

# **Introduction to Logistic Regression**

**I. When to use it**

**II. Why not to use OLS and the general linear model**

**III. The model for a binary outcome**

**IV. Assumptions and model fit**

**V. Interpretation: odds ratios and predicted probabilities**

**VI. With an ordinal or nominal outcome variable**

**VII. Some references about logistic regression**

## I. When to use it

Predictive relationship (clear “outcome” variable) of interest

Categorical or continuous predictors

Categorical outcome variable

a) binary: yes/no, true/false, right/wrong, etc.

b) ordinal: rankings, Likert-type rating scales  
(strongly agree, agree, neutral, disagree, strongly disagree),  
sum of Likert-type item scores, etc.

c) nominal: religious groups, experimental groups, etc.

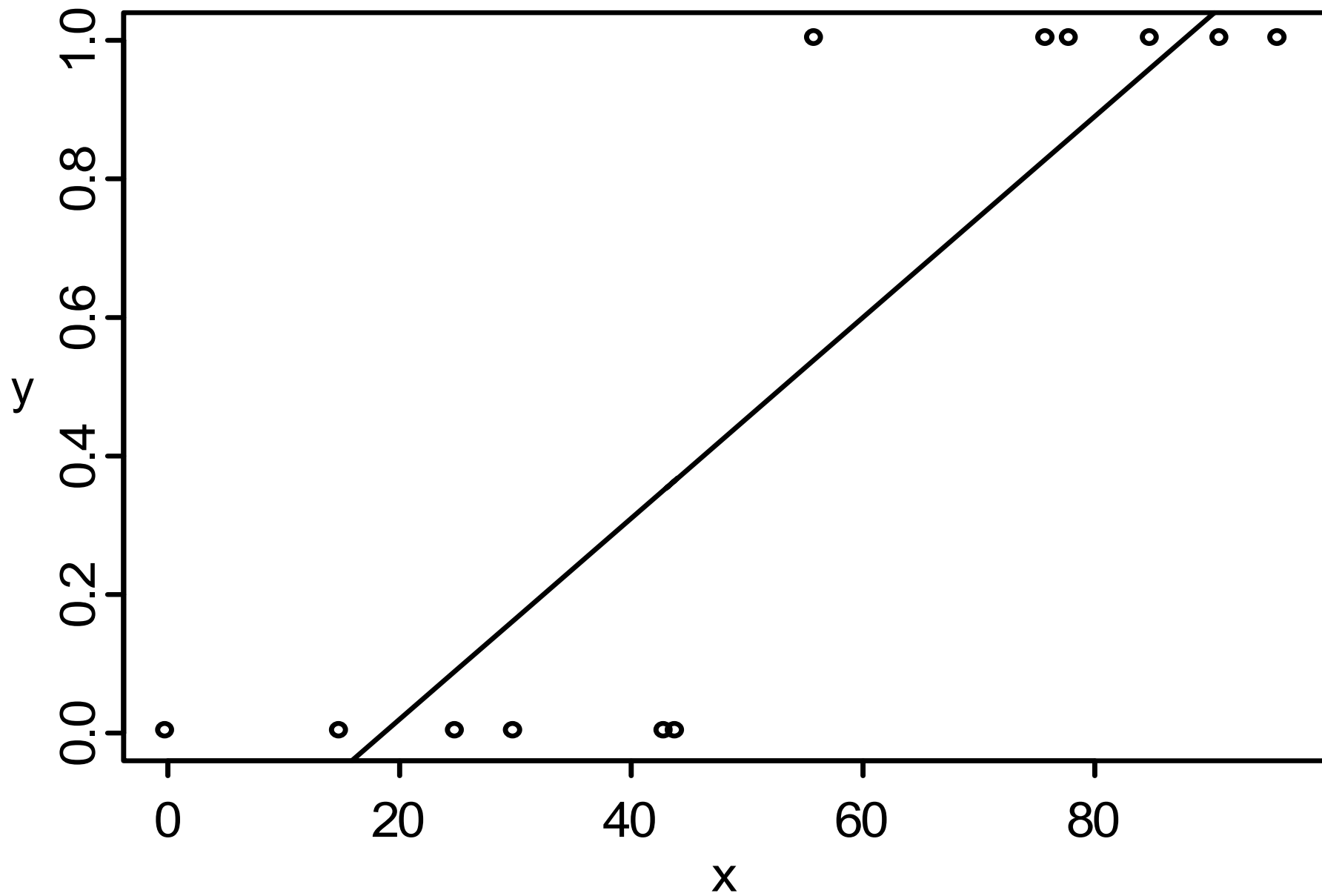
## II. Why not to use OLS and the general linear model

