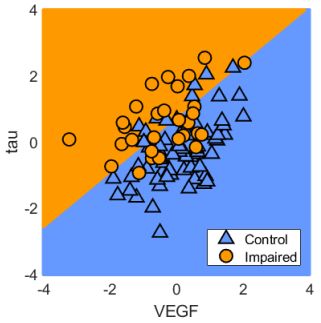


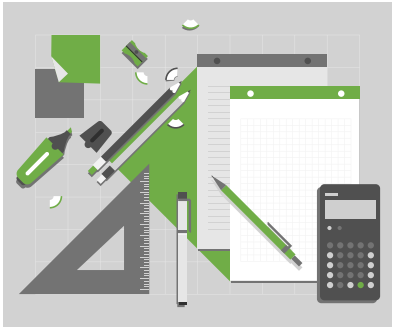
Logistic regression

- Logistic regression models the *propability* of discrete classes Y given input X. The model output is non-linear in X.
- The log odds is linear in X, and the decision surface/-boundary produced by logistic regression is also linear in X.
- The two-class logistic regression is easily extended to more classes (K>2) which is known as *multinomial logistic regression*.
- The logistic regression model is a simple but very useful classification model for binary classification - especially when combined with *regularization* which we will look at later.









In-class exercises

5 Logistic Regression

References

Figures from James et al. *An Introduction to Statistical Learning*, second edition, https://www.statlearning.com/resources-second-edition