Logistic regression

Regression so far ...

At this point we have covered:

Regression so far ...

At this point we have covered:

Simple linear regression

- Relationship between numerical response and a numerical or categorical predictor

Regression so far ...

At this point we have covered:

Simple linear regression

- Relationship between numerical response and a numerical or categorical predictor

Multiple regression

- Relationship between numerical response and multiple numerical and/or categorical predictors

Regression so far ...

At this point we have covered:

Simple linear regression

- Relationship between numerical response and a numerical or categorical predictor

Multiple regression

- Relationship between numerical response and multiple numerical and/or categorical predictors

What we haven’t seen is what to do when we want to predict a categorical response (e.g. “gets cancer”) from numerical predictors (“height”, “weight”, ...)

Odds

Odds are another way of quantifying the probability of an event, commonly used in gambling (and logistic regression).

For some event E (“gets cancer”), odds represent ratio of probability of event occuring to probability of event NOT occurring:

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

Similarly, if we are told the odds of E are x to y then

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

which implies

$$P(E) = x/(x+y), \quad P(E^c) = y/(x+y)$$

Example - Donner Party

In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested rote to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

From *Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*

Example - Donner Party - Data

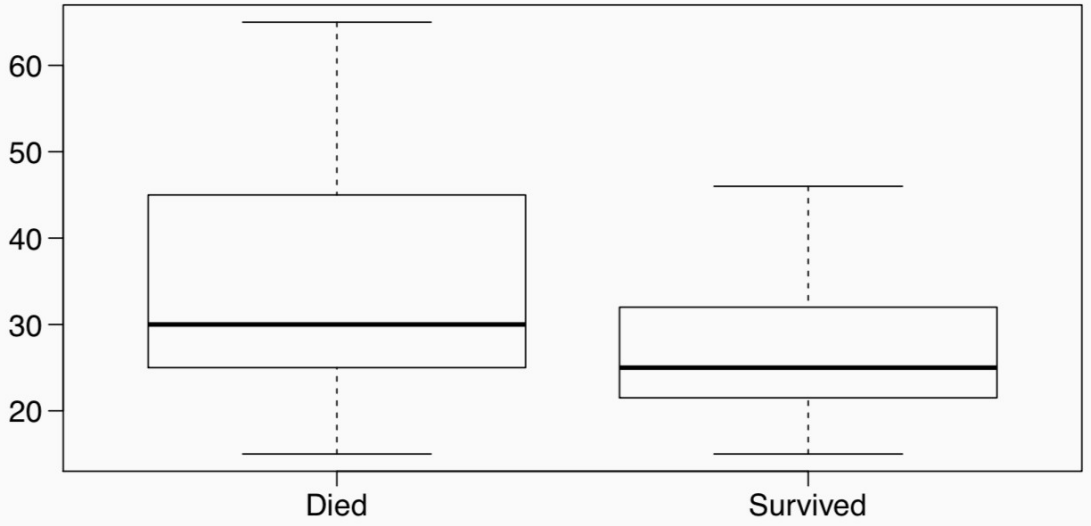
	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Example - Donner Party - EDA

Status vs Gender

	Male	Female
Died	20	5
Survived	10	10

Status vs Age



Example - Donner Party

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, standard linear model approach will predict a number – could be -1000, could be 1,000! How to deal with this?

One solution: treat Survived and Died as successes and failures arising from a binomial distribution where the probability of a success is given by a transformation of a linear model of the predictors.
A mouthful, but we will see in coming slides what this means.

Inverse functions: a refresher

Inverse functions allow for “undoing” equations.

$f(x) = y$ has inverse function $f^{-1}(y)$ if $f(f^{-1}(y)) = y$ for all y, and $f^{-1}(f(x)) = x$ for all x.

For instance, inverse function of $f(x) = x^2$ is $f^{-1}(y) = \text{sqrt}(y)$
Inverse function of $f(x) = \log(x)$ is $f^{-1}(y) = \exp(y)$

If I tell you that $\log(p / (1-p)) = z$, you should be able to solve for p in terms of z. This will give you the inverse function of $\log(p / (1-p))$.

Generalized linear models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

Generalized linear models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable
2. A linear model: $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$.
3. A link function that relates the linear model to the parameter of the outcome distribution: $g(p) = \eta$ or $p = g^{-1}(\eta)$.

