Logistic regression

Jeff Leek

@jtleek

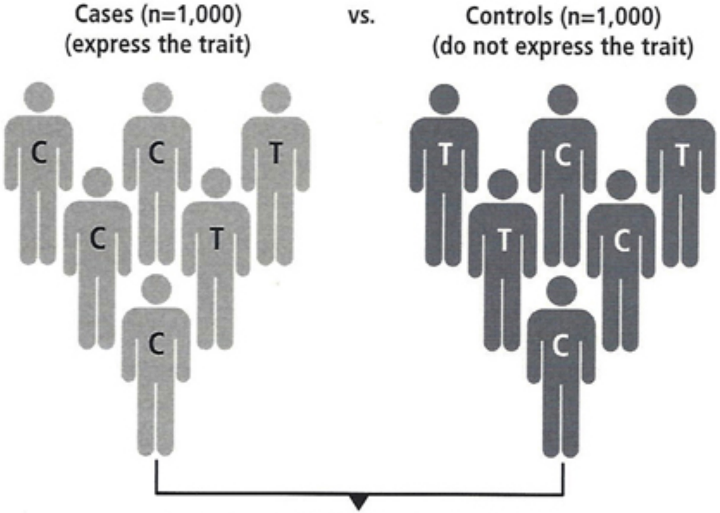
www.jtleek.com

Data aren’t always “Normal”

Case/control status is binary

http://www.discoveryandinnovation.com/BIOL202/notes/lecture25.html

Case-control study for genetic association



	C	T
Cases	62%	38%
Controls	49%	51%

[redacted]

$$C_i = b_0 + b_1 G_i + e_i$$

C = 1 if case, 0 if control

G = 0 if C, 1 if T

Not continuous

$$C_i = b_0 + b_1 G_i + e_i$$

C = 1 if case, 0 if control

G = 0 if C, 1 if T

Between 0 and 1

$$\Pr(C_i = 1) = b_0 + b_1 G_i + e_i$$

C = 1 if case, 0 if control

G = 0 if C, 1 if T

Always less than 0

$$\log(p) = b_0 + b_1 G_i + e_i$$

C = 1 if case, 0 if control

G = 0 if C, 1 if T

p = Pr(C = 1)

Log odds can be any number

$$\log(p/(1-p)) = b_0 + b_1 G_i + e_i$$

C = 1 if case, 0 if control

G = 0 if C, 1 if T

p = Pr(C = 1)

