SESSION 10 LOGISTIC REGRESSION: DETAILS

Review

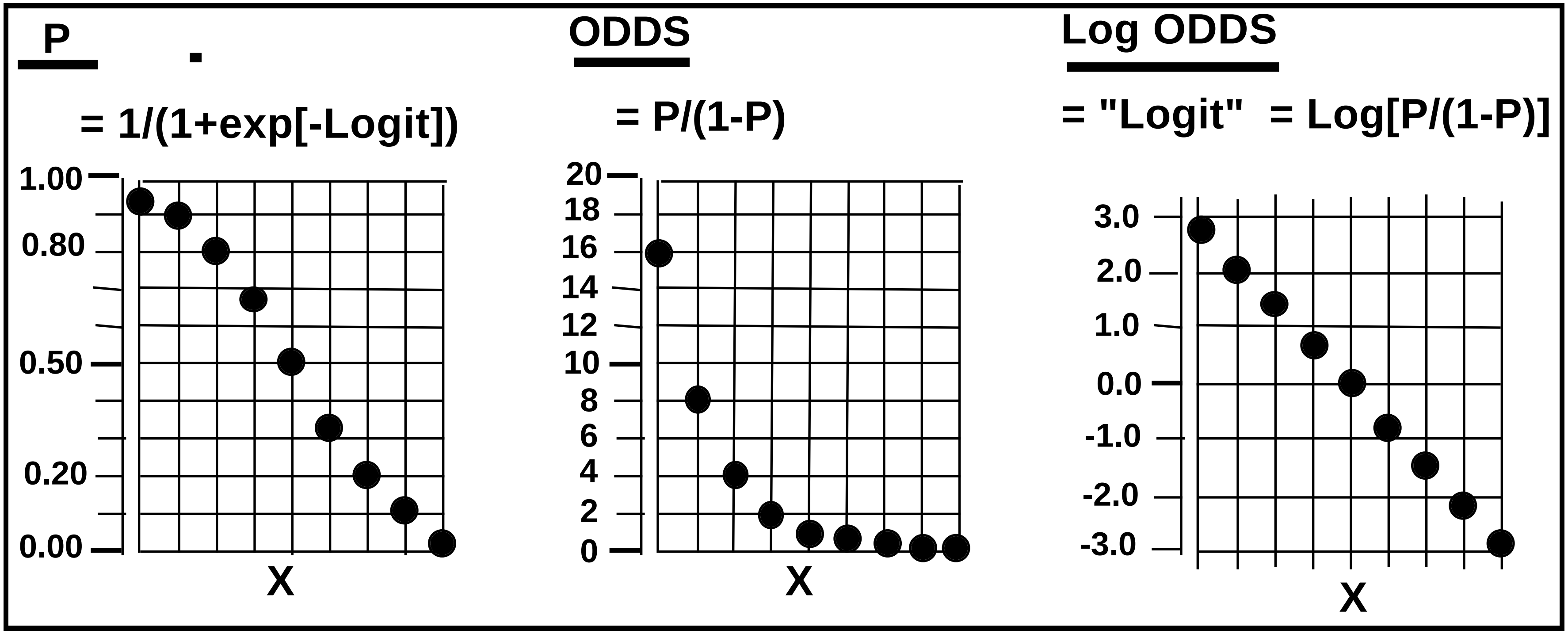
Data: Binary Y's; Parameters of interest: PROPORTIONS (P's)

"regular" regression not usually appropriate
(constraints on range of P: var[Y|X] varies with P, ...)

Logistic regression:

Logit = Log odds = Log[P/(1-P)] {Log is to base e, where e = 2.718...}

Logit[P] linear in X <==> P is S-shaped function of X;
Logit[P] = B_0 + B_1.X <==> P = 1/(1+exp[-(B_0 + B_1.X)])



P

= 1/(1+exp[-Logit])

ODDS

= P/(1-P)

Log ODDS

= "Logit" = Log[P/(1-P)]

Logistic regression

P nearly linear in X if narrow range of P

Logistic regression is one of family of
Generalized Linear Models for Binary (Bernoulli) Y's..
or (if few covariate patterns) Sums of Binary Y's (Binomial counts)
with P's "indexed by", i.e., a function of, X values

```
Identity "LINK": => P      as linear function of X;
Log      "LINK": => log[P] as linear function of X.
```

Fitting of parameters (nowadays, not as in Cornfield 1962):

Parameter values that give the Maximum (log) Likelihood

Key: a probability model (here Binomial) that, for a given value of model parameters, applied to the known X's, yields the probability of observing each observation.

