York SPIDA John Fox

Generalized Linear Models and Related Topics

1. Topics

- Introduction to maximum-likelihood estimation
- Introduction to Bayesian inferenceLogit and probit models for dichotmous data
- The structure of generalized linear models
- Poisson and other generalized linear models for count data
- Diagnostics for generalized linear models (as time permits)
- Logit and Loglinear models for contingency tables (as time permits)
- Implementation of generalized linear models in R

2. Introduction to Maximum-Likelihood Estimation

- The *method of maximum likelihood* provides estimators that have both a reasonable intuitive basis and many desirable statistical properties.
- The method is very broadly applicable and is simple to apply.
- Once a maximum-likelihood estimator is derived, the general theory of maximum-likelihood estimation provides standard errors, statistical tests, and other results useful for statistical inference.
- A disadvantage of the method is that it frequently requires strong assumptions about the structure of the data.

2.1 An Example

- We want to estimate the probability π of getting a head upon flipping a particular coin.
  - We flip the coin 'independently' 10 times (i.e., we sample $n = 10$ flips), obtaining the following result: $HHTHHHTTHH$.
  - The probability of obtaining this sequence — in advance of collecting the data — is a function of the unknown parameter π:

$$\begin{aligned}\Pr(\text{data}|\text{parameter}) &= \Pr(HHTHHHTTHH|\pi)\\ &= \pi\pi(1-\pi)\pi\pi\pi(1-\pi)(1-\pi)\pi\pi\\ &= \pi^7(1-\pi)^3\end{aligned}$$

  - But the data for our particular sample are *fixed*: We have already collected them.
  - The parameter π also has a fixed value, but this value is unknown, and so we can let it vary in our imagination between 0 and 1, treating the probability of the observed data as a function of π.

- This function is called the likelihood function:

$$L(\text{parameter}|\text{data}) = L(\pi|HHTHHHTTHH) = \pi^7(1-\pi)^3$$

- The probability function and the likelihood function are given by the same equation, but the probability function is a function of the data with the value of the parameter fixed, while the likelihood function is a function of the parameter with the data fixed.

- Here are some representative values of the likelihood for different values of π:

π	$L(\pi\|\text{data}) = \pi^7(1-\pi)^3$
0.0	0.0
.1	.0000000729
.2	.00000655
.3	.0000750
.4	.000354
.5	.000977
.6	.00179
.7	.00222
.8	.00168
.9	.000478
1.0	0.0

- The complete likelihood function is graphed in Figure 1.
- Although each value of $L(\pi|\text{data})$ is a notional probability, the function $L(\pi|\text{data})$ is not a probability or density function — it does not enclose an area of 1.
- The probability of obtaining the sample of data that we have in hand, $HHTHHHTTHH$, is small regardless of the true value of π.
  - This is usually the case: *Any specific* sample result — including the one that is realized — will have low probability.
- Nevertheless, the likelihood contains useful information about the unknown parameter π.
- For example, π *cannot* be 0 or 1, and is 'unlikely' to be close to 0 or 1.

▶ Reversing this reasoning, the value of π that is most supported by the data is the one for which the likelihood is largest.
- This value is the *maximum-likelihood estimate (MLE)*, denoted $\widehat{\pi}$.
- Here, $\widehat{\pi} = .7$, which is the sample proportion of heads, 7/10.

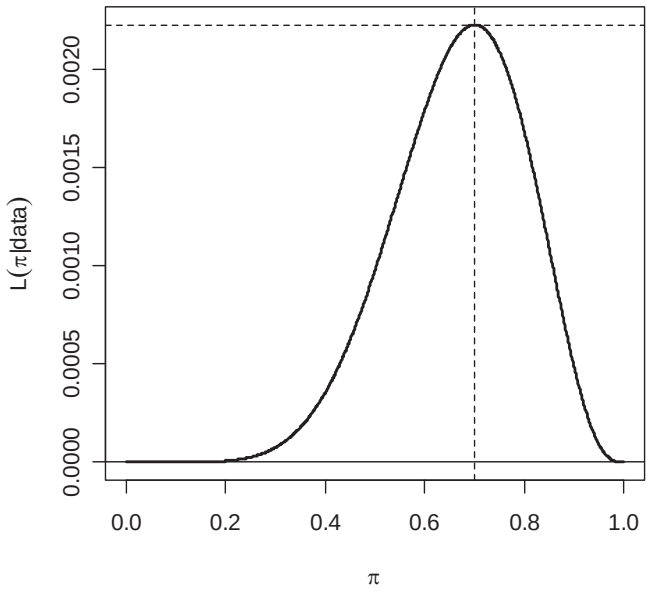

Figure 1. Likelihood of observing 7 heads and 3 tails in a particular sequence for different values of the probability of observing a head, π.

- More generally, for n independent flips of the coin, producing a particular sequence that includes x heads and $n - x$ tails,

$$L(\pi|\text{data}) = \Pr(\text{data}|\pi) = \pi^x(1-\pi)^{n-x}$$

  - We want the value of π that maximizes $L(\pi|\text{data})$, which we often abbreviate $L(\pi)$.
  - It is simpler — and equivalent — to find the value of π that maximizes the log of the likelihood

$$\log_e L(\pi) = x \log_e \pi + (n-x) \log_e (1-\pi)$$

  - Differentiating $\log_e L(\pi)$ with respect to π produces

$$\begin{aligned} \frac{d \log_e L(\pi)}{d\pi} &= \frac{x}{\pi} + (n-x)\frac{1}{1-\pi}(-1) \\ &= \frac{x}{\pi} - \frac{n-x}{1-\pi} \end{aligned}$$

  - Setting the derivative to 0 and solving produces the MLE which, as before, is the sample proportion x/n.
  - The maximum-likelihood *estimator* is $\widehat{\pi} = X/n$.

2.2 Properties of Maximum-Likelihood Estimators

Under very broad conditions, maximum-likelihood estimators have the following general properties:

- Maximum-likelihood estimators are consistent.
- They are asymptotically unbiased, although they may be biased in finite samples.
- They are asymptotically efficient — no asymptotically unbiased estimator has a smaller asymptotic variance.
- They are asymptotically normally distributed.
- If there is a sufficient statistic for a parameter, then the maximum-likelihood estimator of the parameter is a function of a sufficient statistic.
  - A sufficient statistic is a statistic that exhausts all of the information in the sample about the parameter of interest.

- The asymptotic sampling variance of the MLE $\widehat{\alpha}$ of a parameter α can be obtained from the second derivative of the log-likelihood:

$$\mathcal{V}(\widehat{\alpha}) = \frac{1}{-E\left[\dfrac{d^2 \log_e L(\alpha)}{d\alpha^2}\right]}$$

  - The denominator of $\mathcal{V}(\widehat{\alpha})$ is called the *expected* or *Fisher information*

$$\mathcal{I}(\alpha) \equiv -E\left[\frac{d^2 \log_e L(\alpha)}{d\alpha^2}\right]$$

  - In practice, we substitute the MLE $\widehat{\alpha}$ into the equation for $\mathcal{V}(\widehat{\alpha})$ to obtain an *estimate* of the asymptotic sampling variance, $\widehat{\mathcal{V}(\widehat{\alpha})}$.

- $L(\widehat{\alpha})$ is the value of the likelihood function at the MLE $\widehat{\alpha}$, while $L(\alpha)$ is the likelihood for the true (but generally unknown) parameter α.
  - The *log likelihood-ratio statistic*

$$G^2 \equiv -2\log_e \frac{L(\alpha)}{L(\widehat{\alpha})} = 2[\log_e L(\widehat{\alpha}) - \log_e L(\alpha)]$$

    follows an asymptotic chisquare distribution with one degree of freedom.
    - Because, by definition, the MLE maximizes the likelihood for our particular sample, the value of the likelihood at the true parameter value α is generally smaller than at the MLE $\widehat{\alpha}$ (unless, by good fortune, $\widehat{\alpha}$ and α happen to coincide).

2.3 Statistical Inference: Wald, Likelihood-Ratio, and Score Tests

These properties of maximum-likelihood estimators lead directly to three common and general procedures for testing the statistical hypothesis H_0: $\alpha = \alpha_0$.

1. *Wald Test:* Relying on the asymptotic normality of the MLE $\widehat{\alpha}$, we calculate the test statistic

$$Z_0 \equiv \frac{\widehat{\alpha} - \alpha_0}{\sqrt{\widehat{\mathcal{V}(\widehat{\alpha})}}}$$

   which is asymptotically distributed as $N(0, 1)$ under H_0.
2. *Likelihood-Ratio Test:* Employing the log likelihood ratio, the test statistic

$$G_0^2 \equiv -2\log_e \frac{L(\alpha_0)}{L(\widehat{\alpha})} = 2[\log_e L(\widehat{\alpha}) - \log_e L(\alpha_0)]$$

   is asymptotically distributed as χ_1^2 under H_0.

3. *Score Test:* The 'score' is the slope of the log-likelihood at a particular value of α, that is, $S(\alpha) \equiv d \log_e L(\alpha)/d\alpha$.
   - At the MLE, the score is 0: $S(\widehat{\alpha}) = 0$. It can be shown that the *score statistic*
   $$S_0 \equiv \frac{S(\alpha_0)}{\sqrt{\mathcal{I}(\alpha_0)}}$$
   is asymptotically distributed as $N(0, 1)$ under H_0.

