Logistic regression

Overview

Information on final project

Logistic regression basic ideas

Logistic regression in R

Visualizing multiple regression models with ggplot

Final projects!

The final project is a 5-8 page R Markdown report where you analyze your own data to address a question that you find interesting

- It's a chance to practice everything you've learned in class!

Sources for data sets are listed on Canvas

- You can use data you collect as well. If you use data for another class your work must be unique for each class.

An R Markdown template describing sections in the project is on the class GitHub site.

- An R function you can run to download this template is listed on Canvas.

Final projects!

A key challenge is going to be to fit your analyses into 5-8 pages:

- You can include an appendix with additional code that does not count against your 5-8 pages
  - although this might be fully evaluation so do not include critical information there

Project is due at 11:59pm on Sunday December 6th

- i.e., the day before the start of reading period

Final exam is on Wednesday December 16th at 9am

Questions about anything?

Logistic regression

Logistic regression

In **logistic regression** we try to predict whether a case belongs to one of two categories

- Does a case belong to category *a* or category *b*?
- Example: can we predict if a faculty member is an Assistant of Full professor based on the salary level?

Making predictions for a categorical variable is called **classification**

- The field of machine learning has developed many classification methods

In logistic regression we build a conditional probability model:

- Pr(Class = a | x)
- Pr(Assistant Professor | salary = $60,000)

Logistic regression

Question: could we use linear regression to make these predictions?

$$Pr(Y = a \mid x_1) = \beta_0 + \beta_1 x_1$$

Problem: we could get negative probabilities and probabilities greater than 1!

Logistic regression

Question: what if we transformed the probability to odds?

$$\frac{\Pr(Y = a \mid x_1)}{\Pr(Y = b \mid x_1)}$$

Question: what are the range of values odds can take on?

A: 0 to ∞

Logistic regression

Instead we model the log odds as a linear function of our predictors

$$log(\frac{Pr(Y=a|x)}{Pr(Y=b|x)})$$

log-odds or logit

This scales values in the range of [0 1] to values in the range of (-∞ ∞)

