

Simple logistic regression

The logistic regression model is for outcomes Y_i that are Bernoulli (“zero-one”) or binomial with covariate x_i . The probability of success changes smoothly with the covariate x_i , much like the mean changes smoothly in the linear regression model.

Example: From our PubH 7401 census data, we recorded whether or not someone had gone dancing or not and the number of years of post high-school education. The raw data are read into R as follows:

```
e=c(10, 6,10, 6, 8, 8, 9, 5, 7, 6, 7, 4, 8, 7, 7, 7, 8, 7,13,10,10)
d=c( 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0)
plot(e,d,xlab="Educations (years)",ylab="Dancing (0=no, 1=yes)")
```

A plot of the raw data is of limited use. The outcome is zero/one and there are several overlapping data points.

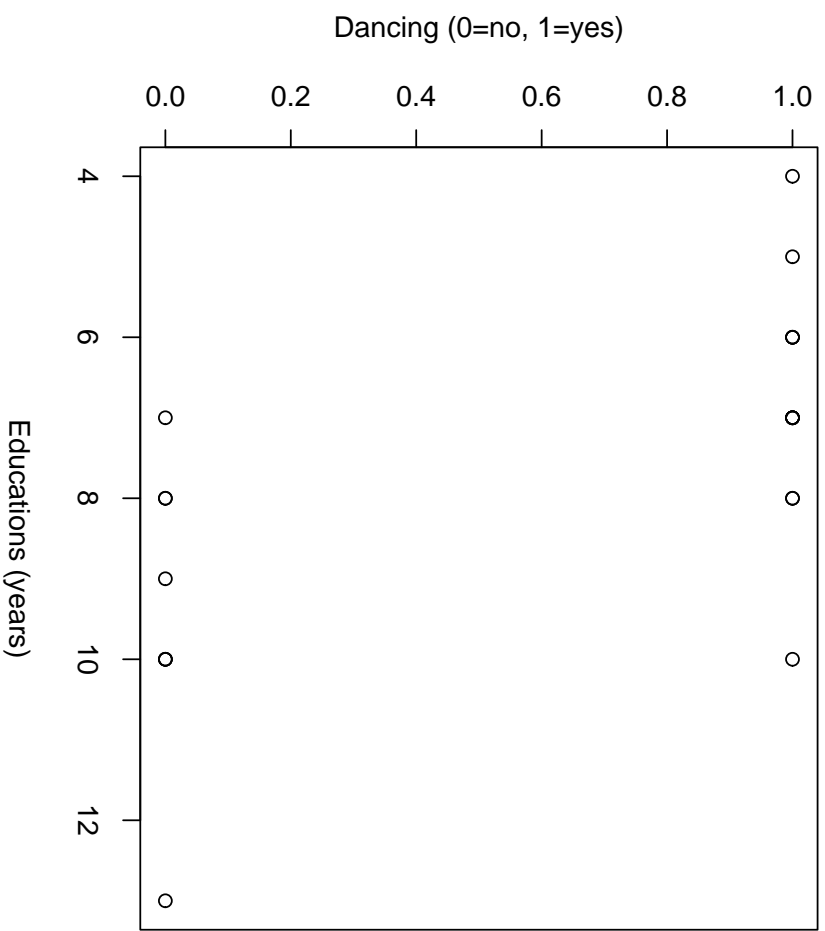


Figure 1: Raw data scatterplot.

We can instead aggregate responses into bins and obtain a plot with a bit more information.

```
x=c(1,2,3,4,5); y=c(2,8,2,1,0); n=c(2,9,5,4,1)
plot(x,y/n,ylab="Proportion (n=21)",axes=FALSE, xlab="Education")
axis(side=1,at=x,labels=c("4-5 yrs","6-7 yrs","8-9 yrs","10-11 yrs","12-13 yrs"),tick=FALSE)
title("Probability of Dancing versus Years Education")
axis(2); box()
```

The above R code defines 5 education categories and computes the sample proportion of those that have gone dancing for each category.

This helps us tease out whether there's a real trend in the probability of dancing with increasing education level.

