

Logistic Regression

February 25, 2020

Biostatistics 208

Outline

- I Review of binary outcomes in prospective and cross-sectional studies
 - A. measures of risk and models for the dependence of risk on predictors
 - B. assessment of associations in 2x2 tables
- II The logistic regression model for one predictor
 - A. motivation, definition & examples
 - B. smoothing for binary outcomes
 - C. maximum likelihood estimation
- III Multiple logistic regression
 - A. fitting and interpreting models
 - B. casual inference: confounding, interaction & mediation
- IV Advanced issues in model construction and interpretation
 - A. measurement scales, transformations, assessing linearity
 - B. risk prediction models
 - 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 outcomes which take on one of two possible values reflecting:

- Occurrence (or not) of an outcome 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			

variable name	storage type	display format	value 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: 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 Probability/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 probability to predictor variables

A common goal of studies of binary outcomes is to investigate the association of outcome probability/risk with specified predictor variables.

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 probability/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 probability 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

$P(0) \times 100$

$P(1) \times 100$

Measures of association for binary outcomes

Example:

- risk difference (RD) : $P(1) - P(0)$ $-1 \leq RD \leq 1$
 - constrained in the range $[-1, 1]$
 - also called the “excess risk”
 - measures "absolute" risk; sometimes may better reflect "importance" than alternative relative measures
 - regression models based on RD may have estimation issues

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 $P(0) = 0.5$, $0 \leq RR \leq (1 / 0.5)$)
 - asymmetric with respect to outcome definition
 - as a relative measure, doesn't reflect absolute risk
 - regression models based on RR may have estimation issues

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, does not address absolute risk (may not always reflect "importance" of a predictor)
- symmetric with respect to outcome definition (i.e. OR for reversed outcome categories is the inverse of the OR for original categories)
- approximates the relative risk when outcomes are rare (to be discussed later for case-control studies)
- generalizes easily to the regression setting

Association between arcus senilis and CHD in the WCGS cohort

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

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

	arcus senilis		
	Exposed	Unexposed	Total
Cases	102	153	255
Noncases	839	2058	2897
Total	941	2211	3152
Risk	<i>P(1)</i> → .1083953	→ .0691995 <i>P(0)</i>	.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			

- $P(1)$ & $P(0)$ estimate 10-year cumulative incidence of CHD among those with/without arcus
- Association measures include:
 - 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			

- $P(1)$ & $P(0)$ estimate prevalence of HIV infection among those with/without AIDS dx
- Association measures include:
 - prevalence odds ratio
 - relative prevalence (prevalence risk ratio)
 - prevalence difference (excess prevalence)

