STAT 8620, Categorical Data Analysis & Generalized Linear Models — Lecture Notes

The goal of this course is to teach methods for the analysis of discrete response data and to develop a general framework for the analysis of discrete data and other data types for which the assumptions of the classical linear model (CLM) do not hold.

Such a framework is provided by a class of models known as **generalized linear models** or **GLMs**.

- GLMs extend the class of CLMs (a.k.a. normal-theory linear models, Gauss-Markov models, or, confusingly, general linear models).

CLMs can be extended in other important ways:

- E.g., by the inclusion of both traditional (fixed) regression parameters and "random" parameters, better known as random effects.
  - Such *linear mixed models* (or LMMs) are very useful for handling correlation and multiple sources of variability (covered in STAT 8630).
- By allowing more general forms of nonlinear regression functions (STAT 8230).
- More general classes are possible: generalized linear mixed models (GLMMs; covered in STAT 8630), nonlinear mixed models (NLMMs; coverd in STAT 8230), etc.

In this class, however, we concentrate on GLMs and extensions of GLMs suitable for the analysis of independent, discrete (or otherwise non-normal) data.

But first, we need to begin by introducing some "pre-model" ideas: descriptive statistics, measures of association, model-free inference in simple data tables, etc.

Non-model-based Concepts and Methods for Discrete Data:

What do we mean by discrete data?

A discrete random variable is a random variable that can taken on a finite or countably infinite number of possible values.

- In practice, all random variables are discrete, due to limitations in the precision of measurement.
- Typically, though, variables that theoretically have an underlying continuous scale are treated as continuous in statistical analyses - unless the scale of measurement is extremely coarse. E.g., weight, height, time elapsed.

The practical basis of the distinction is, does the variable taken on enough values with positive probability to be well approximated by a continuous distribution.

- Therefore, we're concerned with random variables that can assume only a small number of values.

This includes many **qualitative**, or **nominally scaled categorical** variables,

– Religion (Christian, Hindu, Jewish), Gender (male, female), etc.

and also **ordinally scaled categorical** variables.

– Agreement (strongly agree, agree, neutral, disagree, strongly disagree), Pain (mild, moderate, severe), etc.

As presented none of these characteristics (Religion, Agreement, etc.) are even, strictly speaking, random variables. They only become so, and become analyzable, when we assign numbers to their values.

– Religion (1=Christian, 2=Moslem, 3=Jewish), Gender (1=male, 2=female), Agreement (1=strongly agree, 2=agree, 3=neutral, 4=disagree, 5=strongly disagree), etc

The other main types of discrete variables we'll work with are **grouped continuous** variables (e.g., < 1, 1–5, > 5 years of service) and **counts** (e.g., number of hospitalizations).

2-Way Contingency Tables:

A 2−way **contingency table** a.k.a. **cross-tabulation** is simply a two-way array containing the joint distribution of two categorical random variables.

- Contingency tables can be used to display either the joint frequency distribution or the joint probability distribution.

Let X and Y denote two categorical random variables having I and J levels, respectively.

Let $\pi_{ij} = \Pr(X = i, Y = j)$. Then the joint probability distribution for the pair (X, Y) can be displayed as an $I \times J$ table:

The collection of π_{ij}'s $i = 1, \ldots, I$, $j = 1, \ldots, J$, can be written succinctly as $\{\pi_{ij}\}$. $\{\pi_{ij}\}$ is the joint distribution of (X, Y).

The probability distribution of X ignoring Y (i.e., averaging over Y) is called the **marginal distribution** of X because of its natural place in the margin of the contingency table.

- The marginal distribution of X is given by $\{\pi_{i+}\}$ where $\pi_{i+} = \Pr(X = i)$.
- The marginal distribution of Y is given by $\{\pi_{+j}\}$ where $\pi_{+j} = \Pr(Y = j)$.

The conditional distribution of X given $Y = j$ may also be of interest. This distribution is given by the set of probabilities $\{\pi_{1|j}, \ldots, \pi_{I|j}\}$ where

$$\pi_{i|j} = \Pr(X = i|Y = j) = \frac{\Pr(X = i, Y = j)}{\Pr(Y = j)} = \frac{\pi_{ij}}{\pi_{+j}}$$

Joint frequency distributions can also be displayed in contingency tables simply by replacing π_{ij}'s with n_{ij}'s:

- Here, n_{ij} =the number of "subjects" out of a random sample of size n whose response was $(X, Y) = (i, j)$.

The sample analog of the joint probability distribution is the joint relative frequency distribution, obtained from the joint frequency distribution simply by dividing through by n:

- Here, $p_{ij} = n_{ij}/n$ is the proportion of the sample for which the response was $(X, Y) = (i, j)$.

Perhaps the most basic question of interest in a two-way table situation is whether or not X and Y are **independent**. That is, is it true that

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \qquad \text{for all } i, j?$$

Note that under independence,

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} = \frac{\pi_{i+}\pi_{+j}}{\pi_{i+}} = \pi_{+j}$$

for all i, j.

- That is, Y is independent of X iff the conditional distribution of Y is the same as the marginal distribution of Y, for each row of the table.

- This is a natural way to think about independence especially if X is the explanatory variable, and Y is the response.

Another question that often is of interest in two-way tables formed from ordinal variables is, are large values of Y more (less) likely when $X = i$ than when $X = i'$.

I.e., if we think of there being an underlying continuous distribution for Y (that has been grouped or categorized to form Y), do the underlying conditional densities look like:

Or, equivalently, do the underlying c.d.f.s look like this:

Mathematically, this happens when for two rows i and i' we have

$$F_{j|i'} \leq F_{j|i}, \quad \text{for all } j$$

where $F_{j|i}$ denotes the conditional c.d.f. of Y given that $X = i$:

$$F_{j|i} = \sum_{h \leq j} \pi_{h|i}$$

The simplest case of a two-way table results when X and Y are both binary (dichotomous, Bernoulli, quantal). In this case we have a 2×2 table:

There are several ways to compare probabilities in a table like this.

1. Risk Difference.

   In epidemiology and biostatistics, 2×2 tables often involve a response Y =disease status (1=disease present, 2=absent) and an explanatory variable X (e.g., gender: 1=female, 2=male):

   In this case, $\pi_{1|i}$ = probability, or **risk** of being diseased for gender i.

An obvious way to compare the disease risks is to take the **risk difference**:

$$\pi_{1|1} - \pi_{1|2} = (1 - \pi_{2|1}) - (1 - \pi_{2|2}) = -(\pi_{2|1} - \pi_{2|2})$$

- Notice that comparisons of conditional probabilities that $Y = 1$ with the risk difference are equivalent to comparisons of conditional probabilities that $Y = 2$ (differ only by sign).

- In the 2×2 table, X, Y are independent iff

$$\pi_{1|1} - \pi_{1|2} = 0$$

- In the $I \times 2$ table, X, Y are independent iff

$$\pi_{1|i} - \pi_{1|i'} = 0 \qquad \text{for all } i, i'$$

and for the $I \times J$ table, X, Y are independent iff

$$\pi_{j|i} - \pi_{j|i'} = 0 \qquad \text{for all } i, i' \in \{1, \ldots, I\} \text{ and all } j$$

2. Relative Risk.

One feature (drawback) of the risk difference is that it ignores the rarity of the disease; e.g., risk difference is .01 if female, male risks are $(.51, .50)$ and if female, male risks are $(.02, .01)$.

- In the latter case the risk is twice as great for females!

This suggests that for some purposes, the ratio of risks is a better basis of comparison. For 2×2 tables the relative risk is

$$\pi_{1|1}/\pi_{1|2}$$

- The relative risk takes values in $[0, +\infty)$, and a value of 1.0 corresponds to independence.

3. Odds ratio.

Another common way to think about chance is in terms of the odds of an event occurring rather than the probability of the event.

- The **odds** of an event are simply the probability of the event occurring divided by the probability that the event does not occur.

In a 2×2 table, the odds that $Y = 1$ in the first row (given that $X = 1$) is

$$\Omega_1 = \frac{\pi_{1|1}}{\pi_{2|1}} = \frac{\pi_{11}/\pi_{1+}}{\pi_{12}/\pi_{1+}} = \frac{\pi_{11}}{\pi_{12}}$$

and the odds in the second row is $\Omega_2 = \pi_{1|2}/\pi_{2|2} = \pi_{21}/\pi_{22}$.

The **odds ratio** compares the chances of $Y = 1$ at the two levels of X where "chances" are in terms of odds rather than probabilities (risks). We'll denote the odds ratio as θ:

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

- Because of this last representation, θ is sometimes called the **cross-product ratio**.
- Odds ratios take values in $[0, +\infty)$, and, when all cell probabilities are positive, a value of 1.0 corresponds to independence.
- When $\theta > 1$, $Y = 1$ is more likely in the first row, than in the second.
- When $\theta < 1$, $Y = 1$ is more likely in the second row, than in the first.
- Note that θ is not symmetric about 1; e.g., $\theta = 4$ and $\theta = 0.25$ represent equally strong associations between X and Y, but in the opposite direction.

The sample version of θ replaces π_{ij}'s by p_{ij}'s, or, equivalently by n_{ij}'s:

$$\hat{\theta} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{\frac{n_{11}n_{22}}{n^2}}{\frac{n_{12}n_{21}}{n^2}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

The odds ratio has several nice invariance properties, not all of which are possessed by the relative risk and risk difference.

- θ is invariant to interchange of row variable and column variable (if we let Y the row variable and X the column variable we get the same value of θ).

  $\Rightarrow$ Explanatory/response variable distinction doesn't affect θ.

- If we switch either the row order or the column order the result is to invert the value of θ (same strength of association but direction changed to match the row or column change).

- The sample version of θ is unaffected if we multiply through by a constant in either a row or a column.

  – This last result is true of the relative risk and risk difference with respect to rows, but not columns.

It follows from these facts that the odds ratio is the only one of the three association measures that is appropriate for **cross-sectional**, **prospective**, and **retrospective** study designs.

Prospective Study (Clinical Trials, Cohort Studies): exposed and unexposed groups are identified and followed over time to compare incidence of disease.

Retrospective Study (Case-Control Study): diseased and disease-free subjects are identified and their exposure history investigated.

Cross-sectional Study: Sample of n subjects of unknown disease and exposure status is identified and disease status and prior exposure status are assessed simultaneously.

		Disease Status		
		D	$\bar{D}$	
Exposure Status	E	n_{11}	n_{12}	n_{1+}
	$\bar{E}$	n_{21}	n_{22}	n_{2+}
		n_{+1}	n_{+2}	n

	Prospective	Retrospective	Cross-sectional
Row totals:	fixed	random	random
Col. totals:	random	fixed	random
Grand total:	fixed	fixed	fixed

In a cross-sectional design, we fix n, and then measure $n_{11}, n_{12}, n_{21}, n_{22}$, from which we form $p_{11} = n_{11}/n, p_{12} = n_{12}/n, p_{21} = n_{21}/n, p_{22} = n_{22}/n$ which are appropriate estimators of $\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}$.

- Since we can estimate the π_{ij}'s, we can estimate the $\pi_{j|i}$'s $(\pi_{1|1}, \pi_{1|2})$; and since all three association measures can be computed from these quantities, we can estimate risk difference, relative risk, and odds ratio.

Compared to a cross-sectional design, in a prospective design, we artificially inflate or deflate n_{1+}, n_{2+} to fixed sizes. That is, we multiply through each row by a constant.

- We still have information appropriate to estimate the $\pi_{j|i}$'s:

$$\hat{\pi}_{j=1|i=1} = n_{11}/n_{1+}, \qquad \hat{\pi}_{j=1|i2} = n_{21}/n_{2+}$$

  so we can still estimate risk difference, relative risk, and odds ratio.

Compared to a cross-sectional design, in a retrospective design, we artificially inflate or deflate n_{+1}, n_{+2} to fixed sizes. That is, we multiply through each column by a constant.

- We now have information appropriate to estimate the $\pi_{i|j}$'s (the probability of being exposed or not exposed given your disease status): $\hat{\pi}_{i=1|j=1} = n_{11}/n_{+1}$, $\hat{\pi}_{i=1|j=2} = n_{12}/n_{+2}$.

- However, we can't form appropriate estimators of the $\pi_{j|i}$'s, so we can't get at the risk difference or relative risk.

- In contrast, we do still have information from which it is appropriate to estimate θ, because

$$\begin{aligned}\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} &= \frac{\frac{\pi_{11}}{\pi_{+1}}\frac{\pi_{22}}{\pi_{+2}}}{\frac{\pi_{12}}{\pi_{+2}}\frac{\pi_{21}}{\pi_{+1}}} \\ &= \frac{\pi_{i=1|j=1}\pi_{i=2|j=2}}{\pi_{i=1|j=2}\pi_{i=2|j=1}}\end{aligned}$$

  and we can estimate the $\pi_{i|j}$'s.

Examples:

A clinical trial of aspirin use to prevent heart attack: Approximately equal numbers of subjects (11,034 and 11,037) were assigned to placebo and aspirin use groups.

		Heart Attack: Yes	Heart Attack: No	
Aspirin Use	Placebo	189	10,845	11,034
	Aspirin	104	10,933	11,037
		293	21,778	22,071

Risk estimates ($\hat{\pi}_{j|i}$'s):

$$\hat{\pi}_{1|1} = 189/11034 = .0171, \quad \hat{\pi}_{1|2} = 104/11037 = .0094$$

- $\Rightarrow$ risk difference estimate = $.0171 - .0094 = .0077$ (risk is greater on placebo).
- Relative risk estimate = $.0171/.0094 = 1.82$.
- Odds ratio estimate is $(189)(10933)/[10845(104)] = 1.83$.
- The similarity of the odds ratio and relative risk estimates here illustrates a general phenomenon: *for rare events* (small risks), $\theta \approx$relative risk.

A case-control study of oral contraceptive use and heart attack. 58 female heart attack victims were identified and each of these "cases" was matched to three "control" subjects of similar age, etc. who had not suffered heart attacks.

		Heart Attack: Yes	Heart Attack: No	
Contraceptive Use	Yes	23	34	57
	No	35	132	167
		58	166	224

In this case, the probability of heart attack given whether or not oral contraceptives have been used is not estimable. However, we can estimate probabilities of contraceptive use given presence or absence of heart attack:

$$\hat{\pi}_{i=1|j=1} = 23/58 = .397, \qquad \hat{\pi}_{i=1|j=2} = 34/166 = .205$$

And from these quantities we can estimate the odds ratio:

$$\Rightarrow \qquad \hat{\theta} = \frac{\frac{23}{58}\left(1 - \frac{34}{166}\right)}{\frac{34}{166}\left(1 - \frac{23}{58}\right)} = \frac{23(132)}{34(35)} = 2.551$$

The odds ratio (and risk difference and relative risk) quantifies the association between two levels of X (two rows) and two levels of Y (two columns).

- For 2×2 tables, that's the whole story (the odds ratio summarizes all of the association in the table).

For $I \times J$ tables, odds ratios can be constructed for all of the combinations of pairs of rows ($\binom{I}{2}$ of them) combined with pairs of columns ($\binom{J}{2}$ of them).

- Many of these $\binom{I}{2}\binom{J}{2}$ odds ratios are redundant.

A (non-unique) minimal set to describe all of the association in a $I \times J$ table requires only $(I-1)(J-1)$ odds ratios.

Possibilities:

1. Adjacent category odds ratios:

$$\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}}, \quad i = 1, \ldots, I-1, j = 1, \ldots, J-1$$

2. Reference category odds ratios (last category = reference):

$$\theta_{ij} = \frac{\pi_{ij}\pi_{IJ}}{\pi_{i,J}\pi_{I,j}}, \quad i = 1, \ldots, I-1, j = 1, \ldots, J-1$$

3. Reference category odds ratios (first category = reference):

$$\theta_{ij} = \frac{\pi_{ij}\pi_{11}}{\pi_{i,1}\pi_{1,j}}, \quad i = 2, \ldots, I, j = 2, \ldots, J$$

- When the response scale is nominal, it is difficult to summarize the information in a minimal set of $(I-1)(J-1)$ odds ratios as a single number with little loss of information.
- Several such measures exist, however, including the *concentration coefficient* and the *uncertainty coefficient*. See Agresti, §2.4.2.

For ordinal scaled responses, reduction of the odds ratios in an $I \times J$ table down to a single number is easier and more appropriate.

For interval and ratio scaled random variables, bivariate analyses often focus on linear association (Pearson correlation). For ordinal variables, though, linearity is not a meaningful concept. Instead, measures of linear association can be replaced by measures of monotonicity.

- X and Y have a **monotone increasing** (decreasing) relationship if for (X_a, Y_a), (X_b, Y_b) measured on two subjects a and b, $X_a > X_b$ implies $Y_a \geq Y_b$ ($Y_a \leq Y_b$). The relationship is **strictly monotone increasing** if $X_a > X_b$ implies $Y_a > Y_b$ ($Y_a < Y_b$).

There are several measures of the degree to which monotonicity tends to hold; these include **Goodman and Kruskal's** γ, **Kendall's** τ_b , and **Somers'** d.

- All of these parameters are defined in terms of the probabilities of **concordance** and **discordance** and are estimated by counting numbers of concordant and discordant pairs.

- A pair of subjects on whom we've measured X and Y is concordant if the subject with the higher X-value is also the subject with the higher Y-value.

- A pair of subjects is discordant if the subject with the higher X-value is also the subject with the lower Y-value.

- A pair of subjects is tied if the subjects have the same X-value or the same Y-value.

Example — Clothing and Intelligence

The table below was published in 1894 in *Biometrika*. It contains a cross-classification of X ="standard of clothing" and Y ="intelligence" for 1725 school children. Both X and Y were subjectively measured on ordinal scales.

	Intelligence						
Clothing	A,B	C	D	E	F	G	
1=Very badly clad	17	13	22	10	10	1	73
2=Poor but passable	39	58	70	61	33	4	265
3=Well clad	41	100	202	255	138	15	751
4=Very well clad	33	48	113	209	194	39	636
Total	130	219	407	535	375	59	1725

Intelligence was rated as follows: A =mentally deficient, B =slow and dull, C =dull, D =slow but intelligent, E =fairly intelligent, F =distinctly capable, G =very able. Categories A and B are combined in the above table.