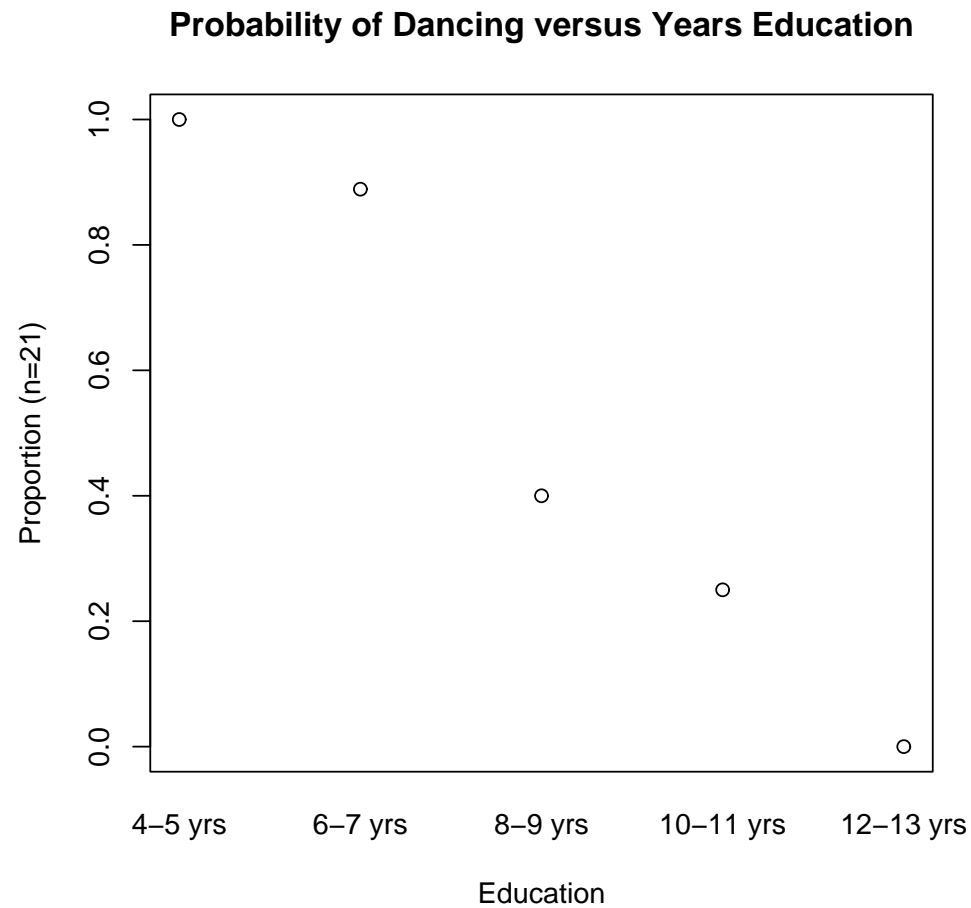


Figure 2: Raw sample proportions versus education level.

The observed proportions have an a somewhat linear trend plotted against education. We could fit a linear regression model to the observed proportions, or to the raw zero/one outcome, but this would allow for dancing probabilities outside the range zero to one. Furthermore, the data are clearly not normal, so modeling assumptions would be invalidated.

A common alternative approach to modeling probabilities of Bernoulli outcomes is to use a non-linear model for the proportions.

One non-linear model gives *logistic regression*.

The simple logistic regression model

The simple logistic regression model expresses the population proportion $\pi(x)$ of individuals with a given attribute (called a success) as a function of a single predictor variable x . The model assumes that π is related to x through

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x,$$

or, equivalently, as

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}.$$

The logistic regression model is a binomial regression model, where the response Y_i for each individual (i.e. a student enrolled in PubH 7401) falls into one of two exclusive and exhaustive categories, often called success (cases with the attribute of interest) and failure (cases without the attribute of interest).

In many biostatistical applications, the success category is presence of a disease, or death from a disease.

We write π as $\pi(x)$ to emphasize that π is the proportion of all individuals with score x that have the attribute of interest. In the dancing data, $\pi = \pi(x)$ is the population proportion of students with education x that have gone dancing within the last year.

Odds of success

The odds of success are $\pi/(1 - \pi)$. For example, the odds of success are 1 (or 1 to 1) when $\pi = 1/2$. The odds of success are 9 (or 9 to 1) when $\pi = 0.9$. The logistic model assumes that the log-odds of success is linearly related to x :

$$O(x) = \frac{\pi(x)}{1 - \pi(x)} = e^{\beta_0 + \beta_1 x}.$$

The logistic regression model is parameterized to easily make inferences about the odds of success.

