2.1 From Tables to Logistic Regression Models

Regression Analysis

Analysis of how one or more independent variables, X, impact the value of a dependent variable Y

Specifically, what can we say about Y if we know X?

1. Is there a relationship between variables X and Y?
2. How does Y change if X changes?
3. What is the best guess for Y for a given value of X?
4. …
5. …

Different types of outcome variables

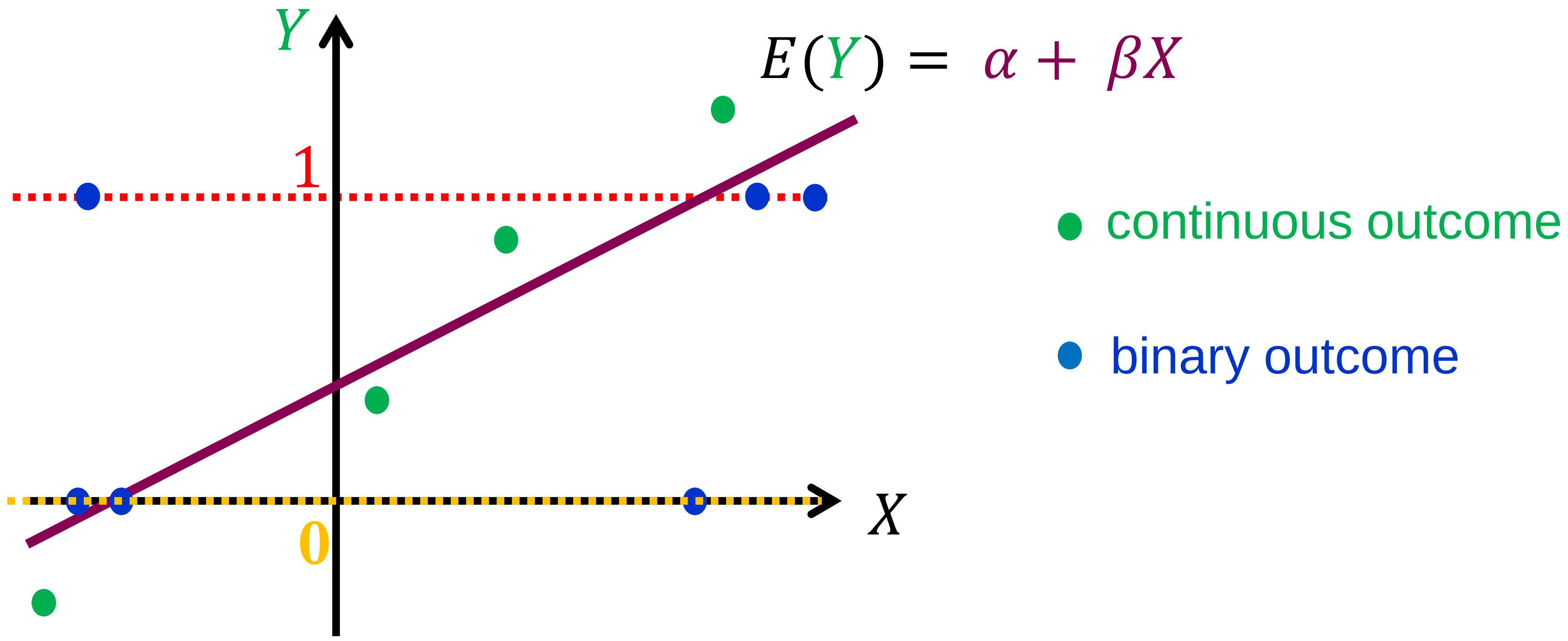
Response variable *Y*	Explanatory variable *X*
Survival after diagnosis	Dosage of drug
p53 expression level	Radiation dose
Diagnosis 0/1	Serum biomarker level
Weight loss	Physical activity/week
Response to drug	Tumor genotype

Outcome (Y)	Regression model
Continuous	Linear regression
Binary	Logistic regression
Count/rate	Poisson regression
Time	Cox regression

For a **continuous outcome** Y and an exposure X

Common model: $Y = \alpha + \beta X + \varepsilon$ (linear regression)

