Chapter 2

Statistical methods and models of genetic risk

2.1 Introduction

The field of complex disease genetics is inherently statistical, both in the sense that it studies a phenomenon (complex disease) that is by definition probabilistic, and in the sense that it relies on statistical methods to make inferences from the data under study. Examples of these statistical methods include risk prediction (either using relative risks or odds ratios), regression analyses (usually using logistic regression) and family analyses (generally using liability threshold models). Each of these methods is built around assumptions, and these assumptions themselves form a model (either explicit or implicit) about the distributions of genetic risk in the population. In

many cases, these methods imply very different and mutually incompatible assumptions.

In the last few years the interest in statistical models of genetic risk has increased dramatically. Recent papers include general discussions of modelling issues arising from GWAS (e.g. Sawcer and Wason (2012)), and detailed examinations of specific models (e.g. Wray et al. (2010)). Two recent reviews (Wray and Goddard, 2010; Clayton, 2012) have made broad comparisons of different models of genetic risk, noting a number of inconsistencies between models and describing different implications for association studies and risk prediction. However, neither provided a systematic survey of the properties of genetic risk models, and in particular neither gave a detailed investigation into the relationships between different models, and between models and statistical methods. The time is thus ripe for a unified analysis that places different statistical methods and models of risk into a single framework.

In this chapter I will lay out a simple framework for classifying such models, and discuss three major models of genetic risk. Together, these three models underlie most standard models and methods used in the field. I will investigate how these models differ, how suitable each is to the tasks that they have been used for, and how their predictions about the distributions of genetic risk differ from each other.

In the introduction, I will formulate a general description of a model of genetic risk, and discuss a specific family of models that are specified in terms of a normally distributed genetic risk score and a link function. In Section 2.2, I will go on to discuss in more detail the relationship between locus-based models of genetic risk (such as those fitted in GWAS) and continuous risk scores. Sections 2.3-2.5 will discuss and critically assess three specific models of risk that correspond to three link functions (the log, probit and

logit models), and in Section 2.6 I will compare how these models differ in their predictions about the distribution of genetic risk. In the final section I will discuss how confusion between these models can generate real problems in statistical genetics, as well as discussion some of the limitations of this approach.

2.1.1 Definition of a genetic risk model

In general, a model of genetic risk has two properties. Firstly, it specifies a distribution of a genetic risk value $p_i \in [0, 1]$ for a randomly selected individual i

$$p_i \sim Distribution(\theta), \tag{2.1}$$

where the probability of an individual developing the disease is equal to p_i, or

$$P(d_i = 1|p_i) = p_i. \tag{2.2}$$

Here, d_i is an indicator variable taking on value $d_i = 1$ if the individual i has the disease (if we are modelling the prevalence) or will develop the disease in their lifetime (if we are modelling the lifetime risk).

Secondly, a model of genetic risk specifies a joint distribution for genetic risk values p_i and p_j for individuals i and j that share a family relationship r_{ij}

$$(p_i, p_j) \sim Distribution(\theta, r_{ij}). \tag{2.3}$$

For a purely genetic model, we make the additional assumption that

disease incidence is independent in families conditional on their genetic risk, i.e.

$$\begin{aligned} P(d_i = 1, d_j = 1|p_i, p_j) &= P(d_i = 1|p_i)P(d_j = 1|p_j) \\ &= p_i p_j. \end{aligned} \tag{2.4}$$

In essence, we assume that relatives have no shared environmental risk. In this chapter we will almost exclusively consider purely genetic models. In the case where environmental and genetic risks act independently, these models can be reasonably interpreted as the behaviour of the genetic component, and are easily extended to include environmental risk (as discussed in Section 2.4.1). In the presence of strong gene-environment interaction, however, these purely genetic models will become inaccurate, and the true model will depend on the form of the interaction.

We refer to p_i as the genetic risk or the genetic disease probability. Its distribution can be discrete or continuous, though we will only consider continuous distributions in this chapter.

2.1.2 Observable parameters of a genetic risk model

While each model of genetic risk has its own set of parameters θ, there are a number of common parameters that we can calculate for any model, which in turn are measurable in real populations.

The first parameter I will consider is the population prevalence of the disease, or the probability that a randomly selected individual has the disease in question. This is equal to

$$\begin{aligned} K &= P(d=1) \\ &= \int_p P(d=1|p)f(p)dp \\ &= \int_p pf(p)dp \\ &= E[p], \end{aligned} \tag{2.5}$$

where $f(\cdot)$ is the probability density function of p.

A more complicated measure is how "genetic" a disease is. This concept is relatively ill defined. The heritability of liability h^2 is often used for this purpose, which is equal to the proportion of variance in the total risk that can be attributed to genetics, where risk is measured on the liability scale (discussed in Section 2.4.1). However, this parameter is model specific.

Instead, for comparison across models we will use the relative recurrence risk, equal to the fold enrichment of disease prevalence in relatives of affected individuals. For relatives of type r_{ij}, this is calculated as

$$\begin{aligned} \lambda_r &= \frac{P(d_i=1|d_j=1,r_{ij})}{P(d_i=1)} \\ &= \frac{P(d_i=1,d_j=1|r_{ij})}{P(d_i=1)^2} \\ &= \frac{\int_{p_i}\int_{p_j} P(d_i=1|p_i)P(d_j=1|p_j)f(p_i,p_j|r_{ij})dp_i dp_j}{K^2}. \end{aligned} \tag{2.6}$$

This can in theory be measured directly from population data, if common environment can be controlled for. Regardless of whether or not it can actually be measured, the definition is model independent, and acts as a useful benchmark to compare across models.

Finally, we will be interested in the distribution of p in cases and controls

$$\begin{aligned} f(p|d=1) &= \frac{P(d=1|p)f(p)}{P(d=1)} \\ &= \frac{p}{K}f(p) \qquad (2.7) \\ f(p|d=0) &= \frac{1-p}{1-K}f(p). \qquad (2.8) \end{aligned}$$

2.1.3 Genetic risk scores and link functions

In this chapter, we will consider a specific family of continuous genetic risk models. These models have two components, firstly a normally distributed genetic risk score

$$\eta \sim N(\mu, \sigma^2) \qquad (2.9)$$

and secondly a link function g that connects this genetic risk score to the genetic risk probability

$$p = g(\eta). \qquad (2.10)$$

We can thus write down the probability density function of p as

$$f(p) = \frac{d\eta}{dg}\frac{1}{\sigma}\phi\left(\frac{\eta-\mu}{\sigma}\right), \qquad (2.11)$$

where ϕ is the density of the standard normal distribution.

In the following section, we will describe the relationship between discrete genotypes $\vec{x}$ and risk scores η. We will then consider three link functions: the log link $g(\eta) = exp(\eta)$, the logit link $g(\eta) = (1 + exp(-\eta))^{-1}$ and the probit link $g(\eta) = \Phi^{-1}(\eta)$.

2.2 From discrete genotypes to continuous risk

The conversion from discrete genotypes to a continuous genetic trait was first outlined by Fisher (1918), who showed not only that a large number of discrete genetic factors can give rise to a continuous trait, but also that certain correlation structures in this continuous trait exist between family members as a consequence of Mendelian inheritance. In this section I will outline the relationship between discrete genetic risk factors and a continuous risk score, and outline the distribution and parameters of this score.

Note that in the following section I will use lowercase x and y to refer to random variables that represent genotype dosages (i.e. $x \in (0, 1, 2)$), uppercase X and Y to refer to general random variables, and lowercase z to refer to a standard normal random variable.

The above described η score is constructed from a combination of genotypes across n loci, $\vec{x} = (x_1, ..., x_n)$. The general form is

$$\eta = t(\vec{x}), \tag{2.12}$$

where t is the function that maps from genotype to score. Note that, in this general formulation, there is no requirement that η be normally distributed (as described in Equation 2.9).

We can simplify this by assuming that the loci are all independent, and each contributes independently to η, i.e.

$$\eta = a_0 + \sum_{l=1}^{n} t_l(x_l). \tag{2.13}$$

As the random variables x_l are independent, and providing that the transformed variables $t_l(x_l)$ have finite means and variances that are independent of the indicator variable l, it follows from the central limit theorum that η

tends to a normal distribution as n increases.

We can modify the score to include interaction terms between genotypes, e.g. by including second-degree interaction

$$\eta = a_0 + \sum_{i=1}^{n} \sum_{j=i}^{n} t_{ij}(x_i, x_j). \tag{2.14}$$

In this section we discuss the particulars of going from a combination of genotypes to a continuous risk score. We will discuss the problem in general in terms of the properties of sums of independent variables, and then discuss the specific case where η is a linear function, i.e. $f_l(x_l) = a_l x_l$. Finally, we will discuss issues with non-linear functions.

2.2.1 Properties of a sum of independent variables

Suppose we have two sets of random variables, X_1 and X_2, and Y_1 and Y_2, such that $X_i \perp Y_j \; \forall(i,j)$.

We construct scores by adding these variables together, i.e. $\eta_i = X_i + Y_i$. The expectation and variance of this score are given by

$$\begin{aligned} E[\eta_i] &= E[X_i + Y_i] \\ &= E[X_i] + E[Y_i] \qquad (2.15) \\ Var[\eta_i] &= Var[X_i + Y_i] \\ &= Var[X_i] + Var[Y_i], \qquad (2.16) \end{aligned}$$

and the covariance are given by

$$\begin{aligned} cov(\eta_1, \eta_2) &= E[\eta_1\eta_2] - E[\eta_1]E[\eta_2] \\ &= E[(X_1 + Y_1)(X_2 + Y_2)] - E[X_1 + Y_1]E[X_2 + Y_2] \\ &= E[X_1X_2] + E[Y_1]E[X_2] + E[X_1]E[Y_2] + E[Y_1Y_2] - \\ &\quad E[X_1]E[X_2] - E[X_1]E[Y_2] - E[X_2]E[Y_1] - E[Y_1]E[Y_2] \\ &= E[X_1X_2] - E[X_1]E[X_2] + E[Y_1Y_2] - E[Y_1]E[Y_2] \\ &= cov[X_1, X_2] + cov[Y_1, Y_2]. \end{aligned} \tag{2.17}$$

We can generalise this to the sum of n variables $\eta_i = \sum_{j=1}^{n} X_{ij}$ such that $X_{ab} \perp X_{cd} \forall a, c, b \neq d$, to give

$$E[\eta_i] = \sum_{j=1}^{n} E[X_{ij}] \tag{2.18}$$

$$Var[\eta_i] = \sum_{j=1}^{n} Var[X_{ij}] \tag{2.19}$$

$$cov(\eta_1, \eta_2) = \sum_{j=1}^{n} cov(X_{1j}, X_{2j}). \tag{2.20}$$

If the X_{ij}'s have finite mean and variance, then when n is large we can approximate (η_1, η_2) as a multivariate normal with

$$\vec{\mu} = (E[\eta_1], E[\eta_1]) \tag{2.21}$$

$$\Sigma = \begin{pmatrix} Var[\eta_1] & cov(\eta_1, \eta_2) \\ cov(\eta_1, \eta_2) & Var[\eta_1]. \end{pmatrix} \tag{2.22}$$

If we imagine that the X_{ij}'s are functions of allele count for independently

segregating genetic risk loci, we can see that to calculate the covariance of a function that is a sum of such functions only requires the calculation of the covariance of each function individually.

2.2.2 General covariance for linear functions of allele count

A linear, or additive, risk score has the form

$$\eta_i = a_0 + \sum_{l=1}^{n} a_l x_{il}. \tag{2.23}$$

Again we will assume that the variants in the score are in linkage equilibrium, and thus the allele counts at different loci are independent ($x_{ia} \perp x_{ib} \forall a \neq b$).

The score η_i has expectation and variance

$$\begin{aligned} E[\eta_i] &= a_0 + \sum_{l=1}^{n} a_l E[x_{il}] \\ &= a_0 + \sum_{l=1}^{n} a_l 2 f_l \qquad (2.24) \\ Var[\eta_i] &= \sum_{l=1}^{n} a_l^2 Var[x_{il}] \\ &= \sum_{l=1}^{n} a_l^2 2 f_l (1 - f_l), \qquad (2.25) \end{aligned}$$

where f_i is the allele frequency of variant l.

To calculate the covariance, suppose two individuals i and j have a coefficient of relatedness ρ_{ij}. This is equal to the probability that any given allele on any given chromosome will be shared IBD (with $\rho_{ij} = 0.5$ for siblings i and j, etc).