Let's look at how the odds of success changes when we increase x by one unit:

$$\begin{aligned}\frac{O(x+1)}{O(x)} &= \frac{\pi(x+1)/[1-\pi(x+1)]}{\pi(x)/[1-\pi(x)]} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1x}} \\ &= \frac{e^{\beta_0+\beta_1x} e^{\beta_1}}{e^{\beta_0+\beta_1x}} \\ &= e^{\beta_1}\end{aligned}$$

When we increase x by one unit, the odds of an event occurring increases by a factor of e^{β_1} , *regardless of the value of x .*

e^{β_1} is an odds ratio.

Data for Simple Logistic Regression

For the formulas below, the data is given in summarized or **aggregate** form:

x	n	Y
x_1	n_1	Y_1
x_2	n_2	Y_2
\cdot	\cdot	\cdot
\cdot	\cdot	\cdot
x_m	n_m	Y_m

where Y_i is the number of individuals with the attribute of interest (number of diseased) among n_i randomly selected or representative individuals with predictor variable value x_i . The subscripts identify the group of cases in the data set.

In many situations, the sample size is 1 in each group, and for this situation Y_i is 0 or 1. The data assumptions are thus:

$$Y_i \stackrel{ind.}{\sim} \text{binomial} \left(n_i, \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right), \text{ for } i = 1, \dots, m.$$

For **raw data** on individual cases, the sample size column n is usually omitted and Y takes on 1 of two coded levels, depending on whether the case at x_i is a success or not. The values 0 and 1 are typically used to identify “failures” and “successes” respectively.

Estimating Regression Coefficients

Maximum likelihood is commonly used to estimate the two unknown parameters in the logistic model. MLE of the regression coefficients are estimated iteratively by maximizing the likelihood function for the responses

$$f(y_1, \dots, y_m | \beta_0, \beta_1) = \prod_{i=1}^m \binom{n_i}{y_i} \left(\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right)^{y_i} \times \left(1 - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right)^{n_i - y_i} .$$

over all possible values of β_0 and β_1 .

Dancing around the subject

A logistic model for these data implies that the probability π of dancing is related to education through

$$\log \left(\frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 \text{ EDUC.}$$

If the model holds, then a slope of $\beta = 0$ implies that π does not depend on EDUC, i.e. the proportion those that have gone dancing in the last year is identical across education levels. However, the power of the logistic regression model is that if the model holds, and if the proportions change with education, then you have a way to quantify the effect of education on the proportion who have gone dancing. This is more appealing and useful than just testing homogeneity across age groups.

A logistic regression model with a single predictor can be fit in R using the generalized linear model function `glm()`. The code and pertinent output for zero/one data (aggregated data requires some additional work) looks like:

```
> fit1 <- glm(d~e,binomial)
> summary(fit1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	9.4177	3.9460	2.387	0.0170 *
e	-1.1258	0.4906	-2.295	0.0217 *

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 27.910 on 20 degrees of freedom
Residual deviance: 17.255 on 19 degrees of freedom
AIC: 21.255

Number of Fisher Scoring iterations: 5

The output tables the MLEs of the parameters: $\hat{\beta}_0 = 9.42$ and $\hat{\beta}_1 = -1.13$. Thus, the fitted or predicted probabilities satisfy:

$$\log \left(\frac{\hat{\pi}}{1 - \hat{\pi}} \right) = 9.42 - 1.13 \text{ EDUC}$$

or

$$\hat{\pi}(\text{EDUC}) = \frac{\exp(9.42 - 1.13 \text{ EDUC})}{1 + \exp(9.42 - 1.13 \text{ EDUC})}.$$

or

$$\hat{O}(\text{EDUC}) = \exp(9.42 - 1.13 \text{ EDUC}).$$

The p -value for testing $H_0 : \beta_1 = 0$ (i.e. the slope for the regression model is zero) based upon large sample approximate normality of MLEs is $\Pr(|Z| > |2.30|) = 0.022$, which leads to rejecting H_0 at the 5% test level (more on this later). Thus, the proportion of those that have gone dancing in the last year is not constant across education level.

The odds of dancing is $\exp(-1.13) = 0.32$ times less for every additional year of education.

Restated, the odds of dancing *increases* by a factor of $\exp(1.13) = 3.1$ for every decrease in the number of years of post high school education. Later on we'll compute a plausible range (confidence interval) for these odds to be in.