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 categorization is imposed

Example: association between age and CHD in WCGS

```
. recode age 35/40=0 41/45=1 46/50=2 51/55=3 56/60=4, g(agec)
. label define agelab 0 "35-40" 1 "41-45" 2 "46-50" 3 "51-55" 4 "56-60"
. label val agec agelab
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
- 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*: score test of 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 (risk) for two groups distinguished by different values of a predictor ($x = 0$ or $x = 1$) equals diff. between log odds (risk) in the two groups:

$$\text{Log odds ratio} \quad \log \left[\frac{P(1)/(1 - P(1))}{P(0)/(1 - P(0))} \right] = \log[P(1)/(1 - P(1))] - \log[P(0)/(1 - P(0))]$$

$$\text{Log risk ratio} \quad \log \left[\frac{P(1)}{P(0)} \right] = \log[P(1)] - \log[P(0)]$$

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

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

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

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 predictor(s) 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 with predictor value x :

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

Similarities with standard linear regression model:

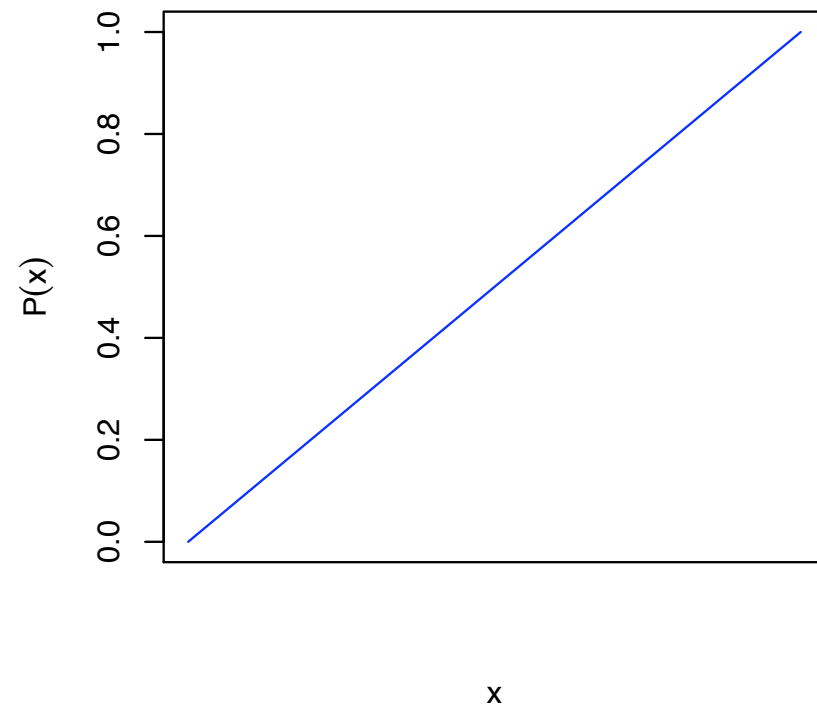
- Conditional mean $P(x)$ linked to predictors via an additive model

Differences with standard linear regression model:

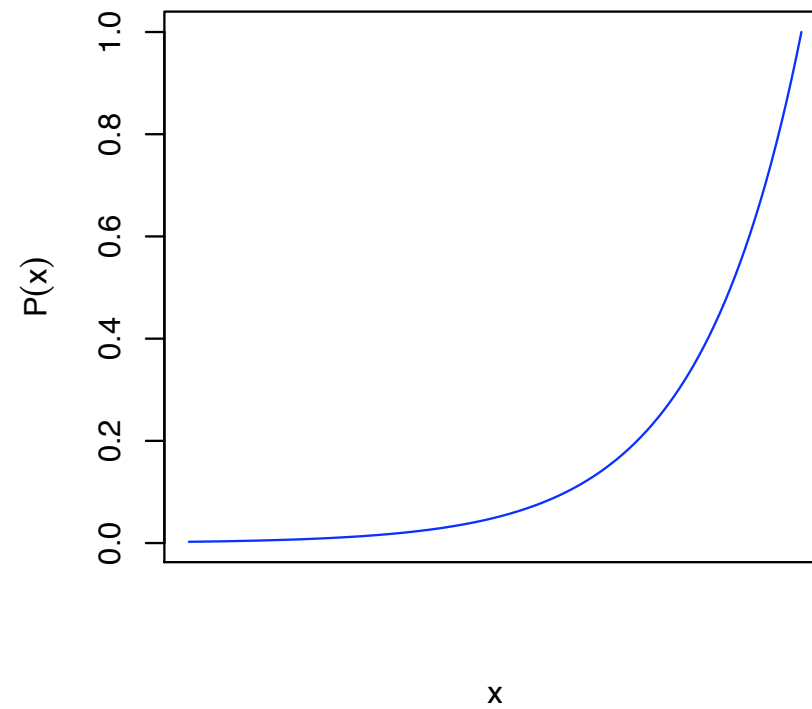
- y does not follow a normal distribution (either marginally or conditional on x)
 - y has a *binomial distribution* (mean P)
- response $P(x)$ represents the probability of outcome occurrence, and is constrained to be between 0 and 1 for all values of x
- different method for estimation of regression coefficients

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

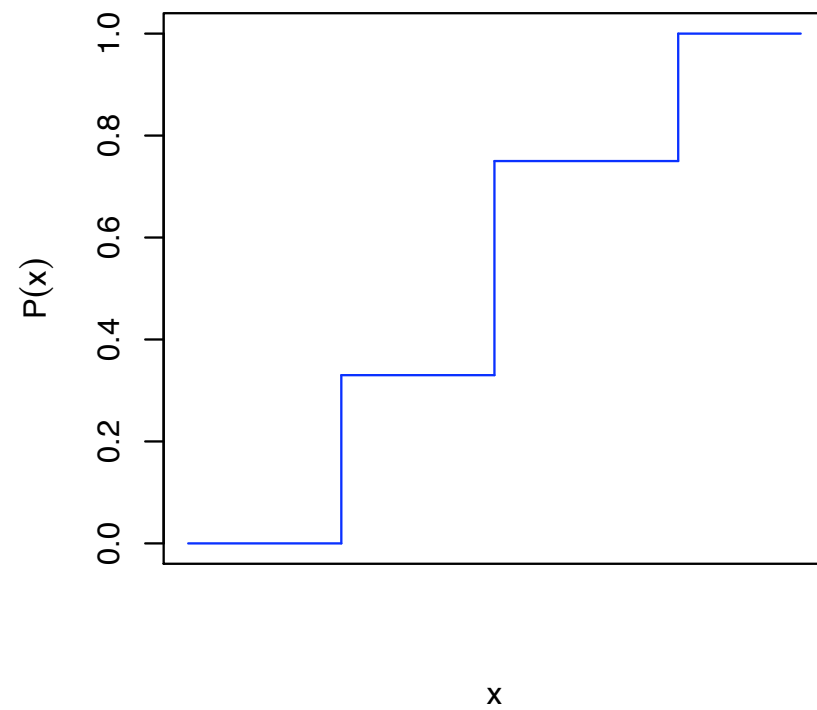
(A) Linear



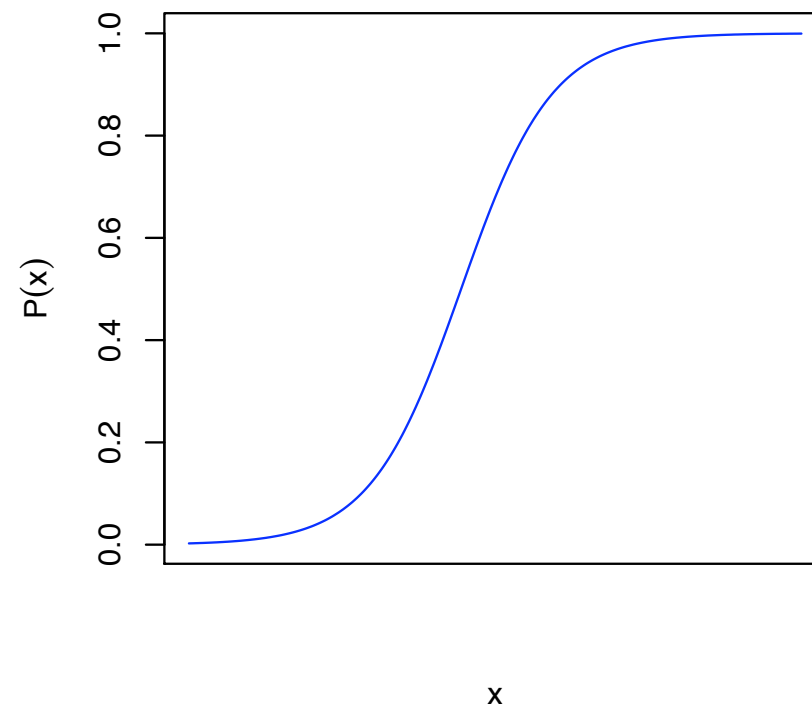
(B) Log linear



(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 probability/risk ($P(0)$) for individuals with $x = 0$
- β_1 gives the change in $P(x)$ for each unit increase in x .
- implies that β_1 is the *risk difference* associated with a unit increase in x .

Limitations:

- $P(x)$ has to be constrained to be between 0 and 1 for all values of x
 - *coefficient estimates have to satisfy this constraint*
- 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 increase in x .
- Implies that β_1 is the *log relative risk* associated with a unit increase in x .
- $\log$ = natural logarithm (i.e. log with base e)

Limitations:

- For this model to make sense, the coefficient estimates need to satisfy the constraint that the outcome ($\log P(x)$) is between $-\infty$ and 0 for all values of x
 - $\log(0) = -\infty$; $\log(1) = 0$

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
 - ─ The log odds transformation is also known as the “logit”
- 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 increase in x .
- Implies that β_1 is the *log odds ratio* associated with a unit increase in x .
- Response probability ($P(x)$) is naturally constrained to be between 0 and 1 for all values of x

