Ongoing Notes - Methods 3

Definitions

Chapter 1

Four Fundamental Notions of Probability:

(1): Laplacian Probability

- "Gambling"/"Classical" probability
- E.g., fair coins, dice, cards
- A material concept of probability
- Equally likely outcomes a typical result
- $Pr(E) \equiv \frac{|E|}{|S|}$, S sample space, E event, |A| size of set A

(2): Relative Frequency

- **"Not hypothetical limiting relative frequency"**
- A material concept of probability
- Probability is a direct consequence of physical realities, particularly for finite populations
- $Pr(A|B) = \frac{|A|}{|B|}$, where there is a finite number of objects in class B, and operation to determine whether object also belongs to a class A.

(3): Hypothetical Limiting Relative Frequency

- "What we usually mean when we refer to relative frequency or frequentist probability"
- A material concept of probability
- Can at least hypothetically repeat operation an infinite number of times
- $Pr(E) = \lim_{n\to\infty}(\frac{E_n}{n})$

(4): Epistemic Probability

- Any concept of probability that cannot be expressed in terms of physical events can be considered epistemic probability.
- Probability $\equiv$ knowledge or belief
- Belief is updated or modified in the light of observed information; mathematical formalism necessary for logical consistency
- $Pr(E|y) = \frac{Pr(y|E)Pr(E)}{Pr(y|E)Pr(E)Pr(y|E^c)Pr(E^c)}$

Note: (1) through (3) use same notion of "operation", "sample space", and "events"

Operation: Observation, measurement, or selection

Sample Space: Set of possible Outcomes of an operation

Events: Subsets of elements in the sample space

Approaches to Statistical Analysis:

(1): Analysis Through Randomization: Predominantly using Laplacian and Relative Frequency probability concepts

(2): Analysis Using Models: Typically uses Hypothetical Limiting Relative Frequency

(3): Bayesian Analysis: Construction of data model; makes use of epistemic probability in the form of prior and posterior distributions

Chapter 2

Abstract: To mean abstruse; but also, to separate, to express a quality apart from an object, or to consider a part as divorced from the whole

Experimental vs. Observational Studies: More like Experimental vs. Non-Experimental

Experimental Approach:

- The *Whole*: consists of all the external conditions to which an experimental unit is subject.
- The *Part*: is to examine fluctuations in a small number of those conditions while holding all others constant.
- Key element is control of all relevant factors.

Statistical Modelling: Depends entirely on the mathematical concepts of random variables and theoretical probability distributions.

- Def: A collection of probabilistic assignments and specified relations among model components that leads to a joint probability distribution for the entire collection of random variables involved in a problem.

Random variables: Mathematical concepts that are attached to the results of observation of the quantities of primary interest on sampling units.

- Sampling units may correspond to experimental units under that approach to analysis and may certainly be physically existing entities, but neither are required for the concepts of random variables and their assumed theoretical distributions to be valid.
- Random variables and theoretical probability distributions are not simply an extension of finite populations to infinite collections of physical units, if even possible; random variables are mathematical beasts.

Statistical Abstraction: Capturing the key elements of a problem in a small set of parameters of a probabilistic model.

A Linear Model:

$$Y_i = \beta_0 + \beta_1 x_i + \sigma\epsilon_i; \epsilon_i \sim iidN(0,1).$$

Systematic Model Component: $\beta_0 + \beta_1 x_i$

Dispersion, variance, or precision parameter quantities the degree to which observed values differ from what is "explained" by the modeled version.": σ^2

Objectives of Analysis:

(1): Data Description

(2): Problem Conceptualization

(3): Examination of Scientific Theory

(4): Estimation and Inference About a Specific Quantity

(5): Prediction or Forecasting: These are two different things!

Prediction: prediction of an unobserved random variable or functional of a distribution that is given (conceptualized) existence within the spatial and/or teporal extent of a set of data.

Forecasting: prediction of random quantities that are given (conceptual) existence outside the spatial and/or temporal extent of the available data.

Distinguishing between *motivation* and *verification*

- Motivation can be provided by elements of the scientific problem, physical realities of data collection, definitiion of random variables, or exploratory analysis of the data themselves.
- The observational process can motivation assumptions of independence (or lack thereof).
- For now we can think of verification as a goodness of fit problem.

Chapter 3

"... [W]e distinguish between classes and families of distributions.":

Family: "Consider the probability mass or density function $f(x \mid \theta)$ with support $x \in \Omega_x$, and where $\theta = (\theta_1, \ldots, \theta_p)^T$ is a parameter such that $\theta \in \Theta \subseteq \mathbb{R}^p$. As θ varies over its parameter space Θ we say that $f(x \mid \theta)$ generates a family of distributions."

- Of interest is (1): Location-Scale families and (2): Exponential families

Class: "Collections of different families of distributions may be grouped into classes of distributions."

Location-Scale Family:

Let U be a random variable with a fixed distribution F. If U is transformed into Y as

$$Y = U + \mu, \quad -\infty < \mu < \infty,$$

then Y has distribution $F(y - \mu)$ since $\Pr(Y \leq y) = \Pr(U \leq y - \mu)$.

The set of distributions generated for a fixed F, as μ varies from $-\infty$ to ∞, is called a **location family** of distributions.

If the resultant distribution is of the same form as F only with modified parameter values, then F forms a location family.

A similar definition of a distribution F forming a **scale family** is if F is unchanged other than parameter values under transformations

$$Y = \sigma U, \quad \sigma > 0,$$

in which case the distribution of Y is $F(y/\sigma)$ since $\Pr(Y \leq y) = \Pr(U \leq y/\sigma)$.

The composition of location and scale transformations results in

$$Y = \mu + \sigma U, \quad -\infty < \mu < \infty, \ \sigma > 0,$$

and Y has distribution $F((y - \mu)/\sigma)$.

