Logistic regression

Department of Statistics, University of South Carolina

Stat 506: Introduction to Experimental Design

Logistic regression

- Sometimes we wish to predict a categorical response Y using a quantitive variable X.
- Consider Y to be binary ($0 =$ failure, $1 =$ success)
- Logistic regression is used to model how the probability of success $\Pr\{Y = 1\}$ depends on X.
- Rather than normally distributed data we now have binomially distributed data.

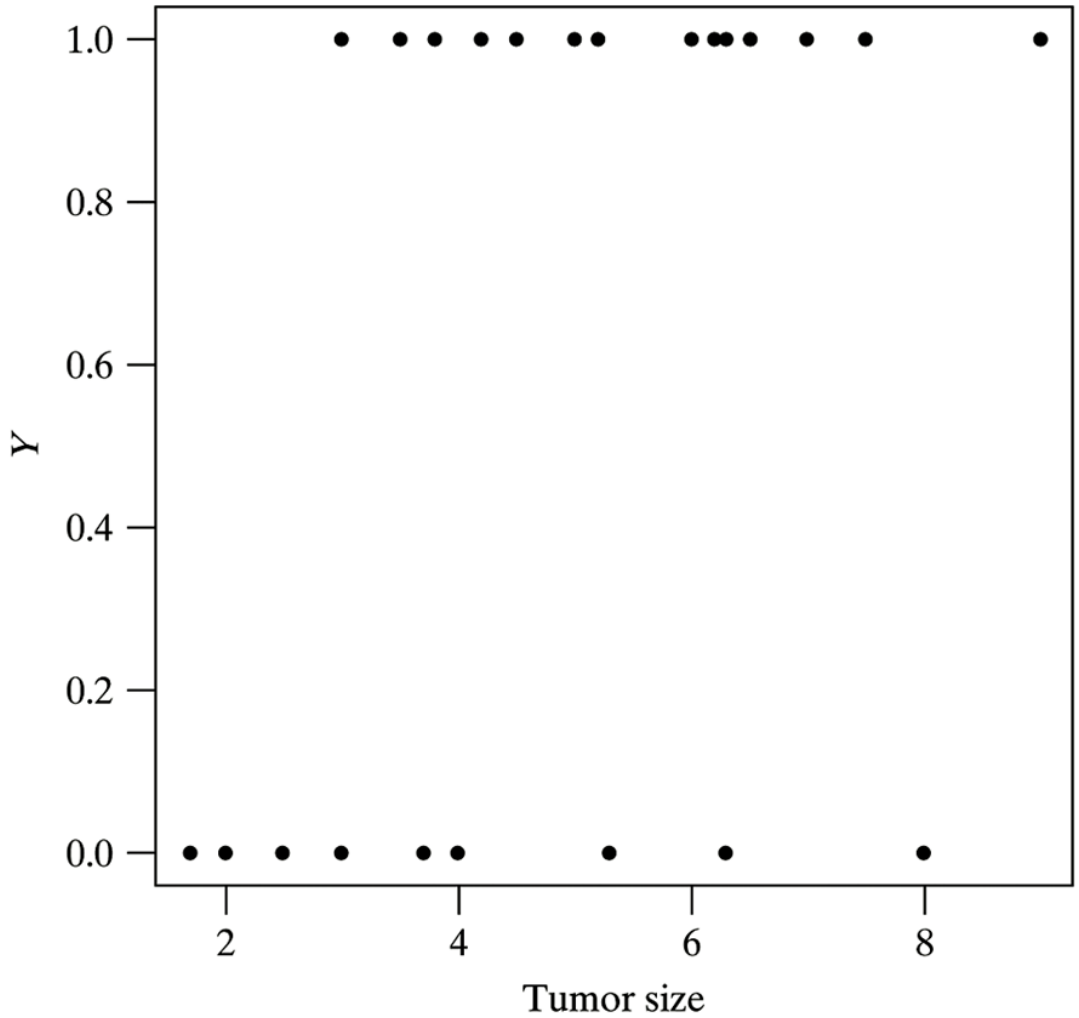
Example: Esophageal Cancer

- Esophageal cancer is a serious and very aggressive disease.
- $n = 31$ patients with esophageal cancer studied; looked at size of patient's tumor & whether cancer had spread (metastasized) to lymph nodes.
- Two variables. $Y = 1$ if cancer spread to lymph notes, $Y = 0$ if not. X is maximum dimension (cm) of esophagus tumor.