Limitations:

- Linearity in log odds is not always plausible; odds ratio interpretation not favored in causal inference applications

Nonlinear form of the logistic regression model

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

Note:

- Alternate (and equivalent) representation of the log-linear form of the model (previous slide)
- You can check that $0 \leq P(x) \leq 1$ for any choice of x and coefficient values

Tolerance distribution motivation for the logistic model

- The logistic model can be motivated by assuming that a binary outcome y is observed for individuals with a latent tolerance level z (e.g. to a toxic substance) not exceeding a level x (e.g. a lethal dose). If the tolerance variable z follows the *logistic* distribution:

$$\Pr [y = 1|x] = \Pr [z \leq x] = \int_{-\infty}^x \frac{\exp[(z-\mu)/\tau]}{\tau(1+\exp[(z-\mu)/\tau])^2} dz = \frac{\exp[(x-\mu)/\tau]}{(1+\exp[(x-\mu)/\tau])}$$

$$\text{If } \beta_0 = -\mu/\tau, \beta_1 = 1/\tau, \Pr [y = 1|x] = \frac{\exp [\beta_0 + \beta_1 x]}{(1 + \exp [\beta_0 + \beta_1 x])}$$

- **Note:** a *Normal* tolerance distribution yields the *probit regression model* for binary outcomes (Φ denotes the normal distribution function; the *probit* transformation):

$$\Pr [y = 1|x] = \Phi(\beta_0 + \beta_1 x), \quad \Phi^{-1} (\Pr [y = 1|x]) = \beta_0 + \beta_1 x$$

Generalized linear models (GLMs) for binary outcomes

- From lecture 6 (and chapter 8 of text): GLMs are a family of models for a range of different outcome distributions that link the mean to predictors via a *link function*
- Binary outcomes follow a Binomial distribution (mean P), and different links yield a range of possible binary regression models:

<i>link function</i>	<i>regression model</i>
identity (P)	risk difference
log (log (P))	relative risk
log odds (log [$P / (1-P)$])	logistic
log-log (log [-log (P)])	log-log
complementary log-log (log [-log ($1-P$)])	complementary log-log
probit ($\Phi^{-1}(P)$)	probit

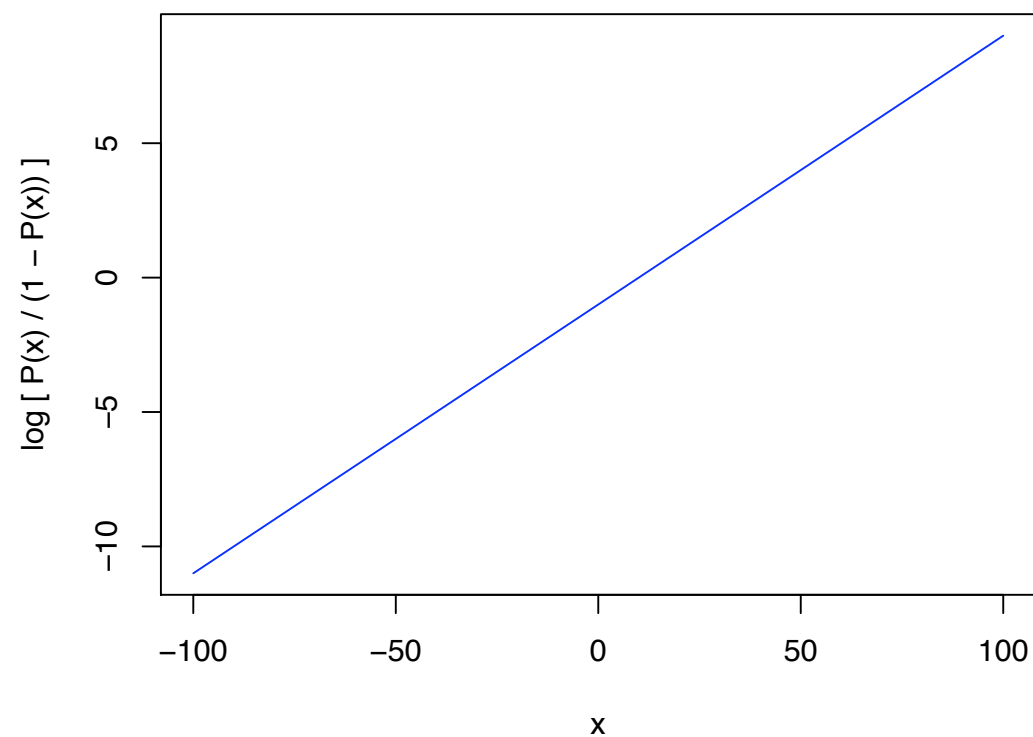
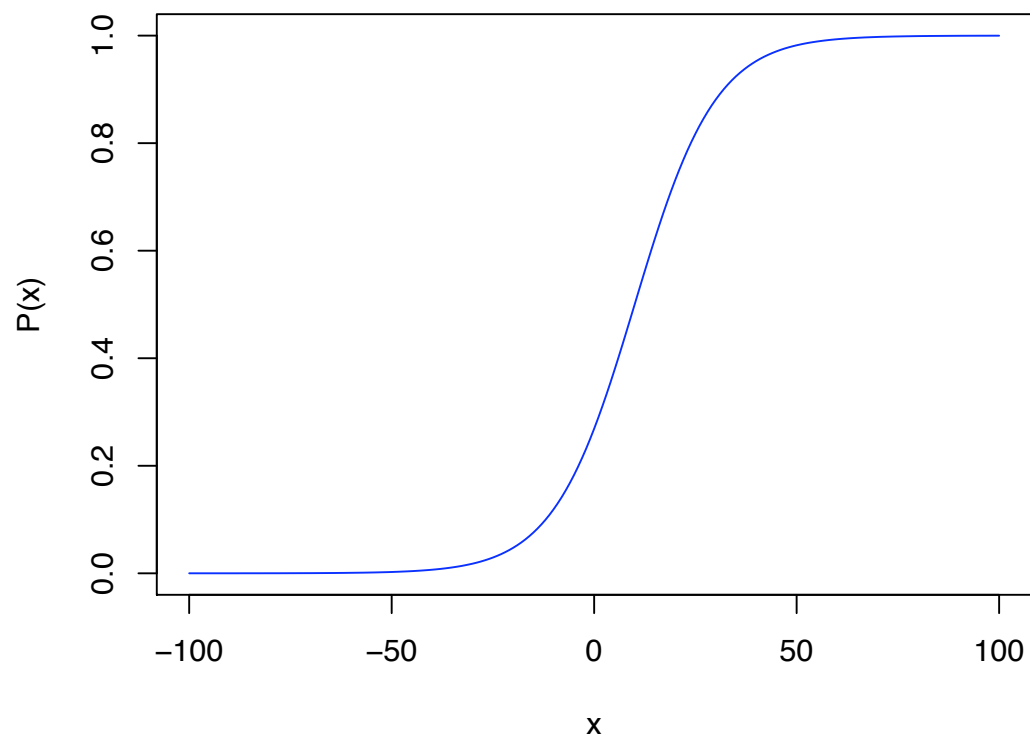
- These models can all be fitted in Stata using the `glm` command

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 for a binary outcome

Primary assumptions:

- model provides an appropriate representation for the dependence of the outcome probability 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 advantages over (constrained) alternatives based on RD or RR

Disadvantages:

- based on the odds ratio (a *non-collapsible* ratio measure) that has interpretation issues in certain contexts (and is 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 (or exposure) x :

$x = 1$ for “exposed” individuals

$x = 0$ for “unexposed” individuals

- β_0 is the log odds of the outcome for $x = 0$.
- β_1 is the log odds ratio comparing outcome odds in exposed ($x = 1$) to unexposed ($x = 0$) individuals:

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

- $\exp(\beta_1)$ is therefore the *odds ratio* comparing outcome odds between the exposed and unexposed

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 i.arcus
```

