Short Course
Categorical Data Analysis

Alan Agresti, Distinguished Professor Emeritus, University of Florida, USA

Presented at AUEB, Athens, Greece
May 18, 19, 20, 21, 2026

Course Outline

Focus of short course

- Overview of most important methods for analyzing categorical response data. Emphasis on concepts, examples of use, interpretations, rather than theory, derivations, technical details.
- Course is based on *Categorical Data Analysis* (3rd ed. 2013), referred to in notes by *CDA*. A less technical version, with `R` examples, is *An Introduction to Categorical Data Analysis* (3rd ed. 2019).
- Examples of analyses use `R` software. For more details, and also SAS, SPSS, and Stata, see

  `https://alanagresti.com/cda/cda.html`

  and the detailed `R` tutorial by Laura Thompson linked there.

CONTINGENCY TABLE ANALYSIS

- Measurement scales and distributions
- Three measures for comparing proportions
- Odds ratios and their properties
- Testing independence — Chi-squared test, standardized residuals
- Loglinear model for two-way tables, odds ratio connection
- Testing independence for small n — Fisher's exact test

Measurement and Distributions

Choice of model, analysis, and interpretation depends on

- Response – explanatory variable distinction
- Measurement scales:
  - **binary** - (favor, oppose), (yes, no), (success, failure)
  - **nominal** - *unordered* categories; for example, choice of transport (car, bus, subway, bike, walk), religious affiliation, choice of product brand, choice of where to shop
  - **ordinal** - *ordered* categories; for example, political ideology (very liberal, somewhat liberal, moderate, somewhat conservative, very conservative), patient recovery (complete, partial, none), quality of life (excellent, good, fair, poor)

Contrasts with **continuous** scales for regression, ANOVA

Measurement	**Probability Distribution for Response**
Continuous	Normal
Categorical	Binomial, multinomial

Binomial : Binary response (0, 1)

Multinomial : Multicategory (nominal or ordinal) response

Cell counts in a contingency table

Data files when explanatory variables are solely categorical can be *grouped* (as in contingency tables) or *ungrouped* (i.e., at the subject level).

Comparing Proportions (*CDA* Sec. 2.2)

ex. Aspirin use and heart attacks (grouped data)

Source: Harvard physicians health study

	Heart Attack		
	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

The data are counts in a 2×2 *contingency table*

Sample proportions of heart attack

$$\begin{aligned} \text{Placebo:} \quad & 189/11{,}034 = 0.0171 \\ \text{Aspirin:} \quad & 104/11{,}037 = 0.0094 \end{aligned}$$

We treat the data as two independent binomial samples.

Three descriptive measures:

1. **Difference of proportions**
   $0.0171 - 0.0094 = 0.0077$
   estimates $P(Y = 1 \mid X = 1) - P(Y = 1 \mid X = 2)$

2. **Risk ratio** (relative risk)
   $0.0171/0.0094 = 1.82$
   estimates $\frac{P(Y=1|X=1)}{P(Y=1|X=2)}$

3. **Odds ratio** (cross-product ratio)
   $\hat{\theta} = \frac{0.0171/0.9829_{\leftarrow(1-0.0171)}}{0.0094/0.9906_{\leftarrow(1-0.0094)}} = 1.83 = \frac{189 \times 10{,}933}{104 \times 10{,}845}$

Odds of heart attack were (number 'yes')/(number 'no')

$$\frac{189}{10{,}845} = 0.0174 \quad \text{for placebo}$$
(i.e., 174 yes for every 10,000 no)

$$\frac{104}{10{,}933} = 0.0095 \quad \text{for aspirin}$$
(i.e., 95 yes for every 10,000 no)

Odds of attack for placebo group were $174/95 = 1.83$ times odds of attack for aspirin group (i.e., 83% higher for placebo group).

For conditional probabilities $\{\pi_{j|i} = P(Y = j \mid X = i)\}$,
joint cell probabilities $\{\pi_{ij} = P(X = i, Y = j)\}$,
cell counts $\{n_{ij}\}$,

$$\begin{array}{lcc} \text{Odds ratio} & \theta & = \frac{\pi_{1|1}/\pi_{2|1}}{\pi_{1|2}/\pi_{2|2}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \\ \text{Sample estimate} & \hat{\theta} & = \frac{n_{11}n_{22}}{n_{12}n_{21}} \end{array}$$

Properties of odds ratio

- $0 \leq \theta \leq \infty$
- $\theta = 1 \leftrightarrow$ no effect

  $(\log \theta = 0)$ ("independence")
- Interchange rows (columns)

  $\theta \to 1/\theta \quad (\log \theta \to -\log \theta)$

  $\theta = 4 \quad (\log \theta = 1.39)$ same strength as

  $\theta = \frac{1}{4} \quad (\log \theta = -1.39)$

- Denote $X =$ row variable, $Y =$ column variable

$$\begin{aligned}\theta &= \frac{P(Y=1 \mid X=1)/P(Y=2 \mid X=1)}{P(Y=1 \mid X=2)/P(Y=2 \mid X=2)} \\ &= \frac{P(X=1 \mid Y=1)/P(X=2 \mid Y=1)}{P(X=1 \mid Y=2)/P(X=2 \mid Y=2)}\end{aligned}$$

$\Rightarrow$ odds ratio θ applicable for *prospective* (e.g., cohort) studies, or *retrospective* (e.g., case-control) studies that observe past behavior on X for matched subjects at the levels of Y.

- Risk ratio $= \frac{P(Y=1|X=1)}{P(Y=1|X=2)}$

is *not* applicable for retrospective studies, but it may be approximated by odds ratio when

$P(Y=1 \mid X=x) \approx 0$ for $x = 1, 2$.

Famous example of case-control study

One of the first studies of the link between lung cancer and smoking, based on data from 20 hospitals in London, England.

For each cancer case, they recorded the smoking behavior (yes = at least 1 cigarette/day for at least 1 year) of a noncancer control at same hospital of same gender and within same 5-year grouping on age.

Table: **Cross-Classification of Smoking by Lung Cancer**

	Lung Cancer	
Smoker	Cases	Controls
Yes	688	650
No	21	59
Total	709	709

Source: R. Doll and A. B. Hill, *British Med. J.*, 1950

Sample odds ratio $\hat{\theta} = (688 \times 59)/(650 \times 21) = 3.0$.

Inference for odds ratio (*CDA* Sec. 3.1)

- $\hat{\theta} \rightarrow$ normal sampling distribution for large n, but $\log \hat{\theta} \rightarrow$ normal much more quickly. Can conduct inferences such as CI for $\log \theta$, take antilogs for inferences about θ.
- Standard error of $\log \hat{\theta}$ is $SE = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$.

Example: Aspirin and heart attacks
95% confidence interval for $\log \theta$ is

$$\log 1.83 \pm 1.96 \sqrt{\frac{1}{189} + \frac{1}{10{,}845} + \frac{1}{104} + \frac{1}{10{,}983}}$$
$$= 0.60 \pm 0.24, \text{ or } (0.36, 0.84)$$

$\rightarrow (e^{0.36}, e^{0.84}) = (1.44, 2.33)$ for θ, called the *Wald* CI.

Conclude $\theta > 1$; i.e., odds of heart attack higher for placebo.

We'll see later that the Wald CI also is a by-product of fitting a logistic regression model.

Chi-Squared Tests of Independence

(*CDA*, Sec. 3.2)

Context: $I \times J$ contingency table

Observed frequencies: $\{n_{ij}\}$, $n_{i+} = \sum_j n_{ij}$, $n_{+j} = \sum_i n_{ij}$, $n = \sum_i \sum_j n_{ij}$

Distributional assumption: $\{n_{ij}\} \sim \mathit{Multinomial}(n, \{\pi_{ij}\})$,
with cell probabilities $\{\pi_{ij} = P(X = i, Y = j)\}$ $(\sum_i \sum_j \pi_{ij} = 1)$

Expected frequencies $\{\mu_{ij} = n\pi_{ij}\}$

Maximum likelihood (ML) estimates $\{\hat{\mu}_{ij} = n\hat{\pi}_{ij}\}$ are called *estimated expected frequencies*, or *fitted values*.

Two responses X and Y are *independent* if

$$\pi_{ij} = P(X = i, Y = j) = P(X = i)P(Y = j) \quad \text{for all } i \text{ and } j$$

Let $\hat{P}(X=i) = n_{i+}/n, \quad \hat{P}(Y=j) = n_{+j}/n$ (sample proportions).

Under H_0 : independence,

$$\begin{aligned} \hat{\pi}_{ij} &= \hat{P}(X=i)\hat{P}(Y=j) \\ \hat{\mu}_{ij} &= n\hat{\pi}_{ij} = n\hat{P}(X=i)\hat{P}(Y=j) = \frac{n_{i+}n_{+j}}{n}. \end{aligned}$$

Compare data $\{n_{ij}\}$ to fit $\{\hat{\mu}_{ij}\}$ using

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \quad \text{(Pearson statistic)}$$

or $$G^2 = 2\sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right) \quad \text{(likelihood-ratio statistic)}$$

measure discrepancy between data $\{n_{ij}\}$ and fit $\{\hat{\mu}_{ij}\}$.

Properties

- Larger values provide stronger evidence against H_0: independence.
- $X^2, G^2 \to \chi^2$ (chi-squared) as $\{\mu_{ij}\} \to \infty$.
- $df =$ difference in dimensions of parameter space between general case ($IJ-1$, because cell probabilities add up to 1) and under H_0 ($I-1$ row prob's and $J-1$ column prob's).
  $df = (IJ-1) - [(I-1)+(J-1)] = (I-1)(J-1)$
- To test H_0 : model holds,

$$\begin{array}{ccc} \text{P-value} & = & P_{H_0}[X^2 \geq X^2_{obs}] \\ \text{or} & & P_{H_0}[G^2 \geq G^2_{obs}]. \end{array}$$

- X^2 is preferable for sparse data (sampling distribution close to chi-squared if most $\hat{\mu}_{ij} \geq 5$), but more computationally complex methods exist (pp. 24–30) that do not require distributional approximations or sample-size restrictions.

Example: Attained education (highest degree) and belief in God (*CDA*, p. 77)

Highest Degree	Belief in God: Don't Believe	No Way to Find Out	Some Higher Power	Believe Sometimes	Believe but Doubts	Know God Exists	Total
Less than high school	9 (10.0)[1] (-0.4)[2]	8 (15.9) (-2.2)	27 (34.2) (-1.4)	8 (12.7) (-1.5)	47 (55.3) (-1.3)	236 (206.9) (3.6)	335
High school or junior college	23 (32.5) (-2.5)	39 (51.5) (-2.6)	88 (110.6) (-3.3)	49 (41.2) (1.8)	179 (178.9) (0.0)	706 (669.4) (3.4)	1084
Bachelor or graduate	28 (17.4) (3.1)	48 (27.6) (4.7)	89 (59.3) (4.8)	19 (22.1) (-0.8)	104 (95.9) (1.1)	293 (358.8) (-6.7)	581
Total	60	95	204	76	330	1235	2000

Source: General Social Survey, National Opinion Research Center.

[1]Estimated expected frequency for testing independence

*e.g., $\hat{\mu}_{11} = \frac{335 \times 60}{2000} = 10.0$

$X^2 = 76.15, \quad G^2 = 73.2, \quad df = (3-1)(6-1) = 10$

P-value $= P_{H_0}[X^2 \geq 76.15] = 0.000...$

[2]standardized residual, $z = (n_{ij} - \hat{\mu}_{ij})/SE$ using std. error of $(n_{ij} - \hat{\mu}_{ij})$

R for analyzing data on education and belief in God:

```
------------------------------------------------------------------------------
>  data <- matrix(c(9,8,27,8,47,236,23,39,88,49,179,706,28,48,89,19,104,293),
                  ncol=6,byrow=TRUE)
> chisq.test(data)

        Pearson's Chi-squared test
X-squared = 76.1483, df = 10, p-value = 2.843e-12

> chisq.test(data)$stdres   #  standardized residuals
          [,1]      [,2]      [,3]      [,4]       [,5]      [,6]
[1,] -0.368577 -2.227511 -1.418621 -1.481383 -1.3349600  3.590075
[2,] -2.504627 -2.635335 -3.346628  1.832792  0.0169276  3.382637
[3,]  3.051857  4.724326  4.839597 -0.792912  1.0794638 -6.665195
------------------------------------------------------------------------------
```

Standardized residuals tell us, e.g., that significantly fewer college-educated people "know God exists" than if belief in God were independent of highest degree.

Loglinear model of independence

Independence between two categorical variables can be expressed as a **loglinear model**.

$$\pi_{ij} = P(X = i)P(Y = j)$$

$$\mu_{ij} = nP(X = i)P(Y = j)$$

$$\Updownarrow$$

$$\log\ \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

for effect of being in row i of variable X and in column j of variable Y.

Analogous to main effects model in 2-way ANOVA for two factor explanatory variables. We'll present such models in Section 4 of this course and see that software imposes identifiability constraints, such as $\lambda_1^X = \lambda_1^Y = 0$ in R.

Fisher's Exact Test (*CDA*, Sec. 3.5)

For small n, *exact* test of independence uses the hypergeometric distribution. This results from conditioning on margin totals, the "sufficient statistics" for unknown marginal probabilities.

ex. 2×2 contingency table

n_{11}		n_{1+}
		n_{2+}
n_{+1}	n_{+2}	n

$$p_{H_0}(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}}$$

e.g, for $H_0 : \theta = 1$, $H_a : \theta > 1$ (odds ratio)

P-value $= P_{H_0}[n_{11} \geq$ observed $n_{11}]$.

Example: Fisher's tea taster

	Guess Poured First		
Poured First	Milk	Tea	
Milk	3	1	4
Tea	1	3	4
	4	4	8

$$P = \frac{\binom{4}{3}\binom{4}{1} + \binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = 0.243$$

The attainable P-values for these margins are

0.014, 0.243, 0.757, 0.986, 1.00,

for $n_{11} = 4, 3, 2, 1, 0$.

R for Fisher's exact test with tea tasting data:

```
-------------------------------------------------------------------
> tea <- matrix(c(3,1,1,3), ncol=2, byrow=TRUE)

> fisher.test(tea, alternative="greater")

^^IFisher's Exact Test for Count Data

data:  tea
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1

> fisher.test(tea)

^^IFisher's Exact Test for Count Data

data:  tea
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.2117329 621.9337505
sample estimates:
odds ratio  #  This is the "conditional ML" estimate (CDA, p. 267-268)
  6.408309
-------------------------------------------------------------------
```

Notes

- Fixing both margins makes sampling distribution highly discrete.

  Fisher's exact test is *conservative*; i.e., when H_0 true, may reject H_0 at 0.05 level much less than 5% of time.

  For tea-tasting data, actual $P(\text{Type I error}) = 0.014$, not 0.05.

- Alternatives to reduce conservativeness in 2×2 case include
  - "exact" tests that fix only row marginal totals (*CDA*, Sec. 3.5.6).
  - **mid-P value** (*CDA*, Sec. 1.4.4)
    $P = \frac{1}{2}P(n_{11} = n_{11} \text{ obs.}) + P(n_{11} > n_{11} \text{ obs.})$.

    $\rightarrow$ Satisfies $E_{H_0}(P) = 0.50$.

```
> library(epitools)
> ormidp.test(3, 1, 1, 3, or=1) # enter the four cell counts and H0 value
  one.sided  two.sided
  0.12857    0.25714  # mid P-values for testing independence
```

- Fisher exact test generalizes to $I \times J$ and $I \times J \times K$ tables, with simulation used to precisely approximate the P-value when computations take too long. It also generalizes to "exact" CIs for odds ratios (*CDA*, Sec. 16.5, 16.6).

```
------------------------------------------------------------------------------
>  data <- matrix(c(9,8,27,8,47,236,23,39,88,49,179,706,28,48,89,19,104,293),
           ncol=6, byrow=TRUE) # education and belief in God data
> chisq.test(data)
        Pearson's Chi-squared test
X-squared = 76.1483, df = 10, p-value = 2.843e-12

> fisher.test(data, simulate.p.value=TRUE, B=10000000) # B simulations

p-value = 1e-07 # close approximation for exact P-value
------------------------------------------------------------------------------
```

Summary

- Categorical data: binary, nominal, ordinal
- Binomial and multinomial sampling models for categorical data
- Describing association:
  difference of proportions, relative risk, odds ratio
- Inference for odds ratio
- Independence for $I \times J$ table is a loglinear model.
  G^2 and X^2 chi-squared tests of independence (large n)
  Fisher's exact test (small n)

LOGISTIC REGRESSION FOR BINARY RESPONSE VARIABLES

- Logistic regression as a generalized linear model
- Binary Y, continuous x
- Binary Y, categorical x, multiple predictors
- Likelihood-based inference for model parameters
- Goodness of fit

Notation:

Y_i = response outcome for subject i

x_{ij} = value of explanatory variable j for subject i

Generalized Linear Models (*CDA*, Chap. 4)

1. Random component

   Choose response variable Y and its distribution (normal, binomial, Poisson, gamma, negative binomial, ...).
   Let $E(Y_i) = \mu_i$ for subject i.

2. Linear predictor

   Choose p explanatory variables.

   The linear predictor is $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$

3. Link function

   Monotone function g relating linear predictor to μ_i.

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Deviance

For GLMs, analysis of variance generalizes to analysis of likelihood functions, called analysis of *deviance*.

Let maximized log likelihood $= L_M$ for the model.

L_S = maximized log likelihood for "saturated model"
(perfect fit: parameter for each observation).

Deviance $= 2[L_S - L_M]$, which equals G^2 for models for categorical data.

References on generalized linear models

- J. Nelder and R. Wedderburn, *J. Roy. Stat. Soc., A*, 1972
- P. McCullagh and J. Nelder, *Generalized Linear Models*, 2nd ed., Chapman and Hall, 1989
- A. Agresti, *Foundations of Linear and Generalized Linear Models*, Wiley, 2015

Logistic regression as a GLM

- Random component: Binomial for ungrouped or grouped binary data

  Ungrouped: $p(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i},\ y_i = 0, 1.\ [p(1) = \pi_i,\ p(0) = 1-\pi_i]$

  $$\text{Grouped: } p(y_i) = \binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i},\ y_i = 0, 1, \ldots, n_i.$$

- Link: $g(\pi_i) = \log(\frac{\pi_i}{1-\pi_i})$
  logit link = natural parameter in exponential family representation (called the *canonical* link function).

  Other popular link functions include

  Identity: $g(\pi_i) = \pi_i$ (ok if $0.2 \leq \pi \leq 0.8$)

  Probit: $g(\pi_i) = \Phi^{-1}(\pi_i)$ (underlying normal latent variable)

  Complementary log-log: $g(\pi_i) = \log[-\log(1-\pi_i)]$ (nonsymmetric)

Example: Beetles killed after exposure to carbon disulfide

This is the dataset analyzed in the original binary modeling paper, by Bliss (1935), with an appendix written by R. A. Fisher showing how to use an iterative method, now called "Fisher scoring," to find the ML estimates.

Dosage	Number of Beetles	Number Dead	Proportion Dead
1.691	59	6	0.10
1.724	60	13	0.22
1.755	62	18	0.29
1.784	56	28	0.50
1.811	63	52	0.83
1.837	59	53	0.90
1.861	62	61	0.98
1.884	60	60	1.00

We regard these grouped data as 8 binomial samples.

Tolerance distribution justification for binary regression

Consider binary response $y_i = 1$ (death) or 0 (survive), dosage x of toxic substance.

Suppose subject i has tolerance T_i, with

$$y_i = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \leftrightarrow \begin{matrix} x \geq T_i \\ x < T_i \end{matrix}$$

For population, suppose $P(T \leq t) = G(t)$ (unknown).

Then,
$P(y = 1|x) = P(T \leq x) = G(x)$

$= F(\alpha + \beta x)$ for a "standard" cumulative distribution function (cdf) F.

This suggests a model of form

$F^{-1}[P(y = 1|x)] = \alpha + \beta x$ for some standard cdf F.

Example: F = standard normal cdf Φ

$P(y = 1|x) = \Phi(\alpha + \beta x)$ is called the *probit* model.
(Response curve looks like normal cdf with $\mu = -\alpha/\beta$, $\sigma = 1/\beta$.)

