

Logistic Regression

February 25, 2014

Biostatistics 208

Outline

- I Introduction to binary outcomes in prospective and cross-sectional studies
 - A. definitions
 - B. measures of risk and models for the dependence of risk on predictors
 - C. assessment of associations in 2x2 tables

- II The logistic regression model for one predictor
 - A. motivation
 - B. definition
 - C. examples

- III Multiple logistic regression
 - A. fitting and interpreting models
 - B. confounding, interaction and mediation

- IV Advanced issues in model construction and interpretation
 - A. measurement scales, transformations, assessing linearity
 - B. risk prediction
 - C. model evaluation, outliers, goodness of fit

- V Alternatives to logistic regression
 - A. relative risk and risk difference models for binary outcomes
 - B. pooled logistic regression
 - C. exact logistic regression
 - D. regression for categorical outcomes with >2 levels

- VI Analysis of data from case-control studies

Binary outcomes

Many clinical and epidemiological studies generate responses which take on one of two possible values reflecting:

- Development (or not) of a response within a defined period of observation
 - CHD in men ages 39-59 between 1960 and 1969 (WCGS study)
- Presence/absence of a condition or characteristic at a particular time
 - HIV serostatus among female partners of HIV-infected males (CDC heterosexual transmission study)
- Binary indicator of a specified level of a measured quantity
 - baseline hypertension (SBP > 140) (HERS study)

Binary outcomes

Arise in a variety of study designs:

- Prospective (Incidence)
- Cross-sectional (Prevalence)
- Case-Control (Outcome-based sampling)

Binary outcomes

Example: WCGS study of first occurrence of coronary heart disease (CHD) in 1354 men, ages 39-59 in the period 1960 - 1969.

Study type: prospective cohort

Outcome: chd69, a binary indicator of occurrence of CHD in the period of follow-up.

Variables:

obs:	3,154			
vars:	14			

	storage	display	value	
variable name	type	format	label	variable label

id	float	%9.0g		id
age	float	%9.0g		age
height	float	%9.0g		height (inches)
weight	float	%9.0g		weight (lbs)
sbp	float	%9.0g		systolic BP
dbp	float	%9.0g		diastolic BP
chol	float	%9.0g		total cholesterol
behpat	float	%9.0g	bpat4	behavioral pattern (4 level)
ncigs	float	%9.0g		cigarettes smoked per day
dibpat	float	%9.0g	bpat2	behavioral pattern (A vs. B)
chd69	float	%9.0g	ind	CHD event
typchd69	float	%9.0g	chdtype	type of event
time169	float	%9.0g		follow-up time (days)
arcus	float	%9.0g	pamiss	arcus senilis

Binary outcomes

Example: CDC study of sexual HIV transmission among monogamous female partners of males infected with HIV via transfusion.

Study design: cross-sectional/retrospective

Outcome: hivp, HIV status of female partner at recruitment

Variables:

obs:	31			
vars:	5			

	storage	display	value	
variable name	type	format	label	variable label

hivp	float	%9.0g		female's HIV status (1=pos/0=neg)
cd4_200	float	%9.0g		CD4 in male < 200 (1=yes//0=no)
incumo	float	%9.0g		length (mo) AIDS incubation period in male
aids	float	%9.0g		AIDS diag. in male (1=yes/0=no)
contacts	float	%9.0g		number of unprotected contacts

Assessing Occurrence Risk for Binary Outcomes

Let y denote a typical binary outcome indicator variable:

$$y = \begin{cases} 1 & \text{yes} \\ 0 & \text{no} \end{cases}$$

- The outcome risk, or occurrence probability associated with a binary outcome is estimated by the proportion (P) of individuals in the sample with the outcome (i.e. those with $y=1$):

$$P = \Pr(y = 1) = \frac{1}{n} \sum_{i=1}^n y_i$$

- **Examples:**
 - CDC transmission study (prevalence)
 - $P = 7/31 = 0.226$
 - WCGS (incidence proportion)
 - $P = 257/3154 = 0.081$

Relating outcome risk to predictor variables

A common goal of studies of binary outcomes is to investigate the association of outcome risk with specified predictor variables (AKA “risk factors” or “exposures”).

Examples:

- CDC HIV transmission study
 - association between female serostatus and reported numbers of unprotected sexual contacts
- WCGS
 - association between CHD status and type A or B “behavior pattern”

Measures of association for binary outcomes & categorical predictors

Association measures:

- compare outcome risk in groups defined by predictors
- desirable numerical characteristics
- known statistical properties

For an outcome y and predictor x , both taking on two distinct values (1 for presence of the defining characteristic, 0 for absence), let $P(x)$ be the outcome risk associated with a given value of x :

$$P(x) = \Pr(y = 1 \mid x)$$

$P(x)$ is estimated by the proportion of individuals with the outcome among those with predictor equal to x .

Example from CDC study

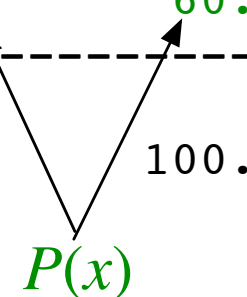
outcome: y = HIV status of female partner

predictor: x = indicator of AIDS diagnosis in male partner)

$P(x)$ = proportion HIV+ for those with partners having AIDS diag. x

```
. tabulate hivp aids, col
```

female's HIV status (1=pos/0=n eg)	AIDS diag. in male (1=yes/0=no)		Total
	0	1	
0	22 84.62	2 40.00	24 77.42
1	4 15.38	3 60.00	7 22.58
Total	26 100.00	5 100.00	31 100.00



$P(x)$

Measures of association for binary outcomes

Example:

- odds ratio (OR) :
$$\frac{\frac{P(1)}{1-P(1)}}{\frac{P(0)}{1-P(0)}} \quad 0 \leq OR < \infty$$
- magnitude of OR is unconstrained
- as a relative measure, may not always reflect "importance" of a predictor
- approximates the relative risk when outcomes are rare (to be discussed later for case-control studies)
- generalizes easily to the regression setting

Measures of association for binary outcomes

Example:

- relative risk (RR) : $P(1) / P(0)$ $0 \leq RR < \infty$
 - magnitude of RR constrained by value of $P(0)$ (e.g. for $P(0) = 0.5$, $0 \leq RR \leq (1 / 0.5)$)
 - as a relative measure, suffers same drawbacks as OR
 - doesn't generalize to regression models as easily as OR

Measures of association for binary outcomes

Example:

- risk difference (RD) : $P(1) - P(0)$ $-1 \leq RD < 1$
 - also called the “excess risk”
 - measures "absolute" risk; sometimes may better reflect "importance" than relative measures
 - doesn't generalize to regression models as easily as OR

Association between arcus senilis and CHD in the WCGS cohort

Use of the `cs` (cohort study) command in Stata with the `or` option:

```
. use wcfgs
. cs chd69 arcus, or
```

	arcus senilis		
	Exposed	Unexposed	Total
Cases	102	153	255
Noncases	839	2058	2897
Total	941	2211	3152
Risk	.1083953	.0691995	.080901
	Point estimate		[95% Conf. Interval]
Risk difference	.0391959		.0166915 .0617003
Risk ratio	1.566419		1.233865 1.988603
Attr. frac. ex.	.3616011		.1895387 .4971343
Attr. frac. pop	.1446404		
Odds ratio	1.63528		1.257732 2.126197 (Cornfield)

chi2(1) = 13.64 Pr>chi2 = 0.0002

- measures of risk are estimates of risk of CHD in the target population
 - odds ratio
 - relative risk (*risk ratio*)
 - risk difference (*excess risk*)
- two individuals omitted because of missing values of arcus