- A pair where subject 1 is in the $(1, D)$ cell and subject 2 is in the $(2, E)$ cell is concordant (subject 2 is higher on both X and Y).
- A pair where subject 1 is in the $(1, D)$ cell and subject 2 is in the $(3, C)$ cell is discordant (subject 1 is lower on X and higher on Y).
- A pair where subject 1 is in the $(1, D)$ cell and subject 2 is in the $(1, F)$ cell is tied.

The measures of monotonicity mentioned above all are defined in terms of Π_c and Π_d, the probabilities of concordance and discordance, respectively, for a randomly selected pair of bivariate observations.

For these parameters, the association is said to be positive (negative) if $\Pi_c - \Pi_d > 0$ ($\Pi_c - \Pi_d < 0$), where

$$\Pi_c = 2\sum_i \sum_j \pi_{ij} \left(\sum_{h>i} \sum_{k>j} \pi_{hk} \right), \quad \Pi_d = 2\sum_i \sum_j \pi_{ij} \left(\sum_{h>i} \sum_{k<j} \pi_{hk} \right)$$

The most important of these measures is G&K's γ. Gamma (γ) is defined as

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d} = \frac{\Pi_c}{\Pi_c + \Pi_d} - \frac{\Pi_d}{\Pi_c + \Pi_d}.$$

- From the last expression above we obtain an interpretation for γ: γ is equal to the difference in the conditional probabilities of concordance and discordance, given that the pair is not tied.

Since γ is a difference in probabilities, it takes values in $[-1, +1]$ with perfect positive (negative) association occurring when $\gamma = 1$ ($\gamma = -1$).

The sample estimate of γ is

$$\hat{\gamma} = \frac{C - D}{C + D}$$

where C =the total number of concordant pairs and D =total number of discordant pairs.

Back to the Example:

$$\begin{aligned} C &= 17(58+70+61+33+4+100+\cdots+39) \\ &\quad + 13(70+61+33+4+202+\cdots+39) + 22(61+\cdots+39) \\ &\quad + \cdots + 138(39) \\ &= 507067 \end{aligned}$$

and

$$\begin{aligned} D &= 13(39+41+33) + 22(39+58+41+100+33+48) + \cdots \\ &\quad + 15(33+48+113+209+194) = 254066 \end{aligned}$$

so

$$\hat{\gamma} = \frac{507067 - 254066}{507067 + 254066} = .3324$$

- Here we have mild positive association. Intelligence tends to increase with quality of clothing (keep in mind, though, that there's a lot wrong with a study of this kind).
- See also handout clothes.sas.

Inference for Two-way Tables

- See Ch. 3 of Agresti.

The data generation mechanism that leads to a given contingency table is usually modelled with one of three sampling models:

1. Poisson Sampling;
2. Multinomial Sampling; or
3. Product Multinomial Sampling.

(Rarely, a fourth sampling model based on the hypergeometric distribution comes up, but we'll talk about that later.)

Example 1 (piston ring failures):

For quality control purposes, an engine manufacturer kept track of the number of piston ring failures that occurred in its engines while under warranty over a certain period of time. Failures were cross-classified by cylinder and position on the piston.

	Position			
CylinderNo.	N	C	S	
1	17	17	12	46
2	11	9	13	33
3	11	8	19	38
4	14	7	28	49
	53	41	72	166

- Notice that in this case all margins are random.

Example 2 (malignant melanoma):

Medical researchers decided to investigate the incidence of malignant melanoma (skin cancer) by tumor type and location (site) of tumor. To do this they examined hospital records from 400 patients treated for malignant melanoma.

	Site			
TumorType	Head/Neck	Trunk	Extremities	
A	22	2	10	34
B	16	54	115	185
C	19	33	73	125
D	11	17	28	56
	68	106	226	400

- Notice that in this case the bottom right table margin (400) is fixed by design, all other margins are random.

Example 3 (flu vaccine):

In a clinical trial for a flu vaccine approximately half of a total sample size of 73 subjects were randomized to each of two groups: placebo, and vaccine. The response of interest was antibody level, which was categorized as low, medium and high.

	AntibodyLevel			
TreatmentGroup	Low	Medium	High	
Placebo	25	8	5	38
Vaccine	6	18	11	35
	31	26	16	73

- Notice that in this case the bottom right margin as well as both row margins are fixed by design. The column margins are random.

These 3 tables illustrate 3 distinct sampling situations $\Rightarrow$ we must have 3 distinct sampling models.

Sampling Models:

Let n_{ij} =count in the $(i,j)^{\text{th}}$ cell of of an $I \times J$ contingency table (we won't distinguish notationally between the random variable n_{ij} and its realized value; the reference will typically be clear from the context). Let $N = IJ$, the total number of cells in the table, and let $m_{ij} = \mathrm{E}(n_{ij})$ $i = 1, \ldots, I$, $j = 1, \ldots, J$, denote the expected cell frequencies.

1. Poisson sampling. Assume $n_{11}, \ldots, n_{IJ}$'s are independent Poisson r.v.'s with means $m_{11}, \ldots, m_{IJ}$, respectively.

   Recall the Poisson distribution: $n_{ij} \sim Poisson(m_{ij})$ means that n_{ij} has probability mass function

$$f(n_{ij}; m_{ij}) = \begin{cases} \frac{\exp(-m_{ij}) m_{ij}^{n_{ij}}}{n_{ij}!} & \text{for } n_{ij} = 0, 1, 2, \ldots; \\ 0 & \text{otherwise} \end{cases}$$

   and $\mathrm{E}(n_{ij}) = \mathrm{var}(n_{ij}) = m_{ij}$.

- The Poisson distribution is appropriate when n_{ij} can be thought of as the count of events that occur according to a Poisson process with rate m_{ij}.

- More generally, the Poisson distribution is useful for counts of events that occur randomly through time or space without upper bound.

Recall that the likelihood function for a random variable Y is equal to the probability density (mass) function for that random variable, but thought of as a function of the parameters of that density rather than as a function of the value of the random variable.

The likelihood function for random variables $Y_1, \ldots, Y_n$ (i.e., for the random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$) is the joint p.d.f. of $Y_1, \ldots, Y_n$ treated as a function the parameters of the density.

Likelihood function for n_{ij}:

$$L(m_{ij}; n_{ij}) = f(n_{ij}; m_{ij}) = \frac{\exp(-m_{ij}) m_{ij}^{n_{ij}}}{n_{ij}!}$$

$\Rightarrow$ the likelihood for $\mathbf{n} = n_{11}, \ldots, n_{IJ}$ is

$$L(\mathbf{m}; \mathbf{n}) = \prod_{i,j} \frac{e^{-m_{ij}} m_{ij}^{n_{ij}}}{n_{ij}!}$$

where $\mathbf{m} = (m_{11}, \ldots, m_{IJ})^T$.

- Notice that if $n_{11}, \ldots, n_{IJ}$ are all independent Poisson random variables, then $n = \sum_{i,j} n_{ij}$ is random, too. In fact, $n \sim Poisson(\sum_{i,j} m_{ij})$.
  $\Rightarrow$ appropriate for a situation like example 1, piston ring failures.

2. Multinomial Sampling.

   What if n is fixed, as in the melanoma example?

   In that case the Poisson sampling model is not appropriate. If $n = 400$, then no single cell count can exceed 400 ($n_{ij} \leq n \ \forall i, j$).

   Instead we can think of the N cells in the contingency table as N distinct, mutually-exclusive, and exhaustive outcomes for a single categorical response variable, with outcome probabilities $\pi_{11}, \pi_{12}, \ldots, \pi_{IJ}$.

   This approach leads to the multinomial sampling model.

The Multinomial Distribution

Let Z be a response variable taking N possible values which we'll label $1, 2, \ldots, N$. In some sense, Z is really a multivariate response, because we can replace Z with $\mathbf{x} = (x_1, \ldots, x_{N-1})^T$ where

$$x_r = \begin{cases} 1, & \text{if } Z = r,\ r = 1, \ldots, N-1, \\ 0, & \text{otherwise} \end{cases}$$

Consideration of Z is equivalent to considering $\mathbf{x}$ because

$$Z = r \qquad \text{if and only if} \qquad \mathbf{x} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$$

and

$$\Pr(Z = r) = \Pr(x_r = 1)$$

Now suppose we have n copies of $\mathbf{x}$: $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Then the sum of these vectors

$$\mathbf{y} = \sum_{i=1}^{n} \mathbf{x}_i$$

follows the multinomial distribution and

$$\Pr(\mathbf{y} = (n_1, \ldots, n_{N-1})^T) = \frac{n!}{\left(\prod_{r=1}^{N-1} n_r!\right)(n - n_+)!} \left(\prod_{r=1}^{N-1} \pi_r^{n_r}\right) (1 - \pi_+)^{n - n_+}$$

where $\pi_r = \Pr(Z_i = r)$, $\pi_+ = \sum_{r=1}^{N-1} \pi_r$, and $n_+ = \sum_{r=1}^{N-1} n_r$.

- We write this as $\mathbf{y} \sim Mult(n, \boldsymbol{\pi})$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{N-1})^T$.

While $\mathbf{y}$ is the vector consisting of the number of units falling in each of the first $N - 1$ categories, $\mathbf{y}/n$ gives the proportion of the sample falling in each of these categories. Moments of $\bar{\mathbf{y}} = \mathbf{y}/n$ are

$$\mathrm{E}(\bar{\mathbf{y}}) = \boldsymbol{\pi}, \qquad \mathrm{var}(\bar{\mathbf{y}}) = \frac{1}{n}(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T)$$

$$= \frac{1}{n} \begin{pmatrix} \pi_1(1 - \pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_{N-1} \\ -\pi_2\pi_1 & \pi_2(1 - \pi_2) & \cdots & -\pi_2\pi_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_{N-1}\pi_1 & -\pi_{N-1}\pi_2 & \cdots & \pi_{N-1}(1 - \pi_{N-1}) \end{pmatrix}$$

For the two-way table where $\sum_{i,j} n_{ij} = n$, any subset of $IJ - 1$ if the IJ cell counts follows a multinomial distribution. Equivalently, if $\mathbf{n} = (n_{11}, \ldots, n_{IJ})^T$ and $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{IJ})^T$, the likelihood function is given by

$$L(\boldsymbol{\pi}; \mathbf{n}) = \Pr(\mathbf{n} = \mathbf{n}; \boldsymbol{\pi}) = n! \prod_{i,j} \frac{\pi_{ij}^{n_{ij}}}{n_{ij}!}$$

- In contrast to the Poisson sampling model, we've parameterized the likelihood in terms of the cell probabilities (the π_{ij}'s) rather than the expected cell counts (the m_{ij}'s), but these quantities are related by

$$m_{ij} = n\pi_{ij}, \quad i = 1, \ldots, I, j = 1, \ldots, J.$$

3. Product Multinomial Sampling.

   In Poisson sampling, no margins were fixed, in multinomial sampling the grand total was fixed. If row (or column) totals are fixed, then its appropriate to think of the table as made up independent multinomial samples corresponding to the rows.

- In the flu vaccine example, two independent samples were taken: one of the 38 placebo patients, and a second of 35 vaccine patients.

- $\Rightarrow$ the two rows of the flu vaccine table can be thought of as independent multinomials.

  First row: $(n_{11}, n_{12})^T \sim Mult\left(38, (\pi_{j=1|i=1}, \pi_{j=2|i=1})^T\right)$.
  Second row: $(n_{21}, n_{22})^T \sim Mult\left(35, (\pi_{j=1|i=2}, \pi_{j=2|i=2})^T\right)$.

  The likelihood for the entire table is the product of the multinomial likelihoods for each row (independence).

  For row totals fixed, the likelihood is given by

$$\prod_{i=1}^{I} n_{i+}! \prod_{j=1}^{J} \frac{\pi_{j|i}^{n_{ij}}}{n_{ij}!}$$

- For row totals fixed, multinomial probabilities are related to expected cell counts via

$$m_{ij} = n_{i+}\pi_{j|i}, \quad i = 1, \ldots, I, j = 1, \ldots, J.$$

Estimation:

We've already utilized the sample proportions (p_{ij}'s where $p_{ij} = n_{ij}/n$) as estimates of the corresponding cell probabilities (the π_{ij}'s) — for example, to form an estimator of the odds ratio in a 2×2 table: $\hat{\theta} = p_{11}p_{22}/(p_{12}p_{21})$.

p_{ij} as an estimator of π_{ij} can be motivated as a method of moments (MOM) estimator — we estimate a population moment (in this case, a mean) with the corresponding sample moment.

This estimator can also be derived as the maximum likelihood estimator (MLE).

Maximum Likelihood Estimation:

> Suppose we have a discrete random variable Y (possibly a vector) with observed value y. Suppose Y has probability mass function $f(y; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.
>
> The **likelihood function**, $L(\boldsymbol{\theta}; y)$ is defined to equal the probabilty mass function (more generally, the density) but viewed as a function of $\boldsymbol{\theta}$, not y:
>
> $$L(\boldsymbol{\theta}; y) = f(y; \boldsymbol{\theta})$$
>
> Therefore, the likelihood at $\boldsymbol{\theta}_0$, say, has the interpretation
>
> $$\begin{aligned} L(\boldsymbol{\theta}_0; y) &= \Pr(Y = y \quad \text{when} \quad \boldsymbol{\theta} = \boldsymbol{\theta}_0) \\ &= \Pr(\text{observing the obtained data when } \boldsymbol{\theta} = \boldsymbol{\theta}_0) \end{aligned}$$
>
> Logic of ML: choose the value of $\boldsymbol{\theta}$ that makes this probability largest $\Rightarrow \hat{\boldsymbol{\theta}}$, the MLE.

Important Property of MLEs: Consider a reparameterization, $\phi = h(\theta)$ where h is one-to-one, $h : \Theta \to \Phi$.

If $\hat{\theta}$ is a MLE of θ, then $\hat{\phi} = h(\hat{\theta})$ is a MLE of ϕ (MLE is invariant to parameterization).

We use the same procedure when Y is continuous with density function $f(y;\boldsymbol{\theta})$: maximize $L(\boldsymbol{\theta};y) = f(y;\boldsymbol{\theta})$

Often, our data come from a random sample so that we observe $\mathbf{y}$ corresponding to $\mathbf{Y}_{n\times 1}$, a vector of independent r.v.'s. In this case

$$L(\boldsymbol{\theta};\mathbf{y}) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\theta})$$

Since its easier to work with sums than products its useful to note that in general

$$\arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta};y) = \arg\max_{\boldsymbol{\theta}} \underbrace{\log L(\boldsymbol{\theta};y)}_{\equiv \ell(\boldsymbol{\theta};y)}$$

It can be shown that under all three sampling models, the log-likelihood function is strictly concave and has a unique maximum (Birch, 1963; Haberman, 1973, 1974; Baker *et al.* 1985).

It follows that the MLEs can be found by solving the **likelihood equations** in which we set partial derivatives of the log-likelihood with respect to each parameter equal to zero.

Multinomial Sampling.

The multinomial loglikelihood is

$$\ell(\boldsymbol{\pi}; \mathbf{n}) = \underbrace{\sum_i \sum_j n_{ij} \log(\pi_{ij})}_{\text{kernel of } \ell} + \underbrace{\log(n!) - \sum_i \sum_j \log(n_{ij}!)}_{\text{doesn't involve } \boldsymbol{\pi}}$$

- Maximizing the kernel of $\ell(\boldsymbol{\pi}; \mathbf{n})$ is equivalent to maximizing $\ell(\boldsymbol{\pi}; \mathbf{n})$.

To find MLE of $\boldsymbol{\pi}$ we must maximize the kernel of $\ell(\boldsymbol{\pi}; \mathbf{n})$ subject to the side constraint $\sum_i \sum_j \pi_{ij} = 1$.

Question: *How do we maximize a function $f(\boldsymbol{\theta})$ subject to equality constraints $g_1(\boldsymbol{\theta}) = 0, \ldots, g_r(\boldsymbol{\theta}) = 0$?*

Answer: Use method of Lagrangian multipliers.

First construct the "Lagrangian" $\mathcal{L}$ where

$$\mathcal{L} = f(\boldsymbol{\theta}) - \lambda_1 g_1(\boldsymbol{\theta}) - \cdots - \lambda_r g_r(\boldsymbol{\theta})$$

Then maximize $\mathcal{L}$ with respect to $\boldsymbol{\theta}$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^T$. The maximizer of the Lagrangian $\hat{\boldsymbol{\theta}}$ is the constrained maximizer of $f(\boldsymbol{\theta})$.

- The λ_i's are called **Lagrange multipliers**.

To obtain multinomial MLE we first construct the Lagrangian ($r = 1$):

$$\mathcal{L} = \sum_i \sum_j n_{ij} \log \pi_{ij} - \lambda(\sum_i \sum_j \pi_{ij} - 1)$$

Then maximize with respect to $\boldsymbol{\pi}$, λ:

$$\frac{\partial \mathcal{L}}{\partial \pi_{ij}} = \frac{n_{ij}}{\pi_{ij}} - \lambda, \quad i = 1, \ldots, I, j = 1, \ldots, J$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = 1 - \sum_i \sum_j \pi_{ij}$$

Setting these partials to zero implies that the maximizers $\hat{\lambda}$ and $\hat{\pi}_{ij}$, $i = 1, \ldots, J$, $j = 1, \ldots, J$, satisfy

$$\hat{\lambda}\hat{\pi}_{ij} = n_{ij}, \quad i = 1, \ldots, I, j = 1, \ldots, J$$
$$\sum_i \sum_j \hat{\pi}_{ij} = 1$$

Adding the first IJ equations together we get

$$\hat{\lambda} \sum_{i,j} \hat{\pi}_{ij} = \sum_{i,j} n_{ij} = n$$

which combined with the last equation yields $\hat{\lambda} = n$.

So, the MLE of π_{ij} subject to the constraint $\sum_{i,j} \pi_{ij} = 1$ is

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n} = p_{ij}, \quad \text{the sample proportion.}$$

Since $m_{ij} = n\pi_{ij}$, the MLE of the expected count in the $(i, j)^{\text{th}}$ cell is

$$\hat{m}_{ij} = n\hat{\pi}_{ij} = n_{ij}, \quad \text{the observed cell count}$$

Poisson Sampling.

For Poisson sampling, obtaining the MLEs of the expected cell counts is a bit easier since there are no constraints on these parameters.