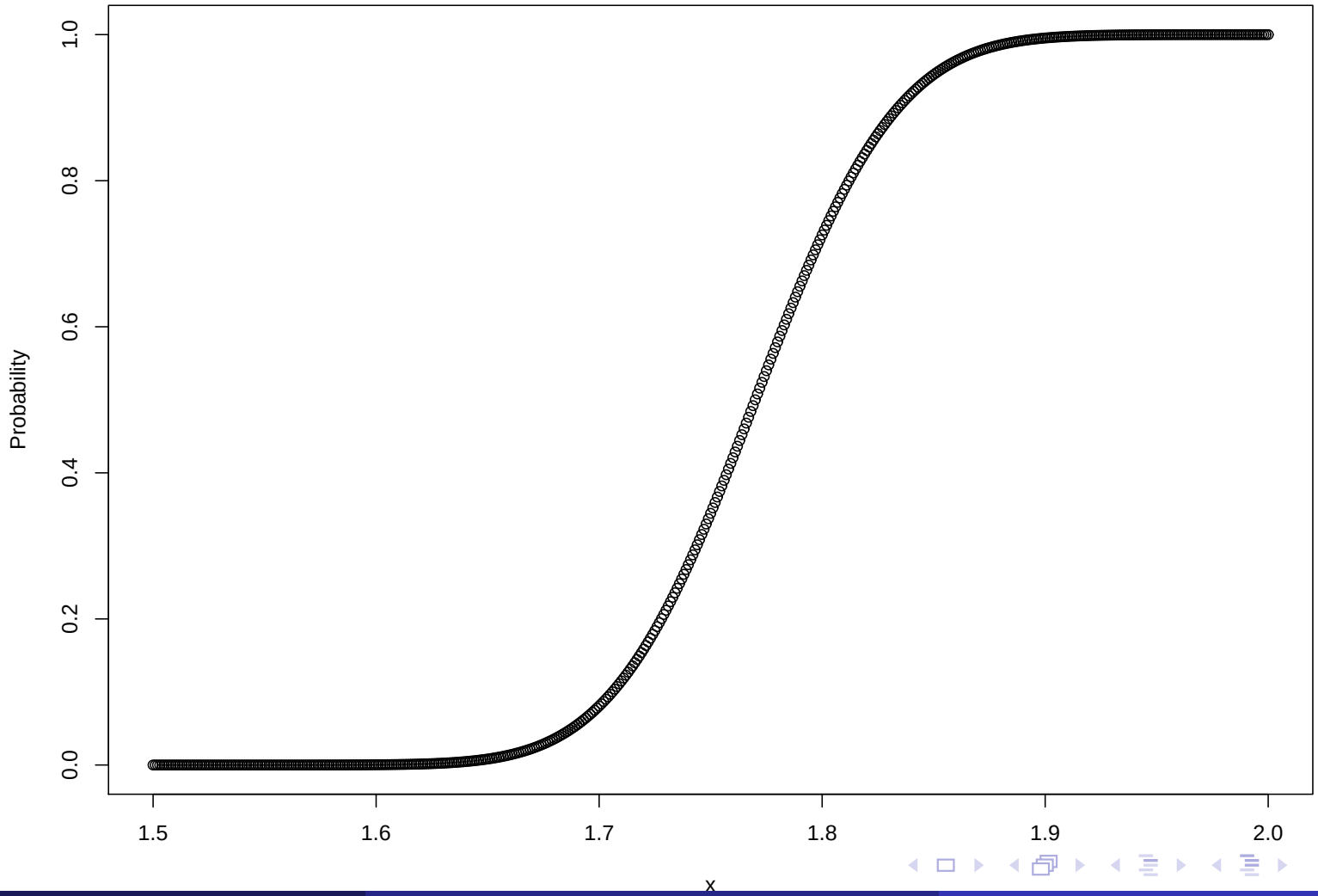
The link Φ^{-1} is the *probit link* (Bliss, 1935).

R software code for the probit model for the beetle mortality data:

```
------------------------------------------------------------------------------
> logdose <- c(1.691, 1.724, 1.755, 1.784, 1.811, 1.837, 1.861, 1.884)
> dead <- c(6, 13, 18, 28, 52, 53, 61, 60)  # numbers of dead beetles
> n <- c(59, 60, 62, 56, 63, 59, 62, 60)    # binomial sample sizes

> fit.probit <- glm(dead/n ~ logdose, family=binomial(link=probit), weights=n)

> summary(fit.probit)
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -34.956       2.649   -13.20    <2e-16
logdose       19.741       1.488    13.27    <2e-16
---
    Null deviance: 284.202 on 7 degrees of freedom
Residual deviance:   9.987 on 6 degrees of freedom # goodness of fit
AIC: 40.185
------------------------------------------------------------------------------
```

The fit looks like a normal cumulative distribution function with mean $(34.956/19.741) = 1.77$ and standard deviation $1/19.741 = 0.05$.

Probit model for the probability of death



Example: $F(x) = \frac{e^x}{1+e^x}$ = standard logistic (bell-shaped density function, mean 0, standard dev. 1.81).

$$P(y = 1|x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

is the *logistic regression* model, for which

$$P(y = 0|x) = 1 - P(y = 1|x) = \frac{1}{1 + e^{\alpha+\beta x}}.$$

Response curve looks like logistic cdf with $\mu = -\alpha/\beta$, $\sigma = 1.81/\beta$.

Probit and logit have similar fits with $\hat{\beta}_{logit} \approx 1.8\hat{\beta}_{probit}$.

R software code for logistic model for beetle mortality data:

```
------------------------------------------------------------------------------
> fit.logit <- glm(dead/n ~ logdose, family=binomial(link=logit), weights=n)
> summary(fit.logit)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.740      5.182  -11.72   <2e-16
logdose       34.286      2.913   11.77   <2e-16
---
    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.116  on 6  degrees of freedom
AIC: 41.314

> cbind(logdose, dead/n, fitted(fit.probit), fitted(fit.logit))
  logdose
1   1.691 0.1016949 0.0577367 0.05937747
2   1.724 0.2166667 0.1781060 0.16366723
3   1.755 0.2903226 0.3780390 0.36162283
4   1.784 0.5000000 0.6032833 0.60490961
5   1.811 0.8253968 0.7866532 0.79440490
6   1.837 0.8983051 0.9045852 0.90405532
7   1.861 0.9838710 0.9626183 0.95546748
8   1.884 1.0000000 0.9873227 0.97925643
------------------------------------------------------------------------------
```

Logistic fit looks like logistic cdf with mean $(60.740/34.286) = 1.77$ and standard deviation $(1.81/34.286) = 0.05$.

Note: Ratio of logistic and probit effects is $34.286/19.741 = 1.74$.

Note: *Complementary log-log* link (beyond our scope) yields non-symmetric curve and slightly better fit.

Interpreting parameters in logistic regression

(*CDA*, Sec. 5.1)

For a single explanatory variable and $\pi(x) = P(y = 1|x)$,

$$\text{odds} = \frac{\pi(x)}{1 - \pi(x)} = e^{\alpha+\beta x}.$$

For two levels x and $x + 1$,

$$\begin{aligned} \text{odds ratio} &= \frac{\pi(x+1)/[1 - \pi(x+1)]}{\pi(x)/[1 - \pi(x)]} \\ &= e^{\alpha+\beta(x+1)} / e^{\alpha+\beta x} \\ &= e^{\beta}. \end{aligned}$$

The odds of a success at $x + 1$ are e^{β} times odds of success at x; i.e., odds multiply by e^{β} for every 1-unit increase in x.

$\beta = 0 \leftrightarrow$ odds ratio $= 1 \leftrightarrow$ no effect.

Linearized form of model uses *logit* transform

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x.$$

This is a GLM with binomial random component and logit link.

Generalizes to multiple logistic regression

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Then, $\exp(\beta_j)$ represents odds ratio between Y and two levels of x_j that are 1-unit apart, adjusting for other explanatory variables in the model.

The term "logit" and introduction of logistic regression modeling originated in series of articles beginning in 1944 by Joseph Berkson, a biostatistician at Mayo Clinic.

- Odds $= \frac{\pi(x)}{1-\pi(x)} = e^{\alpha}(e^{\beta})^{x}$; i.e., multiplicative effect
- Monotone

$$\begin{aligned} \beta > 0: &\quad \pi(x) \uparrow 1 \text{ as } x \to \infty \\ \beta < 0: &\quad \pi(x) \downarrow 0 \text{ as } x \to \infty \\ \beta = 0: &\quad \pi(x) \text{ constant} \end{aligned}$$

- $\frac{\partial \pi(x)}{\partial x} = \beta\pi(x)(1-\pi(x))$,

  so slope is proportional to β and steepest $(\beta/4)$ at x-value where $\pi(x) = 0.50$; this x value is $x = -\alpha/\beta$.
- Valid with retrospective studies, essentially because odds ratio for Y given x is the same as for X given y.

  (*CDA*, Sec. 5.1.4)

Example: Y = cancer remission (1 = yes, 0 = no),
x = labeling index (LI) = percentage of "labeled" cells, describes increased activity of blood cells after injection of tritiated thymidine.

LI	No. cases	No. remissions	
8	2	0	← i.e, Y = 0 for each of
10	2	0	2 observations (grouped data)
12	3	0	
14	3	0	
16	3	0	
18	1	1	
20	3	2	
22	2	1	
24	1	0	
26	1	1	
28	1	1	
32	1	0	
34	1	1	
38	3	2	
	27		

With software, you can enter data as 27 Bernoulli outcomes (*ungrouped* data) or 14 binomials (*grouped* data); likelihood function and ML estimates are the same either way.

Deviance $= 2[L_S - L_M]$ differs for the two types of data files, because saturated model has more parameters for the ungrouped data file.

R for cancer remission data (grouped data file):

```
> Remission <- read.table("https://alanagresti.com/cda/CDA_data/Remission.dat",
+                         header=TRUE) # you can copy data file from this site
> Remission
   LI cases remissions
1   8     2          0
2  10     2          0
3  12     3          0
4  14     3          0
5  16     3          0
6  18     1          1
7  20     3          2
8  22     2          1
9  24     1          0
10 26     1          1
11 28     1          1
12 32     1          0
13 34     1          1
14 38     3          2
> fit <- glm(remissions/cases ~ LI, family=binomial(link=logit), weights=cases, data=Remission)
> summary(fit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.77714    1.37863  -2.740  0.00615
LI           0.14486    0.05934   2.441  0.01464
---
    Null deviance: 23.961  on 13  degrees of freedom
Residual deviance: 15.662  on 12  degrees of freedom
```

R for cancer remission data (ungrouped data file):

```
-------------------------------------------------------------------------
> LI <- c(8,8,10,10,12,12,12,14,14,14,16,16,16,18,20,20,20,22,22,24,26,
+         28,32,34,38,38,38)
> y <- c(0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,1,0,0,1,1,0,1,1,1,0)

> logistic.fit <- glm(y ~ LI, family=binomial(link=logit))
> summary(logistic.fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.77714    1.37862  -2.740  0.00615
LI           0.14486    0.05934   2.441  0.01464
---
    Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 26.073  on 25  degrees of freedom

> confint(logistic.fit)  # to get "profile likelihood" CI's for parameters
                 2.5 %     97.5 %
(Intercept) -6.9951909 -1.4098443
LI           0.0425232  0.2846668

> exp(-3.77714+0.14486*20.1)/(1+exp(-3.77714+0.14486*20.1))
[1] 0.2962011 # estimated probability at mean of LI of 20.1

> cbind(LI, y, fitted(logistic.fit))
   LI y
1   8 0 0.06797405 # fitted() gives model-estimated probabilities
2   8 0 0.06797405
3  10 0 0.08878928
...
26 38 1 0.84911301
27 38 0 0.84911301
-------------------------------------------------------------------------
```

Logistic regression ML fit for π = probability of remission:

$$\log\left[\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right] = -3.777 + 0.145x.$$

Prediction equation

$$\hat{\pi}(x) = \frac{\exp(\hat{\alpha}+\hat{\beta}x)}{1+\exp(\hat{\alpha}+\hat{\beta}x)} = \frac{\exp(-3.777+0.145x)}{1+\exp(-3.777+0.145x)}.$$

$$\text{e.g. at } x = \bar{x} = 20.1,$$

$$\hat{\pi}(x) = \frac{\exp[-3.777+0.14486(20.1)]}{1+\exp[-3.777+0.14486(20.1)]} = 0.296.$$

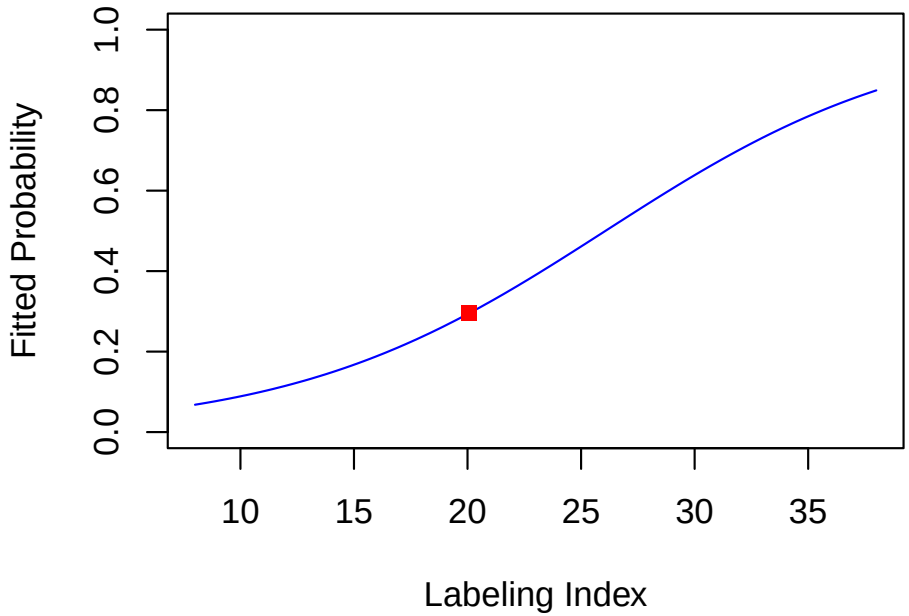
The incremental rate of change at $x = 20.1$ is
$\hat{\beta}\hat{\pi}(x)\,[1-\hat{\pi}(x)] = 0.14486(0.296)(0.704) = 0.030.$

$$\hat{\pi}(x) = 0.50 \leftrightarrow \log\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) = 0 = \hat{\alpha}+\hat{\beta}x$$

$$\leftrightarrow x = -\hat{\alpha}/\hat{\beta} = 26.0$$

Using R to show fitted probabilities and fit at the mean of x = labeling index:

```
> ## To predict probability at mean LI:
> mean(LI)
[1] 20.07407
> # predicted log-odds:
> predict(logistic.fit,
+   newdata=data.frame(LI=20.07))
-0.8697359
> # predicted probability:
> predict(logistic.fit,
+   newdata=data.frame(LI=20.07),
+   type="response")
0.2953093
> ## Plotting
> myLI <- seq(min(LI),max(LI), length.out=100)
> fitted.probs <- predict(logistic.fit,
+   newdata=data.frame(LI=myLI),
+   type="response")
> plot(fitted.probs~myLI, type="l",
+   col="blue", xlab="Labeling Index",
+   ylab="Fitted Probability", ylim=c(0,1))
> points(x=20.07, y=0.295, col="red", pch=15)
```



Odds ratio interpretation of LI effect:

$\hat{\beta} = 0.145$ means that for each unit change in LI, estimated odds of remission are multiplied by $\exp(0.145) = 1.16$.
i.e., 16% increase when LI $\uparrow$ 1.

$$\begin{aligned} \text{e.g., at } x &= 26, \ \hat{\pi}(x) = 0.498 \ (\text{odds} = 0.990) \\ \text{at } x &= 27, \ \hat{\pi}(x) = 0.534 \ (\text{odds} = 1.145 \\ &\qquad = 0.990 \times 1.16). \end{aligned}$$

i.e., odds ratio $= \frac{0.534/(1-0.534)}{0.498/(1-0.498)} = 1.16$.

Simpler probability-based effect measures:

Change in $\hat{\pi}(x)$ from minimum to maximum value of x.

Can be misleading if x has an extreme outlier, in which case report change in $\hat{\pi}(x)$ between lower and upper quartiles of x in ungrouped data file, which is more resistant.

Following R code shows $\hat{\pi}(x)$ increases from 0.07 to 0.85 over range of LI values; also, $LQ = 13$, $UQ = 25$, $\hat{\pi}(x)$ goes from 0.13 to 0.46 over middle half of data.

```
------------------------------------------------------------------------------
> predict(logistic.fit, data.frame(LI=quantile(LI)), type="response")
        0%        25%        50%        75%       100%
0.06797405 0.13079831 0.23692681 0.46118815 0.84911301
------------------------------------------------------------------------------
```

With multiple predictors, can report these measures with other explanatory variables set at their means.

Average marginal effect: A simple summary effect measure

Average the probability rate of change at the n sample values of the explanatory variable (for ungrouped data). Some books and software refer to this measure as an *average marginal effect*.

In R:

```
------------------------------------------------------------------------
> fit <- glm(y ~ LI, family=binomial(link=logit), data=Remission)
> summary(fit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.77714    1.37862  -2.740  0.00615
LI           0.14486    0.05934   2.441  0.01464
---
> library(mfx)
> logitmfx(fit, data=Remission, atmean=FALSE)

Marginal Effects:
      dF/dx Std. Err.     z P>|z|
LI 0.022584  0.014722 1.534 0.125
------------------------------------------------------------------------
```

The average rate of change in the probability of remission per unit change in LI, calculated at the 27 observed LI values, is 0.0226.

Inference for logistic regression: Tests and CIs

ML Fitting: Likelihood equations are nonlinear in β.

$\hat{\beta}$ obtained iteratively using Newton-Raphson method. This provides a sequence of parabolic approximations to log-likelihood L, which is concave. Equivalently, Fisher scoring uses iteratively reweighted least squares.

Test of no effect $H_0 : \beta = 0$: $\hat{\beta} = 0.145$, est. std. error $SE = 0.059$

- $z = (\hat{\beta} - 0)/SE = 2.45$ (like t statistic for normal response)
  ($z^2 = 5.96 \sim \chi^2$, under H_0, called Wald statistic)
  Strong evidence of a positive association ($P = 0.015$).

- Likelihood-ratio test statistic, using L_0 for H_0, L_1 for H_1 is
  $2(L_1 - L_0) = 2(L_S - L_0) - [2(L_S - L_1)]$ = increase in deviance when impose H_0 constraint. Here, equals
  $23.96 - 15.66 = 34.37 - 26.07 = 8.30$, $df = 1$ ($P = 0.004$).

  Likelihood-ratio test is often more powerful than Wald test, especially when $|\beta|$ large, so probability is near 0 or near 1 (*CDA*, Sec. 5.2.6).

Inference: Confidence intervals

Similarly, there are Wald and likelihood-ratio confidence intervals, based on inverting these tests; e.g., 95% confidence interval is the set of β_0 not rejected at the 0.05 level in testing $H_0 : \beta = \beta_0$ against $H_a : \beta \neq \beta_0$.

- Best known is "Wald interval," $\hat{\beta} \pm 1.96(SE)$.

  e.g., for a binomial parameter π, $\hat{\pi} \pm 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$

- The likelihood-ratio test-based profile-likelihood CI is the set of β_0 for which 2(log-likelihood) decreases by less than $(1.96)^2$, which is the 0.95 percentile of chi-squared with $df = 1$.

  For small n or parameter taking value near boundary (and ML estimate possibly infinite, as in example in next section), this is preferable to Wald CI.