- Unless the log-likelihood is quadratic, the three test statistics can produce somewhat different results in specific samples, although the three tests are asymptotically equivalent.
- In certain contexts, the score test has the practical advantage of not requiring the computation of the MLE $\widehat{\alpha}$ (because S_0 depends only on the null value α_0, which is specified in H_0).
- The Wald and likelihood-ratio tests can be 'turned around' to produce confidence intervals for α.

- Figure 2 compares the three test statistics.
- Maximum-likelihood estimation and the Wald, likelihood-ratio, and score tests, extend straightforwardly to simultaneous estimation of several parameters.
- When the log-likelihood function is relatively flat at its maximum, as opposed to sharply peaked, there is little information in the data about the parameter, and the MLE will be an imprecise estimator: See Figure 3.
- Maximum-likelihood estimation and the various tests extend straightforwardly to the simultaneous estimation of several parameters.

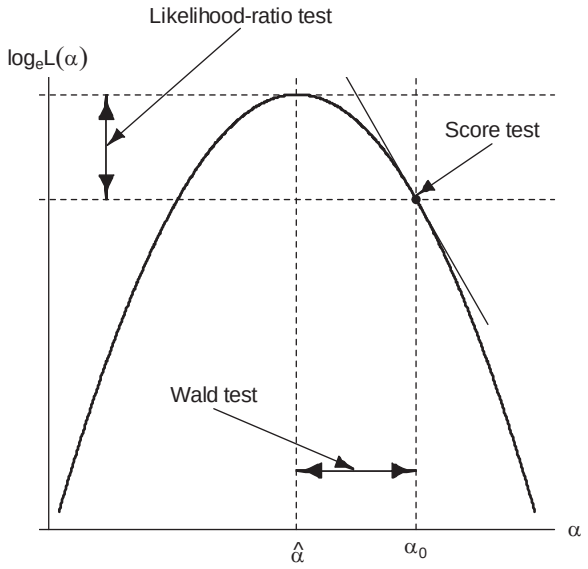

Figure 2. Likelihood-ratio, Wald, and score tests.

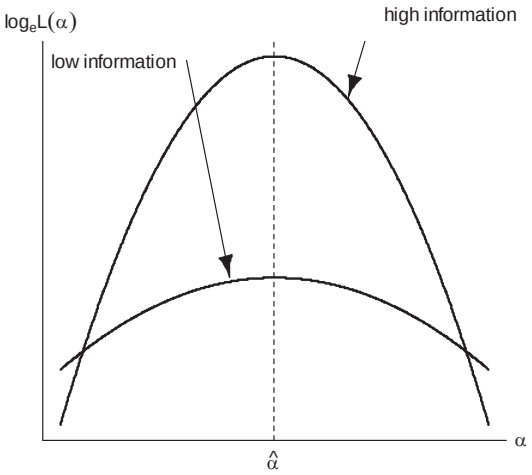

Figure 3. Two imagined log likelihoods: one strongly peaked, providing high information about the the parameter α; and the other flat, providing low information about α.

2.4 Example (Continued)

- Recall the log-likelihood from the coin-flipping example:

$$\log_e L(\pi) = x \log_e \pi + (n - x) \log_e (1 - \pi)$$

where x is the number of heads in n independent flips of a coin.

- The derivative of the log-likelihood is (again, recall)

$$\frac{d \log_e L(\pi)}{d\pi} = \frac{x}{\pi} - \frac{n - x}{1 - \pi}$$

- To get the Fisher information, we need the second derivative of the log likelihood, which is

$$\frac{d^2 \log_e L(\pi)}{d\pi^2} = -\frac{x}{\pi^2} - \frac{n - x}{(1 - \pi)^2}$$

- Noting that the expected number of heads is $E(x) = n\pi$, the Fisher information is

$$\begin{aligned} \mathcal{I}(\pi) &= -E\left[\frac{d^2 \log_e L(\pi)}{d\pi^2}\right] \\ &= \frac{n\pi}{\pi^2} + \frac{n - n\pi}{(1 - \pi)^2} \\ &= \frac{n}{\pi(1 - \pi)} \end{aligned}$$

- Then the asymptotic variance of the sample proportion $\widehat{\pi}$ is

$$\mathcal{V}(\widehat{\pi}) = \frac{1}{\mathcal{I}(\pi)} = \frac{\pi(1-\pi)}{n}$$

and the estimated asymptotic standard error of the sample proportion is

$$\mathsf{SE}(\widehat{\pi}) = \sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}}$$

which is the familiar result for the standard error of a proportion.

2.4.1 Wald Test

▶ Suppose that we want to test the hypothesis that the coin is fair, H_0: $\pi = .5$, and that our sample has $x = 7$ heads in $n = 10$ flips, so that $\widehat{\pi} = 7/10 = .7$.

- Then

$$\mathsf{SE}(\widehat{\pi}) = \sqrt{\frac{.7(1-.3)}{10}} = 0.1449$$

- The Wald test statistic is

$$Z_0 = \frac{\widehat{\pi} - \pi_0}{\mathsf{SE}(\widehat{\pi})} = \frac{.7 - .5}{0.1449} = 1.380$$

for which the two-sided p-value (from the standard-normal distribution) is $p = .167$.

2.4.2 Likelihood-Ratio Test

- The likelihood is
$$L(\pi) = \pi^x(1-\pi)^{n-x}$$
  - So the log-likelihood at the MLE and at the hypothesized value of π are, respectively,
$$\log_e L(\widehat{\pi}) = \log_e[.7^7(1-.7)^3] = -6.1086$$
$$\log_e L(\pi_0) = \log_e[.5^7(1-.5)^3] = -6.9315$$
  - The likelihood-ratio (LR) test statistic is
$$\begin{aligned} G_0^2 &= 2[\log_e L(\widehat{\pi}) - \log_e L(\pi_0)] \\ &= 2(-6.1086 - -6.9315) = 1.646 \end{aligned}$$
  - From the χ^2 distribution with one degree of freedom, the p-value for this test statistic is $p = .199$.
  - Converting the chisquare test statistic to a standard-normal test statistic produces $Z_0 = \sqrt{1.646} = 1.283$, which differs somewhat from the Wald test statistic.

2.4.3 Score Test

- We need the score at the null value
$$\begin{aligned} S(\pi_0) &= \left.\frac{d\log_e L(\pi)}{d\pi}\right|_{\pi=\pi_0} = \frac{x}{\pi_0} - \frac{n-x}{1-\pi_0} \\ &= \frac{7}{.5} - \frac{10-7}{1-.5} = 8.0 \end{aligned}$$
  and the Fisher information at the null value
$$\begin{aligned} \mathcal{I}(\pi_0) &\equiv \frac{n}{\pi_0(1-\pi_0)} \\ &= \frac{10}{.5(1-.5)} = 40.0 \end{aligned}$$

- Then the score statistic is

$$S_0 = \frac{S(\pi_0)}{\sqrt{\mathcal{I}(\pi_0)}} = \frac{8.0}{\sqrt{40.0}} = 1.265$$

  - From the standard-normal distribution, the two-sided p-value is $p = .206$.
  - In this case, therefore, the score statistic is closer to the likelihood-ratio statistic than the Wald statistic is.

2.4.4 An Exact Test

- In this simple setting, an exact binomial test is available (as you likely learned in basic statistics).
- The distribution of the sample proportion $\widehat{\pi}$ and of the number of heads X is

$$\Pr(\widehat{\pi} = \frac{x}{n}) = \Pr(X = x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x} \text{ for } x = 0, 1, 2, \ldots, n$$

- The null distribution for the current example follows from setting $\pi = \pi_0 = .5$ and $n = 10$, and is given in the following table:

x	$\widehat{\pi} = \frac{x}{10}$	$\Pr(X = x)$
0	0.0	.000977
1	.1	.009766
2	.2	.043945
3	.3	.117188
4	.4	.205078
5	.5	.246094
6	.6	.205078
7	.7	.117188
8	.8	.043945
9	.9	.009766
10	1.0	.000977

- Having obtained $x = 7$ heads (i.e., $\widehat{\pi} = .7$), the two-sided p-value for the hypothesis is $\Pr(X \leq 3 \text{ or } X \geq 7) = .3437$, quite different from the results produced by the three asymptotic tests. Of course, $n = 10$ is a very small sample.

2.4.5 Confidence Intervals

- The Wald and LR statistics can be inverted to produce confidence intervals.
- The Wald interval is particularly simple; e.g., for a 95-percent confidence interval:

$$\begin{aligned}\pi &= \widehat{\pi} \pm 1.96 \times \mathsf{SE}(\widehat{\pi}) \\ &= .7 \pm 1.96 \times 0.1449 \\ &= .7 \pm .284 \\ &= (.416, .984)\end{aligned}$$

- The confidence interval based on the LR statistic includes all values of π that cannot be rejected when tested as hypotheses given the observed value of $\widehat{\pi}$.
  - To be statistically significant at the .05 level, a chisquare statistic with one degree of freedom must be at least $1.96^2 = 3.84$.

  - That is, the confidence interval includes all values of π for which

$$\begin{aligned}2[\log_e L(\widehat{\pi}) - \log_e L(\pi)] &\leq 3.84 \\ \log_e L(\widehat{\pi}) - \log_e L(\pi) &\leq 1.92\end{aligned}$$

  - This is illustrated, for the example, in Figure 4:
  - The 95 percent confidence interval for π runs from .394 to .915.
- With only $n = 10$ observations, neither the Wald interval nor the LR interval can be trusted.