Example: Esophageal Cancer

Esophageal cancer data					
Patient number	Tumor size (cm), X	Lymph node metastasis, Y	Patient number	Tumor size (cm), X	Lymph node metastasis, Y
1	6.5	1	17	6.2	1
2	6.3	0	18	2.0	0
3	3.8	1	19	9.0	1
4	7.5	1	20	4.0	0
5	4.5	1	21	3.0	1
6	3.5	1	22	6.0	1
7	4.0	0	23	4.0	0
8	3.7	0	24	4.0	0
9	6.3	1	25	4.0	0
10	4.2	1	26	5.0	1
11	8.0	0	27	9.0	1
12	5.2	1	28	4.5	1
13	5.0	1	29	3.0	0
14	2.5	0	30	3.0	1
15	7.0	1	31	1.7	0
16	5.3	0			

Plot of Y versus X



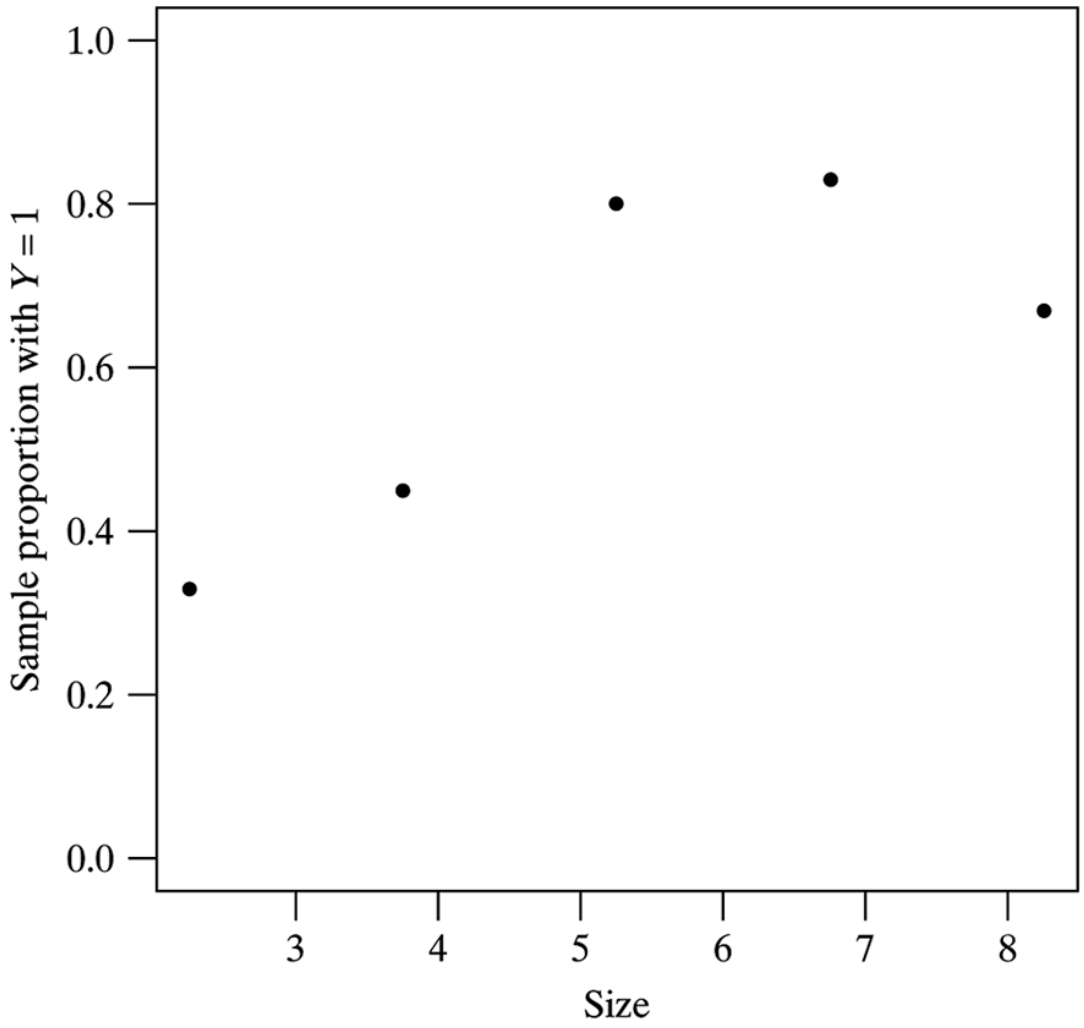
Let's group the predictor "Tumor size" into bins (like a histogram) and compute sample proportions for each bin.