e.g., binary  $y$   $E(y_i) = p_i = b_0 + b_1x_{1i} + b_2x_{2i} \dots + b_kx_{ki}$

### Problems:

- 1) Nonsensical predictions: The linear predictor can equal a value that cannot possibly be a probability (i.e., is outside the range 0 to 1).
  - 2) Heteroscedasticity: The error variance depends on the predictors and is not constant.
  - 3) Normality:  $y$  conditional on  $x$  has two possible values, not a normal distribution
- usual standard errors and significance tests are inaccurate

#### 4) Functional form: not a best fit line



### III. The model for a binary outcome

Logistic link function; “linear in the logit”

$$g(p_i) = b_0 + b_1 x_{1i} + b_2 x_{2i} \dots + b_k x_{ki}$$

$$\log\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_{1i} + b_2 x_{2i} \dots + b_k x_{ki}$$

i counts observations, k counts predictors

p = probability of outcome “1”

1 – p = probability of outcome “0”

x = predictor variable

b = parameter estimated with maximum likelihood

## IV. Assumptions and model fit

### 1) Assumptions

- a) sample size is adequate for goodness-of-fit tests and for ML estimates to be  $\sim N$
- b) data come from a (product) binomial distribution
- c) model is correct and fits the data
- d)  $E(\varepsilon) = 0$
- e) observations are independent

## 2) Model fit

**a) global measures of model fit** compare observed probabilities to probabilities predicted by the model

$H_0$ : model-predicted probabilities are close to observed probabilities, i.e., good fit

fail to reject  $H_0$  for good model fit

when all predictors are categorical:

- i) Pearson chi-square
- ii) likelihood ratio chi-square (deviance)

when one or more predictors is continuous:

- iii) Hosmer-Lemeshow (also a chi-square test)

## **b) diagnostics**

use to evaluate more specific aspects of fit; see where fit is problematic when global fit is poor

many tools analogous to those for GLM are available

i) various types of residuals and ways to plot them

ii) various measures of influence

e.g., influence of an observation on regression parameter estimates or on the chi-square

## V. Interpretation: odds ratios and predicted probabilities

### 1) Example: One binary predictor

Predict the presence (versus absence) of coronary artery disease from male versus female sex.

	CAD	No CAD	
Male	30	15	45
Female	12	21	33
	42	36	78

The model (for one person):

$$\log\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 \text{sex}_i$$

(No chi-square tests of global fit or residuals because model is saturated: number of groups with unique settings on the predictors = number of model parameters.)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5596	0.3619	2.3914	0.1220
sex	1	1.2527	0.4806	6.7944	0.0091

Sex is a significant predictor of the presence/absence of coronary artery disease.

The model with these parameters:

$$\log\left(\frac{p_i}{1-p_i}\right) = -0.56 + 1.25(\text{sex}_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = -0.56 + 1.25(\text{sex}_i)$$

Regression coefficients are not readily interpretable in this form. So, exponentiate both sides, creating a model for the odds:

$$\frac{p_i}{1-p_i} = \exp[-0.56 + 1.25(\text{sex}_i)]$$

$p$  = probability of outcome “1”

$1 - p$  = probability of outcome “0”

$$\frac{p_i}{1-p_i} = \exp[-0.56 + 1.25(\text{sex}_i)]$$

$\text{sex}_i$ : "0" = female, "1" = male

$$\left( \frac{p_0}{1-p_0} \right) = \exp(-0.56) = 0.57 \quad \text{odds of CAD for women}$$

$$\left( \frac{p_1}{1-p_1} \right) = \exp(-0.56 + 1.25) = 2.00 \quad \text{odds of CAD for men}$$

$$\text{Odds ratio: } \frac{2.00}{0.57} = 3.5$$

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5596	0.3619	2.3914	0.1220
sex	1	1.2527	0.4806	6.7944	0.0091

The regression parameter, exponentiated, is the odds ratio.

$$\exp(1.2527) = 3.5$$

The odds of men having coronary artery disease are about 3.5 times the odds of women having coronary artery disease.