CDC HIV transmission study

Association between prevalent HIV status in female partner and a binary indicator of prior AIDS diagnosis in the male partner:

`. cs hivp aids, or exact`

	AIDS diag. in male [1=yes/0=no		
	Exposed	Unexposed	Total
Cases	3	4	7
Noncases	2	22	24
Total	5	26	31
Risk	.6	.1538462	.2258065
	Point estimate	[95% Conf. Interval]	
Risk difference	.4461538	-.0050928	.8974005
Risk ratio	3.9	1.233644	12.32933
Attr. frac. ex.	.7435897	.1893933	.9188926
Attr. frac. pop	.3186813		
Odds ratio	8.25	1.200901	57.1864 (Cornfield)
1-sided Fisher's exact P = 0.0619			
2-sided Fisher's exact P = 0.0619			
chi2(1) = 4.77 Pr>chi2 = 0.0289			

Note: Risk measures are based on prevalent HIV infection and are not necessarily reflective of corresponding estimates derived from a prospective study

- prevalence odds ratio (odds ratio)
- relative prevalence (risk ratio)
- prevalence difference (excess risk)
- attributable risk measures are not applicable in this setting

Why do we need regression models for binary outcomes?

Limitations of frequency table methods:

- awkward for large numbers of predictors
- not appropriate for continuous predictors unless grouping is imposed

Example: association between age and CHD in WCGS

```
. tabulate chd69 agec, col
```

CHD event	agec					Total
	35-40	41-45	46-50	51-55	56-60	
no	512	1,036	680	463	206	2,897
	94.29	94.96	90.67	87.69	85.12	91.85
yes	31	55	70	65	36	257
	5.71	5.04	9.33	12.31	14.88	8.15
Total	543	1,091	750	528	242	3,154
	100.00	100.00	100.00	100.00	100.00	100.00

Example: association between age and CHD in WCGS

```
. tabodds chd69 agec, or
```

agec	Odds Ratio	chi2	P>chi2	[95% Conf. Interval]	
35-40	1.000000	.	.	.	.
41-45	0.876822	0.32	0.5692	0.557454	1.379156
46-50	1.700190	5.74	0.0166	1.095789	2.637958
51-55	2.318679	14.28	0.0002	1.479779	3.633160
56-60	2.886314	18.00	0.0000	1.728069	4.820876

Test of homogeneity (equal odds): chi2(4) = 46.64
Pr>chi2 = 0.0000

Score test for trend of odds: chi2(1) = 40.76
Pr>chi2 = 0.0000

- 4 odds ratios are necessary to summarize relationship
- results indicate increase in risk with age, but do not reveal much about the nature of the change
- alternate age groupings may reveal different patterns
- accompanying tests evaluate hypotheses addressing changes in odds of CHD between age groups:
 - *homogeneity*: no difference in odds of CHD between age groups
 - *trend*: no linear association between log odds of CHD and age

Review of logarithms and exponentials for binary regression models

(1) Log of the ratio of odds or risk for two groups distinguished by different values of the predictor ($x = 0$ or $x = 1$):

$$\log \left[\frac{P(1)/(1 - P(1))}{P(0)/(1 - P(0))} \right] = \log[P(1)/(1 - P(1))] - \log[P(0)/(1 - P(0))]$$

$$\log \left[\frac{P(1)}{P(0)} \right] = \log[P(1)] - \log[P(0)]$$

(2) Exponential of the log ratio of odds or risk for two groups distinguished by different values of the predictor:

$$\exp \left(\log \left[\frac{P(1)/(1 - P(1))}{P(0)/(1 - P(0))} \right] \right) = \frac{P(1)/(1 - P(1))}{P(0)/(1 - P(0))}$$

$$\exp \left(\log \left[\frac{P(1)}{P(0)} \right] \right) = \frac{P(1)}{P(0)}$$

Regression models for binary outcomes

Goal: similar to the linear model, we want to find a model linking the average of the outcome variable y to a predictor x .

Since y is binary, its average among a group of individuals characterized by a specified value of x is just the proportion of individuals with the outcome ($y = 1$) in the group:

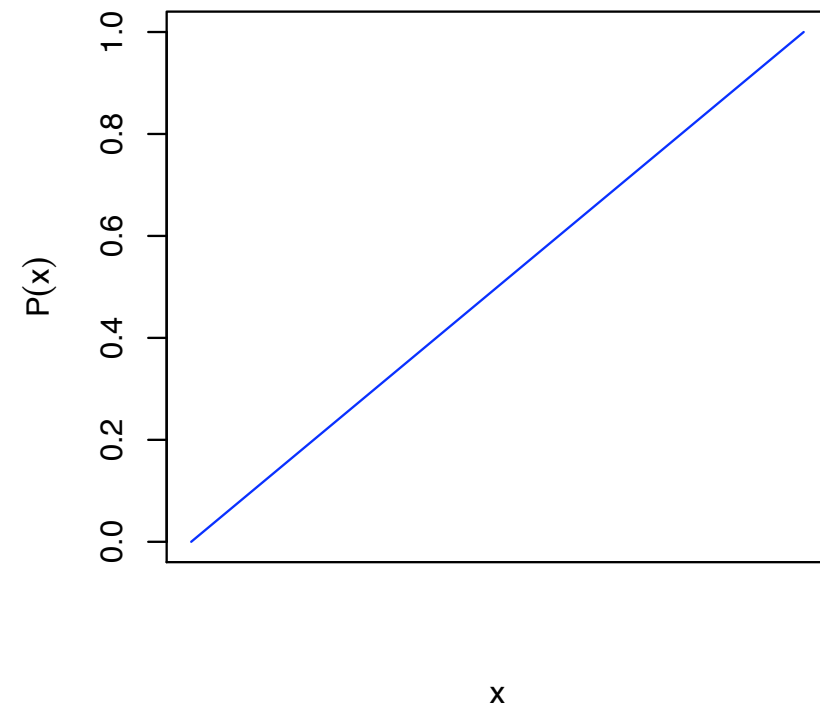
$$E(y \mid x) = P(x) = \frac{\# \text{ with } y = 1}{\# \text{ in group}}$$

Differences with standard linear regression model:

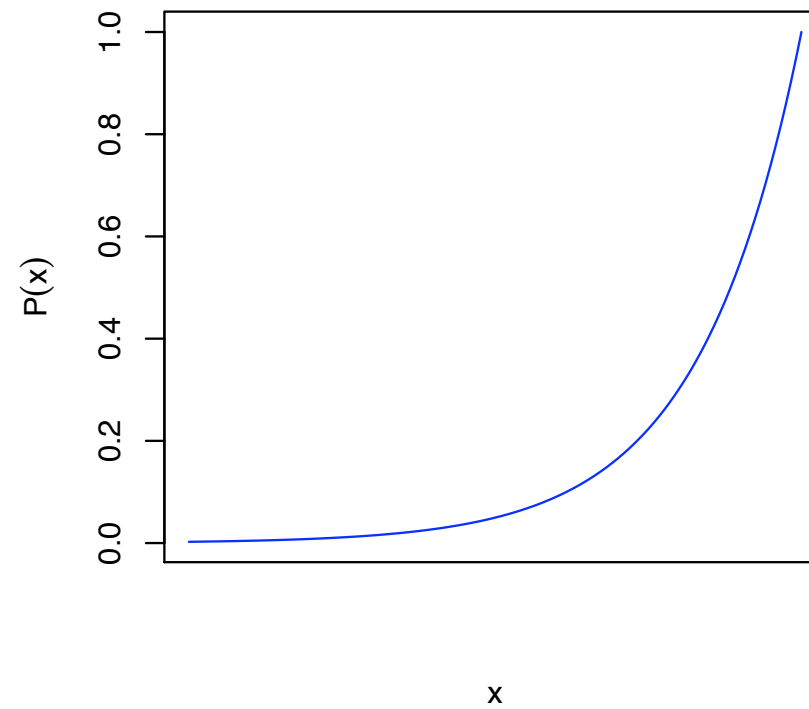
- y does not follow a normal distribution
- response $P(x)$ represents the probability of outcome occurrence, and is constrained to be between 0 and 1 for all values of x

Possible models for the relationship between $P(x)$ and x

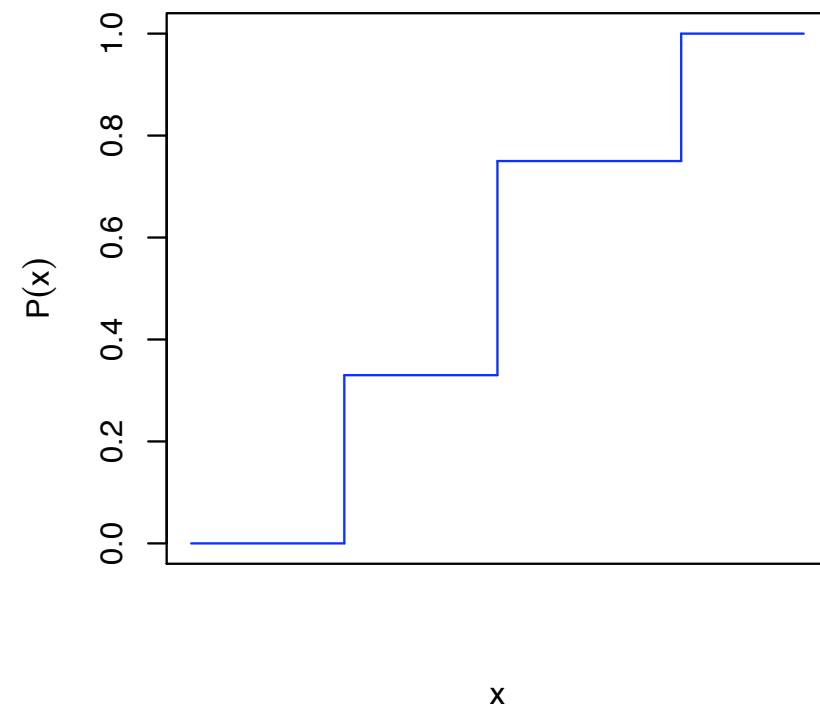
(A) Linear



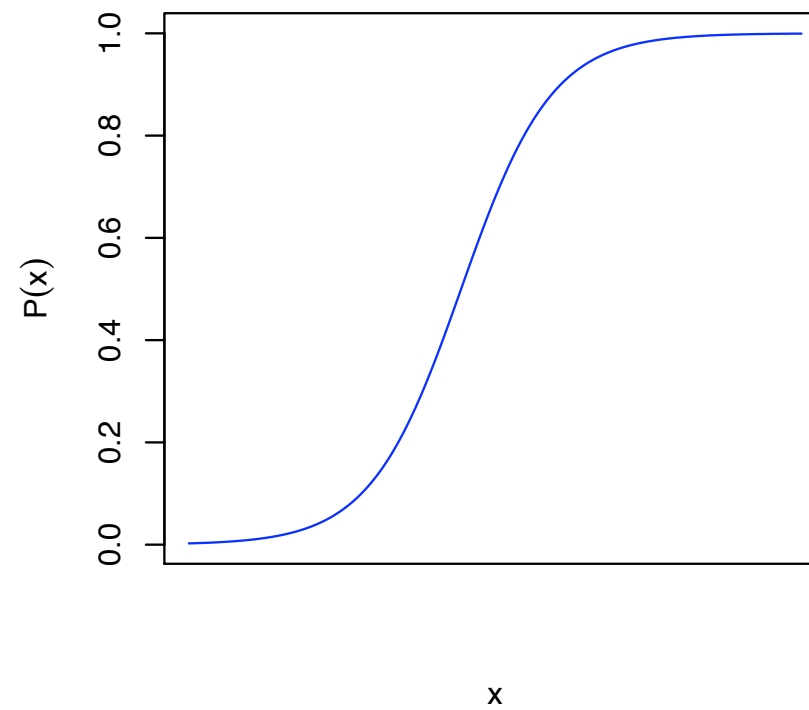
(B) Log linear



(C) Step



(D) Logistic



Linear model for the relationship between $P(x)$ and x

$$P(x) = \beta_0 + \beta_1 x$$

Features:

- β_0 gives the outcome risk ($P(0)$) for individuals with $x = 0$
- β_1 gives the change in $P(x)$ for each unit change in x .
- This is the *risk difference* associated with a unit change in x .

Limitations:

- response has to be constrained to be between 0 and 1 for all values of x
- assumption that $P(x)$ changes linearly throughout the range of x is frequently implausible

Log-linear model for the relationship between $P(x)$ and x

$$\log[P(x)] = \beta_0 + \beta_1 x$$

Features:

- β_0 gives the log of the outcome risk ($P(0)$) for individuals with $x = 0$
- β_1 gives the change in $\log P(x)$ for each unit change in x .
- This is the *log relative risk* associated with a unit change in x .
- $\log$ = natural logarithm (i.e. log with base e)

Limitations:

- response ($\log P(x)$) has to be constrained to be between $-\infty$ and 0 for all values of x

Logistic model for the relationship between $P(x)$ and x

$$\log\left[\frac{P(x)}{1-P(x)}\right] = \beta_0 + \beta_1 x$$

Features:

- The log-odds of the outcome is linear in x , with intercept β_0 and slope β_1
- β_0 gives the log odds of the outcome for individuals with $x = 0$
- β_1 gives the change in log odds for each unit change in x .
- This is the *log odds ratio* associated with a unit change in x .
- Response probability ($P(x)$) is naturally constrained to be between 0 and 1 for

Nonlinear form of the logistic regression model

$$P(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Note:

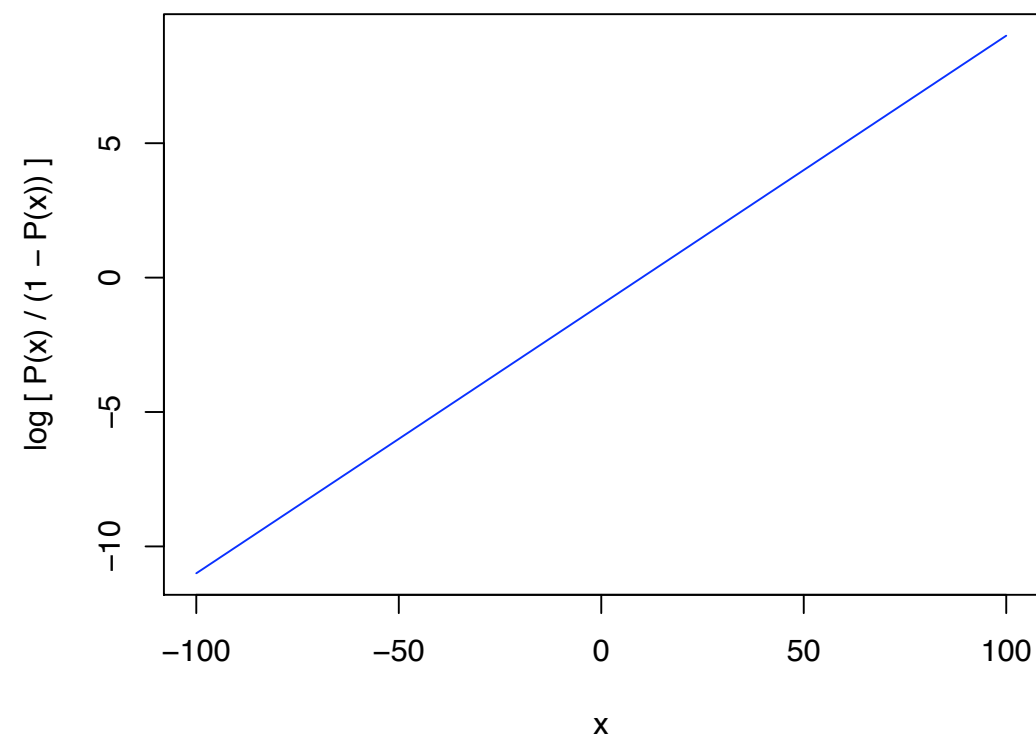
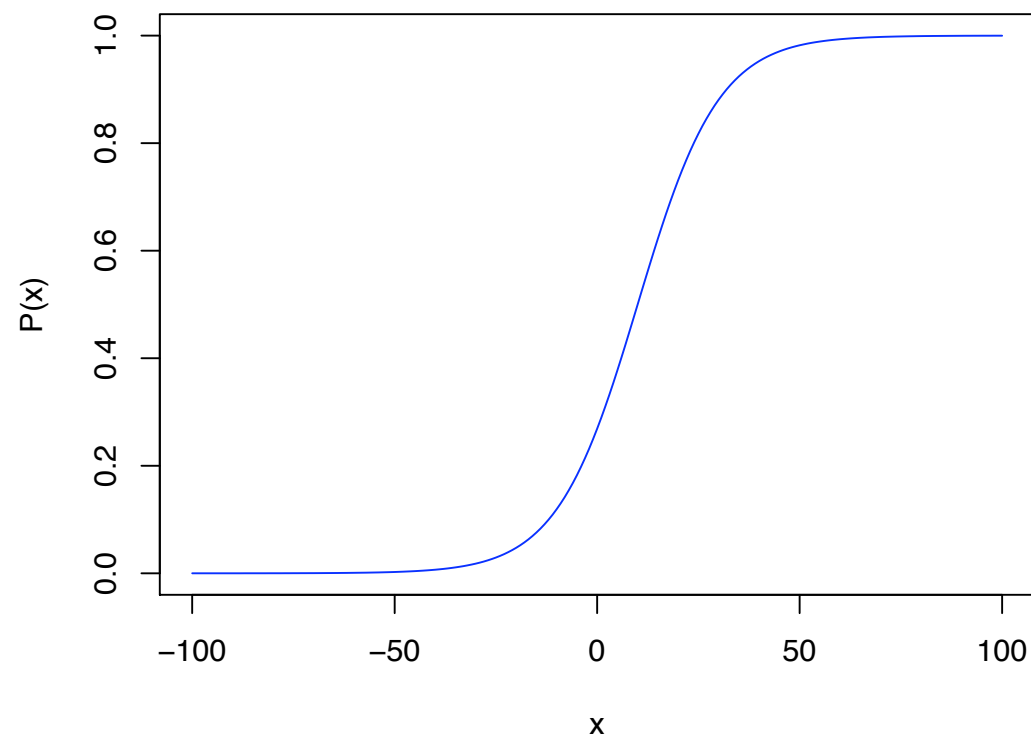
- Alternate representation of the log-linear form of the model (previous slide)
- Notice that $P(x)$ can't be larger than 1 or smaller than 0 for any choice of x

The logistic regression model for a single predictor

Equivalent nonlinear and linear representations:

$$P(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$\log \left[\frac{P(x)}{1 - P(x)} \right] = \beta_0 + \beta_1 x$$



Features of the logistic model

Primary assumptions:

- model provides an appropriate representation for the dependence of risk on predictor(s) (i.e. linearity in log odds)
- individual outcomes are independent

Advantages:

- can be applied to a wide range of study designs (including case-control)
- based on the odds ratio measure of association, which is easily interpretable and valid for case-control studies
- has practical & numerical advantages over (constrained) alternatives based on *RD* or *RR*

Disadvantages:

- based on the odds ratio (a *non-collapsible* ratio measure) that has some interpretation issues (and not universally favored by epidemiologists)

Logistic regression for a single binary predictor

$$\log\left[\frac{P(x)}{1-P(x)}\right] = \beta_0 + \beta_1 x \qquad P(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Convention for coding levels of a binary predictor x :

$x = 1$ for individuals having the predictor

$x = 0$ for individuals not having the predictor

- β_0 is the log odds of the outcome for $x = 0$.
- β_1 is the log odds ratio comparing those with the predictor ($x = 1$) to those without ($x = 0$):

$$\begin{aligned} \log\left[\frac{\frac{P(1)}{1-P(1)}}{\frac{P(0)}{1-P(0)}}\right] &= \log\left[\frac{P(1)}{1-P(1)}\right] - \log\left[\frac{P(0)}{1-P(0)}\right] \\ &= (\beta_0 + \beta_1) - \beta_0 \\ &= \beta_1 \end{aligned}$$

- $\exp(\beta_1)$ is therefore the *odds ratio* comparing those with the predictor ($x = 1$) to those without ($x = 0$)

Logistic regression for a single binary predictor

multiplicative on the odds scale:

- $\exp(\beta_1)$ is the multiplicative increase in odds of outcome occurrence for individuals with $x = 1$ compared to those with $x = 0$:

$$\frac{P(1)}{1 - P(1)} = \left[\frac{P(0)}{1 - P(0)} \right] \times e^{\beta_1} = e^{\beta_0} \times e^{\beta_1}$$

additive on the log odds scale:

- β_1 is the additive increase in the log odds of outcome occurrence for individuals with $x = 1$ compared to those with $x = 0$:

$$\log \left[\frac{P(1)}{1 - P(1)} \right] = \log \left[\frac{P(0)}{1 - P(0)} \right] + \beta_1 = \beta_0 + \beta_1$$

Note: the relative risk regression model is multiplicative on the risk (P) scale, while the risk difference model is inherently additive

logistic command to investigate CHD and corneal arcus association in WCGS

```
.logistic chd69 arcus
```

```
Logistic regression
```

```
Number of obs      =          3152
```

```
LR chi2(1)         =          12.98
```

```
Prob > chi2        =          0.0003
```

```
Pseudo R2         =          0.0073
```

```
Log likelihood = -879.10783
```