If F has a density f, then the density of Y is given by

$$g(y \mid \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{y - \mu}{\sigma}\right).$$

Properties:

Location–scale families have simple properties that stem directly from the transformations.
For example, if Y is produced as a location–scale transformation of U,

$$Y = \mu + \sigma U,$$

then

$$\mathbb{E}(Y) = \mu + \sigma\, \mathbb{E}(U) \quad \text{and} \quad \mathrm{Var}(Y) = \sigma^2 \mathrm{Var}(U).$$

Traditionally, if $\mathbb{E}(U) = 0$ and $\mathrm{Var}(U) = 1$ then the distribution of U is called the **parent distribution** for the family

Exponential Family:

Common (Basic Form of) representation:

$$f(y \mid \theta) \;=\; \exp\left\{\sum_{j=1}^{s} \theta_j T_j(y) - B(\theta) + c(y)\right\}. \tag{3.4}$$

Note: The term $\exp\{c(y)\}$ in the last expression of (3.4) could be absorbed into the relevant measure. This is typically not done so that integrals can be written with respect to the dominating **Lebesgue measure** (for continuous Y) or **counting measure** (for discrete Y).

Properties:

(Property 2 is the case that the sufficient statistic and parameter does not satisfy a linear constraint, suchj that the representation is "minimal" or "full"); the subsequent properties follow fromm this.)

3. For a minimal, regular exponential family, the statistic $T \equiv (T_1, \ldots, T_s)$ is minimal sufficient for θ.

This property is often useful because, as we will see, the joint distribution of *iid* random variables belonging to an exponential family are also of the exponential family form.

4. For an integrable function $h(\cdot)$, dominating measure ν, and any θ in the interior of Θ, the integral

$$\int h(y) \exp\left\{\sum_{j=1}^{s} \theta_j T_j(y) - B(\theta) + c(y)\right\} d\nu(y)$$

is continuous, has derivatives of all orders with respect to the θ_js,
and these derivatives can be obtained by interchanging differentiation and integration (e.g., Lehmann, 1983, Theorem 4.1).

This property does several things for us:
- It can be used to derive additional properties of exponential families such as the form of the moment generating function in property 6.
- It allows us to evaluate expressions needed for estimation and variance evaluation through numerical integration of derivatives, which can be important to actually conduct an analysis with real data.

5. Property 4 can be used to show that (e.g., Lehmann, 1983),

$$\mathbb{E}\{T_j(Y)\} = \frac{\partial}{\partial \theta_j} B(\theta),$$

$$\operatorname{cov}\{T_j(Y), T_k(Y)\} = \frac{\partial^2}{\partial \theta_j \partial \theta_k} B(\theta).$$

These lead directly to $\mathbb{E}(Y)$ and $\operatorname{Var}(Y)$ for what are called **natural exponential families** and **exponential dispersion families**, which will be discussed in the sequel.

They also provide an alternative parameterization of exponential families in general.

6. The moment generating function of an exponential family is defined to be that for the moments of the T_js and may be derived to be,

$$M_T(u) = \frac{\exp\{B(\theta + u)\}}{\exp\{B(\theta)\}}.$$

Exponential Dispersion Family Consider a random variable $Y \sim N(\mu, \sigma_*^2)$ for which σ_*^2 is considered a fixed, known value.
In this case we can write, for $-\infty < \mu < \infty$ and $0 < \sigma^2$,

$$f(y \mid \mu) = \exp\left[\frac{-1}{2\sigma_*^2}(y-\mu)^2 - \tfrac{1}{2}\log(2\pi\sigma_*^2)\right]$$

$$= \exp\left[\frac{1}{\sigma_*^2}\left(\mu y - \tfrac{1}{2}\mu^2\right) - \tfrac{1}{2}\left(\frac{y^2}{\sigma_*^2} - \log(2\pi\sigma_*^2)\right)\right].$$

Letting $\theta = \mu$, $b(\theta) = (1/2)\theta^2$, $\phi = 1/\sigma_*^2$, and $c(y, \phi) = (1/2)[y/\sigma_*^2 - \log(2\pi\sigma_*^2)]$,

this density may be written as what is called an **exponential dispersion family**, which has the general form of

$$f(y \mid \theta, \phi) = \exp\{\phi[y\theta - b(\theta)] + c(y, \phi)\}. \tag{3.7}$$

For a distribution with pdf or pmf of the form (3.7) the properties of s–parameter exponential families may be used to demonstrate that

$$\mathbb{E}(Y) = \frac{d}{d\theta}b(\theta) = b'(\theta),$$

$$\mathrm{Var}(Y) = \frac{1}{\phi}\frac{d^2}{d\theta^2}b(\theta) = \frac{1}{\phi}b''(\theta) = \frac{1}{\phi}V(\mu). \tag{3.8}$$

The function $V(\cdot)$ in (3.8) is often called the **variance function**, which is not the variance except for a few cases in which $\phi = 1$. The variance function is important because it quantifies the relation between the mean and variance of the distribution.

Comments:

(1): What has happened in (3.7) is that we have coerced a two-parameter exponential family to look almost like a natural exponential family (see Example 3.5) but with the addition of an extra parameter ϕ called the **dispersion parameter**. This parameter is a scale factor for the variance (3.8).

(2): Clearly, it will not be possible to write an exponential family in the form of expression (3.7) unless one of the sufficient statistics is given by the identity function (i.e., $T_j(y) = y$ for some j). While this is not, in itself, sufficient for representation of a pdf or pmf as in (3.7), distributions for which one of the sufficient statistics is y and which can subsequently be written in exponential dispersion family form include the **binomial, Poisson, normal, gamma, and inverse Gaussian**. But it is not possible, for example, to write a beta pdf in this form.

(3): Exponential dispersion families of the form (3.7) are the exponential families upon which **generalized linear models** are based (e.g., McCullagh and Nelder, 1989). But, as discussed in Chapter 1, the impetus provided by generalized linear models to consider random model components in a more serious light than mere error distributions has much wider applicability than just these families.