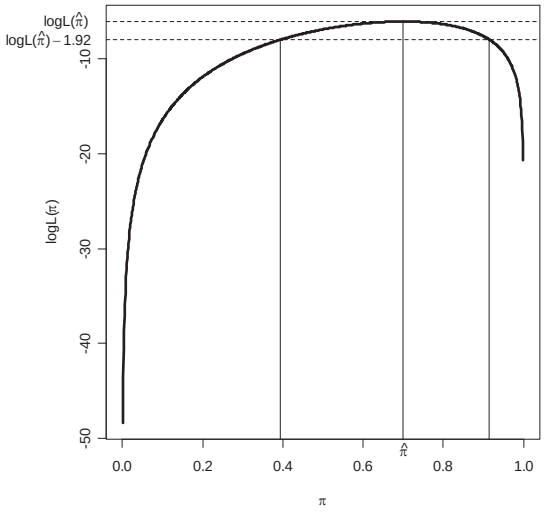

Figure 4. Likelihood-ratio-based 95 percent confidence interval for π when $n = 10$ and $\widehat{\pi} = .7$.

3. Introduction to Bayesian Inference

- Bayesian inference is an alternative to classical statistical inference based on null-hypothesis significance tests and confidence intervals.

3.1 Bayes' Theorem

- The conditional probability of an event A given that another event B is known to have occurred is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

- Likewise, the conditional probability of B given A is

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

- Solving for the joint probability of A and B produces

$$\Pr(A \cap B) = \Pr(B|A)\Pr(A)$$

- Substituting this result into the equation for $\Pr(A|B)$ yields *Bayes' Theorem* (named for Thomas Bayes, 1701–1761):

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

- Bayesian statistical inference is based on the following interpretation of Bayes' Theorem:
  - Let A represent some uncertain proposition whose truth or falsity we wish to establish—for example, the proposition that a parameter is equal to a particular value.
  - Let B represent observed data that are relevant to the truth of the proposition.
    - The unconditional probability $\Pr(A)$, called the *prior probability* of A, is our strength of belief in the truth of A prior to collecting data.
    - $\Pr(B|A)$ as the probability of obtaining the observed data assuming the truth of A—that is, the *likelihood* of the data given A (in the sense of the preceding section).

    - The *unconditional* probability of the data B is

$$\Pr(B) = \Pr(B|A)\Pr(A) + \Pr(B|\overline{A})\Pr(\overline{A})$$

    - Then $\Pr(A|B)$ is called the *posterior probability* of A and represents our revised strength of belief in A in light of the data B.

- Bayesian inference is therefore a rational procedure for updating one's beliefs on the basis of evidence.
- This *subjectivist* interpretation of probabilities contrasts with the *frequentist* interpretation of probabilities as long-run proportions.

3.2 Preliminary Example

- You are given a gift of two "biased" coins, one of which produces heads with probability $\Pr(H) = .3$ and the other with $\Pr(H) = .8$.
- Each of these coins comes in a box marked with its bias, but you carelessly misplace the boxes and put the coins in a drawer; a year later, you do not remember which coin is which.
- To try to distinguish the coins, you pick one arbitrarily and flip it 10 times, obtaining the data $HHTHHHTTHH$—that is, a particular sequence of 7 heads and 3 tails.
- Let A represent the event that the selected coin has $\Pr(H) = .3$; then $\overline{A}$ is the event that the coin has $\Pr(H) = .8$.
- Under these circumstances, it seems reasonable to take as prior probabilities $\Pr(A) = \Pr(\overline{A}) = .5$.

- The likelihood of the data under A and $\overline{A}$ is

$$\Pr(B|A) = .3^7(1-.3)^3 = .0000750$$
$$\Pr(B|\overline{A}) = .8^7(1-.8)^3 = .0016777$$

  - As is typically the case, the likelihood of the observed data is small in both cases, but the data are much more likely under $\overline{A}$ than under A.
- Using Bayes' Theorem, you find the posterior probabilities

$$\Pr(A|B) = \frac{.0000750 \times .5}{.0000750 \times .5 + .0016777 \times .5} = .0428$$
$$\Pr(\overline{A}|B) = \frac{.0016777 \times .5}{.0000750 \times .5 + .0016777 \times .5} = .9572$$

  suggesting that it is much more probable that the selected coin has $\Pr(H) = .8$ than $\Pr(H) = .2$.

3.3 Extending Bayes Theorem

- Bayes' Theorem extends readily to situations in which there are more than two hypotheses A and $\overline{A}$:
- Let the various hypotheses be represented by $H_1, H_2, \ldots, H_k$, with prior probabilities $\Pr(H_i)$, $i = 1, \ldots, k$ that sum to 1; and let D represent the observed data, with likelihood $\Pr(D|H_i)$ under hypothesis H_i.
  - Then the posterior probability of hypothesis H_i is
$$\Pr(H_i|D) = \frac{\Pr(D|H_i)\Pr(H_i)}{\sum_{j=1}^{k}\Pr(D|H_j)\Pr(H_j)}$$
  - The denominator insures that the posterior probabilities for the various hypotheses sum to 1.

  - It is sometimes convenient to omit this normalization, simply noting that
$$\Pr(H_i|D) \propto \Pr(D|H_i)\Pr(H_i)$$
    that is, that the posterior probability of a hypothesis is proportional to the product of the likelihood under the hypothesis and its prior probability.
- Bayes' Theorem is also applicable to random variables:
  - Let α represent a parameter of interest, with prior probability distribution or density $p(\alpha)$, and let $L(\alpha) = p(D|\alpha)$ represent the likelihood function for the parameter α.

- Then
$$p(\alpha|D) = \frac{L(\alpha)p(\alpha)}{\sum_{\text{all } \alpha'} L(\alpha')p(\alpha')}$$
when the parameter α is discrete, or
$$p(\alpha|D) = \frac{L(\alpha)p(\alpha)}{\int L(\alpha')p(\alpha')d\alpha'}$$
when, as is more common, α is continuous.
- In either case,
$$p(\alpha|D) \propto L(\alpha)p(\alpha)$$
- The posterior distribution or density is proportional to the product of the likelihood and the prior distribution or density.

▶ We require a prior distribution $p(\alpha)$ over the possible values of the parameter α (the *parameter space*) to set the machinery of Bayesian inference in motion.

▶ In contrast to classical statistics, we treat the parameter α as a *random variable* rather than as an unknown *constant*.

3.4 Conjugate Priors

▶ The mathematics of Bayesian inference is especially simple when the prior distribution is selected so that the likelihood and prior combine to produce a posterior distribution that is in the same family as the prior.
- In this case, we say that the prior distribution is a *conjugate prior*.

▶ At one time, Bayesian inference was only practical when conjugate priors were employed, limiting its scope of application.
- Advances in computer software and hardware, make it practical to evaluate mathematically intractable posterior distributions by simulated random sampling.
- Such *Markov-chain Monte-Carlo* ("*MCMC*") methods have produced a flowering of Bayesian applied statistics.

3.5 An Example of Bayesian Inference

- Continuing the previous example, suppose more realistically that you are given a coin and wish to estimate the probability π that the coin turns up heads, but cannot restrict π in advance to a small set of discrete values
  - π could, in principle, be any number between 0 and 1.
- To estimate π, you plan to gather data by independently flipping the coin 10 times.
  - We know from our previous work that the likelihood is
    $$L(\pi) = \pi^h(1-\pi)^{10-h}$$
    where h is the observed number of heads.
  - You conduct the experiment, obtaining the data $HHTHHHTTHH$, and thus $h = 7$.

- The conjugate prior for the this Bernouli likelihood is the beta distribution
  $$p(\pi) = \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,b)} \text{ for } 0 \leq \pi \leq 1 \text{ and } a, b \geq 1$$
  where $B(a,b)$ is the beta function.
  - Some beta distributions are shown in Figure 5
- When you multiply the beta prior by the likelihood, you get a posterior density of the form
  $$p(\pi|D) \propto \pi^{h+a-1}(1-\pi)^{10-h+b-1} = \pi^{6+a}(1-\pi)^{2+b}$$
  - This is a beta distribution with parameters $h + a - 1 = 6 + a$ and $10 - h + b - 1 = 2 + b$.
  - Put another way, the prior in effect adds a heads and b tails to the likelihood.
- How should you select a and b?
- One approach would be to reflect your subjective assessment of the likely value of π.

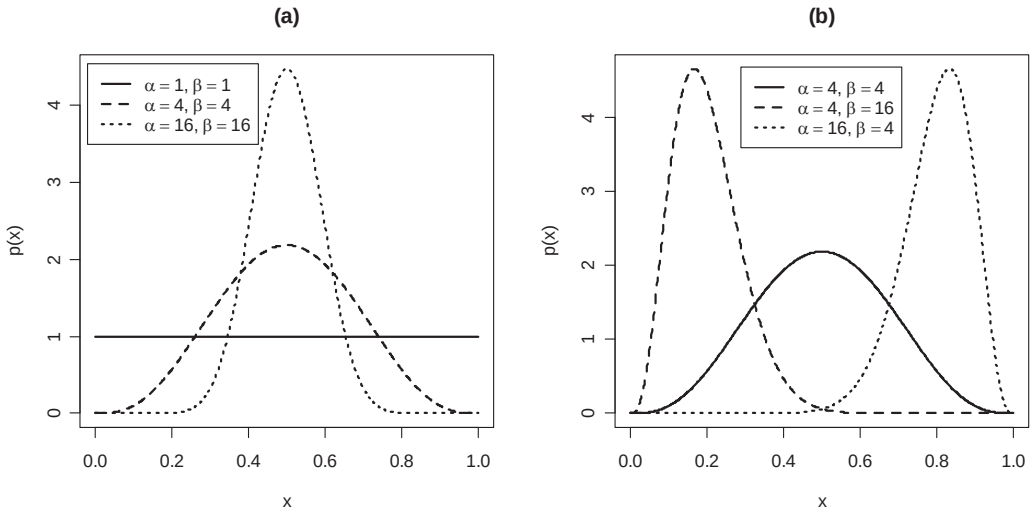

Figure 5. Beta distributions for several combinations of values of the parameters α and β. As is apparent in panel (a), the beta distribution reduces to the rectangular distribution when $\alpha = \beta = 1$.