In this case, the loglikelihood is given by

$$\ell(\mathbf{m};\mathbf{n}) = \underbrace{\sum_i \sum_j \{n_{ij}\log(m_{ij}) - m_{ij}\}}_{\text{kernel of } \ell} - \sum_i \sum_j \log(n_{ij}!)$$

The likelihood equations are given by

$$\left.\frac{\partial \ell}{\partial m_{ij}}\right|_{m_{ij}=\hat{m}_{ij}} = 0, \quad i = 1,\ldots,I, j = 1,\ldots,J$$

or

$$\frac{n_{ij}}{\hat{m}_{ij}} - 1 = 0 \quad \Rightarrow \quad \hat{m}_{ij} = n_{ij}, \quad i = 1,\ldots,I, j = 1,\ldots,J$$

- MLEs are the same under Poisson, multinomial sampling!

Product Multinomial Sampling.

Let $\boldsymbol{\pi}^c = (\pi_{j=1|i=1},\ldots,\pi_{j=J|i=I})^T$ be the vector of conditional probabilities conditioning on the value of X, the row variable. Then the product multinomial loglikelihood is given by

$$\ell(\boldsymbol{\pi}^c;\mathbf{n}) = \sum_i \log n_{i+}! + \sum_i \sum_j (n_{ij}\log \pi_{j|i} - \log n_{ij}!)$$

The $\pi_{j|i}$'s must satisfy $\sum_j \pi_{j|i} = 1$ for each i, so we must maximize $\ell(\boldsymbol{\pi}^c;\mathbf{n})$ subject to these constraints.

Lagrangian:

$$\mathcal{L} = \sum_i \sum_j n_{ij} \log \pi_{j|i} - \sum_i \lambda_i \left(\sum_j \pi_{j|i} - 1 \right)$$

Partials:

$$\frac{\partial \mathcal{L}}{\partial \pi_{j|i}} = \frac{n_{ij}}{\pi_{j|i}} - \lambda_i, \quad i = 1, \ldots, I, j = 1, \ldots, J$$
$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = \sum_j \pi_{j|i} - 1 \quad i = 1, \ldots, I$$

So the MLEs must satisfy

$$n_{ij} = \hat{\lambda}_i \hat{\pi}_{j|i}, \quad i = 1, \ldots, I, j = 1, \ldots, J$$
$$\sum_j \hat{\pi}_{j|i} = 1 \quad i = 1, \ldots, I$$

Summing the first set of equations over j for each i and using the second set of equations we get $\hat{\lambda}_i = n_{i+}$, $i = 1, \ldots, I$.

Plugging $\hat{\lambda}_i = n_{i+}$ back into the first set of equations, we get the MLEs:

$$\hat{\pi}_{j|i} = \frac{n_{ij}}{n_{i+}}, \quad i = 1, \ldots, I, j = 1, \ldots, J$$

- Since $m_{ij} = n_{i+}\pi_{j|i}$, we get $\hat{m}_{ij} = n_{i+}\hat{\pi}_{j|i} = n_{ij}$, under product multinomial sampling.
- Again, we get the same MLEs for expected cell counts.

Goodness of Fit

Suppose we have a multinomial vector $(n_1, \ldots, n_{N-1})^T$ with probabilities $(\pi_1, \ldots, \pi_{N-1})^T$ based on a sample of size n.

Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{N-1}, \pi_N)^T$, $\mathbf{n} = (n_1, \ldots, n_{N-1}, n_N)^T$ where $\pi_N = 1 - \sum_{i=1}^{N-1} \pi_i$ and $n_N = n - \sum_{i=1}^{N-1} n_i$ (i.e., $\sum_i^N \pi_i = 1$ and $\sum_i^N n_i = n$).

Consider the problem of testing $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$ where $\boldsymbol{\pi}_0 = (\pi_{10}, \ldots, \pi_{N0})^T$ and $\sum_i \pi_{i0} = 1$.

Under H_0, the expected category frequencies are $m_i = n\pi_{i0}$, $i = 1, \ldots, N$. A test statistic for comparing observed category frequencies with expected (or estimated expected) frequencies is called a **goodness of fit** statistic.

The two most important goodness of fit statistics are

1. The Pearson chi-squared statistic (Pearson statistic).
2. The likelihood ratio or deviance statistic.

The Pearson statistic is given by

$$X^2 = \sum_{i=1}^{N} \frac{(n_i - m_i)^2}{m_i}$$

or, in the two-way table,

$$X^2 = \sum_{i}^{I} \sum_{j}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}.$$

The likelihood ratio statistic is given by

$$G^2 = 2 \sum_{i}^{N} n_i \log(n_i/m_i)$$

or, in the two-way table,

$$G^2 = 2 \sum_{i}^{I} \sum_{j}^{J} n_{ij} \log(n_{ij}/m_{ij}).$$

Asymptotically, both of these statistics have $\chi^2(N-1)$ distributions ($\chi^2(IJ-1)$ in the two-way table case). This result holds for fixed N as $n \to \infty$.

- The sample size necessary for the limiting χ^2 distribution to provide a good approximation to the exact distributions of X^2 and G^2 depends upon the statistic, n, N, and the degree of "sparseness" in the table.

- For fixed number of cells, the exact distribution of X^2 usually converges to χ^2 more quickly than that of G^2. Agresti recommends a minimum ratio n/N of at least 5 for G^2, while n/N can be as small as 1 for accuracy of the χ^2 approximation to X^2 as long as the table does not contain both very small and moderately large expected frequencies.

- Section 9.8.4 of Agresti provides further guidelines concerning the accuracy of the χ^2 approximations to the distributions of X^2 and G^2.

  Example – Mendel's Genetics Experiments:

  Gregor Mendel is widely considered to be the father of modern genetics. In 1865 he published results of an experiment in which he crossed pea plants of a pure yellow strain with plants of a pure green strain. He hypothesized the existence of genes in the parent plants that controlled color and that came in one of two varieties: dominant and recessive. In this case the color gene was hypothesized to be either y or g with y being dominant. Offspring receive one gene from each parent, so there are four gene-pairs:

  (y, y), (y, g), (g, y) – all resulting in a yellow colored offspring (y is dominant); and
  (g, g) – which results in a green colored offspring.

  Mendel hypothesized that in second generation crosses of pure yellow and pure green plants, these four gene-pairs would be equally likely so that 75% of the second generation plants would be yellow, 25% green.

Observed data:

Color		
Yellow	Green	
6022	2001	8023

Expected data:

Color		
Yellow	Green	
$8023(.75) = 6017.25$	$8023(.25) = 2005.75$	8023

Test statistics:

$$X^2 = \frac{(6022 - 6017.25)^2}{6017.25} + \frac{(2001 - 2005.75)^2}{2005.75} = 0.015$$

$$G^2 = 2\left\{6022 \log(6022/6017.25) + 2001 \log(2001/2005.75)\right\} = 0.015$$

Here $N - 1 = 1$, so each statistic has an approximate $\chi^2(1)$ distribution which gives an approximate $p-$value of $p = .88$ for each statistic. $\Rightarrow$ fail to reject H_0, $\Rightarrow$ data support Mendel's theory.

Mendel reported the results of several experiments of this sort. R.A. Fisher used the the reproductive property of the χ^2 distribution to summarize those results.

The reproductive property says that if $X_1^2, \ldots, X_k^2$ are independent χ^2 random variables with degrees of freedom $\nu_1, \ldots, \nu_k$, respectively, then $\sum_i^k X_i^2 \sim \chi^2(\sum_i^k \nu_i)$.

By combining X^2 statistics in this way across many of Mendel's experiments, Fisher obtained $X^2 = 42$ on d.f.=84 which yields $p = .99996$.

- This result says that under H_0, the probability is .00004 that the observed cell counts should fit the expected cell counts at least as closely as Mendel's data did.
- Did Mendel "cook" his data? Maybe so.

Where do the test statistics X^2 and G^2 come from and why are their (asymptotic) distributions χ^2?

Notice that in the $N = 2$ case (Mendel example), $H_0 : (\pi_1, \pi_2)^T = (\pi_{10}, \pi_{20})^T$ is equivalent to $H_0 : \pi_1 = \pi_{10}$ since $\pi_2 = 1 - \pi_1$ and $\pi_{20} = 1 - \pi_{10}$.

That is, the problem is that of testing a binomial probability equal to some null value. In the Mendel example we observed a binomial: 6022 successes (yellow plants) out of 8023 trials, and we wanted to know if $p = 6022/8023$ was consistent with an underlying probability of success $\pi = \pi_0$ where $\pi_0 = .75$.

By the Central Limit Theorem, $p \dot{\sim} N(\pi, \pi(1-\pi)/n)$ for large n, which leads to a $z-$test for H_0 (STAT 2000):

$$z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \dot{\sim} N(0,1)$$

or, equivalently,

$$z^2 = n(p-\pi_0)[\pi_0(1-\pi_0)]^{-1}(p-\pi_0) \dot{\sim} \chi^2(1)$$

With a little algebra, its easy to show that $z^2 = X^2$ in the $N = 2$ case.

More generally, for any N, let $\mathbf{p} = (n_1/n, \ldots, n_{N-1}/n)^T$, $\boldsymbol{\pi}_0 = (\pi_{10}, \ldots, \pi_{N-1,0})^T$ and $\Sigma_0 = \text{var}(\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}_0))$ evaluated under H_0. Then the Pearson statistic X^2 can be written as a quadratic form:

$$X^2 = n(\mathbf{p} - \boldsymbol{\pi}_0)^T \Sigma_0^{-1} (\mathbf{p} - \boldsymbol{\pi}_0)$$

In the $N = 2$ (binomial) case, $\sqrt{n}(p - \pi_0)$ converged to $N(0, \pi_0(1-\pi_0))$ under H_0 (CLT). In the $N > 2$ (multinomial) case, $\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}_0)$ converges to $N(\mathbf{0}, \Sigma_0)$. Therefore, the quadratic form

$$n(\mathbf{p} - \boldsymbol{\pi}_0)^T \Sigma_0^{-1} (\mathbf{p} - \boldsymbol{\pi}_0)$$

converges in distribution to a $\chi^2(N-1)$.

How about G^2?

Under likelihood-based inference the classical tool for testing hypotheses is the Likelihood Ratio Statistic:

$$\lambda = \frac{L(\tilde{\boldsymbol{\theta}}; \mathbf{y})}{L(\hat{\boldsymbol{\theta}}; \mathbf{y})}, \quad \begin{array}{l} \tilde{\boldsymbol{\theta}} = \text{MLE under } H_0 \cup H_A \\ \hat{\boldsymbol{\theta}} = \text{MLE under } H_0 \end{array}$$

Logic:

If $\lambda \doteq 1 \Rightarrow$ no evidence against H_0.

If $\lambda >> 1$ then assuming H_0 is true has made our data much less likely that it would have been without assuming $H_0 \Rightarrow$ reject H_0.

How large does λ need to be to reject?

Answer: large in comparison to its distribution (above the $100(1-\alpha)^{\text{th}}$ percentile of its distribution).

Wilks is famous for having proved that under mild regularity conditions, $2 \log \lambda$ has a limiting $\chi^2(\nu)$ distribution, as $n \to \infty$. Here $\nu =$the difference between the dimensions of the parameter space for $\boldsymbol{\theta}$ under $H_0 \cup H_A$ and H_0.

- I.e., $\nu =$ the number of restrictions placed on $\boldsymbol{\theta}$ by H_0.

In our problem $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$ places $N-1$ restrictions on $\boldsymbol{\pi}$: (1) $\pi_1 = \pi_{10}$, (2) $\pi_2 = \pi_{20}$, ..., $(N-1)$ $\pi_{N-1} = \pi_{N-1,0}$.

- Under both H_0 and H_A we require $\sum_i^N \pi_i = 1$, so $\pi_N = \pi_{N0}$ is not an additional restriction once restrictions (1)–$(N-1)$ are required.

- Another way to think about the d.f. is that the dimension of the parameter space for $\boldsymbol{\pi}$ under $H_0 \cup H_A$ is $N-1$ (all N of the π_i's can vary freely except for one restriction) and the dimension under H_0 is 0 (none of the π_i's can vary freely. Therefore the degrees of freedom are $N-1-0 = N-1$.

Under multinomial sampling, the likelihood is $n! \prod_i^N \frac{\pi_i^{n_i}}{n_i!}$, so

$$\lambda = \frac{n! \prod_i^N \frac{\tilde{\pi}_i^{n_i}}{n_i!}}{n! \prod_i^N \frac{\hat{\pi}_i^{n_i}}{n_i!}} = \prod_i^N \frac{\tilde{\pi}_i^{n_i}}{\hat{\pi}_i^{n_i}} = \prod_i^N \left(\frac{\tilde{\pi}_i}{\hat{\pi}_i} \right)^{n_i}$$

where $\tilde{\pi}_i = n_i/n$ is the unrestricted MLE, and $\hat{\pi}_i = \pi_{i0}$ is the restricted MLE (under $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$).

Since $\tilde{\pi}_i/\hat{\pi}_i = n_i/m_{i0}$, where $m_{i0} = n\pi_{i0}$ is the expected cell frequency under the null hypothesis, we have

$$\lambda = \prod_i^N \left(\frac{n_i}{m_{i0}} \right)^{n_i}$$

so

$$2 \log(\lambda) = 2 \sum_i n_i \log(n_i/m_{i0}) = G^2 \overset{a}{\sim} \chi^2(N-1)$$

The fact that X^2 and G^2 have the same asymptotic χ^2 distribution says that these tests are asymptotically equivalent.

- In fact, X^2 and G^2 are special cases of a family of tests statistics which are all asymptotically equivalent — the power-divergence family.
- The power divergence family is defined to include statistics of the form

  $$\frac{2}{\phi(\phi+1)} \sum_i^N n_i \left[\left(\frac{n_i}{m_i} \right)^{\phi} - 1 \right]$$

  with special cases given by different choices of $\phi \in (-\infty, +\infty)$
- The special cases corresponding to $\phi = 0$ and $\phi = -1$ are defined as the limits as $\phi \to 0$ and $\phi \to -1$, respectively.
- Note that $\phi = 1$ gives X^2 and $\phi \to 0$ gives G^2, but other important cases include $\phi = -2$ (Neymans' modified chi-square statistic, $\sum_i^N (n_i - m_i)^2/n_i$), $\phi = -1/2$ (the Freeman-Tukey statistic, $4 \sum_i^N (\sqrt{n_i} - \sqrt{m_i})^2$), and $\phi = 2/3$ (Cressie and Read found that the χ^2 approximation in small samples is best for this choice of ϕ).

Goodness of fit with estimated expected frequencies:

One of the most common uses of the Pearson statistic in a two-way table is to test independence between the row and column variables, X and Y.

	Site			
TumorType	Head/Neck	Trunk	Extremities	
A	22	2	10	34
B	16	54	115	185
C	19	33	73	125
D	11	17	28	56
	68	106	226	400

For example, in the above table a natural question to ask is, "Does tumor type depend upon tumor site?"

Under independence,

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \quad i = 1, \ldots, I, j = 1, \ldots, J.$$

If we knew the π_{i+}'s and the π_{+j}'s, then we could test $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ $\forall i, j$ using

$$X^2 = \sum_i^I \sum_j^J \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \dot{\sim} \chi^2(IJ - 1)$$

where $m_{ij} = n\pi_{i+}\pi_{+j}$.

The usual case, however, is one in which we don't know the marginal probabilities. The best we can do is to estimate m_{ij} with $\hat{m}_{ij} = np_{i+}p_{j+}$ and substitute estimates for the expected cell counts in X^2:

$$X^2 = \sum_i^I \sum_j^J \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

How does this affect the distribution of X^2?

Answer: When the m_{ij}'s are estimated by ML (or any best asymptotically normal (BAN) estimator) then $X^2 \overset{a}{\sim} \chi^2(N-1-s)$ ($N = IJ$ in the two-way table) where s =the the number of "nonredundant" parameters estimated to obtain the $\hat{m}_{ij}$'s.

- Proof: see Read and Cressie (1988, Appendix 6), or Bishop *et al.* (1975, §14.9).
- BAN: consistent, asymptotically normal, and asymptotically efficient.
- s "non-redundant" parameters: the Jacobian of the transformation from the model parameters to $\mathbf{m}$ is of rank s.
- Analogous result holds for G^2 and other power-divergence statistics.
- All results on asymptotic distributions of power-divergence statistics assume N fixed, $n \to \infty$ (we can get into trouble when N increases with n) and hold under all three sampling models.

In the two-way table under independence, $\hat{m}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = np_{i+}p_{+j}$. We estimated $\pi_{1+}, \ldots, \pi_{I+}$ with $p_{1+}, \ldots, p_{I+}$ ($I-1$ non-redundant parameters) plus we estimated $\pi_{+1}, \ldots, \pi_{+J}$ with $p_{+1}, \ldots, p_{+J}$ ($J-1$ non-redundant parameters).

So, with expected cell frequencies estimated under independence,

$$X^2, G^2 \dot{\sim} \chi^2(IJ - 1 - [I-1] - [J-1]) = \chi^2([I-1][J-1])$$

Independence in the Melanoma Example: (See melanoma.sas.)

Observed cell counts:

TumorType	Site: Head/Neck	Trunk	Extremities	
A	22	2	10	34
B	16	54	115	185
C	19	33	73	125
D	11	17	28	56
	68	106	226	400

Expected cell counts:

TumorType	Site: Head/Neck	Trunk	Extremities	
A	$34(68)/400 = 5.78$	9.01	19.21	34
B	31.45	49.025	104.525	185
C	21.25	33.125	70.625	125
D	9.52	14.84	31.64	56
	68	106	226	400

So,

$$X^2 = \frac{(22 - 5.78)^2}{5.78} + \cdots + \frac{(28 - 31.64)^2}{31.64}$$
$$= 45.517 + \cdots + .4188 = 65.8129$$

and $G^2 = 51.795$. Both statistics have approximate $\chi^2(12 - 1 - 3 - 2) = \chi^2(6)$ distributions, which yields $p-$values $< .0001$ in each case.

- Tumor type and tumor site are not independent.
- Notice in melanoma.lst that the contributions to the overall chi-square statistic vary considerably over the entire table. This suggests that the dependence between tumor type and site may be limited to (or at least strongest for) certain categories of X and Y. It would be of interest here to *partition* X^2 (or G^2) to better understand the nature of the dependence between tumor type and site. See §3.3 in Agresti.

Large-Sample Confidence Intervals

Suppose we would like to form a $100(1-\alpha)\%$ confidence interval around θ, the odds ratio; or around γ, Goodman and Kruskal's measure of monotonicity; or suppose we want to test a hypothesis on the odds ratio of the form $H_0 : \theta = \theta_0$ or on the relative risk: $H_0 : \pi_{1|1}/\pi_{1|2} = \mathrm{RR}_0$?