Extending Exponential Families for Samples (Random Samples, Multivariate) One additional property of exponential families will be useful in this subsection.
For Y distributed according to an s-parameter exponential family as in (3.4) with $\theta = (\theta_1, \ldots, \theta_s)$, the sufficient statistic

$$T(y) = (T_1(Y), \ldots, T_s(Y))$$

is distributed according to an exponential family with density or mass function

$$g(t \mid \theta) = \exp\left\{ \sum_{j=1}^{s} \theta_j t_j - B(\theta) + k(t) \right\}. \tag{3.9}$$

Note that the dominating measure of the distributions of Y and T may differ, and that $k(t)$ may or may not be easily derived from the original $c(y)$, but θ and $B(\theta)$ are the same as for the original distributions $f_Y(y \mid \theta)$.

Consider now the case of n independent and identically distributed random variables $Y_1, \ldots, Y_n$, with each variable having a pdf or pmf of the form

$$f(y \mid \theta) = \exp\left\{ \sum_{j=1}^{s} \theta_j T_j(y) - B(\theta) + c(y) \right\}.$$

Under the *iid* assumption, the joint distribution of

$$Y \equiv (Y_1, \ldots, Y_n)^T,$$

is

$$f(y \mid \theta) = \exp\left\{ \sum_{j=1}^{s} \theta_j \sum_{i=1}^{n} T_j(y_i) - nB(\theta) + \sum_{i=1}^{n} c(y_i) \right\}. \tag{3.10}$$

Note that expression (3.10) is still in the form of an exponential family, with sufficient statistics given by the sums of the $T_j(\cdot)$.
In particular, let $Y_1, \ldots, Y_n$ be distributed according to a one-parameter exponential family.
Then the joint distribution is again a one-parameter exponential family with the same canonical parameter and sufficient statistic given as the sum $\sum_{i=1}^{n} T(Y_i)$.

Chapter 4

Chapter 5

Positivity Condition: "The positivity condition states that, if a random variable can assume a given value, it can assume that value in combination with any other set of values for the other random variables involved in the problem, that is, there are no forbidden states in the possible joint configurations."

$$L(\theta \mid y) = \Pr(y \mid \theta) \propto \prod_{i=1}^{n} f_i(y_i \mid \theta),$$

Equality if f is pmf, $\propto$ if f is pdf

Assymptotic Normality:

If ψ is a scalar parameter and $\hat{\psi}_n$ denotes a consistent sequence of estimators of ψ, that sequence is asymptotically normal if there exists a sequence of constants σ_n such that,

$$\frac{\hat{\psi}_n - \psi}{\sigma_n} \quad \xrightarrow{d} \quad N(0,1), \tag{5.5}$$

Fisher Information:

The expected or Fisher information plays an important role in the theory of estimation. In the case of iid random variables $Y_1, \ldots, Y_n$ with common distribution depending on a scalar parameter ψ, define the expected information in a single random variable as,

$$I(\psi) = E\left(\left[\frac{d}{d\psi}\log\{f(Y \mid \psi)\}\right]^2\right). \tag{5.8}$$

If the random variables $Y_1, \ldots, Y_n$ are independent but not identically distributed (inid) with distributions that depend on a common parameter ψ, the total expected information is defined as the $p \times p$ matrix $I_n(\psi)$ with jk^{th} element,

$$I_{n,jk}(\psi) = \sum_{i=1}^{n} E\left[\frac{\partial}{\partial\psi_j}\log\{f_i(Y_i \mid \psi)\}\,\frac{\partial}{\partial\psi_k}\log\{f_i(Y_i \mid \psi)\}\right]. \tag{5.10}$$

Efficiency

An estimator for which there is equality in the information inequality is said to be efficient, which is a small sample or exact property

Scalar Parameter Regularity Conditions

R1. The distributions of the response variables are identifiable, meaning that different parameter values result in distinct distributions. We will assume these distributions have a common probability density or mass function $f(y \mid \theta)$; $y \in \Omega$ and that the support Ω does not depend on θ.

R2. The parameter space Θ is an open interval (not necessarily finite).

R3. The common density or mass function $f(y \mid \theta)$ has three continuous derivatives with respect to θ.

R4. With $\mu(y)$ denoting the dominating measure (Lebesgue or counting) the first and second derivatives of the integral $\int f(y \mid \theta)\, d\mu(y)$ can be evaluated by passing the derivative under the integral operator, that is, for $k = 1, 2$,

$$\frac{d^k}{d\theta^k} \int f(y \mid \theta)\, d\mu(y) = \int \frac{d^k}{d\theta^k} f(y \mid \theta)\, d\mu(y).$$

R5. The expected (or Fisher) information in a single random variable,

$$I(\theta) = E\left[\left(\frac{d}{d\theta} \log\{f(y \mid \theta)\}\right)^2\right],$$

is positive and finite, $0 < I(\theta) < \infty$.

R6. For all elements of the support $y \in \Omega$ and in an interval of the true parameter value, $\theta_0 - c < \theta < \theta_0 + c$, the third derivative of $\log\{f(y \mid \theta)\}$ satisfies

$$\left|\frac{d^3}{d\theta^3} \log\{f(y \mid \theta)\}\right| \leq M(y),$$

where

$$E_{\theta_0}[M(y)] < \infty.$$

"It should be noted that the regularity conditions just listed are typicalbut not unique in developing asymptotic results for likelihood estimation"

"Also, not all of these conditions are needed for resolution of each of the individual issues listed previously."

