Statistical Inference

Logistic Regression

Behnam Bahrak
Spring 2020

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 the predictors are weird (nonlinear, complicated dependence structure, etc.) or when the response is weird (categorical, count data, etc.)

Odds

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

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

- 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:

$$P(E) = \frac{x}{x+y} \quad , \qquad P(E^c) = \frac{y}{x+y}$$

Example - Donner Party

- In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon.
- In July, the Donner Party reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested route 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.
- 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.

Example - Donner Party - Data

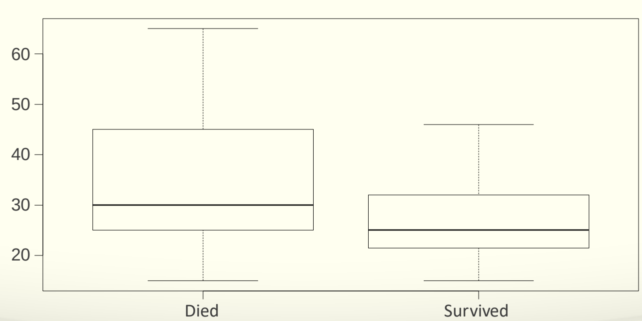
	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Example - Donner Party

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, this isn't something we can transform our way out of - we need something more.
- One way to think about the problem - we can 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.

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
  $$g(p) = \eta \ , \qquad p = g^{-1}(\eta)$$

Logistic Regression

- Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.
- We assume a Bernoulli distribution produced the outcome variable and we therefore want to 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:

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

Properties of the Logit

- The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and ∞.

Inverse logit (logistic) function

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

- The inverse logit function takes a value between $-\infty$ 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.

The logistic regression model

- The three GLM criteria give us:

$$y_i \sim Bernoulli(p_i)$$

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

$$logit(p) = \eta$$

- From which we arrive at:

$$p_i = \frac{e^{\beta_0 + \beta_1 X_{1,i} + \cdots + \beta_n X_{n,i}}}{1 + e^{\beta_0 + \beta_1 X_{1,i} + \cdots + \beta_n X_{n,i}}}$$

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.

```
R
> summary(glm(Status ~ Age, data=donner, family=binomial))
```

	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

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 $= 0$):

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0 \;\Rightarrow\; \frac{p}{1-p} = e^{1.8185} = 6.16 \;\Rightarrow\; p = 0.86$$

Example - Donner Party - Prediction

Model:

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

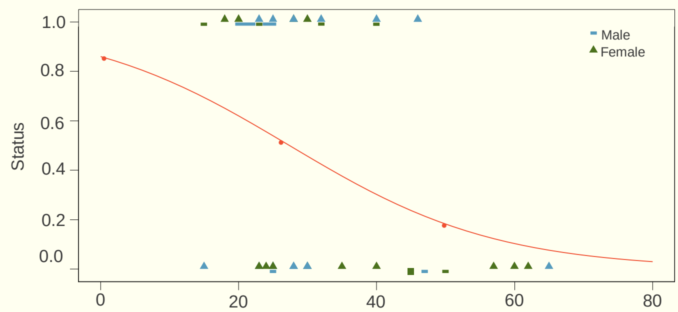
Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25 \;\Rightarrow\; \frac{p}{1-p} = e^{0.156} = 1.17 \;\Rightarrow\; p = 0.539$$

Odds / Probability of survival for a 50 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 50 \;\Rightarrow\; \frac{p}{1-p} = e^{-1.5065} = 0.222 \;\Rightarrow\; p = 0.181$$

Example - Donner Party - Prediction



Example - Donner Party - Interpretation

	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

- Simple interpretation is only possible in terms of log odds and log odds ratios for intercept and slope terms.
- Intercept: The log odds of survival for a party member with an age of 0. From this we can calculate the odds or probability, but additional calculations are necessary.
- Slope: For a unit increase in age (being 1 year older) how much will the log odds ratio change, not particularly intuitive. More often then not we care only about sign and relative magnitude.

Example - Interpretation of Slope

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.8185 - 0.0665(x+1)$$

$$\log\left(\frac{p_2}{1-p_2}\right) = 1.8185 - 0.0665x$$

$$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = -0.0665$$

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = e^{-0.0665} = 0.94$$

Example: Donner Party - Age and Gender

R

```
> summary(glm(Status ~ Age + Sex, family = binomial, data = donner))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
Sex:Female	1.5973	0.7555	2.11	0.0345

Gender slope: When the other predictors are held constant this is the log odds ratio between the given level (Female) and the reference level (Male).

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