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
arcus	1.63528	.2195035	3.66	0.000	1.257	2.127399
_cons	.074344	.0062298	-31.02	0.000	.0630839	.0876141

- *exponentiated coefficients are displayed by default* (i.e. option `or` is standard)
- results are identical to previous analysis based on the 2×2 table (slide #15)
 - `arcus` (β_1) : log odds ratio comparing CHD odds in those with arcus to the corresponding odds in those without arcus
 - `_cons` (β_0) : log odds of outcome occurrence in individuals with `arcus = 0`

logistic command to investigate CHD and corneal arcus association in WCGS

```
.logistic chd69 arcus
```

```
Logistic regression                                Number of obs   =           3152
                                                    LR  chi2(1)     =           12.98
                                                    Prob > chi2     =           0.0003
Log likelihood = -879.10783                        Pseudo R2      =           0.0073
```

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
arcus	1.63528	.2195035	3.66	0.000	1.257	2.127399
_cons	.074344	.0062298	-31.02	0.000	.0630839	.0876141

- the z (Wald) test statistic in the logistic results is the ratio of the estimated regression coefficient (β_1) for arcus to its standard error, and follows (approximately) a standard normal distribution
- the log-likelihood is a measure of support of the data for the model (the larger the likelihood and/or log-likelihood, the better the support).
- the statistic "chi2" is the likelihood ratio statistic for comparing this model including arcus to the simpler one (presented below) containing no predictors. chi2 is derived from the difference between the log-likelihoods (LL) from two models:

$$2 \times [(\text{LL from more complex model}) - (\text{LL from simpler model (on slide 33)})] =$$
$$2 \times [(-879.10783) - (-885.6)] = 2 \times (6.49217) = 12.98434$$

This statistic has a chi-squared distribution with one degree of freedom ($p = 0.0003$)

- Note** the similarity between the default LL test and the corresponding *F* test from linear regression

Using coef option to display regression coefficients

```
. logistic chd69 arcus, coef
```

```
Logistic regression
```

```
Number of obs = 3152
```

```
LR chi2(1) = 12.98
```

```
Prob > chi2 = 0.0003
```

```
Log likelihood = -879.10783
```

```
Pseudo R2 = 0.0073
```

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
arcus	.4918141	.1342299	3.66	0.000	.2287283	.7548999
_cons	-2.599052	.0837965	-31.02	0.000	-2.76329	-2.434814

- coef option prints untransformed regression coefficients:
 - $\beta_0 = -2.599052$ (log outcome odds for indiv. w/o arcus),
 - $\beta_1 = 0.4918141$ (log odds ratio for the effect of arcus on CHD)
- confidence interval for odds ratio ($\exp(\beta_1)$) for arcus can be obtained by exponentiating the limits of the CI for log-odds ratio (β_1)

```
. display exp(.2287283)
```

```
1.2570005
```

```
. display exp(.7548999)
```

```
2.1273986
```

Logistic regression model equations for previous example

$$P(\text{arcus}) = \frac{\exp(-2.599052 + 0.4918141 \times \text{arcus})}{1 + \exp(-2.599052 + 0.4918141 \times \text{arcus})}$$

$$\log \frac{P(\text{arcus})}{1 - P(\text{arcus})} = -2.599052 + 0.4918141 \times \text{arcus}$$

Logistic model for CHD including no predictors

Fit model including no predictors, excluding indiv. with missing values of arcus:

```
. logistic chd69 if arcus!=., coef
Logistic regression
Log likelihood = -885.6
Number of obs = 3152
LR chi2(0) = -0.00
Prob > chi2 = .
Pseudo R2 = -0.0000
```

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
_cons	-2.430167	.0653204	-37.20	0.000	-2.558193 -2.302142

- this model is only of interest for comparison of likelihoods to the previous model including arcus (We'll discuss likelihood ratio testing soon)
- intercept coefficient "_cons" is the log odds of CHD in the sample
- From the logistic equation:
 - $P = \exp(-2.430167) / [1 + \exp(-2.430167)] = 0.0815 =$ proportion with CHD in sample (excluding those with missing arcus)
- z statistic and Wald test in this case address the hypothesis that the actual odds of the outcome are 1.0 (i.e. that the probability of the outcome is 0.5)

Logistic regression for a single continuous predictor

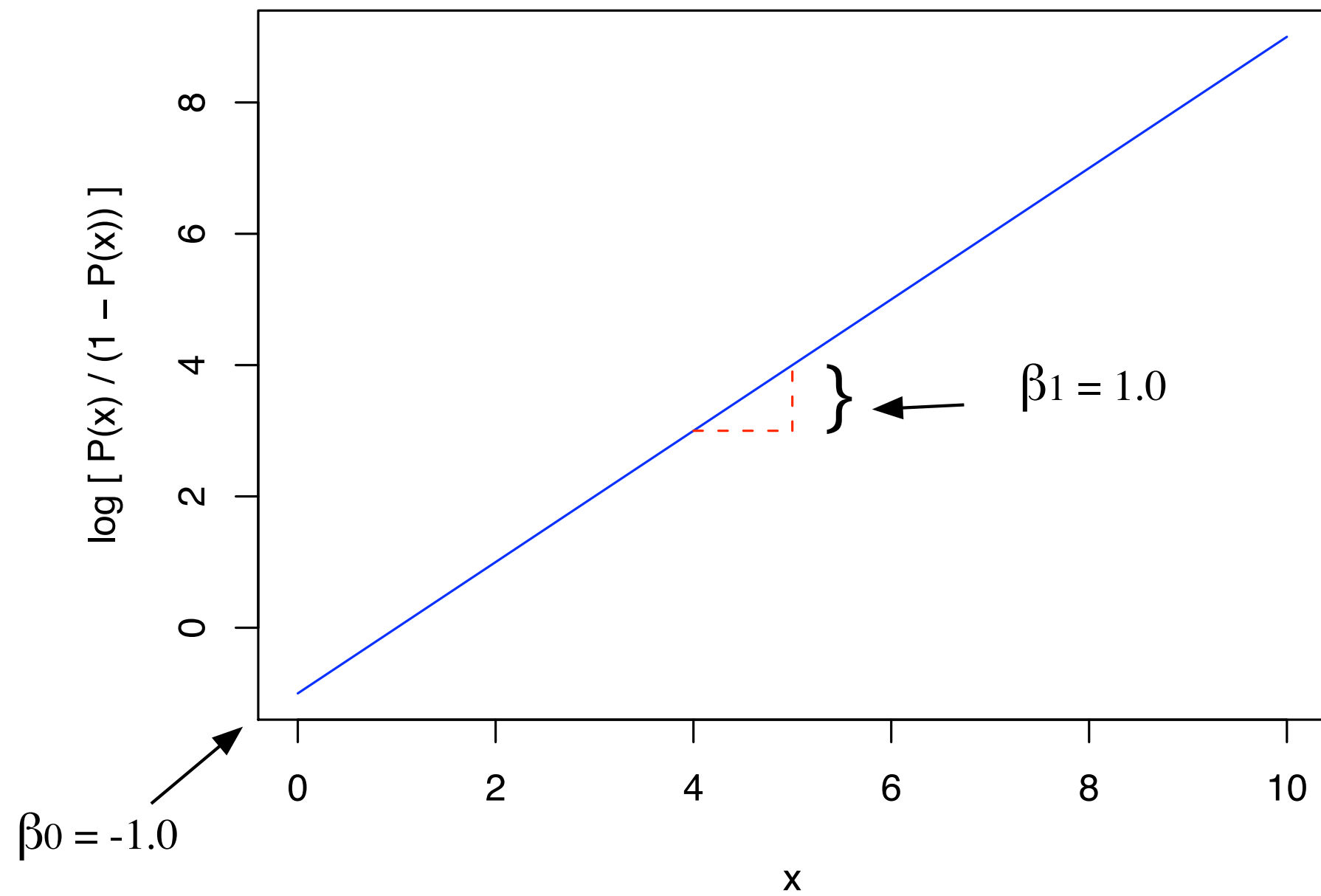
$$\log\left[\frac{P(x)}{1-P(x)}\right] = \beta_0 + \beta_1 x \qquad P(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

