

Unit 5: The General Linear Model

4. The Generalized Linear Model (Chapter 6.4)

4/13/2020

Recap from Last Time

1. When your diagnostics show problems with constant variance, transformations can help to normalize data
2. The log transformation is a common solution to right-skewed data
3. Transformed models are often hard to interpret in linear units, reversing the transformation can help

Key ideas

1. When outcome variables are categorical and binary, we can use logistic regression
2. Logistic regression is one of a family of models called Generalized Linear Models that often apply when the assumptions of linear regression fail
3. This is an extremely general statistical method that you'll be able to use in almost all of the cases you're likely to work with

Regression so far...

So far, we have covered:

- Simple Regression
 - Relationship between numerical response and a numerical or categorical predictor
- Multiple Regression (General Linear Model)
 - 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 ,

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

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 $P(E) = \frac{x}{(x+y)}$ and $P(\sim E) = \frac{y}{(x+y)}$

Example: The 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 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 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.

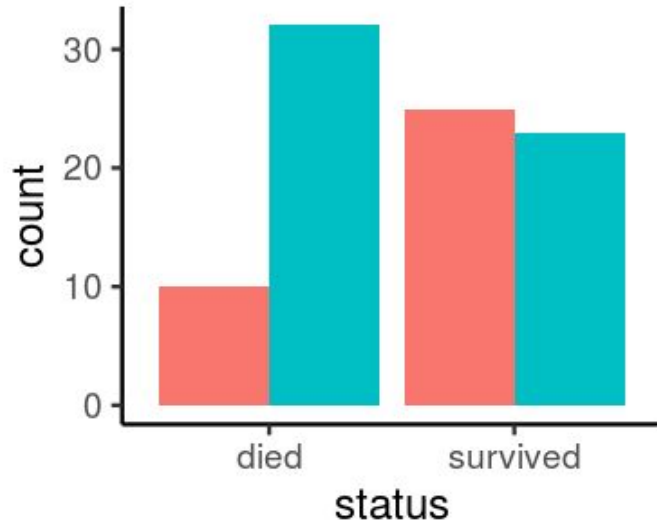
By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

Example: The Donner party - data

family <chr>	age <dbl>	gender <chr>	status <chr>
Other	23	Male	died
Breen	13	Male	survived
Breen	1	Female	survived
Breen	5	Male	survived
Breen	14	Male	survived
Breen	40	Female	survived
Breen	51	Male	survived
Breen	9	Male	survived
Breen	3	Male	survived
Breen	8	Male	survived

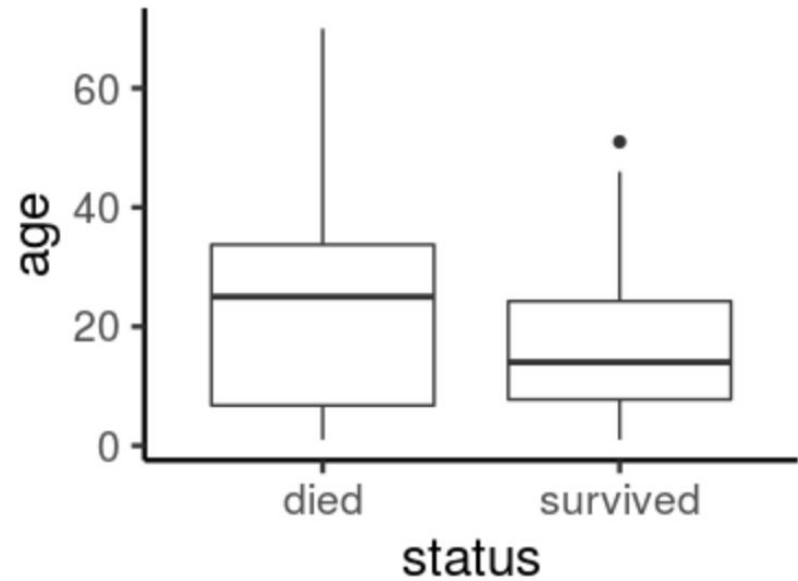
The Donner party - Exploratory Data Analysis

gender



gender Female Male

age



Example: The 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 heads and tails arising from flipping a coin, and try to estimate the weight of the coin using 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: $\eta = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$.
3. A link function that relates the linear model to the parameter of the outcome distribution: $g(\mu) = \eta$ or $\mu = g^{-1}(\eta)$.

The linear regression model


The regression models you've already seen are a special case of the Generalized Linear Model (GLM)

1. A probability distribution describing the outcome:

$$y_i = \text{Normal}(p_i, \sigma^2)$$


constant variance

2. A linear model:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$


linearity

3. A link function that relates the linear model to the parameter of the outcome distribution

$$p = \eta$$

Normal residuals

Generalized linear models

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

We assume a binomial distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors (i.e. whether a coin comes up heads).

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

Properties of the logit function

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

Logit is the inverse of the logistic function

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

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

The logistic regression model

Logistic regression is another instantiation of the General Linear Model

1. A probability distribution describing the outcome:

$$y_i = \text{Binomial}(p_i)$$

2. A linear model:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

3. A link function that relates the linear model to the parameter of the outcome distribution

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

Modeling the Donner party

In **R** we fit a GLM in the same way 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.