=> probability(each observed Y | model)

=> probability(entire set of observed Y's | model)
= PRODUCT of probabilities ("likelihoods") of Y's

PRODUCT small --> use Log(PRODUCT) (= sum of log of components)

maximize log likelihood = Σ log[prob(Y)| β's and X's]

(cf. minimize SS(residuals) = Σ (Y-Yhat)^2 | β's and X's)

Going back and forward between P, Odds and Logit

(e.g. of logistic regression with 2 X's)

$$\text{Logit of } P|X_1 \ \& \ X_2 = \text{Log odds} = \text{Log} \frac{P}{1-P} = \beta_0 + \beta_1.X_1 + \beta_2.X_2$$

$$\text{Odds} = \text{antiLog[Log odds (= Logit)]}$$

$$= \exp[\text{Logit}] = \exp[\beta_0 + \beta_1.X_1 + \beta_2.X_2]$$

(exp[value] =antilog of a value, where log is to base e)

$$P = \frac{\text{odds}}{1+\text{odds}} = \frac{\exp[\beta_0 + \beta_1.X_1 + \beta_2.X_2]}{1+\exp[\beta_0 + \beta_1.X_1 + \beta_2.X_2]}$$ (I prefer this form)

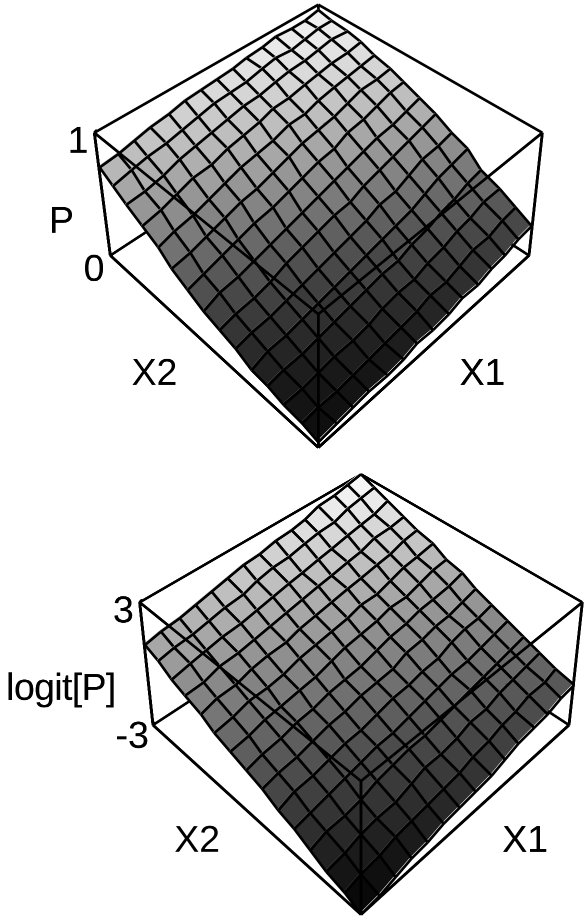
it has form: odds / (1+odds)

$$= \frac{1}{1+\exp[-(\beta_0 + \beta_1.X_1 + \beta_2.X_2)]}$$ (others prefer this form)

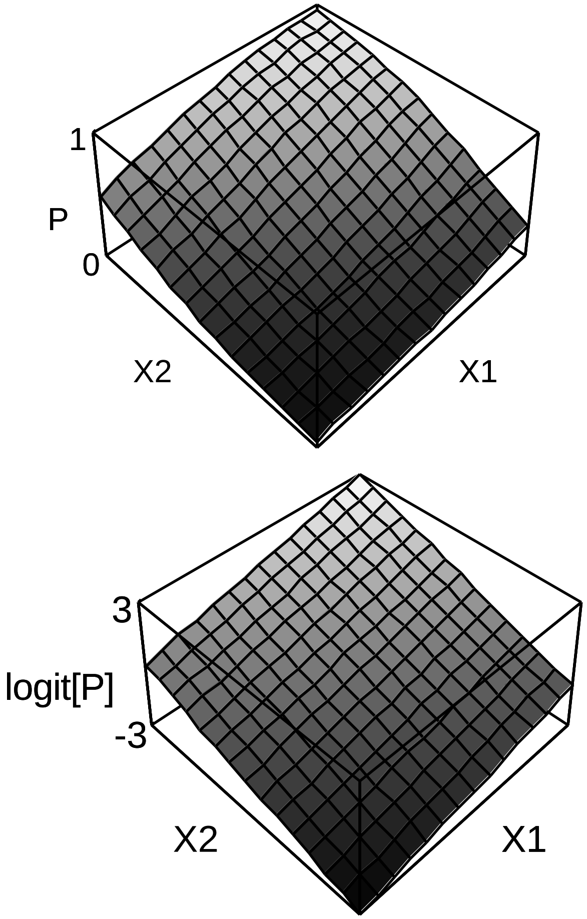
Going back and forward between P, Odds and Logit