- β_0 is the log odds of the outcome for $x = 0$.
- β_1 is the change in the log odds for each unit increase in x :

$$\log\left[\frac{P(x+1)}{1+P(x+1)}\right] - \log\left[\frac{P(x)}{1+P(x)}\right] = \log\left[\frac{\frac{P(x+1)}{1+P(x+1)}}{\frac{P(x)}{1+P(x)}}\right] = \beta_1$$

- $\exp(\beta_1)$ is the *odds ratio* for each unit increase in x
- $\exp(\beta_1)$ is the multiplicative increase in odds associated with a one unit increase in x

Example: $\log\{P(x)/[1 - P(x)]\} = -1.0 + 1.0x$



Example: relationship between log odds of CHD and serum cholesterol in WCGS

```
. logistic chd69 chol, coef
```

Logistic regression

Number of obs = 3142

LR chi2(1) = 76.78

Prob > chi2 = 0.0000

Pseudo R2 = 0.0432

Log likelihood = -851.21151

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
chol	.0124478	.0014385	8.65	0.000	.0096284	.0152672
_cons	-5.359022	.3594727	-14.91	0.000	-6.063575	-4.654468

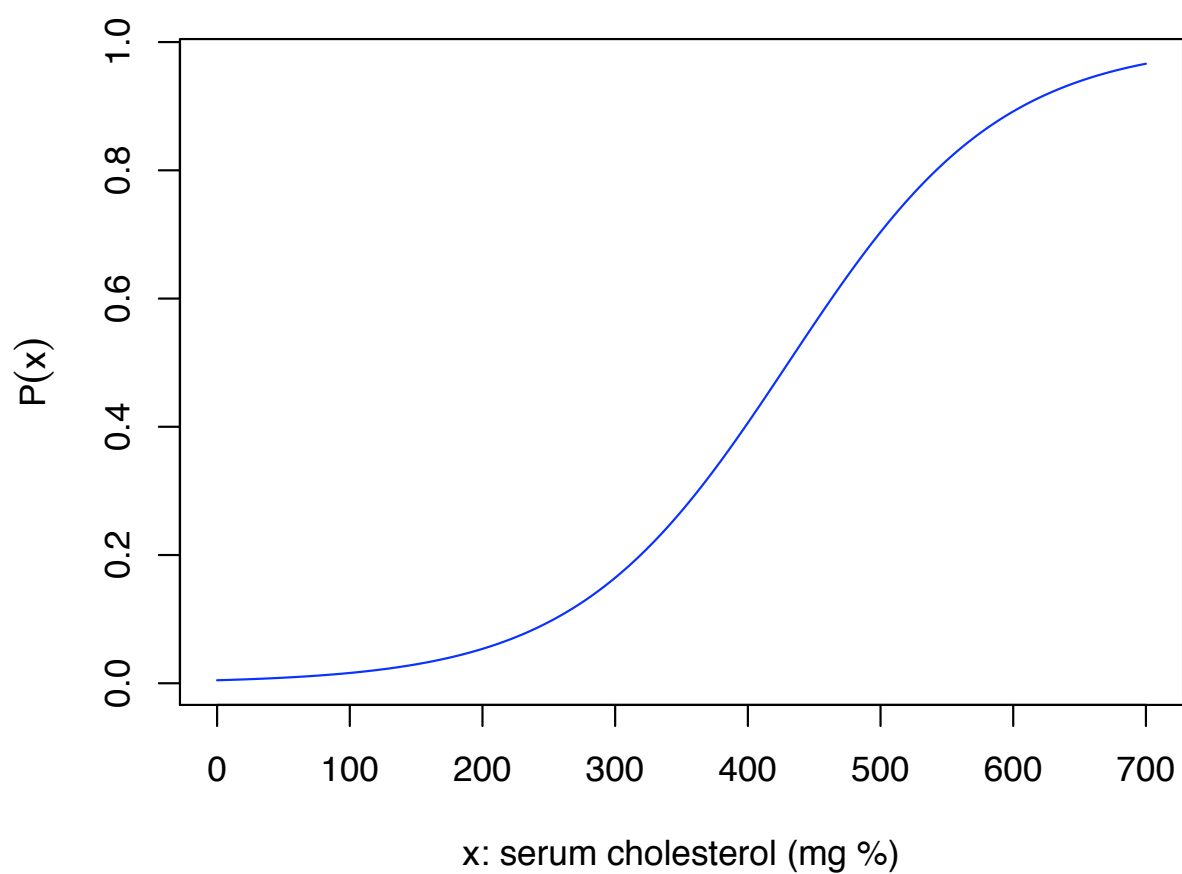
- estimated log odds ratio for each unit increase in cholesterol: 0.0124478
- estimated log odds ratio for a 100-unit increase in cholesterol: 1.24478
- corresponding log odds ratio computed from the definition of the logistic model:

$$\log \left[\frac{\frac{P(100)}{1-P(100)}}{\frac{P(0)}{1-P(0)}} \right] = \log \left[\frac{P(100)}{1-P(100)} \right] - \log \left[\frac{P(0)}{1-P(0)} \right] =$$
$$(-5.359022 + 0.0124478 \times 100) - (-5.359022) =$$
$$0.0124478 \times 100 = 1.24468$$