  R: *confint* function

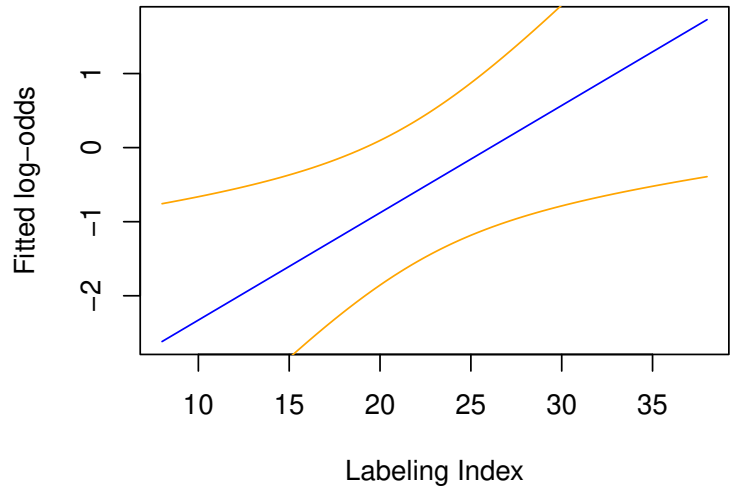
Logistic regression used to get confidence interval for odds ratio for aspirin/placebo group and heart attack outcome (p. 7):

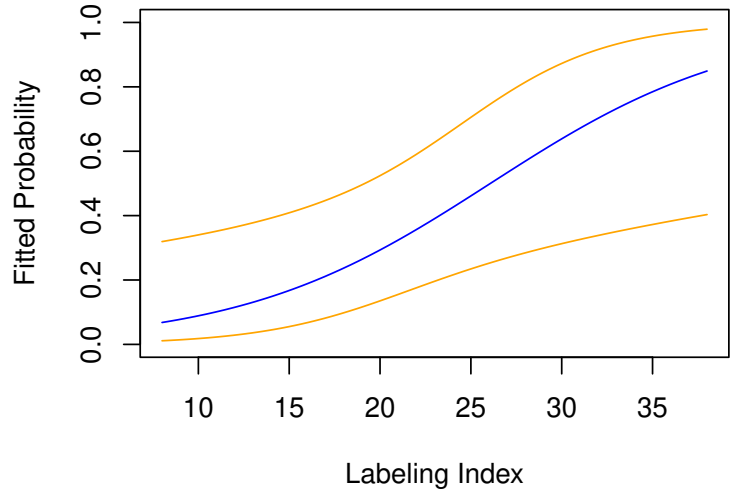
```
> library(epitools) # uses the four cell counts  (i.e., grouped data)
> oddsratio(c(189,10845,104,10933), method=c("wald"), conf=0.95, correct=FALSE)
     odds ratio with 95% C.I.
     estimate    lower   upper
     1.832054 1.440042 2.33078 # Wald CI for odds ratio that we got before

  # also get it by fitting logistic regression model

> group <- c(1,0) # two groups to be compared, 1 = placebo
> attack <- c(189, 104) # binomial outcomes for the two groups
> n <- c(189 + 10845, 104 + 10933) # binomial sample sizes
> fit <- glm(attack/n ~ group, weights=n, family=binomial(link=logit))
> summary(fit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.65515    0.09852 -47.249  < 2e-16
group        0.60544    0.12284   4.929 8.28e-07 # 0.60544 = estimated log OR
---
    Null deviance:  2.5372e+01  on 1  degrees of freedom
Residual deviance: -2.0983e-13  on 0  degrees of freedom

> exp(0.60544 - 1.96*(0.12284)); exp(0.60544 + 1.96*(0.12284)) # Wald CI
[1] 1.440044
[1] 2.330788

> exp(confint(fit))  # profile likelihood CI for odds ratio

                  2.5 %     97.5 %
(Intercept) 0.007792767 0.01147071
group       1.443579310 2.33786091
```

Finding **confidence bands** around predicted probabilities

Example: R code for Remission Data

```
> vcov(logistic.fit)
             (Intercept)           LI
(Intercept)  1.90060430 -0.076525261
LI          -0.07652526  0.003521352
>
> ## Confidence bands for predicted probability
> # First, get them on the log-odds scale!
> fitted.logodds <- predict(logistic.fit,
+   newdata=data.frame(LI=myLI), se.fit=TRUE)
> str(fitted.logodds)
List of 3
 $ fit: -2.62 -2.6 -2.59 -2.57 ...
 $ se.fit: 0.95 0.944 0.939 0.934 ...
> LB.logodds <- fitted.logodds$fit -
+   1.96*fitted.logodds$se.fit
> UB.logodds <- fitted.logodds$fit +
+   1.96*fitted.logodds$se.fit
> plot(fitted.logodds$fit~myLI, type="l", col="blue",
+   xlab="Labeling Index", ylab="Fitted log-odds")
lines(LB.logodds~myLI, col="orange")
lines(UB.logodds~myLI, col="orange")

> # Then, transform back to the probability scale
> LB.prob <- exp(LB.logodds)/(1+exp(LB.logodds))
> UB.prob <- exp(UB.logodds)/(1+exp(UB.logodds))
> plot(fitted.probs~myLI, ylim=c(0,1), type="l",
+   col="blue", xlab="Labeling Index",
+   ylab="Fitted Probability")
> lines(LB.prob~myLI, col="orange")
> lines(UB.prob~myLI, col="orange")
```





As in ANOVA, qualitative factors can be explanatory variables in logistic regression models, using indicator (dummy) variables. (*CDA*, Sec. 5.3)

Example: Model for comparing groups with stratified data

$Y =$ response (success, failure)

$X =$ group (e.g., treatments in a clinical trial)

$Z =$ control variable such as age, gender, race, location, clinic

Main effects model is

$$\log\left[\frac{P(Y=1 \mid X=i, Z=k)}{P(Y=0 \mid X=i, Z=k)}\right] = \alpha + \beta_i^X + \beta_k^Z.$$

For each stratum, the odds of success for group 1 are $\exp(\beta_1^X - \beta_2^X)$ times those for group 2.

For identifiability, software uses constraints (such as $\beta_1^X = \beta_1^Z = 0$ in R), when declare categorical explanatory variable to be a factor.

For two groups, equivalent to set up indicator $x = 0$ for group 1, $x = 1$ for group 2, and use βx in linear predictor, where $\beta = \beta_2^X$.

Example: AIDS symptoms and AZT use, by race
(*CDA*, Sec. 5.4.2)
338 subjects whose immune systems were beginning to falter after infection with HIV were randomly assigned either to receive AZT immediately or to wait until T cells showed severe immune weakness.

$Y =$ whether developed AIDS symptoms during 3-year study.

	AZT	AIDS Symptoms		Proportion
Race	Use	Yes	No	Yes
White	Yes	14	93	0.13
	No	32	81	0.28
Black	Yes	11	52	0.17
	No	12	43	0.22

$\pi_{ij} =$ prob(AIDS symptoms = yes) for AZT use i and race j.

Logit model with categorical explanatory variables

$$\log\left[\frac{\pi_{ij}}{1-\pi_{ij}}\right] = \alpha + \beta_i^A + \beta_j^R$$

for effect β_i^A of cat. i of AZT use and effect β_j^R of cat. j of race.

R for AIDS and AZT use example, which sets parameter = 0 at first category (alphabetized) of a factor

```
> Aids <- read.table("https://alanagresti.com/cda/CDA_data/AIDS.dat", header=TRUE)
> Aids
   race azt yes no
1 white yes  14 93
2 white  no  32 81
3 black yes  11 52
4 black  no  12 43

> fit <- glm(yes/(yes+no) ~ race + azt, family=binomial, weights=yes+no, data=Aids)
> summary(fit)
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.07357    0.26294  -4.083 4.45e-05
racewhite          0.05548    0.28861   0.192  0.84755
aztyes            -0.71946    0.27898  -2.579  0.00991
---
    Null deviance: 8.3499  on 3  degrees of freedom
Residual deviance: 1.3835  on 1  degrees of freedom

> library(car)
> Anova(fit) # provides likelihood-ratio tests of explanatory variables

     LR Chisq Df Pr(>Chisq)
race   0.0371  1   0.847295
azt    6.8709  1   0.008761
```

Using factor command corresponds to expressing the model using indicator variables as

$$\text{logit}[P(\text{AIDS symptoms} = \text{yes})] = \alpha + \beta_1 A + \beta_2 R$$

where

$$A = \begin{array}{ll} 0 & \text{no on AZT use} \\ 1 & \text{yes on AZT use} \end{array}$$

$$R = \begin{array}{ll} 0 & \text{black subject} \\ 1 & \text{white subject} \end{array}$$

As in ordinary linear models, with k categories for a predictor, software forms $k-1$ indicator variables. Then, $df = k-1$ for testing its effect.

Adjusting for race, odds of AIDS symptoms for those using AZT estimated to be $e^{-0.719} = 0.49$ times the odds for those not using it.

95% Wald CI for log odds ratio is $-0.719 \pm 1.96\ (0.279)$, or $(-1.27, -0.17)$, which translates to $(e^{-1.27}, e^{-.17}) = (0.28, 0.84)$ for the odds ratio, which does not contain "no effect" value of 1.0

Similarly, the estimated odds of symptoms when subject was white are $e^{0.055} = 1.06$ times the estimated odds when subject was black, given AZT status.

This estimate is not significantly different from 1.0.

```
-------------------===-----------------------------------------------------
> confint(fit) # profile likelihood CIs for log odds ratios
                 2.5 %     97.5 %
(Intercept) -1.6088054 -0.5734959
racewhite   -0.5022982  0.6334104
aztyes      -1.2773237 -0.1798769 # similar to Wald CI of (-1.27, -0.17)
-------------------------------------------------------------------------
```

Binary regression goodness of fit (*CDA*, Sec. 4.5, 5.2)

For categorical **x** with *grouped* data file and most counts ≥ 5, we can use X^2 or G^2 to test fit of model (i.e., test H_0: model holds).

If n_i subjects at i^{th} setting of **x**,
$\hat{\pi}_i =$ estimated $P(Y = 1)$ at that setting,
$\hat{\mu}_{i1} = n_i\hat{\pi}_i =$ predicted number of $(Y = 1)$ observations.
$\hat{\mu}_{i2} = n_i(1 - \hat{\pi}_i) =$ predicted number of $(Y = 0)$ observations.

Substitute counts of $(Y = 1)$ and $(Y = 0)$ and the fitted values into X^2 or G^2 (the deviance) to get statistic for testing fit.

$$X^2 = \sum(\text{observed} - \text{fitted})^2/\text{fitted}$$

$$G^2 = 2\sum \text{observed} \log\left(\frac{\text{observed}}{\text{fitted}}\right)$$

When $\{\mu_i\}$ large, X^2 and $G^2 \to \chi^2$
$df =$ number of binomial observations $-$ number of model parameters

Example: AIDS symptoms and AZT use

Doing this for all 4 AZT–Race combinations yields sum over 8 cells,

$$G^2 = 1.38, \quad X^2 = 1.39.$$

For the $2 \times 2 \times 2$ table, there are $2 \times 2 = 4$ binomial observations,
3 parameters (nonredundant) in model $\text{logit}(\pi) = \alpha + \beta_1 A + \beta_2 R$
$df = 4 - 3 = 1$, so the fit is adequate.

G^2 and X^2 test H_0: model holds; i.e., logit model with additive A and R effects holds, meaning there is no interaction between A and R in their effects on P(AIDS symptoms = yes).

In GLM literature and `R` output, G^2 is the (residual) *deviance*.

- X^2 and $G^2 \not\to \chi^2$ when data sparse
  - small n
  - large n, lots of cells, small cell counts
  - continuous x

  Can then judge goodness of fit by comparing model to more complex models (e.g., with interaction terms).
- Likelihood-ratio statistic for testing $H_0 : \beta_1 = 0$ is $2(L_1 - L_0) = 2(L_S - L_0) - [2(L_S - L_1)]$ = difference in deviances for models with and without AZT term. This is same for Bernoulli (ungrouped) or binomial (grouped) data file, because although L_S differs for them, L_S cancels in finding $2(L_1 - L_0)$.

```
------------------------------------------------------------------------------
> deviance(glm(yes/(yes+no) ~ race + azt, family=binomial, weights=yes+no, data=Aids))
[1] 1.38353
> deviance(glm(yes/(yes+no) ~ race, family=binomial, weights=yes+no, data=Aids))
[1] 8.254436
------------------------------------------------------------------------------
```

Here, LR stat $= 8.25 - 1.38 = 6.87$, $df = 2 - 1 = 1$ P-value $= 0.009$, as reported in R output using Anova function.

Note: We don't believe any model truly holds, but more complex models than needed may provide poorer estimates of characteristics of interest.

Summary

- **Generalized linear models** useful for discrete or continuous data. Select Y and its probability distribution (*random component*), explanatory variables for *linear predictor*, *link function* of mean to model. Same ML algorithm applies for all such models.
- Logit link useful for *binary* response, continuous or categorical x or both (GLM with binomial response and logit link function).
- Logit effects interpreted multiplicatively using odds, odds ratios.
- ML model fitting requires iterative estimation (details not shown here), but simple with modern software.
- Inference (tests, CI's) use likelihood in various ways, with likelihood-ratio methods preferable to Wald methods when n small or parameters near boundary, such as when $P(Y = 1)$ is near 0 or 1.

LOGISTIC REGRESSION MODEL BUILDING

(*CDA*, Chap. 6)

- Selecting explanatory variables for a model
- Example illustrating a backward elimination process
- Residual analysis: Detecting where model fits poorly
- Infinite estimates and remedies: Penalized likelihood, Bayesian approach
- Remedy with very large number of predictors: Lasso

Selecting Predictors for a Logistic Model (*CDA*, Sec. 6.1)

- Analogs exist of stepwise automatic selection procedures. e.g., backward elimination starts with all predictors, eliminates according to which has the largest P-value for a test of its effect.
- Usual regression considerations apply, such as (1) when some predictors are highly correlated, simpler models may achieve nearly as high a maximized log likelihood, (2) for very large n, an effect (e.g., a complex interaction) may be statistically significant without being practically useful (i.e., model parsimony is a sensible objective).
- Rough guideline – number of outcomes of each type (S, F) should exceed 5×(number of predictors), or may have severe bias in $\hat{\beta}_j$ or infinite estimates, need alternative to ML fitting.
- Analogs of R^2 for binary data are only partially successful (*CDA*, Sec. 6.3), but corr$(Y, \hat{P}(Y = 1))$ useful for comparing models.
- Alternative criterion: Minimize AIC $= -2$(maximized log likelihood) $+ 2$(no. model parameters) to find model with best fit to population. BIC has Bayesian interpretation, can lead to simpler model.

Example: Horseshoe crab data ($n = 173$ female crabs)
(*CDA*, Sec. 4.3.2, Sec. 6.1)

$Y =$ whether female horseshoe crab has 'satellites'
(1 = yes for 111 crabs, 0 = no for 62 crabs)

$C =$ color (factor with 4 categories, light to dark)
$S =$ spine condition (factor with 3 categories)
$W =$ width of carapace shell (cm)
$WT =$ weight of crab (kg)

Weight and width are very highly correlated ($r = 0.89$), and we do not consider weight in this analysis.

Backward elimination results in a model with main effects for width and color (as a qualitative factor).

Further simplification results from reducing color to two categories (dark, other), where 'other' combines lighter colors.

Horseshoe crab mating, with satellites

Results of fitting several logistic regression models to horseshoe crab data

Model	Predictors[a]	Deviance G^2	df	AIC	Models Compared	Deviance Difference	Correlation $r(y, \hat{\mu})$
1	C*S + C*W + S*W	173.67	155	209.7	—	—	
2	C + S + W	186.61	166	200.6	(2)–(1)	12.9 (df = 11)	0.456
3a	C + S	208.83	167	220.8	(3a)–(2)	22.2 (df = 1)	0.314
3b	S + W	194.42	169	202.4	(3b)–(2)	7.8 (df = 3)	0.402
3c	C + W	187.46	168	197.5	(3c)–(2)	0.8 (df = 2)	0.452
4a	C	212.06	169	220.1	(4a)–(3c)	24.5 (df = 1)	0.285
4b	W	194.45	171	198.5	(4b)–(3c)	7.0 (df = 3)	0.402
5	(C = dark) + W	187.96	170	194.0	(5)–(3c)	0.5 (df = 2)	0.447
6	None	225.76	172	227.8	(6)–(5)	37.8 (df = 2)	0.000

[a] *C*, color; *S*, spine condition; *W*, width.

C + S + W denotes the model with main effects of color, spine condition, width.

C*S + C*W + S*W denotes the model that has all the two-factor interactions as well as the "main effects."

Note: In model 3c, with indicator (dummy) variables for the first three colors,

$$\text{logit}[\hat{P}(Y = 1)] = -12.715 + 1.33c_1 + 1.40c_2 + 1.11c_3 + 0.468(\text{width}),$$

the color estimates $(1.33, 1.40, 1.11, 0)$ suggest it may be adequate to replace color as a factor by an indicator of whether a crab is dark (color4).

R for logit model with main effects, and correlation summary:

```
> Crabs <- read.table("https://alanagresti.com/cda/CDA_data/Crabs.dat",header=TRUE)
> Crabs
    crab sat y weight width color spine
1      1   8 1  3.050  28.3     2     3
2      2   0 0  1.550  22.5     3     3
3      3   9 1  2.300  26.0     1     1
...
173  173   0 0  2.000  24.5     2     2
>
> fit3c <- glm(y ~ width + factor(color), family=binomial(link=logit), data=Crabs)
> summary(fit3c)
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -11.38519    2.87346  -3.962 7.43e-05
width            0.46796    0.10554   4.434 9.26e-06
factor(color)2   0.07242    0.73989   0.098    0.922
factor(color)3  -0.22380    0.77708  -0.288    0.773
factor(color)4  -1.32992    0.85252  -1.560    0.119
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.46  on 168  degrees of freedom
AIC: 197.46

> cor(y, fitted(fit3c))
[1] 0.4522131
```

Coding color so *last* category does not have indicator var. seems adequate to replace color as a factor by indicator of whether crab is dark (color4). This is model 5.

```
> c1 <- ifelse(color==1,1,0)
> c2 <- ifelse(color==2,1,0)
> c3 <- ifelse(color==3,1,0)
> c4 <- ifelse(color==4,1,0)
> fit <- glm(y ~ width + c1 + c2 + c3, family=binomial(link=logit))
> summary(fit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.7151     2.7617  -4.604 4.14e-06
width         0.4680     0.1055   4.434 9.26e-06
c1            1.3299     0.8525   1.560   0.1188
c2            1.4023     0.5484   2.557   0.0106
c3            1.1061     0.5921   1.868   0.0617 .
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.46  on 168  degrees of freedom
AIC: 197.46

> fit5 <- glm(y ~ width + c4, family=binomial(link=logit))
> summary(fit5)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.6790     2.6925  -4.338 1.44e-05
width         0.4782     0.1041   4.592 4.39e-06
c4           -1.3005     0.5259  -2.473   0.0134
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.96  on 170  degrees of freedom
AIC: 193.96
```

With several explanatory variables, a *purposeful selection* strategy is described in 3rd edition of *An Introduction to Categorical Data Analysis* (based on text *Applied Logistic Regression* by Hosmer and Lemeshow).

1. Construct initial model using explanatory variables that include known important variables and others that show *any* evidence of being relevant when used as sole predictors (e.g., having P-value < 0.2).
2. Conduct backward elimination, keeping a variable if it is either significant at a somewhat more stringent level or shows evidence of being a relevant confounder, in the sense that the estimated effect of a key variable changes substantially when it is removed.
3. Add any variables not included in step (1) but that are significant when adjusting for variables in model after step (2).
4. Check for plausible interactions among variables in the model after step (3), using significance tests at conventional levels such as 0.05.

Analysis of Residuals (*CDA*, Sec. 6.2)

If $Y_i \sim \text{Bin}(n_i, \pi_i)$, with $\mu_i = n_i\pi_i$, the Pearson residual is

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

This is commonly used for binomial models (e.g., logistic regression)

When summed over success and failure counts, these satisfy $X^2 = \sum_i e_i^2$.

However, substituting $\hat{\mu}_i$ for μ_i reduces variance of numerator, considerably if model has many parameters. Better to *standardize* the Pearson residual:

$$r_i = \frac{y_i - \hat{\mu}_i}{SE(y_i - \hat{\mu}_i)}$$

When model holds, these are asymptotically N(0,1).

We used these after applying the chi-squared test of independence to the data on education and belief in God. There are analogous *deviance* and *standardized deviance* residuals.

Example: Berkeley admissions data (*CDA*, p. 63)

	Admitted, male		Admitted, female	
Department	Yes	No	Yes	No
A	512 (62%)	313	89 (82%)	19
B	353 (63%)	207	17 (68%)	8
C	120 (37%)	205	202 (34%)	391
D	138 (33%)	279	131 (35%)	244
E	53 (28%)	138	94 (24%)	299
F	22 (6%)	351	24 (7%)	317
Total	1198 (45%)	1493	557 (30%)	1278

Percents in table are percent admitted by gender and department.

Data nearly satisfy "Simpson's paradox," whereby marginal association has different direction than in each partial table.

A = Admitted (yes, no)
G = Gender (male, female)
D = Department (A, B, C, D, E, F)

$$2 \times 2 \times 6 \text{ table}$$

Use logit model with A as response variable

Model	G^2	df	Interpretation
$G + D$	20.2	5	no interaction
D	21.7	6	no gender effect

Logit model with P(A = yes) dependent on department but independent of gender (given department) fits poorly, as does more complex model also including gender effect but assuming lack of interaction.

(Note: Similar example with different data set in *CDA*, Sec. 6.2.3.)

R for logit model and standardized residuals with graduate admissions data, assuming no effect for gender:

```
> Berkeley <- read.table("berkeley.dat",header=TRUE)
> Berkeley
   dept gender yes  no
1     A   male 512 313
2     A female  89  19
3     B   male 353 207
4     B female  17   8
...
11    F   male  22 351
12    F female  24 317
> attach(Berkeley)
> n <- yes + no

> fit <- glm(yes/n ~ factor(dept), weights=n, family=binomial)
> summary(fit)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.59346    0.06838   8.679   <2e-16
factor(dept)B -0.05059    0.10968  -0.461    0.645
factor(dept)C -1.20915    0.09726 -12.432   <2e-16
factor(dept)D -1.25833    0.10152 -12.395   <2e-16
factor(dept)E -1.68296    0.11733 -14.343   <2e-16
factor(dept)F -3.26911    0.16707 -19.567   <2e-16
---
    Null deviance: 877.056  on 11  degrees of freedom
Residual deviance:  21.736  on  6  degrees of freedom

> rstandard(fit, type="pearson") # stand. Pearson residuals show fewer males admitted than expected
         1          2          3          4          5          6          7
-4.1530728  4.1530728 -0.5037077  0.5037077  0.8680662 -0.8680662 -0.5458732
         8          9         10         11         12
 0.5458732  1.0005342 -1.0005342 -0.6197526  0.6197526
```

Conditional independence model fits well for all departments except first.

Department	G^2 Contribution	df	Sample odds ratio
A	19.05	1	0.35
B	0.26	1	0.80
C	0.75	1	1.13
D	0.30	1	0.92
E	0.99	1	1.22
F	0.38	1	0.83
	21.7	6	

Infinite Estimates in Logistic Regression

(*CDA*, Sec. 6.5)
For ML estimation of logistic model parameters:

- At least one parameter estimate is infinite if can separate with a plane the x values where $y = 1$ and where $y = 0$.

  Complete separation: No observations fall on that plane.

  Quasi-complete separation: On the plane boundary, both outcomes occur (common in contingency tables).

- Most software does not adequately detect this, as convergence of iterative model-fitting occurs at very large estimate, where log-likelihood looks flat.

- Reported *SE* values also useless, because based on curvature at ML estimate (and log-likelihood essentially flat at convergence).

- Wald inference can be very poor whenever a β is very large; as β increases, *SE* of $\hat{\beta}$ grows faster than β (Hauck and Donner 1977).