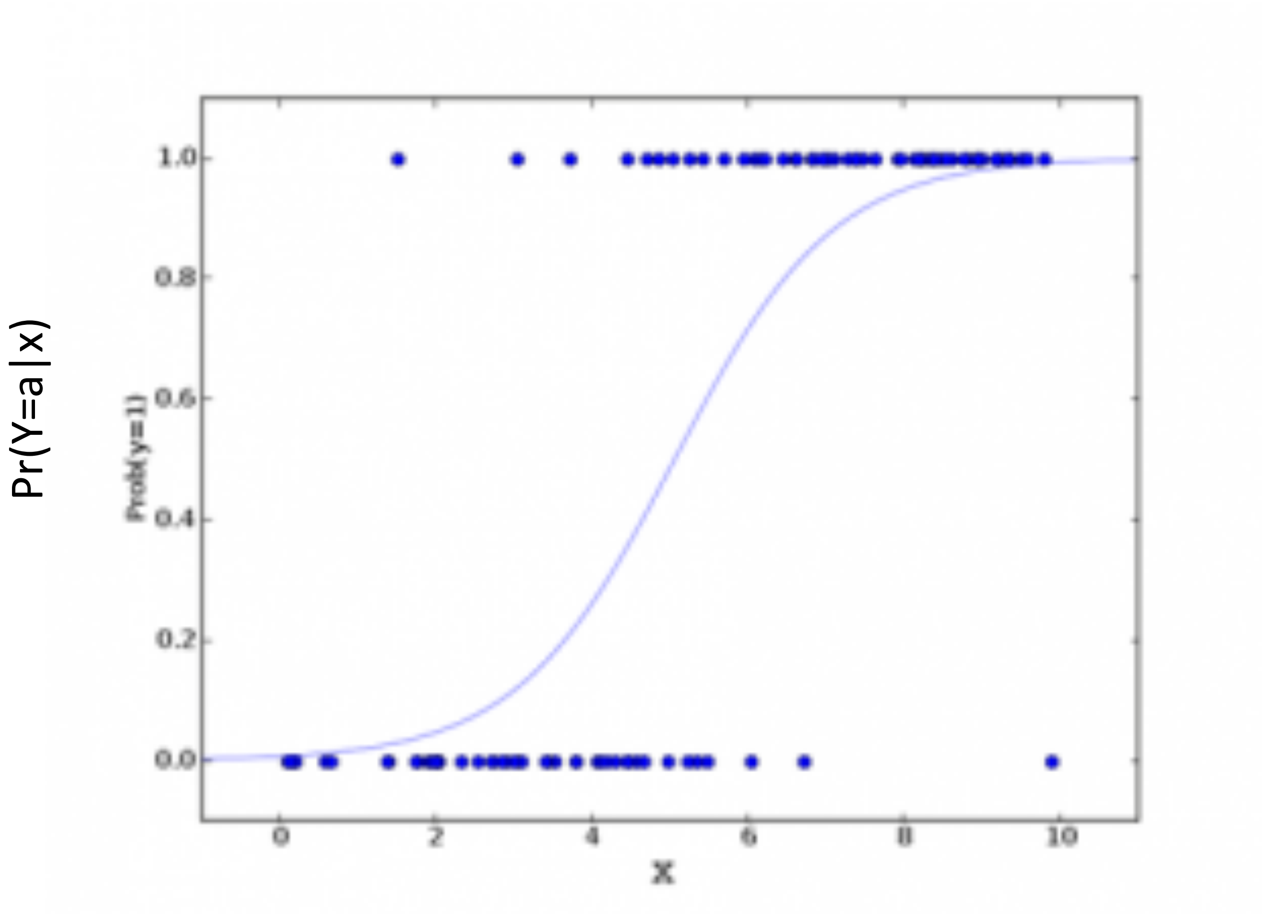
Logistic regression

$$log(\frac{Pr(Y{=}a|x)}{1{-}Pr(Y{=}a|x)}) = \beta_0 + \beta_1 \cdot x$$

Solving for Pr(Y = a| x):

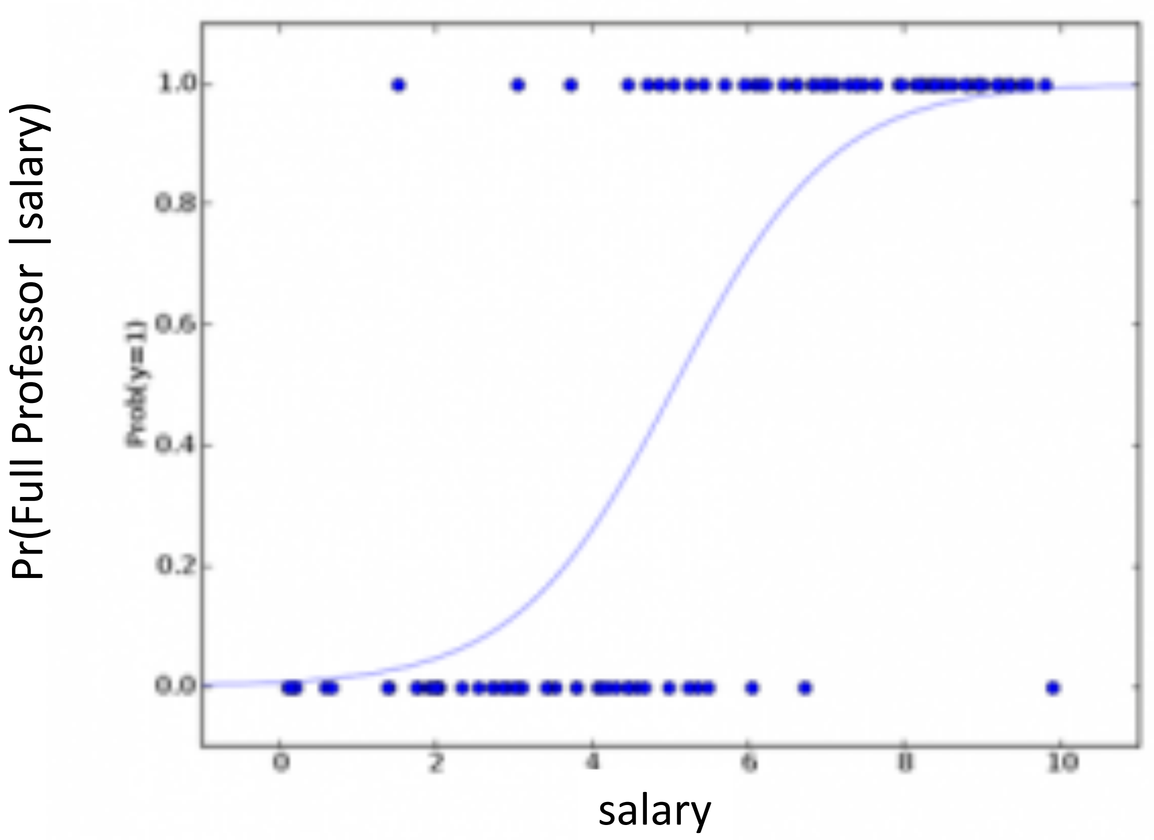
$$Pr(Y = a|x) = \frac{e^{\beta_0 + \beta_1 \cdot x_1}}{1{+}e^{\beta_0 + \beta_1 \cdot x_1}}$$

Plotting Pr(Y=a|x) as a function of x



$$Pr(Y = a|x_1) = \frac{e^{\beta_0 + \beta_1 \cdot x_1}}{1 + e^{\beta_0 + \beta_1 \cdot x_1}}$$

Plotting Pr(Y=a|x) as a function of x



$$Pr(\text{ Full Professor } \mid \text{ salary }) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot \text{salary}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot \text{salary}}}$$

Let’s look at this in R…