In each of these cases we require an exact, or at least approximate, distribution for some function $f(\hat{\boldsymbol{\pi}})$, and a standard error for $f(\hat{\boldsymbol{\pi}})$.

A general method for obtaining the asymptotic distribution (including asymptotic s.e.s) of such a function is provided by the **delta-method.**

$\delta-$Method (see §14.1 of Agresti):

Let X be a r.v. with known mean and variance:

$$\mathrm{E}(X) = \mu, \qquad \mathrm{var}(X) = \mathrm{E}[(X-\mu)^2] = \sigma^2$$

Let $Y = g(X)$ where g is a twice differentiable function.

Suppose that exact calculation of $\mathrm{E}(Y)$ and $\mathrm{var}(Y)$ is difficult. Approximations may be obtained based on a Taylor series expansion of $g(X)$ about μ.

By Taylor's Theorem

$$g(X) = g(\mu) + g'(\mu)(X-\mu) + \frac{1}{2}g''(\mu^*)(X-\mu)^2$$

where μ^* lies between X and μ.

Assuming the 3^{rd} term is small and can be ignored, we have the approximations

$$\begin{aligned}
\mathrm{E}(Y) &\doteq \mathrm{E}[g(\mu) + g'(\mu)(X-\mu)] = g(\mu) \\
\mathrm{var}(Y) &= \mathrm{E}[(g(X) - \mathrm{E}\{g(X)\})^2] \\
&\doteq \mathrm{E}[(g(\mu) + g'(\mu)(X-\mu) - g(\mu))^2] = \mathrm{E}[\{g'(\mu)(X-\mu)\}^2] \\
&= [g'(\mu)]^2 \mathrm{E}((X-\mu)^2] = [g'(\mu)]^2 \mathrm{var}(X)
\end{aligned}$$

Now let X_n be a sequence of r.v.'s such that the asymptotic distribution of $\sqrt{n}(X_n - \mu)$ is $N(0, \sigma^2(\mu))$. I.e., suppose

$$\sqrt{n}(X_n - \mu) \xrightarrow{\mathrm{d}} N(0, \sigma^2(\mu))$$

For a sequence of (non-random) variables $\{x_n\}$, by Taylor's Theorem

$$\begin{aligned} g(x_n) &= g(\mu) + (x_n - \mu)g'(\mu) + \frac{1}{2}(x_n - \mu)^2 g''(\mu^*) \\ &= g(\mu) + (x_n - \mu)g'(\mu) + \mathrm{O}(|x_n - \mu|^2) \end{aligned}$$

Substituting the random sequence X_n for x_n and rearranging,

$$\begin{aligned} \sqrt{n}(g(X_n) - g(\mu)) &= \sqrt{n}(X_n - \mu)g'(\mu) + \sqrt{n}\underbrace{\mathrm{O_p}(|X_n - \mu|^2)}_{=\mathrm{O_p}(n^{-1})} \\ &= \sqrt{n}(X_n - \mu)g'(\mu) + \underbrace{\mathrm{O_p}(n^{-1/2})}_{=\mathrm{o_p}(1)} \end{aligned}$$

$\Rightarrow \sqrt{n}(g(X_n) - g(\mu))$ has the same limiting distribution as $\sqrt{n}(X_n - \mu)g'(\mu)$. Or,

$$\Rightarrow \qquad \sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{\mathrm{d}} N(0, \sigma^2(\mu)[g'(\mu)]^2).$$

Here, if $g'(\cdot)$ and $\sigma^2(\cdot)$ are continuous at μ, the asymptotic variance $\sigma^2(\mu)[g'(\mu)]^2$ can be consistently estimated by $\sigma^2(X_n)[g'(X_n)]^2$, so (e.g.)

$$g(X_n) \pm 1.96\sigma(X_n)|g'(X_n)|/\sqrt{n}$$

is an approximate (large-sample) 95% confidence interval for $g(\mu)$.

Multivariate Case:

Now let $\mathbf{X} = (X_1, \ldots, X_p)^T$ be a random vector with known mean and variance-covariance matrix:

$$\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}, \qquad \mathrm{var}(\mathbf{X}) = \Sigma$$

Let $Y = g(X_1, \ldots, X_p)$ where g is a continuous function with first and second partial derivatives.

A Taylor series expansion of $g(\mathbf{X})$ about $\boldsymbol{\mu}$ is given by

$$g(\mathbf{X}) = g(\boldsymbol{\mu}) + \sum_{i=1}^{p} \frac{\partial g}{\partial \mu_i}(X_i - \mu_i) + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial^2 g}{\partial \mu_i \partial \mu_j}(X_i - \mu_i)(X_j - \mu_j) + \cdots$$

where

$$\frac{\partial g}{\partial \mu_i} = \frac{\partial g(\mathbf{X})}{\partial X_i}\Big|_{\mathbf{X}=\boldsymbol{\mu}}, \quad \frac{\partial^2 g}{\partial \mu_i \partial \mu_j} = \frac{\partial^2 g(\mathbf{X})}{\partial X_i \partial X_j}\Big|_{\mathbf{X}=\boldsymbol{\mu}}$$

Let $\partial g(\boldsymbol{\mu})/\partial \boldsymbol{\mu}^T = (\frac{\partial g}{\partial \mu_1}, \ldots, \frac{\partial g}{\partial \mu_p})$ be the row vector of partials with respect to the elements of $\boldsymbol{\mu}$. Then, ignoring 3rd and higher-order terms,

$$\Rightarrow \qquad Y = g(\mathbf{X}) \doteq g(\boldsymbol{\mu}) + \left(\frac{\partial g(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)(\mathbf{X} - \boldsymbol{\mu})$$

$$\Rightarrow \qquad \mathrm{E}(Y) \doteq \mathrm{E}(g(\boldsymbol{\mu})) + \left(\frac{\partial g(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)\mathrm{E}(\mathbf{X} - \boldsymbol{\mu}) = g(\boldsymbol{\mu})$$

and

$$\begin{aligned}
\mathrm{var}(Y) &= \mathrm{E}[\{g(\mathbf{X}) - \mathrm{E}(g(\mathbf{X}))\}^2] \\
&\doteq \mathrm{E}[\{g(\boldsymbol{\mu}) + \left(\frac{\partial g(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)(\mathbf{X} - \boldsymbol{\mu}) - g(\boldsymbol{\mu})\}^2] = \mathrm{E}[\{\left(\frac{\partial g(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)(\mathbf{X} - \boldsymbol{\mu})\}^2] \\
&= \mathrm{E}[\left(\frac{\partial g(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T \left(\frac{\partial g(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)^T] \\
&= \left(\frac{\partial g(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)\mathrm{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]\left(\frac{\partial g(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)^T = \left(\frac{\partial g(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)\Sigma\left(\frac{\partial g(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)^T
\end{aligned}$$

Now let $\mathbf{Y} = (Y_1, \ldots, Y_u)^T$, where

$$Y_i = g_i(X_1, \ldots, X_p) = g_i(\mathbf{X}), \qquad i = 1, \ldots, u$$

From the univariate results above,

$$\mathrm{E}(Y_i) \doteq g_i(\boldsymbol{\mu}), \qquad \mathrm{var}(Y_i) \doteq \left(\frac{\partial g_i(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right) \Sigma \left(\frac{\partial g_i(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)^T, \qquad i = 1, \ldots, u$$

The off-diagonal elements in $\mathrm{var}(\mathbf{Y})$ can also be approximated:

$$\begin{aligned}
\mathrm{cov}(Y_i, Y_j) &= \mathrm{E}[(Y_i - \mathrm{E}(Y_i))(Y_j - \mathrm{E}(Y_j))] \\
&\doteq \mathrm{E}[\left(\frac{\partial g_i(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)(\mathbf{X} - \boldsymbol{\mu})\left(\frac{\partial g_j(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)(\mathbf{X} - \boldsymbol{\mu})] \\
&= \mathrm{E}[\left(\frac{\partial g_i(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\left(\frac{\partial g_j(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)^T] \\
&= \left(\frac{\partial g_i(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)\mathrm{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]\left(\frac{\partial g_j(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)^T \\
&= \left(\frac{\partial g_i(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)\Sigma\left(\frac{\partial g_j(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)^T
\end{aligned}$$

Now let $\left(\frac{\partial \mathbf{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)$ denote the $u \times p$ matrix with i^{th} row equal to $\frac{\partial g_i(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}$. That is, $\left(\frac{\partial g(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)$ has $(i,j)^{\text{th}}$ element $\left.\frac{\partial g_i(\mathbf{X})}{\partial X_j}\right|_{\mathbf{X}=\boldsymbol{\mu}}$.

Then the results above can be summarized as

$$\mathrm{E}(\mathbf{Y}) \doteq \mathbf{g}(\boldsymbol{\mu}), \qquad \mathrm{var}(\mathbf{Y}) \doteq \left(\frac{\partial \mathbf{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)\Sigma\left(\frac{\partial \mathbf{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)^T$$

Now let $\mathbf{X}_n$ denote a $p-$dimensional sequence of random vectors such that

$$\sqrt{n}(\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{\text{d}} N_p(\mathbf{0}, \Sigma(\boldsymbol{\mu}))$$

Let $\mathbf{g}(\mathbf{X}_n) = (g_1(\mathbf{X}_n), \ldots, g_u(\mathbf{X}_n))^T$ be a $u-$vector-valued function admitting the following expansion:

$$g_i(\mathbf{x}) = g_i(\boldsymbol{\mu}) + \sum_{j=1}^{p}(x_j - \mu_j)\frac{\partial g_i(\mathbf{x})}{\partial x_j}\Big|_{\mathbf{x}=\boldsymbol{\mu}} + \text{O}(||\mathbf{x} - \boldsymbol{\mu}||) \quad i = 1, \ldots, u$$

Then the $\delta-$method's distributional result is

$$\sqrt{n}\left(\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\boldsymbol{\mu})\right) \overset{a}{\sim} N_u\left(\mathbf{0}, \left(\frac{\partial \mathbf{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)\Sigma(\boldsymbol{\mu})\left(\frac{\partial \mathbf{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}\right)^T\right)$$

- Again, $\partial \mathbf{g}(\boldsymbol{\mu})/\partial \boldsymbol{\mu}^T$ and Σ can be evaluated at $\mathbf{X}_n$ to obtain a consistent estimate for the asymptotic variance of $\mathbf{g}(\mathbf{X}_n)$ for large-sample confidence regions and hypothesis tests.

$\delta-$Method for log-odds ratio:

The MLE of the odds ratio in the 2×2 table is

$$\hat{\theta} = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

Recall that $\hat{\theta}$ is not symmetric around the independence value of 1. Furthermore, $\hat{\theta}$ is a multiplicative function of the p_{ij}'s.

In contrast,

$$\log(\hat{\theta}) = \log p_{11} + \log p_{22} - \log p_{12} - \log p_{21}$$

is symmetric around zero and is an additive function. The result is that $\log(\hat{\theta})$ converges in distribution to normality much faster than $\hat{\theta}$.

- For this reason, inference concerning the odds ratio is often done through the log-odds ratio. For example, we construct a 95% CI for $\log(\theta)$ and then exponentiate the endpoints to obtain a 95% interval for θ.

Let $\mathbf{p} = (p_{11}, p_{12}, p_{21})^T$ and $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{21})^T$. Notice that $\log(\hat{\theta})$ can be written as a function of $\mathbf{p}$:

$$\log(\hat{\theta}) = \log p_{11} + \log(1 - p_{11} - p_{12} - p_{21}) - \log p_{12} - \log p_{21} = g(\mathbf{p})$$

where

$$\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{\text{d}} N(\mathbf{0}, \Sigma(\boldsymbol{\pi}))$$

and

$$\Sigma(\boldsymbol{\pi}) = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T = \begin{pmatrix} \pi_{11}(1-\pi_{11}) & -\pi_{11}\pi_{12} & -\pi_{11}\pi_{21} \\ -\pi_{11}\pi_{12} & \pi_{12}(1-\pi_{12}) & -\pi_{12}\pi_{21} \\ -\pi_{11}\pi_{21} & -\pi_{12}\pi_{21} & \pi_{21}(1-\pi_{21}) \end{pmatrix}$$

Therefore, we can apply the multivariate $\delta-$method. In the case of the log odds ratio, $g(\mathbf{p}) = \log(\hat{\theta})$ is a scalar-valued function, so $u = 1$ and

$$\begin{aligned}\frac{\partial g(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}^T} &= \left(\frac{\partial g(\boldsymbol{\pi})}{\partial \pi_{11}}, \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_{12}}, \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_{21}}\right) \\ &= \left(\frac{1}{\pi_{11}} - \frac{1}{\pi_{22}}, -\frac{1}{\pi_{22}} - \frac{1}{\pi_{12}}, -\frac{1}{\pi_{22}} - \frac{1}{\pi_{21}}\right) \\ &= \left(\frac{\pi_{22} - \pi_{11}}{\pi_{11}\pi_{22}}, \frac{-\pi_{12} - \pi_{22}}{\pi_{12}\pi_{22}}, \frac{-\pi_{21} - \pi_{22}}{\pi_{21}\pi_{22}}\right)\end{aligned}$$

It follows that

$$\sqrt{n}\{\log(\hat{\theta}) - \log(\theta)\} \xrightarrow{\text{d}} N\left(0, \left(\frac{\partial g(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}^T}\right) \Sigma(\boldsymbol{\pi}) \left(\frac{\partial g(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}^T}\right)^T\right)$$

or, for large n,

$$\log(\hat{\theta}) \dot{\sim} N\left(\log(\theta), \frac{1}{n}\left(\frac{\partial g(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}^T}\right) \Sigma(\boldsymbol{\pi}) \left(\frac{\partial g(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}^T}\right)^T\right)$$

The asymptotic variance of $\log(\hat{\theta})$ can be consistently estimated by substituting $\mathbf{p}$ for $\boldsymbol{\pi}$ which leads to an estimated asymptotic variance of

$$\begin{aligned}&\frac{1}{n}\left(\frac{\partial g(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}^T}\bigg|_{\boldsymbol{\pi}=\mathbf{p}}\right) \Sigma(\mathbf{p}) \left(\frac{\partial g(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}^T}\bigg|_{\boldsymbol{\pi}=\mathbf{p}}\right)^T = \frac{1}{n}\left(\frac{p_{22} - p_{11}}{p_{11}p_{22}}, \frac{-p_{12} - p_{22}}{p_{12}p_{22}}, \frac{-p_{21} - p_{22}}{p_{21}p_{22}}\right) \\ &\times \begin{pmatrix} p_{11}(1-p_{11}) & -p_{11}p_{12} & -p_{11}p_{21} \\ -p_{11}p_{12} & p_{12}(1-p_{12}) & -p_{12}p_{21} \\ -p_{11}p_{21} & -p_{12}p_{21} & p_{21}(1-p_{21}) \end{pmatrix} \begin{pmatrix} \frac{p_{22}-p_{11}}{p_{11}p_{22}} \\ \frac{-p_{12}-p_{22}}{p_{12}p_{22}} \\ \frac{-p_{21}-p_{22}}{p_{21}p_{22}} \end{pmatrix} \\ &= \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \qquad \text{(after some algebra).}\end{aligned}$$

- $\Rightarrow$ For large n, $\log(\hat{\theta})$ is approximately normally distributed with asymptotic standard error (ASE) $\left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{1/2}$.

Oral Contraceptive Use and Heart Attack:

Recall the results of this case-control study:

		Heart Attack: Yes	Heart Attack: No	
Contraceptive Use	Yes	23	34	57
	No	35	132	167
		58	166	224

which yielded $\hat{\theta} = 2.551 \Rightarrow \log(\hat{\theta}) = \log(2.551) = 0.937$.

The ASE for $\log(\hat{\theta})$ is

$$\left(\frac{1}{23} + \frac{1}{34} + \frac{1}{35} + \frac{1}{132}\right)^{1/2} = \sqrt{.109} = .330$$

So an approximate (large-sample) 95% confidence interval for $\log(\theta)$ is given by

$$0.937 \pm 1.96(.330) = (0.289, 1.584)$$

which leads to an approximate 95% CI for θ given by

$$(e^{0.289}, e^{1.584}) = (1.336, 4.873)$$

- See oralcon.sas.
- Agresti recommends adding .5 to each cell of the table to improve the estimator of θ and to improve the estimator of its asymptotic standard deviation.
- Inference for relative risk, risk difference, gamma, other measures of association is conducted using the same approach that we have illustrated here with θ.

Fisher's Exact Test

In addition to the Poisson, multinomial and product multinomial sampling models, occasionally a two-way contingency table arises in such a way so that both sets of margins are fixed. In this case the appropriate sampling is based on the **hypergeometric distribution**.

Recall the hypergeometric: Consider a population of n red and black balls, of which n_1 are red and $n_2 = n - n_1$ are black. Suppose a random sample of size m is drawn from the population. Let X be the number of red balls in the sample. Then X is a random variable whose possible values are $0, 1, 2, \ldots, m$. The probability function of X is given by

$$\Pr(X = x) = \frac{\binom{n_1}{x}\binom{n-n_1}{m-x}}{\binom{n}{m}}$$

This situation can be summarized by a 2×2 table:

		Sampled: Yes	Sampled: No	
Ball Color	Red	X	$n_1 - X$	n_1
	Black	$m - X$	$n_2 - (m - X)$	n_2
		m	$n - m$	n

- Notice here that the margins of the table are fixed by the population distribution of red vs. black (row margins) and the sample size m (column margins).

Therefore, in a 2×2 table like so:

n_{11}	n_{12}	n_{1+}
n_{21}	n_{22}	n_{2+}
n_{+1}	n_{+2}	n

with fixed row and column margins, n_{11} (any cell count actually) follows a hypergeometric distribution with probability function

$$\frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}}, \quad n_{11} = m_-, \ldots, m_+$$

where

$$m_- = max(0, n_{1+} + n_{+1} - n), \qquad m_+ = min(n_{+1}, n_{1+}).$$

($n_{11} \geq n_{1+} + n_{+1} - n$ iff $n \geq n_{1+} + n_{+1} - n_{11}$, so this condition just ensures that n_{11} is not so small that it forces n_{22} to be negative to satisfy the fixed margins.)