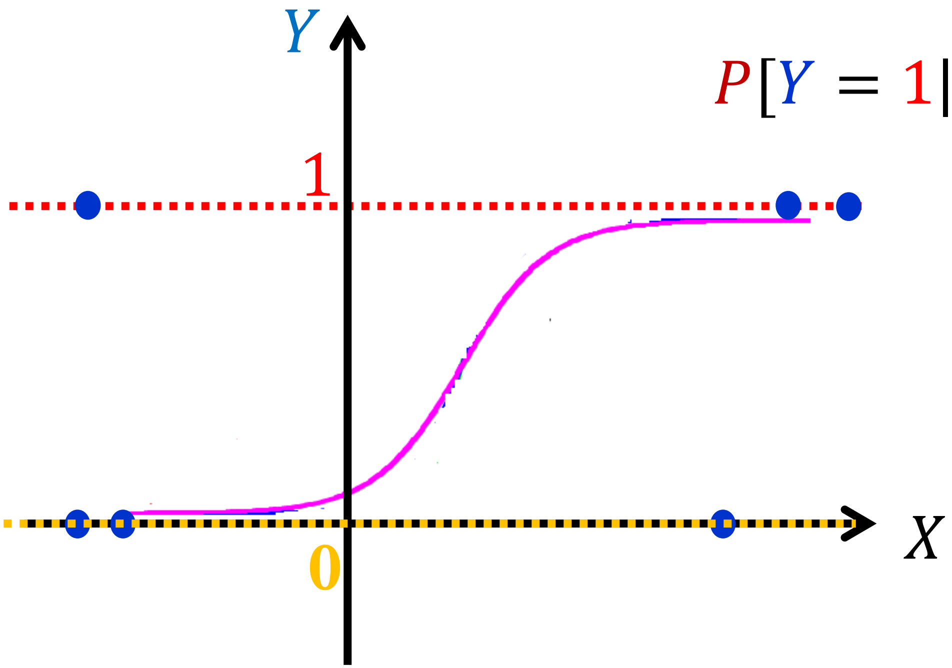
For **binary outcome** Y **(yes=1, no= 0)**,
linear model unreasonable (as Y has only 2 values)



For a **continuous outcome** Y and an exposure X

Common model: $Y = \alpha + \beta X + \varepsilon$ (linear regression)

For **binary outcome** Y **(yes=1, no= 0)**,
model the *probability* that Y=1 for a given X as:



$$P[Y = 1|X] = \frac{exp^{\alpha+\beta X}}{1 + exp^{\alpha+\beta X}}$$

Logistic regression model

Logistic regression
Binary outcome (Y: yes=1, no= 0)

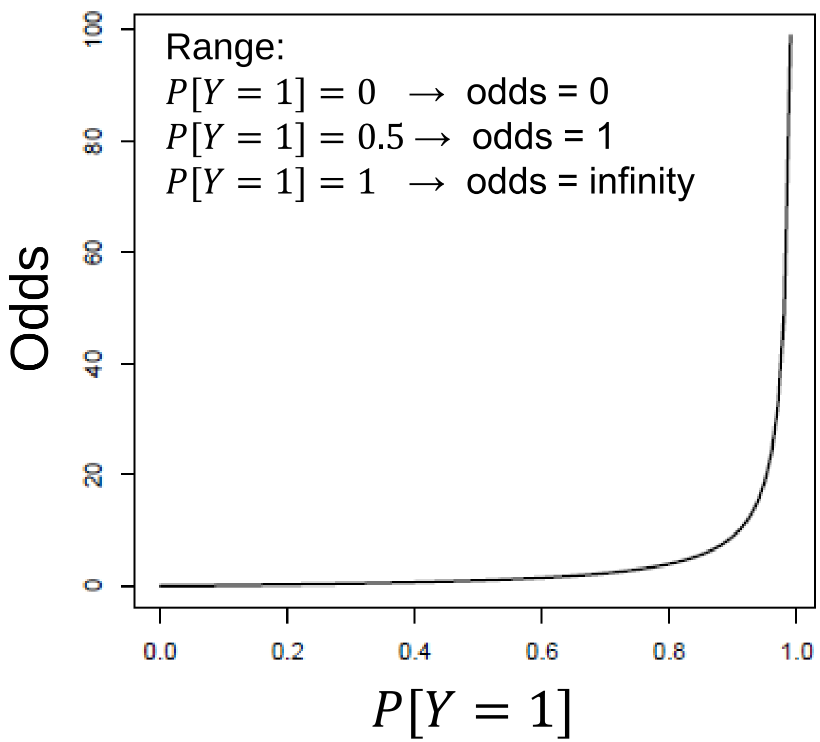
Model $P[Y = 1|X]$ as $P[Y = 1] = \frac{exp^{\alpha+\beta X}}{1+exp^{\alpha+\beta X}}$ (logistic regression):

