Logistic Regression

November 29, 2022

Objectives

- Understand Logistic Regression as a Generalized Linear Model;
- Develop intuition for ideas behind Maximum Likelihood Estimation;
- Build ability to interpret Logistic Regression results;

Binary data

We have learned about and worked with linear regression:

- model parameter interpretation;
- hypothesis tests and p-values;
- model fit, e.g. residuals.

This is great when the dependent variable is **continuous** (i.e. interval or ratio data).

Binary data

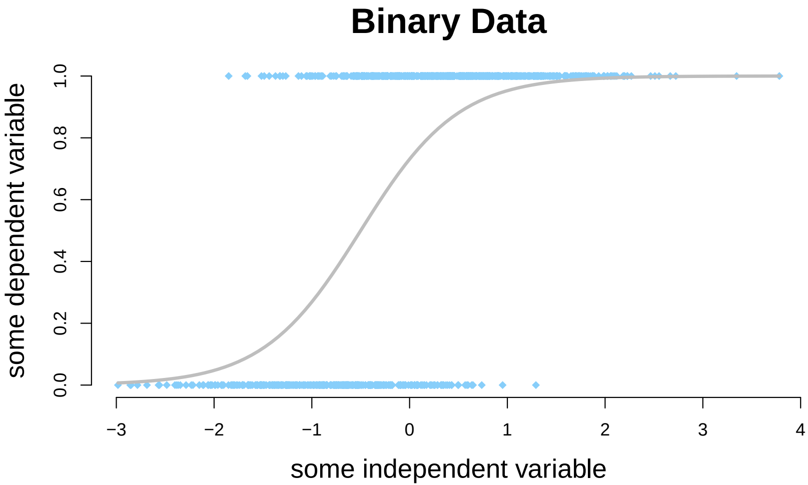
We have learned about and worked with linear regression:

- model parameter interpretation;
- hypothesis tests and p-values;
- model fit, e.g. residuals.

This is great when the dependent variable is **continuous** (i.e. interval or ratio data).

But what if the dependent variable is **weird**, e.g. binary?

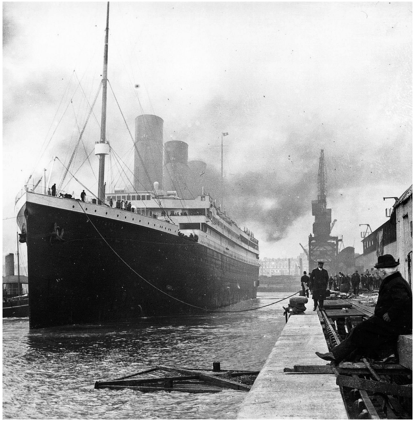
Binary data


Binary Data
some dependent variable
0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3
4
some independent variable

A specific example...

After colliding with an iceberg late on 14 April 1912 RMS *Titanic* sank with huge loss of life:

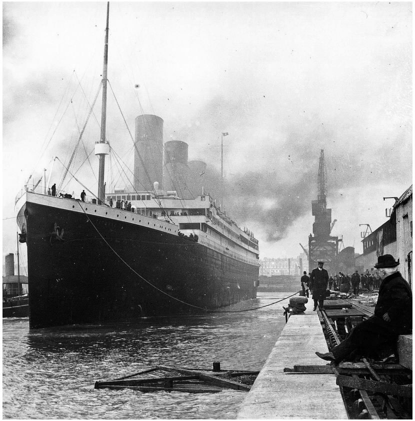
- Estimated that 2,224 passengers and crew were aboard;
- Carried 20 lifeboats with sufficient capacity for 1,178;



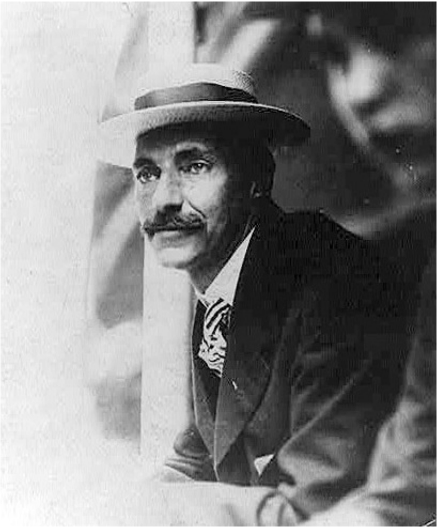
A specific example...

After colliding with an iceberg late on 14 April 1912 RMS *Titanic* sank with huge loss of life:

- Estimated that 2,224 passengers and crew were aboard;
- Carried 20 lifeboats with sufficient capacity for 1,178;
- Even with perfect efficiency 1,046 people were doomed.



A specific example... who lived and who died?



