Stat 205: Elementary Statistics for the Biological and Life Sciences

Logistic Regression (Section 12.7, pp. 582–585)

April 2011

Simple logistic regression

A 2×2 table provides probabilities of an event across two groups $\hat{p}_1$ and $\hat{p}_2$. Often we do not have distinct groups, but rather a *continuous* covariate to predict probabilities. For example we might want to predict the probability of having a heart attack given our age $\hat{p}(x)$ where x is age in years.

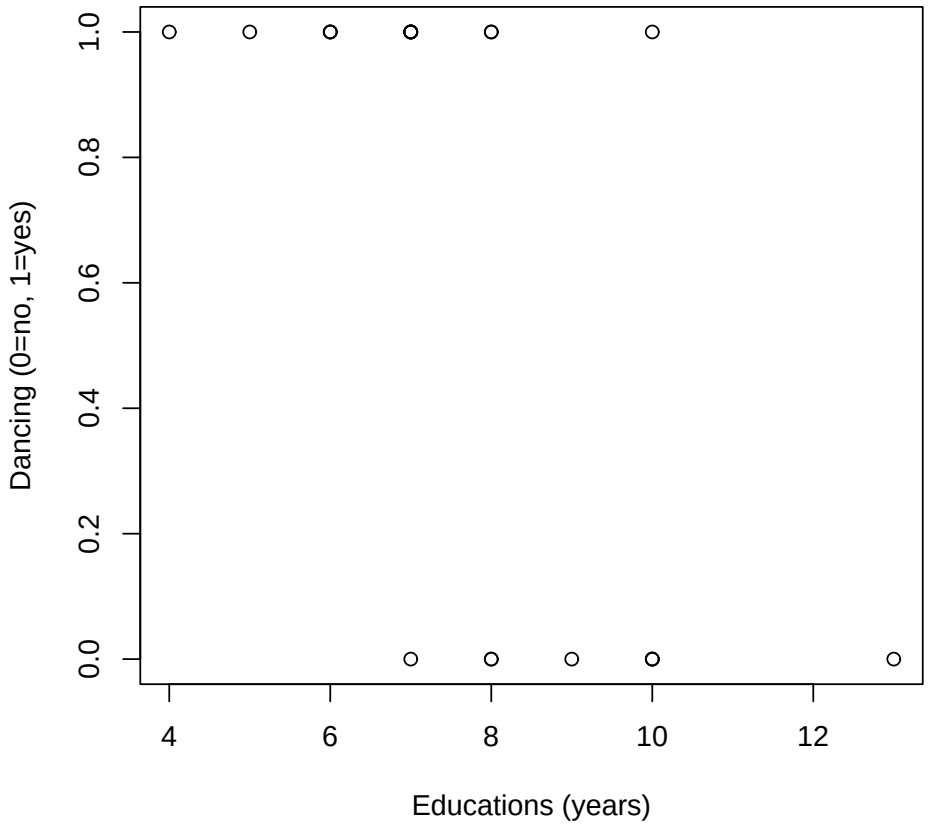
The logistic regression model is for outcomes Y_i that are Bernoulli ("zero-one") with covariate x_i. The probability of success (a "one") changes smoothly with the covariate x_i, much like the mean $\mu_{Y|X}$ changes smoothly in the linear regression model.

Example: From a statistics course I taught to Public Health students at the University of Minnesota, I recorded whether or not someone had gone dancing or not during the semester and the number of years of post high-school education. Here's the data:

Ed. x_i	Danc. Y_i	Ed. x_i	Danc. Y_i	Ed. x_i	Danc. Y_i
10	0	6	1	10	1
6	1	8	0	8	1
9	0	5	1	7	1
6	1	7	0	4	1
8	0	7	1	7	1
7	1	8	1	7	1
13	0	10	0	10	0

$Y_i = 1$ for having gone dancing, and $Y_i = 0$ otherwise. x_i is the number of years of post-high school education.