   In GLMs, we use a linear model to predict g(p), rather than try to predict p.

Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume the outcome variable is a “success” with probability p, and we model p the probability of success for a given set of predictors.

Logistic Regression

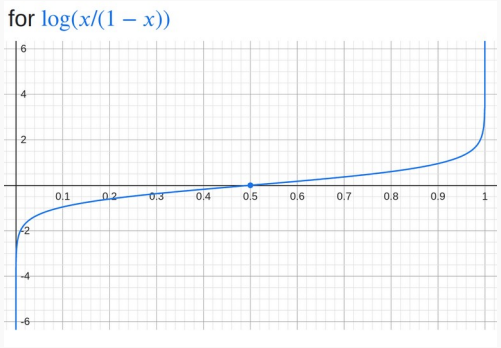
Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume the outcome variable is a “success” with probability p, and we model p the probability of success for a given set of predictors.

To finish specifying the Logistic model we just need to establish a reasonable link function that connects η to p. There are a variety of options but the most commonly used is the logit function.

Logit function: think of “probability p”.

$$logit(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \leq p \leq 1$$

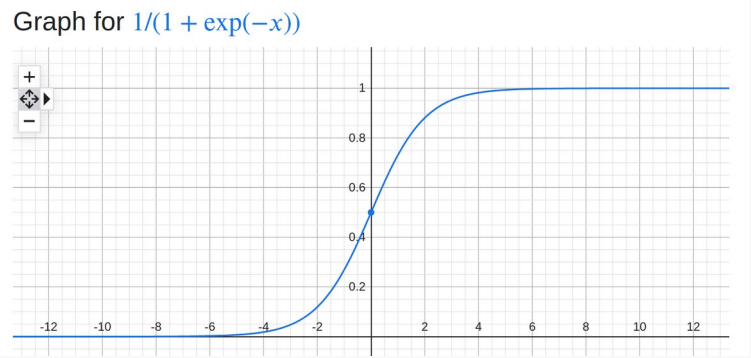


Properties of the Logit

The logit function takes a value between 0 and 1 and maps it to a value between −∞ and ∞.

Inverse logit (logistic) function

$$g^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$



The inverse logit function takes a value between −∞ and ∞ and maps it to a value between 0 and 1.

This formulation also has some use when it comes to interpreting the model as logit can be interpreted as the log odds of a success, more on this later.

The logistic regression model

The three GLM criteria give us:

$$y_i \sim \text{Binom}(p_i)$$

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

$$\text{logit}(p) = \eta$$

Binom(p_i): “probability of success = p_i”. From which we arrive at,

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}$$

Example - Donner Party - Model

In **R** we fit a GLM in the same was as a linear model except using **glm** instead of **lm** and we must also specify the type of GLM to fit using the **family** argument.

```
summary(glm(Status ~ Age, data=donner, family=binomial))
## Call:
## glm(formula = Status ~ Age, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.81852    0.99937   1.820   0.0688 .
## Age         -0.06647    0.03222  -2.063   0.0391 *
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 56.291  on 43  degrees of freedom
## AIC: 60.291
```

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

$$\frac{p}{1-p} = \exp(1.8185) = 6.16$$

$$p = 6.16/7.16 = 0.86$$

Example - Donner Party - Prediction

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

Odds / Probability of survival for a 50 year old:

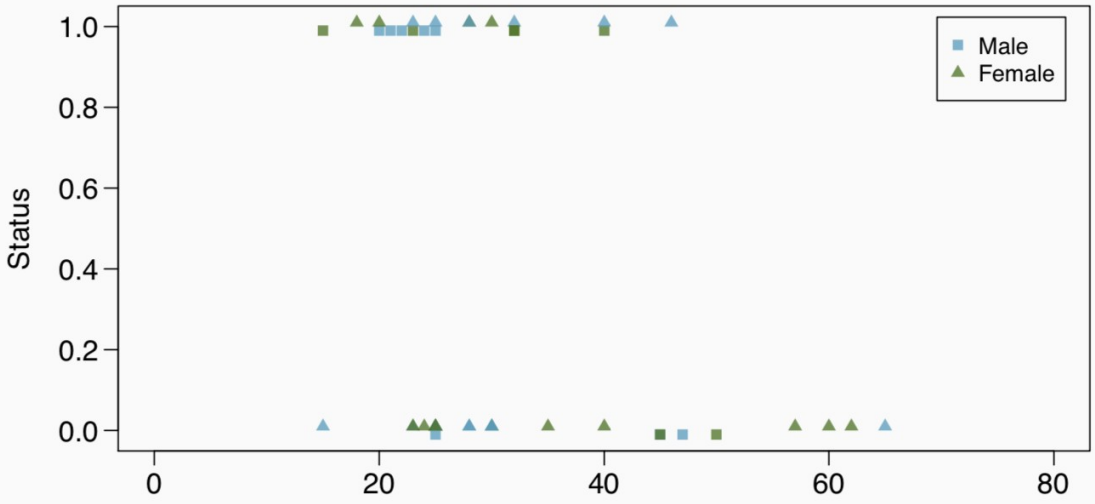
$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 50$$