95% Confidence interval for the odds ratio:

Parameter	DF	Estimate	SE
sex	1	1.2527	0.4806

$$1.2527 \pm (.4806 * 1.96) = .3107 \text{ and } 2.1947$$

$$\exp(.3107) = 1.36 \quad \text{and} \quad \exp(2.1947) = 8.98$$

Plausible values for the odds ratio range from 1.36 to 8.98

95% of the intervals we could have computed in this way contain the true population odds ratio

Gives the significance test: Interval does not contain 1 (would correspond to 0 for the parameter estimate b/c  $\exp(0) = 1$ )

$$H_0: b = 0 \quad \leftrightarrow \quad H_0: \text{odds ratio} = 1$$

Can also interpret in terms of model-predicted probabilities.

Equivalent expression for the model:

$$p_i = \frac{\exp(b_o + b_1 x_{1i} + b_2 x_{2i} \dots + b_k x_{ki})}{1 + \exp(b_o + b_1 x_{1i} + b_2 x_{2i} \dots + b_k x_{ki})}$$

For this example, model-predicted probability of CAD for men:

$$p_1 = \frac{\exp(-0.56 + 1.25)}{1 + \exp(-0.56 + 1.25)} = 0.67$$

Model-predicted probability of CAD for women:

$$p_0 = \frac{\exp(-0.56)}{1 + \exp(-0.56)} = 0.36$$

## 2) Example: One nominal predictor

Predict whether urinary tract infection (UTI) was cured following treatment (drug A, B, or C).

	cured	not cured	
Drug A	118	33	151
Drug B	155	16	171
Drug C	102	52	154
	375	101	476

(Also saturated; 3 groups with 3 parameters.)

As with GLM, coding scheme used for nominal predictor (e.g., reference-cell or deviation-from-the-mean).

Reference-cell coding used here. Reference group is drug C.

	$x_1$	$x_2$
Treatment = drug A:	1	0
drug B:	0	1
drug C:	0	0

	cured	not cured	
Drug A	118	33	151
Drug B	155	16	171
Drug C	102	52	154
	375	101	476

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6737	0.1704	15.6334	<.0001
treatment A	1	0.6004	0.2604	5.3167	0.0211
treatment B	1	1.5969	0.3130	26.0280	<.0001

Exponentiate parameter estimates for odds-ratio interpretations.

The odds of being cured are  $\exp(.6004) = 1.82$  times larger for those taking drug A versus drug C.

The odds of being cured are  $\exp(1.5969) = 4.94$  times larger for people taking drug B versus drug C.

SAS exponentiates and computes CIs for you:

### Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
treatment A vs C	1.823	1.094	3.037
treatment B vs C	4.938	2.674	9.119

Intercept	1	0.6737	0.1704	15.6334	<.0001
treatment A	1	0.6004	0.2604	5.3167	0.0211
treatment B	1	1.5969	0.3130	26.0280	<.0001

Comparison between drug B and drug A is obtained by subtraction:

$$\exp(1.5969 - 0.6004) = \exp(.9965) = 2.71$$

The odds of being cured are 2.71 times larger for people taking drug B versus drug A.

Effect	Point Estimate	95% Wald Confidence Limits	
treatment B vs A	2.709	1.424	5.154

Intercept	1	0.6737	0.1704	15.6334	<.0001
treatment A	1	0.6004	0.2604	5.3167	0.0211
treatment B	1	1.5969	0.3130	26.0280	<.0001

Model-predicted probability of being cured with drug A, B, or C:

$$p_A = \frac{\exp(.6737 + .6004)}{1 + \exp(.6737 + .6004)} = 0.78$$

$$p_B = \frac{\exp(.6737 + 1.5969)}{1 + \exp(.6737 + 1.5969)} = 0.91$$

$$p_c = \frac{\exp(.6737)}{1 + \exp(.6737)} = 0.66$$

### 3) Example: One continuous predictor

Predict the presence (versus absence) of coronary artery disease from age, measured in years.

