

17

Parametric and Non-Parametric Linkage Analysis

17.1: Learning Objectives

By the end of this chapter, the reader should be able to state the assumptions of parametric (LOD-score, model-based) linkage analysis, compute a pedigree likelihood and LOD score for a simple family, describe non-parametric (model-free) linkage methods based on allele sharing, and compare the two approaches in the context of Nigerian family studies of Mendelian versus complex disease.

17.2: From Two-Point Recombination to Full Pedigree Linkage Analysis

Chapter 8 introduced the recombination fraction, θ , and its estimation from double-backcross data. Real human family (pedigree) data are rarely as simple as a double backcross: families vary in size and structure, not every individual's genotype or phase is directly observable, and the mode of inheritance of the disease itself (dominant, recessive, penetrance, phenocopy rate) must typically be specified or estimated. **Parametric linkage analysis** extends the LOD-score approach of Chapter 8 to general pedigrees, under an assumed genetic model for the disease.

17.3: Assumptions of Parametric Linkage Analysis

A typical parametric linkage analysis specifies, in advance (or estimates via segregation analysis, Chapter 7):

1. The **mode of inheritance** (e.g., autosomal dominant, autosomal recessive, X-linked).
2. The **disease allele frequency** in the population.
3. The **penetrance** for each genotype (probability of expressing the disease phenotype given the genotype) — ideally complete penetrance (=1) and no phenocopies (unaffected individuals never sharing the disease genotype), though reduced-penetrance models are also used.
4. A specific pedigree structure with observed marker and (partially observed) disease phenotype data.

17.4: The Pedigree Likelihood

For a pedigree of n individuals, let X_i denote the observed phenotype of the i th individual and G_i their (unobserved, or partially observed) genotype at the disease locus and/or marker locus jointly. The general pedigree likelihood, as a function of the recombination fraction θ between the disease locus and a genetic marker, sums (integrates) over all genotype configurations consistent with the observed data:

$$\begin{aligned}
 L(\theta) = & \sum_{G_1} \cdots \sum_{G_n} \left[\prod_i \text{pen}(X_i | G_i) \right] \\
 & \times \left[\prod_{j \in \text{founders}} \text{prior}(G_j) \right] \\
 & \times \left[\prod_{m,k,l} \text{trans}(G_m | G_k, G_l) \right]
 \end{aligned}$$

where $\text{pen}(X_i | G_i)$ is the penetrance function (probability of the observed phenotype given genotype), the product over founders (individuals with no parents in the pedigree) uses the population prior probability of each founder genotype (typically from Hardy–Weinberg proportions, Chapter 3), and the product over parent–offspring transmissions (m = offspring, k, l = parents) applies Mendelian transmission probabilities, which depend on θ when the disease locus and marker locus are jointly tracked. This general formula, though notationally dense, is simply the systematic bookkeeping of Mendelian segregation probabilities (Chapter 2) and penetrance across an entire family tree; in practice it is evaluated by efficient recursive algorithms (the **Elston–Stewart algorithm**, well suited to pedigrees without loops, and the **Lander–Green algorithm**, well suited to many markers analysed jointly) rather than by brute-force summation.

17.5: Two-Point LOD Score Calculation: Procedure

For a simple pedigree in which recombinant and non-recombinant meioses can be directly counted (as in Chapter 8’s double backcross, or its human-pedigree analogue), the LOD score at a trial value of θ is:

$$Z(\theta) = \log_{10}[\theta^R(1 - \theta)^{N_R} / 0.5^{R+N_R}]$$

where R is the number of recombinant meioses and N_R the number of non-recombinant meioses observed across the pedigree(s).

Worked Example 17.1.

A Nigerian family linkage study of an autosomal dominant disorder finds, across several informative pedigrees, $R = 1$ recombinant and $N_R = 7$ non-recombinant meioses ($n=8$ total).

(a) Compute $Z(\theta)$ at $\theta = 0.1$. $\theta^{R(1-\theta)}\{N_R\} = (0.1)^{1(0.9)}7 = 0.1 \times 0.4783 = 0.04783$. $0.5^8 = 0.003906$. $Z(0.1) = \log_{10}(0.04783/0.003906) = \log_{10}(12.246) = 1.088$.