For a variant with allele frequency f, we can denote the allele count for individual i as $x_i = x_{i1} + x_{i2}$, where x_{ik} are allele counts on individual chromosomes $k = 1, 2$ for individual i. We will use $S_k^{ij} = 1$ to denote that this allele is shared IBD between individuals i and j on chromosome k, with $P(S_k^{ij} = 1) = \rho_{ij}$. For now I will assume $S_1^{ij} \perp S_2^{ij}$, i.e. that the IBD sharing states for the two chromosomes are independent (as is the case for siblings, for example). The next section will generalize this to arbitrary IBD distributions.

The joint distribution of genotypes on a particular chromosome k for two individuals i and j with coefficient of relatedness ρ_{ij} is given by

$$P(x_{ik}, x_{jk}) = \rho_{ij} P(x_{ik}, x_{jk} | S_k^{ij} = 1) + (1 - \rho_{ij}) P(x_{ik}, x_{jk} | S_k^{ij} = 0), \quad (2.26)$$

where

$$P(x_{ik}, x_{jk} | S_k^{ij} = 1) = \begin{cases} P(x_{ik}) & \text{if } x_{ik} = x_{jk}; \\ 0 & \text{otherwise}, \end{cases} \quad (2.27)$$

(2.28)

and

$$P(x_{ik}, x_{jk} | S_k^{ij} = 0) = P(x_{ik}) P(x_{jk}). \quad (2.29)$$

We can calculate the covariance in allele counts between two individuals of x_{ik} and x_{jk} by first calculating

$$\begin{aligned} E[x_{ik}x_{jk}] &= \sum x_{ik}x_{jk}P(x_{ik},x_{jk}) \\ &= P(x_{ik}=1,x_{jk}=1) \\ &= \rho_{ij}P(x_{ik}=1,x_{jk}=1|S_k^{ij}=1)+(1-\rho_{ij})P(x_{ik}=1,x_{jk}=1|S_k^{ij}=0) \\ &= \rho_{ij}f+(1-\rho_{ij})f^2, \end{aligned} \tag{2.30}$$

and then by using this to calculate the covariance

$$\begin{aligned} cov[x_{ik},x_{jk}] &= E[x_{ik}x_{jk}]-E[x_{ik}]E[x_{jk}] \\ &= \rho_{ij}f+(1-\rho_{ij})f^2-f^2 \\ &= \rho_{ij}f(1-f). \end{aligned}$$

We can use Equation 2.17 to give $cov[x_i,x_j]=2\rho_{ij}f(1-f)$. Note that $var[x_i]=2f(1-f)$, so $cor[x_i,x_j]=\rho_{ij}$. This means that, as well as being the probability of sharing any given allele IBD, the coefficient of relatedness is also equal to the correlation in genotype counts.

We can therefore give the covariance of η_i and η_j as

$$\begin{aligned} cov[\eta_i,\eta_j] &= cov[\sum_{l=1}^{n}a_lx_{il},\sum_{l=1}^{n}a_lx_{jl}] \\ &= \sum_{l=1}^{n}cov[a_lx_{1l},a_lx_{2l}] \\ &= \rho_{ij}\sum_{l=1}^{n}a_l^2 2f_l(1-f_l) \\ &= \rho_{ij}var[\eta_j]. \end{aligned} \tag{2.31}$$

Again, note that $cor[\eta_i, \eta_j] = \rho_{ij}$.

When I refer to "additive" genetic risk throughout this thesis, I refer a to risk score which can be expressed thus on some scale. This assumption of additivity is important because it allows us to assume that the correlation between individuals on this scale is equal to their coefficient of relatedness.

2.2.3 Covariance for non-linear functions of allele count

The coefficient of relatedness is not sufficient to give the full joint genotype distribution for two individuals. For instance, while full siblings and parent-offspring pairs both have the same coefficient of relatedness ($\rho = 0.5$), they have distinct patterns of allele sharing due to the fact that parent-offspring always share exactly one allele IBD, but siblings can share zero, one or two.

We write the proportion of alleles shared IBD 1 and 2 as p_1, p_2 (with $1 - p_1 - p_2$ with IBD 0). We can calculate the coefficient of relatedness from the IBD probabilities as $\rho = \frac{1}{2}p_1 + p_2$. Parent-offspring pairs have $p_1 = 1$ and $p_2 = 0$, siblings have $p_1 = 0.5$ and $p_2 = 0.25$.

The table below shows the joint genotype distributions depending on IBD status.

Genotype (x_i,x_j)	IBD = 0	IBD = 1	IBD = 2
0,0	$(1-f)^4$	$(1-f)^3$	$(1-f)^2$
0,1	$2f(1-f)^3$	$f(1-f)^2$	0
0,2	$f^2(1-f)^2$	0	0
1,1	$4f^2(1-f)^2$	$f(1-f)$	$2f(1-f)$
1,2	$2f^3(1-f)$	$f^2(1-f)$	0
2,2	f^4	f^3	f^2

Note that certain genotype combinations can occor multiple ways. For in-

stance, there are two possible ways of having one individual with two alleles and one with zero, $(x_i,x_j) = (0,2)$ and $(x_i,x_j) = (2,0)$. This means that the probability of being in either of these two states is equal to $2f^2(1-f)^2$.

We can then calculate the expected values of various non-linear functions of genotype count. For instance, the product of genotype values has expectation

$$\begin{aligned} E[x_i x_j|\text{IBD} = 0] &= 4f^2(1-f)^2 + 8f^3(1-f) + 4f^4 \\ &= 4f^2 \qquad (2.32) \\ E[x_i x_j|\text{IBD} = 1] &= f(1-f) + 4f^2(1-f) + 4f^3 \\ &= f(1+3f) \qquad (2.33) \\ E[x_i x_j|\text{IBD} = 2] &= 2f(1-f) + 4f^2 \\ &= 2f(1+f) \qquad (2.34) \\ E[x_i x_j] &= (1-p_1-p_2)E[x_i x_j|\text{IBD} = 0] \\ &\quad + p_1 E[x_i x_j|\text{IBD} = 1] + p_2 E[x_i x_j|\text{IBD} = 2] \\ &= (1-p_1-p_2)4f^2 + p_1(f(1+3f)) + p_2 2f(1+f) \\ &= f(p_1+2p_2) + f^2(4-p_1-2p_2) \\ &= 2f\rho + 2f^2(2-\rho). \qquad (2.35) \end{aligned}$$

As we saw above, the expectation of the product is dependent only on ρ, and not on the specific IBD distribution.

The expectation of $x_i x_j^2$ is given by

$$\begin{aligned} E[x_i x_j^2|\text{IBD}=0] &= 4f^2(1-f)^2 + 12f^3(1-f) + 8f^4 \\ &= 2f^2(2(1-f)^2 + 6f(1-f) + 4f^4) \\ &= 4f^2(1+f) \qquad (2.36) \\ E[x_i x_j^2|\text{IBD}=1] &= f(1-f) + 6f^2(1-f) + 8f^3 \\ &= f(1+5f+2f^2) \qquad (2.37) \\ E[x_i x_j^2|\text{IBD}=2] &= 2f(1-f) + 8f^2 \\ &= 2f(1+3f) \qquad (2.38) \end{aligned}$$

$$\begin{aligned} E[x_i x_j^2] &= (1-p_1-p_2)E[x_i x_j^2|\text{IBD}=0] + p_1 E[x_i x_j^2|\text{IBD}=1] \\ &\quad + p_2 E[x_i x_j^2|\text{IBD}=2] \\ &= (1-p_1-p_2)4f^2(1+f) + p_1 f(1+5f+2f^2) + p_2 2f(1+3f) \\ &= (p_1+2p_2)f + (4+p_1+2p_2)f^2 + 2(4-p_1-2p_2)f^3 \\ &= 2\rho f + 2(2+\rho)f^2 + 4(2-\rho)f^3. \qquad (2.39) \end{aligned}$$

Again, this expression is only dependent on ρ. Finally, the expectation of $x_i^2 x_j^2$ is given by

$$
\begin{aligned}
E[x_i^2 x_j^2|\text{IBD}=0] &= 4f^2(1-f)^2 + 16f^3(1-f) + 16f^4 \\
&= 4f^2(1+f)^2 \qquad (2.40) \\
E[x_i^2 x_j^2|\text{IBD}=1] &= f(1-f) + 8f^2(1-f) + 16f^3 \\
&= f(1+7f+8f^2) \qquad (2.41) \\
E[x_i^2 x_j^2|\text{IBD}=2] &= 2f(1-f) + 16f^2 \\
&= 2f(1+7f) \qquad (2.42)
\end{aligned}
$$

$$
\begin{aligned}
E[x_i^2 x_j^2] &= (1-p_1-p_2)E[x_i^2 x_j^2|\text{IBD}=0] \\
&\quad +p_1 E[x_i^2 x_j^2|\text{IBD}=1] + p_2 E[x_i^2 x_j^2|\text{IBD}=2] \\
&= (1-p_1-p_2)4f^2(1+f)^2 + p_1 f(1+7f+8f^2) + p_2 2f(1+7f) \\
&= f(p_1+2p_2) + f^2(4+3p_1+10p_2) \\
&\quad +8f^3(1-p_2) + 4f^4(1-p_1-p_2). \qquad (2.43)
\end{aligned}
$$

And these in turn allow to us to calculate covariance and correlations of non-linear functions of allele count between relatives. For instance, consider the non-linear function $\eta_i = x_i + bx_i^2$ with dominance term b.

$$
\begin{aligned}
E[\eta_i] &= E[x_i] + bE[x_i^2] \\
&= 2f(1+b) + 2bf^2 && (2.44) \\
E[\eta_i]^2 &= E[x_i] + bE[x_i^2] \\
&= 4f^2(1+b)^2 + 8f^3b(1+b) + 4b^2f^4 && (2.45) \\
E[\eta_i^2] &= E[(x_i + bx_i^2)^2] \\
&= E[x_i^2] + 2bE[x_i^3] + b^2E[x_i^4] \\
&= 2f(1+b)^2 + 2f^2(1+6b+7b^2) && (2.46) \\
Var[\eta_i] &= E[\eta_i^2] - E[\eta_i]^2 \\
&= 2f(1+b)^2 - 2f^2(1-2b-5b^2) \\
&\quad -8f^3b(1+b) - 4f^4b^2 && (2.47) \\
E[\eta_i\eta_j] &= E[(x_i + bx_i^2)(x_j + bx_j^2)] \\
&= E[x_ix_j] + 2bE[x_ix_j^2] + b^2E[x_i^2x_j^2] && (2.48) \\
cov[\eta_i, \eta_j] &= E[\eta_i\eta_j] - E[\eta_i]E[\eta_j] && (2.49) \\
cor[\eta_i, \eta_j] &= \frac{cov[\eta_i, \eta_j]}{var[\eta_i, \eta_j]}. && (2.50)
\end{aligned}
$$

We can find the maximum and minimum values of the correlation by differentiating $cor[\eta_i, \eta_j]$ with respect to b. We find that the correlation takes on the minimum value of p_2 when $b = \frac{-1}{1+2f}$, and a maximum value of ρ when $b = 0$.

As Figure 2.1 shows, low frequency variants show very little drop off in correlation until very high degrees of dominance, whereas higher frequency variants show a smoother drop off in correlation. Dominance effects thus have a stronger impact on the risk score correlation when the variants have higher frequency.

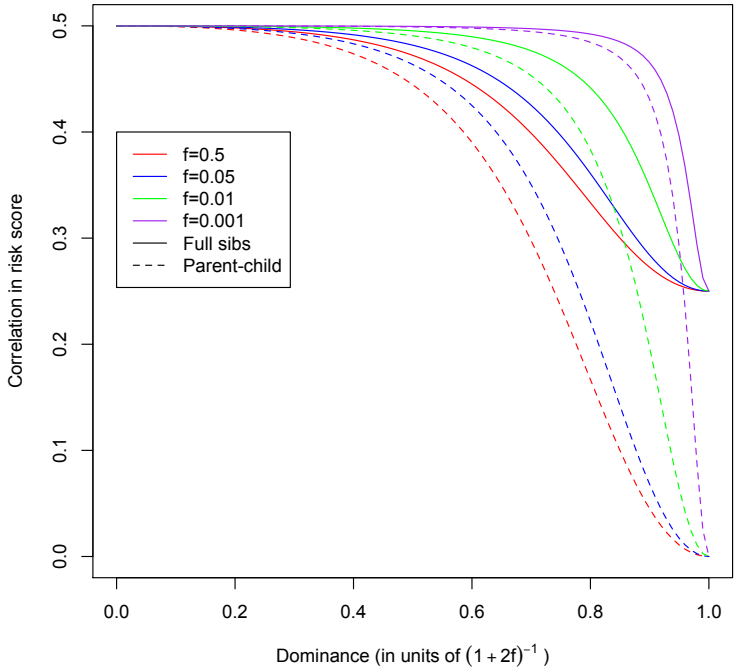


Figure 2.1: The decrease in correlation in risk score for siblings and parent-child pairs with increasing value of the dominance term (normalised such that the maximum value is $\frac{-1}{1+2f}$). Different colour lines represent variants with different allele frequencies.