Increase in log odds of case status given genotype

$$\log(p/(1-p)) = b_0 + b_1 G_i + e_i$$

$C = 1$ if case, 0 if control

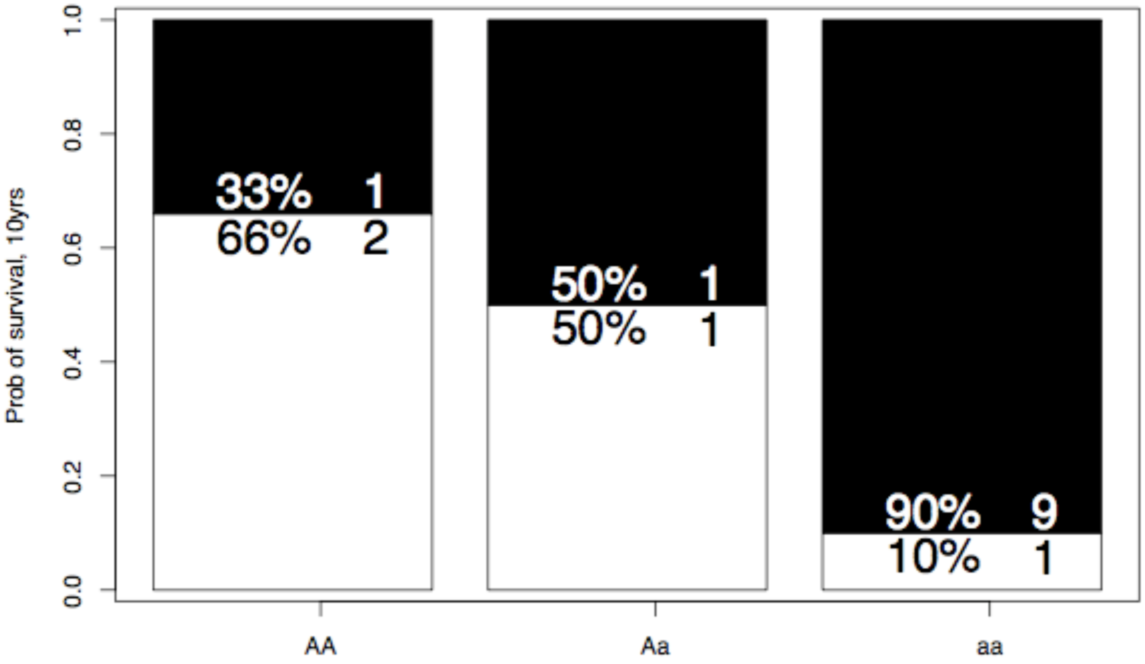
$G = 0$ if C, 1 if T

$p = \Pr(C = 1)$

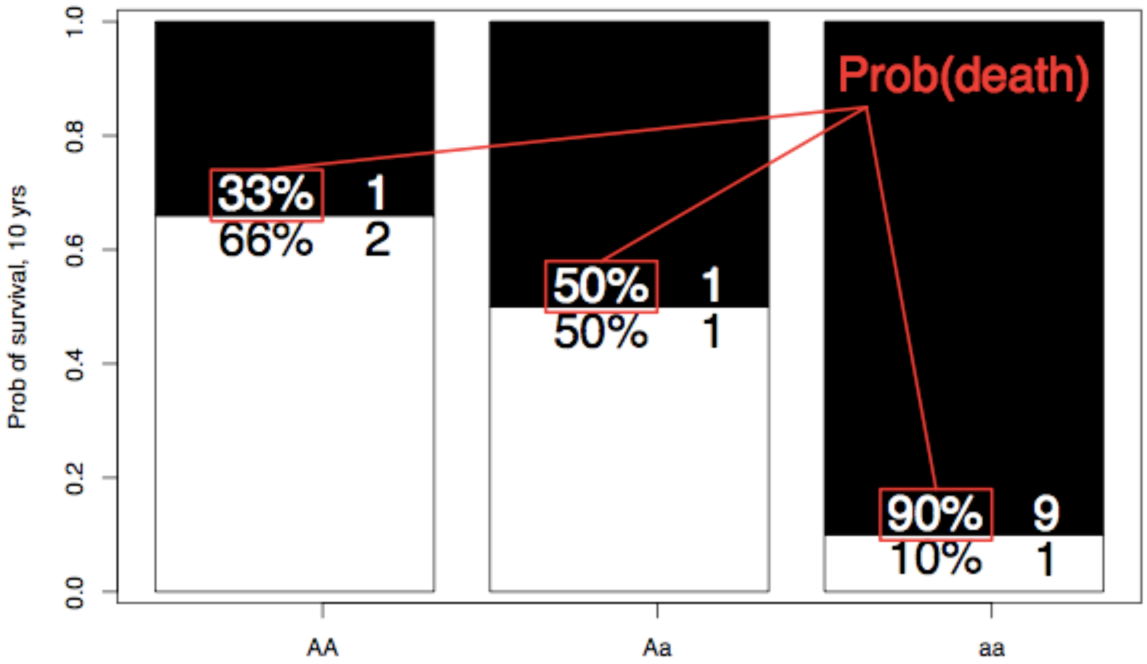
Odds/log odds

Quantity	Log Odds	Odds
Definition	log(p/(1-p))	p
In logistic regression	b	exp(b)
Definition of “no” effect	0	1