Call:

```
glm(formula = status ~ age, family = "binomial", data = donner_data)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.86843	0.37159	2.337	0.0194	*
age	-0.03533	0.01467	-2.408	0.0161	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Prediction from a logistic regression

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.86843	0.37159	2.337	0.0194	*
age	-0.03533	0.01467	-2.408	0.0161	*

Model: $\log\left(\frac{p}{1-p}\right) = .86843 - .03533 \cdot \text{age}$

Probability of survival for a newborn (age=0):

$$\log\left(\frac{p}{1-p}\right) = .86843 - .03533 \cdot 0$$

$$\frac{p}{1-p} = e^{.86843} = 2.38317$$

$$p = \frac{2.38317}{2.38317 + 1} = .704$$

Prediction from a logistic regression

$$\text{Model: } \log\left(\frac{p}{1-p}\right) = .86843 - .03533 \cdot \textit{age}$$

Probability of survival for a 25-year-old :

$$\log\left(\frac{p}{1-p}\right) = .86843 - .03533 \cdot 25$$

$$\frac{p}{1-p} = e^{-0.01482} = 0.98528$$

$$p = \frac{0.98528}{0.98528 + 1} = .496$$

Prediction from a logistic regression

$$\text{Model: } \log\left(\frac{p}{1-p}\right) = .86843 - .03533 \cdot \textit{age}$$

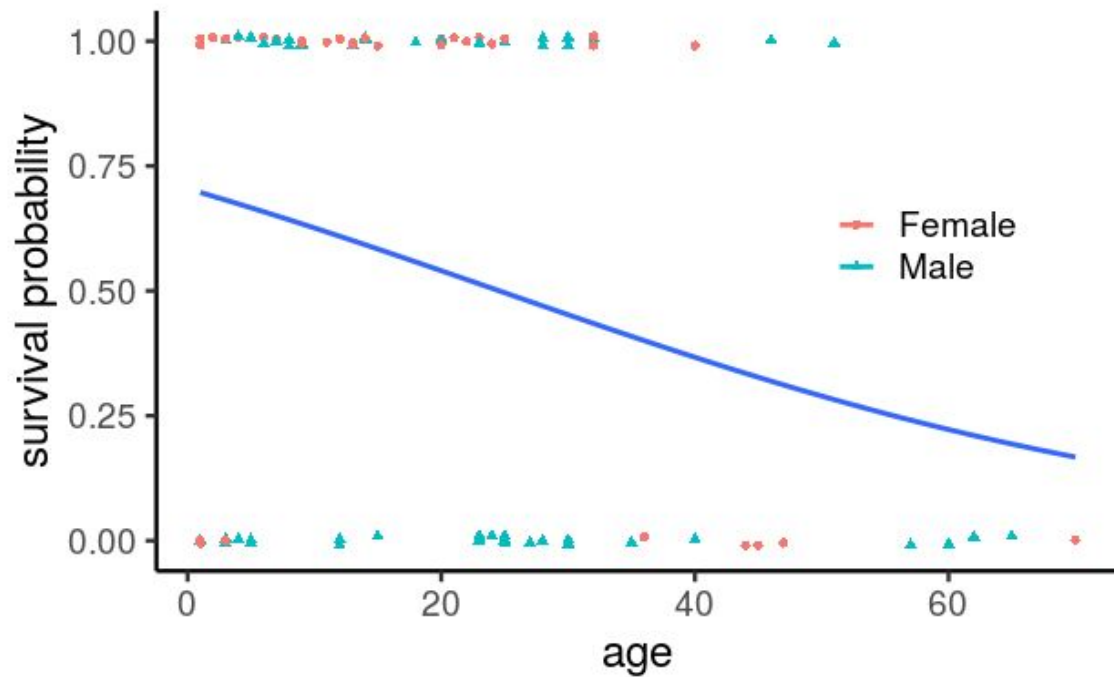
Probability of survival for a 50-year-old :

$$\log\left(\frac{p}{1-p}\right) = .86843 - .03533 \cdot 50$$

$$\frac{p}{1-p} = e^{-0.89807} = 0.40735$$

$$p = \frac{0.98528}{0.98528 + 1} = 0.289$$

Plotting the model



Interpreting the model

Remember, we interpreted log-transformed linear regression models by thinking about the slope as a ratio

$$\begin{aligned}\log(\text{price at year } x + 1) - \log(\text{price at year } x) &= 0.137 \\ \frac{\text{price at year } x + 1}{\text{price at year } x} &= 1.15\end{aligned}$$

For each additional year the car is newer we would expect the price of the car to increase on average by a **factor of 1.15**.

We can do the same thing here.

The slope in a logistic regression model is a ratio too--a log odds ratio.

Interpreting the model

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.86843	0.37159	2.337	0.0194	*
age	-0.03533	0.01467	-2.408	0.0161	*

For each additional year of age,