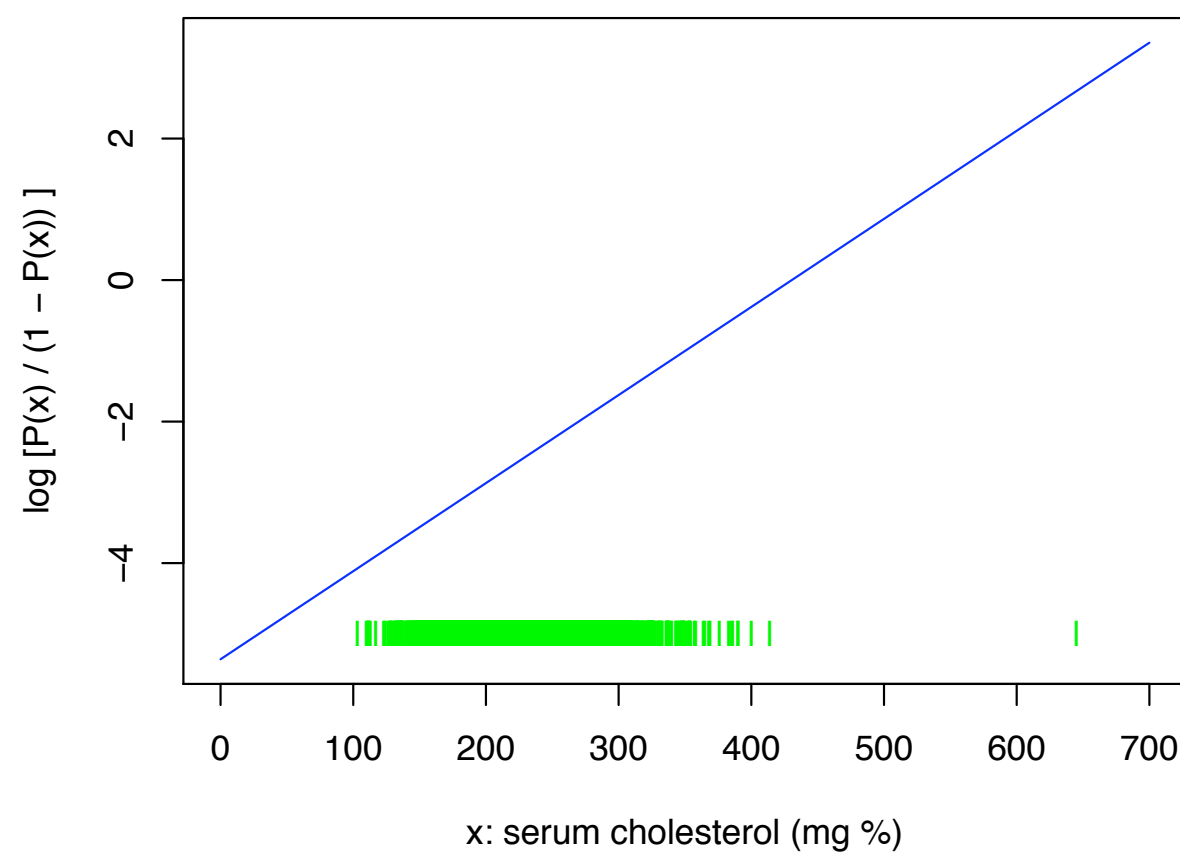
- the model is only interpretable for the range of cholesterol values present in the data
- the model assumes that the change in log odds for any unit change in cholesterol is independent of cholesterol level (linearity)
- 12 individuals have missing cholesterol values and are excluded

Plots of fitted logistic model for the probability of CHD as a function of serum cholesterol

Probability scale



Log odds scale



Example: relationship between CHD and log cholesterol level

```
. logistic chd69 lchol, coef
```

Logistic regression

Number of obs = 3141

LR chi2(1) = 75.59

Prob > chi2 = 0.0000

Log likelihood = -849.30099

Pseudo R2 = 0.0426

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lchol	3.117886	.3700441	8.43	0.000	2.392612	3.843159
_cons	-19.4107	2.032859	-9.55	0.000	-23.39503	-15.42637

- estimated odds ratio for a one log increase in cholesterol:

```
. dis exp(_b[lchol])
22.598545
```

- estimated log odds ratio for a 20% increase in cholesterol:

```
. nlcom _b[lchol]*log(1.2)
```

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_nl_1	.5684577	.067467	8.43	0.000	.4362248	.7006907

- estimated odds ratio for a 20% increase in cholesterol:

```
. nlcom exp(_b[lchol]*log(1.2))
```

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_nl_1	1.765542	.1191159	14.82	0.000	1.532079	1.999005

Logistic model for a multi-level categorical predictor

Example: type A / type B behavior pattern indicator in WCGS

- reported behavior pattern is classified in a single 4-level variable x as being either $A1, A2, B1$ or $B2$.

Three indicator variables are necessary to represent behavior pattern in a logistic model:

<i>Indicator variable</i>			
Level	$I(x=A2)$	$I(x=B1)$	$I(x=B2)$
$A1$	0	0	0
$A2$	1	0	0
$B1$	0	1	0
$B2$	0	0	1

- group $A1$ is chosen as the "baseline" level
- resulting logistic model has 3 indicator variables and 3 regression coefficients to represent the 4-level predictor (see next page)

$$\log \frac{P(x)}{1-P(x)} = \beta_0 + \beta_1 \times I(x = A2) + \beta_2 \times I(x = B1) + \beta_3 \times I(x = B2)$$

Example: relationship between CHD and behavior pattern

```
. logistic chd69 i.behpat
```

Logistic regression

Log likelihood = -870.16237

Number of obs = 3154

LR chi2(3) = 40.92

Prob > chi2 = 0.0000

Pseudo R2 = 0.0230

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
behpat						
2	.9807986	.2085511	-0.09	0.927	.6465252	1.487902
3	.4119481	.0964927	-3.79	0.000	.2602922	.6519642
4	.4241692	.1315496	-2.77	0.006	.2309681	.7789799
_cons	.1282051	.0248622	-10.59	0.000	.087667	.1874884

- likelihood ratio statistic tests difference between this model for behavior pattern (represented by 3 indicator variables) and corresponding model with no predictors
- odds ratio for pattern 2 (A2) not signif. different from pattern 1 (A1, baseline)
- odds ratios for patterns 3 & 4 (B1 & B2) very similar
- suggests possibly grouping 1 & 2 and 3 & 4

Example: relationship between CHD and behavior pattern

The variable `dibpat` is a binary indicator of A-type behavior patterns:

```
. logistic chd69 dibpat
```

Logistic regression	Number of obs	=	3154
	LR chi2(1)	=	40.90
	Prob > chi2	=	0.0000
Log likelihood = -870.17212	Pseudo R2	=	0.0230

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
dibpat	2.372929	.3326993	6.16	0.000	1.802773	3.123406
_cons	.0531629	.0061382	-25.41	0.000	.0423963	.0666636

- odds ratio comparing odds of CHD for individuals reporting type A behavior to those reporting type B behavior: 2.37 (95% CI: 1.80, 3.12)
- comparison of log-likelihood with previous model confirms that the binary representation of behavior pattern provides a reasonable description of the effect of behavior (we'll use a *likelihood ratio* test in a subsequent lab to formally justify that the binary version works as well)

Logistic regression for multiple predictors

Example: For two predictors x_1 and x_2 and assuming no interaction, the logistic model is defined:

$$\log \frac{P(x_1, x_2)}{1 - P(x_1, x_2)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Interpretation:

- β_1 is the change in log-odds of the outcome for every unit increase in x_1 holding x_2 fixed
 - β_1 is said to be *adjusted* for x_2
 - the fact that the effect of x_1 does not depend on the level of x_2 implies that *no interaction* between these variables is assumed

Example: For two predictors x_1 and x_2 , and allowing for *multiplicative interaction*, the logistic model is defined:

$$\log \frac{P(x_1, x_2)}{1 - P(x_1, x_2)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- in this case, the effect of x_1 and x_2 cannot be separately estimated

Maximum likelihood estimation for binary outcomes

Definition: for a sample of observations of a binary outcome, the *likelihood* of the sample is the probability of the observed outcomes, expressed as a function of any unknown parameter(s) in the chosen risk model.