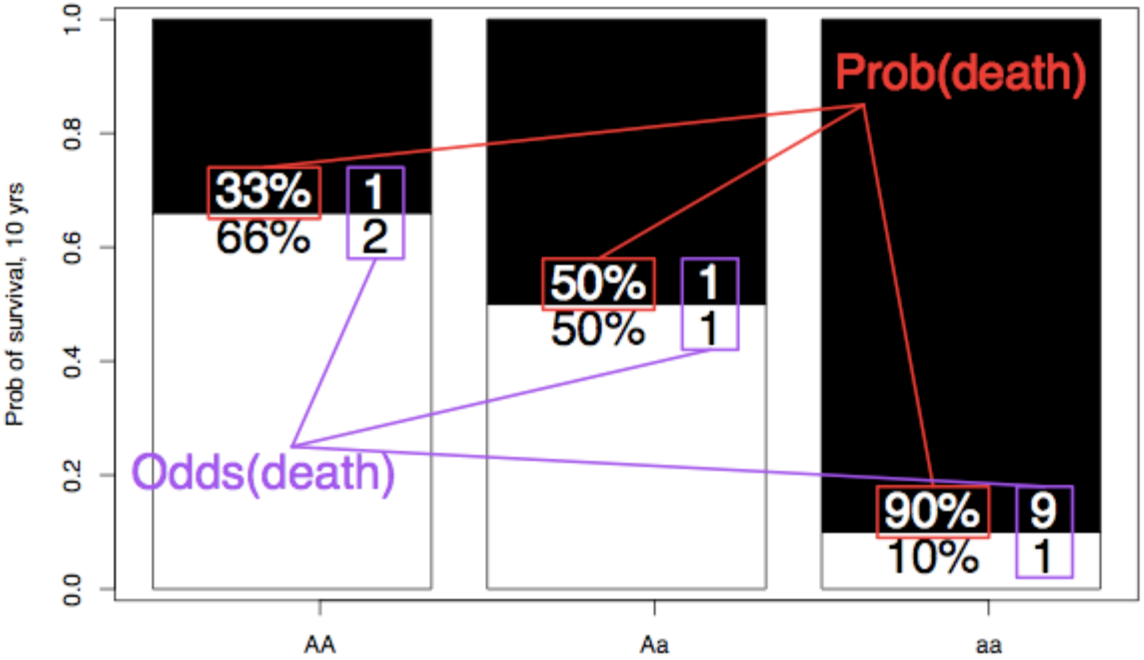
Slide courtesy: Ken Rice



Slide courtesy: Ken Rice

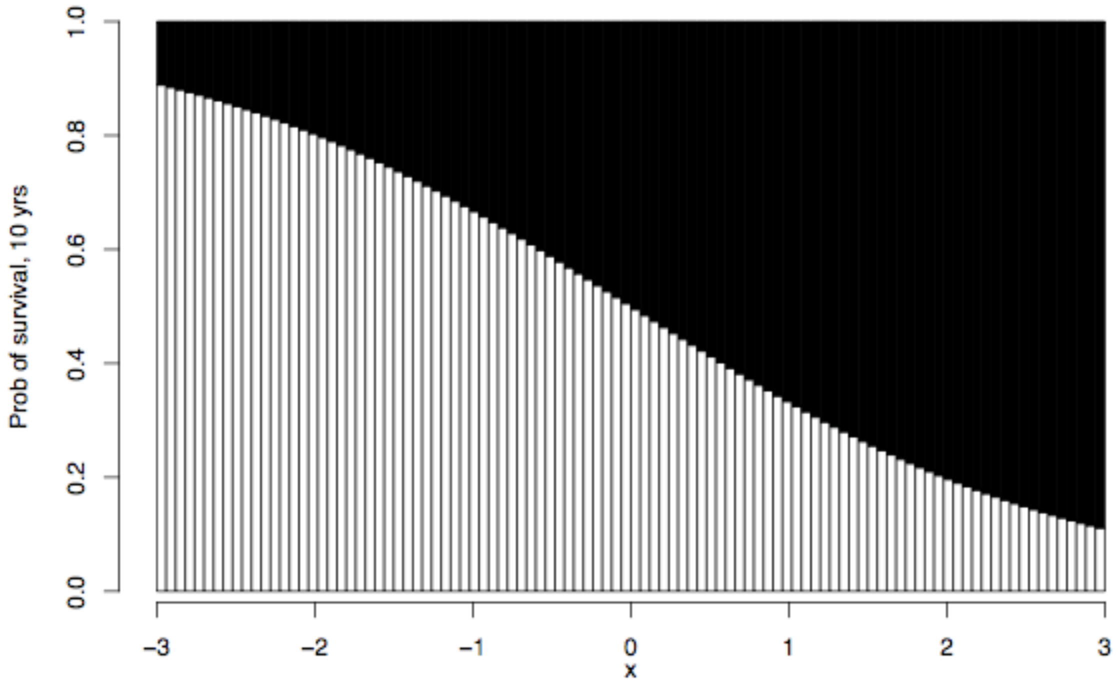


Slide courtesy: Ken Rice

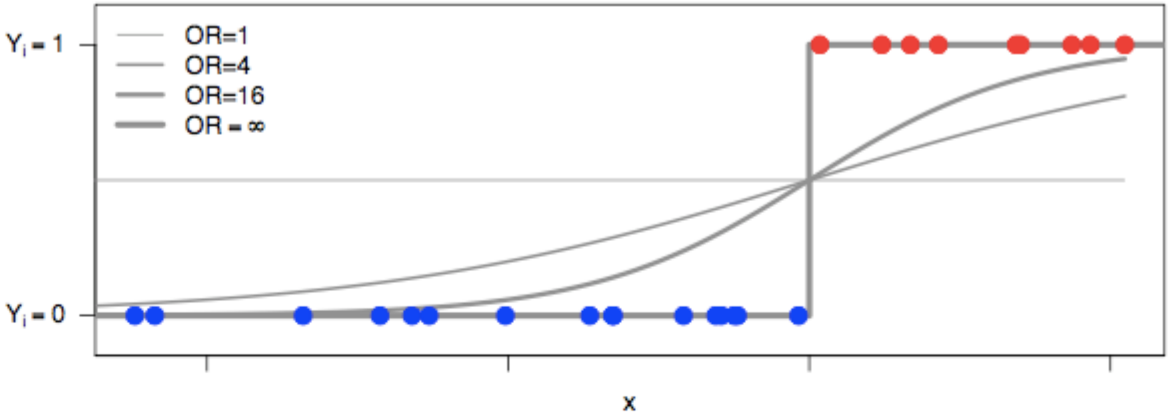


Slide courtesy: Ken Rice

Odds ratio of 2



Slide courtesy: Ken Rice



Notes and further reading

- Logistic regression is a “generalized linear model”
  - https://en.wikipedia.org/wiki/Generalized_linear_model
- A nice set of lecture notes
  - http://data.princeton.edu/wws509/notes/
- This is again a huge topic and we have only scratched the surface.