Example: R for data where ML estimate is actually infinite

```
-----------------------------------------------------------------------
> x <- c(10, 20, 30, 40, 60, 70, 80, 90)
> y <- c(0, 0, 0, 0, 1, 1, 1, 1)  # complete separation

> fit <- glm(y ~ x, family=binomial(link=logit))

> summary(fit)

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -118.158 296046.187       0        1
x               2.363   5805.939       0        1

 Null deviance: 1.1090e+01  on 7  degrees of freedom
 Residual deviance: 2.1827e-10  on 6  degrees of freedom # perfect fit

> library(car)
> Anova(fit)
Analysis of Deviance Table (Type II tests)
Response: y
  LR Chisq Df Pr(>Chisq) # likelihood-ratio test of effect of x
x    11.09  1  0.0008678
-----------------------------------------------------------------------
```

Wald test says absolutely no evidence of effect!!

Yet data suggest strong evidence by other criteria, such as likelihood-ratio test which has test statistic $= 11.09$ ($df = 1$) and P-value $= 0.0009$.

How can this happen with a categorical explanatory variable?

Complete separation:

	Success	Failure
Group 1 ($x = 0$)	10	0
Group 2 ($x = 1$)	0	10

Quasi-complete separation:

	Success	Failure
Group 1 ($x = 0$)	10	10
Group 2 ($x = 1$)	0	10

If model contains interaction of x with other variable, quasi-complete separation occurs if this happens at any particular value of other variable.

- With complete separation, model fit should give perfect predictions for Y. So, warning sign is log-likelihood $= 0$, deviance $= 0$.
- Warning sign of complete or quasi-complete separation: Enormous SE values (because of flat log-likelihood function, at convergence).
- For contingency table data, ML estimates necessarily exist if all counts positive. If at least one count $= 0$, the ML estimates may or may not exist.
- In principle, if estimated odds ratio equals 0 or ∞, Wald test or CI is useless, but one can construct LR test and profile likelihood CI of form [0, u] or [l, ∞].

Example: Endometrial cancer (*CDA*, Sec. 7.2.2, 7.4.8)

y = histology grade (HG: 1 = high, 0 = low)
x_1 = neovasculation (NV: 1 = present, 0 = absent)
x_2 = pulsatility index of arteria uterina (PI: 0 to 49)
x_3 = endometrium height (EH: 0.27 to 3.61).

HG	NV	PI	EH	HG	NV	PI	EH
0	0	13	1.64	0	0	16	2.26
...							
1	1	21	0.98	1	0	5	0.35

Entire data file has $n = 79$ patients, 49 with HG = 0, 30 with HG = 1.
Consider model

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

For all 13 patients having $x_1 = 1$, get $y = 1$.
There is quasi-complete separation. ML estimate $\hat{\beta}_1 = \infty$.

R for logistic regression analysis of endometrial cancer data, with standardized quantitative explanatory variables:

```
> Endometrial <- read.table("https://alanagresti.com/glm/data/Endometrial.dat",header=TRUE)

> Endometrial
   NV PI   EH HG
1   0 13 1.64  0
2   0 16 2.26  0
...
79  0 33 0.85  1

> attach(Endometrial)

> PI2 <- (PI - mean(PI))/sd(PI) # standardize to compare effects of the quantitative
> EH2 <- (EH - mean(EH))/sd(EH) # variables, which have very different std. dev.'s

> table(NV, HG)  #  quasi-complete separation when NV predicts HG

   HG
NV   0  1
  0 49 17
  1  0 13

> fitML <- glm(HG ~ NV + PI2 + EH2, family=binomial) # ordinary ML

> summary(fitML)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.2517     0.3688  -3.394 0.000688
NV            18.1856  1715.7509   0.011 0.991543 # true ML est. is infinity
PI2           -0.4217     0.4432  -0.952 0.341333
EH2           -1.9219     0.5599  -3.433 0.000597 # EH effect stronger than PI
```

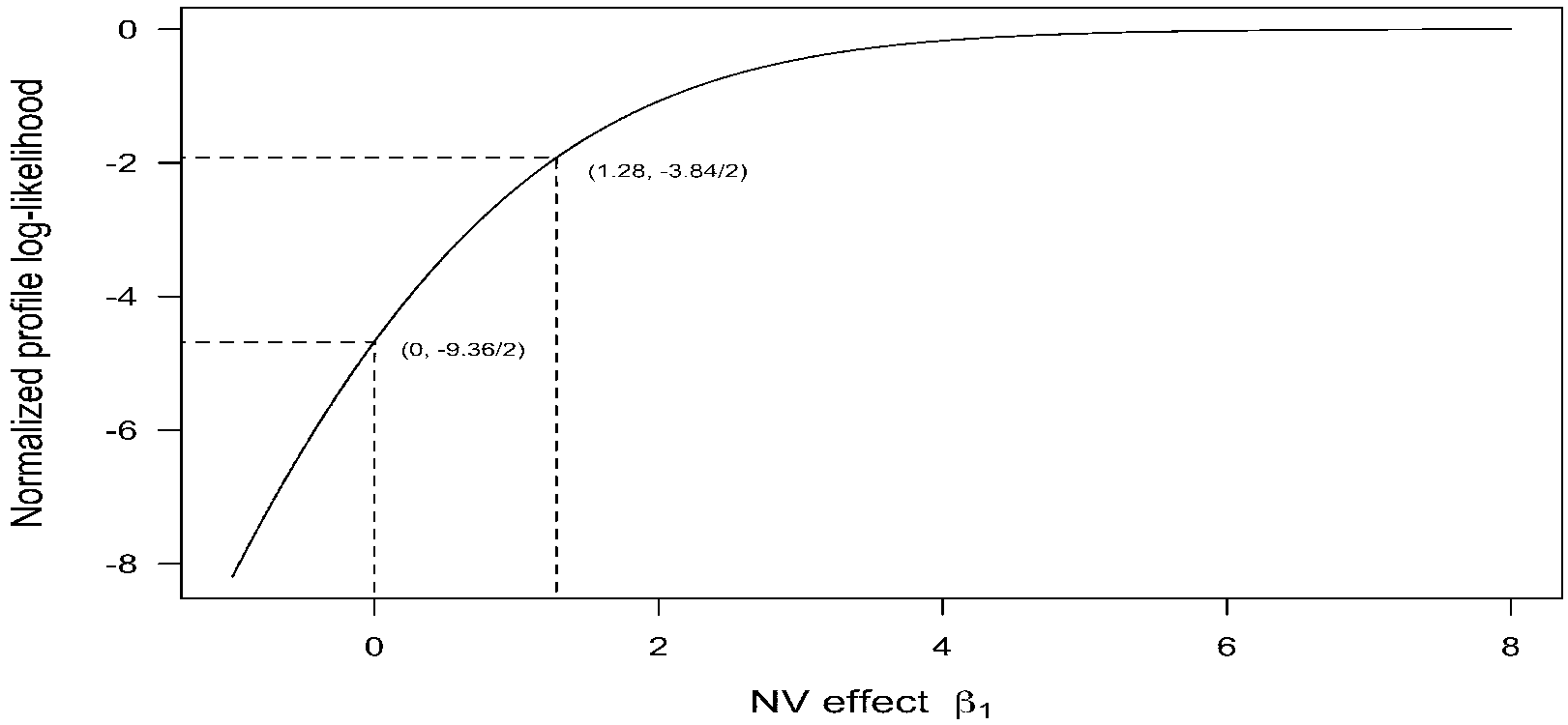
The likelihood function does give information about plausible values for the NV effect:

Likelihood-ratio statistic for testing H_0: $\beta_1 = 0$ is 9.36.

95% profile likelihood CI for NV effect β_1 is (1.28, ∞).

```
------------------------------------------------------------------------------
> logLik(fitML) # not exactly 0 because separation is quasi, not complete
'log Lik.' -27.69663 (df=4)

> library(car)
> Anova(fitML)
    LR Chisq  Df  Pr(>Chisq) # likelihood-ratio tests
NV    9.3576   1     0.00222 # compare to Wald P-value = 0.9915 for NV effect
PI2   0.9851   1     0.32093
EH2  19.7606   1   8.777e-06

> library(profileModel)  # ordinary confint function fails for infinite est.
                         # this library by Ioannis Kosmidis is useful
> confintModel(fitML, objective="ordinaryDeviance", method="zoom",
+              endpoint.tolerance = 1e-08)
         Lower      Upper
NV     1.28411        Inf # 95% profile likelihood CI for beta1
PI2   -1.37047  0.3817637
EH2   -3.16891 -0.9510794

> library(detectseparation) # from Ioannis Kosmidis
> detect_separation(x=cbind(Endo$NV,Endo$PI2,Endo$EH2),y=Endo$HG,family=binomial(link=logit))
Separation: TRUE  # detects complete or quasi-complete separation
Existence of maximum likelihood estimates
 X1  X2  X3
Inf   0   0       # 1st explanatory variable has infinite ML estimate
0: finite value, Inf: infinity, -Inf: -infinity
```

Figure: Normalized profile log-likelihood function $L(\beta_1) - L(\hat{\beta}_1)$ for NV effect in main-effects logistic model. Double the log-likelihood increases by 9.36 (the likelihood-ratio test statistic) between $\beta_1 = 0$ and $\hat{\beta}_1 = \infty$ and by $1.96^2 = 3.84$ (the test statistic value that yields chi-squared P-value $= 0.05$) between $\beta_1 = 1.28$ and $\hat{\beta}_1 = \infty$.



Remedies for infinite estimates?

- *Bayesian approach*: Influence of prior distribution smooths data and results in finite posterior mean estimates.
  (*CDA*, Sec. 7.2; *Foundations of Bayesian Statistics for Data Scientists, with R and Python* by Agresti, Kateri et al. (2026))

- *Penalized likelihood approach*: Add a penalty term to likelihood function, to reduce bias of ML estimates. Maximizing penalized likelihood results in estimates shrunk toward 0.
  (Ref: Firth 1993, *CDA*, Sec. 6.5.3)

  R: Package *logistf* can do the Firth approach.

  The penalized likelihood approach is non-Bayesian, but actually results in estimates similar to modes of Bayesian posterior distribution for a particular type of prior distribution (Jeffreys' prior).

R using Firth penalized likelihood with endometrial cancer data:

```
------------------------------------------------------------------------------

> fitML <- glm(HG ~ NV + PI2 + EH2, family=binomial)

> summary(fitML)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.2517     0.3688  -3.394 0.000688
NV            18.1856  1715.7509   0.011 0.991543
PI2           -0.4217     0.4432  -0.952 0.341333
EH2           -1.9219     0.5599  -3.433 0.000597
---


> install.packages("logistf")

> library("logistf")

> fitpenalized <- logistf(HG ~ NV + PI2 + EH2, family=binomial)

> summary(fitpenalized) # The Firth penalized likelihood approach

Confidence intervals and p-values by Profile Likelihood
                  coef  se(coef) lower 0.95 upper 0.95      Chisq            p
(Intercept) -1.1566145 0.3477055 -1.9166518 -0.5207233 13.8173127 2.014712e-04
NV           2.9292733 1.5507634  0.6097274  7.8546317  6.7984572 9.123668e-03
PI2         -0.3474419 0.3956953 -1.2443165  0.4044667  0.7468285 3.874822e-01
EH2         -1.7243007 0.5138261 -2.8903284 -0.8162243 17.7593175 2.506867e-05

Likelihood ratio test=43.65582 on 3 df, p=1.78586e-09, n=79
------------------------------------------------------------------------------
```

R using *Bayesian inference* with very disperse normal priors ($\mu = 0, \sigma = 10$) for the endometrial cancer data:

```
-------------------------------------------------------------------------------
> PI2 <- (PI-mean(PI))/sd(PI); EH2 <- (EH-mean(EH))/sd(EH); NV2 <- NV-0.5
> library(brms) # excellent package for wide variety of models
> fit.bayes <- brm(HG ~ NV2 + PI2 + EH2, family=bernoulli(link=logit), data=Endo,
+   prior=prior(normal(0,10), class=Intercept) + prior(normal(0, 10), class=b))
> summary(fit.bayes) # Bernoulli is binomial distribution for n=1 trial
Regression coefficients:
          Estimate Est.Error l-95% CI u-95% CI
Intercept     3.54      2.81    -0.30    10.29
NV2           9.77      5.60     2.20    23.21 # conclude 2.2 < beta1 < 23.2
PI2          -0.47      0.45    -1.41     0.37
EH2          -2.13      0.59    -3.38    -1.08
-------------------------------------------------------------------------------
```

Results of maximum likelihood (ML) fitting and Bayesian fitting, with normal prior distributions, of logistic regression model.

Analysis	$\hat{\beta}_1$ (*SD*)	Interval[a] for β_1	$\hat{\beta}_2$ (*SD*)	$\hat{\beta}_3$ (*SD*)	$P(\beta_1 \leq 0)$[b]
ML	∞ (—)	(1.28, ∞)	-0.42 (0.44)	-1.92 (0.56)	0.0011
Bayes, $\sigma = 100$	80.7 (59.0)	(5.9, 222.3)	-0.49 (0.46)	-2.13 (0.60)	< 0.0001
Bayes, $\sigma = 10$	9.8 (5.6)	(2.2, 23.2)	-0.47 (0.45)	-2.13 (0.59)	0.0001
Bayes, $\sigma = 1$	1.6 (0.7)	(0.3, 3.1)	-0.22 (0.33)	-1.75 (0.43)	0.009

[a] Profile likelihood confidence interval and Bayes equal-tail posterior interval

[b] In ML row, this is classical one-sided *P*-value for likelihood-ratio test of H_1: $\beta_1 > 0$

Note: MCMC fitting requires huge number of iterations for convergence when σ large.

Lasso: Handling large numbers of predictors

The lasso maximizes

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) - \lambda \sum_j |\beta_j|,$$

for a smoothing parameter λ chosen using cross-validation to minimize average prediction error for y using remaining data.

- Equivalently, it maximizes likelihood subject to constraint that $\sum_j |\beta_j| \leq \lambda^*$ for some constant λ^*.
  (Larger λ corresponds to smaller λ^*)
- Get ordinary ML when $\lambda = 0$. As λ increases, lasso fitting forces estimates to decrease to 0.
- At each possible value for λ, sample mean prediction error (e.g., using cross-validation) is a random variable. To avoid overfitting, common to pick λ with *one-standard-error rule*, in which chosen λ gives mean prediction error one standard error above minimum, in direction of greater regularization.

Example: Predicting Opinion on Abortion (*Intro CDA*, Sec. 11.5.3)

Data from student survey ($n = 60$) at U. of Florida, with y = support legal abortion within 3 months (1 = yes, 0 = no), 14 explanatory variables

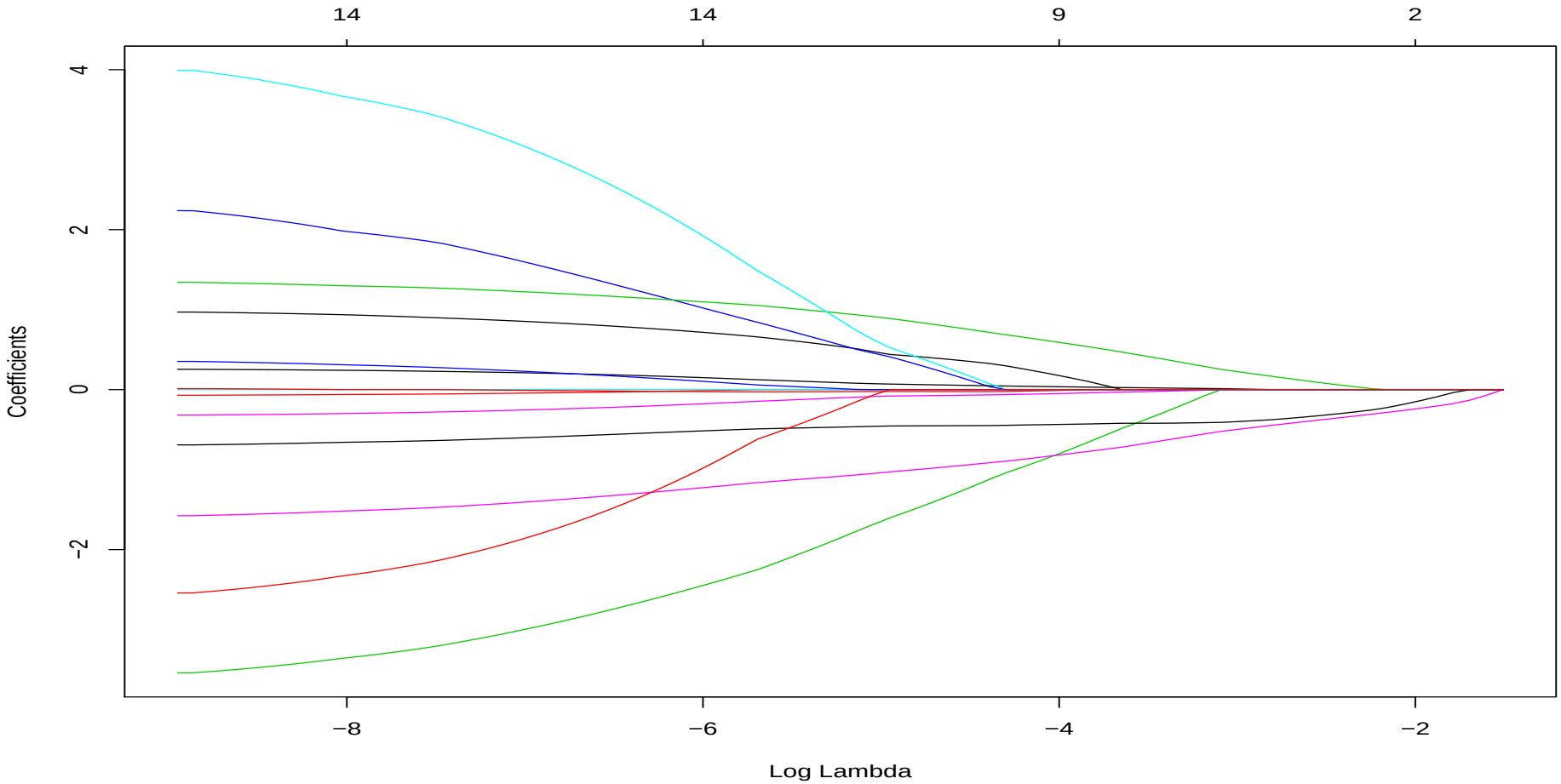
```
> Students <- read.table("https://alanagresti.com/cda/CDA_data/Students.dat",
+                        header=TRUE)
> fit <- glm(abor ~ gender + age + hsgpa + cogpa + dhome + dres + tv + sport +
+  news + aids + veg + ideol + relig + affirm, family=binomial, data=Students)
> summary(fit)
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  10.10142    10.89140    0.927    0.3537
gender        1.00216     1.86553    0.537    0.5911
age          -0.07834     0.12748   -0.615    0.5389
hsgpa        -3.73445     2.80932   -1.329    0.1837
cogpa         2.51127     3.73991    0.671    0.5019
dhome         0.00056     0.00068    0.821    0.4116
dres         -0.33882     0.29538   -1.147    0.2514
tv            0.26598     0.25316    1.051    0.2934
sport         0.02721     0.25515    0.107    0.9151
news          1.38688     0.69868    1.985    0.0471 # like. ratio test P-value = 0.0003
aids          0.39668     0.56637    0.700    0.4837
veg           4.32135     3.86146    1.119    0.2631
ideol        -1.63779     0.78925   -2.075    0.0380 # like. ratio  test P-value = 0.0010
relig        -0.72457     0.78207   -0.926    0.3542
affirm       -2.74815     2.68988   -1.022    0.3069
---
    Null deviance: 62.719  on 59  degrees of freedom
Residual deviance: 21.368  on 45  degrees of freedom

> 1 - pchisq(62.719 - 21.368, 59 - 45) # like. ratio test that all 14 betas = 0
[1] 0.0001566051
```

```
------------------------------------------------------------------------------
> attach(Students)
> x <- cbind(gender, age, hsgpa, cogpa, dhome, dres, tv, sport, news, aids,
>            veg, ideol, relig, affirm) # explanatory var's for lasso
> library(glmnet)
> fit.lasso <- glmnet(x, abor, alpha=1, family="binomial") # alpha=1 is lasso
> plot(fit.lasso, "lambda")
> set.seed(1) # a random seed to implement cross-validation
> cv.glmnet(x, abor, alpha=1, family="binomial", type.measure="class")
  $lambda.min # best lambda by 10-fold cross-validation
  [1] 0.06610251 # this is a random variable, and changes from run to run
  $lambda.1se # lambda suggested by one-standard-error rule, also a r.v.
  [1] 0.1267787
> coef(glmnet(x, abor, alpha=1, family="binomial", lambda=0.1267787))
                    s0
(Intercept)  2.36711 # all 12 lasso estimates that are not shown equal 0
ideol       -0.25994
relig       -0.18311
------------------------------------------------------------------------------
```

At $\lambda = 0.1267787$ (suggested by one-standard-error rule), only explanatory variables are *ideol* (political ideology; 1 = very liberal to 7 = very conservative) and *relig* (how often attend religious services; 0 = never, 1 = occasionally, 2 = most weeks, 3 = every week).

If we use $\lambda = 0.066$ (which minimizes mean prediction error), *news* also enters model.

Figure: Plot of lasso model parameter estimates for predicting opinion on legalized abortion using student survey data, as function of log of smoothing parameter λ.