Example — Fisher's Tea Taster

R.A. Fisher motivated his exact test with the following, now famous, example. A British woman of Fisher's acquaintance claimed to be able to distinguish between tea with milk that was prepared by adding milk to the cup first and tea with milk that was prepared by adding tea to the cup first. To test her claim, she was given 8 cups of tea, in four of which milk was added first. Since she knew that there were four cups of each type, she made four guesses of each type. The results of this experiment appear below.

True Poured First		Guess PouredFirst: Milk	Tea	
	Milk	3	1	4
	Tea	1	3	4
		4	4	8

The question of interest here is whether or not there is a positive association between the row and column variables (between the truth and her guesses).

That is the question of interest can be expressed as a choice between the following hypotheses on the odds ratio θ:

$$H_0 : \theta = 1 \quad \text{vs.} \quad H_A : \theta > 1$$

In this one-sided alternative situation, the $p-$value is defined to be the probability of observing a result at least as extreme of more extreme in the direction of H_A computed under H_0.

Under H_0, the observed result, $\hat{\theta} = 9$ is extreme in the sense that the woman guessed surprisingly well. The probability of the observed result is

$$\Pr(\hat{\theta} = 9) = \Pr(n_{11} = 3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = 0.229$$

The observed result is that she guessed 3 out of 4 of each type correctly. The only more extreme result would be that she guessed 4 out of 4 correctly. I.e., the only more extreme result is $n_{11} = 4$. This result has probability of occurring under H_0 given by

$$\Pr(n_{11} = 4) = \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = 0.014$$

Therefore, an exact $p-$value for our test is $p = .229 + .014 = .243$.
$\Rightarrow$ there is insufficient evidence to support the woman's claim.

- Notice that the discreteness of the data allows only a few exact $p-$values to occur. In particular, the only result that would have been conclusive evidence against H_0 at $\alpha = .05$ would have been if she had been 100% accurate in her guesses.

- In the example above, the discreteness of the null distribution makes it impossible to achieve a significance level of $\alpha = 0.05$; if we test at $\alpha = 0.05$, we only reject when $n_{11} = 4$, which happens with probability 0.014 under H_0. So the type I error is .014, not 0.05, and we end up with a conservative test.

- There are several ways to try to "fix" this problem, but a simple approach to reduce the conservatism is to use the **mid-p-value** rather than the p-value. For a one-sided alternative, the mid-p-value is defined to be one half the observed data probability plus the probabilities of all results more extreme than that observed.

  - For the lady tasting tea example, the mid-p-value is

$$\frac{1}{2}\Pr(n_{11} = 3) + \Pr(n_{11} = 4) = \frac{1}{2}0.2286 + 0.014 = 0.129.$$

- Note that the mid-p-value is not guaranteed to yield Type I error rate $\leq \alpha$, but it usually does, and tends to perform much better in practice than the ordinary p-value for highly discrete distributions.

- See also teadrinker.sas.

In the previous example both margins were fixed by design, providing a clear justification for Fisher's exact test.

More generally, however, Fisher's exact test is often used under multinomial, product multinomial, and Poisson sampling models by basing the significance of a test of association on the conditional distribution of the test statistic given both of the margins rather than on the unconditional distribution of the test statistic.

It is straight-forward to show that if we assume any one of the three usual sampling models and then condition on both margins, we obtain the hypergeometric distribution for the conditional distribution of $\{n_{ij}\}$.

Why would we want to use the conditional distribution rather than the unconditional distribution?

1. Other methods for testing association we have discussed (e.g., X^2, G^2 tests of independence) are asymptotic, and the accuracy of the χ^2 approximations to the distributions of these test statistics may be poor when table contains small cell counts.

   $\Rightarrow$ want an exact distributional result rather than a large-smaple approximation in small cell count situations.

2. The exact (unconditional) distribution is desirable, but it is usually unavailable because it depends upon unknown nuisance parameters.

   – A standard way of eliminating nuisance parameters in statistical inference is to condition on sufficient statistics for them. In two-way tables this leads to conditioning on the margins.

3. Basing inference on the conditional distribution rather than the marginal distribution typically involves some loss of information. However, often this loss of information is small (margins contain little information about the association in a table) and this disadvantage is the lesser of two evils compared with using an inaccurate asymptotic approximation.

Fisher's exact test generalizes from the 2×2 case to the $I \times J$ case. In the $I \times J$ case, the conditional distribution of the cell counts $\{n_{ij}\}$ given the cell margins is the multiple hypergeometric distribution:

$$\frac{\left(\prod_i n_{i+}!\right)\left(\prod_j n_{+j}!\right)}{n!\prod_i\prod_j n_{ij}!}$$

and $p-$values can be obtained by summing multiple hypergeometric probabilities for tables at least as extreme as the one obtained.

- We will discuss conditional vs. unconditional inference including the topic of exact inference more thoroughly later in the course.

Statistical Modelling

Questions:

(i.) What is the purpose of the analysis?

- Is it probabilistic or descriptive?

  Descriptive – methods in which probability models don't explicitly enter into the analysis

  – Tabulations of means, quantiles, etc.
  – Graphical representations (histograms, boxplots, scatterplots, etc.).

(ii.) Which is the response variable and which the explanatory variable? Is such a distinction appropriate?

- Are we trying to predict?
- Are we trying to describe dependence of Y('s) on X('s)?
- Are we trying to describe interrelationships among Y('s)?

(iii.) Are outliers present?

- graphical techniques, influence diagnostics (residuals, leverages, Cook's distance, etc.)
- can outliers be omitted?
- analyze with and without outliers?

(iv.) Are the observations independent?

- How were the data collected? (through time? clustering? repeated measures?)
- unaccounted for correlation among observations can lead to invalid estimates of error and incorrect inferences.

(v.) Is transformation of the data appropriate?

- Under given scale do model assumptions hold? (E.g., constant variance? normality? additive error?)

(vi.) How complex is the problem?

- often the problem appear most complex initially but simplifies after further inspection
- very often its useful to think about the problem in terms of simplest (STAT6210, STAT6220) methods first.

(vii.) Is more than one source of variability present?

- random effects vs. fixed effects?
  (are factor levels observed of interest in themselves or as a representative random sample from some population of levels of interest?)

Models

- Often data can be thought of as **signal** and **noise.**

  Transmission and reception of information (e.g., radio) involves a message (signal) which is distorted by static (noise, or error).

  signal – deterministic
  noise – random

- A **statistical model** of the data accounts for both signal and noise

  - (the latter makes it probabilistic).

- Often the signal and noise metaphor is right on the mark. In other cases the signal component is a mathematical description of the main features, noise component is "everything else" (unexplained component).

- Many models (parametric models) involve unknown constants called **parameters**.

  - In such cases, "Fitting the model" amounts to estimating the parameters of the model.

Models should be

1. As simple as conveniently possible (**parsimony**)

2. Valid over an appropriate range of conditions (**scope,** or generalizability)

3. Consistent with the data.

- Items 1–3 above are often opposed to one another
  - By including enough parameters, we can make a model fit the data as well as we please, but we sacrifice parsimony, generalizability.

Generalized Linear Models (GLMs) are the main subject of this course. They are generalizations of:

Classical Linear Models (CLMs)

Y: Response Variable

$X_1, \ldots, X_p$: Explanatory variables.

We typically observe n independent copies (a random sample):

$$\begin{matrix} Y_i \\ X_{i1}, \ldots, X_{ip}, \end{matrix} \quad i = 1, \ldots, n \quad \text{the random variables}$$

$$\begin{matrix} y_i \\ x_{i1}, \ldots, x_{ip}, \end{matrix} \quad i = 1, \ldots, n \quad \text{the observations}$$

In the CLM (and most other "regression" contexts, including GLMs), the explanatory variables are fixed by the design governing the data collection, or, if not fixed by design, the explanatory variables are conditioned on. That is, we seek to describe, or make inference about the conditional distribution of the response *given that the explanatory variables are equal to their observed values.*

- Thus in all that follows (unless explicitly stated otherwise) the explanatory variables can be regarded as non-random.

Model:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n,$$

where e_i is a random variable with

1. $\mathrm{E}(e_i) = 0$ for all i.
2. $\mathrm{var}(e_i) = \sigma^2$ (constant) for all i.
3. $e_1, \ldots, e_n$ are i.i.d.

- Our model says that $\mathrm{E}(Y_i) = \mu_i$ is a linear function of the β's (for all i).

Matrix Notation:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{12} & \cdots & X_{1p} \\ 1 & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & \cdots & X_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

X is known as the **design matrix** or **model matrix.**

Most common examples:

1. Regression
2. Analysis of variance (fixed effects only).
3. Analysis of covariance.

Benefits of CLMs

1. Model is very flexible.
2. Easy to fit.

   OLS estimate of $\boldsymbol{\beta}$ (denoted $\hat{\boldsymbol{\beta}}$) can be found by solving

   $$(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y} \qquad \text{(normal equations)}$$

   If $(\mathbf{X}^T\mathbf{X})^{-1}$ exists, $\hat{\boldsymbol{\beta}}$ is given by

   $$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

   which is the B.L.U.E. according to the Gauss-Markov Theorem (irrespective of the distribution of the e_i's).
3. Model is relatively easy to interpret.

Limitations of CLMs

1. $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ is unlimited in range, but in many problems the range is restricted. E.g., when $\mathbf{Y}$ is a vector of binary responses, $\boldsymbol{\mu}$ is a vector of probabilities which should be in $[0, 1]$.
2. Most inference (C.I.'s, hypothesis tests) in CLMs assumes a normal distribution for e_i's.
3. Additive error.

GLMs (Nelder and Wedderburn, 1972, *JRSSB*) solve problems 1–3.

Examples of GLMs:

a. Suppose we have Bernoulli responses: $Y_i = \begin{cases} 0 \\ 1 \end{cases}$.

$$\Rightarrow \qquad \begin{array}{c} \mu_i = \mathrm{E}(Y_i) = \Pr(Y_i = 1) \in [0,1] \\ \mathrm{var}(Y_i) = \mu_i(1-\mu_i) \text{ (not constant)} \end{array}$$

Suppose we have a covariate X (e.g., dose level) and response Y (e.g., live(0)/die(1)). We'd like to model the mean response (probability of death) as a function of dose.

One possibility: A CLM with $\mu = \beta_1 + \beta_2 X$.

With this linear model we run into problem 1:

For especially large or small values, we will be predicting μ to be outside of $[0,1]$.

Solution: use a nonlinear model where we map $\beta_1 + \beta_2 X$ into $[0,1]$.

E.g., $\mu = F(\beta_1 + \beta_2 X)$ where F is some c.d.f.

Choices:

1. $F = \Phi$ (standard normal c.d.f.)

$$\Rightarrow \qquad \mu = \Phi(\beta_1 + \beta_2 X)$$

$$\Rightarrow \qquad \Phi^{-1}(\mu) = \beta_1 + \beta_2 X$$

which is known as the **Probit Model**.

- Notice that this model is nonlinear in a very specific way. We map a linear function of the β's ($\beta_1 + \beta_2 X$) into the permissible range of μ via a nonlinear function (in this case Φ).

- We call the linear function of the β's the **linear predictor** of the GLM and it is denoted η (here, $\eta = \beta_1 + \beta_2 X$).

- We call the function that relates μ to η the **link function** (here it is Φ^{-1}).

2. The unit logistic c.d.f.

   The *logistic distribution* has density function

$$f_Y(y; \mu, \sigma) = \frac{\exp\{(y-\mu)/\sigma\}}{\sigma\{1 + \exp\{(y-\mu)/\sigma\}\}^2}$$

   for $-\infty < y < \infty$, where μ, σ are location and scale parameters, respectively.

   The unit logistic distribution has $\mu = 0$, $\sigma = 1$, and has c.d.f. $F(x) = \frac{e^x}{1+e^x}$. This gives

$$\mu = \frac{\exp(\beta_1 + \beta_2 X)}{1 + \exp(\beta_1 + \beta_2 X)}.$$

   Solving for $\eta = \beta_1 + \beta_2 X$, we get

$$\underbrace{\log\left(\frac{\mu}{1-\mu}\right)}_{\text{“logit” link}} = \beta_1 + \beta_2 X = \eta$$

- This model is called a **logistic regression model.**

Note that here and in the previous probit example we haven't fully specify the model yet, just the "signal" part. Our models must also account for noise as well.

- In the CLM we do this by specifying a distribution for an error term. This is an indirect way of implying a distribution for Y.

- In a GLM we drop the device of an error term and specify the distribution for Y directly.

Both the probit and logistic models are completed by specifying the **error distribution** to be Bernoulli.

b. Another example for 0,1 responses:

Suppose that the random variable of interest, Z, is Poisson with mean λ. However, we don't observe Z, only whether or not $Z = 0$. That is, we observe

$$Y = \begin{cases} 0, & \text{if } Z = 0 \\ 1, & \text{if } Z > 0 \end{cases}$$

- For example, Z might be the number of infected samples, but because several samples were combined before presence/absence of infection was measured, we only observe Y.

If we had $Z \sim \text{Poisson}(\lambda)$, we might model λ with a loglinear model relating λ to some linear predictor η involving relevant covariates:

$$\log(\lambda) = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p \qquad (*)$$

- Why loglinear? Because with $\log(\lambda) = \eta$ or, equivalently, $\lambda = e^{\eta}$ we map the unconstrained η to $[0, \infty)$ with the exponential function.

We don't have Z, though, we only observe Y. What do we know about Y?:

$$\mu = \text{E}(Y) = \Pr(Y = 1) = 1 - \underbrace{\Pr(Z = 0)}_{\text{a Poisson Probability}} = 1 - e^{-\lambda}$$

Solving for λ,

$$-\log(1 - \mu) = \lambda$$

or

$$\log(-\log(1 - \mu)) = \log(\lambda)$$

Adding (*) we get

$$\underbrace{\log(-\log(1 - \mu))}_{\text{complementary log-log link}} = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p = \eta$$

- This model comes up in **Dilution Bioassay**.

Example

- Suppose that there is an unknown concentration ρ_0 of an organism in solution.
- To determine ρ_0 we proceed by diluting the solution in powers of 2, so that the x^{th} dilution has concentration

$$\rho_x = \frac{\rho_0}{2^x}, \qquad x = 1, 2, \ldots, n.$$

- After each dilution we examine a plate treated with the solution. If plate is streaked $\Rightarrow$ at least one organism is present.

In this example there is an underlying random variable Z which is a Poisson count of the number of species present that varies with x so that the Poisson mean depends on x. Instead of observing Z, though, we observe

$$Y = \begin{cases} 0, & Z = 0 \text{ (plate is not streaked)} \\ 1, & Z \geq 1 \text{ (plate is streaked)} \end{cases}$$

Let $\mu = \Pr(\text{plate is streaked})$. Then

$$\mu = \Pr(Y = 1) = 1 - \Pr(Z = 0) = 1 - e^{-\rho_x}$$

$$\Rightarrow \qquad \log(-\log(1 - \mu)) = \log(\rho_x) = \underbrace{\log(\rho_0)}_{\beta_1} - \underbrace{(\log 2)}_{\beta_2} x = \eta$$

- Therefore we can estimate ρ_0 as $e^{\hat{\beta}_1}$ where $\hat{\beta}_1$ is the intercept from a simple regression of Y on X using a complementary log-log link and Bernoulli error distribution.

3. Log-linear Models

Suppose we measure two categorical variables, one with r levels and the other with c levels, on each of n subjects where n is fixed (non-random).

We may summarize the resulting data (n bivariate responses) in a two-way **contingency table** or **cross-classification**:

Here, Y_{ij} = the count of how many of the n subjects responded with category i on the first variable and category j on the second variable.

- With n fixed, it is reasonable to assume that the Y_{ij}'s are multinomial with probabilities π_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, c$, and number of trials n. In this case

$$\mathrm{E}(Y_{ij}) = n\pi_{ij}$$

Under an assumption that the two variables are independent, $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ where $\pi_{i\cdot}$ is the marginal probability of responding with category i to variable 1 and $\pi_{\cdot j}$ is the marginal probability of responding with category j to variable 2.

$$\Rightarrow \qquad \mu_{ij} = \mathrm{E}(Y_{ij}) = n\pi_{i\cdot}\pi_{\cdot j}$$

$$\Rightarrow \qquad \log(\mu_{ij}) = \log n + \log \pi_{i\cdot} + \log \pi_{\cdot j} = \eta_{ij}$$

We can rewrite this model in the GLM form as

$$\log \mu_{ij} = \beta_1 + \beta_2 1_{\{i=1\}} + \beta_3 1_{\{i=2\}} + \cdots + \beta_{r+1} 1_{\{i=r\}} + \beta_{r+2} 1_{\{j=1\}} + \cdots + \beta_{r+c+1} 1_{\{j=c\}}$$

and we can fit the model with log link and multinomial error distribution.

- We'll see later that we can obtain the same fit by using the Poisson distribution in place of the multinomial as the error distribution.

Components of a GLM

In a CLM we assume:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n$$
$$e_1, \ldots, e_n \overset{iid}{\sim} 0, \sigma^2 (N(0, \sigma^2), \text{ if we want inference, ML})$$

Here there are three parts:

Sytematic Component: $\eta_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$ or $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ in matrix notation.

Random Component: $Y_1, \ldots, Y_n \overset{ind}{\sim} N(\mu_i, \sigma^2)$ (usually assumed through the e_i's)

In addition we assume $\mu_i = \eta_i \ \forall i$.

More generally, in GLMs we have

1. Systematic Component:
$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

2. Random Component: Y_i's are independent r.v.'s each with $\mathrm{E}(Y_i) = \mu_i$ and each with density
$$f_Y(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}, \qquad (*)$$

   Here, ϕ is a (constant) scale parameter (typically a nuisance param.) and θ_i is a location parameter (typically of interest) and can be expressed as some function of the mean, μ_i.

- (*) denotes the Exponential Dispersion (E.D.) Family of distributions
  - A generalization of the linear exponential family.

How generalized?

- If ϕ is known, then (*) is the linear exponential family with canonical parameter θ_i.
- If ϕ is unknown, (*) may or may not be a 2-parameter exponential family.

In GLMs there is a third component:

3. The **link** between the R.C. and the S.C.:

$$g(\mu_i) = \eta_i$$

All we require of g is that it be one-to-one and differentiable.

Examples of E.D. Family Distributions:

1. Normal Distribution:

$$\begin{aligned} f(y_i; \mu_i, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(y_i - \mu_i)^2}{2\sigma^2} \right\} \\ &= \exp\left\{ \left[y_i\mu_i - \frac{\mu_i^2}{2} \right] \frac{1}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right\} \end{aligned}$$

Here, $\theta_i = \mu_i$, $\phi = \sigma^2$, $b(\theta_i) = \theta_i^2/2$, $a_i(\phi) = \phi$, and $c(y_i, \phi) = -[y_i^2/\phi + \log(2\pi\phi)]/2$.