- For example, you might examine the coin and note that it seems to be reasonably well balanced, suggesting that π is probably close to $.5$.
- Picking $a = b = 16$ would in effect confine your estimate of π to the range between $.3$ and $.7$.
- If you are uncomfortable with this restriction, then you could select smaller values of a and b.
- In the extreme, $a = b = 1$, and all values of π are equally likely—a so-called *flat prior distribution*, reflecting complete ignorance about the value of π.

▶ Figure 6 shows the posterior distribution for π under these two priors.

- Under the flat prior, the posterior is proportional to the likelihood, and therefore if you take the mode of the posterior as your estimate of π, you get the MLE, $\widehat{\pi} = .7$.
- The *informative prior* $a = b = 16$, in contrast, has a mode at $\pi \approx .55$, which is much closer to the mode of the prior distribution, $\pi = .5$.

- As the sample size grows, the likelihood comes to dominate the posterior distribution, and the influence of the prior distribution fades.
  - In the example, if the coin is flipped n times, then the posterior distribution takes the form
$$p(\pi|D) \propto \pi^{h+a-1}(1-\pi)^{n-h+b-1}$$
  - It is intuitively sensible that your prior beliefs should carry greater weight when the sample is small than when it is large.

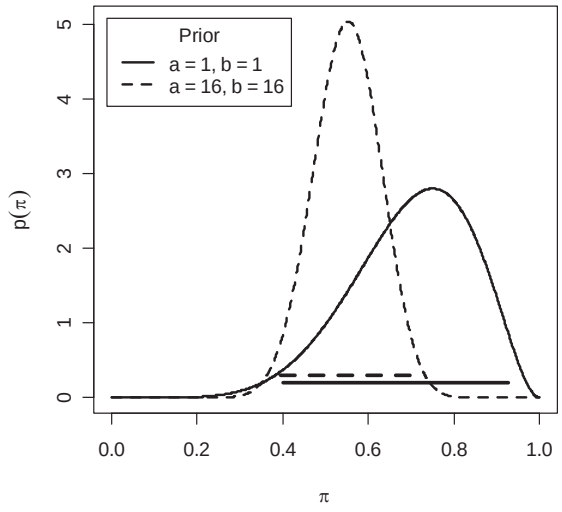

Figure 6. Posterior distributions and 95% posterior intervals for the probability of a head π and data 7 heads in 10 flips under two prior distributions: the flat beta prior with $a = 1, b = 1$, and the informative beta prior with $a = 16, b = 16$.

3.6 Bayesian Interval Estimates

- As in classical inference, it is desirable not only to provide a point estimate of a parameter but also to express uncertainty in the estimate.
  - The posterior distribution of the parameter expresses statistical uncertainty in a direct form.
  - From the posterior distribution, one can compute various kinds of Bayesian interval estimates, which are analogous to classical confidence intervals.
- A very simple choice is the *central posterior interval*:
  - The 100α percent central posterior interval runs from the $(1 - \alpha)/2$ to the $(1 + \alpha)/2$ quantile of the posterior distribution.
  - Unlike a classical confidence interval, a Bayesian posterior interval has a simple interpretation as a probability statement: The probability is $.95$ that the parameter is in the 95-percent posterior interval.

- 95 percent central posterior intervals for the example are shown for the two posterior distributions in Figure 6.

3.7 Bayesian Inference for Several Parameters

- Bayesian inference extends straightforwardly to the simultaneous estimation of several parameters $\boldsymbol{\alpha} \equiv [\alpha_1, \alpha_2, \ldots, \alpha_k]'$.
- In this case, it is necessary to specify a *joint prior distribution* for the parameters, $p(\boldsymbol{\alpha})$, along with the *joint likelihood*, $L(\boldsymbol{\alpha})$.
- Then, as in the case of one parameter, the *joint posterior distribution* is proportional to the product of the prior distribution and the likelihood:

$$p(\boldsymbol{\alpha}|D) \propto p(\boldsymbol{\alpha})L(\boldsymbol{\alpha})$$

- Inference typically focusses on the *marginal posterior distribution* of each parameter, $p(\alpha_i|D)$.

4. Logit and Probit Models for Dichotomous Data

- To understand why logit and probit models for qualitative data are required, let us begin by examining a representative problem, attempting to apply linear regression to it:
  - In September of 1988, 15 years after the coup of 1973, the people of Chile voted in a plebiscite to decide the future of the military government. A 'yes' vote would represent eight more years of military rule; a 'no' vote would return the country to civilian government. The no side won the plebiscite, by a clear if not overwhelming margin.

- Six months before the plebiscite, FLACSO/Chile conducted a national survey of 2,700 randomly selected Chilean voters.
  - Of these individuals, 868 said that they were planning to vote yes, and 889 said that they were planning to vote no.
  - Of the remainder, 558 said that they were undecided, 187 said that they planned to abstain, and 168 did not answer the question.
  - I will look only at those who expressed a preference.
- Figure 7 plots voting intention against a measure of support for the status quo.
  - Voting intention appears as a dummy variable, coded 1 for yes, 0 for no.
  - Support for the status quo is a scale formed from a number of questions about political, social, and economic policies: High scores represent general support for the policies of the miliary regime.

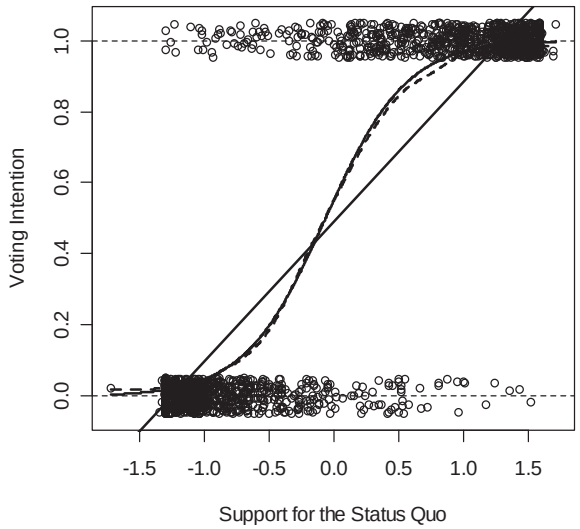

Figure 7. The Chilean plebiscite data: The solid straight line is a linear least-squares fit; the solid curved line is a logistic-regression fit; and the broken line is from a nonparametric kernel regression with a span of .4.The individual observations are all at 0 or 1 and are vertically jittered.

- Does it make sense to think of regression as a conditional average when the response variable is dichotomous?
  - An average between 0 and 1 represents a 'score' for the dummy response variable that cannot be realized by any individual.
  - In the population, the conditional average $E(Y|x_i)$ is the proportion of 1's among those individuals who share the value x_i for the explanatory variable — the conditional probability π_i of sampling a 'yes' in this group:

    $$\pi_i \equiv \Pr(Y_i) \equiv \Pr(Y = 1|X = x_i)$$

    and thus,

    $$E(Y|x_i) = \pi_i(1) + (1 - \pi_i)(0) = \pi_i$$

- If X is discrete, then in a sample we can calculate the conditional proportion for Y at each value of X.
  - The collection of these conditional proportions represents the sample nonparametric regression of the dichotomous Y on X.
  - In the present example, X is continuous, but we can nevertheless resort to strategies such as local averaging, as illustrated in the figure.

4.1 The Linear-Probability Model

- Although non-parametric regression works here, it would be useful to capture the dependency of Y on X as a simple function, particularly when there are several explanatory variables.
- Let us first try linear regression with the usual assumptions:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

  where $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$, and ε_i and ε_j are independent for $i \neq j$.
  - If X is random, then we assume that it is independent of ε.
- Under this model, $E(Y_i) = \alpha + \beta X_i$, and so

$$\pi_i = \alpha + \beta X_i$$

  - For this reason, the linear-regression model applied to a dummy response variable is called the *linear probability model*.
- This model is untenable, but its failure points the way towards more adequate specifications:

- *Non-normality:* Because Y_i can take on only the values of 0 and 1, the error ε_i is dichotomous as well — not normally distributed:
  - If $Y_i = 1$, which occurs with probability π_i, then

$$\begin{aligned} \varepsilon_i &= 1 - E(Y_i) \\ &= 1 - (\alpha + \beta X_i) \\ &= 1 - \pi_i \end{aligned}$$

  - Alternatively, if $Y_i = 0$, which occurs with probability $1 - \pi_i$, then

$$\begin{aligned} \varepsilon_i &= 0 - E(Y_i) \\ &= 0 - (\alpha + \beta X_i) \\ &= 0 - \pi_i \\ &= -\pi_i \end{aligned}$$

  - Because of the central-limit theorem, however, the assumption of normality is not critical to least-squares estimation of the normal-probability model.