(b) Find $\hat{\theta}$ (the MLE of θ) and $Z(\hat{\theta})$. By Chapter 8's binomial MLE result, $\hat{\theta} = R/(R+N_R) = 1/8 = 0.125$. $\hat{\theta}^{1(1-\hat{\theta})}7 = (0.125)(0.875)^7 = 0.125 \times 0.3927 = 0.04909$. $Z(0.125) = \log_{10}(0.04909/0.003906) = \log_{10}(12.569) = 1.099$.

(c) Interpret. The maximum LOD score across pedigrees, $Z(\hat{\theta}) \approx 1.10$, falls well short of the $Z \geq 3$ threshold conventionally required to declare significant linkage (Chapter 8), so with only 8 informative meioses this single-family (or small-multi-family) study provides, at best, weak/suggestive evidence and would need to be extended — either by recruiting additional Nigerian families segregating the disorder, or by combining LOD scores additively across independently ascertained pedigrees at the same fixed θ (a key practical advantage of parametric LOD scores: unlike many non-parametric statistics, Section 17.7, LOD scores at a fixed θ are simply additive across independent families, greatly simplifying multi-family/multi-centre linkage studies).

17.6: LOD Score Interpretation Thresholds (Recap and Extension)

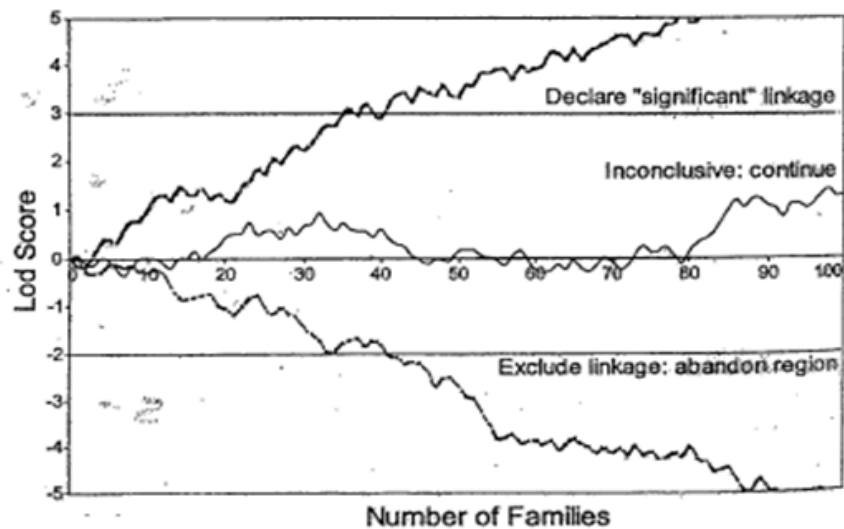


Figure 7.19 Hypothetical realizations of the accumulation of lod scores over time, assuming true linkage (upper curve), a gene of some given magnitude not linked to the marker (bottom curve), or no gene (middle curve). The conventional stopping rules of Morton (1955) are shown as horizontal dotted lines.

LOD score curves accumulating evidence for linkage as more families are added to a genome scan.

As introduced in Chapter 8, $Z \geq 3$ is the conventional threshold for “significant” evidence of linkage in a genome-wide parametric linkage scan (chosen to control the genome-wide false-positive rate given the very large number of loci effectively tested across the genome, an early precursor of the even more stringent genome-wide significance thresholds used in GWAS, Chapter 18); $Z \geq 2$ is often termed “suggestive”; $Z \leq -2$ excludes linkage at that θ . Maximising $Z(\theta)$ over θ (rather than testing at a single fixed θ) also constitutes a formal likelihood ratio test (Chapter 10) of $H_0: \theta=0.5$ vs $H_1: \theta<0.5$, subject to the boundary caveat noted in Section 10.7.

Theorem 17.1 (Additivity of LOD Scores Across Independent Families)

If $Z_1(\theta), \dots, Z_K(\theta)$ are the LOD scores computed from K independently ascertained families at the same trial value θ , then the combined LOD score for the pooled data is $Z_{combined}(\theta) = \sum_{k=1}^K Z_k(\theta)$.