Sample proportions

Esophageal cancer data in groups				
Size range	Points with $Y = 1$	Points with $Y = 0$	Fraction $Y = 1$	Proportion $Y = 1$
(1.5, 3.0]	2	4	2/6	0.33
(3.0, 4.5]	5	6	5/11	0.45
(4.5, 6.0]	4	1	4/5	0.80
(6.0, 7.5]	5	1	5/6	0.83
(7.5, 9.0]	2	1	2/3	0.67

Probability of metastization roughly increases with tumor size.
Let's look at a plot...

Plot of sample proportions



Forms a "lazy S" curve.

Building a model

- The logistic regression model for the probability of success is

$$\Pr\{Y = 1\} = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

- R can give us estimates b_0 (for β_0) and b_1 (for β_1), as well as standard errors SE_{b_0} and SE_{b_1} using the function `glm` (instead of `lm` as in regular regression).
- Recall: $\exp(x) = e^x$ where $e \approx 2.718282$, and $\log(x)$ is the natural logarithm; also $\log(e^x) = x$.

R code for cancer data

```
size=c(6.5,6.3,3.8,7.5,4.5,3.5,4.0,3.7,6.3,4.2,8.0,5.2,
       5.0,2.5,7.0,5.3,6.2,2.0,9.0,4.0,3.0,6.0,4.0,4.0,
                4.0,5.0,9.0,4.5,3.0,3.0,1.7)
Y= c(1,0,1,1,1,1,0,0,1,1,0,1,1,0,1,0,1,0,1,0,1,1,
             0,0,0,1,1,1,0,1,0)
> fit=glm(Y~size,family=binomial)
> summary(fit)

Call:
glm(formula = Y ~ size, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0657  -1.1288   0.5657   0.9844   1.4185

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0858     1.2256  -1.702   0.0888
size          0.5117     0.2561   1.998   0.0457
```

Building a model

- The estimated probability of whether cancer metastisizes is

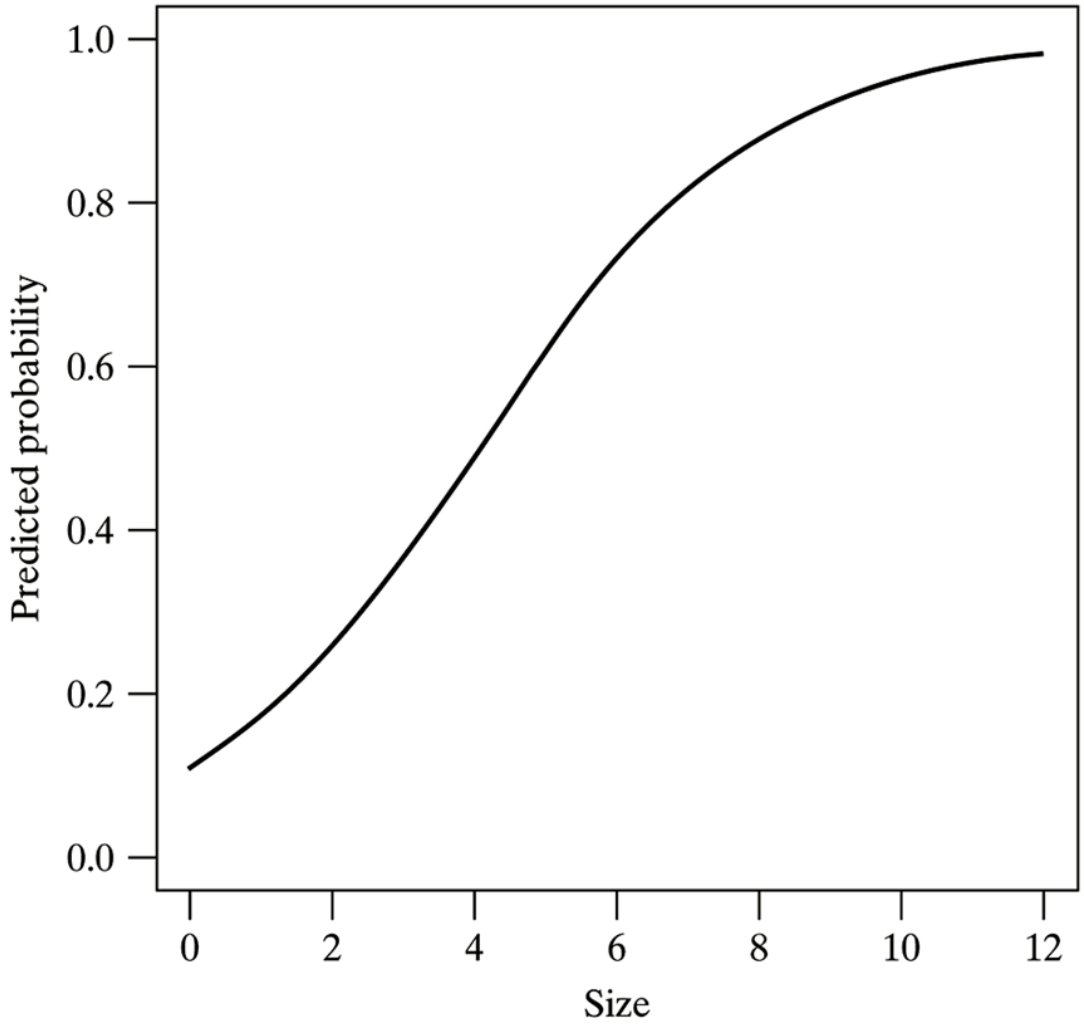
$$\Pr\{Y = 1\} = \frac{e^{-2.086+0.5117 \text{ size}}}{1 + e^{-2.086+0.5117 \text{ size}}}.$$