- *Non-constant error variance:* If the assumption of linearity holds over the range of the data, then $E(\varepsilon_i) = 0$.
  - Using the relations just noted,
$$\begin{aligned} V(\varepsilon_i) &= \pi_i(1-\pi_i)^2 + (1-\pi_i)(-\pi_i)^2 \\ &= \pi_i(1-\pi_i) \end{aligned}$$
  - The heteroscedasticity of the errors bodes ill for ordinary-least-squares estimation of the linear probability model, but only if the probabilities π_i get close to 0 or 1.
- *Nonlinearity:* Most seriously, the assumption that $E(\varepsilon_i) = 0$ — that is, the assumption of linearity — is only tenable over a limited range of X-values.
  - If the range of the X's is sufficiently broad, then the linear specification cannot confine π to the unit interval $[0, 1]$.
  - It makes no sense, of course, to interpret a number outside of the unit interval as a probability.

  - This difficulty is illustrated in the plot of the Chilean plebiscite data, in which the least-squares line produces fitted probabilities below 0 at low levels and above 1 at high levels of support for the status-quo.

▶ Dummy *regressor* variables do not cause comparable difficulties because the general linear model makes no distributional assumptions about the X's.

▶ Nevertheless, if π doesn't get too close to 0 or 1, the linear-probability model estimated by least-squares frequently provides results similar to those produced by more generally adequate methods.

▶ One solution — though not a good one — is simply to constrain π to the unit interval:
$$\pi = \begin{cases} 0 & \text{for } 0 > \alpha + \beta X \\ \alpha + \beta X & \text{for } 0 \leq \alpha + \beta X \leq 1 \\ 1 & \text{for } \alpha + \beta X > 1 \end{cases}$$

- The *constrained linear-probability* model fit to the Chilean plebiscite data by maximum likelihood is shown in Figure 8. Although it cannot be dismissed on logical grounds, this model has certain unattractive features:
  - *Instability:* The critical issue in estimating the linear-probability model is identifying the X-values at which π reaches 0 and 1, since the line $\pi = \alpha + \beta X$ is determined by these two points. As a consequence, estimation of the model is inherently unstable.
  - *Impracticality:* It is much more difficult to estimate the constrained linear-probability model when there are several X's.
  - *Unreasonableness:* Most fundamentally, the abrupt changes in slope at $\pi = 0$ and $\pi = 1$ are unreasonable. A smoother relationship between π and X, is more generally sensible.

4.2 Transformations of π: Logit and Probit Models

- To insure that π stays between 0 and 1, we require a positive monotone (i.e., non-decreasing) function that maps the 'linear predictor' $\eta = \alpha + \beta X$ into the unit interval.
  - A transformation of this type will retain the fundamentally linear structure of the model while avoiding probabilities below 0 or above 1.
  - Any cumulative probability distribution function meets this requirement:
    $$\pi_i = P(\eta_i) = P(\alpha + \beta X_i)$$
    where the CDF $P(\cdot)$ is selected in advance, and α and β are then parameters to be estimated.
  - If we choose $P(\cdot)$ as the cumulative rectangular distribution then we obtain the constrained linear-probability model.
  - An *a priori* reasonable $P(\cdot)$ should be both smooth and symmetric, and should approach $\pi = 0$ and $\pi = 1$ as asymptotes.

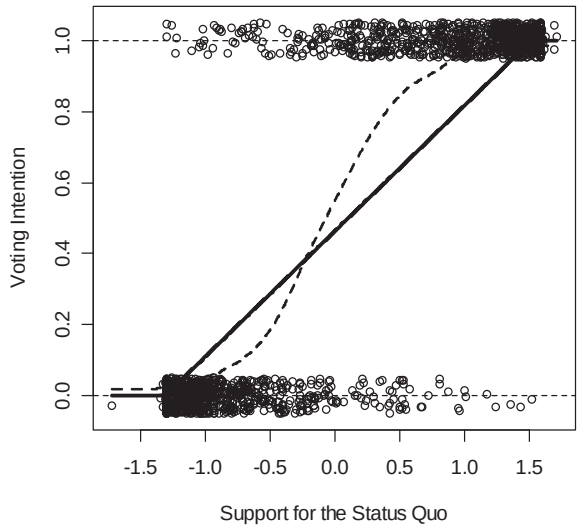

Figure 8. The solid line shows the constrained linear-probability model fit by maximum likelihood to the Chilean plebiscite data; the broken line is for a nonparametric kernel regression.

- Moreover, it is advantageous if $P(\cdot)$ is strictly increasing, permitting us to rewrite the model as
$$P^{-1}(\pi_i) = \eta_i = \alpha + \beta X_i$$
where $P^{-1}(\cdot)$ is the inverse of the CDF $P(\cdot)$, i.e., the quantile function.
  - Thus, we have a linear model for a transformation of π, or — equivalently — a nonlinear model for π itself.

▶ The transformation $P(\cdot)$ is often chosen as the CDF of the unit-normal distribution
$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}Z^2} dZ$$
or, even more commonly, of the *logistic distribution*
$$\Lambda(z) = \frac{1}{1 + e^{-z}}$$
where $\pi \approx 3.141$ and $e \approx 2.718$ are the familiar mathematical constants.

- Using the normal distribution $\Phi(\cdot)$ yields the *linear probit model*:

$$\begin{aligned}\pi_i &= \Phi(\alpha + \beta X_i) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta X_i} e^{-\frac{1}{2}Z^2} dZ\end{aligned}$$

- Using the logistic distribution $\Lambda(\cdot)$ produces the *linear logistic-regression* or *linear logit model*:

$$\begin{aligned}\pi_i &= \Lambda(\alpha + \beta X_i) \\ &= \frac{1}{1 + e^{-(\alpha+\beta X_i)}}\end{aligned}$$

- Once their variances are equated, the logit and probit transformations are so similar that it is not possible in practice to distinguish between them, as is apparent in Figure 9.
- Both functions are nearly linear between about $\pi = .2$ and $\pi = .8$. This is why the linear probability model produces results similar to the logit and probit models, except when there are extreme values of π_i.

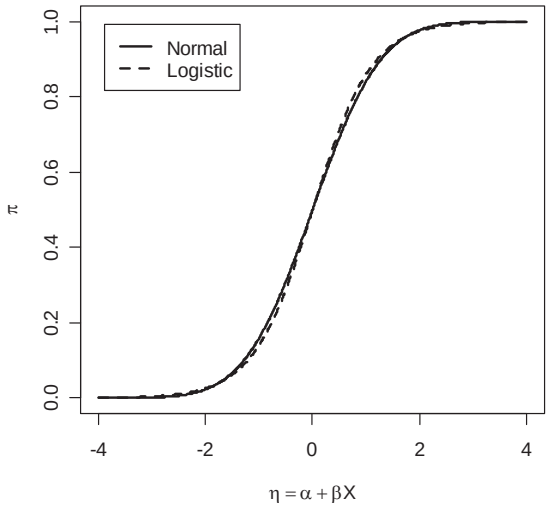

Figure 9. The normal and logistic cumulative distribution functions (as a function of the linear predictor and with variances equated).

- Despite their similarity, there are two practical advantages of the logit model:

1. *Simplicity:* The equation of the logistic CDF is very simple, while the normal CDF involves an unevaluated integral.
   - This difference is trivial for dichotomous data, but for polytomous data, where we will require the *multivariate* logistic or normal distribution, the disadvantage of the probit model is more acute.
2. *Interpretability:* The inverse linearizing transformation for the logit model, $\Lambda^{-1}(\pi)$, is directly interpretable as a *log-odds*, while the inverse transformation $\Phi^{-1}(\pi)$ does not have a direct interpretation.
   - Rearranging the equation for the logit model,
   $$\frac{\pi_i}{1-\pi_i} = e^{\alpha+\beta X_i}$$
   - The ratio $\pi_i/(1-\pi_i)$ is the *odds* that $Y_i = 1$, an expression of relative chances familiar to gamblers.

   - Taking the log of both sides of this equation,
   $$\log_e \frac{\pi_i}{1-\pi_i} = \alpha + \beta X_i$$
   - The inverse transformation $\Lambda^{-1}(\pi) = \log_e[\pi/(1-\pi)]$, called the *logit* of π, is therefore the log of the odds that Y is 1 rather than 0.
   - The logit is symmetric around 0, and unbounded both above and below, making the logit a good candidate for the response-variable side of a linear model:

Probability π	Odds $\frac{\pi}{1-\pi}$	Logit $\log_e \frac{\pi}{1-\pi}$
.01	$1/99 = 0.0101$	-4.60
.05	$5/95 = 0.0526$	-2.94
.10	$1/9 = 0.1111$	-2.20
.30	$3/7 = 0.4286$	-0.85
.50	$5/5 = 1$	0.00
.70	$7/3 = 2.333$	0.85
.90	$9/1 = 9$	2.20
.95	$95/5 = 19$	2.94
.99	$99/1 = 99$	4.60

- The logit model is also a multiplicative model for the odds:

$$\begin{aligned}\frac{\pi_i}{1-\pi_i} &= e^{\alpha+\beta X_i} = e^{\alpha}e^{\beta X_i} \\ &= e^{\alpha}\left(e^{\beta}\right)^{X_i}\end{aligned}$$

  - So, increasing X by 1 changes the logit by β and multiplies the odds by e^{β}.
  - For example, if $\beta = 2$, then increasing X by 1 increases the odds by a factor of $e^2 \approx 2.718^2 = 7.389$.
- Still another way of understanding the parameter β in the logit model is to consider the slope of the relationship between π and X.
  - Since this relationship is nonlinear, the slope is not constant; the slope is $\beta\pi(1-\pi)$, and hence is at a maximum when $\pi = 1/2$, where the slope is $\beta/4$:

π	$\beta\pi(1-\pi)$
.01	$\beta \times .0099$
.05	$\beta \times .0475$
.10	$\beta \times .09$
.20	$\beta \times .16$
.50	$\beta \times .25$
.80	$\beta \times .16$
.90	$\beta \times .09$
.95	$\beta \times .0475$
.99	$\beta \times .0099$

  - The slope does not change very much between $\pi = .2$ and $\pi = .8$, reflecting the near linearity of the logistic curve in this range.