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



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

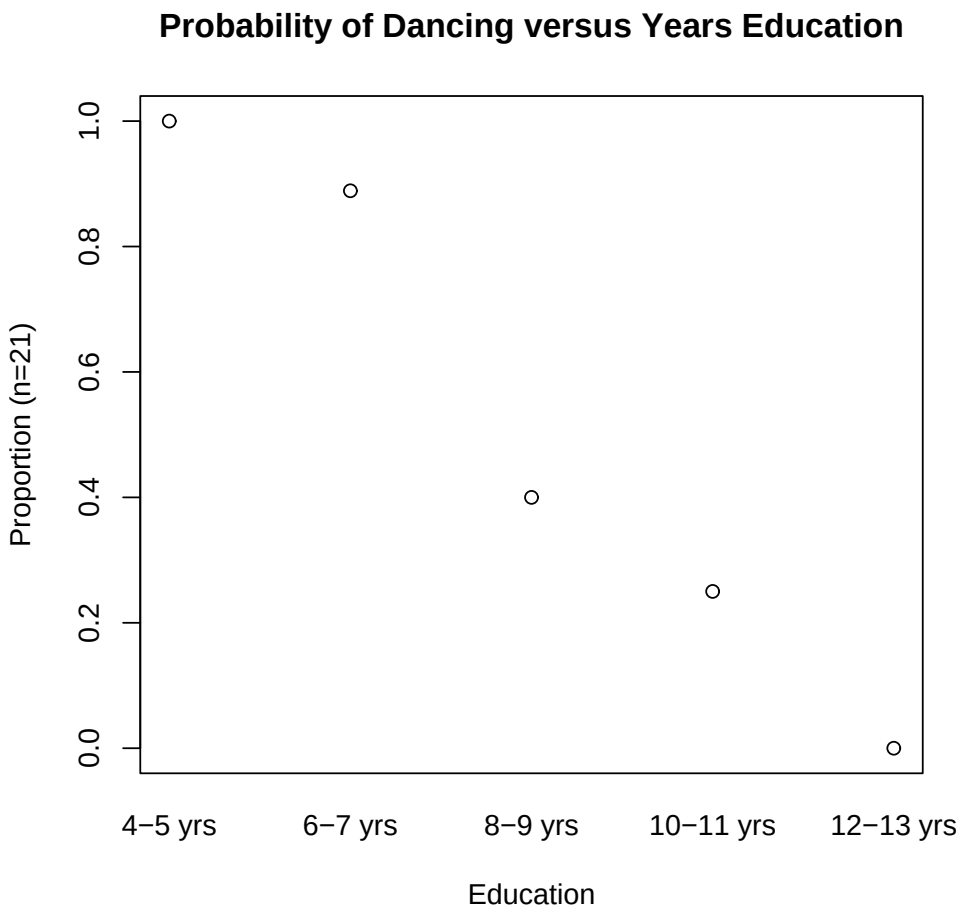
Let's define 5 education categories and compute the sample proportion of those that have gone dancing for each category.

Category	Danced	Total	Proportion
4–5 years	2	2	1.00
6–7 years	8	9	0.89
8–9 years	2	5	0.40
10-11 years	1	4	0.25
12–13 years	0	1	0.00

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

The proportion of those that have gone dancing decreases with age category. A plot helps to see this...

Sample proportions versus education level:



The observed proportions have an a somewhat linear trend plotted against education. We could fit a linear regression model (Chapter 12) 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 $p(x)$ of individuals with a given attribute (called a success) as a function of a single predictor variable x. The model assumes that p is related to x through

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

or, equivalently, as

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

The response Y_i for each individual (i.e. a student enrolled in Stat 205) 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 p as $p(x)$ to emphasize that p is the proportion of all individuals with score x that have the attribute of interest. In the dancing data, $p = p(x)$ is the population proportion of students with education x that have gone dancing within the last year.

We will estimate this function from the data to get $\hat{p}(x)$.

Odds of success

The odds of success are $p/(1-p)$. For example, the odds of success are 1 (or 1 to 1) when $p = 1/2$. The odds of success are 9 (or 9 to 1) when $p = 0.9$. The logistic model assumes that the log-odds of success is linearly related to x (slide 8). Exponentiating both sides:

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

The logistic regression model allows easy inference about the odds of success.