Proof. By definition (Section 8.5), $Z_k(\theta) = \log_{10}[L_k(\theta)/L_k(0.5)]$, where $L_k(\theta)$ is family k 's likelihood at recombination fraction θ . Because the K families are ascertained and genotyped independently, their joint likelihood is the product of the individual likelihoods (a basic consequence of independence, Chapter 10): $L_{combined}(\theta) = \prod_{k=1}^K L_k(\theta)$. Hence:

$$\begin{aligned} Z_{combined}(\theta) &= \log_{10} \left[\frac{L_{combined}(\theta)}{L_{combined}(0.5)} \right] \\ &= \log_{10} \left[\frac{\prod_k L_k(\theta)}{\prod_k L_k(0.5)} \right] = \log_{10} \prod_{k=1}^K \frac{L_k(\theta)}{L_k(0.5)} \\ &= \sum_{k=1}^K \log_{10} \left[\frac{L_k(\theta)}{L_k(0.5)} \right] = \sum_{k=1}^K Z_k(\theta), \end{aligned}$$

using the logarithm identity $\log(\prod_k x_k) = \sum_k \log(x_k)$. ■

This theorem is the rigorous justification for the “combining families” practice used throughout Chapters 8 and 17: because LOD scores are logarithms of likelihood ratios, and independent likelihoods multiply, LOD scores simply add — no re-analysis of the pooled raw data is even needed if each family's LOD curve has already been computed separately.

Worked Example 17.3 (Combining three Nigerian extended families for a Mendelian ataxia linkage study)

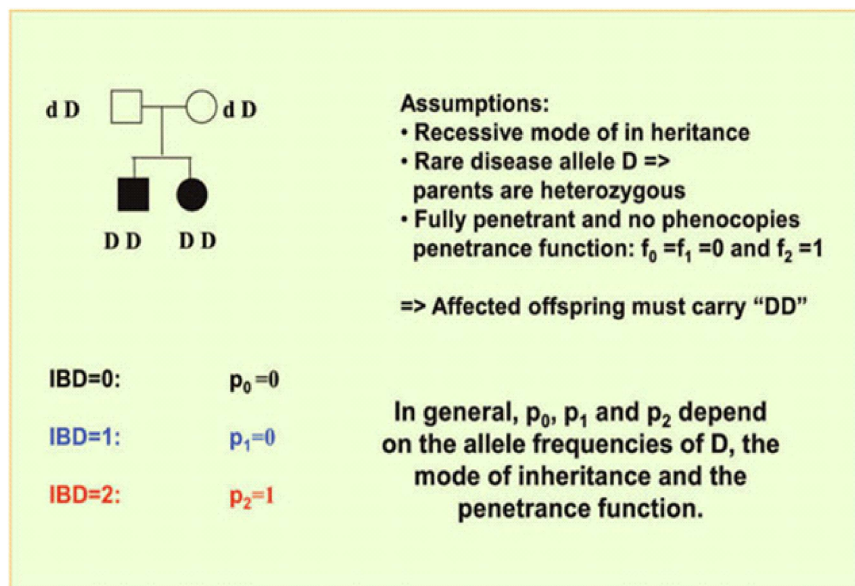
A neurogenetics collaboration pools linkage data on a rare autosomal dominant ataxia from three extended Nigerian families, each genotyped independently at a candidate microsatellite marker. At $\theta = 0.10$, Family A gives $Z_A(0.10) = 1.8$, Family B gives $Z_B(0.10) = 0.9$, and Family C, unfortunately uninformative at this marker by chance, gives $Z_C(0.10) = -0.3$ (weak evidence against linkage at this specific θ , though not necessarily against linkage overall). By Theorem 17.1, the combined LOD score is $Z_{combined}(0.10) = 1.8 + 0.9 - 0.3 = 2.4$ — still short of the formal significance threshold of 3 (Section 17.6) but comfortably in the “suggestive” range, illustrating how even one weakly discordant family does not necessarily overturn a combined signal from several concordant families, and motivating the collaboration’s plan to recruit a fourth family before drawing a final conclusion.