2.3 The log risk model

The log risk model was defined by Pharoah et al. (2002) and more recently elaborated on by Clayton (2009). It has most commonly been used to make inferences about the utility of genetic risk prediction (Clayton, 2009; Sawcer et al., 2010; Chatterjee et al., 2011), though it has also been used to estimate sibling recurrence ratios in twin studies (Clayton, 2009).

As we will see, the model is asymptotically equivalent to the Risch multilocus model of genetic risk. It is also equivalent to the assumption of multiplicative combination of relative risk that is often used in genetic risk pre-

diction (e.g. by the genetic testing company deCODEme).

This model is the least realistic of the models that I will consider, due to the fact that the probability is not bounded, though it is also one of the more widely used, probably due to its analytic tractability.

The link function for the log risk model is

$$p = \exp(\eta). \tag{2.51}$$

Substituting this into Equation 2.11, the density function for p is given by

$$f(p) = \frac{1}{p\sigma}\phi\left(\frac{log(p) - \mu}{\sigma}\right). \tag{2.52}$$

2.3.1 Calculating parameters

The prevalence parameter K is given by

$$\begin{aligned} K &= E[p] \\ &= \int \exp(\mu + \sigma x)\phi(x)dx \\ &= \int \frac{1}{\sqrt{2\pi}} \exp(\mu + \sigma x - \frac{1}{2}x^2)dx \\ &= \int \frac{1}{\sqrt{2\pi}} \exp(\mu + \sigma^2/2 - \frac{1}{2}(x-\sigma)^2)dx \\ &= \exp(\mu + \frac{\sigma^2}{2}) \int \phi(x - \sigma)dx \\ &= \exp(\mu + \frac{\sigma^2}{2}), \end{aligned} \tag{2.53}$$

i.e. the expectation of the log-normal distribution (Johnson et al., 1994).

As we saw in Section 2.2.2, under additivity the correlation in log risk score is equal to their coefficient of relatedness ρ. We can thus express the genetic risk for relatives p_1 and p_2 as

$$p_1 = \exp(\mu + \sigma z_1) \quad (2.54)$$

$$p_2 = \exp(\mu + \rho\sigma z_1 + \sqrt{1-\rho}\sigma z_2), \quad (2.55)$$

where z_i are standard normal variables.

The probability of both relatives developing the disease given z_is is

$$\begin{aligned} P(d_1 = 1, d_2 = 1|\eta_1, \eta_2) &= p(d_1|\eta_1)p(d_2|\eta_2) \\ &= p_1 p_2 \\ &= \exp(\mu + \sigma z_1 + \mu + \rho\sigma z_1 + \sqrt{1-\rho^2}\sigma z_2) \\ &= \exp(2\mu + \sigma(1+\rho)z_1 + \sigma\sqrt{1-\rho^2} z_2). \end{aligned} \quad (2.56)$$

The mean rate of co-occurrence is thus

$$
\begin{aligned}
E[p_1 p_2] &= \int p_1 \phi(z_1) p_2 \phi(z_2) dz_1 dz_2 \\
&= \exp(2\mu) \int \exp(\frac{1}{2}(z_1^2 - 2\sigma(1+\rho))) p_2 \phi(z_2) dz_1 dz_2 \\
&= \exp(2\mu) \int \exp(\frac{1}{2}((z_1 - \sigma(1+\rho))^2 - \sigma^2(1+\rho)^2)) p_2 \phi(z_2) dz_1 dz_2 \\
&= \exp(2\mu + \frac{1}{2}\sigma(1+\rho))^2) \int \phi(z_1 - \sigma(1+\rho)) p_2 \phi(z_2) dz_1 dz_2 \\
&= \exp(2\mu + \frac{1}{2}\sigma(1+\rho)^2) \\
&\quad \times \int \phi(z_1 - \sigma(1+\rho)) \exp(\frac{1}{2}(z_2^2 - 2\sigma\sqrt{1-\rho^2})) dz_1 dz_2 \\
&= \exp(2\mu + \frac{1}{2}\sigma(1+\rho)^2) \\
&\quad \times \int \phi(z_1 - \sigma(1+\rho)) \exp(\frac{1}{2}((z_2 - \sigma\sqrt{1-\rho^2})^2 - \sigma^2(1-\rho^2))) dz_1 dz_2 \\
&= \exp(2\mu + \frac{1}{2}\sigma(1+\rho)^2 + \frac{1}{2}\sigma^2(1-\rho^2))) \\
&\quad \times \int \phi(z_1 - \sigma(1+\rho)) \phi(z_2 - \sigma\sqrt{1-\rho^2}) dz_1 dz_2 \\
&= \exp(2\mu + \frac{1}{2}\sigma(1+\rho)^2 + \frac{1}{2}\sigma^2(1-\rho^2))) \\
&= \exp(2\mu + \sigma(1+\rho)). \qquad (2.57)
\end{aligned}
$$

The recurrence ratio in relatives is thus

$$
\begin{aligned}
\lambda_r &= \frac{E[p_1 p_2]}{K^2} \\
&= \exp(\rho\sigma). \qquad (2.58)
\end{aligned}
$$

We can rearrange equations 2.53 and 2.58 to give parameters μ and σ, given a prevalence K and a sibling recurrence ratio λ_s

$$\sigma^2 = 2log(\lambda_s) \tag{2.59}$$

$$\mu = log(K) - \sigma^2. \tag{2.60}$$

2.3.2 Case and control distributions

The distribution of η in cases is given by the probability density function

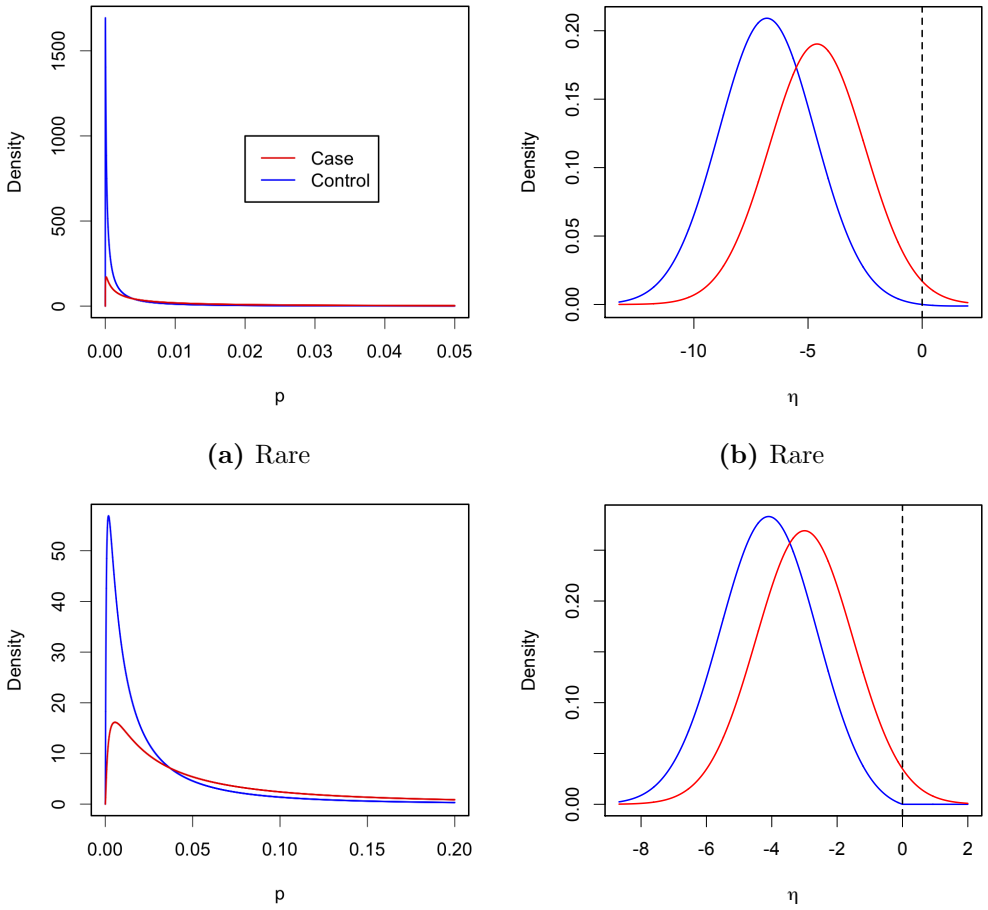
$$\begin{aligned} P(\eta|d=1) &= \frac{P(d=1|\eta)P(\eta)}{P(d=1)} \\ &= \frac{e^{\eta}\phi(\frac{\eta-\mu}{\sigma})}{\sigma K}. \end{aligned} \tag{2.61}$$

This can be simplified to

$$\begin{aligned} P(\eta|d=1) &= \frac{\exp(\eta)\frac{1}{\sigma\sqrt{2\pi}}\exp(\frac{-(\eta-\mu)^2}{2\sigma^2})}{\exp(\mu+\sigma^2)} \\ &= \frac{1}{\sigma\sqrt{2\pi}}\exp(\eta + \frac{-(\eta-\mu)^2}{2\sigma^2} - \mu - \sigma^2) \\ &= \frac{1}{\sigma\sqrt{2\pi}}\exp(\frac{-(\eta-(\mu+\sigma^2))^2}{2\sigma^2}) \\ &= \frac{1}{\sigma}\phi(\frac{\eta-(\mu+\sigma^2)}{\sigma}), \end{aligned} \tag{2.62}$$

i.e. normally distributed with a mean $\mu + \sigma^2$ and a variance σ^2. Thus, the distribution of log risk for cases is the same as for the population as a whole, but shifted upwards by σ^2.

The distribution for risk in controls is given by



(a) Rare **(b)** Rare

(c) Common **(d)** Common

Figure 2.2: The case and control distributions of probability p and risk score η for a rare disease ($K = 0.01, \lambda_s = 9$) and a common disease ($K = 0.05$, $\lambda_s = 3$)

$$P(\eta|d = 0) = \frac{(1 - e^{\eta})\phi(\eta)}{\sigma(1 - K)}. \tag{2.63}$$

The distribution of probability and risk score in cases and controls is shown for example parameters (simulating a common and rare disease) in Figure 2.2. Note that, in both parameter sets, a not insignificant number of cases have a value of $\eta > 0$ and therefore $p > 1$ (see Section 2.3.5 for more on this issue).

2.3.3 Relationship to Risch model

The log risk model can be seen as an approximation to the Risch multilocus model, introduced by Risch (1990), that has been used to make inferences about genetic risk prediction (Wray et al., 2007). The Risch model assumes that n loci exist, each with the same relative risk r and a risk allele frequency f. An individual's disease probability is based on the number of risk alleles they carry x, and is given by

$$\begin{aligned} p &= p_0 r^x \\ &= \exp\left[log(p_0) + xlog(r)\right], \end{aligned} \tag{2.64}$$

where p_0 is the disease probability in individuals with zero risk alleles.

x is binomially distributed, with $x \sim Binom(2n, f)$. As we saw above, as n grows larger, x tends in distribution to $N(2nf, 2nf(1-f))$, and thus

$$p \to \exp(\eta) \text{ where } \eta \sim N\left(log(p_0) + 2nf, 2nf(1-f)log(r)^2\right), \tag{2.65}$$

i.e. the Risch model is asymptotically equivalent to the log risk model with $\mu = log(p_0) + 2nf$ and $\sigma^2 = 2nf(1-f)log(r)^2$.

2.3.4 Relationship to multiplicative relative risk model and log-linked regression

A commonly used risk prediction method is the multiplicative relative risk model (also known as the log-linear relative risk model). This is the most widely used of the relative risk models in epidemiology (Breslow and Storer,

1985), and has been used in genetics, as a model for genetic risk prediction (Lu and Elston, 2008). Notably, it is the model used by the genetic testing company deCODEme to produce individual disease probabilities given a customer's genotypes (deCODEme, 2012).

Under the multiplicative relative risk model, we have n loci, with each having a frequency f_i and a genotypic relative risk r_i. The probability for an individual who has allele counts x_i is given by

$$\begin{aligned} p &= f_0 \prod r_i^{x_i} \\ &= \exp\left[log(f_0) + \sum_{i=1}^{n} x_i log(r_i) \right]. \end{aligned} \tag{2.66}$$

Note that this can be seen as a generalisation of the Risch model, with identical $f = f_i$ and $r = r_i$ for all i, and $x = \sum_i x_i$.