Sample size requirement for deviance and Pearson chi-square tests not met with a continuous predictor. Alternative test of fit:

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
10.4505	7	0.1644

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.6431	1.4184	6.5972	0.0102
age	1	0.0801	0.0299	7.1548	0.0075

Age significantly predicts presence/absence of CAD.

$$\frac{p_i}{1-p_i} = \exp[-3.64 + .08(age_i)]$$

For every one year that age increases, multiply the odds of CAD by  $\exp(.0801) = 1.08$ .

The odds of CAD increase by  $\exp(10*.0801) = 2.23$  for every 10 years that age increases.

Confidence interval:

### Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	1.083	1.022	1.149

Intercept	1	-3.6431	1.4184	6.5972	0.0102
age	1	0.0801	0.0299	7.1548	0.0075

Model-predicted probability of having CAD requires plugging in particular values for age:

e.g., age 21:

$$p_{21} = \frac{\exp[-3.6431 + .0801(21)]}{1 + \exp[-3.6431 + .0801(21)]} = 0.12$$

e.g., age 60:

$$p_{60} = \frac{\exp[-3.6431 + .0801(60)]}{1 + \exp[-3.6431 + .0801(60)]} = 0.76$$

## 4) Interpretation with 2 or more predictors

Analogous to the move from simple linear regression to multiple regression.

### Example: Two binary predictors

ECG	Sex	CAD	no CAD	Total
low	Female	4	11	15
low	Male	9	9	18
Total		13	20	33
high	Female	8	10	18
high	Male	21	6	27
Total		29	16	45

Predict CAD from sex and whether electrical activity of the heartbeat (electrocardiogram, ECG) is low or high.

Model:

$$\log\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 \text{sex}_i + b_2 \text{ecg}_i$$

Model fit:

### Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	0.2141	1	0.2141	0.6436
Pearson	0.2155	1	0.2155	0.6425

Fail to reject  $H_0$ : model-predicted probabilities are close to observed probabilities, i.e., good fit

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.1747	0.4854	5.8571	0.0155
sex	1	1.2770	0.4980	6.5750	0.0103
ecg	1	1.0545	0.4980	4.4844	0.0342

Odds ratios are now partial odds ratios. Relationships are with all other variables in the model statistically controlled (held constant).

Sex is a significant predictor of CAD controlling for ECG.  
ECG is a significant predictor of CAD controlling for sex.

## Odds ratio interpretations:

Parameter	DF	Estimate	SE	Chi-Square	Pr > ChiSq
Intercept	1	-1.1747	0.4854	5.8571	0.0155
sex	1	1.2770	0.4980	6.5750	0.0103
ecg	1	1.0545	0.4980	4.4844	0.0342

Statistically controlling for ECG, the odds of coronary artery disease for men are about  $\exp(1.2770) = 3.59$  times the odds for women.

Statistically controlling for sex, the odds of coronary artery disease for people with high ECG are about  $\exp(1.0545) = 2.87$  times the odds for people with low ECG.

Odds and predicted probabilities of CAD for each group:

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.1747	0.4854	5.8571	0.0155
sex	1	1.2770	0.4980	6.5750	0.0103
ecg	1	1.0545	0.4980	4.4844	0.0342

ECG-sex group

values for the variables

	x <sub>1</sub> (sex)	x <sub>2</sub> (ECG)
(1) female, low ECG	0	0
(2) female, high ECG	0	1
(3) male, low ECG	1	0
(4) male, high ECG	1	1

(1) female, low ECG:  $\left( \frac{p_{00}}{1-p_{00}} \right) = \exp(-1.1747) = .31$