17.7: Non-Parametric (Model-Free) Linkage Analysis

Non-parametric linkage (NPL) analysis avoids specifying a genetic model (mode of inheritance, penetrance, disease allele frequency) altogether — an important practical advantage for **complex diseases** (e.g., type 2 diabetes, hypertension, or asthma, all of major public-health importance in Nigeria) for which the mode of inheritance is unknown, likely polygenic, and not well described by a simple Mendelian model. Instead, NPL methods test whether **affected relatives** (e.g., affected sibling pairs) share genetic material **identical by descent (IBD)**, Chapter 6, at a candidate marker **more often than expected by chance** under Mendelian segregation alone.

For a pair of full siblings, the number of alleles shared IBD at a given autosomal locus (0, 1, or 2) follows, under the null hypothesis of no linkage to any disease locus, the standard

Mendelian distribution: $P(\text{IBD}=0) = 1/4$, $P(\text{IBD}=1) = 1/2$, $P(\text{IBD}=2) = 1/4$ (a direct consequence of independent segregation at the two parental meioses producing each sibling — Chapter 2). If the marker is linked to a disease locus, **affected** sibling pairs are expected to share IBD status **more** than this null expectation (because they are more likely to have jointly inherited the same disease-predisposing chromosomal region from a parent), and the mean IBD sharing among affected sib pairs is compared statistically (e.g., via a one-sided test of mean $\text{IBD} > 1$, or via the widely used **S_pairs** and related NPL statistics implemented in software such as GENEHUNTER and MERLIN) to this null expectation.



An affected sibling pair pedigree with parental genotypes ($4D \times 4d$), showing the observed vs. expected number of alleles shared identical by descent ($\text{IBD} = 0, 1$, or 2) under the null hypothesis of no linkage — the basic data structure for a non-parametric affected-sib-pair linkage test.

The maximum LOD score (MLS) statistic. The formal test summarising evidence for excess IBD sharing across many affected sibling pairs is the **maximum LOD score (MLS)**, constructed, in a spirit paralleling the parametric LOD score of Section 17.5, as a likelihood ratio comparing the IBD-sharing probabilities that best fit the observed data against their null (no-linkage) Mendelian values:

Maximum lod score (MLS)

Maximum likelihood estimates:

$$\hat{p}_0 = \frac{n_0}{n} \quad \hat{p}_1 = \frac{n_1}{n} \quad \hat{p}_2 = \frac{n_2}{n}$$

Number of Alleles Shared IBD

Likelihood ratio test:

$$\frac{L(\hat{p}_0, \hat{p}_1, \hat{p}_2)}{L(1/4, 1/2, 1/4)} = \frac{\hat{p}_2^{n_2} \hat{p}_1^{n_1} \hat{p}_0^{n_0}}{0.25^{n_2} 0.5^{n_1} 0.25^{n_0}}$$

$$2\log\left(\frac{L(\hat{p}_0, \hat{p}_1, \hat{p}_2)}{L(1/4, 1/2, 1/4)}\right) \sim \chi^2_2$$

	0	1	2	Total
Observed	n_0	n_1	n_2	n
Expected	$n/4$	$n/2$	$n/4$	n

The maximum LOD score (MLS) is defined by:



$$\begin{aligned} MLS &= \log_{10}(L(\hat{p}_0, \hat{p}_1, \hat{p}_2)/L(1/4, 1/2, 1/4)) \\ &= \log_{10}(L(\hat{p}_0, \hat{p}_1, \hat{p}_2)) - \log_{10}(L(1/4, 1/2, 1/4)) \end{aligned}$$

The maximum LOD score (MLS) framework: maximum likelihood estimates of the IBD-sharing probabilities z_0, z_1, z_2 are compared against their null values (1/4, 1/2, 1/4) via a likelihood ratio test, asymptotically distributed as a 50:50 mixture of a point mass at zero and a chi-square(1) distribution under the null.

Because the MLS statistic, like the ordinary LRT of Chapter 10, tests a null hypothesis ($z_0 = 1/4, z_1 = 1/2, z_2 = 1/4$) that lies at the *boundary* of the parameter space allowed under the alternative (excess sharing can only push z_2 upward and z_0 downward, not the reverse, under a genuine disease-linked locus), its null distribution is the same non-standard 50:50

mixture distribution encountered in Chapter 10's discussion of boundary-parameter likelihood ratio tests, rather than a simple χ^2 distribution.