As long as the values r_i are finite, the terms $x_i log(r_i)$ will have finite mean and variance, and thus the central limit theorum states that the summation above will tend towards a normal distribution as n grows, giving

$$p \to \exp(\eta) \text{ where } \eta \sim N(\mu, \sigma), \tag{2.67}$$

where

$$\mu = log(f_0) + \sum_{i=1}^{n} 2 f_i log(r_i) \tag{2.68}$$

$$\sigma^2 = \sum_{i=1}^{n} 2 f_i (1 - f_i) log(r_i)^2, \tag{2.69}$$

i.e. equivalent to the log risk model with μ and σ.

2.3.5 Problem with probabilities greater than 1

Wray and Goddard (2010) noted a problem with the log risk and Risch models, in that they can predict probabilities greater than one. The authors suggest a modified version of the model, where probabilities are capped at 1. In practice, capping at 1 may not be conservative enough: the genetic testing company deCODEme cap their genetic risk probabilities at 90% (deCODEme, 2012). In contrast, Clayton (2012) argued that this is not a major problem with the model, as for relatively uncommon diseases probabilities greater than 1 are relatively rare in the general population.

However, I will show that this is a real problem with the model in many circumstances. It is true that unless the disease is very common, the total number of individuals with $p > 1$ is small. For a disease with $K = 0.01$ and $\lambda_s = 9$, less than 0.1% of individuals have $p > 1$, and even for a disease with $K = 0.05$ and $\lambda_s = 3$ only 0.3% of individuals have this property (Figure 2.2). However, these values rise dramatically if we only consider cases, to 0.5% and 2.2% respectively, and if we consider identical twins where both are affected, 7% and 23% of twin pairs have a probability greater than 1.

So, while probabilities for randomly selected individuals are unlikely to suffer from this problem, the individuals in those groups we are often most concerned with (i.e. those with a family history and those who will go on to develop the disease) are far more likely to. In particular, the very high proportion of doubly-affected twin pairs with probabilities greater than 1 is concerning given that the expectation of the product of these probabilities is used to calculate the sibling recurrence ratio in Equation 2.58. Because this expectation is likely to be overestimated due to the greater-than-one

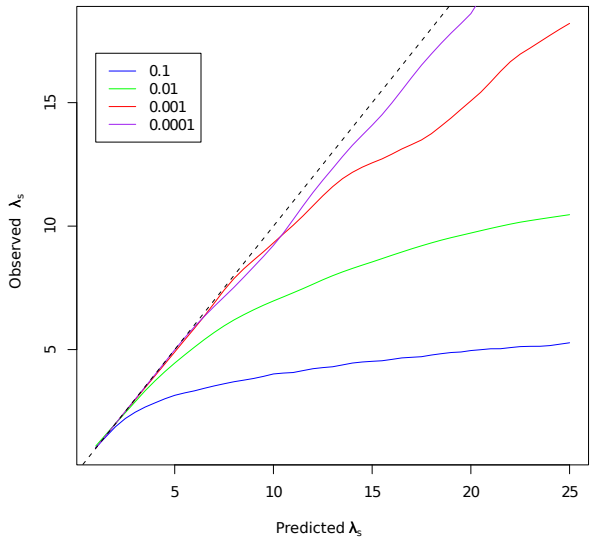


Figure 2.3: The λ_s predicted under the log risk model compared to the observed value under the truncated log model with all probabilities greater than 1 set to 1, for varying prevalence.

probabilities, it will follow that the value of λ_s could be greatly overestimated, and likewise the size of the genetic variance and parameter σ could be underestimated given a value λ_s.

To investigate the degree to which this will lead to errors, I simulated families under a truncated model (i.e. setting all $p > 1$ to $p = 1$), and compared the observed λ_s values to those predicted by Equation 2.58. Figure 2.3 shows that the log risk model significantly overestimates virtually all values of λ_s when $K = 0.1$, all values of $\lambda_s > 5$ for $K = 0.01$, and values of $\lambda_s > 10$ for $K = 0.001$. Only for very rare diseases ($K < 0.0001$) does the log risk model perform well regardless of the value of λ_s.

2.4 The probit risk model

The probit model of risk, also called the liability threshold model, was introduced by Falconer (1965), and further refined by Reich et al. (1972) and Falconer and Mackay (1996). Due to its compatibility with structural equation modelling and the popularity of the Mx program (Neale and Cardon, 1992), it has come to be used as the dominant model for twin studies of binary traits (Rijsdijk and Sham, 2002). Outside of family studies, it has also been used to study the potential limits of genetic risk prediction (Wray et al., 2010), and has even been important in influencing how many non-statisticians develop their theories of disease (see for instance Haegert (2004)).

The link function for the probit risk model is

$$p = \Phi(\eta), \tag{2.70}$$

where Φ is the cumulative distribution function of the standard normal distribution. Substituting this into Equation 2.11 gives a probability density of

$$f(p) = \frac{1}{\sigma\phi\left(\Phi^{-1}(p)\right)}\phi\left(\frac{\Phi^{-1}(p) - \mu}{\sigma}\right), \tag{2.71}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution (or quantile) function of the standard normal distribution.

2.4.1 Relationship to the liability threshold model

The probit risk distribution in Equation 2.70 is derived from the liability threshold model. The liability threshold model assumes that individuals have a liability score $L \sim N(0, 1)$, and an individual is assumed to have the

disease if L is larger than some threshold T. A simple form of the liability model assumes that L can be expressed in terms of an additive genetic component A and an environmental component E as

$$L = A + E, \tag{2.72}$$

where $A \sim N(0, h^2)$, $E \sim N(0, 1 - h^2)$ and $A \perp E$.

We can express $A = hz$ where $z \sim N(0, 1)$, and thus the distribution of genetic disease probabilities is

$$\begin{aligned} p &= P(A + E > T) \\ &= P(E > T - A) \\ &= \Phi\left(-\frac{T - hz}{\sqrt{1 - h^2}}\right) \\ &= \Phi(\eta). \end{aligned} \tag{2.73}$$

We thus see that the liability threshold model is equivalent to the probit model with

$$\mu = -\frac{T}{\sqrt{1 - h^2}} \tag{2.74}$$

$$\sigma = \sqrt{\frac{h^2}{1 - h^2}}, \tag{2.75}$$

and likewise

$$T = -\frac{\mu}{\sqrt{1+\sigma^2}} \tag{2.76}$$

$$h^2 = \frac{\sigma^2}{1+\sigma^2} \tag{2.77}$$

A note on the ACDE liability model

Liability threshold modelling is often extended to partition the liability in more detail. A general formulation is the "ACDE" model, where

$$L = A + C + D + E, \tag{2.78}$$

and where A is an additive genetic risk score, D is a dominant genetic risk score, C is an environmental risk shared between family members and E is non-shared environmental risk. All these terms have their own individual variances σ^2_X, and $\sum_X \sigma^2_X = 1$.

As we have already seen, the correlation in additive risk score A is ρ_{ij}, and as we saw in Section 2.2.3 the correlation in a fully dominant risk score is $p_2 = p(IBD = 2)$. The correlation in common environment is by definition 1. It is this formulation that is generally used in twin studies, where the model is fitted (ideally by maximum likelihood, though often by approximate methods) to a set of identical twins (i.e. $\rho_{ij} = 1$ and $p_2 = 1$) and non-identical twins (i.e. ρ_{ij} =0.5 and p_2 =0.25). In practice, having only two distinct levels of relatedness means that only two parameters can be fitted, so in general we either set $D = 0$ (the "ACE" model), or $C = 0$ (the "ADE" model, generally used for twins reared apart). Note that this formulation is not specific to the liability threshold model, similar covariance relationships can be defined for

any model that is expressed in terms of a normally distributed risk score.

2.4.2 Calculating parameters

By definition, the threshold T is selected such that a proportion K of individuals have a value greater than T, i.e. $T = \Phi^{-1}(1-K)$. We can thus write K in terms as μ and σ as

$$K = 1 - \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right). \tag{2.79}$$

The heritability, provided by Wray et al. (2010) using equations derived by Reich et al. (1972), is given by

$$h^2 = 2\frac{T - T_s\sqrt{1-(T^2-T_s^2)(1-T/z)}}{z + T_s^2(i-T)}, \tag{2.80}$$

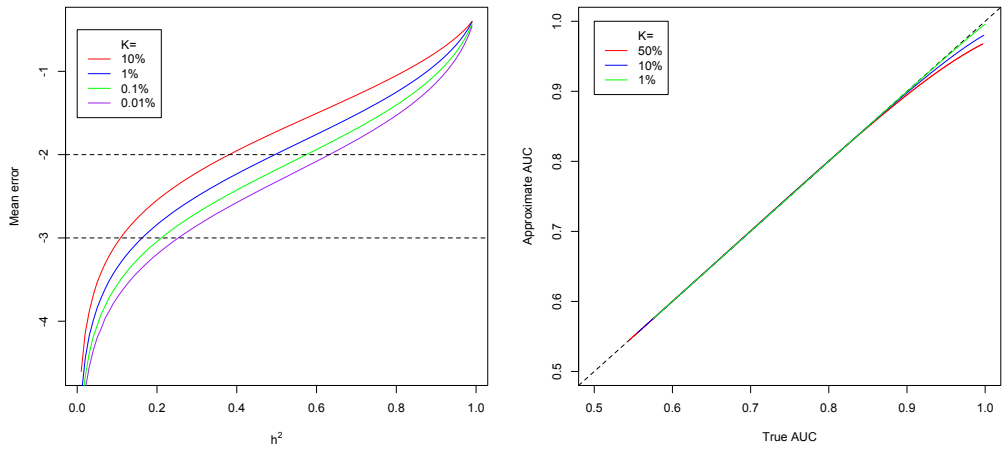
where $T_s = \Phi^{-1}(1-\lambda_s K)$, and $z = \frac{\phi(T)}{K}$.

2.4.3 Case and control distributions

Wray et al. (2010) calculated an approximate normal density for the genetic liability A in cases as

$$P(A|d=1) \approx \frac{1}{\sqrt{h^2(1-h^2z(z-T))}}\phi\left(\frac{zh^2 - A}{\sqrt{h^2(1-h^2z(z-T))}}\right). \tag{2.81}$$

This is an approximation to the exact density



(a) **(b)**

Figure 2.4: a) The $\log_{10}$ mean error (average squared distrance from the true value) of the normal density approximation to the genetic liability in cases $P(A|d = 1)$, as a function of prevalence K and heritability h^2. b) The Area Under the ROC Curve calculated using the exact and approximate equations, as a function of K and h^2

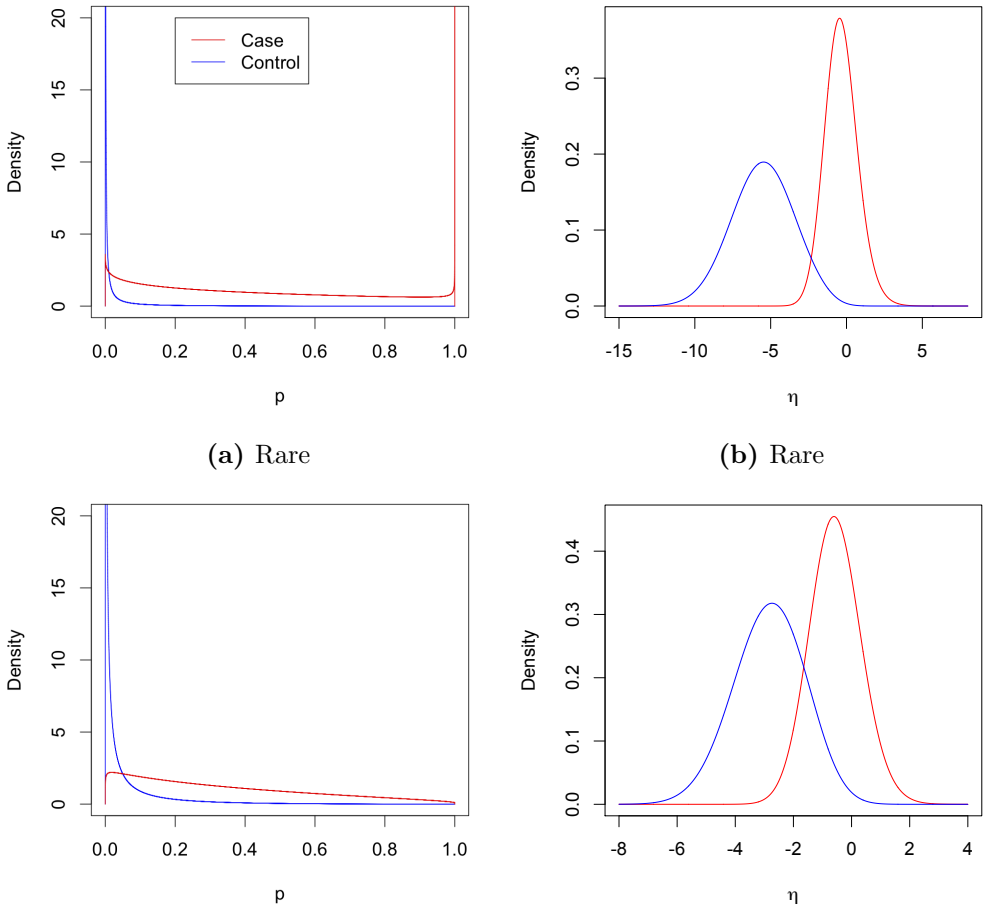
$$P(A|d=1) = \frac{P(A+E>T|A)P(A)}{P(A+E>T)} \quad (2.82)$$

$$= \frac{1}{K}\Phi\left(\frac{A-T}{\sqrt{1-h^2}}\right)\phi\left(\frac{A}{\sqrt{h^2}}\right). \quad (2.83)$$

Similar expressions exist for the genetic liability in controls.