- The least-squares line fit to the Chilean plebescite data has the equation

$$\widehat{\pi}_{\text{yes}} = 0.492 + 0.394 \times \text{ Status-Quo}$$

  - This line is a poor summary of the data.

- The logistic-regression model, fit by the method of maximum-likelihood, has the equation

$$\log_e \frac{\widehat{\pi}_{\text{yes}}}{\widehat{\pi}_{\text{ no}}} = 0.215 + 3.21 \times \text{ Status-Quo}$$

  - The logit model produces a much more adequate summary of the data, one that is very close to the nonparametric regression.
  - Increasing support for the status-quo by one unit multiplies the odds of voting yes by $e^{3.21} = 24.8$.
  - Put alternatively, the slope of the relationship between the fitted probability of voting yes and support for the status-quo at $\widehat{\pi}_{\text{yes}} = .5$ is $3.21/4 = 0.80$.

4.3 Logit and Probit Models for Multiple Regression

- To generalize the logit and probit models to several explanatory variables we require a linear predictor that is a function of several regressors.
  - For the logit model,

$$\begin{aligned}\pi_i &= \Lambda(\eta_i) = \Lambda(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}) \\ &= \Lambda(\mathbf{x}_i'\boldsymbol{\beta}) \\ &= \frac{1}{1 + e^{-\mathbf{x}_i'\boldsymbol{\beta}}}\end{aligned}$$

    or, equivalently,

$$\begin{aligned}\log_e \frac{\pi_i}{1-\pi_i} &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} \\ &= \mathbf{x}_i'\boldsymbol{\beta}\end{aligned}$$

  - For the probit model,

$$\pi_i = \Phi(\eta_i) = \Phi(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$$

- The X's in the linear predictor can be as general as in the general linear model, including, for example:

  - quantitative explanatory variables;
  - transformations of quantitative explanatory variables;
  - polynomial regressors formed from quantitative explanatory variables;
  - dummy regressors representing qualitative explanatory variables; and
  - interaction regressors.

- Interpretation of the partial regression coefficients in the general logit model is similar to the interpretation of the slope in the logit simple-regression model, with the additional provision of holding other explanatory variables in the model constant.
  - Expressing the model in terms of odds,

$$\begin{aligned}\frac{\pi_i}{1-\pi_i} &= e^{(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})} \\ &= e^{\alpha}\left(e^{\beta_1}\right)^{X_{i1}} \cdots \left(e^{\beta_k}\right)^{X_{ik}}\end{aligned}$$

  - Thus, e^{β_j} is the multiplicative effect on the odds of increasing X_j by 1, holding the other X's constant.

- Similarly, $\beta_j/4$ is the slope of the logistic regression surface in the direction of X_j at $\pi = .5$.

▶ The general linear logit and probit models can be fit to data by the method of maximum likelihood.

▶ Hypothesis tests and confidence intervals follow from general procedures for statistical inference in maximum-likelihood estimation.

- For an individual coefficient, it is most convenient to test the hypothesis H_0: $\beta_j = \beta_j^{(0)}$ by calculating the Wald statistic

$$Z_0 = \frac{B_j - \beta_j^{(0)}}{\mathsf{SE}(B_j)}$$

where $\mathsf{SE}(B_j)$ is the asymptotic standard error of B_j.

· The test statistic Z_0 follows an asymptotic unit-normal distribution under the null hypothesis.

- Similarly, an asymptotic $100(1-a)$-percent confidence interval for β_j is given by

$$\beta_j = B_j \pm z_{a/2}\mathsf{SE}(B_j)$$

where $z_{a/2}$ is the value from $Z \sim N(0,1)$ with a probability of $a/2$ to the right.

- Wald tests for several coefficients can be formulated from the estimated asymptotic variances and covariances of the coefficients.
- It is also possible to formulate a likelihood-ratio test for the hypothesis that several coefficients are simultaneously zero, H_0: $\beta_1 = \cdots = \beta_q = 0$. We proceed, as in least-squares regression, by fitting two models to the data:

· The full model (model 1)

$$\mathsf{logit}(\pi) = \alpha + \beta_1 X_1 + \cdots + \beta_q X_q + \beta_{q+1} X_{q+1} + \cdots + \beta_k X_k$$

- and the null model (model 0)

$$\begin{aligned}\text{logit}(\pi) &= \alpha + 0X_1 + \cdots + 0X_q + \beta_{q+1}X_{q+1} + \cdots + \beta_k X_k \\ &= \alpha + \beta_{q+1}X_{q+1} + \cdots + \beta_k X_k\end{aligned}$$

- Each model produces a maximized likelihood: L_1 for the full model, L_0 for the null model.
- Because the null model is a specialization of the full model, $L_1 \geq L_0$.
- The generalized likelihood-ratio test statistic for the null hypothesis is

$$G_0^2 = 2(\log_e L_1 - \log_e L_0)$$

- Under the null hypothesis, this test statistic has an asymptotic chisquare distribution with q degrees of freedom.

- A test of the omnibus null hypothesis H_0: $\beta_1 = \cdots = \beta_k = 0$ is obtained by specifying a null model that includes only the constant, $\text{logit}(\pi) = \alpha$.

- The likelihood-ratio test can be inverted to produce confidence intervals for coefficients.
- The likelihood-ratio test is less prone to breaking down than the Wald test.

- An analog to the multiple-correlation coefficient can also be obtained from the log-likelihood.
  - By comparing $\log_e L_0$ for the model containing only the constant with $\log_e L_1$ for the full model, we can measure the degree to which using the explanatory variables improves the predictability of Y.
  - The quantity $G^2 \equiv -2\log_e L$, called the *residual deviance* under the model, is a generalization of the residual sum of squares for a linear model.
  - Thus,

$$R^2 = 1 - \frac{G_1^2}{G_0^2} = 1 - \frac{\log_e L_1}{\log_e L_0}$$

    is analogous to R^2 for a linear model.

5. The Structure of Generalized Linear Models

- A synthesis due to Nelder and Wedderburn, generalized linear models (GLMs) extend the range of application of linear statistical models by accommodating response variables with non-normal conditional distributions.
- Except for the error, the right-hand side of a generalized linear model is essentially the same as for a linear model.

- A generalized linear model consists of three components:

1. A *random component*, specifying the conditional distribution of the response variable, Y_i, given the explanatory variables.
   - Traditionally, the random component is a member of an "exponential family" — the normal (Gaussian), binomial, Poisson, gamma, or inverse-Gaussian families of distributions — but generalized linear models have been extended beyond the exponential families.
   - The Gaussian and binomial distributions are familiar.
   - Poisson distributions are often used in modeling count data. Poisson random variables take on non-negative integer values, $0, 1, 2, \ldots$. Some examples are shown in Figure 10.

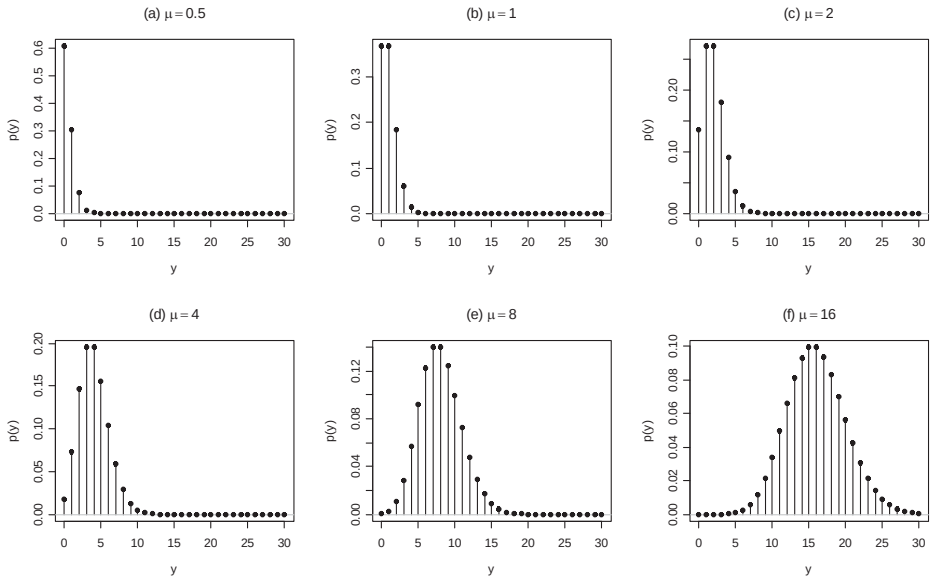

Figure 10. Poisson distributions for various values of the "rate" parameter (mean) μ.

- The gamma and inverse-Gaussian distributions are for positive continuous data; some examples are given in Figure 11.

2. A linear function of the regressors, called the *linear predictor*,

$$\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} = \mathbf{x}_i' \boldsymbol{\beta}$$

on which the expected value μ_i of Y_i depends.

- The X's may include quantitative predictors, but they may also include transformations of predictors, polynomial terms, contrasts generated from factors, interaction regressors, etc.

3. An invertible *link function* $g(\mu_i) = \eta_i$, which transforms the expectation of the response to the linear predictor.