Summary

- Selection of explanatory variables can use algorithm such as backward elimination (cautiously), or such an algorithm together with a purposeful selection process.
- To compare nested models, can use difference of deviances, which is the likelihood-ratio test.
- Likelihood-ratio methods preferable to Wald methods when n small or parameters near boundary, such as when $P(Y = 1)$ is near 0 or 1.
- With complete or quasi-complete separation in space of explanatory variable values, infinite estimates occur.
- Remedies with infinite estimates include penalized likelihood and a Bayesian approach.
- When number of explanatory variables is very large, lasso shrinks ML estimates, many of them to 0.

LOGLINEAR MODELS FOR CONTINGENCY TABLES

- Association structure in multi-way contingency tables for several categorical response variables
- Types of independence/dependence for categorical variables
- Loglinear model formulas for various patterns of association
- Goodness of fit
- Correspondence with logit models for categorical explanatory variables

Illustrate with loglinear models for 3-way tables:
(*CDA*, Sec. 9.2)

$I \times J \times K$ table for categorical response variables X, Y, Z

Multinomial cell probabilities $\{\pi_{ijk}\}$ in table

Expected frequencies $\{\mu_{ijk} = n\pi_{ijk}\}$ for the cells

Cell counts $\{n_{ijk}\}$

ML fitting treats counts as independent Poisson variates
(GLM with Poisson random component, log link).

Conditional on n, Poisson $\rightarrow$ multinomial, get same ML parameter estimates for each.

Types of Independence / Dependence

(a) Mutual independence

$$P(X = i, Y = j, Z = k) = P(X = i)P(Y = j)P(Z = k)$$

for all i, j, k. $(i.e., \pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k})$

- Corresponds to loglinear model

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

- Denote model by (X, Y, Z)
- Almost always too simple to be of use in practice.

Example: Drug use in survey of high school seniors
(*CDA*, Sec. 9.2.4)

Alcohol use (A)
Cigarette use (C)
Marijuana use (M)

Alcohol Use	Cigarette Use	Marijuana Use Yes	Marijuana Use No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Source: Prof. Harry Khamis, Wright State Univ.

Loglinear models describe the association structure, treating all three variables as response variables.

R for loglinear modeling of high school survey data: Mutual independence model

```
> Drugs <- read.table("https://alanagresti.com/cda/CDA_data/Substance_use.dat", header=TRUE)
> Drugs
  alcohol cigarettes marijuana count # data file has 8 rows, for 8 cell counts
1 yes yes yes   911
2 yes yes  no   538
3 yes  no yes    44
4 yes  no  no   456
5  no yes yes     3
6  no yes  no    43
7  no  no yes     2
8  no  no  no   279

> A <- Drugs$alcohol; C <- Drugs$cigarettes; M <- Drugs$marijuana

> fit <- glm(count ~ A + C + M, family=poisson, data=Drugs)
> summary(fit)  #  loglinear model (A, C, M)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.17254    0.06496  64.234  < 2e-16
Ayes         1.78511    0.05976  29.872  < 2e-16
Cyes         0.64931    0.04415  14.707  < 2e-16
Myes        -0.31542    0.04244  -7.431 1.08e-13
---
    Null deviance: 2851.5  on 7  degrees of freedom
Residual deviance: 1286.0  on 4  degrees of freedom # tests goodness-of-fit of model
AIC: 1343.1
Number of Fisher Scoring iterations: 6
```

(b) Y **jointly independent** of X and Z

$$P(X=i, Y=j, Z=k) = P(X=i, Z=k)P(Y=j)$$

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$$

- X and Z may be associated
- Denote by (XZ, Y)
- Corresponds to ordinary independence in two-way $IK \times J$ table cross classifying Y with all combinations of values of X and Z.
- $(X, Y, Z) \Rightarrow (XZ, Y), (YZ, X), (XY, Z)$
  e.g, (X, Y, Z) is special case of (XZ, Y) with $\lambda_{ik}^{XZ} = 0$.
- Usually too simplistic to be very useful.

(c) X and Y **conditionally independent**, given Z

$$P(X = i, Y = j \mid Z = k) =$$

$$P(X = i \mid Z = k)P(Y = j \mid Z = k)$$

$$P(X = i, Y = j, Z = k) = \frac{P(X{=}i, Z{=}k)P(Y{=}j, Z{=}k)}{[P(Z{=}k)]^2}$$

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

- Allows conditional association between X and Z and between Y and Z
- Denote by (XZ, YZ)
- $(X, Y, Z) \Rightarrow (XZ, Y) \Rightarrow (XZ, YZ)$
- XY conditional independence $\not\Rightarrow$ XY marginal independence.

(d) **Homogeneous associations for each pair (but no three-factor interaction)**

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

- Denote by (XY, XZ, YZ)
- Each pair of variables may be associated, conditionally and marginally.
- Implies odds ratio between two variables identical at each level of third variable (*homogeneous association*).
- Independence models are special cases $\{\lambda_{ij}^{XY} = 0\}$ is XY conditional independence.

(e) **Three-factor interaction:** Adds λ_{ijk}^{XYZ}

Goodness of Fit for Loglinear Models

With the fitted values, one can test fit of model by comparing $\{\hat{\mu}_{ijk}\}$ to $\{n_{ijk}\}$ with chi-squared statistics X^2 or G^2.

$df =$ number of Poisson counts $-$ number of parameters

Goodness of fit of some models for student survey data:

Model	G^2	df
(A, C, M)	1286.0	4
(AC, AM)	497.4	2
(AC, CM)	92.0	2
(AM, CM)	187.8	2
(AC, AM, CM)	0.37	1

Loglinear model two-factor parameters relate to odds ratios.
e.g., when (XY, XZ, YZ) or simpler model holds for $2 \times 2 \times k$ table, there is *homogeneous association*.

$$\begin{aligned} \theta_{XY|Z=1} &= \theta_{XY|Z=2} = \cdots = \theta_{XY|Z=k} \\ \hat{\theta}_{XY|Z} &= \exp(\hat{\lambda}^{XY}) \end{aligned}$$

Software output shows estimated conditional odds ratios are:
$e^{2.05} = 7.8$ for AC association.
$e^{2.99} = 19.8$ for AM association.
$e^{2.85} = 17.3$ for CM association.
(95% Wald CI is $\exp[2.85 \pm 1.96(0.164)] = (12.5, 23.8)$.)

R continued for loglinear modeling of high school survey data:

```
> homo.assoc <- glm(count ~ A + C + M + A:C + A:M + C:M, family=poisson, data=Drugs)
> summary(homo.assoc) # homogeneous association model
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  5.6334      0.0597    94.36   < 2e-16
Ayes         0.4877      0.0758     6.44  1.22e-10
Cyes        -1.8867      0.1627   -11.60   < 2e-16
Myes        -5.3090      0.4752   -11.17   < 2e-16
Ayes:Cyes    2.0545      0.1741    11.80   < 2e-16 # AC log odds ratio = 2.0545
Ayes:Myes    2.9860      0.4647     6.43  1.31e-10
Cyes:Myes    2.8479      0.1638    17.38   < 2e-16
---
Residual deviance:    0.37399  on 1  degrees of freedom # model fits well

> pearson <- resid(homo.assoc, type="pearson") #  Pearson residuals
> sum(pearson^2)  # Pearson goodness-of-fit statistic
[1] 0.4011006
> std.resid <- rstandard(homo.assoc, type="pearson") # standardized
> expected <- fitted(homo.assoc) # estimated expected frequencies
> cbind(count, expected, pearson, std.resid)
  count   expected     pearson  std.resid
1   911 910.383170  0.02044342  0.6333249
2   538 538.616830 -0.02657821 -0.6333249
3    44  44.616830 -0.09234564 -0.6333249
4   456 455.383170  0.02890528  0.6333249
5     3   3.616830 -0.32434090 -0.6333251
6    43  42.383170  0.09474777  0.6333249
7     2   1.383170  0.52447895  0.6333251
8   279 279.616830 -0.03688791 -0.6333249
```

Pearson residuals take 8 different values, even though df = 1. Standardized residuals reflect same difference between observed and fitted in each cell.

Correspondence with logit models

Model (AC, AM, CM) for drug data

$$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{AC} + \lambda_{ik}^{AM} + \lambda_{jk}^{CM}$$

Treat M as response, A and C as explanatory (i.e., 4 binomials).

$$\log\left(\frac{P(M=1)}{P(M=2)}\right) = \log\left(\frac{\mu_{ij1}}{\mu_{ij2}}\right) = \log \mu_{ij1} - \log \mu_{ij2}$$

$$= [\lambda+\lambda_i^A+\lambda_j^C+\lambda_1^M+\lambda_{ij}^{AC}+\lambda_{i1}^{AM}+\lambda_{j1}^{CM}]-[\lambda+\lambda_i^A+\lambda_j^C+\lambda_2^M+\lambda_{ij}^{AC}+\lambda_{i2}^{AM}+\lambda_{j2}^{CM}]$$

$$= (\lambda_1^M - \lambda_2^M) + (\lambda_{i1}^{AM} - \lambda_{i2}^{AM}) + (\lambda_{j1}^{CM} - \lambda_{j2}^{CM})$$

$$= \alpha + \beta_i^A + \beta_j^C.$$

Residual $df = 8 - 7$ (no. Poisson counts − no. loglinear para.)
$= 4 - 3$ (no. binomial logits − no. logit parameters) $= 1$.

i.e., we get same results for the association between M and each of A and C if we treat the data as four binomials (instead of eight Poissons) and model the logit for marijuana use in terms of additive effects for alcohol use and cigarette use.

Illustration using R for the corresponding logistic model:

```
------------------------------------------------------------------
> Drugs2 <- read.table("Drugs_binomial.dat", header=TRUE)
> Drugs2
    A    C  M_yes  M_no     n
1 yes  yes    911   538  1449
2 yes   no     44   456   500
3  no  yes      3    43    46
4  no   no      2   279   281

> fit.logistic <- glm(M_yes/n ~ A + C, weights=n,
+                     family=binomial(link=logit), data=Drugs2)

> summary(fit.logistic)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.3090     0.4752 -11.172  < 2e-16
Ayes          2.9860     0.4647   6.426 1.31e-10
Cyes          2.8479     0.1638  17.382  < 2e-16
---
    Null deviance: 843.82664  on 3  degrees of freedom
Residual deviance:   0.37399  on 1  degrees of freedom
------------------------------------------------------------------
```

Summary

- Hierarchy of loglinear models describes association patterns.
- Conditional independence vs. marginal independence
- X^2 and G^2 goodness of fit for non-sparse tables
- Interpret conditional associations using odds ratios.
- Logistic models with categorical predictors correspond to loglinear models with general association structure among explanatory variables
- Iterative methods required for ML fitting is simple, because special case of GLM for Poisson random component with log link.
- Not discussed: Loglinear models that account for ordered categories of any variable (*CDA*, Sec. 10.4)

LOGISTIC MODELS FOR MULTICATEGORY RESPONSES

- *Nominal response models*
  e.g., model choice of how get to work (walk, drive, bus, subway), product brand (A, B, C. ...), where choose to shop (central city, suburban mall, online).
  Standard modeling: Apply logit to all pairs of categories.
- *Ordinal response models*
  e.g., patient quality of life (excellent, good, fair, poor),
  political philosophy (very liberal, slightly liberal, moderate, slightly conservative, very conservative).
  Standard modeling: Apply logits to cumulative probabilities.

In both cases, focus is on modeling how

$$\pi_j = P(Y = j), \quad j = 1, 2, \ldots, c,$$

depends on explanatory variables (categorical and/or quantitative).

The models treat observations on Y at fixed x as *multinomial*.

Nominal response: Baseline-category logits

Baseline category logits for c categories are (*CDA*, Sec. 8.1)

$$\log\left(\frac{\pi_1}{\pi_c}\right), \ \log\left(\frac{\pi_2}{\pi_c}\right), \ \ldots, \ \log\left(\frac{\pi_{c-1}}{\pi_c}\right).$$

Model with p explanatory variables is

$$\log\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p, \quad j = 1, \ldots, c-1.$$

These $c-1$ equations determine parameters for logits with other pairs of response categories, since

$$\log\left(\frac{\pi_a}{\pi_b}\right) = \log\left(\frac{\pi_a}{\pi_c}\right) - \log\left(\frac{\pi_b}{\pi_c}\right) = (\alpha_a - \alpha_b) + \sum_{k=1}^{p}(\beta_{ak} - \beta_{bk})x_k.$$

Corresponding model for response probabilities $\{\pi_j\}$ is

$$\pi_j = \frac{\exp(\alpha_j + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p)}{1 + \sum_{h=1}^{c-1}\exp(\alpha_h + \beta_{h1}x_1 + \cdots + \beta_{hp}x_p)}. \text{ with } \ \beta_{ck} = 0 \text{ for all } k.$$

Example: Alligator food choice in stomachs of captured alligators (*CDA*, Sec. 8.1.2)

Primary Food Choice (in volume) of Alligators, by Lake and Size of the Alligator

	Size	Primary Food Choice				
Lake	(meters)	Fish	Invertebrate	Reptile	Bird	Other
Hancock	≤ 2.3	23	4	2	2	8
	> 2.3	7	0	1	3	5
Oklawaha	≤ 2.3	5	11	1	0	3
	> 2.3	13	8	6	1	0
Trafford	≤ 2.3	5	11	2	1	5
	> 2.3	8	7	6	3	5
George	≤ 2.3	16	19	1	2	3
	> 2.3	17	1	0	1	3

We'll model primary food choice using *Fish* as the baseline category, using the *vglm* (vector generalized linear models) function in the *VGAM* package of Thomas Yee.

See also Yee's book *Vector Generalized Linear and Additive Models* (Springer, 2015).

R for baseline-category logit model with alligator data:

```
> Alligators <- read.table("https://alanagresti.com/cda/CDA_data/Alligators.dat", header=TRUE)
> Alligators
  lake size y1 y2 y3 y4 y5 # food choices are y1, y2, y3, y4, y5
1    1    1 23  4  2  2  8 # size = 1 for small alligators, 0 for large
2    1    0  7  0  1  3  5
3    2    1  5 11  1  0  3
4    2    0 13  8  6  1  0
5    3    1  5 11  2  1  5
6    3    0  8  7  6  3  5
7    4    1 16 19  1  2  3
8    4    0 17  1  0  1  3

> library(VGAM)
> fit <- vglm(formula = cbind(y2,y3,y4,y5,y1) ~ size + factor(lake),
+   family=multinomial, data=Alligators) # fish=1 is baseline category
> fitted(fit)
          y2         y3          y4         y5        y1
1 0.09309880 0.04745657 0.070401523 0.25373963 0.5353035
2 0.02307168 0.07182461 0.140896287 0.19400964 0.5701978
3 0.60189675 0.07722761 0.008817482 0.05387208 0.2581861
4 0.24864518 0.19483742 0.029416085 0.06866281 0.4584385
5 0.51683852 0.08876722 0.035894709 0.17420051 0.1842990
6 0.19296122 0.20239954 0.108225068 0.20066164 0.2957525
7 0.41285579 0.01156654 0.029671169 0.09380245 0.4521040
8 0.13967784 0.02389871 0.081067366 0.09791362 0.6574425

> anova(fit) # vglm function for likelihood-ratio tests of conditional effects
Analysis of Deviance Table (Type II tests)
             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
size          4   21.087        16     38.167 0.0003043
factor(lake) 12   49.133        24     66.213 1.983e-06
```

R continued for baseline-category logit model:

```
-----------------------------------------------------------------------
> summary(fit)
                  Estimate Std. Error  z value
(Intercept):1    -3.20738    0.63873 -5.02147
(Intercept):2    -2.07176    0.70672 -2.93150
(Intercept):3    -1.39796    0.60852 -2.29731
(Intercept):4    -1.07808    0.47091 -2.28932
size:1            1.45820    0.39595  3.68285
size:2           -0.35126    0.58003 -0.60559
size:3           -0.63066    0.64248 -0.98160
size:4            0.33155    0.44825  0.73966
factor(lake)2:1   2.59558    0.65971  3.93442
factor(lake)2:2   1.21610    0.78602  1.54716
factor(lake)2:3  -1.34833    1.16353 -1.15882
factor(lake)2:4  -0.82054    0.72956 -1.12471
factor(lake)3:1   2.78034    0.67122  4.14220
factor(lake)3:2   1.69248    0.78045  2.16860
factor(lake)3:3   0.39265    0.78177  0.50226
factor(lake)3:4   0.69017    0.55967  1.23317
factor(lake)4:1   1.65836    0.61288  2.70586
factor(lake)4:2  -1.24278    1.18543 -1.04837
factor(lake)4:3  -0.69512    0.78126 -0.88974
factor(lake)4:4  -0.82620    0.55754 -1.48186

Residual deviance: 17.07983 on 12 degrees of freedom
Log-likelihood: -47.5138 on 12 degrees of freedom

> 1 - pchisq(17.07983, df=12) # P-value for goodness-of-fit test
[1] 0.146619
-----------------------------------------------------------------------
```

Size of alligator has a noticeable effect in first equation using size.
i.e., prediction equation for log odds of selecting *invertebrates* (e.g., snails, crayfish) instead of *fish* is

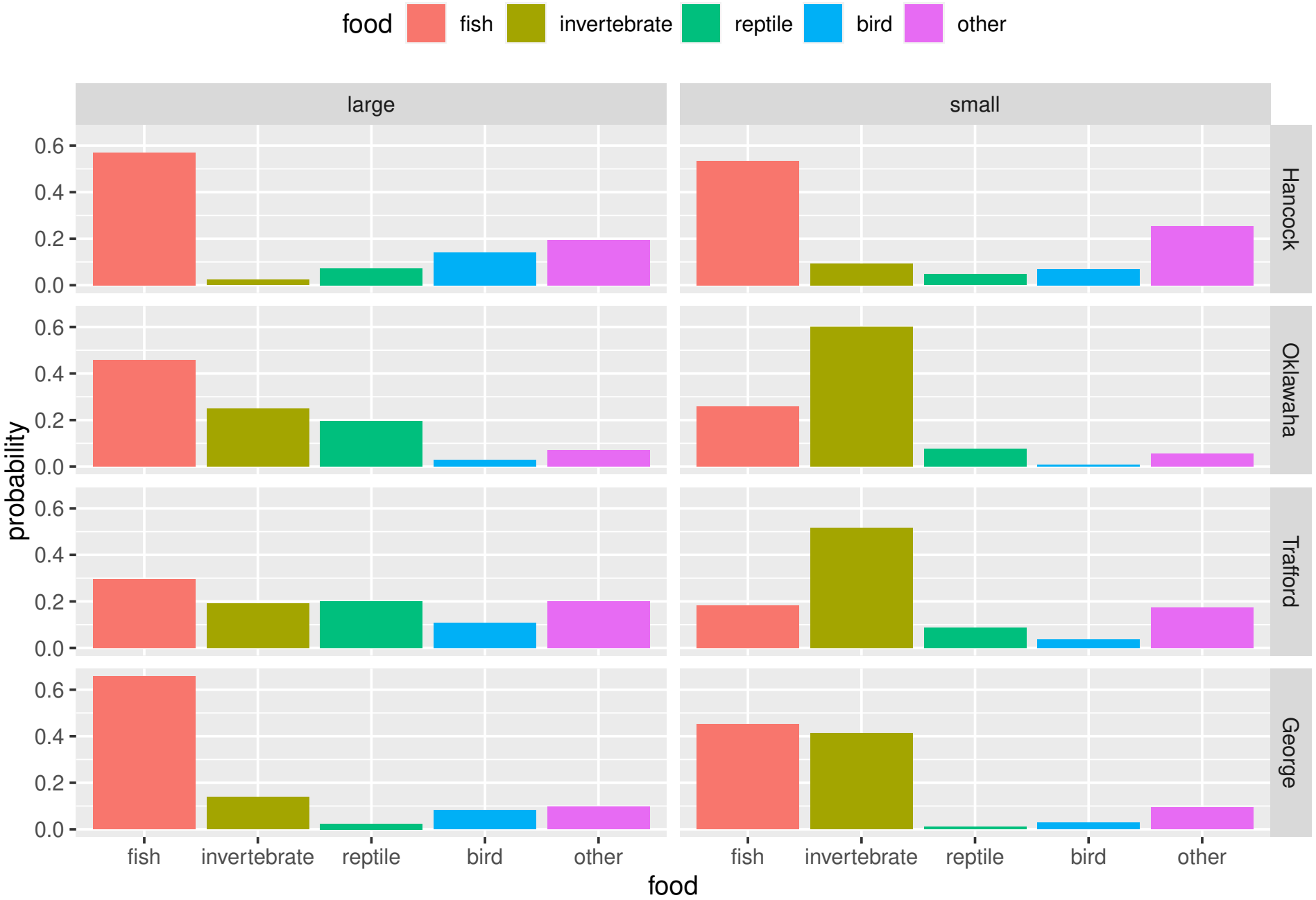
$$\log(\hat{\pi}_I/\hat{\pi}_F) = -3.207 + 1.458s + 2.596z_O + 2.780z_T + 1.658z_G.$$

For a given lake, for small alligators ($s = 1$), estimated odds that primary food choice was invertebrates instead of fish are $\exp(1.458) = 4.30$ times estimated odds for large alligators ($s = 0$).

Estimated effect is imprecise, as Wald 95% CI is
$\exp[1.458 \pm 1.96(0.396)] = (1.98, 9.34)$.