(1) female, low ECG:

$$p_{00} = \frac{\exp(-1.1747)}{1 + \exp(-1.1747)} = .24$$

(2) female, high ECG:

$$\left( \frac{p_{01}}{1 - p_{01}} \right) = \exp(-1.1747 + 1.0545) = .89$$

$$p_{01} = \frac{\exp(-1.1747 + 1.0545)}{1 + \exp(-1.1747 + 1.0545)} = .47$$

(3) male, low ECG:

$$\left(\frac{p_{10}}{1-p_{10}}\right) = \exp(-1.1747+1.2770) = 1.11$$

$$p_{10} = \frac{\exp(-1.1747+1.2770)}{1+\exp(-1.1747+1.2770)} = .53$$

(4) male, high ECG:

$$\left(\frac{p_{11}}{1-p_{11}}\right) = \exp(-1.1747+1.2770+1.0545) = 3.18$$

$$p_{11} = \frac{\exp(-1.1747+1.2770+1.0545)}{1+\exp(-1.1747+1.2770+1.0545)} = .76$$

## VI. With an ordinal or nominal outcome variable

ordinal outcome:

- \*proportional odds
- adjacent category
- continuation ratio
- partial proportion odds
- stereotype logistic

nominal outcome:

generalized logits (multinomial logit)

$c$  response categories  $\rightarrow c-1$  logits fitted simultaneously

## Proportional odds model for ordinal outcome

e.g., 3 ordered response categories → 2 logits

$$\log \left( \frac{p_{1i}}{p_{2i} + p_{3i}} \right) = b_{o1} + b_1 x_{1i} + b_2 x_{2i} \dots + b_k x_{ki}$$

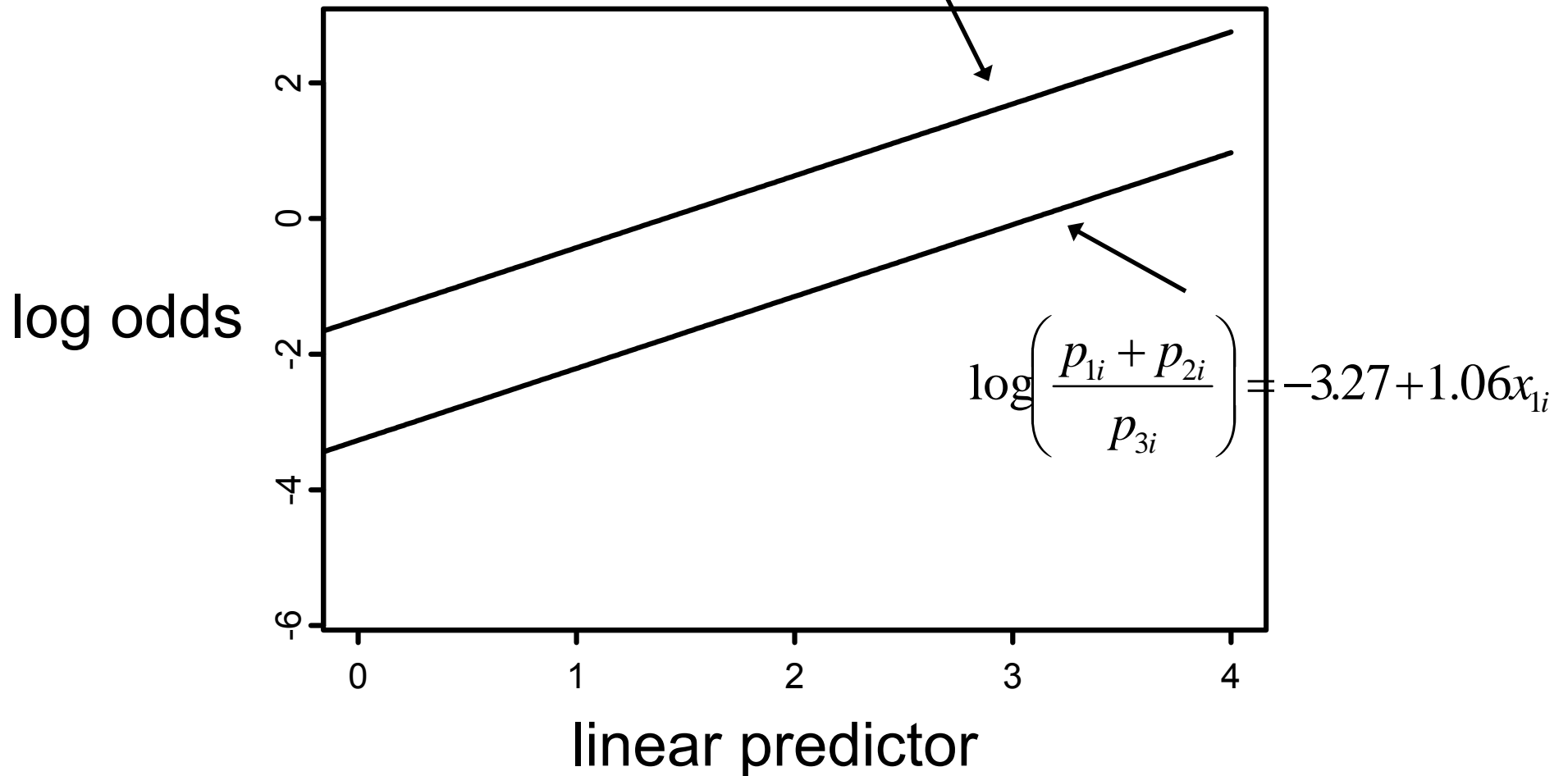
$$\log \left( \frac{p_{1i} + p_{2i}}{p_{3i}} \right) = b_{o2} + b_1 x_{1i} + b_2 x_{2i} \dots + b_k x_{ki}$$