Examples of logistic regression with 2 X's ranging from 0 to 1

$$\text{Logit}[P|X_1 \,\&\, X_2\,] = -3 + 2\,.X_1 + 4\,.X_2$$



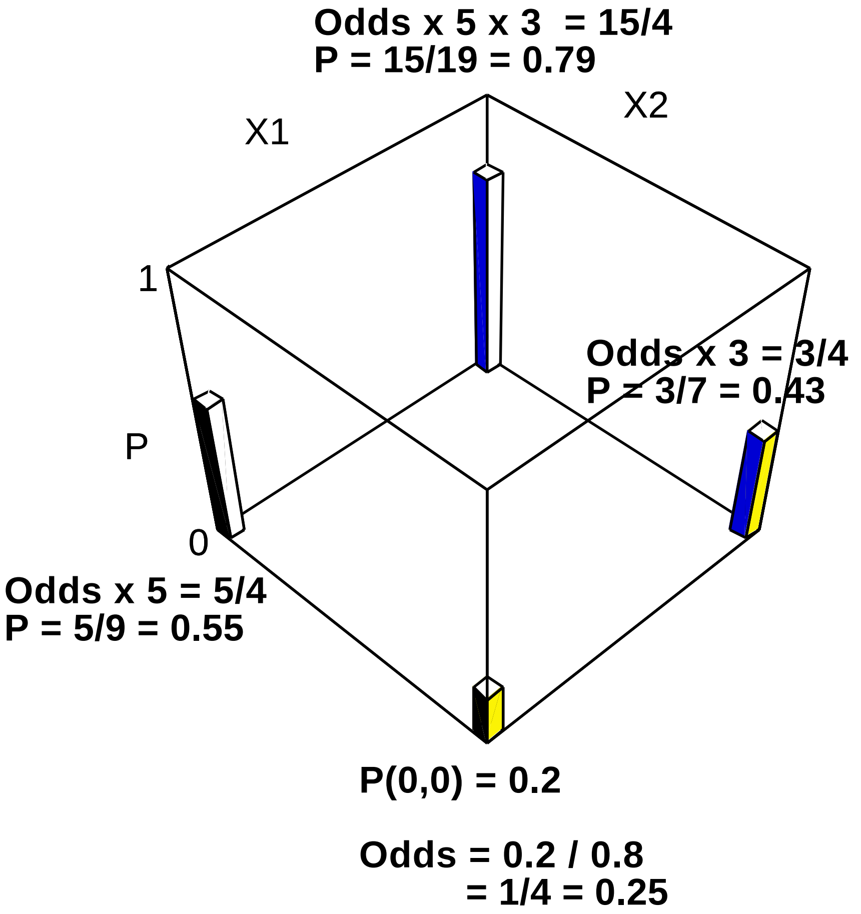
$$\text{Logit}[P|X_1 \,\&\, X_2\,] = -3 + 2\,.X_1 + 3\,.X_2 + 1\,.X_1.X_2$$