Properties/Results When Regularity Conditions Hold

Suppose that conditions R1–R6 are satisfied. If the sequence of solutions to the likelihood equation given by Lemma 1 is unique for all n and y, then it is a sequence of maximum likelihood estimators $\hat{\theta}_n$ and,

$$\sqrt{n}\,[I(\theta)]^{1/2}(\hat{\theta}_n - \theta) \quad \xrightarrow{d} \quad N(0, 1), \tag{5.19}$$

where

$$I(\theta) = -E\left[\frac{d^2}{d\theta^2} \log\{f(Y \mid \theta)\}\right].$$

If $Y_1, ..., Y_n$ in Likelihood Theorem 1 follow a common distribution that constitutes an exponential family, then solutions to the likelihood equations, if they exist, are unique.

Multiple Parameter Regularity Conditions

R1, R2. The first two regularity conditions remain identical to those given previously.

RM3. The common density or mass function $f(y \mid \theta)$ has continuous partial derivatives up to order three with respect to the elements of θ.

RM4. Typically, R4 is restated as the direct consequence of what that condition implies for the single parameter case, generalized to multiple parameters so that, for $j, k = 1, \ldots, p$,

$$E\left[\frac{\partial}{\partial\theta_j}\log\{f(y\mid\theta)\}\right]=0$$

and

$$I_{j,k}(\theta)=E\left[\frac{\partial}{\partial\theta_j}\log\{f(y\mid\theta)\}\frac{\partial}{\partial\theta_k}\log\{f(y\mid\theta)\}\right]=-E\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log\{f(y\mid\theta)\}\right].$$

RM5. Each element of the information matrix $I_{j,k}(\theta)$ is positive and finite and the matrix itself $I(\theta)$ is positive definite.

RM6. The smoothness condition R6 is generalized to hold for third partial derivatives as, for $j,k,\ell=1,\ldots,p$,

$$\left|\frac{\partial^3}{\partial\theta_j\,\partial\theta_k\,\partial\theta_\ell}\log\{f(y\mid\theta)\}\right|\le M_{j,k,\ell}(y),$$

such that for $j,k,\ell=1,\ldots,p$,

$$E_{\theta_0}[M_{j,k,\ell}(\theta)]<\infty.$$

Additional Propoerties of Likelihood-based Estimators

1. If a given scalar parameter θ (which may be an element of the parameter vector $\boldsymbol{\theta}$) has a single sufficient statistic $T(y)$, then the maximum likelihood estimator must be a function of that sufficient statistic.

If that sufficient statistic is minimal and complete, then the maximum likelihood estimator is unique. If the maximum likelihood estimator is unbiased then it is the UMVU

2. Likelihood-based estimators, determined as solutions to the likelihood equations, possess a property called *invariance* that is very useful but is not, in general, a property of other types of estimators, such as unbiased or least squares estimators.

The invariance property can be stated as: if $\tilde{\theta}_n$ is a consistent sequence of solutions to the likelihood equations, and $g(\theta)$ is a continuous, real-valued function of θ, then $g(\tilde{\theta}_n)$ is a consistent sequence of solutions to the likelihood equations when the likelihood is reparameterized in terms of $g(\theta)$.

If $\hat{\theta}_n$ is the maximum likelihood estimator of θ, then $g(\hat{\theta}_n)$ is the maximum likelihood estimator of $g(\theta)$.

Invariance is particularly useful in considering alternate parameterizations of random model components.

Wald Theory

If $\{\hat{\theta}_n\}$ is a sequence of consistent, asymptotically normal and efficient estimators of $\theta\equiv(\theta_1,\ldots,\theta_p)^T$, then

$$(\hat{\theta}_n-\theta)^T I_n(\hat{\theta}_n)(\hat{\theta}_n-\theta)\ \xrightarrow{d}\ \chi^2_p, \tag{5.23}$$

Delta Method

Let $g(\psi) = [g_1(\psi), \ldots, g_r(\psi)]^T$, $r \leq p$, where each component function $g_k(\psi)$ is continuously differentiable in a neighborhood of θ.

Then,

$$g(\hat{\psi}_n) \;\dot{\sim}\; AN\big[g(\psi),\, a_n^2 D\Sigma D^T\big], \tag{5.26}$$

where, D is an $r \times p$ matrix with k, j^{th} element

$$\frac{\partial}{\partial \psi_j} g_k(\psi).$$

In both the covariance Σ and in the matrix D, $\hat{\psi}_n$ may be used as a plug-in estimator of ψ. Consistency of $\hat{\psi}$ allows the asymptotic result to be applied without modification.

Properties of Log Likelihood

To set the stage, consider two models of the same form (i.e., the same random component) but of differing parameter spaces. Specifically, suppose we have a *full model* of the form

$$\ell_n(\theta) = \log f(y \mid \theta), \quad \theta \in \Theta,$$

and a *reduced model* of the form

$$\ell_n(\theta_0) = \log f(y \mid \theta_0), \quad \theta_0 \in \Theta_0,$$

where $\Theta_0 \subset \Theta$. This last condition is crucial, and is called the condition of *nested parameter spaces*.

The procedure we are about to discuss only applies to the comparison of nested models. What results in nested parameter spaces is not simply $\Theta_0 \subset \Theta$, but that the parameter θ is the same for both full and reduced models. In particular, models with different random components or response distributions are not amenable to comparison using the procedures of this subsection.

Let $\dim\{\Theta\} = p$ and $\dim\{\Theta_0\} = r$, and

$$\hat{\theta}_n = \sup_{\theta \in \Theta} \ell_n(\theta), \qquad \tilde{\theta}_n = \sup_{\theta \in \Theta_0} \ell_n(\theta).$$

Then, assuming that $\theta \in \Theta_0$ (the reduced model),

$$T_n \equiv -2\left\{\ell_n(\tilde{\theta}_n) - \ell_n(\hat{\theta}_n)\right\} \;\xrightarrow{d}\; \chi^2_{p-r}. \tag{5.27}$$

"As a final comment, we will point out that the likelihood region (5.28) is invariant to parameter transformation, while the Wald theory region of (5.25) is not. This is because the likelihood and log likelihood functions are invariant to parameter transformation."