- This is the same as:

$$\log\left(\frac{\Pr\{Y = 1\}}{1 - \Pr\{Y = 1\}}\right) = -2.086 + 0.5117 \text{ size},$$

  the log–odds of metastization.
- Here $b_0 = -2.086$ estimates β_0 and $b_1 = 0.5117$ estimates β_1.
- We test $H_0 : \beta_1 = 0$ using the P-value from the table; here P-value$= 0.0457 < 0.05$ so we reject $H_0 : \beta_1 = 0$ at the 5% level. There is a significant, positive ($b_1 > 0$) association between tumor size and metastization.

Smooth curve for probability of 'success'

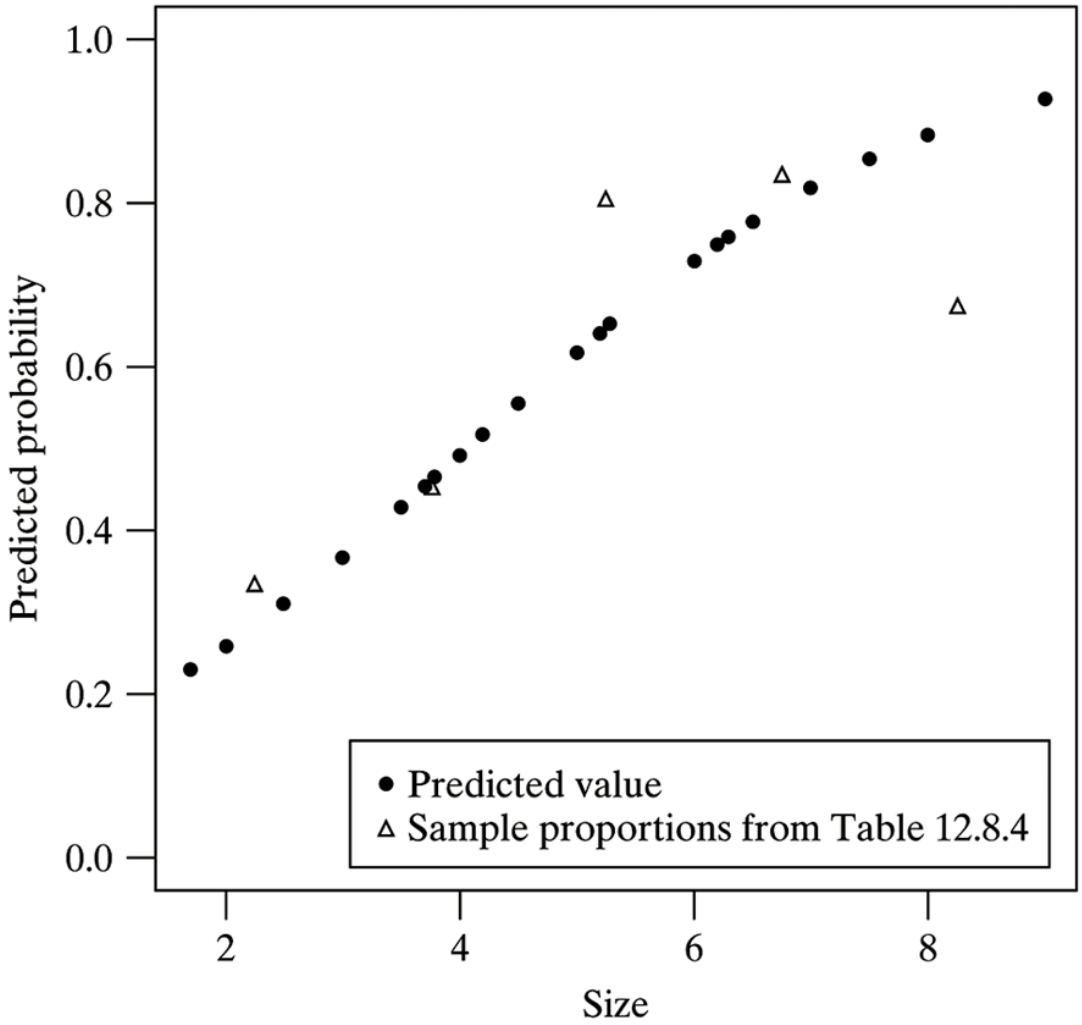


$\Pr\{Y = 1\}$ as a function of tumor size.

Interpretation in terms of odds

- Using the log-odds formula on slide 10, we can show that e^{b_1} is how the odds of success changes when X is increased by one unit.
- For example, when we increase the tumor size by 1 *cm*, the odds of metastization increases by a factor of $e^{0.5117} = 1.668$, i.e. increases by 67%.
- i.e. $e^{0.5117} \approx 1.7$ is an odds ratio.
- If we increase tumor size by 2 *cm* then the odds of metastization increases by $1.7^2 \approx 2.8$ times, or 180%.

Predicted values



Predicted probability at each X-value and sample proportions from windows. Model fits okay.

R code for cancer data (again)

```
size=c(6.5,6.3,3.8,7.5,4.5,3.5,4.0,3.7,6.3,4.2,8.0,5.2,
       5.0,2.5,7.0,5.3,6.2,2.0,9.0,4.0,3.0,6.0,4.0,4.0,
                4.0,5.0,9.0,4.5,3.0,3.0,1.7)
Y= c(1,0,1,1,1,1,0,0,1,1,0,1,1,0,1,0,1,0,1,0,1,1,
             0,0,0,1,1,1,0,1,0)
> fit=glm(Y~size,family=binomial)
> summary(fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0858     1.2256  -1.702   0.0888
size          0.5117     0.2561   1.998   0.0457