Figure 2.4A shows the mean accuracy of this normal approximation as a function of the heritability and prevalence. Note that there is significant error in this approximation at high heritabilities, particularly if the prevalence is also high.

This approximation is used by Wray et al. (2010) to calculate the maximum possible predictive capacity of genetic risk prediction for various dis-



(a) Rare **(b)** Rare

(c) Common **(d)** Common

Figure 2.5: The case and control distributions of probability p and risk score η for a rare disease ($K = 0.01, \lambda_s = 9$) and a common disease ($K = 0.05$, $\lambda_s = 3$), under the probit model.

eases. The error in this function for highly heritable common diseases suggests that these values could be in error. However, Figure 2.4B shows that, in practice, this error only serves to slightly underestimate the very largest AUCs for very common $K > 0.1$ diseases, which does not substantially change the conclusions drawn from these results.

Examples of the distributions of η and p in cases and controls are shown in Figure 2.5.

2.4.4 Relationship to probit regression and latent variable modelling

Probit regression is a form of latent variable regression introduced by Bliss (1935) in 1935 as a model for bio-assay analysis. It was the dominant method of analysis for dichotomous traits until the 1960s, when the logistic regression model began to overtake it (see discussion of the logistic model below).

The probit model is a latent variable model, based on a continuous score

$$y = \beta_0 + \sum_i \beta_i x_i + e, \tag{2.84}$$

where $\vec{\beta}$ are parameters of the model, $\vec{x}$ are observed variables, and $e \sim N(0,1)$ is an unobserved (or latent) variable. The observed outcome is a binary indicator variable

$$d(y) = \begin{cases} 1 & \text{if } y > 0; \\ 0 & \text{otherwise.} \end{cases} \tag{2.85}$$

The probit regression model is fitted to determine the values of $\vec{\beta}$.

We write $X = \beta_0 + \sum_i \beta_i x_i$, which, given a large number of predictors, can be approximated as $X \sim N(\mu_x, \sigma_x^2)$, where

$$\mu_x = \beta_0 + \sum_i 2 f_i \beta_i \tag{2.86}$$

$$\sigma_x^2 = \sum_i 2 f_i (1 - f_i) \beta_i^2. \tag{2.87}$$

We can write the probability of $d = 1$ as

$$\begin{aligned} P(d=1) &= p(y>0) \\ &= P(X+e>0) \\ &= P\left(\frac{X-\mu_x}{\sqrt{1+\sigma_x^2}} + \frac{e}{\sqrt{1+\sigma_x^2}} > -\frac{\mu_x}{\sqrt{1+\sigma_x^2}}\right) \\ &= P(A+E>T), \end{aligned} \tag{2.88}$$

i.e. equivalent to the liability threshold model where

$$\begin{aligned} h^2 &= \frac{\sigma_x}{\sqrt{1+\sigma_x^2}} \\ &= \frac{\sum_i 2f_i(1-f_i)\beta_i^2}{\sqrt{1+\sum_i 2f_i(1-f_i)\beta_i^2}} \end{aligned} \tag{2.89}$$

$$\begin{aligned} T &= -\frac{\mu_x}{\sigma_x^2} \\ &= -\frac{\beta_0 + \sum_i 2f_i\beta_i}{\sqrt{1+\sum_i 2f_i(1-f_i)\beta_i^2}}. \end{aligned} \tag{2.90}$$

We can use this to fit the liability threshold or probit model directly from the results of probit regression, and thus calculate the variance explained by a set of genetic markers. While this is generally not used as a method for calculating heritability, if the liability threshold model is, in fact, the true model of genetic risk, this method should give the best approximation to the true variance explained by a set of genetic predictors.

2.5 The logit risk model

The general logit-normal (or logistic-normal) distribution was first defined by Mead (1965) in 1965, who noted that its moments have no analytic closed form, and its parameters can only be estimated iteratively (and even then only with some difficulty). However, the logit link itself is much older, having been used in bio-assay since the 1930s (see discussion of logistic regression below).

The logistic-normal distribution has been used previously to model serial observations under a random effects model (Stiratelli et al., 1984), but I believe has only been used directly in quantitative genetics once. Commenges et al. (1995) used a logistic-normal model to test hypotheses about familial aggregation in Alzheimer's disease conditional on known risk factors, much like the standard use of the probit model described above.

The implicit importance of the logit model is much larger than its lack of direct application may suggest. The most common methods used in modern statistical genetics, multiplicative odds ratio analysis and logistic regression, both implicitly assume the existence of logit-normally distributed risk. In essence, a model of genetic risk in the population is implicitly assumed by the methodology of almost all human complex disease genetics, but almost never directly investigated. This disconnect between the common usage of the regression technique and the infrequent use of the limiting normal has been noted in other fields (Frederic and Lad, 2003).

The link function for the logit risk model is

$$p = (1 + \exp(-\eta))^{-1} : \eta \sim N(\mu, \sigma), \tag{2.91}$$

and the density is

$$f(p) = \frac{1}{\sigma p(1-p)}\phi\left(\frac{1}{\sigma}log\left(\frac{p}{1-p}\right) - \frac{\mu}{\sigma}\right). \tag{2.92}$$

Note that η is equal to the log-odds of disease

$$\begin{aligned} log(O) &= log\left(\frac{p}{1-p}\right) \\ &= \eta. \end{aligned} \tag{2.93}$$

2.5.1 Calculating parameters

As with the moments of the logit-normal, none of the parameters of the logit normal have closed-form analytic solutions. Instead, they must be calculated by numeric integration.

The prevalence is given by

$$\begin{aligned} K &= E[p] \\ &= \int_0^1 pf(p)dp. \end{aligned} \tag{2.94}$$

To calculate the relative recurrence ratio, we need to look at the bivariate distribution. Suppose we have two individuals with a relatedness coefficient ρ. We model their genotypic risks as

$$p_i = \frac{1}{1+\exp(-\eta_i)}, \tag{2.95}$$

where $\vec{\eta} = (\eta_1, \eta_2)$ are jointly normally distributed with a mean of μ and a covariance

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{2.96}$$

We can transform $\vec{\eta}$ into independent standard normals $\vec{x}$ by noting that

$$\vec{\eta} = \mu + B\vec{x}, \tag{2.97}$$

where B is the Cholesky decomposition of Σ, such that $BB' = \Sigma$ and thus

$$B = \sigma \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix}. \tag{2.98}$$

From this, we can transform p_i, giving

$$x_1 = \frac{1}{\sigma}\left[log\left(\frac{p_1}{1-p_1}\right) - \mu\right] \tag{2.99}$$

$$x_2 = \frac{1}{\sigma\sqrt{1-\rho^2}}\left[log\left(\frac{p_2}{1-p_2}\right) - \mu - \sigma\rho x_1\right]. \tag{2.100}$$

The determinant of the Jacobian of this transformation is

$$\left|\frac{d\vec{x}}{d\vec{y}}\right| = \frac{1}{\sigma^2\sqrt{1-\rho^2}\prod_{i=1}^{2} p_i(1-pi)}, \tag{2.101}$$

thus the joint density of risk is given by

$$g(p_1, p_2) = \frac{\phi(\vec{x})}{\sigma^2\sqrt{1-\rho^2}p_1(1-p_1)P_2(1-p_2)} \tag{2.102}$$

From this we can calculate λ_R

$$\lambda_R = \frac{\int\int p_1 p_2 f(p_1, p_2) dp_1 dp_2}{K^2}. \tag{2.103}$$

2.5.2 Fitting the logit risk model numerically

To find parameters μ and σ given parameters K and λ_s, we find values that minimize the error function

$$Error(\mu,\sigma) = \left(\sqrt{E[p_1 p_2|\mu,\sigma]} - \sqrt{\lambda_s K^2}\right)^2 + \left(E[p|\mu,\sigma] - K\right)^2. \quad (2.104)$$

I use the Nelder-Mead algorithm (Nelder and Mead, 1965) implemented in the statistical language R. Note that the convergence speed and reliability of this procedure can be very dependent on the initial values of μ and σ. We can get a good initial guess by expressing the logit risk in terms of the probit model

We can express the probit model on the logit scale

$$\eta_{probit} = \Phi^{-1}((1+e^{-\eta_{logit}})^{-1}) \quad (2.105)$$

$$\frac{d\eta_{probit}}{d\eta_{logit}} = \left[\phi(\Phi^{-1}((1+e^{-\eta_{logit}})^{-1})(1+e^{-\eta_{logit}})(1+e^{\eta_{logit}})\right]^{-1} \quad (2.106)$$

We can then get the density of the logit risk score given the probit model

$$f(\eta_{logit}|\mu_{probit},\sigma_{probit}) = f(\eta_{probit}|\mu_{probit},\sigma_{probit})\frac{d\eta_{probit}}{d\eta_{logit}}, \quad (2.107)$$

which can in turn give us the expectation and variance of the logit risk variable under the probit model, which we use as an initial guess for the parameters μ and σ under the logit risk model

$$\mu_{init} = \int_{\eta} \eta f(\eta|\mu_{probit}, \sigma_{probit}) d\eta \tag{2.108}$$

$$\sigma_{init} = \int_{\eta} (\eta - \mu_{init})^2 f(\eta|\mu_{probit}, \sigma_{probit}) d\eta. \tag{2.109}$$

2.5.3 Case and Control Distributions

There is no particularly elegant way of describing the distribution of the probability p and the risk score η in cases and controls. Instead we can only use the general equations given in Section 2.1.2.

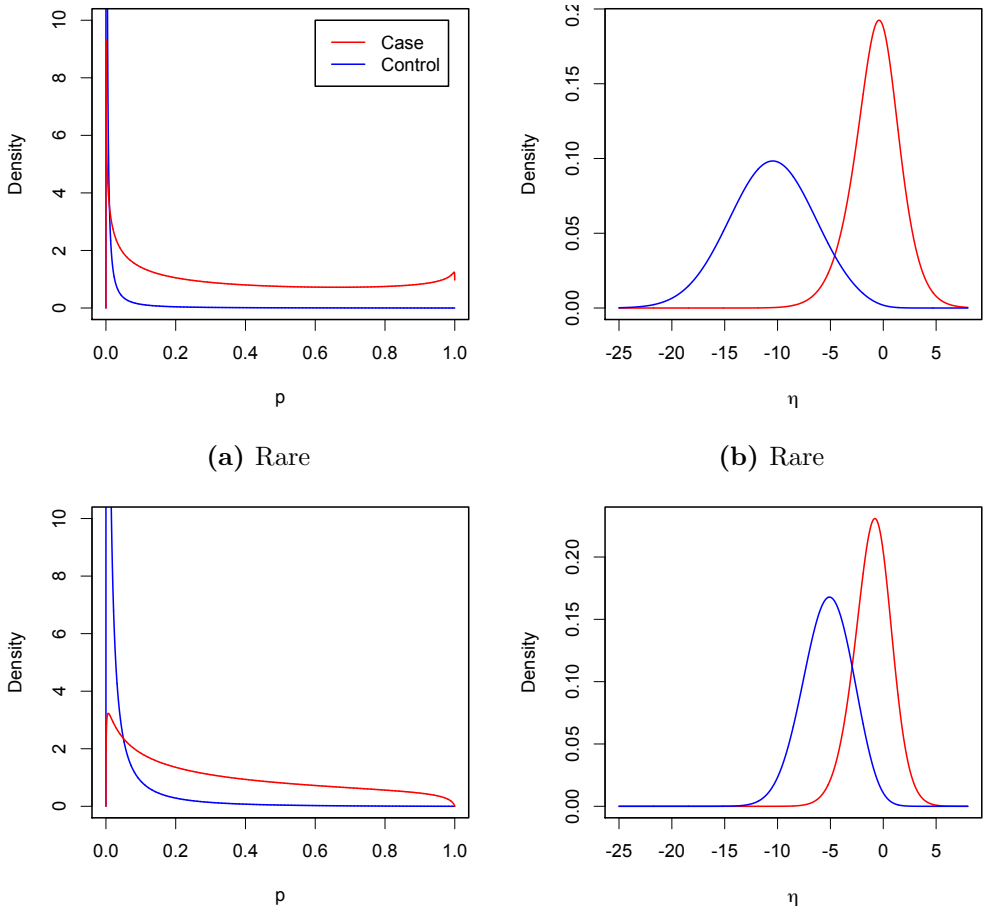
Examples of the distributions of η and p in cases and controls are shown in Figure 2.6.