Numerical Algorithms

"Basic numerical algorithms for optimization can be divided into three broad categories of (1) direct search algorithms, (2) gradient-based algorithms, and (3) Newton-type algorithms."

Equal Interval Search (A Direct Search Method): Typically taken for single dimensions iteratively

Newton-Raphson Algorithm (Newton-Type): "The Newton-Raphson algorithm is most easily understood as a simple application of Newton's method for finding the roots of equations, where that method is applied to a function that is already a first derivative. This algorithm is useful in regular problems in which estimates may be determined by solving the likelihood equations."

Fisher Scoring (Newton-Type): "Equal interval search and Newton-Raphson algorithms have many applications outside of statistical estimation. A Fisher Scoring algorithm, in contrast, is specific to the problem of estimation and, in particular, maximization of a log likelihood function"

Note some terms:

Likelihood Equations

Analytical definition

For a parametric model with density (or pmf) $f(y \mid \theta)$, given independent observations $Y_1, \ldots, Y_n$:

$$L(\theta) = \prod_{i=1}^{n} f(y_i \mid \theta), \quad \ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(y_i \mid \theta).$$

The **likelihood equations** are:

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0.$$

Description

- These equations are solved to obtain the maximum likelihood estimator (MLE) $\hat{\theta}$.
- Sometimes solvable in closed form, often require numerical methods.
- They are the **first-order optimality conditions** for MLE.

Score Function

Analytical definition

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}.$$

For vector parameters, $U(\theta)$ is a gradient vector.

Description

- Measures sensitivity of the log-likelihood to changes in θ.

- At the MLE, $U(\hat{\theta}) = 0$.
- Its variance is the Fisher information, a key quantity in asymptotic inference.

Distinct role

- The **score function** is the actual derivative being set equal to zero in the likelihood equations.
- It is the *bridge* between the log-likelihood and the likelihood equations.

Objective Function

Analytical definition

An **objective function** is the function being optimized (maximized or minimized).

- For MLE:

$$\text{Objective function} = L(\theta) \quad \text{or equivalently} \quad \ell(\theta).$$

- For other estimation methods: could be least squares, penalized likelihood, Bayesian posterior, etc.

Description

- Determines the estimation method.
- For MLE, the objective function is the (log-)likelihood.
- In optimization language:
  - Objective function = target to maximize.
  - Likelihood equations = first-order conditions.
  - Score function = gradient of the objective function.

General Summary Table:

Term	Analytical Form	Description	Distinct Role
Likelihood Equations	$\frac{\partial \ell(\theta)}{\partial \theta} = 0$	System solved for MLE	First-order conditions
Score Function	$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$	Gradient of log-likelihood	Used in likelihood equations
Objective Function	$L(\theta)$ or $\ell(\theta)$	Function being maximized	Target of optimization

Generally:

- **Objective function** = what we optimize.
- **Score function** = derivative of the objective function.
- **Likelihood equations** = set the score function equal to zero to solve for the MLE.

Chapter 6

General Least Squares

Equation (6.1):

$$\min_{\beta \in \mathbb{R}^p} (y - X\beta)^T W (y - X\beta).$$

Equation (6.2):

$$\begin{aligned} (y - X\beta^*)^T W X \beta^* &= 0 \\ \Rightarrow \beta^{*T} X^T W y - \beta^{*T} X^T W X \beta^* &= 0 \\ \Rightarrow X^T W X \beta^* &= X^T W y \\ \Rightarrow (X^T W X)^{-1} X^T W y &= \beta^*. \end{aligned}$$

Equation (6.3):

$$\min_{\beta \in \mathbb{R}^p} [y - g(X, \beta)]^T W [y - g(X, \beta)].$$

OLS (Ordinary)

Equation (6.4):

$$Y_i = x_i^T \beta + \sigma \epsilon_i, \quad i = 1, \ldots, n,$$

where $\beta \in \mathbb{R}^p$, $\sigma > 0$, and $\epsilon_i \sim$ iid F.

Objective function:

$$Q = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2,$$

Equation (6.5):

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

WLS (Weighted)

Equation (6.6):

$$Y_i = x_i^T \beta + \left(\frac{\sigma}{\sqrt{c_i}} \right) \epsilon_i, \quad i = 1, \ldots, n,$$

Objective function:

$$Q = \sum_{i=1}^{n} w_i \, (y_i - x_i^T \beta)^2,$$

Equation (6.7):

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y.$$

GLS (Generalized)

Inference, specifically.

Equation (6.9):

$$\beta_n^{(j)} \overset{AN}{\sim} \mathcal{N}\left(\beta, \frac{\sigma^2}{n}\Sigma_\beta^{-1}\right),$$

Equation (6.10):

$$\Sigma_\beta = \frac{1}{n}\sum_{i=1}^{n} \frac{v(x_i,\beta)v(x_i,\beta)^T}{w_i^2}.$$

Equation (6.11):

$$\hat{\sigma}_n^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left\{\frac{Y_i - g(x_i,\hat{\beta}_n)}{w_i}\right\}^2.$$

Note that except in the special cases leading to models (6.4) or (6.6) this estimator no longer possesses any small sample properties, despite the divisor of $n-p$ which suggests it might be unbiased (it is not). It is, however, consistent as long as $\hat{\beta}_n$ is consistent, which was a condition of the theorem.

Equation (6.12):

$$\hat{C}(\hat{\beta}_n) = \frac{\hat{\sigma}_n^2}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{v(x_i,\hat{\beta}_n)v(x_i,\hat{\beta}_n)^T}{\hat{w}_i}\right]^{-1}.$$

Equation (6.13):

$$\hat{\beta}_{n,k} \pm t_{1-\alpha/2;\,n-p}\left\{\hat{C}(\hat{\beta}_n)_{k,k}\right\}^{1/2}$$

or

$$\hat{\beta}_{n,k} \pm z_{1-\alpha/2}\left\{\hat{C}(\hat{\beta}_n)_{k,k}\right\}^{1/2},$$

where $\hat{C}(\hat{\beta}_n)_{k,k}$ is the k^{th} diagonal element of the estimated covariance matrix given in (6.12), $t_{1-\alpha/2;\,n-p}$ is the $1-\alpha/2$ quantile of a t-distribution with $n-p$ degrees of freedom, and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution.