$$\frac{p}{1-p} = \exp(-1.5065) = 0.222$$

$$p = 0.222/1.222 = 0.181$$

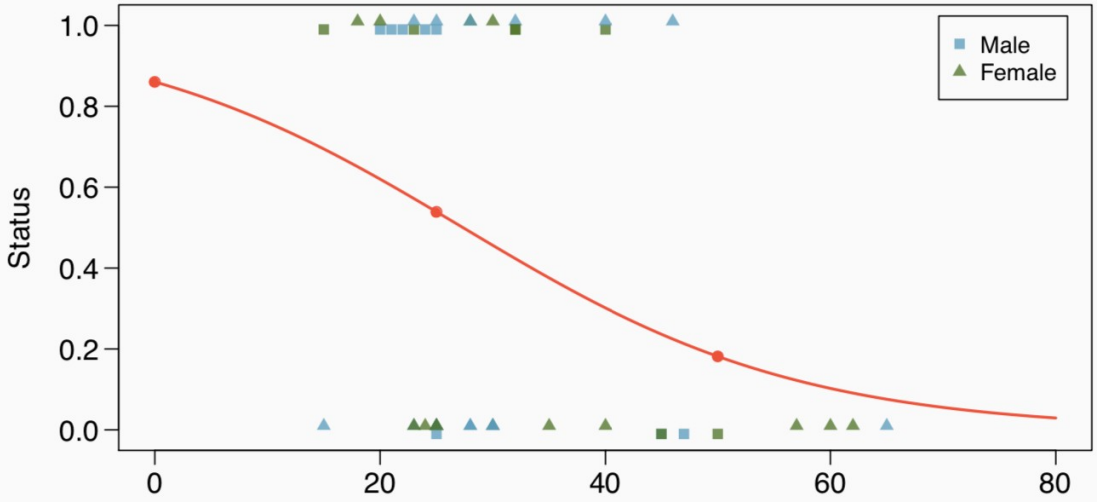
Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Age and Gender

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))
## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Example - Donner Party - Gender Models

Just like MLR we can plug in gender to arrive at two status vs age models for men and women respectively.