2.5.4 Relationship to the multiplicative odds ratio model

Odds ratios are widely used to quantify differences between groups, and to make probabilistic predictions for individuals given group membership (see discussion in Morgan and Teachman (1988) for example). Odds ratios are the most widely used summary statistic in medical studies (Bland and Altman, 2000), mostly due to their utility in meta-analyses, though they are not without their detractors (Sackett et al., 1996). In genetics, the odds ratio has become the dominant method for summarising disease associations, largely due to its connection with logistic regression.

Given an exposure $a \in \{0, 1\}$, and an outcome $d \in \{0, 1\}$, we can define the probability conditional on exposure status $a = i$ as $p_i = P(d = 1|a = i)$. The odds ratio for exposure a is then defined as

$$r_a = \frac{p_1}{1 - p_1} \frac{1 - p_0}{p_0}. \tag{2.110}$$



(a) Rare **(b)** Rare

(c) Common **(d)** Common

Figure 2.6: The case and control distributions of probability p and risk score η for a rare disease ($K = 0.01, \lambda_s = 9$) and a common disease ($K = 0.05$, $\lambda_s = 3$)

Odds ratios in genetics

Throughout this thesis, I will refer to the effect size of a genetic association in terms of the odds ratios r_{het} and r_{hom}, where

$$r_{het} = \frac{p_1(1-p_0)}{p_0(1-p_1} \tag{2.111}$$

$$r_{hom} = \frac{p_2(1-p_0)}{p_0(1-p_2}, \tag{2.112}$$

where $p_x = P(d = 1|g = x)$ are the disease probabilities conditional on risk allele count x. We will sometimes refer to the genotypic odds ratio $r = r_{het} = \sqrt{r_{hom}}$ (also called the additive odds ratio).

We can rearrange the odds ratio definitions to give expressions for the disease probabilities for non-wild type genotypes in terms of the wild-type disease probability

$$p_1 = \frac{p_0 r_{het}}{1 - p_0 + p_0 r_{het}} \tag{2.113}$$

$$p_2 = \frac{p_0 r_{hom}}{1 - p_0 + p_0 r_{hom}}. \tag{2.114}$$

Given a prevalence K we can get the value of p_0 by solving the equation

$$p_0(1-f)^2 + p_1 2f(1-f) + p_2 f^2 = K, \tag{2.115}$$

which can be solved analytically (but messily), or numerically (counterintuitively, the numeric method is likely to be more accurate (Nievergelt, 2003)).

A common analytic approximation to calculate odds ratios is to normalize the odds ratios such that their population mean is equal to 1

$$\frac{(1-f)^2}{\bar{r}} + \frac{2f(1-f)r_{het}}{\bar{r}} + \frac{f^2 r_{hom}}{\bar{r}} = 1, \tag{2.116}$$

i.e.

$$\bar{r} = (1-f)^2 + 2f(1-f)r_{het} + f^2 r_{hom}. \tag{2.117}$$

We can then set

$$\hat{r}_0 = \frac{1}{\bar{r}} \tag{2.118}$$
$$\hat{r}_1 = \frac{r_{het}}{\bar{r}} \tag{2.119}$$
$$\hat{r}_2 = \frac{r_{hom}}{\bar{r}}, \tag{2.120}$$
$$\tag{2.121}$$

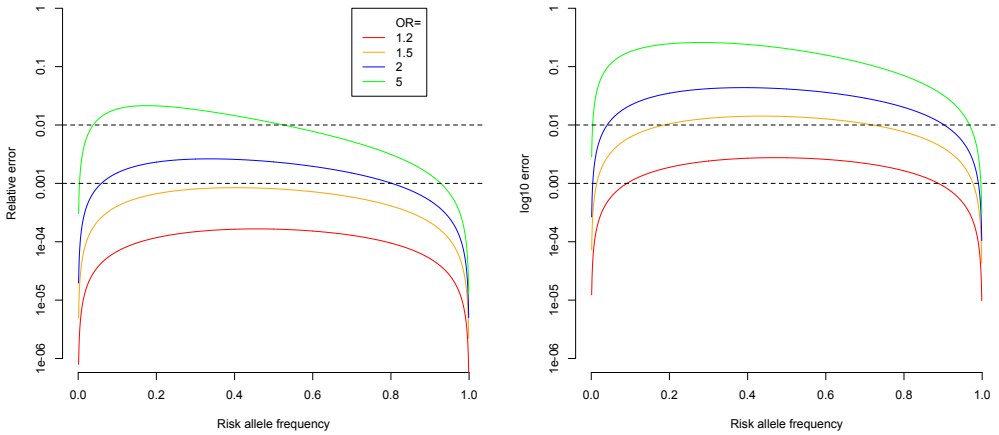
or, given a genotypic odds ratio, $\hat{r_x} = \frac{r^x}{\bar{r}}$.

We can then set the disease probabilities using these normalised odds ratios as

$$p_x = \frac{1}{1 + \frac{1-K}{\hat{r}_x K}}. \tag{2.122}$$
$$\tag{2.123}$$

This is the method used by, for instance, the genetic testing company 23andMe (Macpherson et al., 2007).

The accuracy of this approximation varies depending on the prevalence of the disease in question, and the size of the odds ratio (Figure 2.7). For a rare disease ($K = 0.01$) the approximation is accurate to within 1% for all realistic odds ratios and frequencies (and accurate to within 0.1% or less for $OR < 1.5$). For a more common disease ($K = 0.2$) the approximation is only accurate to within 1% for lower odds ratios ($OR < 1.5$). However, for odds ratios typically found within GWAS (generally $OR < 1.3$) the approximation



(a) K = 0.01 **(b)** K = 0.2

Figure 2.7: The accuracy of the odds ratio normalisation approach to genetic risk prediction, for rare and common diseases, as a function of odds ratio and risk allele frequency.

holds across prevalence and allele frequencies.

Combining independent odds ratios

Suppose we have two exposures a and b, with $p_{ij} = P(d = 1|a = i, b = j)$. A reasonable definition for these two exposures having an independent effect is if the odds ratio r_a does not depend on the value of b, and vice versa, i.e.

$$\begin{aligned} r_a &= \frac{p_{10}}{1-p_{10}} \frac{1-p_{00}}{p_{00}} \\ &= \frac{p_{11}}{1-p_{11}} \frac{1-p_{01}}{p_{01}}, \end{aligned} \tag{2.124}$$

and

$$\begin{aligned} r_b &= \frac{p_{01}}{1-p_{01}}\frac{1-p_{00}}{p_{00}} \\ &= \frac{p_{11}}{1-p_{11}}\frac{1-p_{10}}{p_{10}}. \end{aligned} \tag{2.125}$$

We can then calculate the joint odds ratio for both exposures, r_{ab}, as

$$\begin{aligned} r_{ab} &= \frac{p_{11}}{1-p_{11}}\frac{1-p_{00}}{p_{00}} \\ &= \left(\frac{p_{11}}{1-p_{11}}\frac{1-p_{01}}{p_{01}}\right)\left(\frac{p_{01}}{1-p_{01}}\frac{1-p_{00}}{p_{00}}\right) \\ &= r_a r_b, \end{aligned} \tag{2.126}$$

i.e. to combine independent odds ratios, multiply them together. Note that this justifies the genotypic odds ratio $r^2 = r_{hom} = r_{het}^2$, as it represents both alleles acting independently at a single locus.

We can generalise this to make a combined odds ratio given genotypes $\vec{x} = \{x_l\}$ across n loci with odds ratios $\vec{r} = \{r_l\}$

$$r_{\vec{x}} = \prod_{l=1}^{n} r_l^{x_l}. \tag{2.127}$$

The disease probability is thus given as

$$\begin{aligned} p_{\vec{x}} &= \frac{r_{\vec{x}} p_0}{1 - p_0 + r_{\vec{x}} p_0} \\ &= \frac{1}{1 + \frac{1}{r_{\vec{x}}} \frac{1-p_0}{p_0}} \\ &= \frac{1}{1 + \exp(-\eta)}, \end{aligned} \tag{2.128}$$

where

$$\begin{aligned} \eta &= log\left(\frac{p_0}{1-p_0} r_{\vec{x}}\right) \\ &= log\left(\frac{p_0}{1-p_0}\right) + log\left(r_{\vec{x}}\right) \text{Fixed brackets} \\ &= log\left(\frac{p_0}{1-p_0}\right) + \sum_{l=1}^{n} x_l log(r_l). \end{aligned} \tag{2.129}$$

Again, by the central limit theorem η tends towards a normal distribution with parameters

$$\mu = log\left(\frac{p_0}{1-p_0}\right) + \sum_{l=1}^{n} 2 f_l log(r_l) \tag{2.130}$$

$$\sigma^2 = \sum_{l=1}^{n} 2 f_l (1 - f_l) log(r_l)^2 \tag{2.131}$$

Thus the logit risk model is asymptotically equivalent to the assumption that odds ratios act independently.

2.5.5 Relationship to logistic regression

The logistic function, $g = (1+\exp(-\eta))^{-1}$, has been used since the 19th century as a description of population growth given limited resources (Verhulst, 1838), and in the early 20th century was found to accurately model many physiochemical responses (Reed and Berkson, 1929). It was first used as a regression model by Berkson (1944), who introduced it as an alternative to probit regression (and also introduced the name "logit"). Berkson later laid down in some detail the theoretical and empirical arguments underlying logit and probit link functions (Berkson, 1951).

In the last few decades the logit link has succeeded the probit link as the dominant form of regression model for binary outcomes(Cramer, 2003). It is very widely used in medical literature (though often imperfectly (Bagley et al., 2001)), and is the dominant method for performing genome-wide association studies under the presence of confounding factors, particularly with the rise of principal component methods to control population stratification (Price et al., 2006).

The logistic regression model has the form

$$p = (1 + \exp(-\eta))^{-1} \text{ where} \tag{2.132}$$

$$\eta = \beta_0 + \sum_i x_i \beta_i. \tag{2.133}$$

This is equivalent to equations 2.128 and 2.129 with parameters $\beta_0 = \frac{p_0}{1-p_0}$ and $\beta_i = log(r_i)$. We can thus see that, given an arbitrarily large number of predictors, the logistic regression model is approximated by the logit-normal risk model. This also provides us with a method of fitting the logit risk model from genetic data, using the results of logistic regression.

2.6 Comparing models of risk

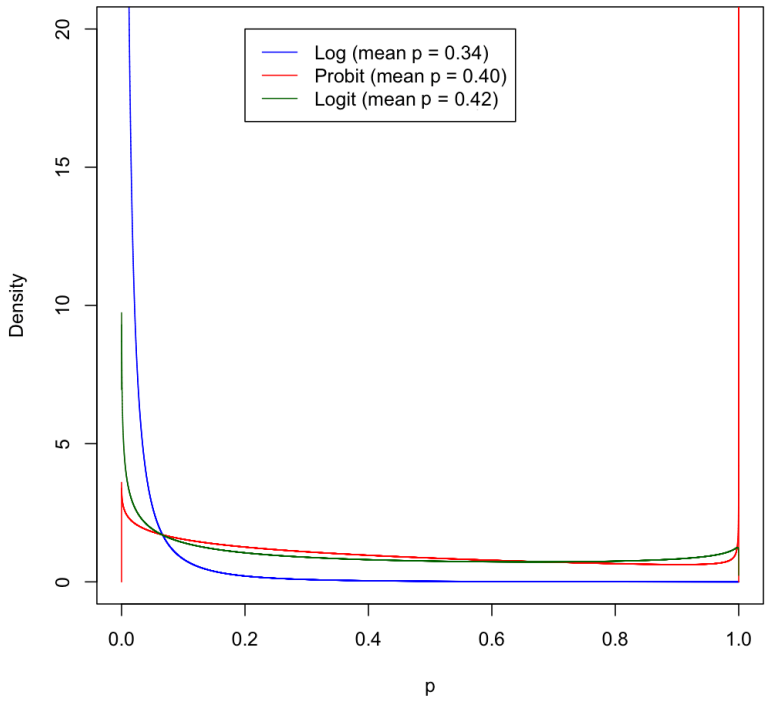
In the previous sections I outlined three continuous models of genetic risk and noted the different assumption that underlie them. In this section I will examine the ways in which these models differ in their predictions about the distribution of genetic risk in the population.