Lake effects indicate that estimated odds that primary food choice was invertebrates instead of fish are relatively higher at Lakes Oklawaha, Trafford and George than at Lake Hancock (baseline category for lakes).

Viewing effect estimates in all equations, size has greatest impact in terms of whether invertebrates rather than fish are primary food choice.


food
fish
invertebrate
reptile
bird
other
large
small
Hancock
Oklawaha
Trafford
George
probability
0.0
0.2
0.4
0.6
fish
invertebrate
reptile
bird
other
food

Ordinal response: Cumulative logits

(*CDA*, Sec. 8.2)

Y an ordinal response (c categories)
x an explanatory variable

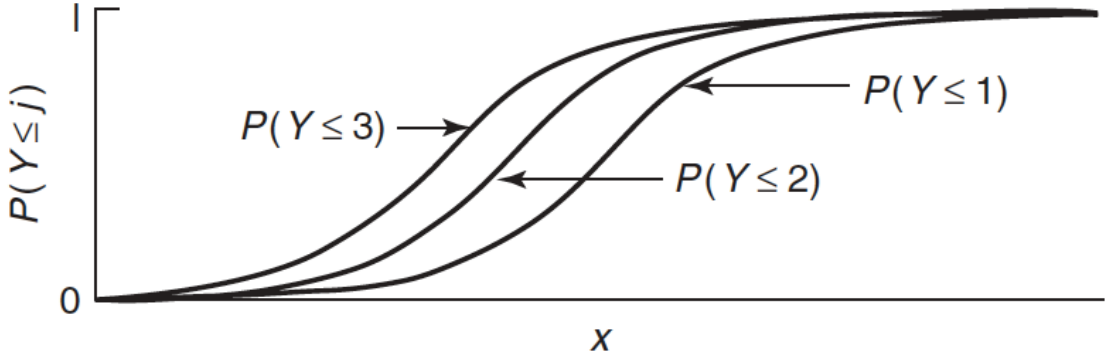
Model $P(Y \leq j)$, $j = 1, 2, \cdots, c-1$, using logits

$$\begin{aligned} \text{logit}[P(Y \leq j)] &= \log[P(Y \leq j)/P(Y > j)] \\ &= \alpha_j + \beta x, \quad j = 1, \ldots, c-1. \end{aligned}$$

This is called a *cumulative logit* model.

As in ordinary logistic regression, effects described by odds ratios. Here, we compare odds of being below vs. above any point on the response scale (*cumulative odds ratios*).

For fixed j, looks like ordinary logistic regression for binary response (below j, above j). See figure on next page for $c = 4$ categories.



At values x_1 and x_2 of the explanatory variable x, model satisfies

$$\log\left[\frac{P(Y \le j \mid x_1)/P(Y > j \mid x_1)}{P(Y \le j \mid x_2)/P(Y > j \mid x_2)}\right] = [\alpha_j + \beta x_1] - [\alpha_j + \beta x_2] = \beta(x_1 - x_2)$$

for all j (called *proportional odds* property).

- β = *cumulative log odds ratio* for 1-unit increase in predictor.
- Model assumes effect β is identical for every "cutpoint" for cumulative probability, $j = 1, \cdots, c-1$.
- Logistic regression is special case $c = 2$.
- Software for ML fitting in R includes *vglm* in VGAM library and *polr* (proportional odds logistic regression) in MASS library).

Properties of cumulative logit models

- Model extends to multiple explanatory variables; for subject i,

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p$$

that can be qualitative or quantitative
(use indicator variables for qualitative explanatory var's).

- Estimated conditional distribution function is

$$\hat{P}(Y \leq j) = \frac{\exp(\hat{\alpha}_j + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)}{1 + \exp(\hat{\alpha}_j + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)}.$$

Estimated probability of outcome j is

$$\hat{P}(Y = j) = \hat{P}(Y \leq j) - \hat{P}(Y \leq j - 1).$$

- Can motivate proportional odds structure by a regression model for underlying continuous *latent variable*.

$$\begin{aligned} Y &= \text{observed ordinal response} \\ Y^* &= \text{underlying continuous latent variable,} \end{aligned}$$

$Y^* = \sum_k \beta_k x_k + \epsilon = \boldsymbol{\beta}^T \mathbf{x} + \epsilon$ where ϵ has cdf G with mean 0. Thresholds (cutpoints) $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_c = \infty$ such that

$$Y = j \text{ if } \alpha_{j-1} < Y^* \leq \alpha_j$$

Then, at fixed $\mathbf{x}$ (see figure on next page)

$$P(Y \leq j) = P(Y^* \leq \alpha_j) = P(Y^* - \boldsymbol{\beta}^T \mathbf{x} \leq \alpha_j - \boldsymbol{\beta}^T \mathbf{x})$$

$$= P(\epsilon \leq \alpha_j - \boldsymbol{\beta}^T \mathbf{x}) = G(\alpha_j - \boldsymbol{\beta}^T \mathbf{x})$$

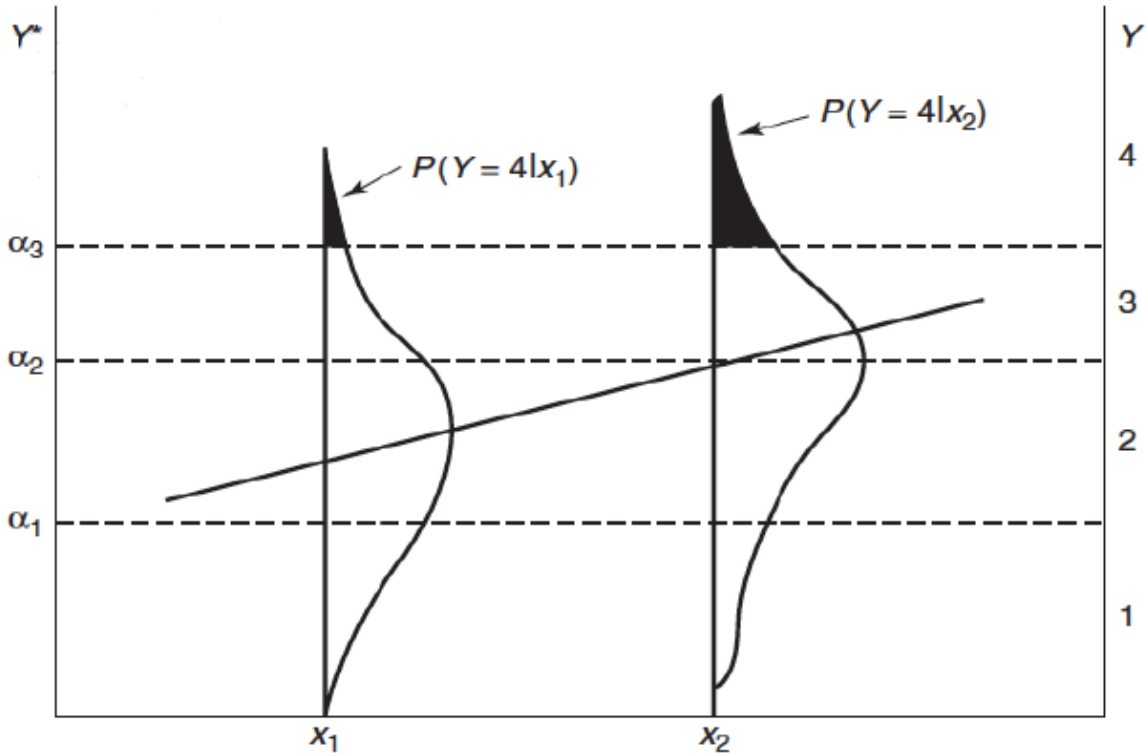
$\rightarrow$ Model $\quad G^{-1}[P(Y \leq j \mid \mathbf{x})] = \alpha_j - \boldsymbol{\beta}^T \mathbf{x}$

with G^{-1} a *link function*.
Get cumulative logit model when $G =$ logistic cdf $\quad (G^{-1} = \text{logit})$.

Note: Model often expressed (e.g., in Stata, SPSS, *polr* in R) as

$$\text{logit}[P(Y \leq j)] = \alpha_j - \boldsymbol{\beta}^T \mathbf{x}.$$

Same fit, estimates, as using $\alpha_j + \boldsymbol{\beta}^T \mathbf{x}$, except sign of $\hat{\boldsymbol{\beta}}$.



Left vertical axis: values for latent variable Y^*

Right vertical axis: values for observed ordinal response Y

Curves: conditional distribution of Y^* at values x_1 and x_2 of x

Line connecting means: regression model for Y^*

Note: This derivation suggests such models detect *location* effects (i.e., shifts in center), not dispersion effects (spread).

Example: Effect of intravenous medication doses on patients with subarachnoid hemorrhage trauma

Treatment Group (x)	Glasgow Outcome Scale (Y)				
	Death	Vegetative State	Major Disability	Minor Disability	Good Recovery
Placebo	59 (28%)	25	46	48	32 (15%)
Low dose	48 (25%)	21	44	47	30 (16%)
Medium dose	44 (21%)	14	54	64	31 (15%)
High dose	43 (22%)	4	49	58	41 (21%)

Some indication that chance of death decreases as dose increases.

Model with linear effect of dose on cumulative logits for outcome (assigning scores $x = 1, 2, 3, 4$ to ordinal x),

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta x$$

has ML estimate $\hat{\beta} = -0.176$ $(SE = 0.056)$.

Likelihood-ratio test of H_0 $\beta = 0$ has test statistic $= 9.61$ $(df = 1,\ P = 0.002)$.

R for modeling dose-response data, using `vglm` in VGAM library

```
> Trauma <- read.table("https://alanagresti.com/glm/data/Trauma.dat", header=TRUE)
> Trauma
  dose y1 y2 y3 y4 y5
1    1 59 25 46 48 32
2    2 48 21 44 47 30
3    3 44 14 54 64 31
4    4 43  4 49 58 41
> library(VGAM)
> fit <- vglm(cbind(y1,y2,y3,y4,y5) ~ dose, family=cumulative(parallel=TRUE), data=Trauma)
> summary(fit)  # use "parallel=TRUE" to impose proportional odds structure
                 Value Std. Error z value
(Intercept):1 -0.71917    0.15881 -4.5285
(Intercept):2 -0.31860    0.15642 -2.0368
(Intercept):3  0.69165    0.15793  4.3796
(Intercept):4  2.05701    0.17369 11.8429
dose          -0.17549    0.05632 -3.1159
Residual Deviance: 18.18245 on 11 degrees of freedom
Log-likelihood: -48.87282 on 11 degrees of freedom
Exponentiated coefficients:     dose
                            0.8390501
> fitted(fit)  #  estimated multinomial response probabilities
         y1         y2        y3        y4        y5
1 0.2901506 0.08878053 0.2473198 0.2415349 0.1322142
2 0.2553767 0.08321565 0.2457635 0.2619656 0.1536786
3 0.2234585 0.07701184 0.2407347 0.2808818 0.1779132
4 0.1944876 0.07043366 0.2325060 0.2975291 0.2050436
> anova(fit) # vglm function to give LR test of effect
Analysis of Deviance Table (Type II tests)
     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
dose  1   9.6124        12     27.795 0.001933
```

Note: propodds() is another possible family for vglm; it defaults to cumulative(reverse = TRUE, link = "logit", parallel = TRUE)

R for modeling dose-response data using `polr` in MASS library, for which response must be an ordered factor:

```
-----------------------------------------------------------------
> Trauma2 <- read.table("trauma2.dat", header=TRUE)
> Trauma2
   dose response count
1     1        1    59
2     1        2    25
...
20    4        5    41
> y <- factor(Trauma2$response)
> fit.clogit <- polr(y ~ dose, data=Trauma2, weight=count)
> summary(fit.clogit)
Coefficients:
         Value Std. Error  t value  # actually, z statistics
dose 0.1754816 0.05671224 3.094245
Intercepts:
    Value   Std. Error t value
1|2 -0.7192  0.1589    -4.5256
2|3 -0.3186  0.1569    -2.0308
3|4  0.6917  0.1597     4.3323
4|5  2.0570  0.1751    11.7493
Residual Deviance: 2461.349
> fitted(fit.clogit)
           1          2         3         4         5
1  0.2901467 0.08878330 0.2473217 0.2415357 0.1322126
2  0.2901467 0.08878330 0.2473217 0.2415357 0.1322126
...
20 0.1944866 0.07043618 0.2325084 0.2975294 0.2050394
-----------------------------------------------------------------
```

Note: `polr` uses the model formula $\text{logit}[P(y \leq j)] = \alpha_j - \boldsymbol{\beta}^T \mathbf{x}$.

Goodness-of-fit statistics:

Pearson $X^2 = 15.8$, *deviance* $G^2 = 18.2$ $(df = 16 - 5 = 11)$

P-values $= 0.15$ and 0.18; model seems to fit adequately.

Interpretation of $\hat{\beta}$: For dose $i+1$, estimated odds of outcome $\leq j$ (instead of $> j$) equal $\exp(-0.176) = 0.84$ times estimated odds for dose i.

Equivalently, for dose $i+1$, estimated odds of outcome $\geq j$ (instead of $< j$) equal $\exp(0.176) = 1.19$ times estimated odds for dose i.

95% Wald confidence interval for $\exp(-\beta)$ is $e^{0.176 \pm 1.96(0.056)} = (1.07, 1.33)$.

With a `vglm` fit, profile likelihood CI's are also available:

```
-------------------------------------------------------------------------------
fit <- vglm(cbind(y1,y2,y3,y4,y5) ~ dose, family=cumulative(parallel=TRUE), data=Trauma)
> confint(fit, method="profile")
                   2.5 %      97.5 %
dose          -0.2868832 -0.06449357

> exp(0.06449); exp(0.28688)
[1] 1.066615
[1] 1.332264
-------------------------------------------------------------------------------
```

Cumulative odds ratio for doses 1 and 4 equals $e^{(4-1)0.176} = 1.69$.

- Any equally-spaced scores (e.g. 0, 10, 20, 30) for dose provide same fitted values and same test statistics (different $\hat{\beta}$, SE).
- Unequally-spaced scores more natural in many cases (e.g., doses may be 0, 125, 250, 500). "Sensitivity analysis" usually shows substantive results don't depend much on that choice.
- The cumulative logit model uses ordinality of Y without assigning category scores.
- Alternative analysis treats dose as factor, using indicator variables. Double the log-likelihood increases by only 0.13, $df = 2$.

  With $\beta_1 = 0$ (R coding):
  $\hat{\beta}_2 = -0.12$, $\hat{\beta}_3 = -0.32$, $\hat{\beta}_4 = -0.52$ ($SE = 0.18$ each).

  With $\beta_4 = 0$ (SAS coding):
  $\hat{\beta}_1 = 0.52$, $\hat{\beta}_2 = 0.40$, $\hat{\beta}_3 = 0.20$ ($SE = 0.18$ each).

  Testing H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4$ gives likelihood-ratio (LR) stat. $= 9.745$ ($df = 3$, $P = 0.02$).

  Using ordinality often increases power (focused on $df = 1$).

R for modeling dose-response data, with dose as a factor, using the *vglm* function in the *VGAM* library:

```
------------------------------------------------------------------------------
> library(VGAM)
> fit2 <- vglm(cbind(y1,y2,y3,y4,y5) ~ factor(dose),
+   family=cumulative(parallel=TRUE), data=Trauma)
> summary(fit2)
Coefficients:
              Estimate Std. Error  z value
(Intercept):1 -0.91880    0.13204 -6.95875
(Intercept):2 -0.51826    0.12856 -4.03122
(Intercept):3  0.49215    0.12841  3.83255
(Intercept):4  1.85785    0.14527 12.78927
factor(dose)2 -0.11756    0.17843 -0.65885
factor(dose)3 -0.31740    0.17473 -1.81649
factor(dose)4 -0.52077    0.17795 -2.92657
----
Residual deviance: 18.04959 on 9 degrees of freedom
Log-likelihood: -48.80638 on 9 degrees of freedom

> anova(fit2) # vglm function for likelihood-ratio test of factor dose effect
Analysis of Deviance Table (Type II tests)
             Df Deviance Resid. Df Resid. Dev Pr(>Chi)
factor(dose)  3   9.7453        12     27.795  0.02086
------------------------------------------------------------------------------
```

Note that the factor effects $(0, -0.12, -0.32, -0.52)$ are monotone decreasing, not far from linear in scores (1,2,3,4).

Other properties of cumulative logit models

- With nonsparse contingency table data, can check goodness of fit using Pearson X^2, deviance G^2. With sparse data, can use deviance to compare nested models.
- Can use similar model with alternative "cumulative link"

$$\text{link}[P(Y \leq j)] = \alpha_j - \boldsymbol{\beta}^T \mathbf{x} \text{ (McCullagh 1980)}$$

  e.g., *cumulative probit* model results from underlying normal Y^*.
- Effects $\boldsymbol{\beta}$ invariant to choice and number of response categories (If model holds, same $\boldsymbol{\beta}$ when response scale collapsed in any way).
- Log likelihood is concave. McCullagh (1980) provided Fisher scoring iterative algorithm for cumulative link models.
- Inference uses standard methods for testing H_0: $\beta_j = 0$ (likelihood-ratio, Wald tests) and inverting tests of H_0: $\beta_j = \beta_{j0}$ to get confidence intervals for β_j.

Alternative ways of summarizing effects

- Survey paper on this by Agresti and Tarantola (2018)
- Can compare $\hat{P}(Y = 1)$ or $\hat{P}(Y = c)$ at maximum and minimum values of a predictor (at means of other predictors), or find "average marginal effects" for these probabilities.
- Summary measures of predictive power include R^2 for regression model for underlying latent response variable shown in next example.

Example: Modeling mental impairment

Y = mental impairment (1 = well, 2 = mild symptom formation, 3 = moderate symptom formation, 4 = impaired)

x_1 = life events index = composite measure of severity of important life events within past three years. In this sample of $n = 40$, it varied between 0 and 9, with mean = 4.3 and standard deviation = 2.7.

x_2 = socioeconomic status (SES) index (0 = low, 1 = high).

Here, we use the *latent variable* induced form of the cumulative logit model with proportional odds form,

$$\text{logit}[P(Y \leq j)] = \alpha_j - \beta_1 x_1 - \beta_2 x_2,$$

because we will approximate R^2 for the underlying latent variable model for Y^*, and because the `polr` (*proportional odds logistic regression*) function in `R` uses it and easily provides estimated probabilities at fixed settings of explanatory variables:

```
> Mental <- read.table("https://alanagresti.com/cda/CDA_data/Mental.dat", header=TRUE)
> Mental
   impair ses life
1       1   1    1
...
40      4   0    9
> library(MASS)
> y <- factor(Mental$impair) # polr function requires response to be a factor
> fit <- polr(y ~ life + ses, method="logistic", data=Mental)
> summary(fit) # not showing the 3 intercept parameter estimates
        Value  Std. Error  t value # these are actually z statistics, not t
life   0.3189      0.1210    2.635
ses   -1.1112      0.6109   -1.819
```

$\hat{\beta}_2 = -1.111$ suggests that mental impairment tends to decrease at higher level of SES. At mean life events, $\hat{P}(Y = 4) = \hat{P}(\text{impaired})$ is 0.300 at low SES and 0.124 at high SES, while $\hat{P}(Y = 1) = \hat{P}(\text{well})$ is 0.162 at low SES and 0.370 at high SES.

```
> predict(fit, data.frame(ses=0, life=mean(Mental$life)), type="probs")
     1       2       3       4
0.1618  0.3007  0.2373  0.3002 # predicted outcome prob's, 1=well, 4=impaired

> predict(fit, data.frame(ses=1, life=mean(Mental$life)), type="probs")
     1       2       3       4
0.3696  0.3537  0.1530  0.1237
```

$\hat{\beta}_1 = 0.319$ suggests that mental impairment tends to be worse with higher life events. At low SES (ses $= 0$), $\hat{P}(Y = 4)$ changes from 0.099 to 0.659 as life events increases from its minimum to maximum values.

At high SES (ses $= 1$), $\hat{P}(Y = 4)$ changes from 0.035 to 0.389.

```
-------------------------------------------------------------------------------
> predict(fit, data.frame(ses=0, life=min(Mental$life)), type="probs")
     1       2       3       4
0.4300  0.3408  0.1303  0.0989

> predict(fit, data.frame(ses=0, life=max(Mental$life)), type="probs")
      1       2       3       4
0.04103  0.1191  0.1805  0.6593



> predict(fit, data.frame(ses=1, life=min(Mental$life)), type="probs")
     1       2       3       4
0.6962  0.2146  0.0543  0.0349

> predict(fit, data.frame(ses=1, life=max(Mental$life)), type="probs")
     1       2       3       4
0.1150  0.2518  0.2440  0.3892
-------------------------------------------------------------------------------
```

R^2 based on the latent variable model

Let y_i^* = latent variable for subject i. Then

$$R_L^2 = \frac{\sum_i (y_i^* - \bar{y}^*)^2 - \sum_i (y_i^* - \hat{y}_i^*)^2}{\sum_i (y_i^* - \bar{y}^*)^2} = \frac{\sum_i (\hat{y}_i^* - \bar{y}^*)^2}{\sum_i (y_i^* - \bar{y}^*)^2}$$

equals estimated variance of $\hat{y}^*$ divided by the estimated variance of y^*.

Can estimate var($\hat{y}^*$) by variance of linear predictor, without intercept terms. Cannot observe latent variable, but can approximate its variance by estimated var($\hat{y}^*$) plus var(residual error) in latent variable model ($\pi^2/3 = 3.29$ for standard logistic dist. that yields logit link function).