Note: There is absolutely no justification for using the quantile of a t-distribution in these intervals rather than that of a standard normal distribution. These intervals are constructed on the basis of an asymptotic result, but you will unfortunately see such things done in the literature.

Reading Notes

Chapter 1

"Probability, no matter how it is conceptualized, obeys certain rules of behavior and, hence, mathematical results developed for probability do not depend on what exactly one believes it is."

"Probability is **not a thing** but a **concept**."

Question: For a given type of statistical analysis, is the "type" of probability concept you use "locked in place" throughout the analysis?

1. Direct Search Algorithms.

Direct search algorithms are characterized by requiring computation of only the relevant objective function and not any derivatives of that function. Thus, direct search algorithms are useful in problems for which derivatives of the objective function are difficult to compute, or in which we do not need derivatives for the purpose of calculating inferential quantities.

2. Gradient-Based Algorithms.

Gradient algorithms provide the same information as direct search algorithms along with the value of the gradient, which should be zero at the maximum. They are generally more efficient than direct search algorithms in terms of the number of function evaluations needed, at the cost of requiring computer functions to be written for evaluation of the first derivatives of the objective function, and typically need the first derivatives to be continuous as well.

3. Newton-type Algorithms. Newton-type algorithms make use of information provided by not onlythe objective function and gradient, but also the second derivatives of the objective function.

In the case that the objective function is a full log likelihood, they provide the benefit of including the observed information matrix as part of the output, which can make inference easier if an approach based on Wald theory is to be used.

Chapter 2

"A model is a set of invented assumptions regarding invented entities such that, if one treats these invented entities as representations of appropriate elements of the phenomena studied, the consequences of the hypotheses constituting the model are expected to agree with observations." - Neyman, 1957, allegedly

Focus of the course will be on *parametric modelling*

"All that is needed is for the actual assignment of treatments to experimental units to be one of a set of a known number of equally likely arrangements."

Question: XU applied to treatments vs. Treatments applied to XU?

"The construct of weight (a covariate of a statistical model) does not depend on the particular measurement tool used to observe it, nor does it depend on a set of physically real outcomes of an observation process." "The majority of scientific investigations (experiments or otherwise) are based on the concept that there exists a mechanism that underlies the production of observable quantities."

"In many, if not most areas of science, mechanisms are not fully understood."

"The upshot of... for the purpose of statistical modeling is that scientific mechanisms or repeatable phenomena represent the key elements of a problem to be captured in a small set of model parameters... by this we do not mean a direct translation of a mechanism into mathematical terms."

"A criticism that is sometimes leveled at the modeling approach is that a model"doesn't care where the data come from", or "how the data were obtained"... however... no model can properly operate beyond the context given it by the process of statistical abstraction which, in a proper application, must have been given careful consideration."

"Observed values of random variables" is not strictly speaking, a valid notion; what we actually mean is that "data represent possible values that might be assumed by random variables."

"Distributions within a class share various statistical properties (which properties these are depends on which class)."

"The major impact of generalized linear models was to promote consideration of random and systematic model components rather than signal plus noise."

Chapter 3

"To adequately model the probabilistic behaviors of random variables, we must have access to a variety of theoretical probability distributions."

"Thus, what we usually refer to as the normal distribution constitutes a family of distributions, as does the Poisson distribution and many others"

"It is important to understand the context being used in a conditional statement so that, for example, the distinction between $E(Y|x)$ and $E(Y|X)$ is clear" (One is an observation of the R.V. X, the other is conditional on that random variable)

(Regarding Location-Scale Families): "What can be called the *standard form* of a distribution is the distribution that results from eliminating parameters."

(Regarding Location-Scale Families): "location-scale families of distributions that have standard forms with expectation 0, variance 1, and support on the entire line are the traditional building blocks for models formulated in terms of what we will come to call a *signal plus noise structure.*"

(On signal-plus-noise, spec. Normality assumptionms for inference): "But we may desire estimation of certain quantiles for responses, or in the probability that a response will exceed some regulatory threshold at a given covariate value, and those questions depend on more than expectations alone."

In the final expression of (3.4) the parameters denoted as θ_j, $j = 1, \ldots, s$ are called **canonical** or sometimes **natural** parameters for the exponential family.

While the canonical parameterization usually leads to the easiest derivation of properties such as those just given, it is not always the best parameterization for purposes of estimation, inference, or model interpretation.

(On Mean Value Parameterization 1): It may be the case that none of the canonical parameters θ_j in (3.4) correspond to the expected value of the random variable Y.
A **mean value parameterization** can be accomplished by a transformation

$$(\theta_1, \ldots, \theta_s) \quad \longrightarrow \quad (\mu, \phi_1, \ldots, \phi_{s-1}),$$

where $\mu \equiv \mathbb{E}(Y)$ and $\phi_1, \ldots, \phi_{s-1}$ are arbitrarily defined.
We will still need s parameters because we are assuming the canonical representation is minimal, as defined previously.

Note, however, that the reparameterized family may no longer be in canonical form.

(Mean Value Parameterization 2): In the canonical parameterization for exponential families there is a clear association between parameters θ_j and sufficient statistics T_j.

It is perhaps natural then to attempt to parameterize families using the expected values of the T_j, which are given in property 5 of the previous section as first derivatives of the function $B(\theta)$.