General model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \text{Sex}$$

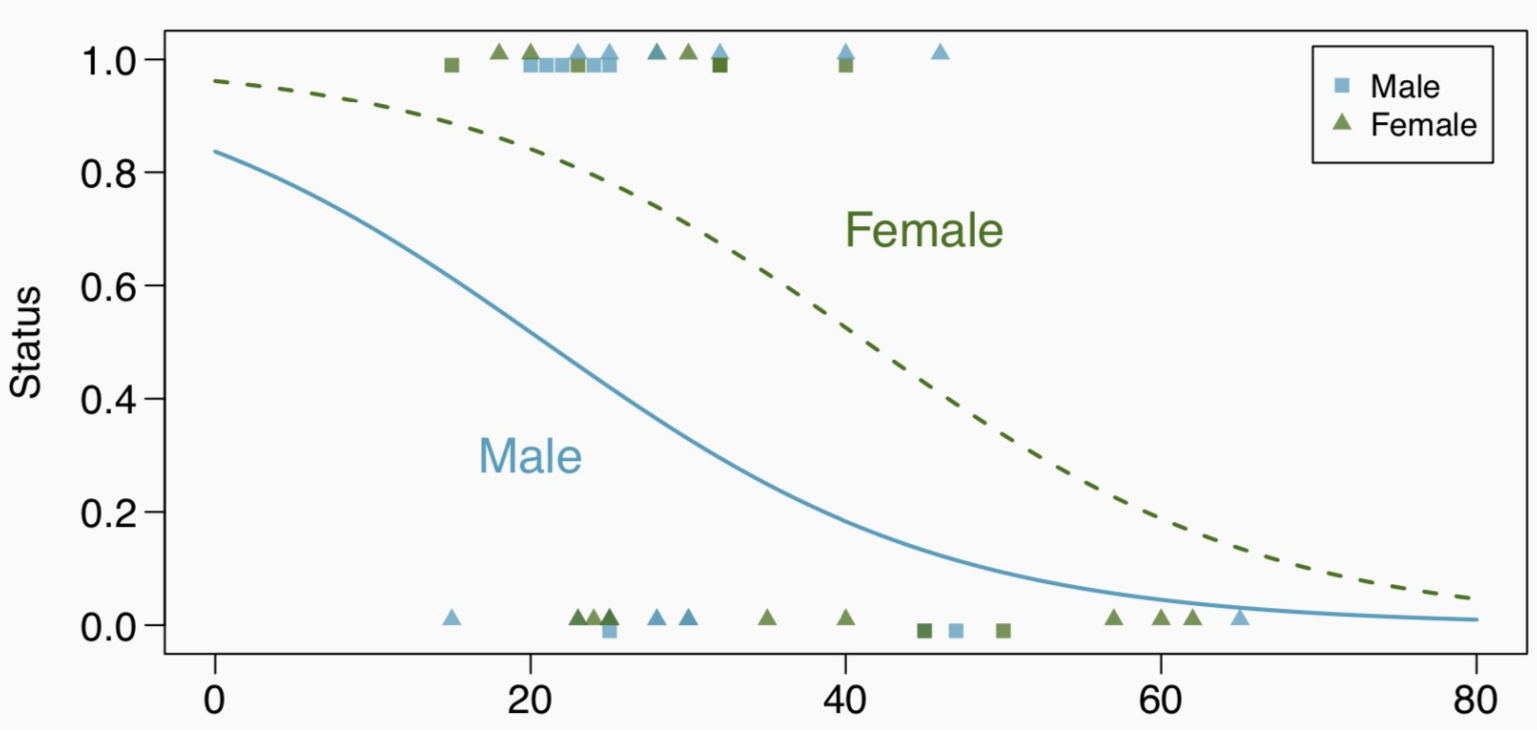
Male model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times 0$$
$$= 1.63312 + -0.07820 \times \text{Age}$$

Female model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times 1$$
$$= 3.23041 + -0.07820 \times \text{Age}$$

Example - Donner Party - Gender Models (cont.)



Practice

* Let’s suppose we have a logistic regression model for predicting the categorical “Goes to grad school” using GPA as a predictor variable.

(a) Describe the equation for this logistic regression model. What is the response? What is the predictor?

Practice

* Let’s suppose we have a logistic regression model for predicting the categorical “Goes to grad school” using GPA as a predictor variable.

(a) Describe the equation for this logistic regression model. What is the response? What is the predictor?

Practice

* Let’s suppose we have a logistic regression model for predicting the categorical “Goes to grad school” using GPA as a predictor variable.

(a) Describe the equation for this logistic regression model. What is the response? What is the predictor?

$\log(p / (1-p)) = a + b*GPA$

Response is log-odds ratio; predictor is GPA.

(b) Suppose we had a=-4 and b=1. Suppose the GPA is 3.5.
What is the estimated probability of going to grad school?

Practice

* Let’s suppose we have a logistic regression model for predicting the categorical “Goes to grad school” using GPA as a predictor variable.

(a) Describe the equation for this logistic regression model. What is the response? What is the predictor?

log(p / (1-p)) = a + b*GPA

Response is log-odds ratio; predictor is GPA.

(b) Suppose we had a=-4 and b=1. Suppose the GPA is 3.5.
What is the estimated probability of going to grad school?

(c)What happens if we increase GPA by 0.5 to 4.0. Does the probability increase by 0.5?

Practice

1) True/False (+explain).
If we have a logistic regression model, the log-odds can be any number,
Not just between 0 and 1.

2) True/False (+explain)
If we have a logistic regression model, the estimated probability can be any number,
Not just between 0 and 1.