- Notice, that (for example) $b(\theta_i)$ is a function of θ_i, so we notice $b(\theta_i) = \mu_i^2/2$ and then make it a function of its argument by using the relationship between θ_i and μ_i.

2. Gamma Distribution

$$f(y_i;\mu_i,\nu) = \left(\frac{\nu}{\mu_i}\right)^{\nu} \frac{y_i^{\nu-1}e^{-\nu y_i/\mu_i}}{\Gamma(\nu)}$$
$$= \exp\left\{\left[-\frac{y_i}{\mu_i} - \log\mu_i\right]\nu + (\nu-1)\log(y_i) + \nu\log\nu - \log\Gamma(\nu)\right\}$$

Here, $\theta_i = -\mu_i^{-1}$, $b(\theta_i) = \log(-\theta_i)$, $\phi = \nu$, $a_i(\phi) = \phi^{-1}$ and $c(y_i;\phi) = (\phi-1)\log y_i + \phi\log\phi - \log\Gamma(\phi)$.

3. Poisson Distribution

$$f(y_i;\mu_i) = \frac{\mu_i^{y_i}e^{-\mu_i}}{y_i!}$$
$$= \exp\{y_i\log\mu_i - \mu_i - \log(y_i!)\}$$

Here, $\theta_i = \log\mu_i$. $b(\theta_i) = e^{\theta_i}$, $\phi = 1$, $a_i(\phi) = \phi$, and $c(y_i,\phi) = -\log(y_i!)$.

4. Binomial Distribution

$$f(y_i;\pi_i) = \binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}$$
$$= \exp\left\{y_i\log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i\log(1-\pi_i) + \log\binom{n_i}{y_i}\right\}$$

Here, $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, $b(\theta_i) = n_i\log(1+e_i^{\theta})$, $\phi = 1$, $a_i(\phi) = \phi$, and $c(y_i,\phi) = \log\binom{n_i}{y_i}$.

Mean and Variance in E.D. Family

In the exponential dispersion family a special relationship exists between the mean and variance. To demonstrate this relationship we first need to review some results about likelihood and score functions.

Likelihood and Score Functions

Consider a single observation of a random variable Y which has density $f_Y(y;\theta)$ which involves the scalar parameter θ.

Likelihood function: $L(\theta;y) = f(y;\theta)$

Log-likelihood function: $\ell(\theta;y) = \log L(\theta;y)$

The first derivative of ℓ with respect to θ,

$$U = U(\theta) = \frac{\partial \ell(\theta;y)}{\partial \theta} = \frac{\partial f(y;\theta)/\partial\theta}{f(y;\theta)}$$

is called the **score** function (or, sometimes, the **efficient score** function).

Results:*

1. $\mathrm{E}(U) = 0$
2. $\mathrm{var}(U) = E(U^2) = -\mathrm{E}(\partial U/\partial\theta)$.

Proofs:

1.

$$\begin{aligned}
\mathrm{E}(U) &= \int \frac{\partial \ell(\theta;y)}{\partial\theta} f(y;\theta)dy, \quad \text{def. of exp. val.} \\
&= \int \frac{\partial f(y;\theta)/\partial\theta}{f(y;\theta)} f(y;\theta)dy = \int \frac{\partial f(y;\theta)}{\partial\theta}dy \\
&\overset{*}{=} \frac{\partial}{\partial\theta}\int f(y;\theta)dy \quad \text{assuming reg. conditions to switch diff. and int.} \\
&= \frac{\partial}{\partial\theta}1 = 0
\end{aligned}$$

---

* Under mild regularity conditions that hold for E.D. family

2. $\text{var}(U) = \text{E}(U^2) - \underbrace{[\text{E}(U)]^2}_{=0} = \text{E}(U^2)$. In addition,

$$\frac{\partial U}{\partial \theta} = \frac{\partial^2 \ell(\theta; y)}{\partial \theta^2} = \frac{\frac{\partial^2 f(y;\theta)}{\partial \theta^2} f(y;\theta) - \left(\frac{\partial f(y;\theta)}{\partial \theta}\right)^2}{[f(y;\theta)]^2}$$

So, using the definition of expected value,

$$\begin{aligned}
-\text{E}\left(\frac{\partial U}{\partial \theta}\right) &= -\int \left(\frac{\frac{\partial^2 f(y;\theta)}{\partial \theta^2} f(y;\theta) - \left(\frac{\partial f(y;\theta)}{\partial \theta}\right)^2}{[f(y;\theta)]^2} \right) f(y;\theta) dy \\
&= -\int \frac{\partial^2 f(y;\theta)}{\partial \theta^2} dy + \int \frac{(\partial f(y;\theta)/\partial \theta)^2}{f(y;\theta)} dy \\
&\overset{*}{=} -\underbrace{\frac{\partial^2}{\partial \theta^2} \int f(y;\theta) dy}_{=0} + \int \underbrace{\left[\frac{\partial f(y;\theta)/\partial \theta}{f(y;\theta)}\right]^2}_{=U^2} f(y;\theta) dy \\
&= E(U^2)
\end{aligned}$$

So we've shown that $\text{var}(U) = -\text{E}\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) \equiv I(\theta)$, which is known as the **Fisher information** about θ contained in y, or, more simply, the information.

Example – $N(\mu, \sigma^2)$**,** μ **unknown,** σ^2 **known:**

$$f(y;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(y-\mu)^2}{2\sigma^2} \right\}$$

$$\Rightarrow \quad \ell(\mu; y) = \frac{y\mu - \mu^2/2}{\sigma^2} + \text{stuff not involving } \mu$$

It follows that

$$U = \frac{\partial \ell(\mu; y)}{\partial \mu} = (y - \mu)/\sigma^2$$

$$\Rightarrow \quad \mathrm{E}(U) = (\mathrm{E}(y) - \mu)/\sigma^2 = (\mu - \mu)/\sigma^2 = 0$$

and

$$I(\mu) = -\mathrm{E}\left(\frac{\partial^2 \ell}{\partial \mu^2}\right) = -\mathrm{E}\left(\frac{\partial}{\partial \mu}\left(\frac{y-\mu}{\sigma^2}\right)\right) = -\mathrm{E}\left(-\frac{1}{\sigma^2}\right) = \frac{1}{\sigma^2}$$

and

$$\mathrm{var}(U) = \mathrm{var}\left(\frac{y-\mu}{\sigma^2}\right) = \mathrm{var}\left(\frac{y}{\sigma^2}\right) = \frac{1}{\sigma^4}\mathrm{var}(y) = \frac{1}{\sigma^2}$$

Now we'll use results 1 and 2 about score functions to reveal the relationship between the mean and variance in E.D. family.

For E.D. family,

$$\ell(\theta_i, \phi; y_i) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)$$
$$\Rightarrow \qquad U_i = \frac{\partial \ell}{\partial \theta_i} = \frac{y_i - \partial b(\theta_i)/\partial\theta_i}{a_i(\phi)}$$

$\mathrm{E}(U_i) = 0$ then implies $\mathrm{E}(y_i) = \partial b(\theta_i)/\partial\theta_i$ or

$$\mu_i = b'(\theta_i)$$

In addition,

$$\frac{\partial U_i}{\partial \theta_i} = -\frac{\partial^2 b(\theta_i)/\partial\theta_i^2}{a_i(\phi)} \quad \text{and} \quad \mathrm{var}(U_i) = \frac{\mathrm{var}(y_i)}{a_i^2(\phi)}$$

So, $\mathrm{var}(U_i) = -\mathrm{E}\left(\frac{\partial U_i}{\partial \theta_i}\right)$ implies

$$\frac{\mathrm{var}(y_i)}{a_i^2(\phi)} = \frac{\partial^2 b(\theta_i)/\partial\theta_i^2}{a_i(\phi)}$$
$$\Rightarrow \qquad \mathrm{var}(y_i) = \frac{\partial^2 b(\theta_i)}{\partial\theta_i^2} a_i(\phi) = \left(\frac{\partial}{\partial\theta_i}\mu_i\right) a_i(\phi)$$
$$\Rightarrow \qquad \mathrm{var}(y_i) = v(\mu_i) a_i(\phi)$$

where $v(\mu_i) = \partial\mu_i/\partial\theta_i = b''(\theta_i)$ is known as the **variance function**.

- We've established that in a GLM, the E.D. family distributional assumption implies that the response variance is a function of the response mean.
  - Note that $\mathrm{var}(y_i)$ may be a function of the mean in a trivial way, as in the Normal distribution where $v(\mu) = 1$, $\phi = \sigma^2$, $a_i(\phi) = \phi$ so $\mathrm{var}(y_i) = (1)\sigma^2$.

In most GLMs, $a_i(\phi)$ is of the form

$$a_i(\phi) = \phi/w_i$$

where w_i is a known prior weight specific to the i^{th} observation.

- In what follows, **we will restrict attention to this special case.**
- Notice that $a_i(\phi) = \phi/w_i$ implies

$$\text{var}(y_i) = \phi v(\mu_i)/w_i$$

- We will sometimes use the notation

$$y \sim ED(\mu, \phi, w)$$

to denote that y has an E.D. distribution with mean parameter $\mu = \mu(\theta)$, scale parameter ϕ and known weight w. (Note that only μ and ϕ are unknown parameters).

Example – Binomial Distribution:

Recall (p.66) that we were able to write the binomial frequency function in the following form:

$$f(y_i; \pi_i) = \exp\left\{ y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i} \right\} \quad (*)$$

and we said that this was an example of an E.D. distribution with $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, $b(\theta_i) = n_i \log(1+e_i^{\theta})$, $\phi = 1$, $a_i(\phi) = \phi$, and $c(y_i, \phi) = \log\binom{n_i}{y_i}$.

- This is correct, but it is slightly inconvenient for $b(\theta_i)$ to depend upon n_i as well as θ_i.
  - Notice that this inconvenience disappears when $n_i = 1$, that is when y_i is Bernoulli.

- When $n_i > 1$, another way to think about the binomial distribution's relationship to the exponential dispersion family is not $y_i \sim ED(\mu_i, \phi = 1, w_i = 1)$ where $b(\theta_i) = n_i \log(1 + e_i^\theta)$ and $\mu_i = b'(\theta_i) = n_i e^{\theta_i}/(1 + e^{\theta_i}) = n_i \pi_i$, but instead $y_i/n_i \sim ED(\mu_i, \phi = 1, w_i = n_i)$, where $b(\theta_i) = \log(1 + e_i^\theta)$ and $\mu_i = b'(\theta_i) = e^{\theta_i}/(1 + e^{\theta_i}) = \pi_i$.

  Notice we can rewrite the RHS of (*) as

$$\exp\left\{\left[\frac{y_i}{n_i}\log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)\right] n_i + \log\binom{n_i}{y_i}\right\}$$

  where now $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, $b(\theta_i) = \log(1 + e_i^\theta)$, $\phi = 1$, $w_i = n_i$, $a_i(\phi) = \phi/w_i$, and $c(y_i, \phi) = \log\binom{n_i}{y_i}$.

Variance Functions:

Distribution	Variance Function
Poisson	$v(\mu) = \mu = e^\theta$
Binomial	$v(\mu) = n\mu(1-\mu) = ne^\theta/(1+e^\theta)^2$
Binomial/n	$v(\mu) = \mu(1-\mu) = e^\theta/(1+e^\theta)^2$
Normal	$v(\mu) = 1$
Gamma	$v(\mu) = \mu^2 = (-1/\theta)^2$

Example – Binomial Distribution

Let Y = the number of successes out of n trials, so $Y \sim Bin(n, \pi)$, where π = probability of success on each trial, $\mu = \mathrm{E}(Y)$. What is μ?

From ED distributional form,

$$\theta = \log\left(\frac{\pi}{1-\pi}\right) \quad \Rightarrow \quad \pi = \frac{e^\theta}{1+e^\theta}$$

and $b(\theta) = n\log(1+e^\theta)$, $a(\phi) = \phi = 1$ so

$$\mu = b'(\theta) = n\frac{e^\theta}{1+e^\theta} = n\pi \quad \surd$$

and

$$\mathrm{var}(Y) = a(\phi)b''(\theta) = n\frac{e^\theta}{1+e^\theta}\frac{1}{1+e^\theta} = n\pi(1-\pi) \quad \surd$$

Another Example

Now let $P = Y/n$ = the proportion of successes, so $P \sim (1/n)Bin(n, \pi)$. Then

$$f(p;\pi) = \Pr(P = p|n,\pi) = \Pr(Y = np|n,\pi) = \binom{n}{np}\pi^{np}(1-\pi)^{n-np}$$
$$= \exp\left\{np\log\left(\frac{\pi_i}{1-\pi_i}\right) + n\log(1-\pi) + \log\binom{n_i}{y_i}\right\}$$

where $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, $b(\theta) = \log(1-\pi) = \log(1+e^\theta)$, and $a(\phi) = \phi/w = 1/n$. So,

$$\mu = \mathrm{E}(P) = b'(\theta) = \frac{e^\theta}{1+e^\theta} = \pi, \quad \surd$$

and

$$\mathrm{var}(P) = a(\phi)b''(\theta) = \frac{1}{n}\pi(1-\pi) \quad \surd$$

Link Functions:

When considering models for binary random variables, we considered inverse c.d.f. links. *Why?*

For a binary random variable Y we want to map the real line (range of $\eta = \mathbf{x}^T\boldsymbol{\beta}$) to $[0,1]$, the range of $\mu = \mathrm{E}(Y)$ (a probability).

- All c.d.f.'s do just that.

So, it made sense to consider models of the form

$$\mu = F(\eta)$$

for some c.d.f. F, or equivalently,

$$F^{-1}(\mu) = \eta$$

1. Logit link (inverse unit logist c.d.f.) $\log\left(\frac{\mu}{1-\mu}\right)$
2. Probit link (inverse std. normal c.d.f.) $\Phi^{-1}(\mu) = \eta$.
3. Complementary log-log link $\log(-\log(1-\mu)) = \eta$ where μ = probability of success. This model is equivalent to the model with log-log link, $\log(-\log(\mu)) = \eta$ where now μ = probability of failure. The log-log link is the inverse c.d.f. for the extreme value or Gumbel distribution.

For count data the inverse link function should again map the range of η $(-\infty, +\infty)$ to the range of μ which is now $[0, \infty)$. A natural choice is the exponential function:

$$\mu = e^{\eta}, \quad \text{or, equivalently,} \quad \log(\mu) = \eta$$

For each error distribution a variety of link functions are possible beyond those mentioned above. However, for each error distribution there is one link that is particularly convenient: the **canonical link** function.

- Canonical links are the links for which $\theta = \eta$.
- For canonical links there exists a sufficient statistic equal in dimension to $\boldsymbol{\beta}$.

Both of these properties make models with canonical links mathematically convenient.

Canonical Links:

Error Distribution	Canonical Link
Normal	$g(\mu) = \mu$
Poisson	$g(\mu) = \log \mu$
Binomial/n	$g(\mu) = \log(\mu/(1-\mu))$
Gamma	$g(\mu) = \mu^{-1}$

- For canonical links the sufficient statistic is $\underbrace{\mathbf{X}^T}_{p\times n}\underbrace{\mathbf{Y}}_{n\times 1}$ a $p \times 1$ vector with components $\sum_{i=1}^{n} x_{ij} Y_i$, $j = 1, \ldots, p$.
- Canonical links are convenient, but that alone is not enough to dictate their use. Choice of link should be made on

1. model fit
2. model interpretability
3. whether or not link is canonical (convenience)

Grouped vs. Ungrouped Data

In the CLM setting we typically consider ungrouped data where $(y_i, \mathbf{x}_i)$ is the original observation on unit i

$$\begin{matrix}\text{Unit } 1\\ \vdots\\ \text{Unit } i\,,\\ \vdots\\ \text{Unit } n\end{matrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

- For example, deaths in a clinical trial: we observe the binary (0=dead,1=alive) variables $y_1, \ldots, y_n$ on the n subjects participating in the trial.

If some of the rows of the $\mathbf{X}$ matrix are identical, it is often convenient to group the data.

In this case index i refers to the i^{th} group of units with the same covariate vector. We also record n_i = the number of units in the i^{th} group.

$$\begin{matrix}\text{Group } 1\\ \vdots\\ \text{Group } i\,,\\ \vdots\\ \text{Group } g\end{matrix} \quad \begin{bmatrix} n_1 \\ \vdots \\ n_i \\ \vdots \\ n_g \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_i \\ \vdots \\ \bar{y}_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots \\ x_{g1} & \cdots & x_{gp} \end{bmatrix}$$

- Here we analyze $\bar{y}_i$ the i^{th} group mean (proportion dead, in example).
  - Alternatively, we could analyze the i^{th} group sum (number dead, in example).

If ungrouped data $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, are modelled with a GLM with

$$\mathrm{E}(y_i) = \mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}), \quad \mathrm{var}(y_i) = \phi v(\mu_i) \quad (w_i = 1)$$

then the group averages are also in E.D. family so an equivalent GLM models the group averages with the same error distribution and

$$\begin{aligned} \mathrm{E}(y_i) &= g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}) \quad \text{(same mean)} \\ \text{and} \quad \mathrm{var}(y_i) &= \phi v(\mu_i)/n_i \quad (w_i = n_i) \end{aligned}$$

Why look at grouped data?

1. Some statistical procedures are valid only for grouped data.
    - E.g., goodness of fit tests like the deviance, Pearson chi-square statistics, inappropriate for binary and count data that can't be grouped.

2. Two types of asymptotics are available for grouped data: i.) asymptotics as $n_i \to \infty \quad \forall i = 1, \ldots, g$ or ii.) asymptotics as $n \to \infty$ without requiring $n_i \to \infty \quad \forall i$. In ungrouped data asymptotics are necessarily as $n \to \infty$.

3. Computing and memory savings can be substantial with grouped data.

Methods of Estimation in GLMs

1. Least Squares (OLS, WLS/GLS)
2. Maximum Likelihood (ML)
3. Conditional ML
4. Quasi-likelihood (QL)
5. Estimating Functions
6. Other methods – Extended QL, Pseudo ML, Pseudo likelihood, etc.