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

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

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.

$\boxed{e^{\beta_1} \text{ is an odds ratio.}}$

Is dancing associated with education?

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

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\ edu.$$

A slope of $\beta_1 = 0$ implies that p does <u>not</u> depend on edu, i.e. the proportion those that have gone dancing in the last year is identical across education levels.

However, if $\beta_1 \neq 0$ then you *quantify* the effect of education on the probability of dancing.

This is more appealing and useful than just testing homogeneity of proportions across education groups (as in Chapter 10).

The logistic regression model can be fit in many statistical packages, including R, Minitab, SAS, etc. Here is some R output:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   9.4177     3.9460   2.387   0.0170
edu          -1.1258     0.4906  -2.295   0.0217
```

We see $\hat{\beta}_0 = 9.42$ and $\hat{\beta}_1 = -1.13$.

The $p-$value for testing $H_0 : \beta_1 = 0$ is 0.0217. Since $0.0217 < 0.05$ we reject $H_0 : \beta_1 = 0$ at the 5% level: there is a significant negative association between education and dancing.

You can also find free software to fit logistic regression (and other analyses such as linear regression, t-tests, etc.) on the web. One nice set of tools is available at `http://www.stattucino.com/`. I used this web-based software to read in `dancing.csv` (comma-separated text file) and fit a logistic regression model:

```
Results:
Logistic Regression Analysis
Summary of Statistical Analysis
Number of observations: 21
Observations with Missing values:   0
Response Variable:  dancing

           Coefficient Standard Error  Wald Test   Chi Square > p
intercept  9.41772     3.94614         5.6957      0.01701
educ      -1.12584     0.4906          5.26619     0.02174
```

These are exactly the same as from R.

Confidence interval for β_1 and e^{β_1}

The main thing we are interested in are the estimated coefficients.
$b_1 = -1.13$ is the estimated slope and $se_{b_1} = 0.49$.

We can get a 95% CI for β_1 as $\boxed{b_1 \pm 1.96 se_{b_1}}$, here

$$-1.13 \pm 1.96(0.49) = (-2.1, -0.2).$$

We exponentiate everything to get odds:

$$e^{-1.13} = 0.32,$$

is the estimate of how the odds of dancing changes with education.

$$(e^{-2.1}, e^{-0.2}) = (0.12, 0.84),$$

is a 95% CI for how the odds change.

The fitted or predicted probability functions are:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 9.42 - 1.13\ edu$$

or

$$\hat{p}(edu) = \frac{\exp(9.42 - 1.13\ edu)}{1 + \exp(9.42 - 1.13\ edu)}.$$

or

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

As mentioned, the p-value for testing $H_0 : \beta_1 = 0$ is 0.022; we reject $H_0 : \beta_1 = 0$ at the 5% test level. The proportion of those that have gone dancing in the last year is not constant across education level. *The probability of dancing is not independent of the amount of education one has had.*

The odds of dancing are one third, $\exp(-1.13) = 0.32$, 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.

On average, the older you get the less fun you have! Or maybe what's fun changes?

Example: From 80 students polled in Stat 205 Spring 2011, whether the student lived on campus or off, and their age in years was recorded. A logistic regression model was fit to the data in Minitab yielding the following output:

```
Variable   Value  Count
residence  on        36  (Event)
           off       44
           Total     80