Logistic regression

Number of obs = 3,152

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							
present		1.635281	.2195036	3.66	0.000	1.257001	2.127399
_cons		.074344	.0062298	-31.02	0.000	.0630839	.0876141

Note: _cons estimates baseline odds.

- *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 #14)
- `arcus` is coded as a binary indicator so use of “`i.`” is optional
 - `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 without arcus

logistic command to investigate CHD - arcus association in WCGS

```
.logistic chd69 i.arcus
```

Logistic regression

Number of obs = 3,152

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						
present	1.635281	.2195036	3.66	0.000	1.257001	2.127399
_cons	.074344	.0062298	-31.02	0.000	.0630839	.0876141

- Note: _cons estimates baseline odds.
- 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:

$$2 \times [(\text{LL from more complex model}) - (\text{LL from simpler model (on slide 35)})] =$$
$$2 \times [(-879.10783) - (-885.6)] = 2 \times (6.49217) = 12.98434$$

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 i.arcus, coef
```

```
Logistic regression
```

```
Number of obs      =      3,152
```

```
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						
present	.4918144	.1342299	3.66	0.000	.2287286	.7549002
_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),
 - $\beta_1 = 0.4918144$ (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(.2287286)
```

```
1.2570008
```

```
. display exp(.7549002)
```

```
2.1273992
```

Logistic regression model equations for previous example

$$P(\text{arcus}) = \frac{\exp[-2.599052 + 0.4918144 \times \text{arcus}]}{(1 + \exp[-2.599052 + 0.4918144 \times \text{arcus}])}$$

$$\log \left[\frac{P(\text{arcus})}{1 - P(\text{arcus})} \right] = -2.599052 + 0.4918144 \times \text{arcus}$$

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

Number of obs = 3152

LR chi2(0) = -0.00

Prob > chi2 = .

Pseudo R2 = -0.0000

Log likelihood = -885.6

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

- this model is only of interest for comparison of log-likelihoods to the previous model (slide 32) including `arcus` (We'll discuss likelihood ratio testing soon)
- intercept coefficient "`_cons`" is the log odds of CHD in the entire 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` status)
- z statistic for the 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

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

linear on log odds scale

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

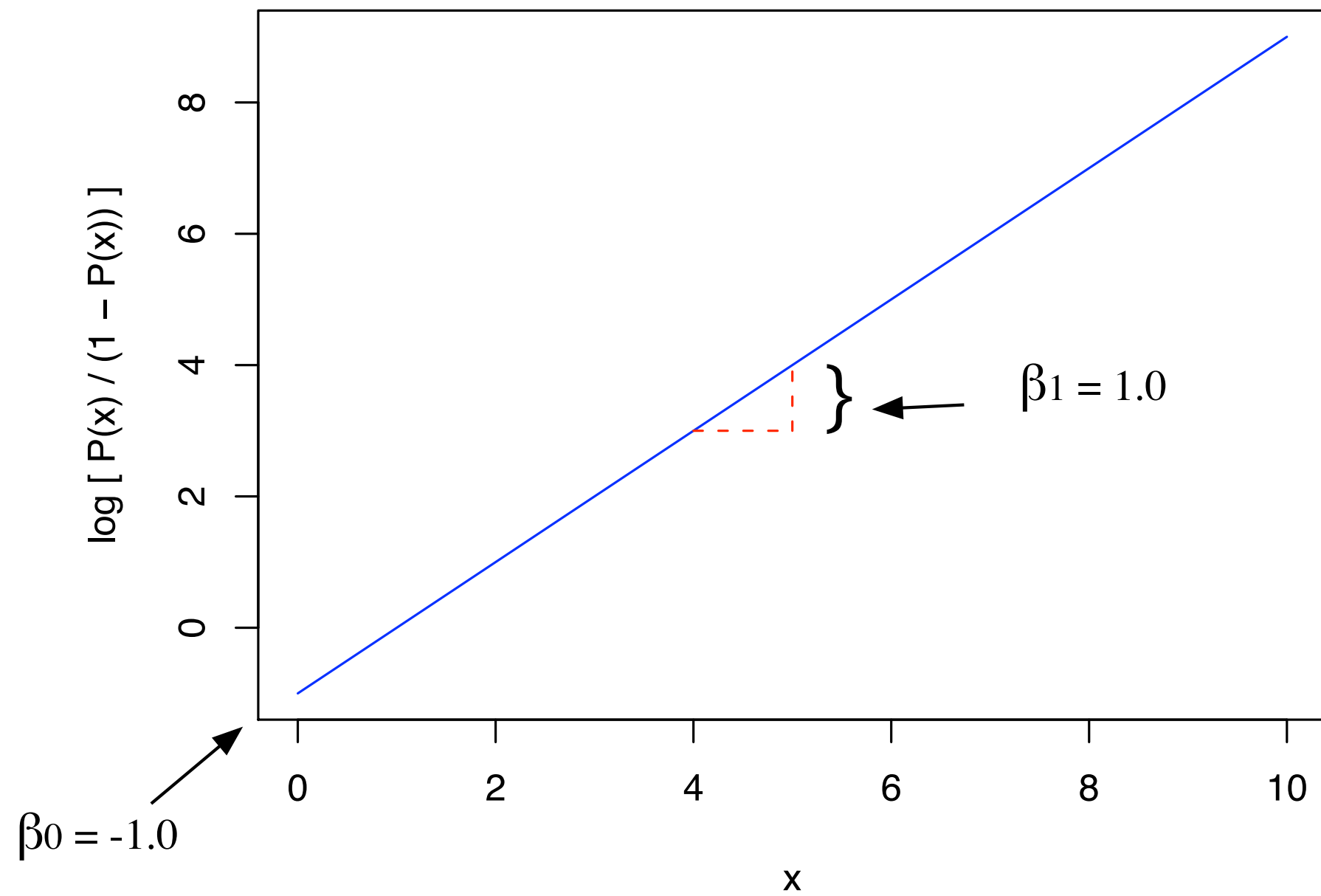
nonlinear on probability scale

- 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	.3594749	-14.91	0.000	-6.06358 -4.654464

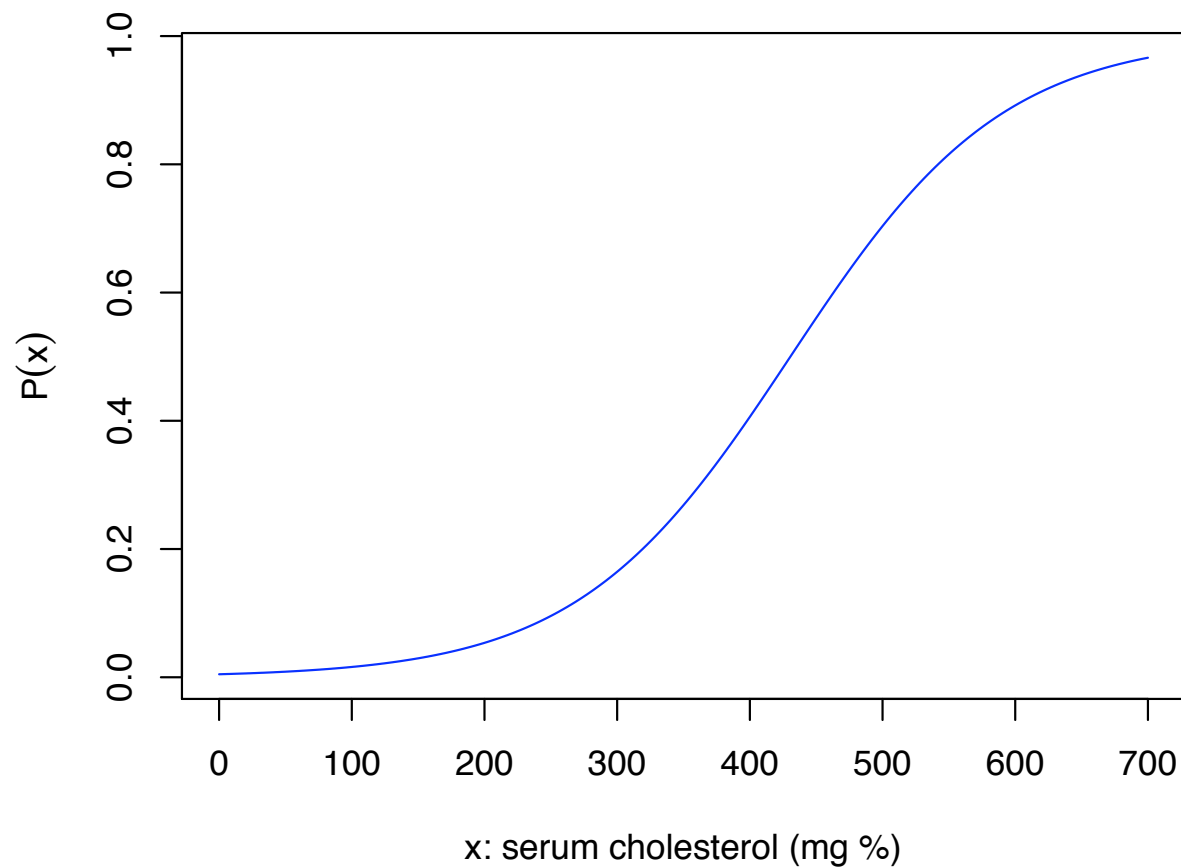
- estimated log odds ratio for each unit increase in cholesterol: 0.0124478
- estimated log odds ratio for a 100-unit increase in cholesterol: 1.24478
- corresponding log odds ratio computed from the definition of the logistic model:

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

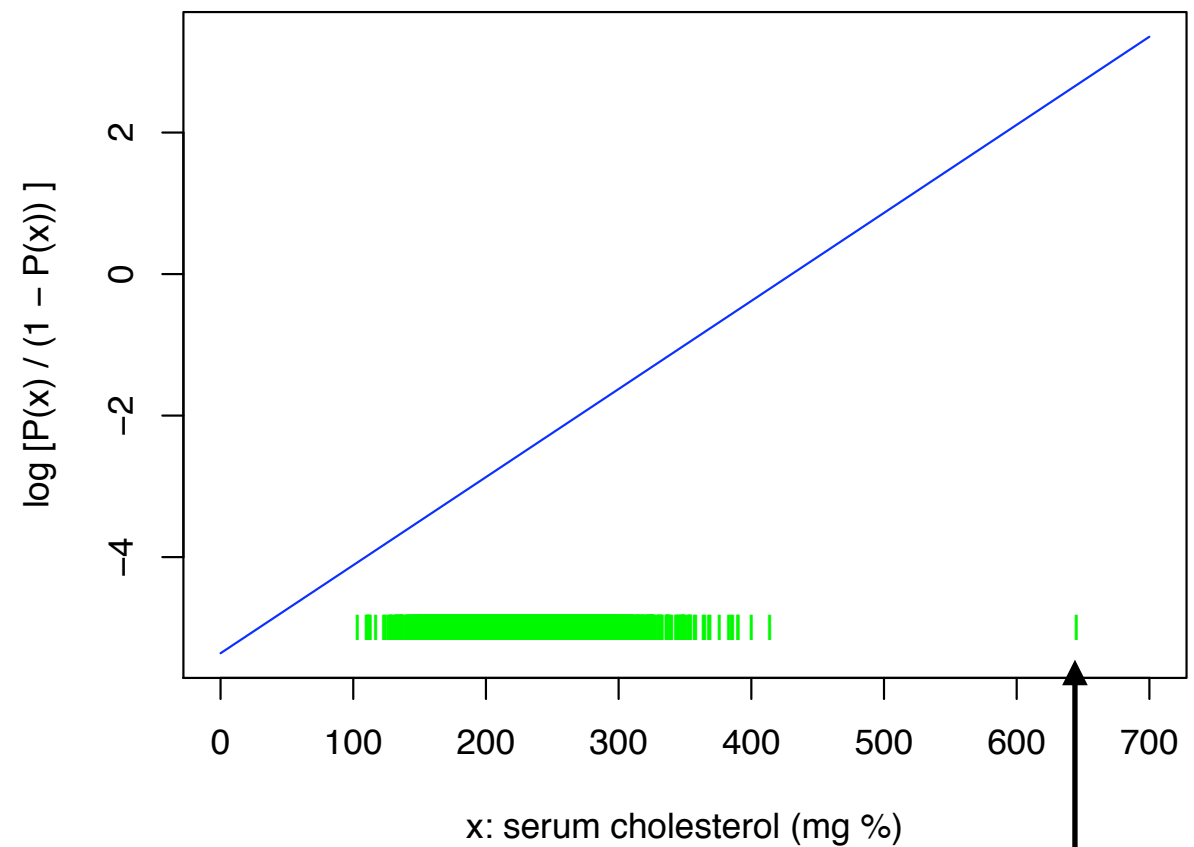
- the model is only interpretable for the range of cholesterol values present in the data
- the model assumes that the change in log odds for any unit change in cholesterol is independent of cholesterol level (linearity)
- 12 individuals have missing cholesterol values and are excluded

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

Probability scale



Log odds scale



Possible outlier?

Example: relationship between CHD and log cholesterol level