Example of logistic regressions with 2 BINARY X's

$$\text{Logit}[P|X_1 \,\&\, X_2\ ] = \beta_0 + \beta_1.X_1 + \beta_2.X_2$$

-1.39 1.10 1.61



Odds x 5 x 3 = 15/4
P = 15/19 = 0.79

Odds x 3 = 3/4
P = 3/7 = 0.43

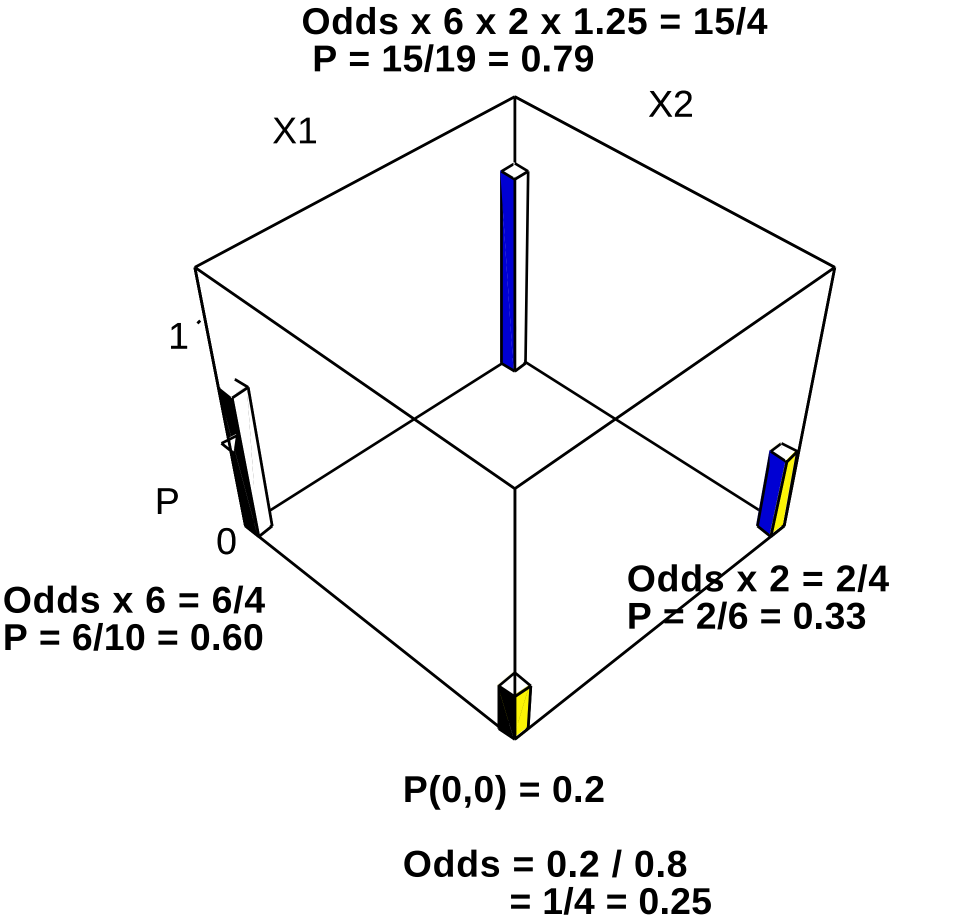
Odds x 5 = 5/4
P = 5/9 = 0.55

P(0,0) = 0.2

Odds = 0.2 / 0.8
= 1/4 = 0.25

$$\text{Logit}[P|X_1 \,\&\, X_2\ ] = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \beta_3.X_1.X_2$$

-1.39 0.69 1.79 0.22



Odds x 6 x 2 x 1.25 = 15/4
P = 15/19 = 0.79

Odds x 6 = 6/4
P = 6/10 = 0.60

Odds x 2 = 2/4
P = 2/6 = 0.33

P(0,0) = 0.2

Odds = 0.2 / 0.8
= 1/4 = 0.25

Note: with a two-point (BINARY) X, there is no issue of “linearity” of the logit with respect to X.

Interpretation of β coefficients in Logistic Regression

First: Interpretation of β in Generalized Linear Models

β = difference in LINK function of μ[Y|X]
for a difference of 1 unit in X,
with other X's in model held constant

μ = Proportion (P) of 1's when dealing with 0/1 Y data

so, if LINK is...

IDENTITY β = P | X + 1 MINUS P | X

LOG β = log{ P | X + 1 } MINUS log{ P | X }

LOGIT

("LOG ODDS") β = logit{ P | X + 1 } MINUS logit{ P | X }

Converting β's in Generalized Linear Models to useful parameters

IDENTITY   β =   P | X + 1   MINUS   P | X

"as is" ... represents RISK DIFFERENCE (RD)
for a 1 unit difference in X

LOG   β =   log{ P | X + 1 }   MINUS   log{ P | X }

Taking Antilogs of both sides
(remember: difference of logs of 2 values = log of their ratio)

ANTI-LOG

$$\exp[\beta] = \frac{P \mid X + 1}{P \mid X}$$

.. represents RISK RATIO (RR)
for 1 unit difference in X

Converting β's in Logistic Models to useful parameters

REMEMBER that "logit" is "LOG ODDS"
so LOGISTIC Regression is LOG ODDS Regression

LOGIT LINK

β = logit{ P | X + 1 } MINUS logit{ P | X }

Remembering that a logit is a LOG ODDS ...

β = log ODDS | X + 1 MINUS log ODDS | X

Again, taking Antilogs of both sides, and remembering that a difference of logs of 2 values = log of their ratio...

$$\exp[\beta] = \frac{\text{ODDS} \mid X + 1}{\text{ODDS} \mid X}$$

... represents <u>ODDS RATIO</u> (OR) for 1 unit difference in X

Point and (Confidence) Interval Estimate of Odds Ratio

For simplicity, use B instead of β, & b instead of β_hat ...

& use “OR” for Odds Ratio PARAMETER and or for estimate ...

b = log[or] ; large sample CI for B : b +/- z.SE[b]

Thus ...