I will look at the predicted distribution of disease probability in cases across different models, and look in more detail at the differences between the logit and probit models. I will then consider the predicted relative recurrence risks and predicted ROC curves for the different models.

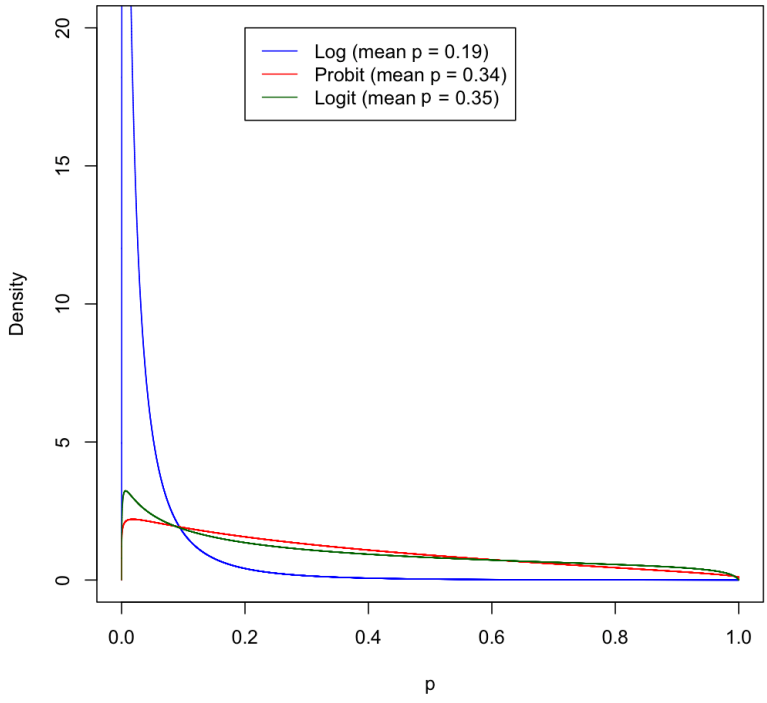
2.6.1 Comparing disease probability distributions in cases

Figure 2.8 shows the distribution of p in affected individuals under the three different models. In both cases, the log model produces a smaller mean p and a left-shifted distribution relative to the log and probit models. Additionally, in both scenarios the logit and probit models give relatively similar distributions, with approximately the same mean value of p. Disregarding a sharp peak near $p = 1$ for the probit model in 2.8a, the logit model seems to show slightly more density towards the ends of the distribution, and the probit model shows more density towards the middle.

In these comparisons, the log model stands out as clearly underestimating both the degree of enrichment of genetic risk in cases, predicting very few cases to have a high risk compared to the other two models. On the scale that we have examined, however, the logit and probit models appear similar, and it is difficult to infer the significance of these deviations. We can look at the differences between these two models in more detail by producing values of p given the probit model, and projecting them onto logit space using



(a) Rare disease



(b) Common disease

Figure 2.8: The distribution of genetic disease probabilities in randomly selected cases under the three different risk models, for a relatively rare, highly heritable disease ($K = 0.01$, $\lambda_s = 9$), and a more common, mildly heritable disease ($K = 0.05$, $\lambda_s = 3$). The legend gives the mean value of p in cases.

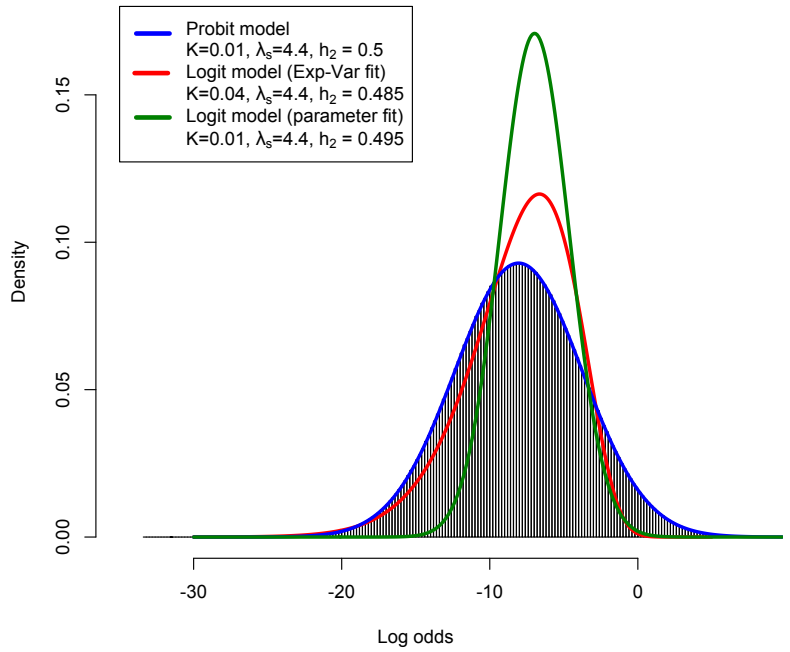


Figure 2.9: Different logistic approximations to a probit distribution. The exact distribution of the logit score under the probit model for $K = 0.01$ and $h^2 = 0.5$ is shown in blue (with bars representing a histogram of samples from the distribution). The red line shows a logistic normal fitted to have the same the mean and variance as the probit model, and the green line shows a logistic normal fitted to have the same K and λ_s values as the probit model.

$$p = \Phi(\eta_{probit}) = (1 + \exp(-\eta_{logit}))^{-1} \quad (2.134)$$

$$\eta_{logit} = log\left(\frac{\Phi(\eta_{probit})}{1 - \Phi(\eta_{probit})}\right). \quad (2.135)$$

Figure 2.9 shows this projection for a probit model with $h^2 = 0.5$ and $K = 0.01$ (bars and blue line). The red and green lines show two logit models: the red line showing the logit model with the same mean and variance on the logit scale as the probit model, and the green line showing the logit model with the same K and λ_s values as the probit model.

We can see that no logit model accurately models the projected probit distribution, due to the high kurtosis of the projection. A model with the same mean and variance, while having similar values of λ_s and h^2, predicts too high a prevalence. The model that has the same K and λ_s also has a similar h^2, but follows a very different curve with a much smaller variance.

This highlights clearly the ambiguity involved in comparing models or results parameterised on these difference scales. Furthermore, we can see that a logit model designed to closely mimic the probit model's risk distribution produces divergent parameters. Despite their superficial similarity, these models cannot be viewed as approximations to each other.

2.6.2 Comparing relative recurrence risk

None of the above distributions reflect any quantities that can be observed in the population. One long measured and studied property in the genetics of disease is the increase in disease risk in relatives of affected individuals, estimations of which are often used to draw conclusions about the genetic architecture of the disease (Compston and Coles, 2008; Sawcer, 2009; Brown et al., 2000).

As we saw in equation 2.58, under the log risk model

$$\lambda_r = \exp(\sigma\rho). \tag{2.136}$$

Substituting $\sigma = 2log(\lambda_s)$ gives

$$\begin{aligned} \lambda_r &= \exp(2log(\lambda_s)\rho) \\ &= \lambda_s^{2\rho}. \end{aligned} \tag{2.137}$$

This means that given the log risk model (and thus also given multiplicative relative risk), the recurrence ratio in relatives λ_r falls off with the logarithm of the coefficient of relatedness ρ. Deviation from this log-linear relationship is often interpreted as evidence of genetic non-additivity (Brown et al., 2000). However, deviations from this relationship could also be evidence that a different model is at play.

Figure 2.10a shows the fall-off in λ_r as a function of ρ for the three models (all with $K = 0.05$ and $\lambda_s = 3$). All models give very similar predictions, though there are slight differences between the models (Figure 2.10b). This includes up to a 6% increase in the risk ratio for probit and logit relative to the log model for highly related individuals ($\rho > 0.5$, including identical twins and siblings of consanguineous parents), and a corresponding decrease in risk for more distance relatives (peaking at a 3% difference at $\rho = 0.25$, or avuncular relationships).

These differences are on the limit of what can be detected in family studies: for 80% power to detect a 3% deviation from $\lambda_s = 3$ at $p < 0.05$ would require over 38 000 avuncular pairs. In addition, even if the log risk model could be rejected, we would not be able to say whether this difference was due to a different additive model applying, or merely a non-additive model. In theory measurements for a large range of different relative types could resolve this question, but in practice an even larger number of relatives would be required. In short, there is no plausible family study that could distinguish

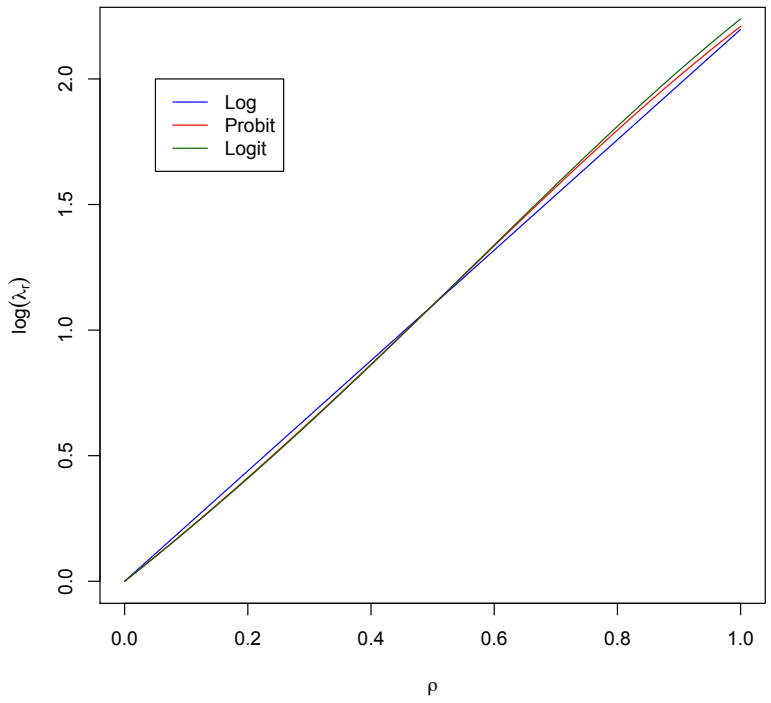
between these three models of genetic risk.

2.6.3 Comparing ROC curves for risk prediction

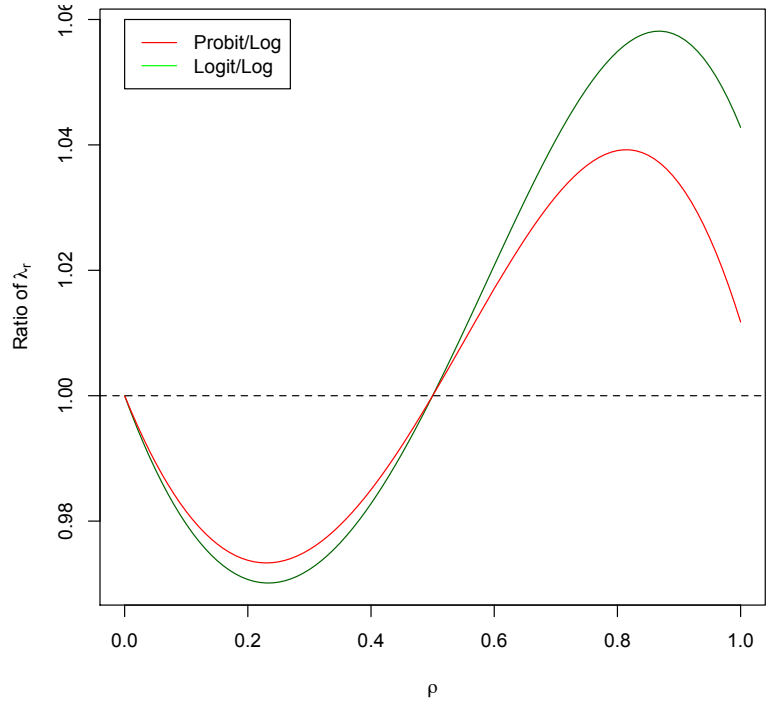
Many authors have attempted to make predictions about how useful genetic risk prediction could be if we managed to account for the total load of genetic risk predicted to exist by family studies. However, the results have been in many cases divergent, even when authors apply their methods to the same datasets. Some authors draw the conclusion that genetic risk prediction is unlikely to ever be of high utility (Clayton, 2009), while others conclude that genetic risk prediction could be of great use (Wray et al., 2010). I discussed the general question of how and when genetic risk prediction could be useful in the introduction, but here I will focus more specifically on how the model used can change your conclusions about the utility of genetic risk prediction.