Logistic Regression Table

Predictor       Coef   SE Coef      Z      P
Constant     18.7198   5.24220   3.57  0.000
age        -0.961319  0.269191  -3.57  0.000
```

For every year increase in age, we estimate that the odds of living on campus changes by $e^{-0.961} \approx 0.38$. *For every year increase in age, the odds of living on campus decreases by more than half.* We can compute a 95% confidence interval as

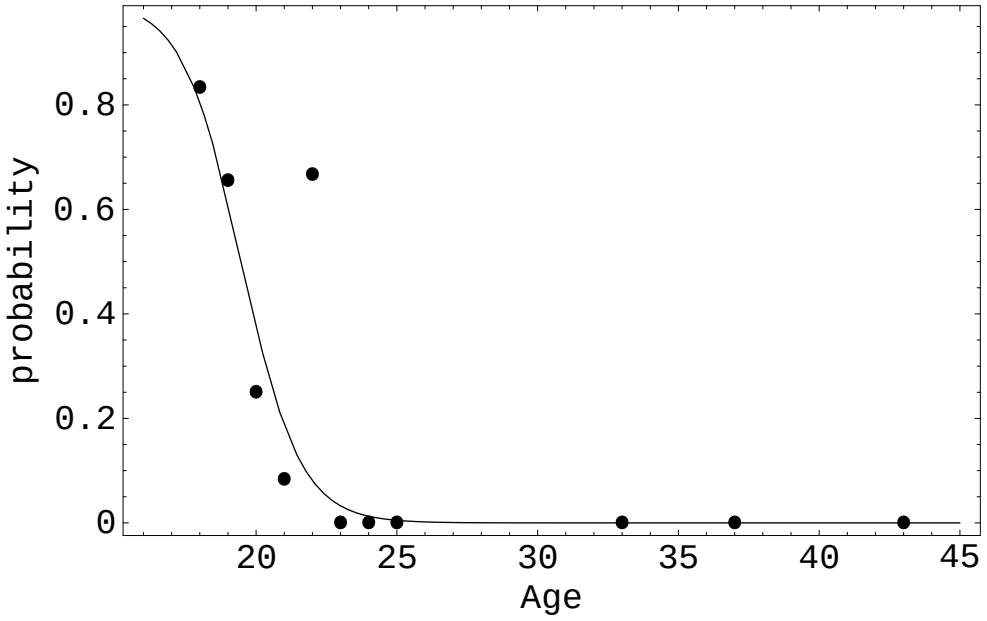
$$\exp(-0.961 - 1.96(0.269)) \approx 0.38 \text{ and } \exp(-0.961 + 1.96(0.269)) \approx 0.65.$$

With 95% confidence, the odds of living on campus change by a factor ranging from 0.38 to 0.65 for every year increase in age.

The probability of living off campus as a function of age is given by

$$\hat{p}(x) = \frac{\exp(18.72 - 0.96\ x)}{1 + \exp(18.72 - 0.96\ x)}.$$

This function and the raw proportions at ages 18, 19, 20, 21, 22, 23, 24, 25, 33, 37, 43 are plotted below:



Homework:

A study was performed to determine the effect of a carcinogen on the survival of rats. Thirty rats were injected with varying levels of a carcinogen x (mg). For rats who survived one week, $Y_i = 1$ was recorded; for rats who died within one week, $Y_i = 0$ was recorded. A logistic regression was fit, and SAS computer output is shown below:

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	6.6656	2.3119	8.3129	0.0039
x	1	-0.2506	0.0826	9.2070	0.0024

1. Does the carcinogen increase or decrease the odds of survival? Explain.

2. Is the carcinogen effect significant? That is, do we accept or reject $H_0 : \beta_1 = 0$ at the 5% level? Explain.

3. Find, *and interpret* a 95% confidence interval for how the odds of survival changes when the amount of carcinogen is increased by one milligram.

Answers:

1. The regression coefficient is negative, so increasing the carcinogen decreases the odds of survival. This makes intuitive sense. The odds of surviving are decreased by a factor of $e^{-0.2506} = 0.78$ for every mg increase in carcinogen.

2. The effect is significant, we reject $H_0 : \beta_1 = 0$ at the 5% level because $0.0024 < 0.05$.

3. A 95% confidence interval for the log odds ratio is

$$b_1 \pm 1.96 \times SE_{b_1} = -0.2506 \pm 1.96 \times 0.0826 = (-0.412, -0.089).$$

Exponentiating gives the 95% confidence interval for how the odds change when increasing the carcinogen by 1 mg:

$$(e^{-0.412}, e^{-0.089}) = (0.66, 0.91).$$