We then divide estimated var($\hat{y}^*$) by approx. var(y^*) to approximate R_L^2 (available with vglm in VGAM package).

```
-------------------------------------------------------------------------------
> fit.vglm <- vglm(impair ~ life + ses, family=cumulative(parallel=TRUE), data=Mental)
> R2latvar(fit.vglm) # using vglm function in VGAM
[1] 0.2237626

> fit <- polr(y ~ life + ses, method="logistic", data=Mental) # do it ourselves
> var(fit$lp)/(var(fit$lp) + (pi^2)/3) # lp = linear predictor, pi = 3.14...
[1] 0.22376 # R-squared based on logistic latent var. model for cumul. logit
-------------------------------------------------------------------------------
```

This measure better reflects the underlying latent variable model than the more commonly used generalization of R^2 (McFadden "pseudo R^2"). It is based on the proportional reduction in *deviance* compared to the null model with only intercept terms.

That measure reduces to R^2 in the normal case where deviance measures variability around predicted values, but in more general case it is unclear how to interpret this on the scale of the log-likelihood function.

```
-------------------------------------------------------------------------------
> fit0 <- polr(y ~ 1, method="logistic", data=Mental) # null model

> (fit0$deviance - fit$deviance)/fit0$deviance
[1]   0.0912  # McFadden pseudo R-squared, based on reduction in deviance

# Which is more similar to what we get with ordinary linear modeling?
# (we got 0.224 with latent variable approach)

> summary(lm(impair ~ life + ses, data=Mental)) # impair scores 1, 2, 3, 4
Multiple R-squared:  0.2315
-------------------------------------------------------------------------------
```

Checking fit (general case) and selecting a model

- Lack of fit may result from omitted predictors (e.g., interaction), non-proportional odds effects, wrong link function. Often, lack of fit reflects effects of dispersion as well as location.
- Can check particular aspects of fit using likelihood-ratio test (change in deviance) to compare to more complex models.
- Likelihood-ratio test compares model to more general "non-proportional odds model" with effects $\{\boldsymbol{\beta}_j\}$, but fitting of more general model fails when cumulative probabilities out-of-order.
- When model with proportional odds structure fits poorly, can use $\{\hat{\boldsymbol{\beta}}_j\}$ in non-proportional odds model (e.g., after fitting binary logistic to each collapsing) to describe effects more fully.
- Even if proportional odds model has lack of fit, it may usefully summarize "first-order effects" and have good power for testing H_0: no effect, because of its parsimony.

R for modeling *dose-response* data without proportional odds, using vglm() in VGAM library without parallel=TRUE option:

```
> Trauma
  dose y1 y2 y3 y4 y5
1    1 59 25 46 48 32
2    2 48 21 44 47 30
3    3 44 14 54 64 31
4    4 43  4 49 58 41
> library(VGAM)
> fit3 <- vglm(cbind(y1,y2,y3,y4,y5) ~ dose, family=cumulative, data=Trauma)
> summary(fit3)
                  Value Std. Error  z value
(Intercept):1 -0.864585   0.194230 -4.45133
(Intercept):2 -0.093747   0.178494 -0.52521
(Intercept):3  0.706251   0.175576  4.02248
(Intercept):4  1.908668   0.238380  8.00684
dose:1        -0.112912   0.072881 -1.54926
dose:2        -0.268895   0.068319 -3.93585
dose:3        -0.182341   0.063855 -2.85555
dose:4        -0.119255   0.084702 -1.40793
Residual Deviance: 3.85163 on 8 degrees of freedom
Log-likelihood: -41.70741 on 8 degrees of freedom

> lrtest(fit3, fit) # vglm function for likelihood-ratio test comparing models
  #Df   LogLik  Df   Chisq  Pr(>Chisq)
1   8  -41.707
2  11  -48.873   3  14.331    0.002488
```

This test compares models with and without common β_j. The improvement in fit here is statistically significant ($P = 0.002$), but perhaps not substantively significant. Effect of dose is moderately negative for each cumulative probability, which simpler model summarizes by $\hat{\beta} = -0.175$.

Summary

- Logit models for nominal response pair each category with a baseline, imply logits for each pair of categories.
- Logit models for ordinal responses use logits of cumulative probabilities, with same effect for each logit (proportional odds). Interpret using odds ratios for binary collapsings of response, which use cumulative probability and its complement.
- More simply, for cumulative logit model, report estimated probabilities in highest and lowest categories of Y at lowest and highest values of a predictor (e.g., at mean of other explanatory var's), and report R^2 analog to summarize goodness of predictions.
- Much more detail on methods for ordinal data is in my separate book, *Analysis of Ordinal Categorical Data* (2nd ed. 2010)

MARGINAL MODELS FOR MULTIVARIATE DISCRETE RESPONSES

- Multivariate responses: Handling correlation due to repeated measurement and other forms of clustering
- Matched pairs: Comparing two *dependent* proportions; connections between a model and classical McNemar test
- Marginal logit models: Comparing several dependent proportions, while adjusting for covariates
- Generalized estimating equations (quasi-likelihood) method of estimating model parameters (GEE)

Comparing proportions for matched pairs

(*CDA*, Sec. 11.1)

Example: Prime minister's performance rated in successive months

	Survey 2		
Survey 1	Approve	Disapprove	
Approve	794	150	944
Disapprove	86	570	656
	880	720	1600

The standard methods for comparing proportions are based on *independent* samples, but the row and column margins here are *dependent* samples.

Comparisons of p_{1+} and p_{+1} with test or CI must take into account that they're calculated from same sample.

Consider $H_0 : \pi_{1+} = \pi_{+1}$ for binary case

	S	F	
S	π_{11}	π_{12}	π_{1+}
F	π_{21}	π_{22}	
	π_{+1}		

Note $\pi_{1+} = \pi_{+1} \Leftrightarrow \pi_{12} = \pi_{21}$

Marginal homogeneity $\Leftrightarrow$ Symmetry (but only when $I = 2$)

For large n, we can test H_0 using $z = \frac{p_{1+} - p_{+1}}{\sqrt{\widehat{var}(p_{1+} - p_{+1})}}$, with $N(0, 1)$ dist.

This uses a "non-null" SE, so we can construct a CI,

$$(p_{1+} - p_{+1}) \pm 1.96\sqrt{\widehat{Var}(p_{1+} - p_{+1})}.$$

Treating $\{n_{ij}\}$ as multinomial $\{\pi_{ij}\}$, then

$\text{Var}(p_{1+} - p_{+1}) =$

$$\frac{\pi_{1+}(1-\pi_{1+}) + \pi_{+1}(1-\pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})}{n}.$$

For matched samples, usually

$$\pi_{11}\pi_{22} >> \pi_{12}\pi_{21}$$

$\rightarrow$ variance is smaller than for independent samples.

(Recall variance has form $\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$ for indep. samples, n_1 with probability π_1 and n_2 with probability π_2.)

ex.

	Survey 2		
Survey 1	Approve	Disapprove	
Approve	794	150	944
Disapprove	86	570	656
	880	720	1600

Marginal proportions are $944/1600 = 0.59$, $880/1600 = 0.55$.
Standard error of difference of marginal proportions is 0.00955

$$z = \frac{0.59 - 0.55}{0.00955} = 4.19 \text{ for } H_0 : \pi_{1+} = \pi_{+1}$$

95% CI for true difference is $0.04 \pm 1.96(0.00955)$, or (0.02, 0.06).

$SE = 0.0175$ for independent samples of sizes 1600 each having proportions 0.59 and 0.55.

To test $H_0 : \pi_{1+} = \pi_{+1}$ $(\pi_{12} = \pi_{21})$, we can condition on $n_{12} + n_{21}$ and use $n_{12} \sim$ binomial $(n_{12} + n_{21}, \frac{1}{2})$ under H_0.

Normal approximation to binomial suggests test statistic

$$z = \frac{n_{12} - (\frac{1}{2})(n_{12} + n_{21})}{\sqrt{(n_{12} + n_{21})(\frac{1}{2})(\frac{1}{2})}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \to N(0, 1),$$

$$z^2 \to \chi^2_1 \quad (\underline{\text{McNemar's}} \text{ test, using } \textit{null} \text{ SE})$$

ex. $z = \frac{150-86}{\sqrt{150+86}} = 4.16$, $z^2 = 17.36$, P-value $= 0.00003$.

```
-----------------------------------------------------------------------
> performance <- matrix(c(794,150,86,570), ncol=2, byrow=TRUE,
+   dimnames = list("Survey 1" = c("Approve","Disapprove"),
+       "Survey 2" = c("Approve","Disapprove")))
> performance
            Survey 2
Survey 1     Approve Disapprove
  Approve        794        150
  Disapprove      86        570
> mcnemar.test(performance,correct=FALSE)  # no continuity correction
        McNemar's Chi-squared test
data:  performance
McNemar's chi-squared = 17.356, df = 1, p-value = 3.099e-05
-----------------------------------------------------------------------
```

McNemar's test *ignores* the main-diagonal counts. One justification for this uses a logistic model for the matched pairs $(Y_{i1},\ Y_{i2})$, $i = 1, \ldots, n$.

$$\log\left[\frac{P(Y_{it}=1)}{P(Y_{it}=0)}\right] = \alpha_i, \quad \text{time 1}$$
$$= \alpha_i + \beta, \quad \text{time 2}$$

Here, β compares margins as $\exp(\beta) = \frac{P(Y_{i2}=1)/P(Y_{i2}=0)}{P(Y_{i1}=1)/P(Y_{i1}=0)}$, with $\beta = 0$ corresponding to marginal homogeneity.

After conditioning to eliminate $\{\alpha_i\}$ (*nuisance parameters*, one for each subject), conditional ML estimate of odds ratio is (Cox 1958)

$\exp(\hat{\beta}) = n_{12}/n_{21} = 86/150 = 0.57$. (*CDA*, Sec. 11.2.2, 11.2.3)
By the delta method, $\hat{\beta}$ has $SE = \sqrt{\frac{1}{n_{21}} + \frac{1}{n_{12}}}$.

For example, we obtain 95% CI for odds ratio $\exp(\beta)$ by exponentiating endpoints of $\log(86/150) \pm 1.96\sqrt{\frac{1}{86} + \frac{1}{150}}$, which is (0.44, 0.75).

This is a *conditional* (subject-specific) model; i.e., effect β is conditional on the subject, as model refers to data in form of n 2×2 tables, where partial table i shows pair of responses for subject i.

		Response		
Subject	Time	Failure	Success	Total
1	1	1	0	1
	2	0	1	1
2	1	0	1	1
	2	1	0	1
3	1	1	0	1
	2	1	0	1
⋮				
n	1	0	1	1
	2	0	1	1

By contrast, a *marginal* (population-averaged) model refers to the overall proportions after collapsing over subjects. The marginal binary logit model for a subject selected at random at time t is

$$\log\left[\frac{P(Y_t=1)}{P(Y_t=0)}\right] = \alpha, \qquad t=1$$
$$= \alpha+\beta, \qquad t=2.$$

Marginal model refers to 2×2 table that collapses 2×2×n subject-specific table over subjects, yielding:

	Response		
Time	Failure	Success	
1	a	b	n
2	c	d	n

The cells in this time×response table form the row and column margins of $Y_1 \times Y_2$ table (as in repeated survey example).

$Y_1 \times Y_2$ (i.e., response 1 $\times$ response 2) table

	Time 2		
Time 1	Failure	Success	
Failure			a
Success			b
	c	d	n

For this table $\hat{\beta}$ = log odds ratio of marginal proportions

$$\exp(\hat{\beta}) = [p_{+1}/(1-p_{+1})]/[p_{1+}/(1-p_{1+})]$$

$$= [880/720]/[944/656] = 0.85 \text{ in survey example } \quad (\hat{\beta} = -0.163).$$

Note: For model with identity link function

$$\begin{aligned} P(Y_t = 1) &= \alpha, & t = 1 \\ &= \alpha + \beta, & t = 2 \end{aligned}$$

$\hat{\beta} = p_{+1} - p_{1+}$, i.e. difference of marginal proportions
$= 880/1600 - 944/1600 = -0.04$ comparing survey 1 with survey 2

R: marginal logit model (fitted using GEE) has estimated marginal odds ratio $e^{-0.163295} = 0.85$.

```
------------------------------------------------------------------------------------------
> Opinions <- read.table("https://alanagresti.com/glm/data/Opinions.dat", header=TRUE)
> Opinions # subject-specific data file
     person survey y
1         1      1 1
2         1      2 1

3199   1600      1 0
3200   1600      2 0

> library(gee)
> fit <- gee(y ~ survey, id=person, family=binomial, data=Opinions)
> summary(fit)
 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial

Coefficients:
             Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept)  0.5272601 0.11343784  4.648009  0.07542110  6.990882
survey      -0.1632947 0.07149937 -2.283862  0.03902673 -4.184176
------------------------------------------------------------------------------------------
```

The 'naive' results here are what we get by treating the binomials as independent instead of dependent, in which case we overestimate the SE.

R: marginal model with identity link function (fitted using GEE methodology), for which estimated difference of marginal proportions is -0.04, with $SE = 0.00955$, as quoted earlier in notes.

```
> Opinions
     person survey y
1         1      1 1
2         1      2 1
3         2      1 1
4         2      2 1

3199   1600      1 0
3200   1600      2 0

> library(gee)
> fit <- gee(y~survey, id=person, family=binomial(link="identity"), data=Opinions)
> summary(fit)

GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 Link:                      Identity
 Variance to Mean Relation: Binomial
 Correlation Structure:     Independent

Coefficients:
            Estimate Naive S.E.  Naive z Robust S.E.  Robust z
(Intercept)     0.63 0.02756651 22.85382 0.018167622 34.677075
survey         -0.04 0.01749475 -2.28640 0.009549215 -4.188826
```

The 'naive' results here are what we get by treating the binomials as independent instead of dependent, in which case we overestimate the SE.

Marginal model fitting

(*CDA*, Sec. 12.1)

A model for marginal logits is a multivariate model that simultaneously forms logits for each margin.

Awkward to apply ML, especially for large tables, because model applies to marginal distributions but likelihood is in terms of joint multinomial probabilities.

Common to use "GEE", a multivariate generalization of **quasi-likelihood** methods. Estimates are solutions of "generalized estimating equations" similar to likelihood equations, without fully specifying distribution.

Generalized Estimating Equations (GEE) Approach

(*CDA*, Sec. 12.2, 12.3)

GEE useful when primary interest is modeling marginal dist. of Y_t as fn. of x's, rather than modeling association among $(Y_1, Y_2, \ldots, Y_T)$ as loglinear models do. Not necessary to assume full multivariate distribution.

Steps of GEE Methodology:

- Assume marginal regression model, "working" covariance structure (e.g., exchangeable, autoregressive, independence).
- Estimates are consistent even if covariance structure misspecified (if marginal model correct).
- Method generates robust estimates of standard errors that are valid even if covariance structure misspecified. Uses "sandwich" covariance matrix incorporating empirical variability.

Some motivation, first for univariate QL:

- The ML estimates $\hat{\boldsymbol{\beta}}$ for a GLM are solutions of likelihood equations

$$\mathbf{u}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^T \frac{(y_i - \mu_i)}{v(\mu_i)} = \mathbf{0},$$

  for variance function $v(\mu)$, such as $v(\mu) = \mu$ for Poisson, $v(\mu) = \mu(1-\mu)$ for binary data.

- The ML estimates depend on the choice of distribution only through the mean and variance!

- For GEE method, with $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$,

$$\text{var}(\hat{\boldsymbol{\beta}}) \approx n \left[\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \text{var}(\mathbf{y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i\right] \left[\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i\right]^{-1}.$$

  where $\mathbf{V}_i$ is working covariance matrix for observation i.

- To obtain estimated covariance matrix, we replace parameters by their estimates and replace $\text{var}(\mathbf{Y}_i)$ by $(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T$ to get an empirical sandwich covariance matrix that yields more robust SE values.

- Originally specified (Liang and Zeger 1986) for univariate Y_t (e.g., binomial, Poisson), but extensions exist for cumulative logit models.
- Some versions base working correlation on structure for odds ratios (e.g., exchangeable), more natural for categorical responses.
- No likelihood function, since do not fully specify joint dist. of $(Y_1, Y_2, \ldots, Y_T)$.
- In GEE methodology, cannot use standard likelihood-ratio tests, model comparison, tests of fit.
- R: *gee* package has *gee* function that handles binary data. Touloumis et al. (2013) use odds ratios for multinomial data (ordinal or nominal) with *multgee* R package.

Example (binary data): Support for legalizing abortion in three situations (*CDA*, Sec. 13.3.2)

	General Social Survey Sequence of Responses in Three Situations							
Gender	(1,1,1)	(1,1,0)	(0,1,1)	(0,1,0)	(1,0,1)	(1,0,0)	(0,0,1)	(0,0,0)
Male	342	26	6	21	11	32	19	356
Female	440	25	14	18	14	47	22	457

Situations are (1) if the family has a very low income and cannot afford any more children, (2) when the woman is not married and does not want to marry the man, and (3) when the woman wants it for any reason. 1, yes; 0, no.

Note: Overwhelming majority of responses are (0,0,0) and (1,1,1), suggesting strong pairwise associations.

Marginal model for response Y_{ij} of subject i for situation j:

$$\text{logit}[P(Y_{ij} = 1)] = \alpha + \beta_j + \gamma x_i.$$

where $x_i = 1$ for females and 0 for males and situation effects $\{\beta_j\}$ satisfy constraint such as $\beta_3 = 0$.

R fitting, assuming exchangeable working correlation structure:

```
------------------------------------------------------------------------------
> Abortion <- read.table("https://alanagresti.com/glm/data/Abortion.dat",header=TRUE)
> Abortion
     gender response situation person
1         1        1         1      1
2         1        1         2      1
3         1        1         3      1
...
5548      0        0         1   1850
5549      0        0         2   1850
5550      0        0         3   1850

> sit <- factor(Abortion$situation, levels=c(3,1,2))
> library(gee)
> fit.gee <- gee(response ~ sit + gender, id=person, family=binomial,
+                corstr="exchangeable", data=Abortion) # cluster on "id" variable
> summary(fit.gee)
             Estimate  Naive S.E.  Naive z  Robust S.E.  Robust z
(Intercept)  -0.12533    0.06783    -1.848     0.06758    -1.854
sit1          0.14935    0.02814     5.307     0.02974     5.022
sit2          0.05202    0.02815     1.848     0.02705     1.923
gender        0.00344    0.08791     0.039     0.08784     0.039
Working Correlation
        [,1]    [,2]    [,3]
[1,] 1.00000 0.81733 0.81733
[2,] 0.81733 1.00000 0.81733
[3,] 0.81733 0.81733 1.00000
------------------------------------------------------------------------------
```

R fitting, assuming independence working correlation structure:

```
-------------------------------------------------------------------------------
> fit.gee2 <- gee(response ~ sit + gender, id=person, family=binomial,
+             corstr="independence", data=Abortion)
> summary(fit.gee2)
              Estimate  Naive S.E.  Naive z  Robust S.E.  Robust z
(Intercept)   -0.12541    0.05562   -2.255      0.06758    -1.856
sit1           0.14935    0.06585    2.268      0.02974     5.022
sit2           0.05202    0.06587    0.790      0.02705     1.923
gender         0.00358    0.05416    0.066      0.08784     0.041
Working Correlation
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
-------------------------------------------------------------------------
```

Because of strong positive correlation between pairs of responses (estimated exchangeable correlation $= 0.817$), naive *SE*'s based on independence working correlations are very misleading.

Naive *SE*'s tend to be too small for between-subject effects and too large for within-subject effects.

Example: Randomized Clinical Trial for Treating Insomnia

Randomized, double-blind clinical trial compared hypnotic drug with placebo in patients with insomnia problems. (*CDA*, Sec. 12.1.3)

		Time to Falling Asleep			
	Initial	Follow-up			
Treatment		<20	20–30	30–60	>60
Active	<20	7	4	1	0
	20–30	11	5	2	2
	30–60	13	23	3	1
	>60	9	17	13	8
Placebo	<20	7	4	2	1
	20–30	14	5	1	0
	30–60	6	9	18	2
	>60	4	11	14	22

Y_t = time to fall asleep
x = treatment (0 = placebo, 1 = active)
t = occasion (0 = initial, 1 = follow-up after 2 weeks)

$$\text{Model:} \quad \text{logit}[P(Y_t \leq j)] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3 (t \cdot x), \quad j = 1, 2, 3.$$

GEE estimates (with independence working equations and robust SE):

$\hat{\beta}_1 = 1.04$ ($SE = 0.17$), occasion effect only for placebo,
$\hat{\beta}_2 = 0.03$ ($SE = 0.24$), treatment effect only initially,
$\hat{\beta}_3 = 0.71$ ($SE = 0.24$), interaction (significant).

Considerable evidence that distribution of time to fall asleep decreased more for treatment than placebo group.

Occasion effect = 1.04 for placebo, 1.04 + 0.71 = 1.75 for active.
Occasion odds ratios $e^{1.04} = 2.8$ for placebo, $e^{1.75} = 5.7$ for active.

Treatment odds ratio $e^{0.03} = 1.03$ initially,
$e^{0.03+0.71} = 2.1$ follow-up.