A specific example... who lived and who died?

Survived	Pclass	Name	Sex	Age	Siblings Spouses	Parents Children	Fare
0	3	Owen Harris Braund	M	22	1	0	7.25
1	1	Mrs. John Cumings	F	38	1	0	71.28
1	3	Miss. Laina Heikkinen	F	26	0	0	7.92
1	1	Mrs. Jacques Futrelle	F	35	1	0	53.10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Logistic Regression

Logistic Regression (logit) is a Generalized Linear Model (GLM) used to model a binary dependent variable using numerical and categorical predictors.

Logistic Regression

Logistic Regression (logit) is a Generalized Linear Model (GLM) used to model a binary dependent variable using numerical and categorical predictors.

Recall that GLMs have three required components:

1. A probability distribution that describes the dependent variable;
2. A linear model $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$;
3. A link function that relates the linear model to the dependent variable distribution.

A probability distribution that describes the dependent variable

- For logit assume that for a given set of predictor variables (say x):

$$y = 1 \text{ w/ probability } p(1)$$
$$y = 0 \text{ w/ probability } 1 - p(1);$$

- We want to model $p(1)$ and figure out how it depends on x.

A linear model $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$

- The x's are independent variables that come from the data and are the input to logit;
- The β's are the effects of the independent variables and are the output of logit;
- In the Titanic example, the x's could be any of Pclass, Sex, Age, Siblings, Parents, and Fare.

A link function that relates the linear model to the dependent variable distribution

This is where the rubber meets the road...

Let's talk a little about **odds**:

A link function that relates the linear model to the dependent variable distribution

This is where the rubber meets the road...

Let's talk a little about **odds**:

- In the Titanic context we can ask **what are the odds that a passenger lives?**

$$Odds(\text{Lived}) = \frac{p(\text{Lived})}{p(\text{not Lived})} = \frac{p(\text{Lived})}{1 - p(\text{Lived})}.$$

A link function that relates the linear model to the dependent variable distribution

This is where the rubber meets the road...

Let's talk a little about **odds**:

- In the Titanic context we can ask **what are the odds that a passenger lives?**

$$Odds(\text{Lived}) = \frac{p(\text{Lived})}{p(\text{not Lived})} = \frac{p(\text{Lived})}{1 - p(\text{Lived})}.$$

- The **Logistic Function** is defined as the inverse of:

$$\ln\left(Odds(Lived)\right) = \ln\left(\frac{p(\text{Lived})}{1 - p(\text{Lived})}\right).$$

It is used because:

A link function that relates the linear model to the dependent variable distribution

This is where the rubber meets the road...

Let's talk a little about **odds**:

- In the Titanic context we can ask **what are the odds that a passenger lives?**

$$\mathit{Odds}(\text{Lived}) = \frac{p(\text{Lived})}{p(\text{not Lived})} = \frac{p(\text{Lived})}{1 - p(\text{Lived})}.$$

- After some algebra for a number γ:

$$\text{Logistic}(\gamma) = \frac{\exp\{\gamma\}}{1 + \exp\{\gamma\}}.$$

It is used because:

A link function that relates the linear model to the dependent variable distribution

This is where the rubber meets the road...

Let's talk a little about **odds**:

- In the Titanic context we can ask **what are the odds that a passenger lives?**

$$Odds(\text{Lived}) = \frac{p(\text{Lived})}{p(\text{not Lived})} = \frac{p(\text{Lived})}{1 - p(\text{Lived})}.$$

- In logit we will use the **Logistic Function** as follows:

$$p(\text{Lived}) = \text{Logistic}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\}}$$

It is used because:

A link function that relates the linear model to the dependent variable distribution

This is where the rubber meets the road...

Let's talk a little about **odds**:

- In the Titanic context we can ask **what are the odds that a passenger lives?**

$$Odds(\text{Lived}) = \frac{p(\text{Lived})}{p(\text{not Lived})} = \frac{p(\text{Lived})}{1 - p(\text{Lived})}.$$

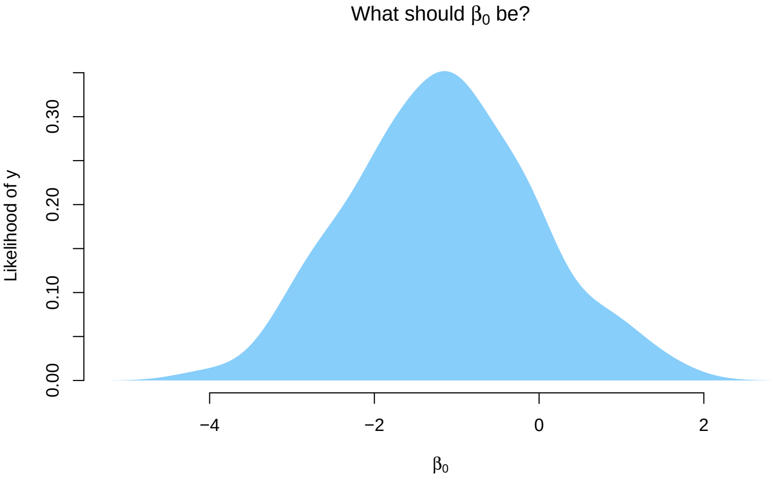
- In logit we will use the **Logistic Function** as follows:

$$p(\text{Lived}) = \text{Logistic}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\}}$$