- For independent outcomes, the likelihood is the product of the individual outcome probabilities specified by the assumed risk model.
- The *maximum likelihood estimate* of the parameter(s) is the most likely value for the parameter given the data in the sample (i.e. the value that yields the maximum value of the likelihood).

known quantities: outcomes (and predictors if applicable)

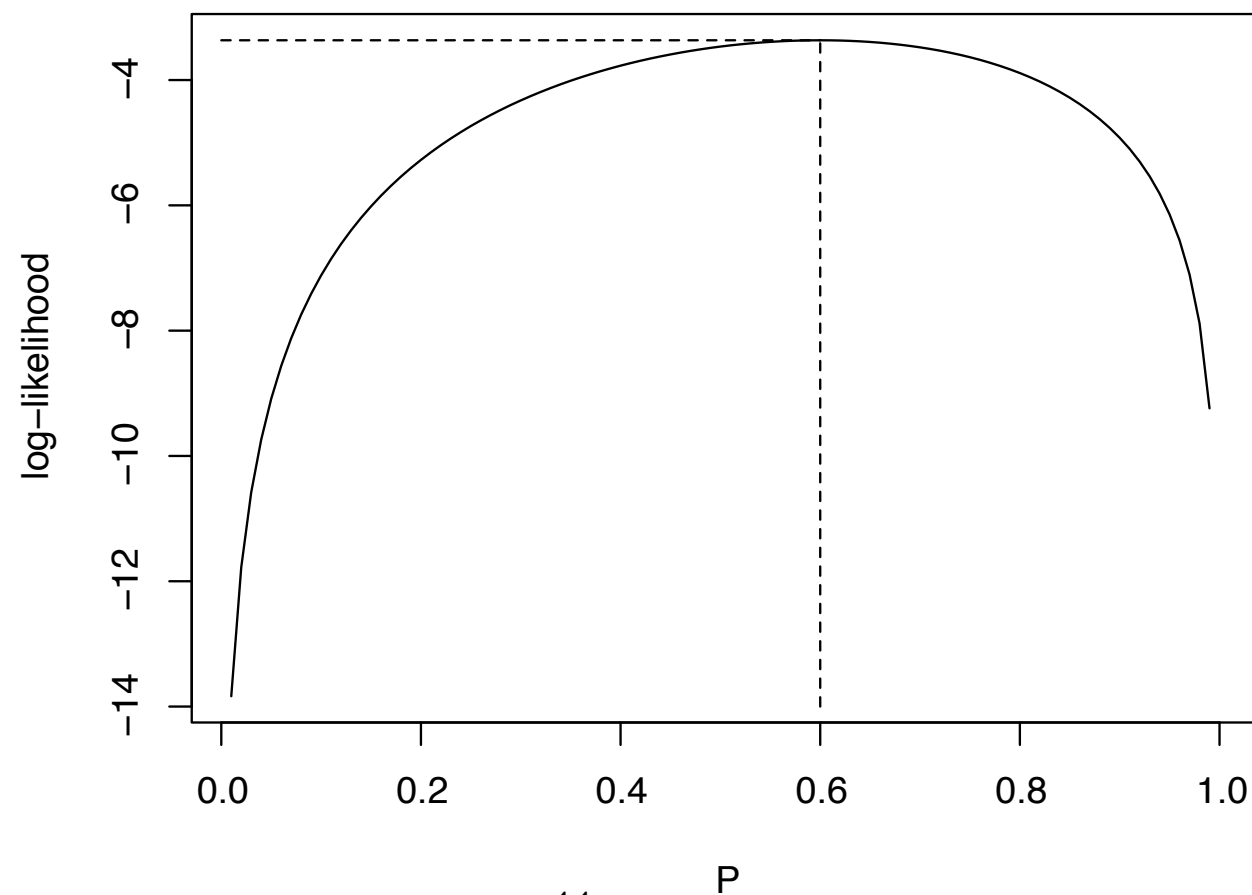
unknown quantities: parameter(s) of interest specified by the model

Maximum likelihood estimation for binary outcomes

Example:

likelihood for the probability of heads in 5 tosses of a biased coin

- **problem:** estimate the probability of throwing a head given 5 tosses
- **model:** 5 indep. observations (tosses) with unknown prob. P of throwing head (H)
- **observed data:** H T H H T
- **likelihood:** $P \times (1-P) \times P \times P \times (1-P) = P^3 \times (1-P)^2$
- **maximum likelihood estimate:** # heads / # tosses = $3/5 = 0.6$



Maximum likelihood estimation for binary outcomes

Example: dependence of CHD risk on a binary predictor in WCGS

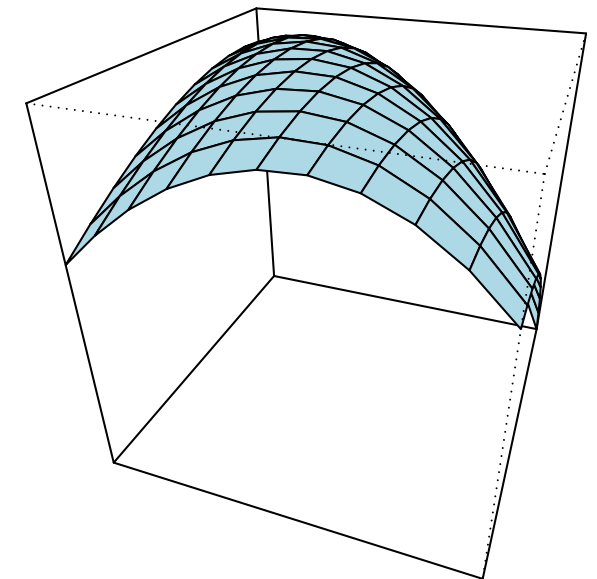
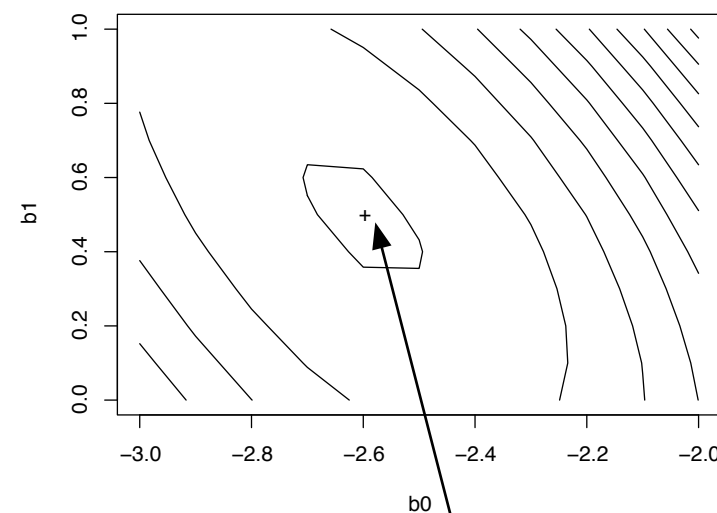
- **problem:** estimate the dependence of CHD risk on a binary indicator of arcus (x)
- **model:** 3154 indep. observations of CHD with prob. P given by the logistic model

$$P(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

- **observed data:**

	chd69(y)	arcus(x)
1.	No CHD ($x=0$)	0
2.	No CHD ($x=1$)	1
3.	No CHD ($x=0$)	0
4.	No CHD ($x=0$)	0
5.	CHD ($x=1$)	1
6.	No CHD ($x=0$)	0
7.	No CHD ($x=0$)	0
	.	.
	.	.
	.	.

- **contours of log-likelihood function**



- **likelihood:** $(1-P(0)) \times (1-P(1)) \times (1-P(0)) \times (1-P(0)) \times P(1) \times (1-P(0)) \times (1-P(0)) \dots$

- **maximum likelihood estimates:** $\beta_0 = -2.59905; \beta_1 = 0.49181$