1. Least Squares (OLS, WLS, GLS):

OLS

For a model of the form

$$Y_i = h_i(\beta_1, \ldots, \beta_p) + e_i, \quad i = 1, \ldots, n,$$

where h_i's are known functions of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathcal{B}$ and $e_1, \ldots, e_n$ satisty

i. $E(e_i) = 0$
ii. $\text{var}(e_i) = \sigma^2 > 0$ (constant)
iii. $\text{cov}(e_i, e_j) = 0$ for $i \neq j$.

we would like to choose $\boldsymbol{\beta}$ so that $h_i(\boldsymbol{\beta})$ is close to Y_i for all i.

One approach: Choose $\boldsymbol{\beta}$ to minimize the sum of squared deviations between the Y_i's and the h_i's.

$$\text{Let} \quad S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - \underbrace{h_i(\boldsymbol{\beta})}_{=\mu_i})^2 \quad \text{(Least Squares Criterion)}$$

Estimate $\boldsymbol{\beta}$ by minimizing $S(\boldsymbol{\beta})$.

If the h_i's are differentiable and $\mathcal{B}$ is open, then $\hat{\boldsymbol{\beta}}$ must satisfy

$$\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \beta_j} = 0, \quad j = 1, \ldots, p,$$

which are called the **normal equations.**

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_j} = -2 \sum_i (Y_i - h_i(\boldsymbol{\beta})) \frac{\partial h_i(\boldsymbol{\beta})}{\partial \beta_j}$$

WLS

- The assumption of constant variance can be relaxed. This leads to **weighted least squares (WLS)**.

  Assume
  $$Y_i = h_i(\boldsymbol{\beta}) + \tilde{e}_i, \quad i = 1, \ldots, n, \qquad (*)$$
  where now $\tilde{e}_i$'s satisfy

  i. $\mathrm{E}(\tilde{e}_i) = 0$
  ii. $\mathrm{var}(\tilde{e}_i) = \sigma^2/w_i$ (no longer constant, but instead changing with a known weight w_i).
  iii. $\mathrm{cov}(\tilde{e}_i, \tilde{e}_j) = 0$ for $i \neq j$.

  Model $(*)$ does not fit into the OLS framework. However, notice that the following transformed model does:
  $$\sqrt{w_i}Y_i = \sqrt{w_i}h_i(\boldsymbol{\beta}) + \underbrace{\sqrt{w_i}\tilde{e}_i}_{=e_i}$$

  This model has an error term e_i which satisfies the OLS assumptions. Therefore we estimate $\boldsymbol{\beta}$ by minimizing
  $$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_i (\sqrt{w_i}Y_i - \sqrt{w_i}h_i(\boldsymbol{\beta}))^2 \\ &= \sum_i w_i(Y_i - h_i(\boldsymbol{\beta}))^2 \quad \text{(the Weighted LS Criterion)} \end{aligned}$$

  In matrix notation, we want
  $$\arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{h}(\boldsymbol{\beta}))^T \mathbf{W} (\mathbf{Y} - \mathbf{h}(\boldsymbol{\beta})),$$

  where
  $$\mathbf{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

  where $w_i \propto 1/\mathrm{var}(y_i)$.

GLS

More generally, suppose $\mathbf{V} = \text{var}(\mathbf{Y})$ where now $\mathbf{V}$ is not necessarily diagonal. (That is, let's now relax assumption iii.)

Using $\mathbf{V}^{-1}$ in place of $\mathbf{W}$ we obtain the **Generalized Least Squares Criterion,** and the GLS estimators of $\boldsymbol{\beta}$ are given by

$$\arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{h}(\boldsymbol{\beta}))^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{h}(\boldsymbol{\beta}))$$

- If $\mathbf{V}$ is known, then $\hat{\boldsymbol{\beta}}_{GLS}$ is BLUE in the linear version of this model (i.e., when $\mathbf{h}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$), but it is typically necessary to estimate $\mathbf{V}$ to obtain the GLS estimator, in which case it is not necessarily optimal in any sense.
- GLS is most useful in a linear model where $\mathbf{h}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$ in which case $\hat{\boldsymbol{\beta}}$ is the solution of

$$(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})\boldsymbol{\beta} = \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y} \quad \text{(the normal equations)}$$

2. Maximum Likelihood:

Suppose we have a discrete random variable Y (possibly a vector) with observed value y. Suppose Y has frequency function $f(y; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.

The **likelihood function**, $L(\boldsymbol{\theta}; y)$ is defined to equal the frequency function (more generally, the density) but viewed as a function of $\boldsymbol{\theta}$, not y:

$$L(\boldsymbol{\theta}; y) = f(y; \boldsymbol{\theta})$$

Therefore, the likelihood at $\boldsymbol{\theta}_0$, say, has the interpretation

$$\begin{aligned} L(\boldsymbol{\theta}_0; y) &= \Pr(Y = y \quad \text{when} \quad \boldsymbol{\theta} = \boldsymbol{\theta}_0) \\ &= \Pr(\text{observing the obtained data when } \boldsymbol{\theta} = \boldsymbol{\theta}_0) \end{aligned}$$

Logic of ML: choose the value of $\boldsymbol{\theta}$ that makes this probability largest $\Rightarrow \hat{\boldsymbol{\theta}}$, the MLE.

We use the same procedure when Y is continuous with density function $f(y;\boldsymbol{\theta})$: maximize $L(\boldsymbol{\theta};y) = f(y;\boldsymbol{\theta})$

Often, our data come from a random sample so that we observe $\mathbf{y}$ corresponding to $\mathbf{Y}_{n\times 1}$, a vector of independent r.v.'s. In this case

$$L(\boldsymbol{\theta};\mathbf{y}) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\theta})$$

Since its easier to work with sums than products its useful to note that in general

$$\arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta};y) = \arg\max_{\boldsymbol{\theta}} \underbrace{\log L(\boldsymbol{\theta};y)}_{\equiv \ell(\boldsymbol{\theta};y)}$$

Therefore, we define a MLE of $\boldsymbol{\theta}$ as a $\hat{\boldsymbol{\theta}}$ so that

$$\ell(\hat{\boldsymbol{\theta}},y) \geq \ell(\boldsymbol{\theta};y) \quad \forall \boldsymbol{\theta} \in \Theta$$

If Θ is an open set, then $\hat{\boldsymbol{\theta}}$ must satisfy (if it exists)

$$\frac{\partial \ell(\hat{\boldsymbol{\theta}})}{\partial \theta_j} = 0, \quad j = 1,\ldots,\dim(\theta)$$

or in vector form

$$\frac{\partial \ell(\hat{\boldsymbol{\theta}};y)}{\partial \boldsymbol{\theta}} = \mathbf{0}, \quad \text{(the } \textbf{score equation}\text{)}$$

- Difficulty: usually the score equation can't be solved explicitly – in such cases we need to use a numerical method (Newton-Raphson, Fisher-Scoring, etc.).

In the GLM context, for a single observation

$$\ell(\theta_i; y_i) = \log f(y_i; \theta_i)$$

which we can write as a function of μ_i since $\theta_i = \theta_i(\mu_i)$:

$$\ell(\mu_i; y_i) = \log(f(y_i; \theta_i(\mu_i)) = \frac{y_i\theta_i(\mu_i) - b(\theta_i(\mu_i))}{a_i(\phi)} + c(y_i; \phi)$$

For all n observations,

$$\ell(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i(\mu_i) - b(\theta_i(\mu_i))}{a_i(\phi)} + c(y_i; \phi) \right)$$

Equivalent to maximizing $\ell(\boldsymbol{\mu}; \mathbf{y})$ with respect to $\boldsymbol{\mu}$, we can minimize the **scaled deviance:**

$$D^*(\mathbf{y}; \boldsymbol{\mu}) = 2\left[\ell(\mathbf{y}; \mathbf{y}) - \ell(\boldsymbol{\mu}; \mathbf{y})\right]$$

Here, we have written the scaled deviance as a function of $\boldsymbol{\mu}$. Its minimizing value w.r.t. $\boldsymbol{\mu}$ (the MLE of $\boldsymbol{\mu}$) we denote $\hat{\boldsymbol{\mu}}$.

For some types of GLMs the value of D^* at $\hat{\boldsymbol{\mu}}$ for a given model gives an important measure of **goodness of fit** for that model.

Why? Because

$$D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\left[\ell(\mathbf{y}; \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}; \mathbf{y})\right]$$

compares the quality of the fit of the current model (given by $\ell(\hat{\boldsymbol{\mu}}, \mathbf{y})$) with the quality of the fit of the best-fitting model possible, the model with $\boldsymbol{\mu} = \mathbf{y}$ (given by $\ell(\mathbf{y}; \mathbf{y})$).

Notice that

$$D^*(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\log\lambda$$

where

$$\lambda = \frac{L(\hat{\boldsymbol{\mu}}_{\max};\mathbf{y})}{L(\hat{\boldsymbol{\mu}};\mathbf{y})}, \quad \textbf{(likelihood ratio)}$$

where $\hat{\boldsymbol{\mu}}_{\max} = \mathbf{y}$ is the MLE under a model with no restrictions (the saturated model), and $\hat{\boldsymbol{\mu}}$ is the MLE under the current model (which contains certain restrictions of the μ_i's).

- λ is the likelihood ratio for comparing the current model with the saturated model.

Notice in the CLM with normally distributed errors, and known σ^2:

$$f(\mathbf{y};\boldsymbol{\mu}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\sum_i \frac{(y_i-\mu_i)^2}{2\sigma^2} \right\}$$

$$\Rightarrow \quad \ell(\boldsymbol{\mu};\mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_i \frac{(y_i-\mu_i)^2}{2\sigma^2}$$

and

$$\ell(\mathbf{y};\mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - 0$$

$$\Rightarrow \quad D^*(\mathbf{y};\boldsymbol{\mu}) = 2\left[\ell(\mathbf{y};\mathbf{y}) - \ell(\boldsymbol{\mu};\mathbf{y})\right] = \frac{1}{\sigma^2}\sum_i (y_i-\mu_i)^2$$

so that $D \equiv \sigma^2 D^*$ (the (unscaled) **deviance**) is identical to SS_E.

- Therefore, in the CLM,

$$\text{Minimum Deviance} = \text{ML} = \text{OLS}$$

Asymptotics of ML

Inference in GLMs relies on the asymptotic properties of MLEs. Under certain "regularity conditions" (see Fahrmeir and Tutz, §2.2.1) we have the following results for a MLE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, the regression parameter in a GLM.

Asymptotic Existence and Uniqueness: The probability that $\hat{\boldsymbol{\beta}}$ exists and is (locally) unique tends to 1 as $n \to \infty$.

- For many important models, the log-likelihood $\ell(\boldsymbol{\beta}; \mathbf{y})$ is concave so that local and global maxima coincide. For strictly concave log-likelihoods the MLE is even unique whenever it exists.

Consistency: As $n \to \infty$, $\hat{\boldsymbol{\beta}}_n \xrightarrow{\text{p}} \boldsymbol{\beta}$ (weak consistency) and $\hat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}$ with probability 1 (strong consistency). Here, $\hat{\boldsymbol{\beta}}_n$ denotes the sequence of MLEs based on samples of size n.

Asymptotic Normality:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{\text{d}} N(\mathbf{0}, n\mathbf{I}(\boldsymbol{\beta})^{-1})$$

where $I(\boldsymbol{\beta})$ denotes the Fisher information matrix. This result implies that for n large

$$\hat{\boldsymbol{\beta}} \dot{\sim} N(\boldsymbol{\beta}, \mathbf{I}(\boldsymbol{\beta})^{-1}).$$

basic idea for proof: (one-dimensional case)

(This proof applies to any MLE $\hat{\beta}$ for some parameter β under the assumed regularity conditions, not just the MLE of a GLM regression parameter. See Fahrmeir and Kaufmann, 1985, for more details.)

Since $\hat{\beta}_n$ maximizes $\ell(\beta) = \sum_i^n \log f(y_i;\beta)$ (I've dropped the argument $\mathbf{y}$ here to simplify the notation), it follows that $\ell'(\hat{\beta}_n) = 0$. By Taylor's Theorem,

$$0 = \ell'(\hat{\beta}_n) = \ell'(\beta) + \ell''(\beta)(\hat{\beta}_n - \beta) + \frac{1}{2}\ell'''(\beta_n^*)(\hat{\beta}_n - \beta)^2,$$

where β_n^* is a point that lies between $\hat{\beta}_n$ and β and which therefore goes in probability to β (by the consistency of $\hat{\beta}_n$).

$$\Rightarrow \qquad (\hat{\beta}_n - \beta) = \frac{-\ell'(\beta)}{\ell''(\beta) + \frac{1}{2}\ell'''(\beta_n^*)(\hat{\beta}_n - \beta)}$$

$$\Rightarrow \qquad \sqrt{n}(\hat{\beta}_n - \beta) = \frac{-\frac{1}{\sqrt{n}}\ell'(\beta)}{\frac{1}{n}\ell''(\beta) + \frac{1}{2n}\ell'''(\beta_n^*)(\hat{\beta}_n - \beta)}$$

Now let's consider the terms of this ratio one at a time:

$$\frac{1}{\sqrt{n}}\ell'(\beta) = \frac{1}{\sqrt{n}}U(\beta)$$

where $U(\beta) = \sum \frac{\partial}{\partial\beta} \log f(y_i;\beta)$ is the score function, a sum of independent r.v.'s each with mean 0, where $\text{var}(U) = I(\beta)$. Therefore, a CLT for non-identically distributed r.v.s yields

$$\frac{1}{\sqrt{n}}U(\beta) \xrightarrow{\text{d}} N(0, n^{-1}I(\beta)) \qquad \text{(CLT)}$$

In addition, a weak law of large numbers gives

$$\frac{1}{n}\ell'' = \frac{1}{n}\sum \underbrace{\frac{\partial^2 \log f(y_i;\beta)}{\partial\beta^2}}_{\text{independent}} \xrightarrow{\text{p}} -\frac{1}{n}I(\beta) \quad \text{(its expected value)}$$

and because

$$\frac{1}{n}\ell'''(\beta_n^*) = \text{mean of } n \text{ indep. r.v.'s} \xrightarrow{\text{p}} \text{some constant}$$

and

$$(\hat{\beta}_n - \beta) \xrightarrow{\text{p}} 0 \quad \text{(consistency)}$$

it follows that

$$\frac{1}{n}\ell'''(\beta_n^*)(\hat{\beta}_n - \beta) \xrightarrow{\text{p}} 0$$

So we now have that

$$\sqrt{n}(\hat{\beta}_n - \beta) = \frac{\frac{1}{\sqrt{n}}U(\beta)}{-\frac{1}{n}\ell''(\beta) + o_p(1)}$$

where

$$\frac{1}{\sqrt{n}}U(\beta) \xrightarrow{\text{d}} N(0, n^{-1}I(\beta))$$

and $-\frac{1}{n}\ell''(\beta) \xrightarrow{\text{p}} n^{-1}I(\beta)$. So, by Slutzky's Theorem,

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n - \beta) &\xrightarrow{\text{d}} nI^{-1}(\beta) \times N(0, n^{-1}I(\beta)) \qquad \text{as } n \to \infty \\ &= N(0, nI^{-1}(\beta))\end{aligned}$$

or $\hat{\beta}_n \dot{\sim} N(\beta, I^{-1}(\boldsymbol{\beta}))$ in large samples. ∎.

- Note that $I(\beta)$ can be estimated consistently by $I(\hat{\beta})$.
- In addition, the expected information $I(\beta) = -\text{E}\{\ell''(\beta)\}$ can be replaced by the observed information, or negative Hessian, $-\ell''(\beta)$, without changing the asymptotic result.

Important Property: Functional Invariance of MLE

Let Y have density $f(y;\theta)$, $\theta \in \Theta$ and consider a reparameterization, $\phi = h(\theta)$ where h is one-to-one, $h : \Theta \to \Phi$.

If $\hat{\theta}$ is a MLE of θ, then $\hat{\phi} = h(\hat{\theta})$ is a MLE of ϕ (MLE is invariant to parameterization).

Computation of MLE for $\boldsymbol{\beta}$ in GLMs: (Fitting GLMs)

Assume for now that $\mathbf{X}$ is of full rank (p).

Need to solve

$$\frac{\partial \ell(\boldsymbol{\beta}; \mathbf{y})}{\partial \beta_j} = 0, \quad j = 1, \ldots, p,$$

where

$$\ell(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}; y_i) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i(\boldsymbol{\beta}) - b(\theta_i(\boldsymbol{\beta}))}{a_i(\phi)} + c(y_i, \phi) \right\}.$$

By the chain rule,

$$\begin{aligned} \frac{\partial \ell_i}{\partial \beta_j} &= \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ &= \frac{y_i - b'(\theta_i)}{a_i(\phi)} \frac{1}{b''(\theta_i)} \frac{1}{g'(\mu_i)} x_{ij} \end{aligned}$$

Let $d_i = g'(\mu_i) = \partial \eta_i / \partial \mu_i$, $u_i = \frac{1}{\mathrm{var}(y_i) d_i^2} = \frac{1}{a_i(\phi) b''(\theta_i) d_i^2}$. Then we can write

$$\frac{\partial \ell}{\partial \beta_j} = \sum_i \frac{\partial \ell_i}{\partial \beta_j} = \sum_i (y_i - \mu_i) u_i d_i x_{ij}$$

So, we need to solve

$$\sum_{i=1}^{n} (y_i - \mu_i) u_i d_i x_{ij} = 0 \quad j = 1, \ldots, p$$

where μ_i, u_i, and d_i all depend on $\boldsymbol{\beta}$. Notice that this is a nonlinear problem for which (in general) no closed-form solution exists.

How do we solve such a problem?

Newton-Raphson: For $f : \mathcal{R} \to \mathcal{R}$ we want to solve $f(x) = 0$. Need to find the x^* satisfying $f(x^*) = 0$. We require $|f'(x^*)| > 0$.