It is used because:

- It takes the linear model and ensures that it will give us something between 0 and 1;
- It makes comparing the odds of two events really easy.

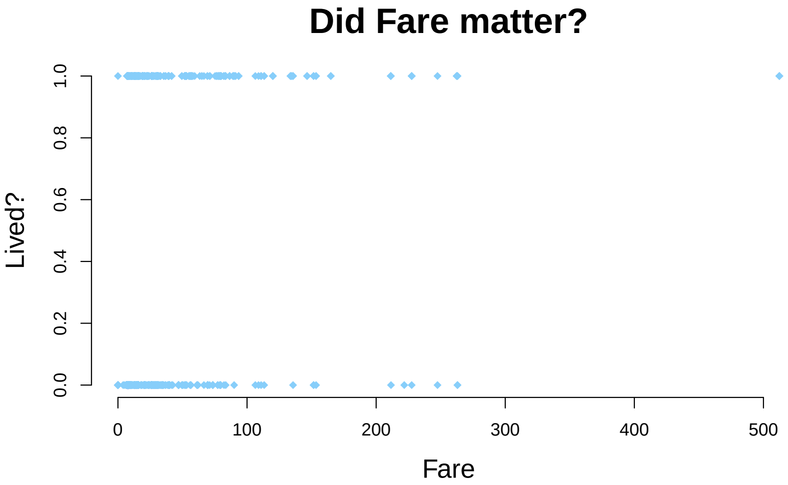
Maximum Likelihood Estimation



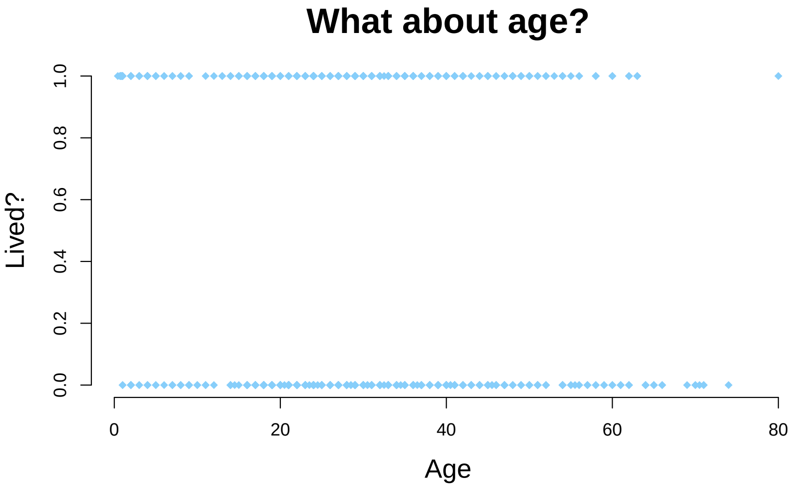
Let's return to the Titanic...

Survived	Pclass	Name	Sex	Age	Siblings Spouses	Parents Children	Fare
0	3	Owen Harris Braund	M	22	1	0	7.25
1	1	Mrs. John Cumings	F	38	1	0	71.28
1	3	Miss. Laina Heikkinen	F	26	0	0	7.92
1	1	Mrs. Jacques Futrelle	F	35	1	0	53.10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Let's return to the Titanic...



Let's return to the Titanic...



Let's return to the Titanic... interpret this!

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.2973	0.5574	9.50	0.0000
Pclass	-1.1777	0.1461	-8.06	0.0000
Age	-0.0435	0.0077	-5.63	0.0000
Male	-2.7573	0.2004	-13.76	0.0000
Siblings/Spouses	-0.4018	0.1107	-3.63	0.0003
Parents/Children	-0.1065	0.1186	-0.90	0.3691
Fare	0.0028	0.0024	1.17	0.2437

Let's return to the Titanic... interpret this!

So if we were doing linear regression, we would have:

$$
\begin{aligned}
y = &\beta_{intercept} \\
&+ \beta_{Pclass} Pclass \\
&+ \beta_{Age} Age \\
&+ \beta_{Male} Male \\
&+ \beta_{Siblings/Spouses} Siblings/Spouses \\
&+ \beta_{Parents/Children} Parents/Children \\
&+ \beta_{Fare} Fare;
\end{aligned}
$$

Let's return to the Titanic... interpret this!

