logistic regression

a gentle introduction

April 24, 2024

1

why is logistic regression needed?

- with a categorical outcome, a regular linear model will fail in several ways
- (we'll focus here on binary outcomes)

- assumptions of normality and homoscedacity will typically be violated
- the linear model will make nonsensical predictions
- the relationship between predictors and the outcome will quite likely be nonlinear

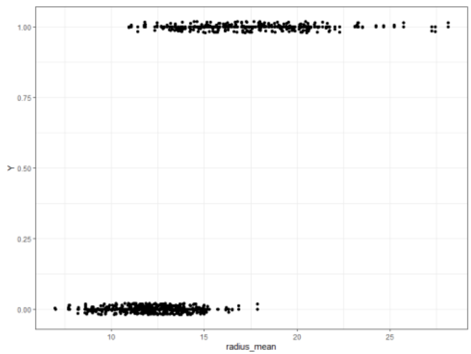
2

some (real? fake?) data

- diagnosis of breast cancer tumors as malignant or benign

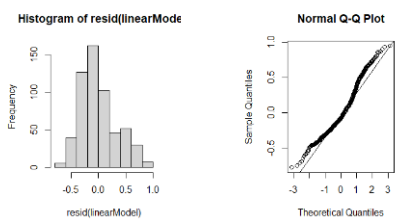
- outcome: malignant, benign (really it's probability of being malignant or benign)
- predictor: mean radius of tumor

3



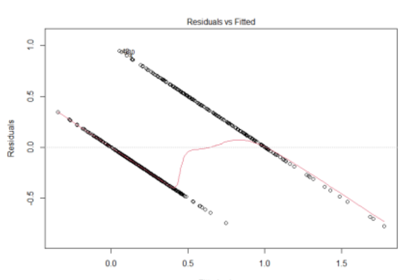
4

linear model assumptions: normality



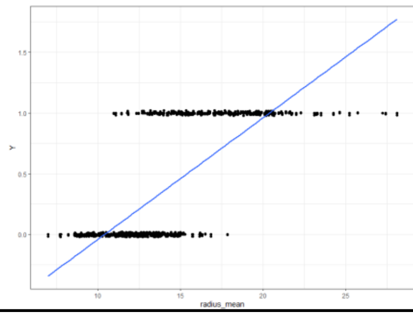
5

linear model assumptions: homoscedasticity



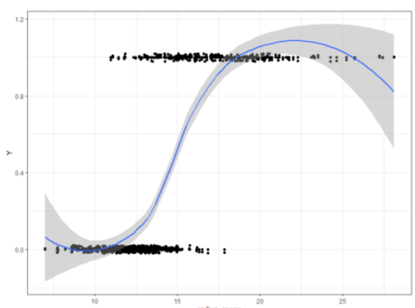
6

nonsensical predictions



7

nonlinearity



8

a little more about nonlinearity

- imagine we want to predict whether someone will buy a house based on household income
- the predicted change in the outcome is not constant per unit increase in the predictor; that is, it depends on the value of the predictor

income	p(buy)
$40K	
$50K	
$60K	
$70K	
$80K	
$90K	
…	
$340K	
$350K	
$360K	
$370K	

9

so what will we do?

- we'll transform the outcome from probability first to odds, and then we'll take the logarithm of the odds
- we'll take this in two steps to talk about why

10

odds

- odds are the ratio of the probability of an event happening to it not happening
- that is

$$odds = \frac{p(A)}{p(\sim A)} = \frac{p(A)}{1 - p(A)}$$

11

some examples

- if p = .5, odds = .5 / .5 = 1
- if p = .25, odds = .25 / .75 = .333
- if p = .75, odds = .75 / .25 = 3
- p > .5 → odds > 1
- p < .5 → odds < 1

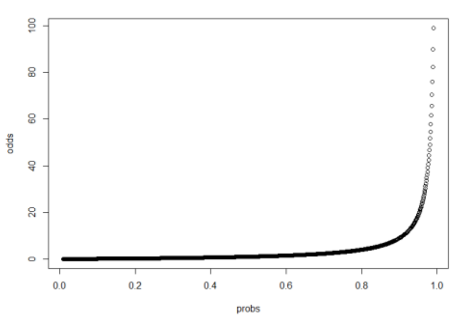
12

odds have no upper bound, but they do have a lower bound

- p = .99 → odds = .99 / .01 = 99
- p = .999 → odds = .999 / .001 = 999

- p = .01 → odds = .01 / .99 = .0101…
- p = .001 → odds = .001 / .999 = .001001…

13



14

logarithms

- the logarithm of a number is the power to which some "base" must be raised to equal the number

- for example, the base 10 logarithm of 100 is 2 because

$$10^X = 100 \rightarrow X = 2$$

- this would be written as $\log_{10}(100) = 2$

15

natural logarithms

- natural logarithms (typically denoted *ln*) are logarithms with *e* as a base

$e \approx 2.718$

- if we take the natural logarithm of odds, some useful things occur

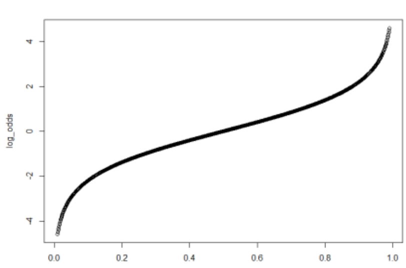
16

log(odds)

- there is no lower or upper bound
- p = .5 → log(odds) = 0
- p > .5 → log(odds) > 0
- p < .5 → log(odds) < 0
- and they're symmetric

17

log(odds)



18

the logit function

- the logit function converts probabilities to logits by taking their odds and finding the natural log

$$logit\ p = \ln \frac{p}{1-p}$$

- if we convert our outcome to logits and fit a regular linear model, we are doing logistic regression

19

the model and its summary

```
glm(Y ~ radius_mean.c, d, family = "binomial")

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.64406    0.13998  -4.601  4.2e-06 ***
radius_mean.c  1.03359    0.09311  11.101  < 2e-16 ***
```

- these parameter estimates are interpretable as usual
- however, they are in logits, which are not very intuitive

20

improving interpretability by exponentiation

- if we exponentiate (i.e., undo the logarithms) the parameter estimates, we can interpret them as odds

```
exp(coef(m))

(Intercept) radius_mean.c
   0.525156      2.811136
```

- the intercept is the predicted odds of a tumor being malignant at the mean of radius_mean
- the slope is the increase in odds for every one unit increase in the radius_mean
- this latter value is not additive! it's multiplicative!

21

multiplicative interpretation

- if the odds of a tumor being malignant at the mean of the predictor are predicted to be 0.525
- at one unit higher (than the mean), the odds are predicted to be

  0.525 × 2.81 = 1.48

- at another unit higher, the odds are predicted to be

  1.48 × 2.81 = 4.15

- at one unit lower than the mean, the odds are predicted to be

  0.524 / 2.81 = 0.187

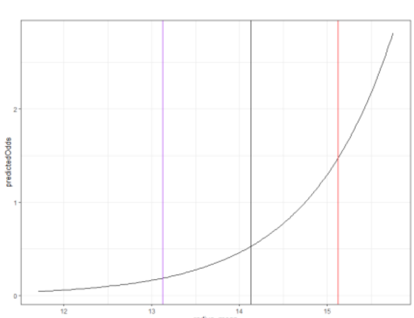
22

interpreting (predicted) odds

- odds = 0.525 → 0.525 less likely to be malignant than benign
- odds = 1.48 → 1.48x more likely to be malignant than benign

23

predicted odds, graphed



24