Thus, we transform

$$(\theta_1, \ldots, \theta_s) \quad \longrightarrow \quad (\mu_1(\theta), \ldots, \mu_s(\theta))$$

where

$$\mu_j(\theta) = \mathbb{E}\{T_j(Y)\} = \frac{\partial}{\partial \theta_j} B(\theta).$$

This parameterization has the potential advantage that each parameter of the density is then the expected value of an element of the complete sufficient statistic, namely $T_j(Y)$, which then immediately give us UMVU estimators for the parameters $\mu_j(\theta)$.

The relevant question is whether such parameters represent quantities that are meaningful for inference.

Families with this structure are among the more commonly used distributions in many types of models such as generalized linear models.

Reasons for Choosing a Particular Parametrization

(1): "It is possible, however, that with estimation by exact theory or least squares one might need to conduct a transformation before estimation to allow inference to be made on the transformed parameters."

(2): "Parameter transformations are sometimes conducted to produce increased stability in numerical estimation procedures."

(3): "[I]n model formulation, a primary goal is to connect the key elements of a scientific problem with parameters of a probabilistic model. It can occur that one parameterization makes this more clearly the case than does an alternative."

(4): "A more easily comprehended goal of parameterization is to sometimes clearly identify how covariate information can appropriately be incorporated into a model."

(5): 5. In the investigation of different parameterizations it is essential that one keep track of possible restrictions on the parameter space, both in terms of allowable values and in terms of restrictions that may be imposed on one parameter component (e.g., θ_2) by the value of another (e.g., θ_1).

Such restrictions (including possibly the lack of such restrictions) can render a parameterization either more or less appropriate to describe a given situation.
From a purely statistical viewpoint, it seems pleasing to have parameter elements that are **variation independent**.

A generic vector-valued parameter $\theta \equiv (\theta_1, \theta_2)$ has variation independent components if the parameter space can be written as the Cartesian product $\Theta = \Theta_1 \times \Theta_2$ where Θ_1 and Θ_2 are sets of possible values for θ_1 and θ_2, respectively.

While having variation independent parameters is probably typical of distributions we are familiar with, it is worth noting this property.
In models that have multiple random components, such as hierarchical models, variation independent parameters in the data model translate into something called the **positivity condition** for modeling random parameter values.

Formulating a proper model can become much more difficult when this condition does not hold.

(Break)

(Exponential Dispersion Family): "This particular subclass of exponential dispersion families is, however, arguably one of the most common forms of exponential family distributions that appear in applications."

(Exponential Dispersion Family):

"An important role is played in both the theory and application of exponential family distributions by one-parameter families for which the sufficient statistic is $T(y) = y$."

"These are often called **natural exponential families**, following the extensive investigation of their behavior by Morris (1982, 1983). If a family of distributions has only one canonical parameter, then both the expectations and variances of those distributions must be functions of the sole parameter."

Chapter 4

Chapter 5

"Recall from Chapter 1 that a statistical model must lead to a joint probability distribution for the entire collection of random variables involved in a problem."

Likelihood Function and Joint Distribution Function: Same Formula, but Different Functions

- Joint Distribution: Fixed parameter, function of Data
- Likelihood: Fixed data, function of Parameter

"While likelihoods are not necessarily equal to probabilities, there is a connection between likelihood and probability."

When the context is that of a given set of observed data, we often write the likelihood without explicit conditioning on those data as simply $L(\theta)$ or $\ell(\theta)$. On the other hand, if the context involves probabilistic behavior of the likelihood we may write the random version as $L(\theta|Y)$ or $\ell(\theta|Y)$.

That is, having the value of the parameter that maximizes the probability of seeing what we saw certainly justifies the maximum likelihood estimate (mle) as a summarization of the available data, but it does not necessarily indicate that maximum likelihood is a good procedure for estimation of the parameter of interest θ.

This provides a connection between a maximum likelihood estimate and the true parameter value in a hypothetical model. That is, as the sample size increases, the parameter value that maximizes the joint distribution not only provides a good value for describing the observations at hand, but also must become close to the true value under a given model.

"Asymptotic normality refers to the convergence in distribution of a suitably centered and scaled sequence of statistics to a standard normal distribution"

"It is important to understand that efficiency is a property of estimators, not estimates. So the fact that an estimator is efficient does not imply that an estimate produced from a set of data somehow has optimal properties. An estimate has no properties at all and cannot be claimed to be accurate or biased, precise or imprecise, it is simply a numerical value whose relation to the true parameter is unknown. Properties such as precision are properties of procedures or estimators only and statements of those properties such as (5.14) and (5.15) involve limiting quantities that can never be used or even approximated in practice. We can have some level of confidence or comfort about a particular estimate only because we know the tool used to produce it had good or optimal properties such as efficiency or asymptotic efficiency."

"We can identify four issues connected with the development of asymptotic results for estimators obtained as solutions to the likelihood equations, (i) consistency of a sequence of likelihood equation solutions, (ii) demonstration that asymptotic normality holds for such a sequence, (iii) verification that the information inequality holds for such a sequence, and (iv) uniqueness of such a sequence."

"Properties of estimators are developed under sets of conditions called regularity conditions. There are a variety of regularity conditions that have been developed, and different sets of conditions are needed to prove different results about likelihood-based estimators."

"In the independent but not identically distributed case, conditions that allow a central limit theorem for non-identically distributed random variables is needed."

"Full maximum likelihood estimation for problems involving spatial and spatio-temporal models is often difficult at best and alternative methods may be employed for estimation and inference, but likelihood-based estimation and inference are practical in a some cases involving Gaussian distributions"

From this (exponential family of distributions) and Corollary 2.1 we have that the MLE of θ, if it exists, is

- (i) unique,

- (ii) uniform minimum variance unbiased (UMVU),
- (iii) asymptotically normal, and
- (iv) asymptotically efficient.

That the MLE is UMVU follows from the fact that the $\{T_j : j = 1, \ldots, s\}$ are minimal sufficient for θ and the MLE is unbiased and a function of these statistics. That the MLE is UMVU gives an optimal exact-theory property, but the sampling distribution is not immediately available, except in the case of normal distributions, for which we typically do not rely on likelihood asymptotics for inference.