we expect the log odds of survival to decrease by .03533

$$\log\left(\frac{p_{survive}(\text{age } x + 1)}{p_{survive}(\text{age } x)}\right) = -.03533$$

$$\frac{p_{survive}(\text{age } x + 1)}{p_{survive}(\text{age } x)} = e^{-.03533} = .965$$

Modeling both age and gender

Call:

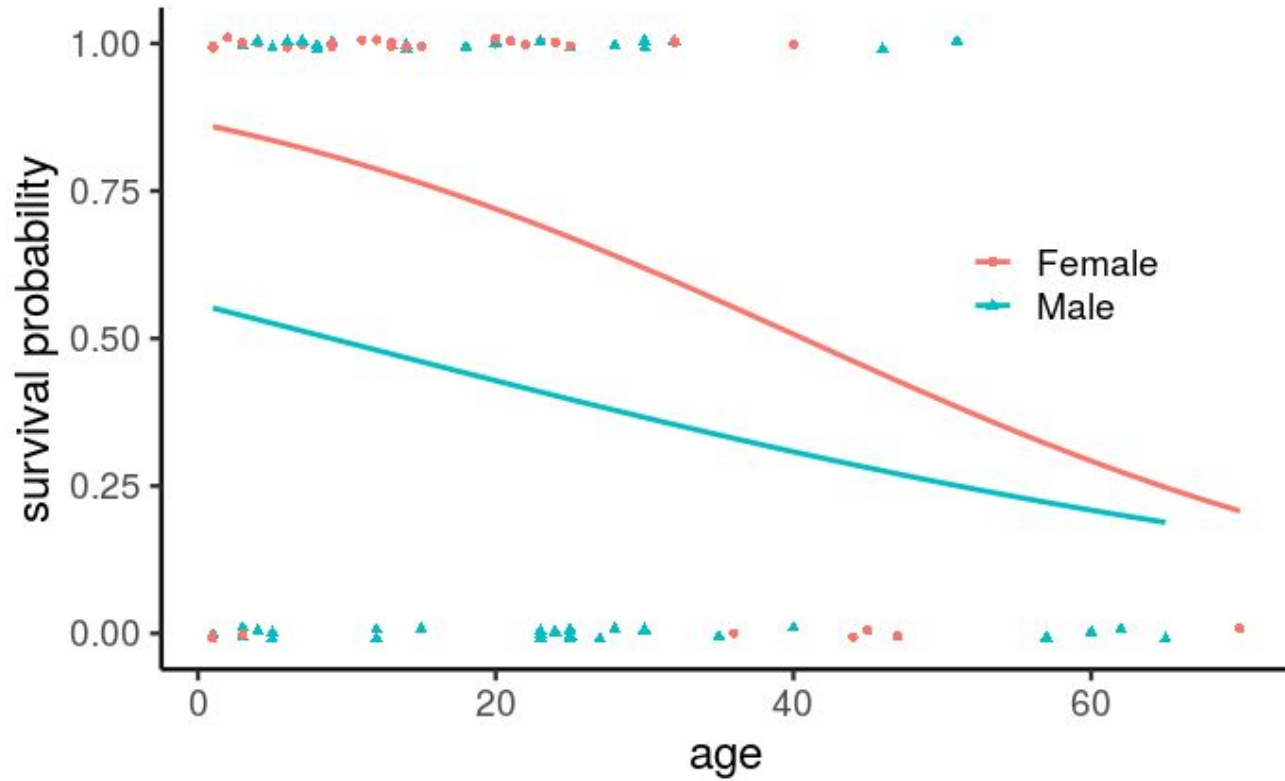
```
glm(formula = status ~ age + gender, family = "binomial",  
     data = donner_data)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	1.5992	0.5041	3.172	0.00151	**
age	-0.0338	0.0151	-2.238	0.02525	*
genderMale	-1.2068	0.4790	-2.519	0.01176	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Plotting the model



No hypothesis test for the whole model

Call:

```
glm(formula = status ~ age + gender, family = "binomial",  
     data = donner_data)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	1.5992	0.5041	3.172	0.00151	**
age	-0.0338	0.0151	-2.238	0.02525	*
genderMale	-1.2068	0.4790	-2.519	0.01176	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 124.37 on 89 degrees of freedom

Residual deviance: 111.13 on 87 degrees of freedom

AIC: 117.13

 **Akaike Information Criterion**

Hypothesis tests for coefficients

Call:

```
glm(formula = status ~ age + gender, family = "binomial",  
     data = donner_data)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	1.5992	0.5041	3.172	0.00151	**
age	-0.0338	0.0151	-2.238	0.02525	*
genderMale	-1.2068	0.4790	-2.519	0.01176	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

We can still perform inference on individual coefficients, the basic setup is exactly the same as what we've seen before except we use a Z-test.

Note: The only tricky bit, which is way beyond the scope of this course, is how the standard error is calculated.

Key ideas

1. When outcome variables are categorical and binary, we can use logistic regression
2. Logistic regression is one of a family of models called Generalized Linear Models that often apply when the assumptions of linear regression fail
3. This is an extremely general statistical method that you'll be able to use in almost all of the cases you're likely to work with