Worked Example 17.2

In a Nigerian family study of hypertension (a complex trait), 100 affected sibling pairs are genotyped at a candidate marker; the observed distribution of IBD sharing is: IBD=0 in 15 pairs, IBD=1 in 45 pairs, IBD=2 in 40 pairs.

(a) Compute the mean IBD sharing observed, and compare with the null expectation.

Mean IBD = $[0(15) + 1(45) + 2(40)] / 100 = (0+45+80)/100 = 125/100 = 1.25$.

Null expected mean IBD (no linkage) = $0(0.25) + 1(0.50) + 2(0.25) = 0 + 0.50 + 0.50 = 1.00$.

(b) **Interpret:** The observed mean IBD sharing (1.25) exceeds the null expectation (1.00), consistent with linkage between this marker and a locus influencing hypertension susceptibility in these Nigerian sibships; a formal significance test (e.g., comparing the observed IBD distribution to the null 1:2:1 distribution via a chi-square or a purpose-built NPL statistic, and typically also correcting for the marker's information content given imperfect parental genotyping) would be required before drawing a definitive linkage conclusion, but this simple mean-IBD comparison already illustrates the basic logic of non-parametric, allele-sharing-based linkage analysis.

17.8: Parametric versus Non-Parametric: When to Use Which

	Parametric (score)	(LOD	Non-parametric (allele sharing)
Requires specified genetic model	Yes		No
Best suited to	Simple Mendelian disorders (e.g., a clearly dominant/recessive Nigerian pedigree disorder)		Complex, polygenic traits (e.g., T2DM, hypertension)
Statistical power (if model correct)	Higher		Lower
Robustness to model misspecification	Lower (misspecified model can mask true linkage)		Higher
Scores additive across families?	Yes, at fixed θ		Generally no (families combined via other pooling methods)

Theorem 17.2: (The LOD Score Is a Monotonic Transformation of the Likelihood Ratio)

The LOD score $Z(\theta) = \log_{10} [L(\theta) / L(0.5)]$ is a strictly increasing function of the likelihood ratio $L(\theta)/L(0.5)$, and consequently the value of θ that maximises the LOD score is identical to the value of θ that maximises the likelihood $L(\theta)$, namely the maximum likelihood estimator $\hat{\theta}$.

Proof. The base-10 logarithm, $\log_{10}(x)$, is a strictly increasing function on $(0, \infty)$, since its derivative $1/(x \ln 10)$ is strictly positive for all $x > 0$. The LOD score is defined as $Z(\theta) = \log_{10}[\text{LR}(\theta)]$, where $\text{LR}(\theta) = L(\theta)/L(0.5)$ is the likelihood ratio comparing linkage at recombination fraction θ to no linkage ($\theta = 0.5$). Because Z is a strictly increasing transformation of $\text{LR}(\theta)$, and $\text{LR}(\theta)$ is itself a strictly increasing transformation of $L(\theta)$ for fixed $L(0.5)$ (a positive constant not depending on θ), it follows that $Z(\theta)$ attains its maximum over θ at exactly the same value of θ at which $L(\theta)$ attains its maximum, namely $\hat{\theta}$, the maximum likelihood estimator from Section 8.4. This confirms that the two-point LOD score procedure (Section 17.5) is not merely a convenient rescaling for interpretation, but is genuinely tied to a Neyman-Pearson-optimal likelihood ratio test at every fixed θ , inheriting the general asymptotic optimality properties of likelihood ratio-based inference established in Chapter 10.

17.9: Chapter Summary

- Parametric linkage analysis extends two-point LOD-score methods (Chapter 8) to full pedigrees via the pedigree likelihood, requiring a specified mode of inheritance, allele frequency, and penetrance.
- LOD scores at fixed θ are additive across independent pedigrees, easing multi-family studies; $Z \geq 3$ remains the conventional significance threshold.
- Non-parametric linkage analysis tests for excess allele sharing (identical by descent) among affected relatives, without assuming a genetic model — well suited to complex Nigerian diseases like hypertension and T2DM.