So if we were doing linear regression, we would have:

$$\begin{aligned} y =& \beta_{intercept} \\ &+ \beta_{Pclass} Pclass \\ &+ \beta_{Age} Age \\ &+ \beta_{Male} Male \\ &+ \beta_{Siblings/Spouses} Siblings/Spouses \\ &+ \beta_{Parents/Children} Parents/Children \\ &+ \beta_{Fare} Fare; \end{aligned}$$

The effect of Age on y would be just β_{Age}.

Let's return to the Titanic... interpret this!

Unfortunately, we are doing logit:

$$p(y = \text{Lived}) = \frac{\exp\{\beta_{intercept} + \beta_{Pclass} Pclass + \beta_{Age} Age + \ldots\}}{1 + \exp\{\beta_{intercept} + \beta_{Pclass} Pclass + \beta_{Age} Age + \ldots\}};$$

Let's return to the Titanic... interpret this!

Unfortunately, we are doing logit:

$$p(y = \text{Lived}) = \frac{\exp\{\beta_{intercept} + \beta_{Pclass} Pclass + \beta_{Age} Age + \dots\}}{1 + \exp\{\beta_{intercept} + \beta_{Pclass} Pclass + \beta_{Age} Age + \dots\}};$$

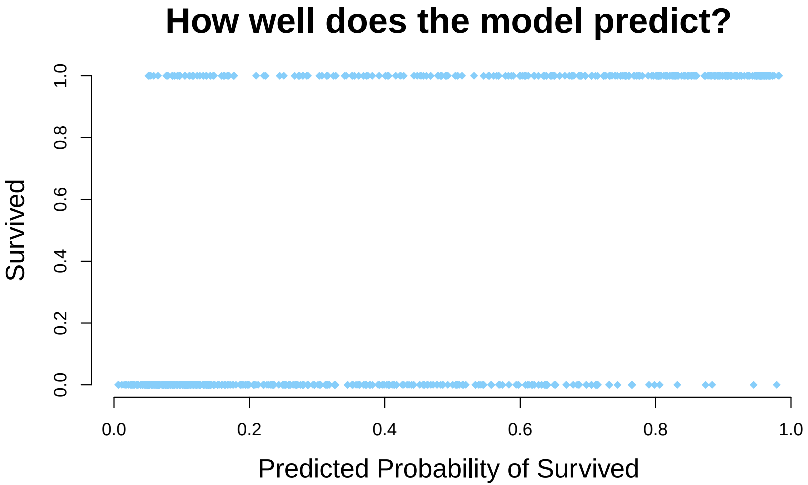
The **effect** of Age on y **depends on the values for all** the other independent vars!

If you want to demonstrate how a variable matters it has to be relative to a particular case:

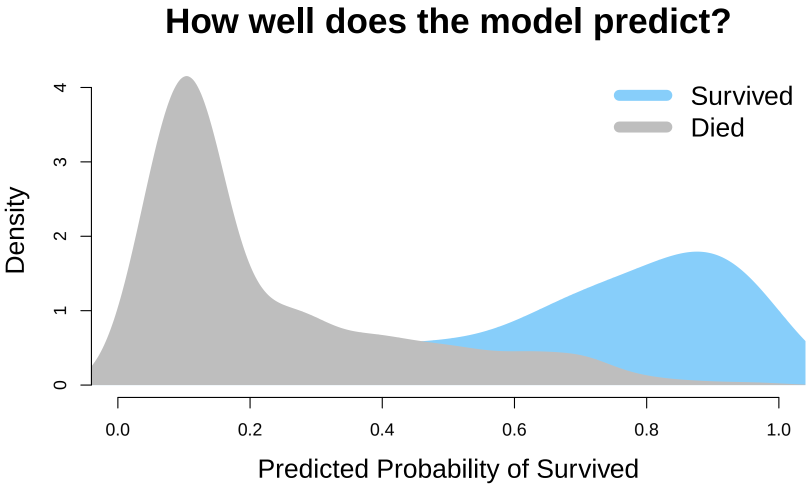
- the average or medians of the rest of the variables;
- a particular substantively interesting case;
- an observation in the dataset.

Let's return to the Titanic... interpret this!

Let's return to the Titanic... interpret this!



Let's return to the Titanic... interpret this!



Summary

Logit is a powerful and ubiquitous method for the relationship between binary dependent variables and numerical or categorical independent variables;

Logit is estimated via Maximum Likelihood Estimation;

Logit is a GLM - there are tons of GLMs built for different tasks.