R for GEE analysis of insomnia data (using *multgee* package):

```
> Insomnia <- read.table("https://alanagresti.com/cda/CDA_data/Insomnia.dat", header=TRUE)
> Insomnia
    case     treat  occasion   response
1      1         1         0          1
2      1         1         1          1
...
478  239         0         1          4
> fit <- vglm(response ~ occasion+treat+occasion:treat,family=cumulative(parallel=TRUE),data=Insomnia)
> summary(fit) # naively treating observations as independent
               Estimate Std. Error z value Pr(>|z|)
(Intercept):1  -2.26709    0.20274 -11.182  < 2e-16
(Intercept):2  -0.95146    0.17848  -5.331 9.78e-08
(Intercept):3   0.35174    0.17269   2.037   0.0417
occasion        1.03808    0.23760   4.369 1.25e-05
treat           0.03361    0.23690   0.142   0.8872
occasion:treat  0.70776    0.33418   2.118   0.0342
---
> library(multgee)
> fit2 <- ordLORgee(response ~ occasion+treat+occasion:treat,id=case,LORstr = "independence",data=Insomnia)
> summary(fit2)
Local Odds Ratios Structure: independence
               Estimate   san.se    san.z Pr(>|san.z|)  # san = sandwich
beta01         -2.26709  0.21876 -10.3633      < 2e-16  # 3 intercepts
beta02         -0.95146  0.18092  -5.2591      < 2e-16
beta03          0.35174  0.17842   1.9714      0.04868
occasion        1.03808  0.16759   6.1943      < 2e-16
treat           0.03361  0.23844   0.1410      0.88790
occasion:treat  0.70776  0.24352   2.9064      0.00366
```

Positive correlation between responses (0.44 with scores (10, 25, 45, 75)), so within-subject estimates have smaller SE's than if observations were independent.

Alternative (*transitional*) model:

$$\text{logit}[P(y_2 \leq j)] = \alpha_j + \beta_1 x + \beta_2 y_1.$$

This is an ordinary univariate model that can be fitted treating y_2 as the response variable and y_1 as a covariate.

```
> Insomnia2 <- read.table("https://alanagresti.com/cda/CDA_data/Insomnia2.dat",
+                         header=TRUE)
> Insomnia2
  treatment initial y1 y2 y3 y4
1         1       10  7  4  1  0
2         1       25 11  5  2  2
3         1       45 13 23  3  1
4         1       75  9 17 13  8
5         0       10  7  4  2  1
6         0       25 14  5  1  0
7         0       45  6  9 18  2
8         0       75  4 11 14 22
> library(VGAM)
> fit <- vglm(cbind(y1,y2,y3,y4) ~ treatment + initial,
+             family=cumulative(parallel=TRUE), data=Insomnia2)
> summary(fit)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  0.581161   0.318824   1.823 0.068330 .
(Intercept):2  2.277442   0.355043   6.415 1.41e-10
(Intercept):3  3.750952   0.399804   9.382  < 2e-16
treatment      0.884679   0.245552   3.603 0.000315
initial       -0.042106   0.005796  -7.264 3.75e-13
```

For any given value for the initial response, the estimated odds of falling asleep by a particular time for the active treatment are $\exp(0.885) = 2.42$ times those for the placebo group.

$\hat{\beta}_1 = 0.885$ ($SE = 0.246$) provides strong evidence that follow-up time to fall asleep is lower for the active drug group.

For any given value for the initial response, the estimated odds of falling asleep by a particular time for the active treatment are $\exp(0.885) = 2.4$ times those for the placebo group.

This approach has advantages over marginal models when initial marginal distributions differ (CDA, Sec. 12.4.5).

RANDOM EFFECTS IN GENERALIZED LINEAR MIXED MODELS

(*CDA*, Chap. 13)

For binary response, "random intercept model" for observation t in cluster (e.g., subject) i is the *generalized linear mixed model* (GLMM)

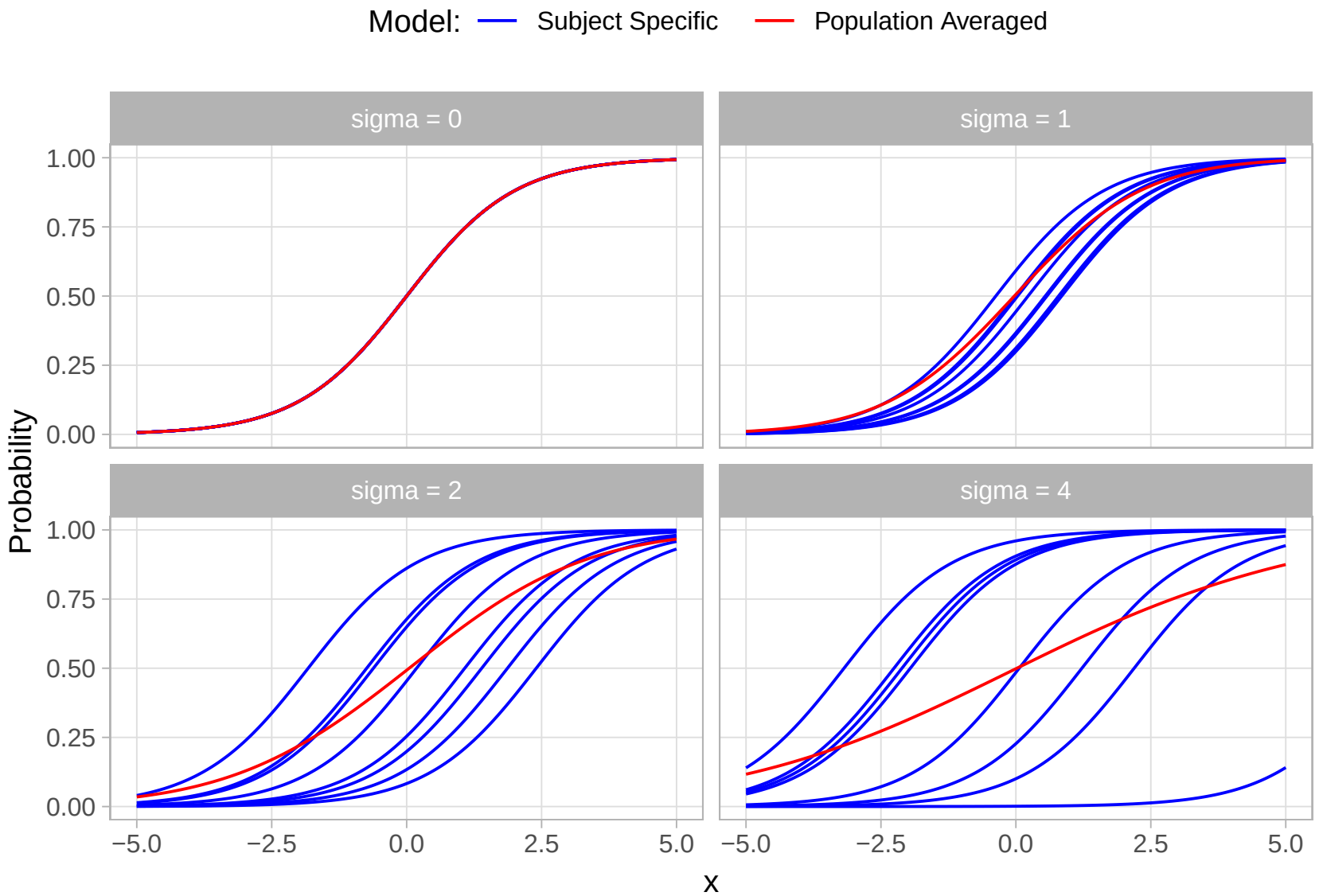
$$\text{logit}[P(Y_{it}=1)] = u_i + \alpha + \boldsymbol{\beta}^T \mathbf{x}_{it}$$

where $\{u_i\}$ are *iid* from $N(0, \sigma^2)$. Introducing random u_i in model induces correlation between repeated responses because of *subject heterogeneity*.

- e.g., Large positive u_i implies high $P(Y_{it}=1)$ for each t (Matched pairs: tend to yield $(Y_{i1}, Y_{i2}) = (1, 1)$ sequences).
- Large negative u_i implies high $P(Y_{it}=0)$ each t (tend to yield $(Y_{i1}, Y_{i2}) = (0, 0)$ sequences of responses).
- As σ increases, correlation between (Y_{i1}, Y_{i2}) increases.
- Recall that if $Y_1 = U + X$, $Y_2 = U + Z$, with U, X, and Z uncorrelated, then $\text{Corr}(Y_1, Y_2) = \dfrac{Var(U)}{\sqrt{[Var(U)+Var(X)][Var(U)+Var(Z)]}}$.
- $\sigma = 0$ corresponds to repeated responses being independent.

- Analogous models exist for other links, multinomial responses.
- Can have differing numbers of observations per cluster and explanatory variables with values varying by observation.
- GLMM approach has disadvantage of adding normal dist. assumption about random effect, but results usually robust to that choice.
- GLMM approach extends to multivariate normal for multivariate random effect (e.g, useful for multilevel models).
- GLMMs are *subject-specific* models; predictors effects are described at subject level rather than *population-averaged* as in marginal models.
- Since effect β is subject-specific, it differs in size from effect in corresponding marginal model.
- As correlation increases among observations in cluster (i.e., var(random effects) increases), subject-specific (conditional) effects in GLMMs tend to increase in magnitude relative to population-averaged (*marginal*) effects.

Logistic random-intercept model showing the *conditional* subject-specific curves, and the corresponding *marginal* (population-averaged) curve that averages over the subject-specific curves, for various levels of heterogeneity



GLMM Model Fitting

- GLMM fitting uses *marginal ML*: Integrate out u_i to obtain marginal likelihood depending on $\boldsymbol{\beta}$ and "variance components."
- Integrating out random effects can be computationally complex. Numerical integration with Gauss-Hermite quadrature approximates likelihood by finite sum; approximation improves as increase number of *quadrature points*.
- Many-dimensional random effects or complex structure requires Monte Carlo methods to approximate likelihood and ML parameter estimates.
- Once obtain likelihood function, software uses Newton-Raphson to maximize wrt $(\boldsymbol{\beta}, \sigma)$, obtain *SE* values.

Example: Support for legalizing abortion in three situations (*CDA*, Sec. 13.3.2)

	General Social Survey Sequence of Responses in Three Situations							
Gender	(1,1,1)	(1,1,0)	(0,1,1)	(0,1,0)	(1,0,1)	(1,0,0)	(0,0,1)	(0,0,0)
Male	342	26	6	21	11	32	19	356
Female	440	25	14	18	14	47	22	457

Situations are (1) if the family has a very low income and cannot afford any more children, (2) when the woman is not married and does not want to marry the man, and (3) when the woman wants it for any reason. 1, yes; 0, no.

Random effects model

$$\text{logit}[P(Y_{ij} = 1 \mid u_i)] = u_i + \alpha + \beta_j + \gamma x_i,$$

where $x_i = 1$ for females and 0 for males and $\{u_i\}$ are assumed to be independent $N(0, \sigma^2)$.

R for random effects modeling of abortion opinion data:

```
> Abortion <- read.table("https://alanagresti.com/glm/data/Abortion.dat",header=TRUE)
> Abortion
     gender response situation person
1         1        1         1      1
2         1        1         2      1
3         1        1         3      1
...
5548      0        0         1   1850
5549      0        0         2   1850
5550      0        0         3   1850

> sit <- factor(Abortion$situation, levels=c(3,1,2))
> library(lme4)
> fit <- glmer(response ~ (1|person) + sit + gender, family=binomial,
+              nAGQ=100, data=Abortion)
> summary(fit)
Random effects:
 Groups Name         Variance  Std.Dev.
 person (Intercept)  76.49     8.746
Fixed effects:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -0.61936     0.37820   -1.638     0.101
sit1          0.83478     0.16004    5.216  1.83e-07
sit2          0.29245     0.15670    1.866     0.062 .
gender        0.01261     0.48955    0.026     0.979
```

Abortion opinions data:

Comparison of estimates with random effects (GLMM) and marginal models (GEE)

Estimate	GLMM	GEE
$\hat{\beta}_1$	0.835	0.149
	(0.160)	(0.066)
$\hat{\beta}_2$	0.292	0.052
	(0.157)	(0.066)

Note the item effects are much stronger with the random effects model (GLMM) than with the marginal model (GEE), which is not surprising because of the strong correlation (large random effects variability).

Cumulative Logit Random Effects Models for Ordinal Data

ex. Insomnia: Y_t = time to fall asleep with treatment x (0 = placebo, 1 = active) at occasion t (0 = initial, 1 = follow-up). Marginal model

$$\text{logit}[P(Y_t \leq j)] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3 (t \cdot x), \quad j = 1, 2, 3.$$

By contrast, random effects model with common random intercept for each logit is

$$\text{logit}[P(Y_{it} \leq j)] = u_i + \alpha_j + \beta_1 t + \beta_2 x + \beta_3 (t \cdot x).$$

Effect	GEE	GLMM
Occasion (t)	1.04 (0.17)	1.60 (0.28)
Treatment (x)	0.03 (0.24)	0.06 (0.37)
Treatment×Occasion ($t \times x$)	0.71 (0.24)	1.08 (0.38)

Results are substantively similar, but random-effects model estimates and SE values about 50% larger. Reflects heterogeneity ($\hat{\sigma} = 1.9$) and moderate (Y_{i1}, Y_{i2}) association (correlation = 0.44).

For R, here we use the *clmm* function in the `ordinal` package, which uses latent variable parameterization:

```
> Insomnia <- read.table("https://alanagresti.com/cda/CDA_data/Insomnia.dat", header=TRUE)
> Insomnia
    case  treat occasion  response
1      1      1        0         1
2      1      1        1         1
...
477  239      0        0         4
478  239      0        1         4
> attach(Insomnia)
> library(ordinal)
> y <- factor(response) #  # response var. for clmm function must be a factor
> fit <- clmm(y ~ (1|case) + occasion + treat + occasion:treat, nAGQ=20)
              # (1|case) puts random intercept for case in model
              # nAGQ = no. points for adaptive Gaussian quadrature.
> summary(fit)
Random effects:
 Groups Name        Variance Std.Dev.
 case   (Intercept) 3.628    1.905
Number of groups:  case 239

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
occasion        -1.60158    0.28336  -5.652 1.58e-08
treat           -0.05785    0.36629  -0.158  0.87450
occasion:treat  -1.08129    0.38046  -2.842  0.00448
---
Threshold coefficients:         # intercept estimates
    Estimate Std. Error z value
1|2  -3.4896     0.3588  -9.727
2|3  -1.4846     0.2903  -5.114
3|4   0.5613     0.2701   2.078
```

Fit Statistics for Conditional Distribution

-2 log L(outcome | r. effects)	789.00

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	case	3.6280	0.8815

Solutions for Fixed Effects

Effect	outcome	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	1	-3.4896	0.3588	237	-9.73	<.0001
Intercept	2	-1.4846	0.2903	237	-5.11	<.0001
Intercept	3	0.5613	0.2702	237	2.08	0.0388
treat		0.05786	0.3663	235	0.16	0.8746
occasion		1.6016	0.2834	235	5.65	<.0001
treat*occasion		1.0813	0.3805	235	2.84	0.0049

Estimated standard deviation of random effects is $\sqrt{3.628} = 1.90$.

Random Effects Modeling with Cluster Random Samples

Random effects can represent *clusters* of subjects that are similar in some way.

For example, some studies sample families and observe variables for every person in each family. We can regard all the people in a particular family as a cluster. For given values of explanatory variables, two people in the same family tend to be more alike than two people in different families.

Identifying the families in the model using a random effect for each family accounts for the correlation among observations within a family.

Multilevel (Hierarchical) Models

Multilevel (hierarchical) models describe observations having a nested nature: Units at one level are contained within units of another level.

Many applications in education: Students nested in schools nested in school districts or regions.

Multilevel models contain terms for the different levels of units. They treat terms for the sampled units on which there are multiple observations as *random effects* rather than fixed effects.

The random effects can enter the model at each level of the hierarchy.

For categorical response variables, we use logistic regression models with random effects, which are also special cases of *generalized linear mixed models*.

Example: Smoking prevention and cessation study

Study of efficacy of programs for discouraging young people from starting or continuing to smoke (Hedeker and Gibbons (2006, p. 9).

Four groups, defined by a 2×2 factorial design according to whether a student was exposed to a school-based curriculum (SC; 1 = yes, 0 = no) and a television-based prevention program (TV; 1 = yes, 0 = no).

Subjects: 1600 seventh-grade students from 135 classrooms in 28 Los Angeles schools. The schools were randomly assigned to the four intervention conditions.

Response variable: tobacco and health knowledge (THK) scale, measured at the end of the study.

THK also observed at beginning of study, and that measure (PTHK = Pre-THK) used as explanatory variable. THK took values between 0 and 7, with mean 2.66 and standard deviation 1.38.

Here, we let $y = 1$ if $THK > 2$, $y = 0$ if $THK \leq 2$.

Table: **Part of Smoking Prevention and Cessation Data File**

School	Class	SC	TV	PTHK	THK	y
403	403101	1	0	2	3	1
403	403101	1	0	4	4	1
...						
515	515113	0	0	3	3	1

Let y_{ijk} denote the follow-up THK score for student i within classroom j in school k. We consider the multilevel model

$$\text{logit}[P(Y_{ijk} = 1 \mid s_k, c_{jk})] = \alpha + \beta_1 \text{PTHK}_{ijk} + \beta_2 \text{SC}_{ijk} + \beta_3 \text{TV}_{ijk} + s_k + c_{jk}.$$

At one level, s_k is a random effect for school k, assumed to have a $N(0, \sigma_s^2)$ distribution for unknown σ_s^2.

At another level, c_{jk} is a random effect for classroom j in school k, assumed to be $N(0, \sigma_c^2)$ with unknown σ_c^2.

R for multilevel modeling of smoking prevention study:

```
> Smoking <- read.table("https://alanagresti.com/cda/CDA_data/Smoking.dat",header=TRUE)
> library(lme4)
> fit <- glmer(y ~ PTHK + SC + TV + (1|school) + (1|class),
               family=binomial(link="logit"), data=Smoking)
> summary(fit)
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )

Random effects:
 Groups Name        Variance Std.Dev.
 class  (Intercept) 0.16728  0.4090
 school (Intercept) 0.06413  0.2532
Number of obs: 1600, groups:  class, 135; school, 28

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.13163    0.17827  -6.348 2.18e-10
PTHK         0.39512    0.04627   8.539  < 2e-16
SC           0.80014    0.16893   4.737 2.17e-06
TV           0.10786    0.16819   0.641    0.521
```

The variance component estimates $\hat{\sigma}_s^2 = 0.064$ and $\hat{\sigma}_c^2 = 0.167$ indicate slightly more variability among classrooms within schools than among schools.

Suppose we ignored the clustering of observations in classrooms and schools and treated the 1600 observations as independent by fitting the ordinary logistic model

$$\text{logit}[P(Y_{ijk} = 1)] = \alpha + \beta_1 \text{PTHK}_{ijk} + \beta_2 \text{SC}_{ijk} + \beta_3 \text{TV}_{ijk}.$$

```
------------------------------------------------------------------------------
> fit.glm <- glm(y ~ PTHK + SC + TV, family=binomial(link="logit"), data=Smoking)
> summary(fit.glm)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.11920    0.13127  -8.526  < 2e-16
PTHK         0.39886    0.04401   9.063  < 2e-16
SC           0.76532    0.10563   7.245 4.32e-13
TV           0.12305    0.10462   1.176     0.24

> fit <- glmer(y ~ PTHK + SC + TV + (1|school) + (1|class),
               family=binomial(link="logit"), data=Smoking)
> summary(fit)
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.13163    0.17827  -6.348 2.18e-10
PTHK         0.39512    0.04627   8.539  < 2e-16
SC           0.80014    0.16893   4.737 2.17e-06
TV           0.10786    0.16819   0.641    0.521
------------------------------------------------------------------------------
```

The estimated fixed effects are similar to those in the multilevel model, but the *SE* values are quite dramatically underestimated for the between-subjects effects (SC and TV).

Choice of marginal vs. random effects models?

- Each approach has pros and cons.
- For certain applications, population-averaged effects (marginal models) or subject-specific effects (conditional models) may be more relevant.
  e.g., population-averaged perhaps more useful if primary focus is on between-subjects effects, subject-specific more useful if primary focus is on within-subjects effects.
- Marginal model has advantage that does not require correct specification of full distribution and its dependence structure, but GEE marginal approach does not yield likelihood inference and sandwich SE may require large n to perform well.
- Conditional model has advantage of likelihood function, but more potential for model misspecification, improper use, especially for the less statistically sophisticated user.

Summary: Modeling Clustered, Correlated Responses

- Multivariate data requires ways of dealing with within-cluster correlation.
- Matched-pairs data: Marginal homogeneity and symmetry, comparing dependent proportions (e.g., McNemar test)
- Marginal modeling of repeated categorical measurement data with covariates, using logits for dependent proportions and quasi-likelihood (GEE) methodology
- Generalized linear mixed models provide a subject-specific approach that is an alternative to marginal models; e.g., random effect for intercept can induce dependence.
- Marginal model often preferred for overall summary of "between-subject" effects (e.g., compare females and males), while GLMM preferred to describe "within-subject" effects and summarize heterogeneity among subjects (random effects σ).

The end!

We hope this short course has been useful, giving you an overview of the most important methods for analyzing categorical response data through several examples with interpretations and software output.

Thanks very much for your time and attention!