- The inverse of the link function is sometimes called the *mean function*: $g^{-1}(\eta_i) = \mu_i$.

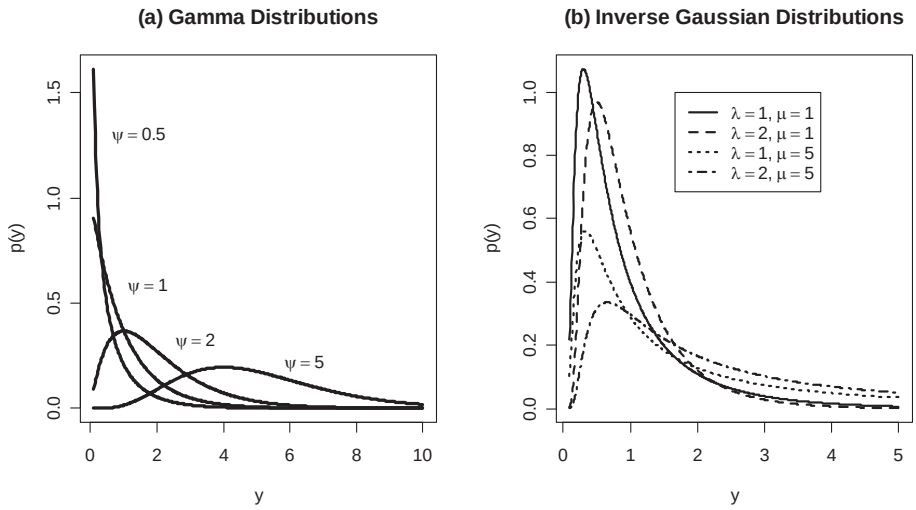

Figure 11. (a) Several gamma distributions for "scale" $\omega = 1$ and various values of the "shape" parameter ψ. (b) Inverse-Gaussian distributions for several combinations of values of the mean μ and "inverse-dispersion" λ.

- Standard link functions and their inverses are shown in the following table:

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
identity	μ_i	η_i
log	$\log_e \mu_i$	e^{η_i}
inverse	μ_i^{-1}	η_i^{-1}
inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
square-root	$\sqrt{\mu_i}$	η_i^2
logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

- The logit, probit, and complementary-log-log links are for *binomial data*, where Y_i represents the observed proportion and μ_i the expected proportion of "successes" in n_i binomial trials — that is, μ_i is the probability of a success.

  - For the probit link, Φ is the standard-normal cumulative distribution function, and Φ^{-1} is the standard-normal quantile function.
  - An important special case is *binary data*, where all of the binomial trials are 1, and therefore all of the observed proportions Y_i are either 0 or 1. This is the case that we examined in the previous session.

▶ Although the logit and probit links are familiar, the log-log and complementary log-log links for binomial data are not.

- These links are compared in Figure 12.
- The log-log or complementary log-log link may be appropriate when the probability of the response as a function of the linear predictor approaches 0 and 1 asymmetrically.

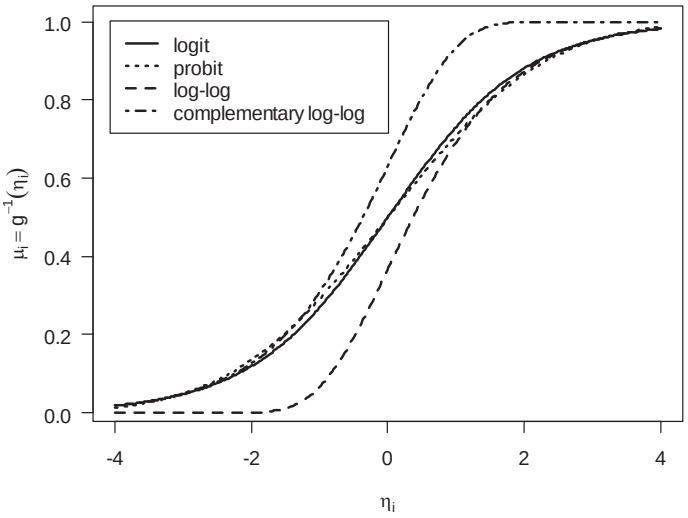

Figure 12. Comparison of logit, probit, and complementary log-log links. The probit link is rescaled to match the variance of the logistic distribution, $\pi^2/3$.

- For distributions in the exponential families, the conditional variance of Y is a function of the mean μ together with a dispersion parameter ϕ (as shown in the table below).
  - For the binomial and Poisson distributions, the dispersion parameter is fixed to 1.
  - For the Gaussian distribution, the dispersion parameter is the usual error variance, which we previously symbolized by σ^2_ε (and which doesn't depend on μ).

Family	*Canonical Link*	*Range of* Y_i	$V(Y_i\|\eta_i)$
Gaussian	identity	$(-\infty, +\infty)$	ϕ
binomial	logit	$\frac{0, 1, ..., n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poisson	log	$0, 1, 2, ...$	μ_i
gamma	inverse	$(0, \infty)$	$\phi\mu_i^2$
inverse-Gaussian	inverse-square	$(0, \infty)$	$\phi\mu_i^3$

- The *canonical link* for each familiy is not only the one most commonly used, but also arises naturally from the general formula for distributions in the exponential families.
  - Other links may be more appropriate for the specific problem at hand
  - One of the strengths of the GLM paradigm — in contrast, for example, to transformation of the response variable in a linear model — is the separation of the link function from the conditional distribution of the response.
- GLMs are typically fit to data by the method of maximum likelihood.
  - Denote the maximum-likelihood estimates of the regression parameters as $\widehat{\alpha}, \widehat{\beta}_1, ..., \widehat{\beta}_k$.
    - These imply an estimate of the mean of the response, $\widehat{\mu}_i = g^{-1}(\widehat{\alpha} + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik})$.

- The log-likelihood for the model, maximized over the regression coefficients, is
$$\log_e L_0 = \sum_{i=1}^{n} \log_e p(\widehat{\mu}_i, \phi; y_i)$$
where $p(\cdot)$ is the probability or probability-density function corresponding to the family employed.
- A "saturated" model, which dedicates one parameter to each observation, and hence fits the data perfectly, has log-likelihood
$$\log_e L_1 = \sum_{i=1}^{n} \log_e p(y_i, \phi; y_i)$$
- Twice the difference between these log-likelihoods defines the *residual deviance* under the model, a generalization of the residual sum of squares for linear models:
$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = 2(\log_e L_1 - \log_e L_0)$$
where $\mathbf{y} = \{Y_i\}$ and $\widehat{\boldsymbol{\mu}} = \{\widehat{\mu}_i\}$.

- Dividing the deviance by the estimated dispersion produces the *scaled deviance*: $D(\mathbf{y}; \widehat{\boldsymbol{\mu}})/\widehat{\phi}$.
- Likelihood-ratio tests can be formulated by taking differences in the residual deviance for nested models.
- For models with an estimated dispersion parameter, one can alternatively use incremental F-tests.
- Wald tests for individual coefficients are formulated using the estimated asymptotic standard errors of the coefficients.

▶ Some familiar examples:

- Combining the identity link with the Gaussian family produces the normal linear model.
  - The maximum-likelihood estimates for this model are the ordinary least-squares estimates.
- Combining the logit link with the binomial family produces the logistic-regression model (linear-logit model).

- Combining the probit link with the binomial family produces the linear probit model.

6. Poisson GLMs for Count Data

- Poisson generalized linear models arise in two common formally identical but substantively distinguishable contexts:

1. when the response variable in a regression model takes on non-negative integer values, such as a count;
2. to analyze associations among categorical variables in a contingency table of counts.

- The canonical link for the Poisson family is the log link.

6.1 Over-Dispersed Binomial and Poisson Models

- The binomial and Poisson GLMs fix the dispersion parameter ϕ to 1.
- It is possible to fit versions of these models in which the dispersion is a free parameter, to be estimated along with the coefficients of the linear predictor
  - The resulting error distribution is not an exponential family; the models are fit by "quasi-likelihood."
- The regression coefficients are unaffected by allowing dispersion different from 1, but the coefficient standard errors are multiplied by the square-root of $\widehat{\phi}$.
  - Because the estimated dispersion typically exceeds 1, this inflates the standard errors
  - That is, failing to account for "over-dispersion" produces misleadingly small standard errors.

- So-called *over-dispersed* binomial and Poisson models arise in several different circumstances.
  - For example, in modeling proportions, it is possible that
    - the probability of success μ varies for different individuals who share identical values of the predictors (this is called "unmodeled heterogeneity");
    - or the individual successes and failures for a "binomial" observation are not independent, as required by the binomial distribution.
- The negative-binomial distribution is also frequently used to model over-dispersed count data.

7. Diagnostics for GLMs

- Most regression diagnostics extend straightforwardly to generalized linear models.
- These extensions typically take advantage of the computation of maximum-likelihood estimates for generalized linear models by iterated weighted least squares (the procedure typically used to fit GLMs).

7.1 Outlier, Leverage, and Influence Diagnostics

7.1.1 Hat-Values

- Hat-values for a generalized linear model can be taken directly from the final iteration of the IWLS procedure
- They have the usual interpretation — except that the hat-values in a GLM depend on Y as well as on the configuration of the X's.