$\text{Odds}(Y = 1) = exp^{\alpha+\beta X}$

$\log_e(\text{Odds}) = \alpha + \beta X$

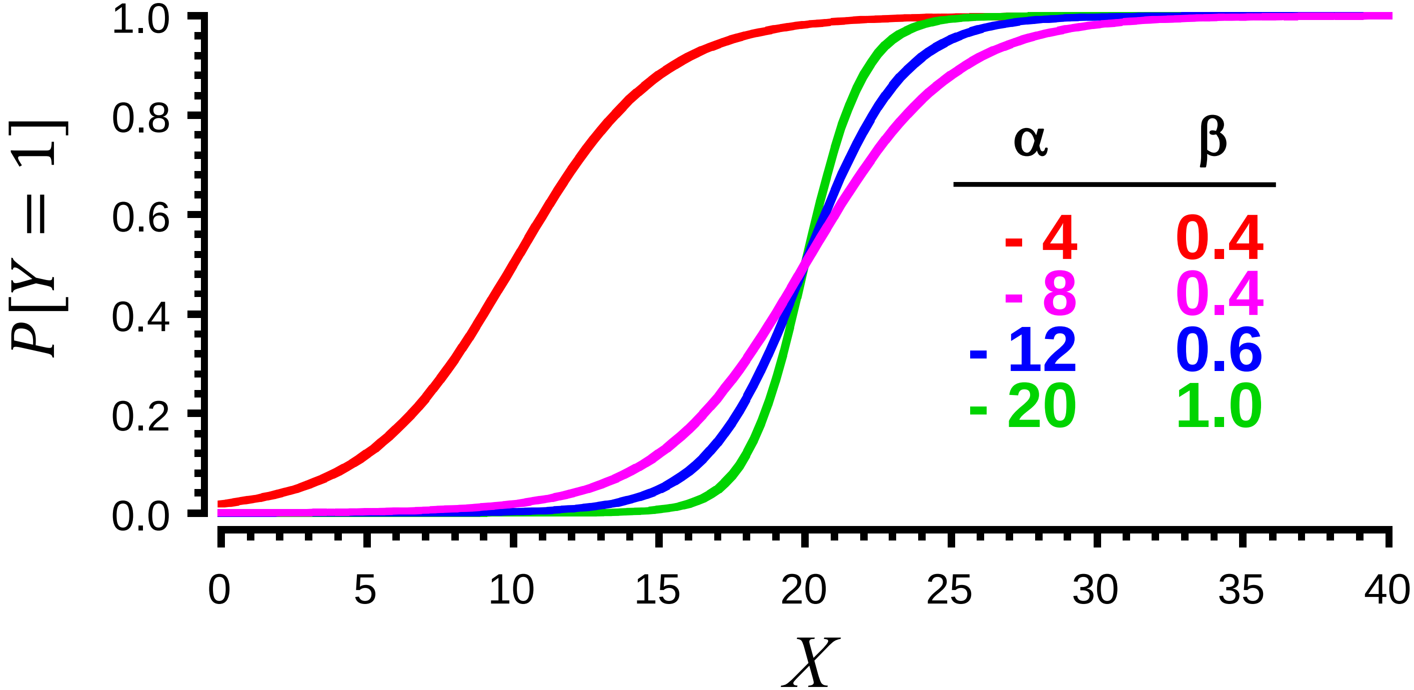
$\log_e(\text{Odds})$ also called:
log-odds, ln(odds),
logit of $P[Y = 1]$

The ln(odds) is linearly related to X



A logistic model of the ***probability*** of the outcome for different X values is a very flexible (sigmoidal) curve:

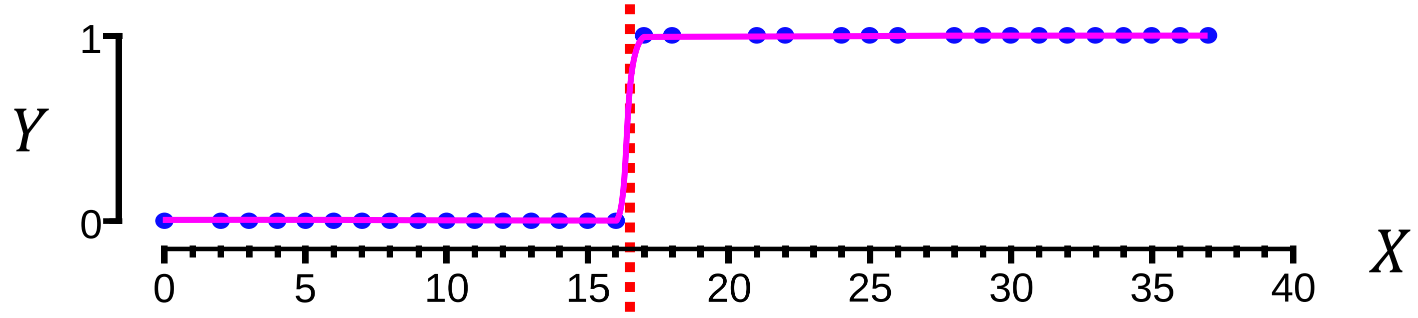
$$P[Y = 1] = \frac{exp^{\alpha+\beta X}}{1 + exp^{\alpha+\beta X}}$$



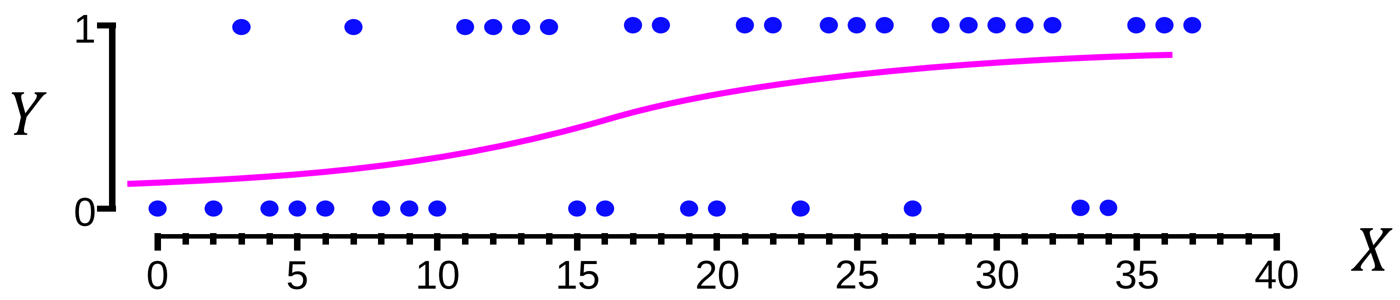
Logistic regression analysis finds the α and β of the curve that ”best fits” the data

(method: ”maximum likelihood”)

Observations with ($Y = 1$) and without the outcome ($Y = 0$) are clearly separated by X (see dotted red line) would have a large value of β



Observations with ($Y = 1$) and without the outcome ($Y = 0$) cannot be separated by X would have a small value of β



Simplest case, binary X
If $X = 1$ (exposed), 0 (unexposed)

The logistic model assumes

$$\text{Prob (outcome)} = \frac{exp^{\alpha+\beta X}}{1+exp^{\alpha+\beta X}}$$

$$\text{i.e., odds (outcome)} = exp^{\alpha+\beta X}$$

If $X = 1$: $\text{odds}_1 = exp^{\alpha+\beta}$

If $X = 0$: $\text{odds}_0 = exp^{\alpha}$

$$\text{odds}_1/\text{odds}_0 = \textbf{OR} = \frac{exp^{\alpha+\beta}}{exp^{\alpha}} = exp^{\beta}$$

β = $\log_e$ of the **OR**

Exposure with more than 2 levels

ln(odds) (Y=1)		odds (Y=1)	
α	for level 0	exp^{α}	for level 0
$\alpha + \beta_1$	for level 1	$exp^{\alpha+\beta_1}$	for level 1
$\alpha + \beta_2$	for level 2	$exp^{\alpha+\beta_2}$	for level 2
$\alpha + \beta_3$	for level 3	$\ldots$etc.	

$\alpha + \beta_K$ for level K

OR (level **1** vs. level **0**) = $\frac{exp^{\alpha+\boldsymbol{\beta}_1}}{exp^{\alpha}} = exp^{\boldsymbol{\beta}_1}$

OR (level **i** vs. **j**) $= exp^{\boldsymbol{\beta}_i - \boldsymbol{\beta}_j}$

Note that the β associated with level 0 (i.e., reference group) is 0, or $\beta_0 = 0$.