Separate intercepts, common slopes.

“The odds of a lower (e.g., less depressed) response are...”

Proportional odds: relation between predictor(s) and log odds is the same strength for each logit.

$$\log\left(\frac{p_{1i}}{p_{2i} + p_{3i}}\right) = -1.49 + 1.06x_{1i}$$



## Generalized logits model for nominal response

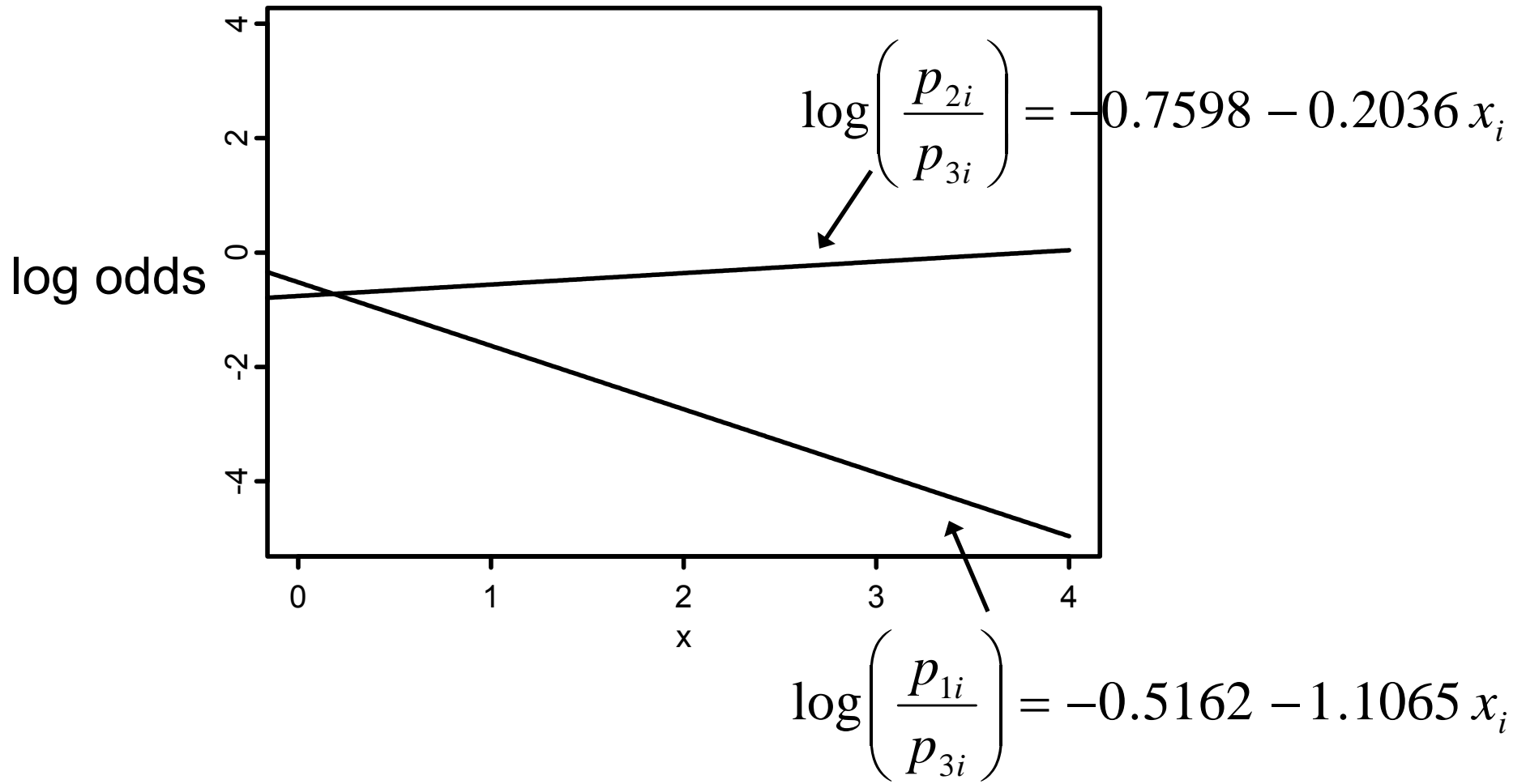
e.g., 3 un-ordered response categories → 2 logits