7.1.2 Residuals

- Several kinds of residuals can be defined for generalized linear models:
  - *Response residuals* are simply the differences between the observed response and its estimated expected value: $Y_i - \widehat{\mu}_i$.
  - *Working residuals* are the residuals from the final WLS fit.
    - These may be used to define partial residuals for component-plus-residual plots (see below).
  - *Pearson residuals* are case-wise components of the Pearson goodness-of-fit statistic for the model:

$$\frac{\widehat{\phi}^{1/2}(Y_i - \widehat{\mu}_i)}{\sqrt{\widehat{V}(Y_i|\eta_i)}}$$

where ϕ is the dispersion parameter for the model and $V(Y_i|\eta_i)$ is the variance of the response given the linear predictor.

- *Standardized Pearson residuals* correct for the conditional response variation and for the leverage of the observations:

$$R_{Pi} = \frac{Y_i - \widehat{\mu}_i}{\sqrt{\widehat{V}(Y_i|\eta_i)(1-h_i)}}$$

.

- *Deviance residuals*, D_i, are the square-roots of the case-wise components of the residual deviance, attaching the sign of $Y_i - \widehat{\mu}_i$.

▶ *Standardized deviance residuals* are

$$R_{Di} = \frac{D_i}{\sqrt{\widehat{\phi}(1-h_i)}}$$

▶ Several different approximations to *studentized residuals* have been suggested.

- To calculate exact studentized residuals would require literally refitting the model deleting each observation in turn, and noting the decline in the deviance.

- Here is an approximation due to Williams:

$$E_i^* = \sqrt{(1-h_i)R_{Di}^2 + h_i R_{Pi}^2}$$

where, once again, the sign is taken from $Y_i - \widehat{\mu}_i$.

- A Bonferroni outlier test using the standard normal distribution may be based on the largest absolute studentized residual.

7.1.3 Influence Measures

- An approximation to Cook's distance influence measure is

$$D_i = \frac{R^2_{Pi}}{\widehat{\phi}(k+1)} \times \frac{h_i}{1-h_i}$$

- Approximate values of dfbeta$_{ij}$ and dfbetas$_{ij}$ (influence and standardized influence on each coefficient) may be obtained directly from the final iteration of the IWLS procedure.
- There are two largely similar extensions of added-variable plots to generalized linear models, one due to Wang and another to Cook and Weisberg.

7.2 Nonlinearity Diagnostics

- Component-plus-residual plots also extend straightforwardly to generalized linear models.
  - Nonparametric smoothing of the resulting scatterplots can be important to interpretation, especially in models for binary responses, where the discreteness of the response makes the plots difficult to examine.
  - Similar effects can occur for binomial and Poisson data.
- Component-plus-residual plots use the linearized model from the last step of the IWLS fit.
  - For example, the partial residual for X_j adds the working residual to B_jX_{ij}.
  - The component-plus-residual plot graphs the partial residual against X_j.

8. Logit and Loglinear Models for Contingency Tables

8.1 The Binomial Logit Model for Contingency Tables

- When the explanatory variables — as well as the response variable — are discrete, the joint sample distribution of the variables defines a contingency table of counts.
- An example, drawn from *The American Voter* (Converse et al., 1960), appears below.
  - This table, based on data from a sample survey conducted after the 1956 U.S. presidential election, relates voting turnout in the election to strength of partisan preference, and perceived closeness of the election:

Perceived Closeness	*Intensity of Preference*	*Turnout*	
		Voted	Did Not Vote
One-Sided	Weak	91	39
	Medium	121	49
	Strong	64	24
Close	Weak	214	87
	Medium	284	76
	Strong	201	25

- The following table gives the *empirical logit* for the response variable,

$$\log_e \frac{\text{proportion voting}}{\text{proportion not voting}}$$

for each of the six combinations of categories of the explanatory variables:

Perceived Closeness	*Intensity of Preference*	$\log_e \frac{\text{Voted}}{\text{Did Not Vote}}$
One-Sided	Weak	0.847
	Medium	0.904
	Strong	0.981
Close	Weak	0.900
	Medium	1.318
	Strong	2.084

  - For example,

$$\begin{aligned}\text{logit(voted}&|\text{one-sided, weak preference)}\\ &= \log_e \frac{91/130}{39/130}\\ &= \log_e \frac{91}{39}\\ &= 0.847\end{aligned}$$

  - Because the conditional proportions voting and not voting share the same denominator, the empirical logit can also be written as

$$\log_e \frac{\text{number voting}}{\text{number not voting}}$$

  - The empirical logits are graphed in Figure 13, much in the manner of profiles of cell means for a two-way analysis of variance.

▶ Logit models are fully appropriate for tabular data.

- When, as in the example, the explanatory variables are qualitative or ordinal, it is natural to use logit or probit models that are analogous to analysis-of-variance models.

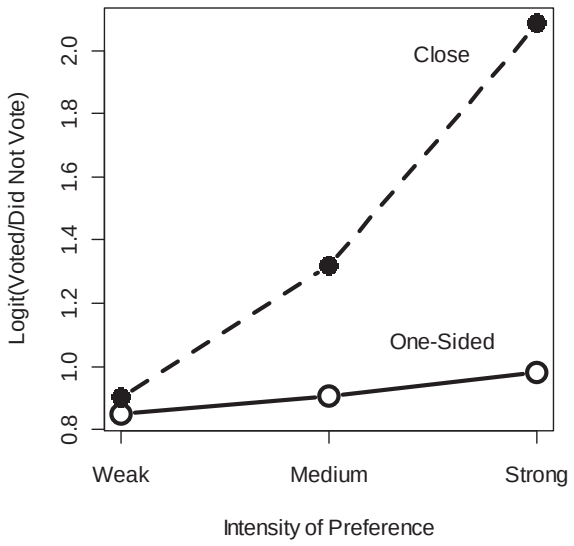

Figure 13. Empirical logits for the *American Voter* data.

- Treating perceived closeness of the election as the 'row' factor and intensity of partisan preference as the 'column' factor, for example, yields the model

$$\text{logit}\, \pi_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk}$$

where

  - π_{jk} is the conditional probability of voting in combination of levels j of perceived closeness and k of preference;
  - μ is the general mean of turnout in the population;
  - α_j is the main effect on turnout of membership in the jth level of perceived closeness;
  - β_k is the main effect on turnout of membership in the kth levels of preference; and
  - γ_{jk} is the interaction effect on turnout of simultaneous membership in levels j of perceived closeness and k of preference.

- Under the usual sigma constraints, this model leads to deviation-coded regressors (`contr.sum` in R), as in the analysis of variance.
- Likelihood-ratio tests for main-effects and interactions can be constructed in close analogy to the incremental F-tests for the two-way ANOVA model.

8.2 Loglinear Models

- Poisson GLMs may also be used to fit loglinear models to a contingency table of frequency counts, where the object is to model association among the variables in the table.
- The variables constituting the classifications of the table are treated as 'explanatory variables' in the Poisson model, while the cell count plays the role of the 'response.'
- We previously examined Campbell et al.'s data on voter turnout in the 1956 U. S. presidential election
  - We used a binomial logit model to analyze a three-way contingency table for turnout by perceived closeness of the election and intensity of partisan preference.
  - The binomial logit model treats turnout as the response.
- An alternative is to construct a log-linear model for the expected cell count.

- This model looks very much like a three-way ANOVA model, where in place of the cell mean we have the log cell expected count:

$$\begin{aligned}\log \mu_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_k \\ &\quad + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}\end{aligned}$$

- Here, variable 1 is perceived closeness of the election; variable 2 is intensity of preference; and variable 3 is turnout.
- Although a term such as $\alpha\beta_{ij}$ looks like an 'interaction,' it actually models the association between variables 1 and 2.
- The three-way term $\alpha\beta\gamma_{ijk}$ allows the association between any pair of variables to be different in different categories of the third variable; it thus represents an interaction in the usual sense of that concept.

▶ In fitting the log-linear model to data, we can use sigma-constraints on the parameters, much as we would for an ANOVA model.

▶ In the context of a three-way contingency table, the loglinear model above is a saturated model, because it has as many independent parameters (12) as there are cells in the table.

▶ The likelihood-ratio test for the three-way term Closeness $\times$ Preference $\times$ Turnout is identical to the test for the Closeness $\times$ Preference interaction in the logit model in which Turnout is the response variable.

▶ In general, as long as we fit the parameters for the associations among the explanatory variable (here Closeness$\times$Preference and, of course, its lower-order relatives, Closeness and Preference) and for the marginal distribution of the response (Turnout), the loglinear model for a contingency table is equivalent to a logit model.
- There is, therefore, no real advantage to using a loglinear model in this setting.
- Loglinear models, however, can be used to model association in other contexts.

9. Implementation of GLMs in R

- The `glm()` function in R is very similar in use to `lm()`,

```
glm(formula, family, data, subset,
    weights, na.action, contrasts)
```

- The `family` argument is one of `gaussian` (the default), `binomial`, `poisson`, `Gamma`, `inverse.gaussian`, `quasi`, `quasibinomial`, or `quasipoisson`.
  - It is possible to write functions for additional families (e.g., the `negative.binomial` family for count data in the **MASS** package).
- The "family-generator" function specified as the value of the `family` argument can itself take a link argument (and possibly other arguments); in each case there is a default link.
  - The available links for each family (○) and the default link (●) are given in the following table:

	link			
family	identity	inverse	sqrt	1/mu^2
gaussian	●	○		
binomial				
poisson	○		○	
Gamma	○	●		
inverse.gaussian	○	○		●
quasi	●	○	○	○
quasibinomial				
quasipoisson	○		○	

	link			
family	log	logit	probit	cloglog
gaussian	○			
binomial	○	●	○	○
poisson	●			
Gamma	○			
inverse.gaussian	○			
quasi	○	○	○	○
quasibinomial		●	○	○
quasipoisson	●			