Continuous X in a logistic model

If we have a continuous X in a logistic model, this assumes

odds (outcome) = $exp^{\alpha+\beta X}$

or the $\log_e$(odds) = $\alpha + \beta X$

i.e. the log odds is *linearly* related to X

β = change in log Odds per unit change in X

exp^{β} = OR for unit change in X.

Also: For a change of **2 units** OR = $exp^{\mathbf{2}\beta}$

For a change of **k units** OR = $exp^{\boldsymbol{k}\beta}$

Interpretation is simple,

But we should first check if the linear assumption is reasonable

Adjusted OR from logistic regression

Assuming a common OR relating Y to X in each stratum (e.g. for 3 strata)

ln(Odds) for stratum 1: $\alpha_1 + \beta X$
ln(Odds) for stratum 2: $\alpha_2 + \beta X$
ln(Odds) for stratum 3: $\alpha_3 + \beta X$

Different α allows the odds to be different in each stratum, but same β represents same OR for $X = 1$ vs. $X = 0$ *regardless of stratum*

Fit logistic model with X and a 3-category stratum variable as predictors: exp^{β} estimate is the Mantel-Haenszel OR!

To assess effect modification

In logistic regression, with binary exposure X and binary confounder Z, we include both as predictors to model:

$$\text{logit}(P[Y=1]) = \alpha + \beta_1 X + \beta_2 Z + \boldsymbol{\gamma} X * Z$$

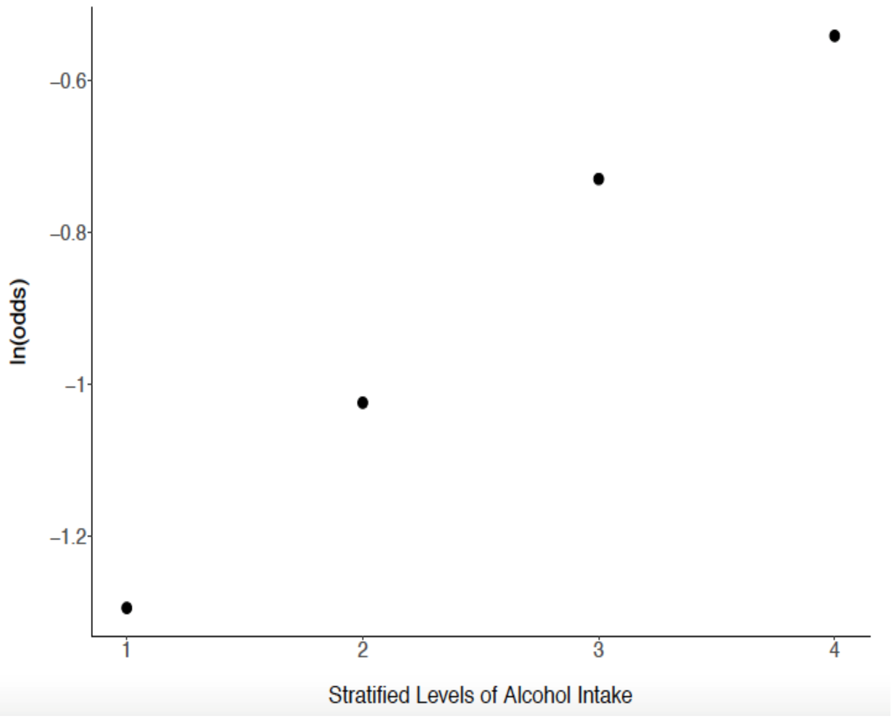
X	Z	odds	OR*
0	0	exp^{α}	OR_{00}=1= $exp^{\alpha}/exp^{\alpha}$
1	0	$exp^{\alpha+\beta_1}$	OR_{10}= exp^{β_1}
0	1	$exp^{\alpha+\beta_2}$	OR_{01}= exp^{β_2}
1	1	$exp^{\alpha+\beta_1+\beta_2+\gamma}$	OR_{11}= $exp^{\beta_1+\beta_2+\gamma}$

*: Reference group corresponds to $X=0$ and $Z=0$

When $\boldsymbol{\gamma}$=0 (no effect modification) $\text{OR}_{X=1 \text{ vs. } 0}$=$exp^{\beta_1}$for all Z

Recall: ln(odds) is linearly related to X in logistic model

		Lung Cancer		
		Yes	No	Odds
Alcohol intake	<1	1090	3976	.274
	1-3.9	806	2244	.359
	4-6.9	378	783	.482
	≥ 7	679	1166	.582
		2953	8169	



