IDS 702 Introduction to logistic regression/GLMs

Agenda

1. Motivation

2. Logistic regression and link functions

1. Motivation



Example data Pumpkins!!!

- Want to characterize differences in two classes of pumpkin seeds
- Outcome: Urgüp sivrisi or çerçevelik pumpkin seeds
- Predictors: area, perimeter, major axis length, minor axis length, convex area, diameter, eccentricity, solidity, extent, roundness, aspect ratio, compactness

KOKLU, M., SARIGIL, S., & OZBEK, O. (2021). The use of machine learning methods in classification of pumpkin seeds (Cucurbita pepo L.). Genetic Resources and Crop Evolution, 68(7), 2713-2726. Doi: https://doi.org/10.1007/s10722-021-01226-0



Why not linear regression for binary data?

2. Logistic regression and link functions

What to do? Start with distribution for Y

What to do? Connect probability to RHS

Math!

GLM Terms

- nonlinear) function of $X\beta$
- explanatory variables

• Generalized linear models: a class of models in which the response variable Y is assumed to follow a distribution*, which is assumed to be some (often

• Link function: connects the distribution of the outcome Y to $X\beta$. Indicates how the expected value of the response relates to the linear combination of



GLM Assumptions

- The data are independently distributed (cases are independent)
- Dependent variable Y follows a specified distribution (e.g., bernoulli for logistic regression)
- Linear relationship between the transformed expected response in terms of the link function and the explanatory variables

Why no error term?