A Taylor series expansion of $f(x^*)$ gives

$$0 = f(x^*) = f(x) + f'(x)(x^* - x) + \cdots$$

which implies that

$$x^* \approx x - \frac{f(x)}{f'(x)}$$

for x close to x^*. This implies the iteration

$$x^{(m+1)} = x^{(m)} - \frac{f(x^{(m)})}{f'(x^{(m)})}$$

or, in multiple dimension

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \left[\frac{\partial}{\partial \mathbf{x}^{(m)T}} \mathbf{f}(\mathbf{x}^{(m)})\right]^{-1} \mathbf{f}(\mathbf{x}^{(m)}).$$

In our case, $\mathbf{f}$ is the score vector, $\frac{\partial \ell}{\partial \boldsymbol{\beta}}$

$$\Rightarrow \qquad \boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \left[\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}\right]^{-1} \left(\frac{\partial \ell}{\partial \boldsymbol{\beta}}\right)_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}$$

The **Fisher scoring** algorithm replaces the "observed information" matrix $-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$ with its expected value, the Fisher information matrix, giving the iteration

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[-\mathrm{E}\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right)_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}} \right]^{-1} \left(\frac{\partial \ell}{\partial \boldsymbol{\beta}} \right)_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}$$

or

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[\mathbf{I}_n^{(m)} \right]^{-1} \mathbf{S}^{(m)}, \qquad (*)$$

where $\mathbf{I}_n^{(m)}$, $\mathbf{S}^{(m)}$ are the Fisher information matrix and score vector, respectively, each evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$.

- Fisher scoring is often used in place of Newton-Raphson because the expected information is easier to compute than observed information.

Fisher scoring may be applied directly via (*) to fit GLMs. However, an *equivalent* more convenient algorithm exists known as **iteratively reweighted least squares** or IRLS.

IRLS:

Pre-multiplying both sides of (*) by $\mathbf{I}_n^{(m)}$ we have

$$\mathbf{I}_n^{(m)}\boldsymbol{\beta}^{(m+1)} = \mathbf{I}_n^{(m)}\boldsymbol{\beta}^{(m)} + \mathbf{S}^{(m)} \qquad (**)$$

In addition, for GLMs $\mathbf{I}_n(\boldsymbol{\beta})$ has $(j,k)^{\text{th}}$ element

$$\begin{aligned}
-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial\beta_j\partial\beta_k}\right] &= \mathrm{E}\left[\left(\frac{\partial\ell}{\beta_j}\right)\left(\frac{\partial\ell}{\beta_k}\right)\right] \\
&= \mathrm{E}\left[\left(\sum_i (y_i-\mu_i)u_i d_i x_{ij}\right)\left(\sum_i (y_i-\mu_i)u_i d_i x_{ik}\right)\right] \\
&= \sum_i \mathrm{E}\left[(y_i-\mu_i)^2 u_i^2 d_i^2 x_{ij}x_{ik}\right]
\end{aligned}$$

since all cross products have mean 0 by the independence of the y_i's. Continuing,

$$\begin{aligned}
-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial\beta_j\partial\beta_k}\right] &= \sum_i \mathrm{var}(y_i) u_i^2 d_i^2 x_{ij}x_{ik} \\
&= \sum_i a_i(\phi) b''(\theta_i) u_i^2 d_i^2 x_{ij}x_{ik} = \sum_i u_i x_{ij}x_{ik}
\end{aligned}$$

Thus,

$$\mathbf{I}_n(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{U}\mathbf{X}, \quad \text{where} \quad \mathbf{U} = \begin{bmatrix} u_1 & 0 & \cdots & 0 \\ 0 & u_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_n \end{bmatrix}$$

The r.h.s. of (**) is the vector with j^{th} element

$$\sum_{i=1}^{n}\sum_{k=1}^{p} u_i^{(m)} x_{ij} x_{ik} \beta_k^{(m)} + \sum_i (y_i - \mu_i^{(m)}) u_i^{(m)} d_i^{(m)} x_{ij}$$
$$= \sum_{i=1}^{n} \left(x_{ij} u_i^{(m)} \left[\sum_{k=1}^{p} x_{ik} \beta_k^{(m)} + (y_i - \mu_i^{(m)}) d_i^{(m)} \right] \right)$$

where the superscript $^{(m)}$ denotes that the quantity is evaluated at $\boldsymbol{\beta}^{(m)}$.

Thus the r.h.s. of (**) can be written as $\mathbf{X}^T\mathbf{U}\mathbf{z}$ where $\mathbf{z}$ is a $n-$dimensional vector with i^{th} element

$$z_i = \sum_{k=1}^{p} x_{ik} \beta_k^{(m)} + (y_i - \mu_i^{(m)}) d_i^{(m)}$$
$$= \eta_i^{(m)} + (y_i - \mu_i^{(m)}) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}$$

Thus value of $\boldsymbol{\beta}$ at the $(m+1)^{\text{st}}$ iteration satisfies

$$\mathbf{X}^T\mathbf{U}\mathbf{X}\boldsymbol{\beta}^{(m+1)} = \mathbf{X}^T\mathbf{U}\mathbf{z}$$

or

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}^{(m+1)} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{z} \qquad (***)$$

where $\mathbf{z}$ and $\mathbf{V} = \mathbf{U}^{-1} = \text{diag}(u_1^{-1}, \ldots, u_n^{-1})$ are evaluated at $\boldsymbol{\beta}^{(m)}$. Notice that (***) is the WLS normal equation for a regression of the **working variate z** on $\mathbf{X}$.

The algorithm for obtaining the MLE of $\boldsymbol{\beta}$ in a GLM can be summarized as follows:

1. Obtain starting values.

   – these could be for $\boldsymbol{\beta}$, but in any given iteration the quantities that depend on the previous iteration's $\boldsymbol{\beta}-$value do so only through $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$. Therefore, we only need starting values for $\boldsymbol{\mu}$ which its convenient to take as the data themselves: $\boldsymbol{\mu}^{(0)} = \mathbf{y}$ (minor adjustments to the original data may be needed to avoid $\log(0)$, other problems).

2. Compute $\boldsymbol{\beta}^{(m+1)}$ via the weighted least squares regression of $\mathbf{z}$ on $\mathbf{X}$ where $z_i = \eta_i^{(m)} + (y_i - \mu_i^{(m)}) \frac{\partial \eta_i}{\partial \mu_i}$ with weight matrix $\mathbf{V} = \text{diag}(u_1^{-1}, \ldots, u_n^{-1})$ where

$$u_i^{-1} = \frac{\phi}{w_i} v(\mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2$$

   (These u_i's are called the **iterative weights**). In this step both $\mathbf{z}$, $\mathbf{V}$ are evaluated at $\boldsymbol{\beta}^{(m)}$ (or at $\boldsymbol{\mu}^{(0)}$ when $m = 0$).

3. Repeat step 2 until some convergence criterion is obtained. For example, stop when

$$\frac{||\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m-1)}||}{||\boldsymbol{\beta}^{(m-1)}||} < \epsilon$$

   for some prechosen $\epsilon > 0$. At convergence, the MLE is $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(m)}$.

- This algorithm is known as iteratively reweighted least squares (IRLS) because it consists of a series of WLS calculations where weights are recomputed at each iteration.
- IRLS = Fisher scoring for E.D. family.
- IRLS = Newton-Raphson for E.D. family with canonical link.
- IRLS implemented in PROC GENMOD, the R function glm(), etc.
- Notice that in the IRLS algorithm, the form of the loglikelihood only enters into the algorithm through the first two moments. Therefore, IRLS can be used to fit more general models outside of the GLM class that are defined by first and second moment assumptions only, rather than full likelihood specifications (Quasi-likelihood models).

  What about ϕ?

Notice that ϕ appears on both sides of the WLS normal equation (***), thus cancelling out of the calculation of $\hat{\boldsymbol{\beta}}$.

Therefore, ϕ need not be updated throughout the IRLS algorithm. ϕ can simply be estimated at convergence. We'll come back to how we estimate ϕ (what method and formulas) later.

By the ML theory we've reviewed we know the asymptotic variance of $\hat{\boldsymbol{\beta}}$:

$$\text{avar}(\hat{\boldsymbol{\beta}}) = \mathbf{I}_n(\boldsymbol{\beta})^{-1}$$

which we can estimate as follows:

$$\hat{\text{avar}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1},$$

where $\hat{\mathbf{V}}$ is $\mathbf{V}$ evaluated at $\hat{\boldsymbol{\beta}}$,$\hat{\phi}$.

Aliasing: (section 3.5, M&N)

So far, we've assumed $\mathbf{X}$ is of full rank. Often this will not be the case.

If columns $\mathbf{x}_1, \ldots, \mathbf{x}_r$ of $\mathbf{X}$ form a linearly independent set, then one or more of $\beta_1, \ldots, \beta_r$ are said to be **aliased.**

In this case $\boldsymbol{\beta}$ is not **estimable** because $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ doesn't uniquely determine $\boldsymbol{\beta}$; however, $\boldsymbol{\eta}$ and hence $\boldsymbol{\mu}$ remains estimable, so it is useful to know how to fit the model and do inference on $\boldsymbol{\mu}$ and other estimable functions of $\boldsymbol{\beta}$ in this situation.

Example:

Suppose that length, breadth and area measurements are made on leaves which have the property that area=constant×length×breadth. Now suppose we form a GLM for calcium concentration in the leaves, say, which has linear predictor

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

involving covariates

$$X_1 = \log \text{ length}$$

$$X_2 = \log \text{ breadth}$$

$$X_3 = \log \text{ area}$$

Since area=constant×length×breadth, we have

$$X_3 = c + X_1 + X_2,$$

where c = log of the original constant in the formula for area. Hence η may be expressed in terms of X_1 and X_2 as

$$\begin{aligned} \eta &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (c + X_1 + X_2) \\ &= \beta_0 + \beta_3 c + (\beta_1 + \beta_3) X_1 + (\beta_2 + \beta_3) X_2. \end{aligned}$$

Thus we can distinguish

$$\beta_0 + \beta_3 c, \quad \beta_1 + \beta_3, \quad \text{and} \quad \beta_2 + \beta_3,$$

but not the four parameters $\beta_0, \beta_1, \beta_2, \beta_3$ separately.

- In this example, the aliasing is intrinsic to the problem. Aliasing will occur whatever the sizes of the leaves (assuming area, length and breadth are all measured without error and the leaves all conform to the relationship area=constant×length×breadth).

Two types of Aliasing:

1. Intrinsic Aliasing – the specification of the linear structure contains redundancy whatever the observed values of $\mathbf{X}$; E.g., the leaves example, or the effects model in the one-way layout.

2. Extrinsic Aliasing – An anomoly of the data makes the columns of $\mathbf{X}$ linearly dependent; E.g., no observations were obtained for one level of a factor yielding a $\mathbf{0}$ column in the $\mathbf{X}$ matrix.

Solutions to the Problem of Aliasing:

1. Use a **generalized inverse** to find one particular (of many) solution $\hat{\boldsymbol{\beta}}$; then estimate and interpret **estimable functions** of $\boldsymbol{\beta}$.

   In solving
   $$(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})\boldsymbol{\beta}^{(m+1)} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{z}$$
   obtain $\boldsymbol{\beta}^{(m+1)}$ as
   $$\boldsymbol{\beta}^{(m+1)} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}^T\mathbf{V}^{-1}\mathbf{z}$$
   where $\mathbf{M}^-$ indicates a generalized inverse of the (perhaps singular) matrix $\mathbf{M}$. That is, $\mathbf{M}^-$ has the property $\mathbf{M}\mathbf{M}^-\mathbf{M} = \mathbf{M}$.

   This approach is satisfactory because, fortunately, statistically important functions of $\boldsymbol{\beta}$ such as $\boldsymbol{\eta}$, $\boldsymbol{\mu}$ are estimable. That is, $\hat{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}})$, $\hat{\boldsymbol{\mu}}(\hat{\boldsymbol{\beta}})$ will be the same regardless of which $\hat{\boldsymbol{\beta}}$ we choose.

- Choice of a particular $\hat{\boldsymbol{\beta}}$ doesn't change our model, it only determines the manner in which we choose to express the linear structure in our model.

2. We can avoid a generalized inverse by imposing constraints on $\hat{\boldsymbol{\beta}}$ so that the solution is unique.

   - Its important to realize that only constraints on the parameter estimates (not the parameters themselves) are necessary to solve the normal equations and to derive most things of interest in fitting the model (e.g., in a linear model setting, the SS_E, the analysis of variance, the error variance estimate, and the b.l.u.e. of any estimable function).

   - However, it is often desirable to restrict model parameters in certain ways for interpretability reasons. In such cases the same constraints are applied to parameter estimates as well to obtain a solution of the normal equations.

   - Common choices of constraints include the *baseline (regression) constraints* in which $p - \text{rank}(\mathbf{X})$ of the elements of $\hat{\boldsymbol{\beta}}$ are set equal to zero corresponding to the $p - \text{rank}(\mathbf{X})$ linear dependencies in the design matrix. The *usual (anova, sum-to-zero, conventional) constraints* are require certain of the parameter estimates to sum to zero (e.g., the estimates of all effects corresponding to the levels of a factor must sum to 0).

   - PROC GENMOD in SAS allows control over the constraints used to identify the model through options on the CLASS statement. In R it is also possible to choose parameter constraints via the options(contrasts=...) function.

Goodness of Fit and Hypothesis Testing:

At this point we consider fit of the linear structure. We assume for now that the error distribution and link specifications are satisfactory.

To talk about model fit and selection its useful to define several cases of the linear structure:

i. If we include n linearly independent explanatory variables $\Rightarrow$ MLE of μ_i's are the y_i's themselves (perfect fit). This situation is known as the **full** or **saturated model.** This model involves no summarization (reduction) of the data.

ii. At the other extreme is the **null model** where we assume $\mu_i = \mu$ $\forall i$. This model is almost always too simple.

iii. The **current model** is the model currently under investigation, lying somewhere in between the full and null models.

Sometimes its useful to identify two more models:

iv. The **minimal model** is the simplest model we wish to consider. Typically its more complex that (ii) because we know (e.g.) that certain effects must be included to account for the experimental design and/or to allow hypothesis tests of *a priori* interest.

v. The **maximal model** is the most complex model we're willing to consider. Typically its less complex than the full model.

Most testing problems of interest in GLMs can be expressed as linear hypotheses of the form

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b} \quad \text{versus} \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{b}$$

where $\mathbf{C}$ is an $s \times p$, full row rank matrix of constants and $\mathbf{b}$ is a $p-$vector of constants.

Most commonly, we want to test

$$H_0 : \boldsymbol{\beta}_r = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\beta}_r \neq \mathbf{0}$$

where $\boldsymbol{\beta}_r$ is an $r-$dimensional sub-vector of $\boldsymbol{\beta}$.

A hypothesis of this form includes as a special case a goodness of fit test for the overall adequacy of the model: for this test the hypotheses are of the form

$$H_0 : \boldsymbol{\beta}_{n-p} = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\beta}_{n-p} \neq \mathbf{0}$$

where $\boldsymbol{\beta}_{n-p}$ is the sub-vector of $\boldsymbol{\beta}_n$, the full model parameter, that must be set to zero to obtain the current model (with $\dim(\boldsymbol{\beta}) = p$).

Under likelihood-based inference the classical tool for testing hypotheses is the Likelihood Ratio Statistic:

$$\begin{aligned} \lambda &= \frac{L(\tilde{\boldsymbol{\beta}}; \mathbf{y})}{L(\hat{\boldsymbol{\beta}}; \mathbf{y})}, \quad \begin{array}{l} \tilde{\boldsymbol{\beta}} = \text{MLE under larger model} \\ \hat{\boldsymbol{\beta}} = \text{MLE under smaller model} \end{array} \\ &= \frac{L(\tilde{\boldsymbol{\mu}}; \mathbf{y})}{L(\hat{\boldsymbol{\mu}}; \mathbf{y})}, \quad \tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}(\tilde{\boldsymbol{\beta}}) \end{aligned}$$

Logic: if $\lambda \doteq 1 \Rightarrow$ no evidence against $H_0 \Rightarrow$ models fit about equally well so should adopt smaller model. If $\lambda >> 1$ then reject H_0, adopt larger model.

How large does λ need to be to reject?

Answer: large in comparison to its distribution under H_0 (above the $100(1-\alpha)^{\text{th}}$ percentile of its null distribution).

The following result is due to Wilks:

Let

$$\lambda = \frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})}, \quad \begin{array}{l} \tilde{\boldsymbol{\theta}} = \text{MLE under model 1} \\ \hat{\boldsymbol{\theta}} = \text{MLE under model 2} \end{array}$$

where model 2 is nested within model 1. Then under mild regularity conditions, we have the following asymptotic result:

$$2 \log \lambda \overset{a}{\sim} \chi^2(t_1 - t_2)$$

where t_i = the number of independent parameters estimated under model i, $i = 1, 2$.

Recall we originally expressed the scaled deviance

$$D^*(\mathbf{y}; \boldsymbol{\mu}) = 2[\ell(\mathbf{y}; \mathbf{y}) - \ell(\boldsymbol{\mu}; \mathbf{y})]$$

as a function of $\boldsymbol{\mu}$ and we noticed that ML=Min Dev.

For a particular model, though, with MLE $\hat{\boldsymbol{\mu}}$, the scaled deviance at $\hat{\boldsymbol{\mu}}$ becomes our primary tool for inference in a GLM context:

$$D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2[\ell(\mathbf{y}; \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}; \mathbf{y})]$$

- Notice that $D^*(\mathbf{y}; \hat{\boldsymbol{\mu}})$ is just $2 \log \lambda$ for comparing the current model to the full model.

Examples – Normal Distribution, σ^2 Known.

We've already seen that

$$D^*(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\left[-\frac{1}{2\sigma^2}\sum_i (y_i - y_i)^2 + \frac{1}{2\sigma^2}\sum_i (y_i - \hat{\mu}_i)^2\right]$$
$$= \frac{1}{\sigma^2}\sum_i (y_i - \hat{\mu}_i)^2 = \frac{1}{\sigma^2} SS_E$$

$\Rightarrow D(\mathbf{y};\hat{\boldsymbol{\mu}}) = SS_E$.

Poisson Distribution

$$f(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad i = 1, \ldots, n$$

$$\Rightarrow \quad \ell(\boldsymbol{\mu};\mathbf{y}) = \log \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \sum_i [-\mu_i + y_i \log \mu_i - \log y_i!]$$

$$\Rightarrow \quad D^*(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\sum_i [-y_i + y_i \log y_i - \log y_i! - (-\hat{\mu}_i + y_i \log \hat{\mu}_i - \log y_i!)]$$
$$= 2\sum_i [y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)] = D(\mathbf{y};\hat{\boldsymbol{\mu}}), \quad \phi = 1 \text{ here}$$

Note that if the residuals $y_i - \hat{\mu}_i$ sum to zero then

$$D^*(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\sum_i y_i \log \frac{y_i}{\hat{\mu}_i} = G^2,$$

the G^2 likelihood ratio chi-square statistic often used for goodness of fit in contingency table analysis.