17.10: Computational Corner: Computing Two-Point LOD Scores

R.

```
# Two-point LOD score from meiosis counts (Section 17.5),
# across a grid of theta
r_recomb <- 6; n_meioses <- 20 # 6 recombinants observed out
# of 20 informative meioses

lod_score <- function(theta, r, n) {
  (r * log10(theta) + (n - r) * log10(1 - theta)) -
  (r * log10(0.5) + (n - r) * log10(0.5))
}
theta_grid <- seq(0.01, 0.49, by = 0.01)
lods <- sapply(theta_grid, lod_score, r = r_recomb, n =
n_meioses)
theta_hat <- theta_grid[which.max(lods)]
cat("theta_hat (grid) = ", theta_hat, " max LOD = ",
max(lods), "\n")
plot(theta_grid, lods, type = "l", xlab = expression(theta),
ylab = "LOD score")
abline(h = 3, col = "red", lty = 2) # conventional linkage
threshold
```

Python.

```
import numpy as np

r_recomb, n_meioses = 6, 20

def lod_score(theta, r, n):
    return (r * np.log10(theta) + (n - r) * np.log10(1 - theta)) - \
           (r * np.log10(0.5) + (n - r) * np.log10(0.5))

theta_grid = np.arange(0.01, 0.5, 0.01)
```

```

lods = lod_score(theta_grid, r_recomb, n_meioses)
theta_hat = theta_grid[np.argmax(lods)]
print(theta_hat, lods.max())

```

Both scripts trace out the full LOD-score curve of Section 17.5 across candidate recombination fractions and mark the conventional genome-wide significance threshold of $\text{LOD} \geq 3$, so the reader can see directly whether a given pedigree's data would, on their own, meet the classical Morton criterion for declaring significant linkage.

Mini-Project

Mini-Project 17: Two-Point Linkage Analysis Across a Chromosome Scan

Compute two-point LOD scores across a simulated chromosomal region and identify the region of strongest linkage evidence.

- Step 1:** Simulate pedigree data with a disease locus linked to one of 10 evenly spaced markers along a chromosome.
- Step 2:** Compute the two-point LOD score (Section 17.5) between the disease phenotype and each of the 10 markers.
- Step 3:** Plot LOD score against marker position to produce a simple linkage scan.
- Step 4:** Identify the marker(s) exceeding the conventional genome-wide significance threshold of 3.0 (Section 17.6).
- Step 5:** Write a short paragraph comparing what this parametric scan would show versus a non-parametric (model-free) analysis of the same data (Section 17.7).

Deliverable: *Simulation and LOD score computation code, a LOD-by-position plot, and a short written comparison of parametric and non-parametric approaches.*

17.11: Practice Questions

- Q1.** A Nigerian pedigree study of an X-linked recessive disorder finds $R = 2$ recombinants and $N_R = 18$ non-recombinants across informative meioses. Compute θ , and $Z(\theta)$.
- Q2.** Two independent Nigerian families studying the same autosomal dominant disorder give LOD scores of $Z_1(0.1) = 1.2$ and $Z_2(0.1) = 0.9$ at $\theta=0.1$. What is the combined (multi-family) LOD score at $\theta=0.1$, and what property of LOD scores allows this simple combination?
- Q3.** In a sib-pair study of asthma, 60 affected sibling pairs show IBD=0 in 10 pairs, IBD=1 in 20 pairs, IBD=2 in 30 pairs at a candidate marker. Compute the mean IBD and compare with the null expectation of 1.0.
- Q4.** Explain why non-parametric linkage analysis is generally preferred over parametric analysis for a trait like type 2 diabetes in a Nigerian family study, but parametric analysis remains preferred for a clearly Mendelian disorder such as sickle-cell disease.
- Q5.** Twenty affected sibling pairs are genotyped at a candidate marker, and the observed mean number of alleles shared IBD is 1.35 (compared with a null expectation of 1.0). Explain, conceptually, whether this observation is consistent with, or provides evidence against, linkage between the marker and a disease locus.
- Q6.** Explain why the null distribution of the maximum LOD score (MLS) statistic (Section 17.7) is a 50:50 mixture of a point mass at zero and a chi-square(1) distribution, rather than a simple chi-square(1) distribution as in an ordinary (non-boundary) likelihood ratio test.