```
. gen lchol = log(chol)
. logistic chd69 lchol, coef
```

Logistic regression

Number of obs = 3,142

LR chi2(1) = 79.84

Prob > chi2 = 0.0000

Log likelihood = -849.67808

Pseudo R2 = 0.0449

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lchol	3.159401	.3653159	8.65	0.000	2.443395	3.875407
_cons	-19.63731	2.007602	-9.78	0.000	-23.57214	-15.70248

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

```
. dis exp(_b[lchol])
23.556477
```

- 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	.5760269	.066605	8.65	0.000	.4454835	.7065702

- 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	1.778956	.1184873	15.01	0.000	1.546725	2.011187

Logistic model for a nominal categorical predictor

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

- reported behavior pattern is classified in a single 4-level variable behpat , with numerical values 1,2,3 & 4 (labelled $A1$, $A2$, $B3$ & $B4$).

Three indicator variables are necessary to represent behavior pattern in a logistic model.
Letting $x = \text{behpat}$:

Indicator variable

Level	$I(x=A2)$	$I(x=B3)$	$I(x=B4)$
$A1$	0	0	0
$A2$	1	0	0
$B3$	0	1	0
$B4$	0	0	1

- group $A1$ (category with smallest numerical value) is chosen as the "baseline" level
- resulting logistic model needs 3 indicator variables and 3 regression coefficients to represent the 4-level predictor

$$\log \left[\frac{P(x)}{1 - P(x)} \right] = \beta_0 + \beta_1 I(x = A_2) + \beta_2 I(x = B_3) + \beta_3 I(x = B_4)$$

Example: relationship between CHD and behavior pattern

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

Logistic regression

Log likelihood = -870.16237

Number of obs = 3,154

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						
A2	.9807986	.2085511	-0.09	0.927	.6465252	1.487902
B3	.4119481	.0964927	-3.79	0.000	.2602922	.6519642
B4	.4241692	.1315496	-2.77	0.006	.2309681	.7789799
_cons	.1282051	.0248622	-10.59	0.000	.087667	.1874884

Note: _cons estimates baseline odds.

- likelihood ratio statistic (LR chi2(3)) 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 (B3 & B4) very similar
- suggests possibly grouping A1 & A2 and B3 & B4

Example: relationship between CHD and behavior pattern

- Use `testparm` to evaluate homogeneity of outcome odds across behavior categories:

```
. testparm i.behpat
```

```
( 1)  [chd69]2.behpat = 0
( 2)  [chd69]3.behpat = 0
( 3)  [chd69]4.behpat = 0
```

```
      chi2( 3) =    38.00
Prob > chi2 =    0.0000
```

- Tests same hypothesis as likelihood ratio test in logistic model including only `behpat` as a predictor; results differ slightly due to the fact that `testparm` uses a Wald test for logistic regression models
- Some difference between these test results is expected, especially in small samples - the likelihood ratio approach (covered later) is preferable in general

Example: relationship between CHD and behavior pattern

The variable `dibpat` is a binary indicator of A-type behavior patterns (*B3* & *B4* are the reference group):

```
. logistic chd69 i.dibpat
Logistic regression                                Number of obs      =       3,154
                                                    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 |
      A1,A2  |      2.372929   .3327001     6.16   0.000     1.802772     3.123408
      _cons  |      .0531629   .0061382    -25.41   0.000     .0423963     .0666636
-----+-----
```

Note: `_cons` estimates baseline odds.

- odds ratio comparing odds of CHD for individuals reporting type A (A1/A2) behavior to those reporting type B (*B3/B4*) 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 is sufficient)

Logistic model for an ordinal categorical predictor

Example: treating age as an ordinal categorical measure in WCGS

- Recode age (in years) as a five-level ordinal categorical measure (agec, defined on slide 16)
- Fit a logistic model for the new measure and evaluate evidence for trend
 - alternate analysis from tabodds (slide 17) provides similar results

```
. logistic chd69 i.agec
```

Logistic regression

Log likelihood = -868.14866

Number of obs

LR chi2(4)

Prob > chi2

Pseudo R2

=

=

=

=

3,154

44.95

0.0000

0.0252

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
+-----+-----+-----+-----+-----+-----+-----						
agec						
41-45	.8768215	.2025406	-0.57	0.569	.5575563	1.378903
46-50	1.70019	.3800504	2.37	0.018	1.097046	2.634935
51-55	2.318679	.5274963	3.70	0.000	1.484545	3.621494
56-60	2.886314	.7462298	4.10	0.000	1.738895	4.790864
_cons	.0605469	.0111989	-15.16	0.000	.0421358	.0870026
-----+-----+-----+-----+-----+-----+-----						

Note: _cons estimates baseline odds.

- Test for linear trend in log odds follows the same approach used for linear regression (covered in this week's lab)

Logistic regression for multiple predictors

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

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

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* or *controlled* 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:

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

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

Logistic models with multiple predictors

Example (two binary predictors): relationship between CHD, type A/B behavior pattern and smoking (yes/no) in WCGS

Research question: Could other factors (e.g. behaviors or factors reflecting "lifestyle") explain part (or all) of the association between CHD and type A behavior pattern?

(The created variable smoke is a binary indicator of any reported smoking):