OR_hat = or = exp[b] = e^b

large sample CI for OR : exp{ b +/- z.SE[b] }

As in any regression, must specify the units for X ...

Wound Infection After Cesarean Delivery

Objective: Studies measuring postoperative infectious morbidity following cesarean delivery (CD) have been limited to inpatient data, leading to an underreporting of surgical site infection rates. The primary objective of this study was to determine the true rate of wound infection in patients undergoing emergency CD at the Royal Alexandra Hospital in 1997 and 1998. The secondary objective was to identify risk factors for wound infection following CD.

Methods: Patients were contacted by phone one month after discharge and a questionnaire was completed to diagnose wound infection. Patient charts were then reviewed for the presence or absence of modifiable risk factors for wound infection: duration of membrane rupture, number of vaginal examinations, obesity, surgical prophylaxis and other antibiotic use in labor. Patients who were preterm or who had not labored were excluded. **Risk factor data were analyzed using logistic regression analysis.**

Results: 62% of patients who underwent CD in 1997 were contacted. Of these 948, 341 were term and had labored prior to their CD. A total of 31 patients developed wound infections giving a cumulative incidence of 9.1%. Obesity was the only significant risk factor in our analysis (OR=1.05, CI=1.01- 1.10). The power of the other risk factors studied was only 55%.

Conclusion: In 1997, the wound infection rate following CD in term patients who have labored is 9.1%. Obesity is the only modifiable risk factor identified using 1997 data; however, additional data from 1998 are currently being collected and analyzed.

Point & Interval Estimate of Odds Ratio for a difference δX

b = difference in log odds for 1 unit difference in X

b.(δX) = difference in log odds for δX unit difference in X

Thus ...

Point Estimate

OR_hat [X + δX relative to X] = exp[b.(δX)]

CI:

OR[X + δX rel. to X]: exp{b.(δX) +/- z.(δX).SE[b]. }

Check on Interval Estimate of OR (often done by calculator)

CI for B=log[OR] is (additively) symmetric
around point estimate b=log[or]

e.g.:

b = 2.5; SE[b] = 0.5 ; Z=2 for 95% (CI -- rounded from 1.96)

CI for B : 2.5 +/- 2.(0.5) = 1.5 to 3.5

So... CI for OR = exp[B] is (multiplicatively) symmetric
around point estimate or=exp[b]

with b = 2.5 and SE[b] = 0.5 ...

point estimate of OR = exp[2.5]= 12.2

CI for OR : exp[1.5 to 3.5] = 4.5 to 33.1

Check: Lower limit (4.5) as many multiples below
Point Estimate (12.2)
as upper limit (33.1) is above.

4.5 x 33.1 = square of 12.2
(apart from rounding) and
12.2/4.5 = 33.1/12.2

Modelling Effect modification and Interpreting the parameters

Links with the more familiar: the 2 x 2 Table

	X=1	X=0
Y=1	a	b
Y=0	c	d
ODDS	a/c	b/d
ODDS RATIO	(a/c) / (b/d) = ad/bc	

As a Fitted Logistic Regression Model:

$$\log[\text{odds}] = B_1 .X + B_0$$

$$\log[\text{odds}] = \log[ad/bc].X + \log[b/d]$$

$$\text{odds} = b/d \qquad \text{if } X=0$$

$$\begin{aligned} \text{odds} &= (b/d).(ad/bc) \\ &= a/c \qquad \text{if } X=1 \end{aligned}$$

Check, using a logistic regression program, that

$$SE[b_1] = \text{sqrt}[\, 1/a + 1/b + 1/c + 1/d]$$

Other Links with the more familiar: several (J) 2 x 2 Tables

stratum			X		
			1	0	
		1	a_1	b_1	
1	Y				
		0	c_1	d_1	
					n_1
...					
		1	a_J	b_J	
J	Y				
		0	c_J	d_J	
					n_J

Dataset for logistic regression

stratum	X	Y	Number
1	1	1	a_1
1	1	0	c_1
1	0	1	b_1
1	0	0	d_1
...	...	...	...
1	1	1	a_J
1	1	0	c_J
1	0	1	b_J
1	0	0	d_{1J}

$OR_{\text{Mantel-Haenszel}}$

$$= \Sigma a_j d_j / n_j \ / \ \Sigma a_j d_j / n_j$$

Logistic regression with indicator variables for strata

or_{MH} and $or = \exp[B_X]$ from logistic regression NOT identical.

.s (unconditional) Logistic regression not appropriate if table margins are extreme -- in such situations, use conditional logistic regression (avoids having to fit 1 B for each stratum).