"The regular theory does not apply in the case where the support of the distribution of response variables depends on the value of the parameter."

(Inference from Properties of thje Log Likelihood) "The title is given, however, to distinguish inference based on the asymptotic normality of maximum likelihood estimates (i.e., Wald Theory) from the topic of this section, which is inference based on asymptotic properties of the log likelihood function itself. The basis of this type of inference is the asymptotic distribution of the likelihood ratio statistic."

(Nunmerical Algorithms): "In the vast majority of problems for which we might choose to conduct analysis based on likelihood theory, the likelihood or log likelihood cannot be maximized analytically"

"If required, maximization can be accomplished by minimizing the negative objective function and it might be noted that literature on numerical analysis usually takes the problem to be minimization rather than maximization."

(Equal Interval Search, a type of Direct Search Algorithm) "Equal interval search algorithms also have application in Bayesian analyses for which we need to quickly locate the mode of a (posterior) distribution in one dimension"

Chapter 6

"The fundamental domain of application for least squares estimation is linear models with constant variance additive errors, for which least squares estimators typically have optimal small sample properties, at least within a restricted class of estimators. Note that linear models and linear estimators refer to different things, models that are linear functions of additive errors versus estimators that are linear combinations of response variables, although the two often go hand in hand."

"We can now state the projection theorem in a more general form. For the majority of statistical applications we can take the Hilbert space in this theorem to be $\mathbb{R}^n$ and the inner product to be the ordinary dot product with respect to a matrix of weights W."

"Take note of the fact that, the exact theory results in this case lead to t-distributions as *results* so that it is entirely appropriate and correct to use quantiles of these distributions for interval estimation."

(WLS): "A studentization of elements of $\hat{\beta}$ then again results in t-distributions which are used to produce intervals and other inferential quantities."

(GLS): "Formulation of the least squares problem will depend on the combination of expectation function and variance model structure under consideration. For example if the expectation function is linear with parameters β and the variance model depends on β as well as σ^2, then the least squares problem becomes similar to (6.1) except with a weight matrix $W(\beta)$ rather than a fixed W that is free of parameters. If the expectation function is nonlinear $g(X, \beta)$ but the variance model is constant depending only on σ^2, then the least squares problem is given by (6.3). If the expectation function is nonlinear and the variance model depends on β then the least squares problem is as in (6.3) except with $W(\beta)$ rather than W."

(GLS Inference):

Thus, inference is approximate, based on a result that has been called the *Fundamental Theorem of Generalized Least Squares*. The context for this theorem is an additive error model such that

$$E(Y_i) = g(x_i, \beta) \quad \text{and} \quad \text{var}(Y_i) = w_i^{-1/2}\sigma^2,$$

where g is a known smooth function and w_i may be a constant or a function of β.

The quantities $\mathbf{V}(\beta)$ and $\mathbf{W}(\beta)$ are as previously defined, and we now include an index for sample size n to emphasize that we are concerned with the behavior of sequences of estimators as sample size grows without bound.

An estimator $\hat{\theta}_n$ is $n^{1/2}$–consistent for a parameter θ if $n^{1/2}(\hat{\theta}_n - \theta)$ is bounded in probability (meaning that $\hat{\theta}_n$ converges to θ at least at the rate $1/n^{1/2}$).

Given this, the stated asymptotic normality holds for estimators that result from *any number* of iterations of the algorithm, and there are proponents for various choices.

Summary of Chapter 6:

1. Least squares is used nearly exclusively with additive error models and is concerned primarily with parameters of the expectation function or systematic model component.

2. Least squares can be motivated in a number of ways, including as the solution to a general minimization problem in linear algebra.

3. For linear models with either constant variance or variances that are proportional to known weights, least squares estimators have exact theory properties. In particular, they are UMVU estimators, and may be used in conjunction with UMVU estimators of the variance parameter σ^2. Inferential procedures in these situations are typically developed under the additional assumption of normally distributed errors so that estimated marginal sampling distributions of expectation function parameters are t–distributions, which are then used to construct intervals and tests.

4. For nonlinear models with constant variance, or for either linear or nonlinear models with variances that depend on parameters of the expectation function and σ^2, but no additional unknown parameters, generalized least squares estimators are asymptotically normal. Generalized least squares estimators are typically used in conjunction with consistent estimators of σ^2 developed from a moment-based approach. Intervals for individual parameter elements may be based on this asymptotic normality. Note that the development of the approximate sampling distribution of expectation function parameters does *not* depend on the additional model assumption of normally distributed errors. It does, however, require reliance on an asymptotic result.

5. Putting together the information in items 3 and 4 immediately above, we arrive at the "no free lunch" conclusion for estimation by least squares methods. Avoiding strong distributional assumptions on model terms is often considered a good thing. Being able to develop exact properties for estimators that do not depend on asymptotic arguments is often considered a good thing. Under models for which we can apply ordinary or weighted least squares we can accomplish both but only up to the point of verifying optimal behavior of *point* estimators. We then generally rely on strong distributional assumptions for *inference*. Under models for which we turn to generalized least squares we can avoid strong distributional assumptions on the model entirely, but must rely on asymptotic results for both properties of point estimators and inferential procedures.

6. The ability to develop properties for least squares estimators, either exact theory for point estimation or asymptotic theory for both point estimation and inference, without assuming a specific parametric form for model distributions is often considered a robustness property or aspect of least squares, and this is true insomuch as robustness refers to small departures from an assumed distributional form. But this concept of robustness is different than what is properly called *resistance*, which refers to the degree to which an estimator is affected by extreme observations. It is well known that, while least

squares estimators have a certain amount of robustness, they are extremely sensitive to the effect of extreme and high leverage observations and thus have low resistance.