```

Questions

- Does the tumor size increase or decrease the odds of having lymph node metastasis Y?
- Is the effect of tumor size significant?
- Find, and interpret a 95% confidence interval for the ratio of odds of Y when the tumor size is increased by 1 cm.

Answers

- The regression coefficient is positive, so increasing the tumor size increases the odds of metastization. This makes intuitive sense. The odds of spreading are increased by a factor of $e^{0.5117} = 1.668$ for every *cm* increase in tumor size.
- The effect is (just) significant, we reject $H_0 : \beta_1$ at the 5% level because $0.0457 < 0.05$.

Confidence interval

- A 95% confidence interval for the log odds ratio is

$$b_1 \pm 1.96 SE_{b_1} = 0.5117 \pm 1.96(0.2561) = (0.010, 1.014).$$

- Exponentiating gives the 95% confidence interval for how the odds change when increasing the size by 1 *cm*: $(e^{0.010}, e^{1.014}) = (1.0097, 2.7557)$.
- Can also get this automatically from R, see next slide...

Elk crossing revisited...

Let's look at an example with two predictors (both factors). Here we fit the logistic regression model using counts of successes and failures instead of zero/one outcomes.

```
success=c(287, 40,237, 57)
failure=c( 57, 42, 52, 12)
location=factor(c("low","low","high","high"))
vehicle=factor(c("car","truck","car","truck"))
f=glm(cbind(success,failure)~location+vehicle,family=binomial)
exp(cbind(OR=coef(f),confint(f))) # OR's w/ CI's
exp(-coef(f)[3]) # compare to estimate via CMH
```

Essentially the same analysis as using Cochran-Mantel-Haenszel!

What is the effect of low vs. high density?