Figure 2.11 shows the predicted ROC curves for diseases with a prevalence of $K = 1/200$ and $K = 1/20$, and a sibling relative risk of $\lambda_S = 9$ and $\lambda_s = 3$ for the three models. For the rarer disease all the models give divergent answers, with the probit model giving an AUC of 0.98, a logit model an AUC of 0.96, and the log model an AUC of 0.89. For the common disease, the logit and probit models agree on an AUC of 0.93, though with a different sensitivity-specificity trade-off, and the log model gives a much lower AUC of 0.84.

The low predictive accuracies for the log model are probably due to the problems mentioned in Section 2.3.5, and I will disregard these values. It therefore seems like a plausible maximum AUC for rare diseases likely lies between 0.96 and 0.98, and common diseases around 0.93, as predicted by the logit and probit models. However, the significant variability, both in the AUC values and in the shape of the ROC curves, highlights the degree



(a) Rare disease



(b) Common disease

Figure 2.10: a) The log relative risk ($log(\lambda_r)$) under the three models as a function of the coefficient of relatedness ρ, Parameters are $K = 0.05$, $\lambda_s = 3$ b) The ratio of probit and logit λ_r values to log λ_r values, as a function of ρ.

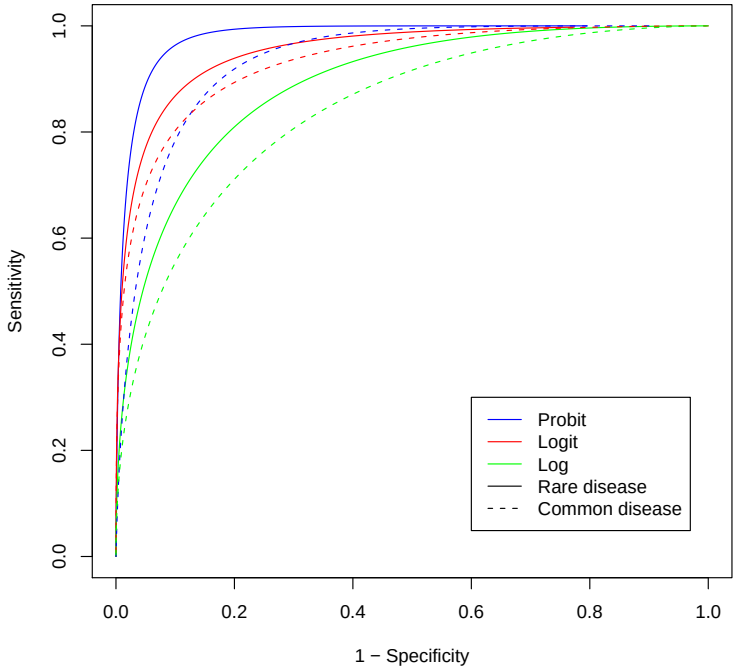


Figure 2.11: The ROC curves for the log, logit and probit models of disease risk for a rare disease with a prevalence $K = 1/200$ and sibling relative risk of $\lambda_S = 9$, and a common disease with $K = 1/20$ and $\lambda_s = 3$, given that all genetic risk has been explained. The corresponding AUCs are 0.89, 0.96 and 0.98 respectively for the rare disease, and 0.84, 0.93 and 0.93 for the common disease.

to which forecasts of the future utility of genetic risk prediction are model specific.

2.7 Conclusion

2.7.1 Summary of models

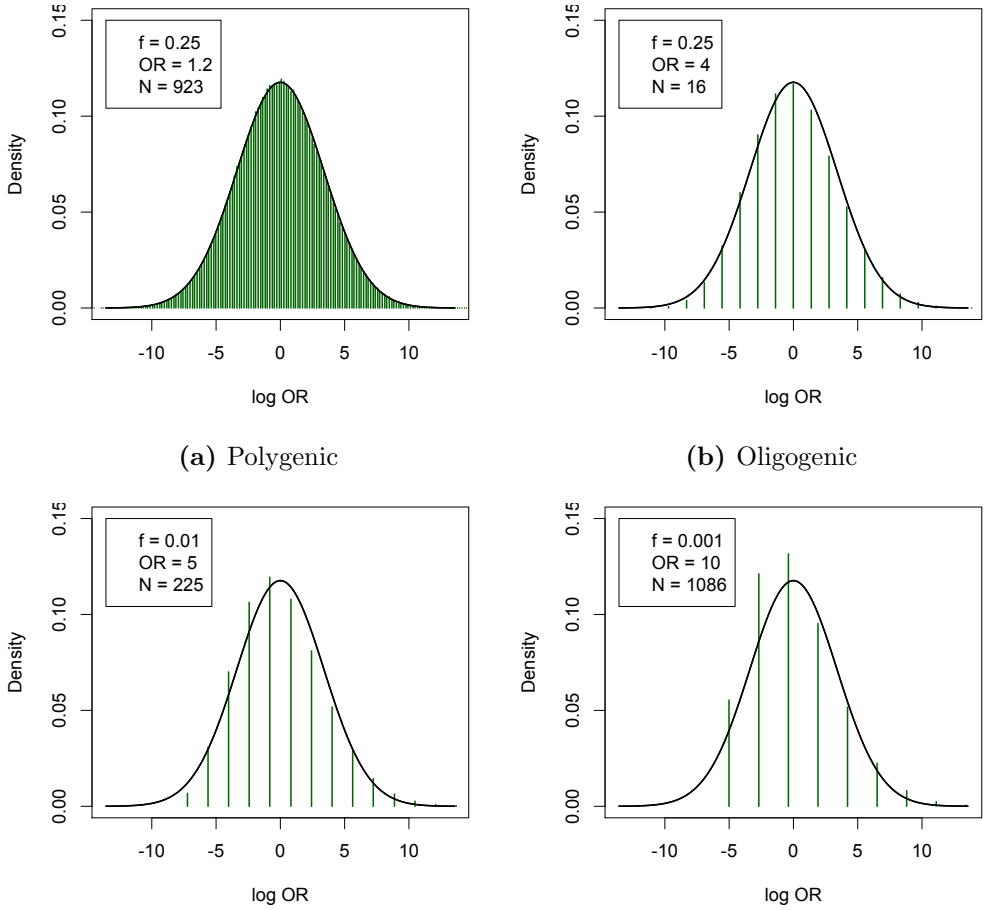
As we have seen, the three models that we have examined can each be seen as the natural result of the assumptions made in one or more major statistical method. We can summarise the three models, and their corresponding methods, using the following table:

Model	Link function	Equivalent models/methods
Log risk	$p = \exp(\eta)$	Risch model, multiplicative relative risk
Probit risk	$p = \Phi(\eta)$	Liability threshold model, latent variable model, probit regression
Logit risk	$p = (1 + \exp(-\eta))^{-1}$	Multiplicative odds ratios, logistic regression

We have seen that these models differ in their predictions about the distribution of risk in populations. Some of these differences are minor (they all have a similar relationship between coefficient of relatedness and relative recurrence risk), but some are large (they give divergent predictions about the maximum utility of genetic risk prediction).

2.7.2 Limitations of this approach

An important caveat is that the analyses of these three models above are all built on two major assumptions. The first is that the risk score η can be approximated by a normal distribution, and the second is that the risk score η is additive.

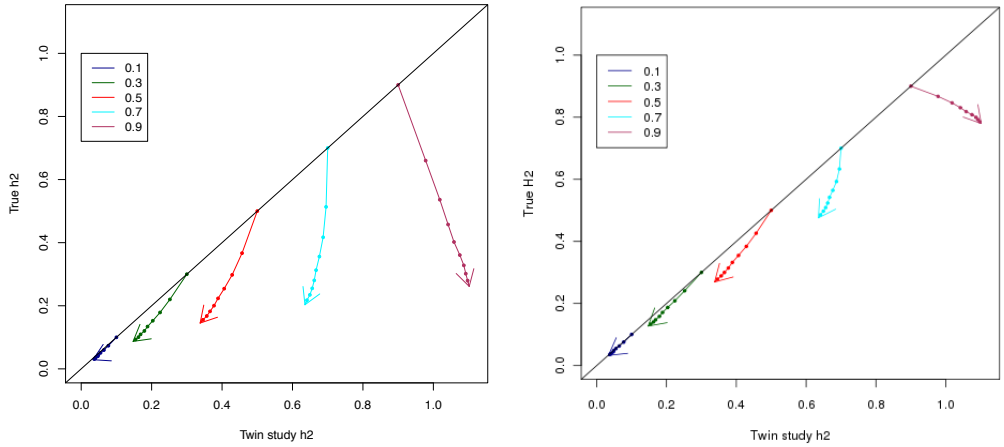


(a) Polygenic **(b)** Oligogenic

(c) Low-frequency **(d)** Rare

Figure 2.12: The closeness of fit to the normal distribution for variants with different frequencies and odds ratios. The black line represents the normal approximation, and the green bars are odds ratios sampled from the model. The number of variants N is chosen to have $\lambda_s = 3$.

Speaking to the first assumption, Figure 2.12 illustrates how well this approximation holds across different architectures, given the same value of λ_s. In fact, the normal approximation holds for almost all plausible genetic architectures; the approximation is very accurate for polygenic and oligogenic models, and is relatively accurate for low-frequency variants. The approximation becomes significantly less accurate for a disease driven purely by rare,



(a) Narrow sense heritability **(b)** Broad sense heritability

Figure 2.13: The affected of epistasis on heritability estimation from twin studies. The epistasis model used is the multiple threshold model of Zuk et al. (2012), in which the risk score is the minimum of N independent liability scales, each with a heritability h_p^2. The dots represent increasing N (starting with 1, increasing in the direction of the arrow), and the colours represent different values of h_p^2. The first panel shows the overestimation of the narrow-sense heritability, and the second shows the overestimation of the broad-sense heritability.

highly penetrant mutations.

As for the second, non-additivity can alter the models in two ways. Firstly, it can lead to non-normality in the risk score. However, as I mentioned in Section 2.2, it seems likely that most forms of pairwise interaction can be approximated as a normal distribution, and even risk scores based on more detailed forms of epistasis can be modelled as normal (see, for example, Zuk et al. (2012)). Secondly, as we saw for single-locus dominance in section 2.2.3, non-linearity can alter the correlation structure of risk scores in related individuals. Specifically, non-additivity reduces correlation such that $cor[\eta_i, \eta_j] < \rho$. This is turn can lead us to overestimate the heritability of the disease.

We use the model of Zuk et al. (2012) to explore this effect. Figure 2.13

examines how serious this effect will be on our estimation of heritability, and thus the results of our models. Zuk et al. (2012) showed that, under epistasis, the narrow sense heritability (i.e. the correlation in additive risk score) will be greatly overestimated by twin studies (as shown in Figure 2.13a). However, for our purposes we are more interested in the overestimation of the full heritability, which is what determines the univariate distribution of the probit score η. Figure 2.13b shows that this value is significantly less prone to overestimation than the narrow sense heritability, and is only seriously overestimated in cases where $H^2 > 0.8$.

2.7.3 Problems generated by model ambiguity

The use of methods with differing underlying models can itself create ambiguity in results. Suppose we have performed a genome-wide association study of a disease with $K = 0.05$, using logistic regression. We have identified 48 loci, each with an estimated odds ratio of 2 and a frequency of 50%. We wish to compare these results with data from of twin studies, which have found that the disease has a heritability of $h^2 = 0.8$, in order to say what proportion of genetic variance has been explained. There are three ways that we could answer this question

1. Fit the log-normal model from the data using equation 2.131, project the result onto the probit scale using Equation 2.134, calculate the variance and convert to h^2 using Equation 2.75. This gives $h^2 \approx 0.586$

2. Fit the log-normal model, calculate the value of λ_s, and use Equation 2.80 to calculate the corresponding h^2. This gives $h^2 \approx 0.634$

3. Perform probit regression on the original genetic data, and use equation 2.89 to calculate h^2. A simulation of this (generated under the logistic

model) gives $h^2 \approx 0.751$.

The technique used can alter the percentage of heritability explained from 74% to 93%. The smallest answer may lead people to invest further to discover the missing quarter of heritability, while the latter will likely lead to a conclusion that the trait is essentially solved. There is no correct answer, as the question we are asking is inherently problematic: the two results we are comparing were generated under different models. Which of the values is correct (if any) will depend on the true model underlying the genetic risk in the first place, which is unknown.