Reasonable to fit alcohol as continuous

If assumption of a linear trend is *not* reasonable

We classify the variable into categories/levels, and choose one of them as the "reference" and fit the effect of different levels as before:

log-odds	$= \alpha$	for level 0
	$= \alpha + \beta_1$	for level 1
	= ….	……
	$= \alpha + \beta_K$	for level K

This means we are modelling a different odds for each level (and not assuming that they follow a linear trend)

The exp^{β} values from the logistic regression are the ORs of each of the levels vs. the reference

Note: You must tell your software that the variable is a **factor !**

Categorization very common in medical research

Especially age groups

Even where there may be a linear trend!

(easier to communicate: OR of level=j vs. reference group)

BUT:

Where a linear trend is reasonable, and we only wish to adjust for the factor (i.e., we are not interested in the magnitude of its effect)

Then: model with linear trend has greater statistical power, especially if some categories have a small number of individuals.

Example of interpreting β coefficients

P is probability of disease (proportion with disease)

$\text{logit}(P) = \alpha + \beta_1 age + \beta_2 sex$

sex is coded 0 for M, 1 for F

age in years (continuous)

OR for F vs M for disease is exp^{β_2} *if both are the same age*

Note this assumes there is a common odds ratio in all age strata

(For categorical exposure and confounder, this is the MH odds ratio!)

exp^{β_1} is odds ratio per one year increase in age

(assuming this is common for males and female)

$\left(exp^{\beta_1}\right)^k = exp^{k\beta_1}$ is the OR for a change in age of 'k' years for individuals of the same sex.

More general logistic model

May have many explanatory variables, both exposure(s) and confounders (maybe frequency matched):

$$\ln(\text{odds}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

So odds = $(exp^{\alpha})(exp^{\beta_1})(exp^{\beta_2})\ldots(exp^{\beta_k})$

= (base odds) OR_1 OR_2 … OR_k

Model is multiplicative on the odds scale

From prospective to retrospective

For cohort or cross-sectional data, logistic model is a ”regression model” for binary outcomes in the sense that X’s can be fixed/chosen but Y random:

$\text{logit}(P[Y = 1]) = \alpha + \beta X$

Equivalent to $P[Y = 1] = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$

$P[Y = 1]$ when $X = 0$ (unexposed) = $\frac{e^{\alpha}}{1+e^{\alpha}}$

So we can estimate prevalence (in unexposed) from α

But for case-control data, we are modeling
$P[Y = 1|X]$ ***conditional on being sampled***

From prospective to retrospective

If probability of being sampled is π_1 for cases and π_0 for controls

Then using Bayes theorem:

$$\mathrm{P}[Y = 1|X, S = 1] = \frac{P[Y=1, S=1, X]}{P[S=1, X]}$$

$$= \frac{P[X]P[\,Y = 1 \mid X\,]P[S = 1 \mid X, Y = 1]}{P[X]P[\,Y = 1 \mid X\,]P[S = 1 \mid X, Y = 1] + P[X]\,P[Y = 0 \mid X]P[S = 1 \mid X, Y = 0]}$$

$$= \frac{P[Y = 1 \mid X]\pi_1}{P[Y = 1 \mid X]\pi_1 + P[Y = 0 \mid X]\pi_0}$$

$$= \frac{e^{\alpha^* + \beta X}}{1 + e^{\alpha^* + \beta X}} \quad \text{where } \alpha^* = \alpha + \ln\left(\frac{\pi_1}{\pi_0}\right)$$

From prospective to retrospective

We know that using 2-by-2 tables the exact same calculations can be used to make inferences on OR from cohort or case-control data.

Now, we see that when

$\text{logit}\{P(Y=1)\} = \alpha + \beta X$
► $\text{logit}\{P(Y=1|X, S=1)\} = \alpha^* + \beta X$

$$\alpha^* = \alpha + \ln\left(\frac{\pi_1}{\pi_0}\right)$$

where π_1 and π_0 are sampling fractions of cases and controls

If we have whole cohort, then $\alpha^* = \alpha$

Prentice & Pyke (1979, Biometrika): same β, α different

So OR has nice properties

Used in cohort studies as well as case-control studies

Logistic regression widely used and adjusted ORs reported

The reported OR often referred to as ”relative risk”: it is a good approximation in many settings when prevalence is low

It is possible to estimate adjusted RR (later in this course)