$$\log \left( \frac{p_{1i}}{p_{3i}} \right) = b_{o1} + b_{11} x_{1i} + b_{21} x_{2i} \dots + b_{k1} x_{ki}$$

$$\log \left( \frac{p_{2i}}{p_{3i}} \right) = b_{o2} + b_{12} x_{1i} + b_{22} x_{2i} \dots + b_{k2} x_{ki}$$

Separate intercepts and slopes (smaller SEs than for two completely separate binary logistic regressions).

Generalized logits model: Relation between predictor(s) and log odds may differ for each logit.



## VII. Some references about logistic regression

Agresti, A. (1996). *An introduction to categorical data analysis*. Hoboken, NJ: Wiley & Sons Inc.

Agresti, A. (2002). *Categorical data analysis, 2nd edition*. Hoboken, NJ: Wiley & Sons Inc.

Allison, P. (2003). *Logistic regression using the SAS system: Theory and Application*. Cary, NC: SAS Institute.

Collett, D. (2003). *Modeling binary data, 2nd edition*. Boca Raton, FL: Chapman & Hall.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied logistic regression, 2nd edition*. New York: Wiley & Sons, Inc.

Long, S.J. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

McCullagh, P., & Nelder, J. A. (1999). *Generalized linear models, 2nd Edition*. Boca Raton, FL: Chapman & Hall/CRC Press.

Stokes, Davis, & Koch (2000). *Categorical Data Analysis Using the SAS System, 2<sup>nd</sup> ed.* Cary, NC: SAS Institute Inc.





Illustration of the equivalence between the two ways of writing the logistic regression model.

$$p_i = \frac{\exp(x\beta)}{1 + \exp(x\beta)} \quad x\beta = b_0 + b_1x_{1i} + b_2x_{2i} \dots + b_kx_{ki}$$

Multiply both sides by  $(1 + \exp(x\beta))$ :

$$p_i [1 + \exp(x\beta)] = \exp(x\beta)$$

Distribute the  $p$ :

$$p_i + p_i \exp(x\beta) = \exp(x\beta)$$

Subtract  $p(\exp(x\beta))$  from both sides:

$$p_i = \exp(x\beta) - p_i \exp(x\beta)$$

From the previous slide:

$$p_i = \exp(x\beta) - p_i \exp(x\beta)$$

Factor out  $\exp(x\beta)$ :

$$p_i = \exp(x\beta)[1 - p_i]$$

Divide both sides by  $(1-p)$ :

$$\frac{p_i}{[1 - p_i]} = \exp(x\beta)$$

Take the log of both sides:

$$\log \left[ \frac{p_i}{(1 - p_i)} \right] = x\beta$$