Logistic Regression

For a binary dependent variable:
1=Yes, 0=No

Binary outcomes are common and important

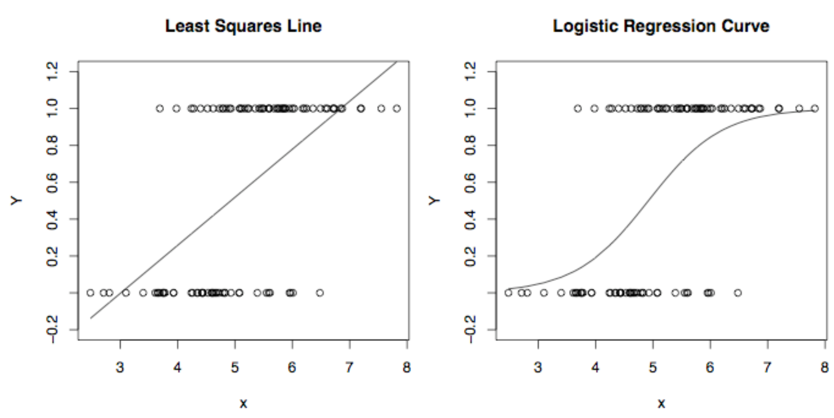
- The patient survives the operation, or does not.
- The accused is convicted, or is not.
- The customer makes a purchase, or does not.
- The marriage lasts at least five years, or does not.
- The student graduates, or does not.

For a binary variable

- The population mean E[Y] is the probability that Y=1
- Make the mean depend on a set of independent variables

- Consider one independent variable. Think of a scatterplot

Least Squares vs. Logistic Regression



The logistic regression curve arises from an indirect representation of the probability of Y=1 for a given set of x values.

Representing the probability of an event by π

$$\text{Odds} = \frac{\pi}{1-\pi}$$

$$\text{Odds} = \frac{\pi}{1-\pi}$$

- If P(Y=1)=1/2, odds = .5/(1-.5) = 1 (to 1)
- If P(Y=1)=2/3, odds = 2 (to 1)
- If P(Y=1)=3/5, odds = (3/5)/(2/5) = 1.5 (to 1)
- If P(Y=1)=1/5, odds = .25 (to 1)

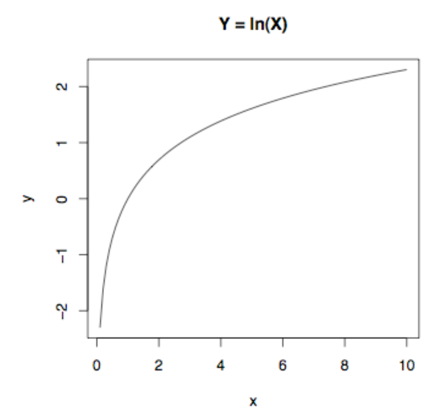
The higher the probability, the greater the odds

$$\text{Odds} = \frac{\pi}{1-\pi}$$

$$0 \leq \text{ Odds } < \infty$$

Linear model for the **log** odds

- Natural log, not base 10
- Symbolized `ln`



Some facts about `ln`

- The higher the probability, the higher the log odds.
- `ln`(*e*)=1, *e* = 2.1728…
- Only defined for positive numbers.
- So logistic regression will not work for events of probability exactly zero or exactly one (why not one?)

The log of a product is the sum of logs

$$\ln(ab) = \ln(a) + \ln(b)$$

$$\ln(\frac{a}{b}) = \ln(a) - \ln(b)$$

This means the log of an odds *ratio* is the difference between the two log odds quantities.

Linear regression model for the log odds of the event Y=1

$$\ln\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

Equivalent Statements

$$\ln\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

$$\begin{aligned}\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})} &= e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}} \\ &= e^{\beta_0} e^{\beta_1 x_1} \cdots e^{\beta_{p-1} x_{p-1}}\end{aligned}$$

$$P(Y=1|x_1, \ldots, x_{p-1}) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}}}$$

In terms of log odds, logistic regression is like regular regression

$$\ln\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

Logistic regression

- X=1 means smoker, X=0 means non-smoker
- Y=1 means dead, Y=0 means alive
- Log odds of death = $\beta_0 + \beta_1 x$
- Odds of death = $e^{\beta_0} e^{\beta_1 x}$