```
. logistic chd69 i.dibpat i.smoke
```

Logistic regression

Log likelihood = -860.36587

Number of obs = 3,154

LR chi2(2) = 60.51

Prob > chi2 = 0.0000

Pseudo R2 = 0.0340

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
dibpat						
A1,A2	2.301319	.3238094	5.92	0.000	1.746657	3.032117
smoke						
smoker	1.8002	.2422866	4.37	0.000	1.382798	2.343597
_cons	.0393656	.0054842	-23.22	0.000	.0299593	.0517251

Note: _cons estimates baseline odds.

Logistic models with multiple predictors (smoking example continued)

```
. logistic chd69 i.dibpat i.smoke
```

Logistic regression

Number of obs = 3,154

LR chi2(2) = 60.51

Prob > chi2 = 0.0000

Log likelihood = -860.36587

Pseudo R2 = 0.0340

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
dibpat						
A1,A2	2.301319	.3238094	5.92	0.000	1.746657	3.032117
smoke						
smoker	1.8002	.2422866	4.37	0.000	1.382798	2.343597
_cons	.0393656	.0054842	-23.22	0.000	.0299593	.0517251

Note: _cons estimates baseline odds.

Interpretation:

- The adjusted odds ratio of 2.3 for type A behavior holds for smokers and non-smokers alike.
- Similarly, the adjusted odds ratio of 1.8 for smokers holds for type A & B behavior patterns.
- Estimated odds for CHD among non-smoker, type-B individuals is 0.39

Logistic models with multiple predictors

Example: CHD risk with systolic blood pressure and total serum cholesterol in the WCGS

Research question: are cholesterol and systolic blood pressure independently predictive of CHD risk?

```
. logistic chd69 chol sbp if chol<600
```

Logistic regression	Number of obs	=	3141
	LR chi2(2)	=	110.21
	Prob > chi2	=	0.0000
Log likelihood = -831.98952	Pseudo R2	=	0.0621

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
chol	1.011765	.0014981	7.90	0.000	1.008833	1.014706
sbp	1.024741	.0039104	6.40	0.000	1.017106	1.032434
_cons	.0002261	.0001393	-13.63	0.000	.0000676	.0007562

Notes:

- Excludes 12 observations with missing chol and 1 obs. with chol ≥ 600.
- Assumption of linearity for both chol and sbp must be checked before model is accepted as reasonable.

Logistic models with multiple predictors

Example: Logistic model including SBP and cholesterol level in WCGS (*continued*)

```
. logistic chd69 chol sbp if chol<600, coef
```

Logistic regression

Number of obs = 3141

LR chi2(2) = 110.21

Prob > chi2 = 0.0000

Pseudo R2 = 0.0621

Log likelihood = -831.98952

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
chol	.0116966	.0014807	7.90	0.000	.0087945	.0145987
sbp	.0244401	.003816	6.40	0.000	.016961	.0319193
_cons	-8.394706	.6160962	-13.63	0.000	-9.602233	-7.18718

Notes:

- 0.0117 is the log-odds ratio for each unit increase in chol holding sbp fixed
- 0.0244 gives the log-odds ratio for each unit increase in sbp holding chol fixed
- -8.395 gives the log-odds of CHD among indiv. with chol=0 & sbp=0
- model specifies linearity of log odds in both predictors:
 - the log-odds of CHD is linear in cholesterol for any specified value of SBP
 - the log-odds of CHD is linear in SBP for any specified value of cholesterol

Logistic models with multiple predictors

Example: Logistic model for centered (by the mean) & scaled (by 25 units) SBP and cholesterol levels

```
. logistic chd69 chol_25 sbp_25 if chol<600
```

Logistic regression

Number of obs = 3141

LR chi2(2) = 110.21

Prob > chi2 = 0.0000

Pseudo R2 = 0.0621

Log likelihood = -831.98952

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
chol_25	1.339658	.0495901	7.90	0.000	1.245905	1.440466
sbp_25	1.842279	.1757513	6.40	0.000	1.5281	2.221054
_cons	.0740423	.0055521	-34.71	0.000	.0639222	.0857646

Questions:

- 1.3397 is the odds ratio for each ? unit increase in chol
- 1.8423 is the lodds ratio for each ? until increase in sbp
- 0.0740 gives the odds of CHD among indiv. with chol=? & sbp=?

Logistic regression for grouped data (useful for data presented in tabular format)

Example: grouped data for CHD outcomes by behavior pattern and smoking status

Note: `denom` is the number of indiv. in each row, and `case` gives number of outcomes

	denom	case	smoke	behpat
1.	130	8	0	1
2.	654	62	0	2
3.	658	20	0	3
4.	210	8	0	4
5.	134	22	1	1
6.	671	86	1	2
7.	558	41	1	3
8.	139	10	1	4

Note: To fit a logistic model, need separate entries for cases and non-cases

```
. gen n0 = denom - case      /* counts of non-cases */
. rename case n1             /* counts of cases */
. gen id = _n                /* reshape needs a group id variable */
. drop denom
. reshape long n, i(id) j(y)
. list , sep(0)
```

	id	y	n	smoke	behpat
1.	1	0	122	0	1
2.	1	1	8	0	1
3.	2	0	592	0	2
4.	2	1	62	0	2

Logistic regression for grouped data (continued)

```
. logistic y i.smoke i.behpat [fw=n]
```

Logistic regression

Number of obs = 3,154

LR chi2(4) = 60.58

Prob > chi2 = 0.0000

Log likelihood = -860.33398

Pseudo R2 = 0.0340

y	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
1.smoke	1.80189	.2426315	4.37	0.000	1.383919	2.346097
behpat						
2	.9813006	.2095303	-0.09	0.930	.6457299	1.49126
3	.4215688	.0991076	-3.67	0.000	.2659239	.6683123
4	.4501848	.1401403	-2.56	0.010	.2445772	.8286397
_cons	.0919746	.0194862	-11.26	0.000	.0607195	.1393181

.* **Note:** results identical to using full (ungrouped) WCGS dataset

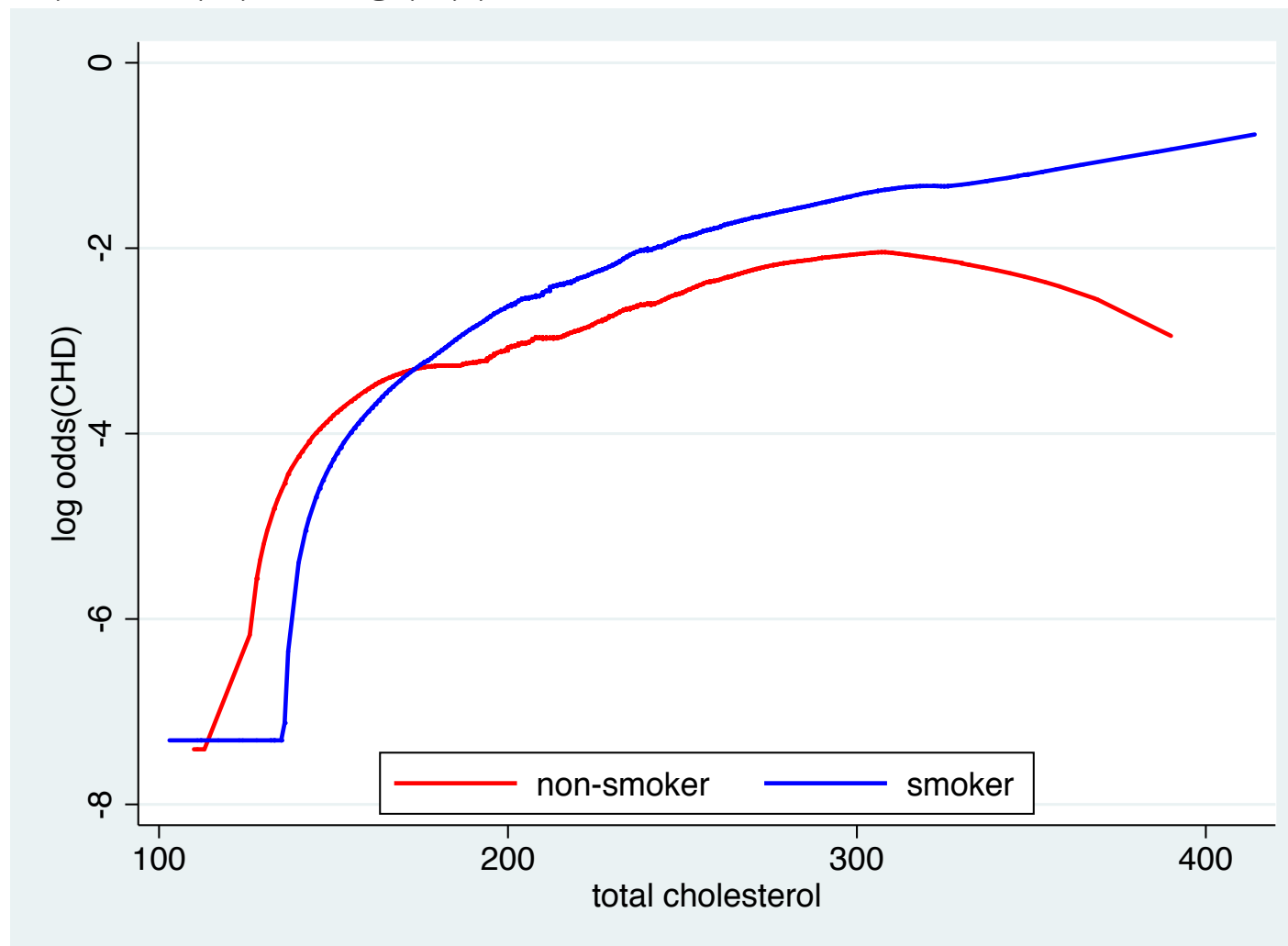
Notes:

- frequency counts in each outcome category accounted for by `fw[ ]` option
- results identical to using full (ungrouped) WCGS dataset

Smoothing for binary outcomes

Example: The `lowess` command in Stata can also be used to explore the relationship between a binary outcome and a continuous predictor. Example: log odds of CHD and cholesterol levels by smoking status in the WCGS:

```
. twoway (lowess chd69 chol if smoke==0 & chol<600, logit sort bw(.8) lcol(red)
lw(medthick) msymbol(none)) (lowess chd69 chol if smoke==1 & chol<600, logit sort bw(.8)
lcol(blue) lw(medthick) msymbol(none)), ytitle("log odds(CHD)") legend(order(1 "non-
smoker" 2 "smoker") pos(6) ring(0))
```



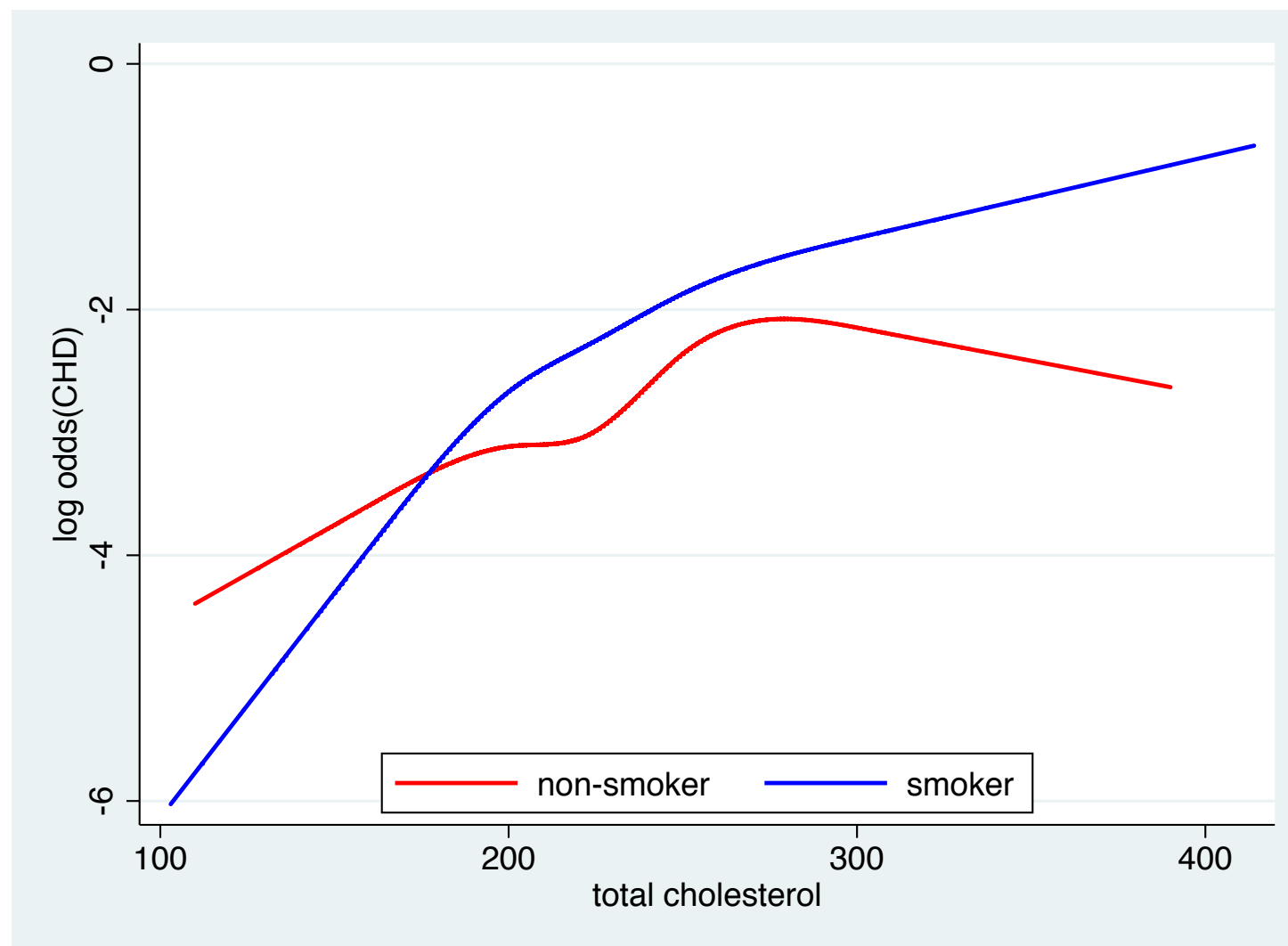
Notes:

- The `logit` option plots the log odds of outcome vs. continuous predictor variable
- Linear and restricted cubic splines can be applied similarly to linear models

Smoothing for binary outcomes

Example: Restricted cubic spline for log odds of CHD and cholesterol levels by smoking status:

```
. mkspline chsp=chol if chol<600, cubic
. quietly logistic chd69 c.chsp*##i.smoke if chol<600
. predict lopred, xb
. twoway (line lopred chol if smoke==0 & chol<600, lcol(red) lw(thick) msymbol(none)
sort) (line lopred chol if smoke==1 & chol<600, lcol(blue) lw(thick) msymbol(none)
sort), ytitle("log odds(CHD)") legend(order(1 "non-smoker" 2 "smoker") pos(6) ring(0))
```



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 model for the outcome probability.

- For independent outcomes, the likelihood is the product of the individual outcome probabilities specified by the assumed 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)

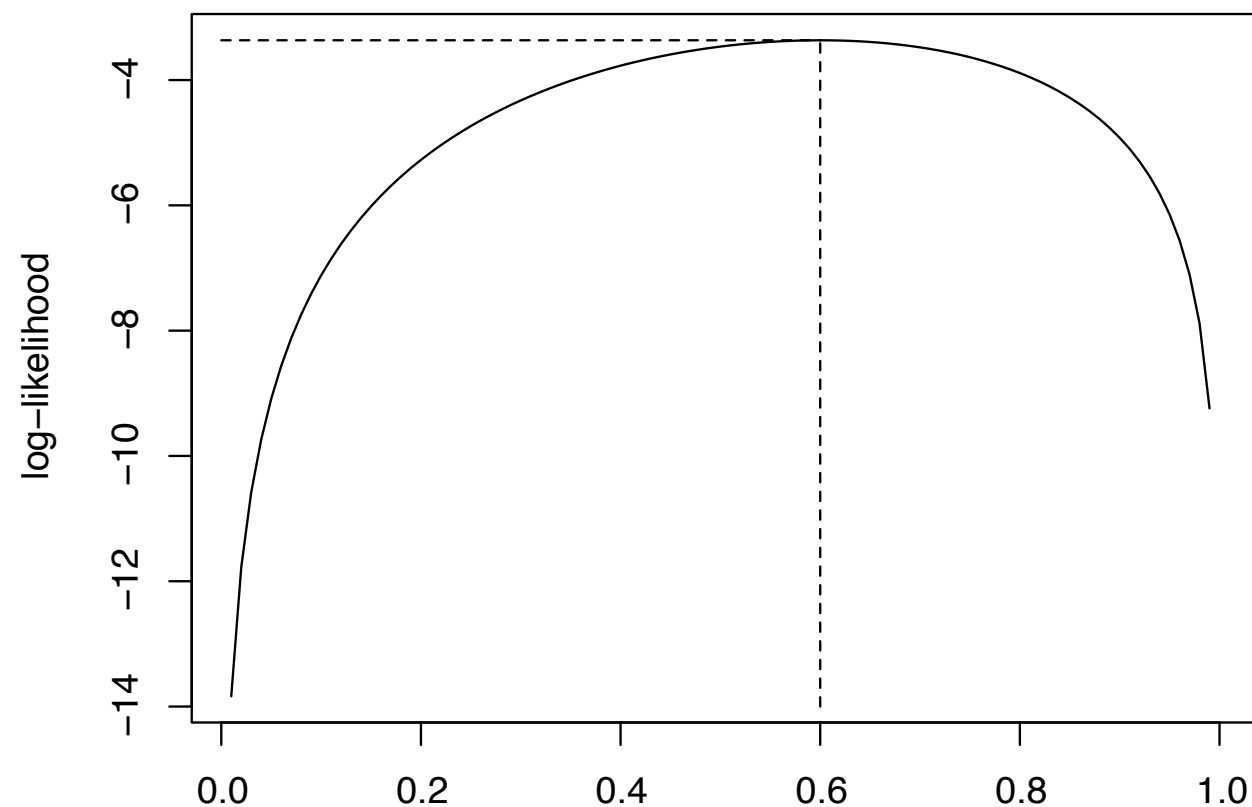
unknown quantities: parameter(s) of interest specified by the model

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:** unknown prob. P of throwing head (H)
- **observed data:** 5 indep. observations (coin tosses): 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

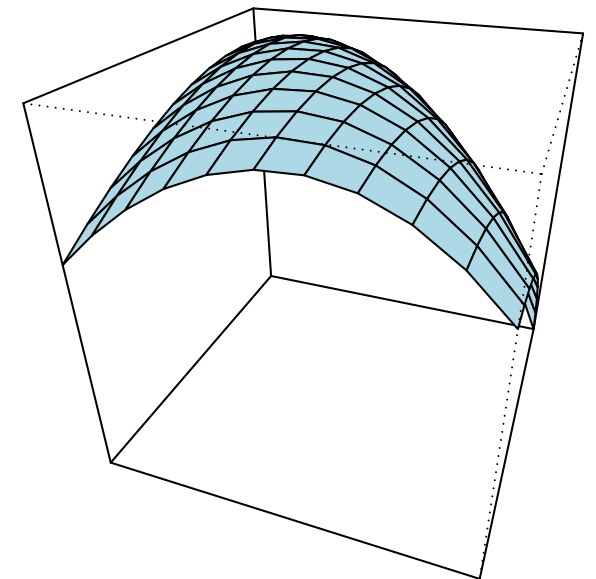
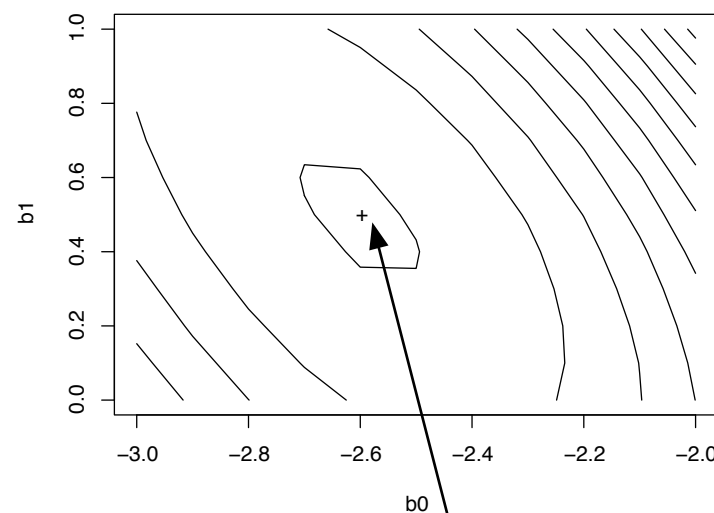
- **problem:** estimate the dependence of CHD risk on a binary indicator of arcus (x)
- **model:** 3154 indep. observations of CHD with prob. P given by the logistic model

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

- **observed data:**

	chd69(y)	arcus(x)
1.	No CHD (y=0)	0
2.	No CHD (y=0)	1
3.	No CHD (y=0)	0
4.	No CHD (y=0)	0
5.	CHD (y=1)	1
6.	No CHD (y=0)	0
7.	No CHD (y=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:** $\beta_0 = -2.59905; \beta_1 = 0.49181$

Maximum likelihood estimation for a continuous outcome

Example: dependence of glucose on BMI in HERS ($n = 2758$)

- **problem:** estimate the linear dependence of mean glucose (y) on BMI (x)
- **model:** mean $\mu(x) = E[y | x] = \beta_0 + \beta_1 x$; norm. distributed errors with variance σ^2

- **likelihood contribution for an indiv. with glucose = y & BMI = x :**

$$\frac{1}{\sigma\sqrt{2\pi}} \exp \left[\frac{-(y - \mu(x))^2}{2\sigma^2} \right]$$

- **likelihood for sample:**

$$\prod_{i=1}^{2758} \left\{ \frac{1}{\sigma\sqrt{2\pi}} \exp \left[\frac{-(y_i - \mu(x_i))^2}{2\sigma^2} \right] \right\} = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^{2758} \times \exp \left[\frac{-\sum_{i=1}^{2758} (y_i - \mu(x_i))^2}{2\sigma^2} \right]$$

- **maximum likelihood estimates:** $\beta_0 = 60.074$; $\beta_1 = 1.822$; $\sigma^2 = 1258.648$

```
. quietly regress glucose BMI
. matrix list e(b)
```

```
e(b) [1,2]
           BMI      _cons
y1  1.8221759  60.074051
```

```
. display e(rss)/(e(N)-1)
1258.6479
```

Properties of maximum likelihood estimates

ML estimates for many standard regression models have desirable properties as sample size increases (i.e. *asymptotically*):

- Consistency
 - ─ estimates correspond to underlying true parameter values (if model is correct)
- Normality
 - ─ estimates are normally distributed
- Efficient
 - ─ variability of estimates is no greater than competing estimators

Notes:

- Variability of ML estimates is directly related to curvature (information) of the likelihood function around the maximum
- Least squares estimate of regression coefficients in a linear model are equivalent to ML estimates when the residuals follow a normal distribution