In terms of plain odds,

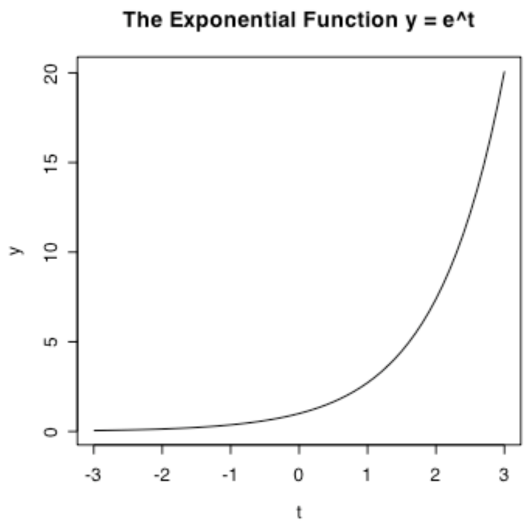
- Logistic regression coefficients represent *odds ratios*
- For example, “Among 50 year old men, the odds of being dead before age 60 are three times as great for smokers.”

$$\frac{\text{Odds of death given smoker}}{\text{Odds of death given nonsmoker}} = 3$$

$$\text{Odds of Death} = e^{\beta_0} e^{\beta_1 x}$$

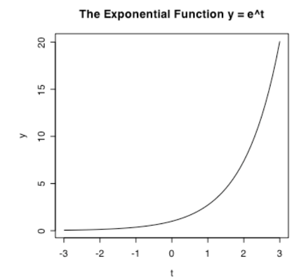
Group	x	**Odds of Death**
Smokers	1	$e^{\beta_0} e^{\beta_1}$
Non-smokers	0	e^{β_0}

$$\frac{\text{Odds of death given smoker}}{\text{Odds of death given nonsmoker}} = \frac{e^{\beta_0} e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$



Exponential function $f(t) = e^t$

- Always positive
- $e^0=1$, so when $\beta_1 = 0$, the odds ratio equals one (50-50).
- $f(t) = e^t$ is increasing



One more example

$$\text{Log Survival Odds} = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 x$$

Treatment	d_1	d_2	**Odds of Survival** $= e^{\beta_0} e^{\beta_1 d_1} e^{\beta_2 d_2} e^{\beta_3 x}$
Chemotherapy	1	0	$e^{\beta_0} e^{\beta_1} e^{\beta_3 x}$
Radiation	0	1	$e^{\beta_0} e^{\beta_2} e^{\beta_3 x}$
Both	0	0	$e^{\beta_0} e^{\beta_3 x}$

For any given disease severity x,

$$\frac{\text{Survival odds with Chemo}}{\text{Survival odds with Both}} = \frac{e^{\beta_0} e^{\beta_1} e^{\beta_3 x}}{e^{\beta_0} e^{\beta_3 x}} = e^{\beta_1}$$

In general,

- When x_k is increased by one unit and all other independent variables are held constant, the odds of Y=1 are multiplied by e^{β_k}
- That is, e^{β_k} is an **odds ratio** --- the ratio of the odds of Y=1 when x_k is increased by one unit, to the odds of Y=1 when everything is left alone.
- As in ordinary regression, we speak of "controlling" for the other variables.

The conditional probability of Y=1

$$P(Y=1|x_1,\ldots,x_{p-1}) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}$$

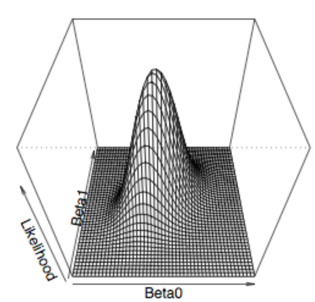
This formula can be used to calculate a predicted P(Y=1)
Just replace betas by their estimates (b)

It can also be used to calculate the probability of getting
The sample data values we actually did observe.

Maximum likelihood estimation

- Likelihood = Probability of getting the data values we did observe
- Viewed as a function of the parameters (betas), it's called the likelihood function
- Those parameter values for which the likelihood function is greatest are called the *maximum likelihood estimates.*
- Thank you again, Mr. Fisher.

Likelihood Function for Simple Logistic Regression



Maximum likelihood estimates

- Must be found numerically
- Lead to nice large-sample chi-square tests
- We will mostly use Wald tests